The time dimension in measurements of population health,
Lauer, Jeremy A (2009).
Thesis Erasmus MC, University Medical Centre Rotterdam,
with summary in English and Dutch.

Cover design: Services Concept, Geneva, Switzerland
Layout: Services Concept, Geneva, Switzerland
Printed by: Ridderprint, Ridderkerk, Netherlands

ISBN/EAN: 978-90-5335-185-7

The time dimension in measurements of population health

Thesis

Jeremy Addison Lauer

The Time Dimension in Measurements of Population Health

De tijdsdimensie in het meten van volksgezondheid

Thesis

to obtain the degree of Doctor from the
Erasmus University Rotterdam
by command of the
rector magnificus

Prof.dr. S.W.J. Lamberts

and in accordance with the decision of the Doctorate Board

The public defence shall be held on
Thursday 16 April 2009 at 11:00 o'clock

by

Jeremy Addison Lauer

born at Chapel Hill, North Carolina, United States of America

Doctoral Committee

***Promotor*:**

Prof.dr.ir. J.D.F. Habbema

Other members:

Prof.dr. E.M.E.H. Lesaffre
Prof.dr. M.G.M. Hunink
Prof.dr. J.W. Hargrove
Prof.dr. F.F.H. Rutten
Prof.dr.ir. F.J.C. Willekens

***Copromotor*:**

Dr. J.J. Barendregt

Table of contents

Chapter 1

Introduction

Introduction

What is population health and why is its measurement important?

As recently attested by the Millennium Declaration (United Nations, 2000), the health of populations is a concern for both governments and civil society: three of the eight Millennium Development Goals are defined in terms of health objectives. It is therefore reasonable to enquire what it means when we claim population health improves or declines. We should also know how to measure such changes. Since the Millennium Declaration does not answer these questions, we infer that concern for population health on the part of governments and civil society is not necessarily accompanied by clear concepts regarding its definition and measurement.

Nevertheless, substantial technical progress on population health measurement has been made recently (see, for example, Murray and Lopez, 1996). In addition, there is a substantial body of older work on life expectancy and related measures (for an overview, see Keyfitz and Caswell, 2005). Life expectancy, in addition to being one of the oldest population health measures, might also be the only one to have successfully established itself in popular and political conceptions. As we argue in the following, this very success conceals subtle technical points that have confounded, at least to some degree, further developments in population health measurement.

The large literature on the "burden of disease", and the widespread use of concepts such as the QALY (see, for example, Drummond et al., 1987), the DALY (Murray and Lopez, 1996), the HeaLY (Hyder et al., 1998) and the healthy life expectancy (Sullivan, 1971) testify to a sustained scientific interest in population health measurement. Notably, this interest has fuelled the development and distribution of a number of related software tools (e.g. Barendregt et al., 2003; Gunning-Schepers, 1989; Gunning-Schepers, 1999; Kruijshaar et al., 2002; Mooy and Gunning-Schepers, 2001; Naidoo et al., 1997; van Baal et al., 2008). However, progress in the technical dimensions of health measurement has not been matched by progress in political and popular concepts.

The central role of time and age

In addition to technical barriers, we propose that a lack of consensus regarding a central issue of population health measurement has limited the diffusion of modern concepts. We believe the principal disagreement centers around how to understand and represent the time and age dimensions of population health. Although human lifespan is limited in time, and age is the most obvious measure of its progress, disagreements about how to represent age and time in health measurement reflect divergent views. A corollary disagreement is whether data collected for the purpose of population health measurement should be time- or age-oriented.

The need for a model

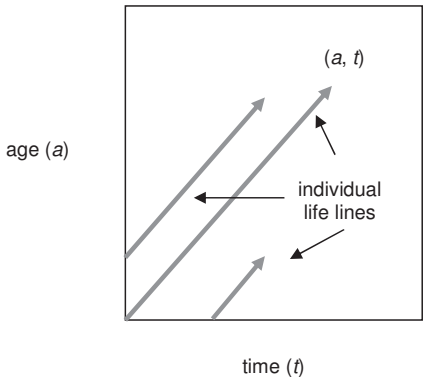
Measurement requires a model: to measure length, a model of length is necessary. Since population health is considerably harder to define than length, a model for its measurement is arguably even more necessary. By "model" we refer to a mathematical structure defining a measurement scale (e.g. a "ruler"). To be a valid model, the scale defined should permit the fundamental algebraic operations of addition, subtraction, multiplication and division (Krantz et al., 1971). The population health model implicit in the life table is technically and conceptually simple. As for the model of length, it may be simple enough even to escape notice as a model.

A model requires an interpretation. The interpretation makes explicit a correspondence between the structures and assumptions of the model and the reality it is intended to represent (the correspondence is usually expressed in terms of a "theory"). Since a simple population health model cannot support a detailed correspondence to complex reality, it requires more elaborate assumptions for its interpretation.

Terminology for population health measurement

We briefly review some terms related to population health measurement that are important for the research presented here. Because the terms implicitly express an orientation towards time and age, they define (or at least suggest) corresponding population health models. We discuss examples in the context of births and mortality.

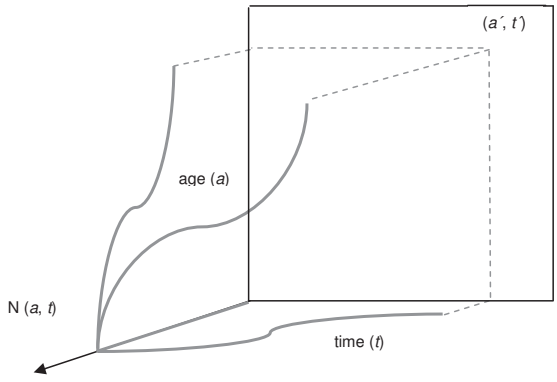
Figure 1: Lexis diagram showing individual life lines.



As noted in Keiding (2000), the years around 1870 witnessed an explosion, in German-speaking Europe, in the use of sophisticated graphical representations for population health measurement as well as in the underlying mathematical concepts. We begin by introducing the Lexis diagram, so called for its inventor W. Lexis (1875). Although there are different versions (Keiding, 2000), the Lexis diagram is fundamentally a two-dimensional graph with age and time dimensions (Figure 1).

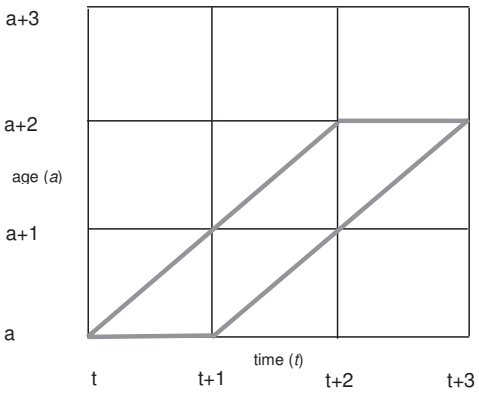
This simple diagram admits several interpretations, not all of them simple. For example, the life line of an individual can be represented as a ray of a certain length. Different individuals will show different rays, always parallel, but of differing lengths reflecting age at death and with different starting points reflecting year of birth. In brief, vital events like birth and death have a unique (*a*, *t*) coordinate in two dimensions.

Figure 2: Zeuner surface in three dimensions (defined by solid blue curves), showing for a population the number of individuals in the Lexis diagram.



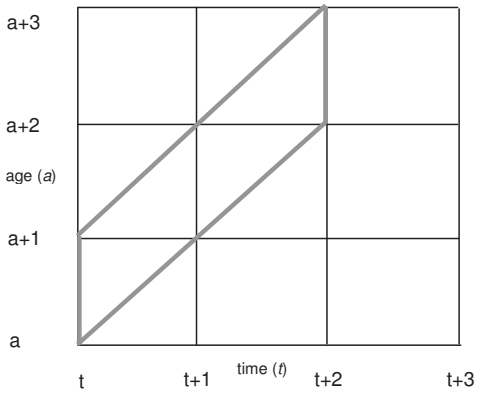
Now, if a third dimension (e.g. orthogonally out of the page) represents the number of individuals alive at coordinate (a, t), or some other related quantity of interest, this quantity defines a surface, called the Zeuner surface, after G. Zeuner (Keiding, 2000) (Figure 2). For simplicity, in the following we usually do not represent the surface, although it is important to understand that it is implicitly present.

Figure 3: The AC set of observations classified by age (age between a and $a + 2$) and cohort (born between $t - a$ and $t + l - a$).



The Lexis diagram also defines three distinct two-factor classification schemes. These schemes, present at the birth of population health measurement, are at the heart of current disagreements about time and age. Notice, for example, that population observations classified by age and birth *cohort* (i.e. year of birth) may not show the precise timing of vital events. Such a classification of observations defines a set in Lexis space that is similar to the parallelogram shown in Figure 3. Conversely, observations classified by cohort and *period* (an interval of time) may not indicate the precise age of the individuals experiencing the events. The set defined by such a classification of observations is similar to the parallelogram shown in Figure 4.

Figure 4: The PC set of observations classified by cohort (born between *t* − *a* and *t* + *1* − *a*) and period (years *t* to *t* + *2*).



Finally, the classification of observations by age and period will generally conceal the exact cohort of the individuals experiencing the events. The set corresponding to such a classification of observations is shown by the rectangle in Figure 5. Since the cohort is concealed, observations classified by age and period, such as those belonging to a set like that shown in Figure 5, can be called *synthetic*. The meaning of synthetic is that such observations combine (literally, "put together", but also, in practice, "confound") the health

Figure 5: The AP set of observations classified by age (age between *a* and *a* + *3*) and period (years t to *t* + *1*).

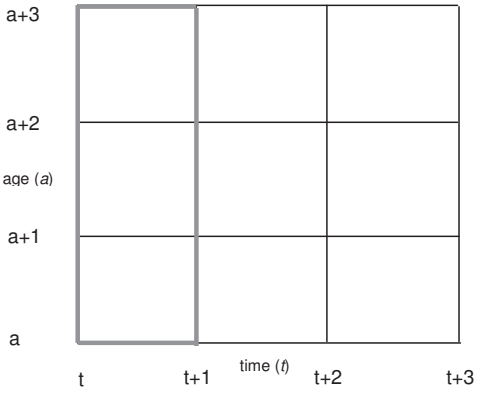
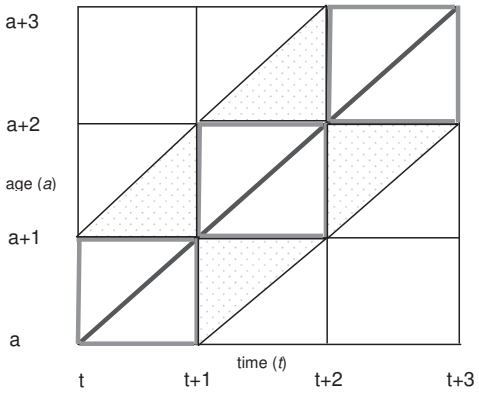


Figure 6: Synthetic cohort based on a classification by age and period. See text for explanation.



experience of more than one birth cohort. One can construct a synthetic cohort from the observations made of two or more "natural" cohorts. A set of observations corresponding to a synthetic cohort, which might be constructed from age–period classified data, is shown in Figure 6 by the diagonal series of grey squares. The synthetic cohort contains part of the vital experience of each of two natural cohorts: one above and the one below the bold diagonal line. Notice that observations of vital events occurring in the light grey-shaded regions, which belong to these two cohorts, are not included in the set corresponding to the synthetic cohort.

Cross-sectional population health measures refer to measures constructed from observations of vital events from a single time period, such as those corresponding to the set shown in Figure 5. Within the measurement interval, period is fixed (i.e. constant). Thus, cross-sectional measures are also called *period* measures. In contrast, age is often variable in period measures, in the sense that the observations of different age classes (cohorts) are referenced (see Figure 5). Cross-sectional measures typically define a population made up of several different cohorts (hence, a synthetic population). Notice that, in this particular terminology, "synthetic population" is identical to what in non-technical language is usually referred to simply as the "population". By showing how health experience varies by age in a given period, cross-sectional measures can be said to promote an "age-oriented" view of population health.

Perhaps the best-known example of a cross-sectional population health measure is the ordinary *period life expectancy*. Period life expectancy is a synthetic measure derived from observations of the vital experience of different birth cohorts. As such, unless special conditions pertain, period life expectancy cannot be said to be the life expectancy of any single member of the population.

In contrast, *longitudinal* population health measures are based on a time (or age) series of vital rates (see Figures 3, 4 and 6). Note that the time series may refer either to a single birth cohort (Figures 3 and 4) or to different ones (Figure 6). In either case, longitudinal series of measures have a *cohort interpretation*. Further, if we assume that the distribution of vital events is uniform in time, longitudinal observations in the cohort–age classification (Figure 3) can be easily transformed into the cohort–period classification (Figure 4). Since, in the latter classification, age is not explicit, longitudinal population health measures of the cohort–period type can be said to promote a "time-oriented" view of population health.

The longitudinal version of period life expectancy is *cohort life expectancy*. Cohort life expectancy is the life expectancy of a particular birth cohort (for example, that of individuals born in 1964). By definition, measuring cohort life expectancy requires information on the full mortality experience of the cohort, which in this example can only be obtained after the last person born in 1964 has died. It will likely be many years before it is possible to compile complete mortality data for this cohort. Thus, cohort life expectancy can be calculated only for cohorts that are no longer members of the population.

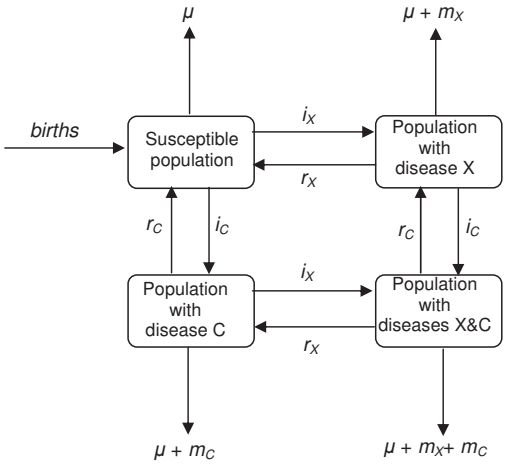
In this respect, cohort life expectancy is like period life expectancy: apparently, neither is an estimate of the life expectancy of any living member of the population. Thus, both measures conceal problems: either of measurement (cohort life expectancy) or of interpretation (both life expectancies). The naive interpretations of life expectancy that are in common use in popular and political discussions are therefore incorrect (since they do not admit a valid correspondence with what we understand about reality); other, potentially valid, interpretations require elaborate and potentially unrealistic assumptions.

As an example of the latter sort of interpretation is known as the *stationary population interpretation*. The interpretation derives its name from the fact that, when the population is "stationary", in the sense that the health experience in all age classes is constant with respect to time, the population's cross-sectional health measures are numerically identical to its longitudinal measures. The stationary population interpretation is an example of a set of assumptions about how a population model corresponds to what we understand about the reality it is supposed to measure.

Multi-state models

The preceding discussion has centered around the observation and modelling of vital events, and the principal focus has been on mortality. Although the history of population health measurement has mostly been concerned with mortality-based measures, in early days, Lexis and Zeuner already presented sophisticated multi-state descriptions of population processes, in which events for marriage and disability, as well as emigration and immigration, were represented (Keiding, 2000). In a single-state model, such as the ordinary life table, transitions are only to an *absorbing state*, such as death. For a model to be a multi-state model, it must represent at least one transition to a non-aborbing state. As in the early models of Zeuner and Lexis, non-absorbing transitions can be events such as marriage (and divorce), emigration (and immigration), or entry to (and exit from) various health conditions.

Figure 7: Multi-state representation of a population.



If transition rates are constant, the multi-state model defines a system of ordinary differential equations, with solutions in terms of standard exponential formulas. If transition rates are stratified by age, the multi-state model defines a multi-state life table. Figure 7 depicts a graphical representation of a multi-state, constant-rate model with four non-absorbing states and one absorbing state. Transition rates are represented as arrows, where i is incidence, r remission, m disease-specific mortality and μ other ("background") mortality. Note that transition to an absorbing state (i.e. death) is equivalent to exiting the model, and births are entries. Since multi-state life tables allow for calculation of quantities such as the average population time spent with disability or disease, they are of considerable interest for population health measurement, and there is a long tradition of their use (see, for example, Keyfitz and Caswell, 2005).

Types of population health measures

The research presented here mainly uses examples of a class of population health measures known as *summary measures*, since they summarize (usually in a single number) the impact of fatal and non-fatal health events on population health (Murray et al,. 2002). Several subfamilies of summary measures have been described. *Health gaps* (e.g. DALYs) compare actual population health to some reference standard. *Health expectancies* (essentially multi-state generalizations of the life expectancy; Robine et al., 1993) summarize population health by means of a weighting function applied to non-fatal outcomes (non-absorbing states). Such a function sometimes purports to quantify the health of the modelled states relative to a state of perfect health (Murray et al., 2002). If the weighting function is simply a tabulation of numerical values by health state, the values are called *disability weights*, or *health-state valuations* (Murray et al., 2002).

Static versus dynamic models

In conventional usage, *dynamic* models (such as longitudinal models) measure change in an outcome variable with respect to time. Similarly, *static* models are understood to be models without a time dimension. Since, as we have noted, time is fixed in cross-sectional models, such models are often understood to be synonymous with static models.

In the context of the age, period and cohort classification, the question arises whether a population health model should not be capable of accounting for all three tabulation variables. A class of models designed with this criterion in mind has been referred to as *dynamic life tables* (Murray et al., 2002). A dynamic life table is usually described as a period life table augmented by an independent time dimension. More directly, in the spirit of Zeuner and Lexis, a dynamic life table is a model of the function $N(a, t)$ indicated in Figure 2.

In terms of that figure, a dynamic life table could be represented by a two-factor (e.g. age/time) table of values for N. Typical summary measures are derived from a dynamic life table are integrals (sums) and weighted integrals (weighted sums) of the function N, for example $\int_a \int_t N(a, t)$. Appropriate limits of integration need to be defined. When N is a tabular function (e.g. because it is computed numerically), the integral is simply a sum $\sum_a \sum_t N(a, t)$, again, taken over appropriate intervals of age and time. In the multi-state case, N itself is a (possibly weighted) sum of other populations, such as S, X, C, and $X\&C$ (cf. Figure 7): $N = S + X + C + X\&C$.

Another common term for a dynamic life table is simply a *population model*. In the ecology literature, however, such a model is usually referred to as a *meta-population model*, meaning a model containing more than one population (Caswell, 2001). When additional precision is required, or to emphasize the presence of an independent time dimension, we shall frequently use the term "longitudinal population model". When we wish, on the other hand, to emphasize the continuity of dynamic population models with classical life table methods, we say "longitudinal population life tables".

A longitudinal population model can be cast in either discrete or continuous time, and the choice of one or the other is often a question of convenience. For example, since demographic data are often reported as proportions (in terms of some initial at-risk population), it may be natural to cast the corresponding population model in discrete time, where the observed proportions are directly used as model parameters. Particularly in animal or plant biology, where mortality and reproduction are limited to specific periods, there may be clear subject-matter reasons to prefer discrete-time over continuous-time models (Caswell, 2001). Nevertheless, at least mathematically, a discrete-time model

can always be recast as a continuous-time one, and vice versa. Sometimes, as here, tradition plays a role; consistent with life-table methods from population demography, the longitudinal population health models discussed in this thesis are all of the continuous-time type.

Requirements of longitudinal population models

As longitudinal population models represent a greater level of complexity than the ordinary period or cohort life tables, it is useful to consider whether they make additional demands in terms of either the technical sophistication required of the user or the kinds of data required for input parameters. A corollary question is whether any additional data requirements can be satisfied by the sort of data that are typically available in practice.

In reply to the first question, the technical demands made by longitudinal population models are modest compared with those of other technical fields, for example engineering or actuarial science. In addition, computational barriers that in the past made a simple one-dimensional single-state life table an attractive option for population health measurement can no longer be considered relevant. Finally, the would-be user of longitudinal population models has access to a body of publications, some of which are referred to here, that explain in detail their methods and concepts. Therefore, technical complexity cannot be considered an important barrier to the use of longitudinal population life tables.

On the other hand, to supply the parameters required by a longitudinal population model, especially the multi-state kind, can be more challenging. To start, population data corresponding to the age classes used in the model are required. Although it is possible in principle to use age classes of any width, it is common practice as well as conceptually simple to define one-year age classes as the unit of analysis (Caswell, 2001). So data on population size and population vital rates are required in terms of one-year age groups. Since most applications typically need to distinguish outcomes by sex, sex-specific data on population size and vital rates by year of age comprise a minimum set of required data. For the multi-state population model, basic information about different disease states of interest is also required, usually in the form of an incidence, prevalence, mortality (IPM) model (Kruijshaar et al., 2002).

Finally, to define the distribution of the total population across subpopulations at the start of the simulation, information on the prevalence of each modelled condition is required.

Some of these data may be routinely estimated and made available in disease registries, or published by health agencies. Some of the additional data required, especially those required for disease-specific outcomes, will either be difficult to obtain or possibly difficult even to observe. For example, disease mortality and remission rates frequently differ according to disease severity, or according to the presence or absence of comorbidities. In addition, there may be several distinct disease stages, each with its own particular transition rates. Such complexities require additional data. If all the required data are not available, it may be possible to obtain the necessary parameter estimates by employing methods that more fully make use of the data that are available. For example, as in the IPM model, it may be possible to assume the existence of systematic relations between disease parameters; these relations can then be used to estimate missing parameters. In a few cases, merely changing denominators appropriately (e.g. to transform population-level mortality estimates to estimates for subpopulations) can yield estimates of parameters that are not directly observed. Yet other parameter estimates may be obtained by unit analysis of observed parameters, such as by obtaining the recovery rate as the inverse of disease duration.

However, all such indirect approaches require assumptions. In particular, the unobserved parameters need to be representable in a model whose form is simple enough to facilitate indirect estimation (e.g. the IPM model) yet complex enough to provide a convincing representation of the relevant process. The assumptions required for inference (characterized mainly by simplicity) may be in conflict with the needs of validity and realism. Moreover, estimates resulting from indirect methods depend on the validity of the assumptions used, such as the degree of simplification or approximation involved.

Given the significant demands for data that are made by population health modelling, especially in a longitudinal and multi-state framework, the availability of the required data can become a central concern. Indeed, embarking on a project of multi-state longitudinal population health modelling can involve a constant struggle to meet the ambitious demands

of the model for parameters using only the scarce supply of data available in practice. Despite the required assumptions and limitations of such methods, the demand for parameters can typically be met only by the creative use of indirect methods of estimation,.

Stationary population interpretation

One of the more powerful methods of indirect estimation is the stationary population interpretation referred to above. As we describe subsequently, applying this interpretation allows us to identify a number of "implicit" dynamic methods. These methods mitigate, to a certain extent, the need for data inherent in multi-state longitudinal population modeling. In addition to the life table, a number of other simple models admit the stationary population interpretation. Unlike life tables, however, these other models are not purpose-built population health models but are rather descriptive and statistical models in wide use in epidemiology, econometrics, ecology and related fields. In the body of the thesis we discuss the implications of some of these models, and show examples of how they can be employed for population health measurement in the absence of complete information about the required model parameters.

Perhaps the most fundamental implicit dynamic model is the logistic model, which includes the exponential model as a special case. Taken together, these two models have extremely wide scope of application in the study of growth processes and epidemics. They can in turn be embedded in a richer set of related models such as generalized logistic growth models (Tsoularis and Wallace, 2002) and models for the so-called GS-distributions (generalized S-distributions; Muiño et al., 2006). With the exception of the models for GS-distributions, which have a more statistical formulation, all of these models are fundamentally longitudinal models of dynamic processes; thus, under the assumption of population equilibrium in the model covariates, the dynamics are reproduced in cross section.

Research questions addressed by this thesis

The research presented in this thesis can be described in terms of a number of related research questions, some general and others more specific. The individual chapters in the body of the thesis are designed to shed light on

these questions, which thus form the unifying themes of the research. By *general* questions we have in mind questions that can be considered of general importance for scientific work in the field of population health measurement. Another feature is that such questions typically do not permit completely determined and discrete answers. The general questions that we examine can be formulated as follows:

G1. Should we care about the longitudinal (i.e. time) dimension of population health? To motivate this question, we have argued above that population time dynamics are an essential concern for the measurement of population health.

G2. Do longitudinal measures allow us to answer questions about population health better than other measures? To address this question we focus, in particular, on an important set of questions of the form "What is the impact on population health of a given course of action?", i.e. we examine how longitudinal measures can be applied in answering questions about quantifying risks to health as well as in quantifying the benefits of interventions designed to improve health.

G3. Are standard models from other scientific fields relevant to questions of population health measurement? In relation to this question, we apply some of these models to addressing questions of population health measurement.

Inherent in the approach adopted here is an effort to illustrate general questions through related subsidiary questions which, as they are more specific in their formulation, offer the possibility of more tightly circumscribed answers. The subsidiary questions we address here concern, on the one hand, the practicability and feasibility of the methods inspired by the general questions. On the other hand, they also concern the nature and usefulness of the results that can be obtained by application of these methods. Two of the specific questions addressed here are as follows:

S1. Is it possible to develop an easy-to-use, reasonably generic, longitudinal model that can be used for population health measurement with generally available epidemiologic and demographic data?

S2. Is it feasibile to apply a generic, longitudinal model to quantify population health risks, or to quantify the benefits of interventions to improve health, and what sort of insights can be obtained through its application?

These specific questions (S1, S2) are related to the first two general questions (G1, G2) identified above. The third general question (G3) is explored through the following specific questions:

S3. Does the logistic family of models offer an approach relevant to longitudinal population health measurement?

S4. Is it feasibile to apply logistic models to quantify real population health risks and what sort of insights can be obtained through their application?

In tabular form, the specific questions relate to the general questions as follows:

$$\begin{array}{lcl} (S1, S2) & \rightarrow & (G1, G2) \\ (S3, S4) & \rightarrow & (G3). \end{array}$$

Contents of thesis and relation to research questions

With the exception of Chapter 8, the chapters in the body of the thesis relate to one of the specific questions. The present introductory chapter, Chapter 1, motivates the research and introduces the research questions. The final chapter, Chapter 9, discusses answers to the research questions, highlights areas for additional research and proposes conclusions and recommendations. We also attach two technical appendices that discuss in detail two of the areas for further research. A description of the chapters in the body of the thesis, and their relation to the research questions, follows.

Chapter 2 looks at the technical aspects of constructing a generic multi-state longitudinal population model. In that chapter, the technical feasibility of such a model is demonstrated, which is relevant for question S1. In addition, some of the issues related to the availability and use of data are also touched upon; however, these questions are also addressed in the subsequent two chapters.

Chapter 3 describes a particular application of the model described in Chapter 2 to the cost-effectiveness analysis of interventions to lower blood pressure

and cholesterol. In much the same spirit, Chapter 4 describes an application of the model to the cost-effectiveness analysis of malaria control strategies. Both chapters respond primarily to question S2, which concerns the feasibility of applying a generic longitudinal model to quantify real population health interventions and risks; these chapters also illustrate questions concerning the use and availability of the sort of data required for a multi-state longitudinal population model.

Chapter 5 examines how the logistic model, one of the most widely used models in epidemiology, ecology and related fields, can be derived from basic processes that are relevant for standard applications of the model. An examination of these processes shows their importance for population health measurement (question S3).

Chapter 6 and Chapter 7 discuss an application of the logistic model to questions of population health measurement; specifically, these chapters demonstrate the feasibility of using the logistic model (relevant to S4) to infer missing epidemiological parameters relevant to the measurement of exposure to a risk factor. Chapter 6 addresses the technical aspects of prevalence estimation; Chapter 7 uses the resulting exposure estimates to estimate the impact of the risk factor on population health using the framework of population attributable fraction. As such, Chapter 7 is the only chapter that does not use an explicitly dynamic (longitudinal) method to estimate population health. Although this was originally done for comparability with similar estimates published in the *World health report 2002* (WHO, 2002), it also allows for a discussion of the advantages and disadvantages of cross-sectional methods.

Chapters 5, 6 and 7 thus all relate to the third general question, regarding the use of models from other fields in population health measurement. In adopting an econometric time-series model for panel data to analyse the determinants of health care utilization in a group of developed countries, Chapter 8 also relates to this question. That chapter provides a detailed opportunity to compare directly longitudinal and cross-sectional estimates of the same phenomenon. Chapter 9 concludes.

References

1. Barendregt JJ, Van Oortmarssen GJ, Vos T, Murray CJ. A generic model for the assessment of disease epidemiology: the computational basis of DisMod II. *Population health metrics*, 1(1):4, 2003.

2. Caswell H. *Matrix population models: construction, analysis, and interpretation.* Sunderland, Massachusetts, Sinauer Associates, 2001.

3. Drummond MF, Stoddart GL, Torrance GW. *Methods for the economic evaluation of health care programmes*, Oxford, Oxford University Press, 1987: 112-148.

4. Gunning-Schepers L. The health benefits of prevention: a simulation approach. *Health policy*, 12(1-2):1-255, 1989.

5. Gunning-Schepers LJ. Models: instruments for evidence based policy. *Journal of epidemiology and community health*, 53(5):263, 1999.

6. Hyder AA, Rotllant G, Morrow RH. Measuring the burden of disease: healthy life-years. *American journal of public health*, 88(2):196-202, 1998.

7. Keiding N. *Graphical representations in mortality measurement: Knapp, Zeuner, Becker, Lexis*. Copenhagen, Denmark, 2000 [Research report 00/8, Department of Biostatistics, University of Copenhagen].

8. Keyfitz N, Caswell H. *Applied mathematical demography*, 3rd ed. Berlin, Springer, 2005.

9. Krantz DH, Luce D, Suppes P, Tversky A. *Foundations of measurement, volume I: additive and polynomial representations*. New York, Academic Press, 1971

10. Kruijshaar ME, Barendregt JJ, Hoeymans N. The use of models in the estimation of disease epidemiology. *Bulletin of the World Health Organization*, 80(8):622-628, 2002.

11. Lexis W. *Einleitung in die Theorie der Bevölkerungsstatistik* [Introduction to the theory of population statistics]. Karl J Trübner, Strassburg, 1875.

12. Mooy JM, Gunning-Schepers LJ. Computer-assisted health impact assessment for intersectoral health policy. *Health policy*, 57(3):169-177, 2001.

13. Muiño JM, Voit EO, Sorribas A. GS-distributions: a new family of distributions for continuous unimodal variables. *Computational statistics and data analysis*, 50(10): 2769-2798, 2006.

14. Murray CJL, Lopez AD (eds.). *Global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*. Cambridge, Massachusetts, Harvard University Press, 1996.

15. Murray CJL, Salomon JA, Mathers CD, Lopez AD (eds.). *Summary measures of population health: concepts, ethics, measurement and applications*. Geneva, World Health Organization, 2002.

16. Naidoo B, Thorogood M, McPherson K, Gunning-Schepers LJ. Modelling the effects of increased physical activity on coronary heart disease in England and Wales. Journal of epidemiology and community health, 51(2):144-150, 1997.

17. Robine, J-M, Mathers C, Bone MR, Romieu I (eds.). *Calculation of health expectancies: harmonization, consensus achieved and future perspectives*. Paris, Institut National de la Santé et de la Recherche Médicale/John Libbey Eurotext, 1993

18. Sullivan DF. A single index of mortality and morbidity. *HSMHA health reports*, 86:347-354, 1971.

19. Tsoularis A, Wallace J. Analysis of logistic growth models. *Mathematical biosciences* 179:21–55, 2002.

20. United Nations. United Nations Millennium Declaration. A/RES/55/2. Adopted by the United Nations General Assembly 18 September 2000 (available at http://www.un.org/millennium/).

21. van Baal PHM, Polder JJ, de Wit GA, Hoogenveen RT, Feenstra TL, et al. (2008) Lifetime medical costs of obesity: prevention no cure for increasing health expenditure. *PLoS medicine* 5(2): e29. doi:10.1371/journal.pmed.0050029.

22. WHO. *World health report 2002: reducing risks, promoting healthy life*. Geneva, World Health Organization, 2002.

List of publications

Chapter 2: (S1) Lauer JA, Roehrich K, Wirth H, Charette C, Gribble S, Murray CJL. PopMod: a longitudinal population model with two interacting disease states. *Cost effectiveness and resource allocation*, 2003, 1:6.

Chapter 3: (S2) Murray CJL, Lauer JA, Hutubessy RCW, Niessen L, Tomijima N, Rodgers A, Lawes CMM, Evans DB. Effectiveness and costs of interventions to lower systolic blood pressure and cholesterol: a global and regional analysis on reduction of cardiovascular-disease risk. *Lancet*, 2003, 361:717-725.

Chapter 4: (S2) Morel CM, Lauer JA, Evans DB. Achieving the millennium development goals for health: Cost effectiveness analysis of strategies to combat malaria in developing countries. *BMJ*, 2005, 331:1299-1302.

Chapter 5: (S3) Lauer JA, Greenland S. Deriving logistic distributions from basic processes (in submission).

Chapter 6: (S4) Lauer JA, Betrán AP, Victora CG, de Onís M, Barros AJD. Breastfeeding patterns and exposure to suboptimal breastfeeding among children in developing countries: review and analysis of nationally representative surveys. *BMC Medicine*, 2004, 2:26.

Chapter 7: (S4) Lauer JA, Betrán AP, Barros AJD, de Onís M. Deaths and years of life lost due to suboptimal breastfeeding among children in the developing world: a global ecological risk assessment. *Public health nutritio*n, 2006, 9(6):673-685.

Chapter 8: (G3) Lauer JA, Betrán AP, Merialdi M, Wojdyla D. Trends in caesarean section in developed countries: supply, demand and opportunities for control (in submission).

Chapter 2

PopMod: a longitudinal population model with two interacting disease states

Based on: Lauer JA, Roehrich K, Wirth H, Charette C, Gribble S, Murray CJL. PopMod: a longitudinal population model with two interacting disease states. *Cost effectiveness and resource allocation*, 2003, 1:6.

Abstract

This article provides a description of the population model PopMod, which is designed to simulate the health and mortality experience of an arbitrary population subjected to two interacting disease conditions as well as all other "background" causes of death and disability. Among population models with a longitudinal dimension, PopMod is unique in modelling two interacting disease conditions; among the life-table family of population models, PopMod is unique in not assuming statistical independence of the diseases of interest, as well as in modelling age and time independently. Like other multi-state models, however, PopMod takes account of "competing risk" among diseases and causes of death.

PopMod represents a new level of complexity among both generic population models and the family of multi-state life tables. While one of its intended uses is to describe the time evolution of population health for standard demographic purposes (e.g. estimates of healthy life expectancy), another prominent aim is to provide a standard measure of effectiveness for intervention and cost-effectiveness analysis. PopMod, and a set of related standard approaches to disease modelling and cost-effectiveness analysis, will facilitate disease modelling and cost-effectiveness analysis in diverse settings and help make results more comparable.

Introduction

Historical background and analytical context

Measuring population health has been inseparable from modelling population health for at least three hundred years. The first accurate empirically based life table – a population model, albeit a simple one – was constructed by Edmund Halley in 1693 for the population of Breslau, Germany.[1] However, the 1662 life table of John Graunt, while less rigorously based on empirical mortality data, represented a reasonably good approximation of life expectancy at birth in the seventeenth century.[2] Indeed, because of Graunt's strong *a priori* assumptions about age-specific mortality, his life table could be said to represent the first population model. Recently, multi-state life tables, which explicitly model several population transitions, have become a common tool for demographers, health economists and others, and a considerable body of theory has been developed for their use and interpretation.[3-5] Despite the substantial complexity of existing multi-state models, a recent publication has highlighted the advantages of so-called "dynamic life tables", in which age and time would be modelled independently.[6]

Mathematical and computational constraints are no longer serious obstacles to solving complex modelling problems, although the empirical data required for complex models are. In particular, multi-state models present data requirements that can rapidly exceed empirical knowledge about real-world parameter values, and in many cases, the input parameters for such models are therefore subject to uncertainty. Nevertheless, even with substantial uncertainty, such models can provide robust answers to interesting questions. Indeed, the work of John Graunt demonstrates the practical value of results obtained with even purely hypothetical parameter values.

PopMod, one of the standard tools of the WHO-CHOICE programme (http://www.who.int/evidence/cea), is the first published example of a multi-state dynamic life table. Like other multi-state models, PopMod takes account of "competing risk" among diseases, causes of death and possible interventions. However, PopMod represents a new level of complexity among both generic population models and the family of multi-state life tables.

Among population models with a longitudinal dimension, PopMod is unique in modelling two distinct and possibly interacting disease conditions; among the life-table family of population models, PopMod is unique in not assuming statistical independence of the diseases of interest, as well as in modelling age and time independently.

While one of PopMod's intended uses is to describe the time evolution of population health for standard demographic purposes (e.g. estimates of healthy life expectancy), another prominent aim is to provide a standard measure of effectiveness for intervention and cost-effectiveness analysis. PopMod, and a related set of standard approaches to disease modelling and cost-effectiveness analysis used in the WHO-CHOICE programme, facilitate disease modelling and cost-effectiveness analysis in diverse settings and help make results more comparable. However, the implications of a tool such as PopMod for intervention analysis and cost-effectiveness analysis is a relatively new area with little published scholarship. Most published cost-effectiveness analysis has not taken a population approach to measuring effectiveness, and when studies have done so they have generally adopted a steady-state population metric.[7] Relatively little published research has noted the biases of conventional approaches when used for resource allocation.[8]

Despite similarities in some of the mathematical techniques,[9] this paper does not consider transmissible disease modelling.

Basic description of the model

PopMod simulates the evolution in time of an arbitrary population subject to births, deaths and two distinct disease conditions. The model population is segregated into male and female subpopulations, in turn segmented into age groups of one-year span. The model population is truncated at 101 years of age. The population in the first group is increased by births, and all groups are depleted by deaths. Each age group is further subdivided into four distinct states representing disease status. The four states comprise the two groups with the individual disease conditions, a group with the combined condition and a group with neither of the conditions. The states are denominated for convenience X, C, XC and S, respectively. The state entirely determines health status and disease and mortality risk for its members. For example, X could

be ischaemic heart disease, C cerebrovascular disease, XC the joint condition and S the absence of X or C.

State members undergo transitions from one group to another, they are born, they get sick and recover, and they die. The four groups are collectively referred to as the total population T, births are represented as the special state B, and deaths as the special state D. A diagram for the first age group is shown in Figure 1 (notation used is explained in the section *Describing states, populations and transitions between states*). In the diagram, states are represented as boxes and flows are depicted as arrows. Basic output consists of the size of the population age-sex groups reported at yearly intervals. From this output further information is derived. Estimates of the severity of the states X, C, XC and S are required for full reporting of results, which include standard life-table measures as well as a variety of other summary measures of population health.

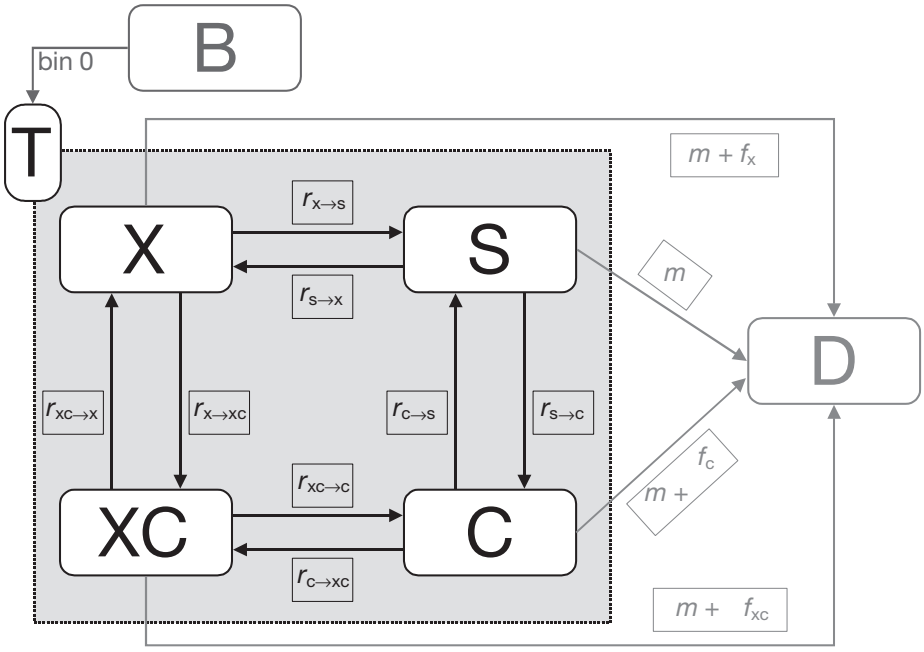
There now follows a more technical description of the model and its components, broken down into the following sections: describing states, populations and transitions between states; disease interactions; modelling mechanics; and output interpretation. The article concludes with a discussion of the relation of PopMod to other modelling strategies, plus a consideration of the implications, advantages and limitations of the approach.

Describing states, populations and transitions between states

Describing states and populations

In the full population model depicted in Figure 1, six age-and-sex specific states (X, C, XC, S, B and D) are distinguished. However, births B and deaths D are special states in the sense that they only feed into or absorb from other states (while the states X, C, XC and S both feed into and absorb from other states). Special states are not treated systematically in the following, which focuses on the "reduced form" of the model consisting of the states X, C, XC, and S.

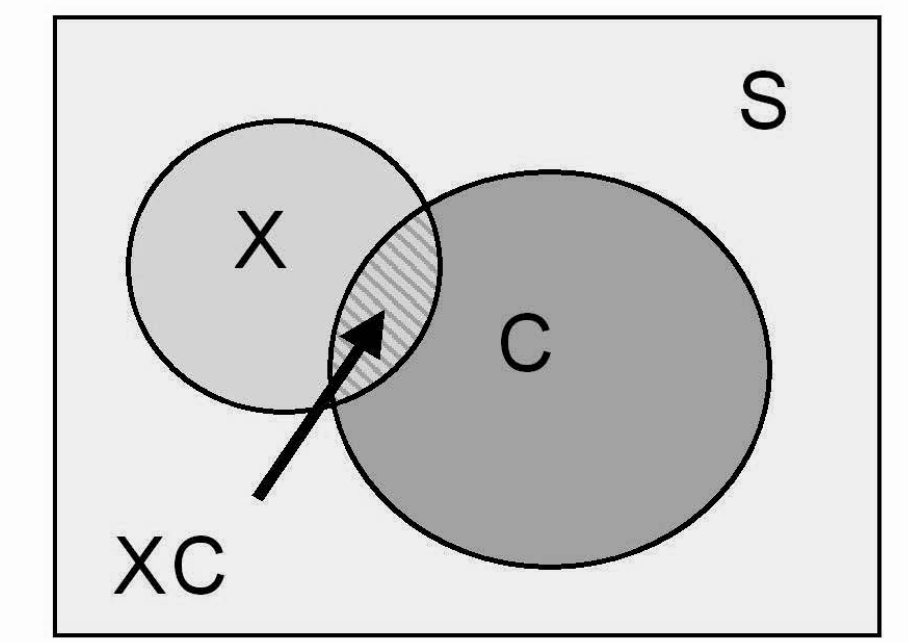
Figure 1: The differential equations model.



States are not distinguished from their members; thus, "X" is used to mean alternatively "disease X" or "the population group with disease X", according to context. The second meaning is equivalent to the prevalence count for the population group.

For the differential equation system, states/groups are always denoted in the strict sense: "X" means "state X only" or "the population group with only X". However, in deriving input parameters (described more fully below in the section *Disease interactions*) from observed populations, it is convenient to describe groups in a way that allows for the possibility of "overlap". For example in Figure 2, the area "X" might be understood to mean either "the population group with X including those members with C as well" (i.e. the entire circle X) or the "the population group with only X" (i.e. the circle minus the region overlapping with circle C).

Figure 2: A schematic for describing observed populations.



Since these two valid meanings imply different uses of notation, the following conventions are adopted:

- The differential equations expressions X, C, XC and S refer only to disjoint states (or groups).
- The logical operator "~"means "not", thus "~X" is the state "not X" (or "the group without X").
- The logical expressions denoted in the left-hand column of Table 1 have the meaning and alternative description indicated in the two right-hand columns.

Table 1: Alternative ways to describe populations.

Logical expression	Meaning	Differential equations expression
~ X~ C	Population group with neither X nor C	S
X~ C	Population group with X but not C, i.e. with X only	X
~ XC	Population group with C but not X, i.e. with C only	C
~ X	Population group without X	S + C
~ C	Population group without C	S + X
X	Total population group with X	X + XC
C	Total population group with C	C + XC
S	Susceptible population	S
XC	Population with both X and C	XC
T	Total population	T

Table 2: Probability of finding members of population groups in PopMod.

Symbol	Description
π_X	Probability of finding a member of T that is a member of X with random selection.
π_C	Probability of finding a member of T that is a member of C with random selection.
π_{XC}	Probability of finding a member of T that is a member of XC with random selection.

Prevalence rates (p) describe populations (i.e. prevalence counts) as a proportion of the total, for example:

$$p_X = n_X/n_T,\ p_C = n_C/n_T,\ p_{XC} = n_{XC}/n_T,\ p_S = n_S/n_T. \quad (1)$$

Here, prevalence is presented in terms of the disjoint populations X, C and XC, and the notation from the right-hand column of Table 1 is used. In the section *Disease interactions*, we discuss the case of overlapping populations.

A prevalence rate is always interpretable as a probability, but a probability is not always interpretable as a prevalence. The lower-case Greek letter pi (π) is

used throughout this article to denote probability. Probabilities can be used to describe populations as noted in Table 2.

Describing transitions between states

In the differential equation system, transitions (i.e. flows) between population groups are modelled as instantaneous rates, represented in Figure 1 as labelled arrows. Instantaneous rates are frequently called hazard rates, a usage generally adopted here (demographers tend to refer to instantaneous rates as "hazards" or as "forces" - e.g. force of mortality - although epidemiologists commonly use the term "rate" with the same meaning). A transition hazard is labelled here h, frequently with subscript arrows denoting the specific state transition.

In PopMod terminology, the transitions X→D, C→D and XC→D are partitioned into two parts, one of which is the cause-specific fatality hazard f due to the condition X, C or XC, and the other which is the non-specific death hazard (due to all other causes), called background mortality m:

$$\begin{aligned} h_{X\to D} &= f_X + m && \text{(2a)}\\ h_{C\to D} &= f_C + m && \text{(2b)}\\ h_{XC\to D} &= f_{XC} + m && \text{(2c)}\\ h_{S\to D} &= m. && \text{(2d)} \end{aligned} \qquad (2)$$

PopMod consequently allows for up to twelve exogeneous hazard parameters (Table 3).

Transition hazards

A time-varying transition hazard is denoted $h(t)$. The hazard expresses the proportion of the at-risk population, P ($\mathrm{d}P/P$), experiencing a transition event (i.e. exiting the population) during an infinitesimal time $\mathrm{d}t$:

$$h(t) = -(1/P) \cdot \mathrm{d}P/\mathrm{d}t. \qquad (3)$$

"Instantaneous rate" means the transition rate obtaining during the infinitesimal interval $\mathrm{d}t$, that is, during the instant in time t. If an instantaneous rate does not vary, or its small fluctuations are immaterial to the analysis, PopMod parameters can be interpreted as average hazards without prejudice to the model assumptions.

Table 3: Transition hazards in the population model.

Hazard	Description	State transition
$h_{S\to X}$	incidence hazard	S→X
$h_{X\to S}$	remission hazard	X→S
$h_{S\to C}$	incidence hazard	S→C
$h_{C\to S}$	remission hazard	C→S
$h_{X\to D}$	case fatality hazard	X→D
$h_{C\to D}$	case fatality hazard	C→D
$h_{XC\to D}$	case fatality hazard	XC→D
$h_{T\to D}$	background mortality hazard	T→D
$h_{C\to XC}$	incidence hazard	C→XC
$h_{XC\to C}$	remission hazard	XC→C
$h_{X\to XC}$	incidence hazard	X→XC
$h_{XC\to X}$	remission hazard	XC→X

Average hazards can be approximated by counting events ΔP during a period Δt and dividing by the population time at risk. If for practical purposes the instantaneous rate does not change within the time span, the approximate average hazard can be used as an estimate for the underlying instantaneous rate:

$$-(1/P) \cdot \mathrm{d}P/\mathrm{d}t \approx -\int \mathrm{d}P \,/\, \int P \mathrm{d}t \approx -\Delta P \,/\, (P \cdot \Delta t), \qquad (4)$$

where $\Delta P = \int \mathrm{d}P$ is the cumulative number of events occurring during the interval Δt, and $\int P \mathrm{d}t \approx P \cdot \Delta t$ is the corresponding population time at risk. Time at risk is approximated by multiplying the mid-interval population (P) by the length of the interval Δt.

For example, if ten deaths due to disease X ($\Delta P = 10$) occur in a population with approximately one million years of time at risk ($P \cdot \Delta t = 1{,}000{,}000$), an approximation of the instantaneous rate $h_{\mathrm{X}\to\mathrm{D}}(t)$ is given by:

$$h_{\mathrm{X}\to\mathrm{D}}(t) \approx \Delta P \,/\, P \cdot \Delta t = 10 \,/\, 1{,}000{,}000 = 0.00001\ . \qquad (5)$$

Note that while eq. (3) and eq. (4) are equivalent in the limit where $\Delta t \to 0$, the approximation in eq. (4) will result in large errors when rates are high. This is discussed in the section *Proportions and hazard rates*, and an alternative formula for deducing average hazard is proposed in eq. (9).

The quantity in eq. (4) has units "deaths per year at risk", and is often called a "cause-specific mortality hazard". For the same population and deaths, but restricting attention to the group with disease X (where, for example, $P \cdot \Delta t = 10{,}000$) the calculated hazard will be larger:

$$h_{\mathrm{X}\to\mathrm{D}}(t) \approx \Delta P \,/\, P \cdot \Delta t = 10 \,/\, 10{,}000 = 0.001. \qquad (6)$$

The quantity in eq. (6) has the same units as that in eq. (5), but is a "case fatality hazard". Note that the same transition events (e.g. "dying of disease X") can be used to define different hazard rates depending on which population group is considered.

Proportions and hazard rates

Integration by parts of eq. (3) shows that the proportion of the population experiencing the transition in the time interval Δt (i.e. the "incident proportion") is given by:

$$\frac{\Delta P}{P(t_0)} = 1 - \exp \int_{t_0}^{t_0+\Delta t} -h(t)\ \mathrm{d}\,t \tag{7}$$

If the hazard is constant, that is, if $h(t) = h\,(t_0) = h$, $\int \mathrm{d}\,t = \Delta t$ and the integral collapses. The incident proportion is then written:

$$\frac{\Delta P}{P} = 1 - \mathrm{e}^{-h\cdot\Delta t} \tag{8}$$

.

The incident proportion can always be interpreted as the average probability that an individual in the population will experience the transition event during the interval (e.g. for mortality, this probability can be written $\pi_{\mathrm{P}\to\mathrm{D}} = \Delta P/P$). The qualification "average" is dropped if individuals in P are homogeneous with respect to transition risk during the interval.

Even if the hazard is not constant, eq. (8) can be rearranged to give an alternative (exact) formula for calculating the equivalent constant hazard h yielding ΔP transitions in the interval Δt:

$$h = -\ln\left(1 - \frac{\Delta P}{P}\right) / \Delta t \tag{9}$$

.

However, if the true hazard is constant during the interval, the "equivalent constant hazard" equals the "average hazard" and the "instantaneous rate". The same identity applies when fluctuations in the underlying hazard are of no practical importance. PopMod requires the assumption that hazards are constant within the unit of its standard reporting interval, defined by convention as one year.

Note that series expansion of $\exp\{-h \cdot \Delta t\}$ or $\ln\{1 - \Delta P/P\}$ shows that, for values of $h \cdot \Delta t < 1$ and $\Delta P / P < 1$, the equivalent constant hazard is well approximated by the time-normalized incident proportion, and vice versa, as in eq. (4):

$$h \approx \frac{1}{\Delta t}\frac{\Delta P}{P}. \tag{10}$$

Case-fatality hazards

Case-fatality hazards f_X, f_C, and f_{XC} are defined with respect to the specific populations X, C and XC, respectively:

$$f_X = -\frac{1}{\Delta t}\ln\left(1 - \frac{\Delta X}{X}\right), \tag{11}$$

$$f_C = -\frac{1}{\Delta t}\ln\left(1 - \frac{\Delta C}{C}\right), \tag{12}$$

$$f_{XC} = -\frac{1}{\Delta t}\ln\left(1 - \frac{\Delta XC}{XC}\right). \tag{13}$$

Mortality hazards

Mortality hazards are defined with respect to the entire population, where cause-specific mortality hazards are conditional on cause of death:

$$m_{tot} = -\frac{1}{\Delta t}\ln\left(1 - \frac{\Delta T}{T}\right), \tag{14}$$

$$m_X = -\frac{1}{\Delta t}\ln\left(1 - \frac{\Delta T}{T}\right)_{T \xrightarrow{X} D}, \tag{15}$$

$$m_C = -\frac{1}{\Delta t}\ln\left(1 - \frac{\Delta T}{T}\right)_{T \xrightarrow{C} D}. \tag{16}$$

The background mortality rate m is defined as the instantaneous rate of deaths due to causes other than X or C.

Disease interactions

PopMod is typically used to simulate the evolution of a population subjected to two disease conditions, where health status, health risk and mortality risk are conditional on disease state. Health status, health risk and mortality risk are plausibly conditional on disease state when the two primary disease conditions X and C interact. Such interactions can be analysed from various perspectives, for example, common risk factors, common treatments, common prognosis; however, the primary perspective adopted here for the pupose of analysis is that of "common prognosis", by which is meant that the two conditions mutually influence prevalence, incidence, remission and mortality risk.

A previously cited example was that of ischaemic heart disease (X) and cerebrovascular disease (C): it is well known that individuals with either heart disease or stroke history have lower health status and higher mortality risk than individuals with neither of these conditions, and that individuals with heart disease are at increased risk for stroke and vice versa.

Furthermore, individuals with history of both heart disease and stroke (XC) are known to have higher mortality risk and lower health status than either individuals with only one of the disease histories or those with neither. However, in this example as in many others, information about the joint condition (heart disease and stroke) is scarce relative to information about the two individual conditions (heart disease or stroke). The obvious reason for this is that the population group with the joint condition is smaller in size and has a lower life expectancy, reducing opportunities for data collection.

The presimulation problem

One of PopMod's guiding principles, therefore, is that while an analyst has access to information about basic parameter values for the conditions X and C (i.e. prevalence rates and incidence, remission and either case-fatality or cause-specific mortality hazards), the same is not generally true for the joint condition XC. Thus, more or less by construction, the modelling situation is one in which data for the joint condition are scarce or unavailable, and must consequently be derived from data known for the individual conditions.

An important implication is that the data available for the individual conditions (X and C) will be reported in terms of overlapping populations. Where specifically noted, therefore, the notation in the left-hand column of Table 1 (Logical expressions) is used in the following, with the particular implication that "X", for example, means "the population group with X including those members with C as well" (i.e. "X + XC" in differential equations terminology).

Once parameter values for the joint condition are determined, the minimum set of parameters required for population simulation are known. This parameter-value problem – referred to here as the presimulation problem, since its solution must precede population simulation per se – can be divided into two principal parts: one concerning the prevalence rates defining the intial conditions (stocks) of the differential equations system, and the other the transition hazards defining its flows. These stocks and flows together make up the initial scenario of the population model. A cross-sectional approach is adopted in which deriving these two kinds of parameters values for the initial scenario are treated as separate problems.

The analytics of these derivations largely depend on which of a range of possible assumptions is made about the interactions of the two principal conditions. The simplest possible assumption is essentially an assumption of non-interaction (statistical independence). Since an understanding of the non-interacting case is an essential starting point for more complex interactions, it is discussed first.

The independence assumption

Prevalence for the joint group

When conditions X and C are statistically independent, the joint prevalence is the product of the individual (marginal) prevalences:

$$p_{XC} = p_X \cdot p_C. \tag{17}$$

Transition hazards for the joint group

Independence implies that the hazards for the group with X or C are equal to the corresponding hazards for the group without X or C (in eq. (18)

populations are denoted in differential equations (disjoint) notation from the right-hand column of Table 1):

$$\begin{aligned} h_{\mathrm{XC}\to\mathrm{C}} &= h_{\mathrm{X}\to\mathrm{S}} \\ h_{\mathrm{XC}\to\mathrm{X}} &= h_{\mathrm{C}\to\mathrm{S}} \\ h_{\mathrm{C}\to\mathrm{XC}} &= h_{\mathrm{S}\to\mathrm{X}} \\ h_{\mathrm{X}\to\mathrm{XC}} &= h_{\mathrm{S}\to\mathrm{C}} \end{aligned} \tag{18}$$

Joint case fatality hazard

The probabilities $\pi_{\mathrm{X}\xrightarrow{\mathrm{X}}\mathrm{D}}$ and $\pi_{\mathrm{C}\xrightarrow{\mathrm{C}}D}$ for an individual in group X or C to die of cause X or C, respectively, during an interval Δt are:

$$\pi_{\mathrm{X}\xrightarrow{\mathrm{X}}\mathrm{D}} = (1-\mathrm{e}^{-f_{\mathrm{X}}\cdot\Delta t}) = \left.\frac{\Delta X}{X}\right|_{X\xrightarrow{\mathrm{X}}D} \text{ and } \pi_{\mathrm{C}\xrightarrow{\mathrm{C}}D} = (1-\mathrm{e}^{-f_{\mathrm{C}}\cdot\Delta t}) = \left.\frac{\Delta C}{C}\right|_{C\xrightarrow{\mathrm{C}}D}. \tag{19}$$

So the joint probability $\pi_{\mathrm{XC}\xrightarrow{\mathrm{X\,or\,C}}D}$ for someone in the group XC dying of either X or C is given by the laws of probability:

$$\begin{aligned} \pi_{\mathrm{XC}\xrightarrow{\mathrm{X\,or\,C}}D} &= \pi_{\mathrm{X}\xrightarrow{\mathrm{X}}\mathrm{D}} + \pi_{\mathrm{C}\xrightarrow{\mathrm{C}}D} - \left(\pi_{\mathrm{X}\xrightarrow{\mathrm{X}}\mathrm{D}}\cdot\pi_{\mathrm{C}\xrightarrow{\mathrm{C}}D}\right) \\ &= (1-\mathrm{e}^{-f_{\mathrm{X}}\cdot\Delta t}) + (1-\mathrm{e}^{-f_{\mathrm{C}}\cdot\Delta t}) - (1-\mathrm{e}^{-f_{\mathrm{X}}\cdot\Delta t})\cdot(1-\mathrm{e}^{-f_{\mathrm{C}}\cdot\Delta t}) \\ &= 1-\mathrm{e}^{-f_{\mathrm{X}}\cdot\Delta t}\mathrm{e}^{-f_{\mathrm{C}}\cdot\Delta t} \\ &= 1-\mathrm{e}^{-(f_{\mathrm{X}}+f_{\mathrm{C}})\cdot\Delta t} \\ &\equiv 1-\mathrm{e}^{-f_{\mathrm{XC}}\cdot\Delta t} \end{aligned} \tag{20}$$

Although individuals in the joint group XC are at risk of death from either X or C, or from other causes, the probability framework requires the assumption that they do not die of simultaneous causes (i.e. there is no cause of death "XC").

The combined case-fatality rate f_{XC} is thus:

$$f_{\mathrm{XC}} = f_{\mathrm{X}} + f_{\mathrm{C}}. \tag{21}$$

This simple addition rule can be generalized to situations with more than two independent causes of death.

Background mortality hazard

The "background mortality hazard" m expresses mortality risk for population T due to any cause of death other than X and C. The "independence assumption" claims m is independent of these causes, in other words, that m acts equally on all groups (in eqs. (22–25) populations are denoted in differential equations notation from the right-hand column of Table 1):

$$m \cdot T = m \cdot (S + X + C + XC) = m \cdot S + m \cdot X + m \cdot C + m \cdot XC \cdot \tag{22}$$

The total ("all cause" or "crude") death hazard for the population is written m_{tot}. The following identity expresses the constraint that deaths in population T equal the sum of deaths in populations S, X, C and XC:

$$m_{\text{tot}} \cdot T = m \cdot S + (m + f_{\text{X}}) \cdot X + (m + f_{\text{C}}) \cdot C + (m + f_{\text{XC}}) \cdot XC \cdot \tag{23}$$

Thus:

$$\begin{aligned} m_{\text{tot}} \cdot T &= m \cdot (S + X + C + XC) + f_{\text{X}} \cdot X + f_{\text{C}} \cdot C + f_{\text{XC}} \cdot XC \\ &= m \cdot T + f_{\text{X}} \cdot X + f_{\text{C}} \cdot C + (f_{\text{X}} + f_{\text{C}}) \cdot XC \\ &= m \cdot T + f_{\text{X}} \cdot (X + XC) + f_{\text{C}} \cdot (C + XC) \end{aligned} \tag{24}$$

Since by definition group X or C contributes no deaths due to cause C or X, respectively, we can write the system of equations:

$$\begin{aligned} f_{\text{C}} \cdot (C + XC) &= m_{\text{C}} \cdot T \,, \\ f_{\text{X}} \cdot (X + XC) &= m_{\text{X}} \cdot T \,, \end{aligned} \tag{25}$$

so:

$$m_{\text{tot}} \cdot T = m \cdot T + m_{\text{X}} \cdot T + m_{\text{C}} \cdot T \cdot \tag{26}$$

and:

$$m = m_{\text{tot}} - m_{\text{X}} - m_{\text{C}} \cdot \tag{27}$$

Likewise, this rule is generalizable to scenarios with more than three (m, X, C) independent causes of death.

Relaxing the independence assumption

As noted in the introduction, one of the primary reasons for the introduction of PopMod was to model disease interactions in a longitudinal population model. Modelling interactions requires relaxing the assumption of independence.

In the presimulation (see p. 36) of the "stocks and flows" required for the initial scenario, three areas of interaction for the health states X and C can be distinguished. Having X (C) may make it more or less likely to:

(1) have C (X),
(2) acquire or recover from C (X),
(3) die from C (X).

Note that while interaction (1) could alternatively be considered the cumulative result of interactions (2) and (3) in the past, this is not the approach adopted here.

Interaction (1): Prevalence of the joint group

In this and subsequent sections except where noted, we revert to the notation from the left-hand column of Table 1. Table 4 shows six possible cases for calculating prevalence of the joint group depending on the type of information known about the disease interaction. The probability notation π is used for prevalence, where $\pi_{X|C}$ is the probability of having disease X among those who have disease C and π_X and π_C are short forms for $\pi_{X|T}$ and $\pi_{C|T}$. Relative risk (RR) is defined here as a ratio of probabilities (risk ratio), for example, $RR_{C|X} = \pi_{C|X} / \pi_{C|-X}$ is the probability of having C if X is present over the probability of having C if X is not present.

Table 4: Options for calculating overlap probability π_{xc}

Case	π_{xc} calculated as	Comment
1	$\pi_C \cdot \pi_X$	C and X are independent
2	$\pi_{C\|X} \cdot \pi_X$	C and X interact and $\pi_{C\|X}$ or $\pi_{X\|C}$ is known.
3	$\pi_{X\|C} \cdot \pi_C$	
4	$\pi_C \cdot \pi_X / [\pi_C + (1 - \pi_C) / RR_{X\|C}]$	C and X are dependent and the relative risk $RR_{X\|C}$ or $RR_{C\|X}$ is known.
5	$\pi_X \cdot \pi_C / [\pi_X + (1 - \pi_X) / RR_{C\|X}]$	X (C) either potentiates, or protects from, C (X).
6	$\pi_X \cdot \pi_C \cdot k$	

Calculations for case 1 follow directly from the assumption of independence. Cases 2 and 3 follow directly from the definition of conditional probability. Cases 4 and 5 are derived as follows. Since the probability of belonging to the joint group is independent of which disease group is conditioned on, it is clear that:

$$\pi_{XC} = \pi_{X|C} \cdot \pi_C = \pi_{C|X} \cdot \pi_X . \tag{28}$$

Using the definition of conditional probability, we write:

$$\begin{aligned} \pi_X &= \pi_{X|C} \cdot \pi_C + \pi_{X|\sim C} \cdot \pi_{\sim C}, \text{and} \\ \pi_C &= \pi_{C|X} \cdot \pi_X + \pi_{C|\sim X} \cdot \pi_{\sim X} . \end{aligned} \tag{29}$$

Now supposing $RR_{X|C}$ or $RR_{C|X}$ is known, solving either for $\pi_{X|C}$ or $\pi_{C|X}$ and substituting the result into eq. (29) and solving again for $\pi_{X|C}$ and $\pi_{C|X}$ yields:

$$\begin{aligned} \pi_{X|C} &= \pi_X \Big/ \left(\pi_C + \pi_{\sim C} \Big/ RR_{X|C}\right), \text{and} \\ \pi_{C|X} &= \pi_C \Big/ \left(\pi_X + \pi_{\sim X} \Big/ RR_{C|X}\right). \end{aligned} \tag{30}$$

So again using the definition of conditional probability:

$$\begin{aligned} \pi_{XC} &= \pi_X \cdot \pi_C \Big/ \left(\pi_C + \pi_{\sim C} \Big/ RR_{X|C},\right) \text{and} \\ \pi_{XC} &= \pi_C \cdot \pi_X \Big/ \left(\pi_X + \pi_{\sim X} \Big/ RR_{C|X}.\right) \end{aligned} \tag{31}$$

Recalling $1 - \pi_X = \pi_{\sim X}$ and $1 - \pi_C = \pi_{\sim C}$, the required expressions in Table 4 are obtained.

The factor *k* in case 6 is an arbitrary multiplier that increases or reduces the prevalence of group XC compared to what would be obtained under independence, and lies between 0 and 1 if having one disease reduces the probability of having the other, and between 1 and MAX($1/\pi_C$, $1/\pi_X$) if having one disease makes it more likely to have the other. Upper bounds on *k* are easy to derive using the fact that $\pi_{XC} = \pi_X = \pi_C$ when X and C are obligate symbiotes.

The six cases span a range of information availability about interaction of X and C on the prevalence of the joint condition:

- Case 1 assumes independence (no interaction).
- Case 2 and 3 assume conditional prevalence is known.
- Case 4 and 5 assume relative risk is known.
- Case 6 assumes a potentiation (or protection) factor can be defined.

Interaction (2): Incidence and remission for the joint group

For incidence hazard, we write i and for remission hazard, r. Consistent with "overlapping populations", unless specifically noted, hazards are understood as "total hazards", that is, i_X includes all incidence to X regardless of whether C is also present in the population at risk. Conditional hazards are denoted $i_{X|\sim C}$ or $i_{X|C}$ to signify "incidence to X in the group without C" and "incidence to X in the group with C", respectively.

Consider total incidence i_X for the initial scenario. The product of total incidence to X and the total population without X (~X) must be equal to the sum of the products of the conditional incidences ($i_{X|\sim C}$, $i_{X|C}$) and the conditional populations (~X~C, ~XC):

$$i_X \cdot (\sim X) = i_{X|\sim C} \cdot (\sim X \sim C) + i_{X|C} \cdot (\sim XC) \cdot \tag{32}$$

Dividing by total population T yields:

$$i_X \cdot \frac{(\sim X)}{(T)} = i_{X|\sim C} \cdot \frac{(\sim X \sim C)}{(T)} + i_{X|C} \cdot \frac{(\sim XC)}{(T)}, \tag{33}$$

and replacing population ratios by the corresponding prevalence rates yields:

$$i_X \cdot \pi_{\sim X} = i_{X|\sim C} \cdot \pi_{\sim X \sim C} + i_{X|C} \cdot \pi_{\sim XC} \cdot \tag{34}$$

Dividing both sides by $\pi_{\sim X}$ yields the following expression for i_X:

$$i_X = i_{X|\sim C} \cdot \frac{\pi_{\sim X \sim C}}{\pi_{\sim X}} + i_{X|C} \cdot \frac{\pi_{\sim XC}}{\pi_{\sim X}}, \tag{35}$$

where:

$$\pi_{\sim X} = \pi_{\sim X \sim C} + \pi_{\sim XC} \cdot \tag{36}$$

It is therefore clear that total incidence to X is a weighted average of the conditional incidences, where the weights are the proportions of the population without X partitioned according to C status.

Recall that, in terms of the differential equations notation from the right-hand column of Table 1, $\pi_{\sim X} = \pi_C + \pi_S$, $\pi_{\sim X \sim C} = \pi_S$ and $\pi_{\sim XC} = \pi_C$, the values of which are determined according to one of the six cases defined above in interaction (1). Thus, when total hazard i_X is known, eq. (34) has only two unknowns ($i_{X|\sim C}$ and $i_{X|C}$). Clearly, if information on one or both conditional hazards is available, interaction (2) with respect to i_X is fully characterized for the initial scenario.

However, the guiding principle of the presimulation problem was that information on the non-overlapping populations (e.g. direct observation of the conditional hazards) is relatively scarce. When this is true, the unknown conditional hazards must remain undetermined unless one of the following three rate ratios (RR) is known or can be approximated:

$$RR(i_X)_1 = \frac{i_{X|C}}{i_{X|\sim C}}, RR(i_X)_2 = \frac{i_{X|\sim C}}{i_X}, \text{ or } RR(i_X)_3 = \frac{i_{X|C}}{i_X}. \qquad (37)$$

A similar situation applies to the total hazards i_C, r_C, and r_X for the initial scenario, that is, eq. (34) is one of a family of equations representing the relation between the total disease hazards and the corresponding conditional hazards for subpopulations:

$$\begin{aligned} i_X \cdot \pi_{\sim X} &= i_{X|\sim C} \cdot \pi_{\sim X \sim C} + i_{X|C} \cdot \pi_{\sim XC} \\ i_C \cdot \pi_{\sim C} &= i_{C|\sim X} \cdot \pi_{\sim X \sim C} + i_{C|X} \cdot \pi_{X \sim C} \\ r_X \cdot \pi_X &= r_{X|\sim C} \cdot \pi_{X \sim C} + r_{X|C} \cdot \pi_{XC} \\ r_C \cdot \pi_C &= r_{C|\sim X} \cdot \pi_{\sim XC} + r_{C|X} \cdot \pi_{XC}. \end{aligned} \qquad (38)$$

Note that, with respect to the initial scenario, eq. (38) forms a simultaneous system with eq. (31) – or one of the other methods of calculating π_{XC} noted in Table 4

– and the system has a unique numerical solution whenever enough parameter values are known, that is, assuming the four total hazards are known, if one of the three following rate ratios (or its inverse) is known for each hazard:

$$\begin{aligned}
&RR(i_{\mathrm{X}})_1 = \frac{i_{\mathrm{X|C}}}{i_{\mathrm{X|\sim C}}},\ RR(i_{\mathrm{X}})_2 = \frac{i_{\mathrm{X|\sim C}}}{i_{\mathrm{X}}},\ \text{or } RR(i_{\mathrm{X}})_3 = \frac{i_{\mathrm{X|C}}}{i_{\mathrm{X}}},\ \text{and}\\
&RR(i_{\mathrm{C}})_1 = \frac{i_{\mathrm{C|X}}}{i_{\mathrm{C|\sim X}}},\ RR(i_{\mathrm{C}})_2 = \frac{i_{\mathrm{C|\sim X}}}{i_{\mathrm{C}}},\ \text{or } RR(i_{\mathrm{C}})_3 = \frac{i_{\mathrm{C|X}}}{i_{\mathrm{C}}},\ \text{and}\\
&RR(r_{\mathrm{X}})_1 = \frac{r_{\mathrm{X|C}}}{r_{\mathrm{X|\sim C}}},\ RR(r_{\mathrm{X}})_2 = \frac{r_{\mathrm{X|\sim C}}}{r_{\mathrm{X}}},\ \text{or } RR(r_{\mathrm{X}})_3 = \frac{r_{\mathrm{X|C}}}{r_{\mathrm{X}}},\ \text{and}\\
&RR(r_{\mathrm{C}})_1 = \frac{r_{\mathrm{C|X}}}{r_{\mathrm{C|\sim X}}},\ RR(r_{\mathrm{C}})_2 = \frac{r_{\mathrm{C|\sim X}}}{r_{\mathrm{C}}},\ \text{or } RR(r_{\mathrm{C}})_3 = \frac{r_{\mathrm{C|X}}}{r_{\mathrm{C}}}.
\end{aligned} \tag{39}$$

Interaction (3): Mortality for the joint group

This interaction concerns causes of death. We assume that the all-cause mortality hazard m_{tot} and the total (i.e. overlapping) case-fatality hazards f_{X} and f_{C} are known. It follows that:

$$\begin{aligned}
&f_{\mathrm{X}} \cdot \pi_{\mathrm{X}} = f_{\mathrm{X|\sim C}} \cdot \pi_{\mathrm{X\sim C}} + f_{\mathrm{X|C}} \cdot \pi_{\mathrm{XC}},\ \text{and}\\
&f_{\mathrm{C}} \cdot \pi_{\mathrm{C}} = f_{\mathrm{C|\sim X}} \cdot \pi_{\mathrm{\sim XC}} + f_{\mathrm{C|X}} \cdot \pi_{\mathrm{XC}}.
\end{aligned} \tag{40}$$

Following a derivation similar to that in eqs (19) – (21), one can show that, given total case-fatality hazards f_{X} and f_{C}, the case-fatality hazard for the joint condition is the sum of the conditional hazards:

$$f_{\mathrm{XC}} = f_{\mathrm{X|C}} + f_{\mathrm{C|X}} \tag{41}$$

Further, since:

$$\begin{aligned}
&m_{\mathrm{X}} \cdot T = f_{\mathrm{X|\sim C}} \cdot (X \sim C) + f_{\mathrm{X|C}} \cdot (XC),\ \text{and}\\
&m_{\mathrm{C}} \cdot T = f_{\mathrm{C|\sim X}} \cdot (\sim XC) + f_{\mathrm{C|X}} \cdot (XC),
\end{aligned} \tag{42}$$

so:

$$m_X = f_{X|\sim C} \cdot \pi_{X\sim C} + f_{X|C} \cdot \pi_{XC}, \quad (43)$$

and:

$$m_C = f_{C|\sim X} \cdot \pi_{\sim XC} + f_{C|X} \cdot \pi_{XC}. \quad (44)$$

In other words, the cause-specific mortality hazards are weighted averages of the conditional case-fatality hazards, where weights are the proportions of the total population according to disease status regarding the other condition.

It remains true that:

$$m = m_{tot} - m_X - m_C, \quad (45)$$

as in eq. (27).

Other interactions

Another interaction might involve relaxing the assumption of independence between background mortality hazard m and case-fatality hazards f_X and f_C. However, in cases where such dependence is suspected or known, it may be possible to "work around" it by choosing appropriate definitions for X and C. For example, to take the ischaemic heart disease (X) and stroke (C) example, suppose it is important for the research question to account for the fact that individuals with X or C are also at increased risk of mortality from other selected causes of death such as cardiac failure. While one approach might be to introduce a new box for cardiac failure, within the current structure of PopMod, the onus is effectively on the analyst to take into account such increased risk of background mortality by modifying the way state C is defined and by adjusting the corresponding incidence and case-fatality rates. For example, state C could be defined as "stroke and all other conditions (including cardiac failure) at increased risk due to heart disease". Another type of exception to the general rule of independence between background mortality and cause-specific mortality would be the existence of any common causal modifiers of m, f_X and f_C, for example, the allocation of health-care expenditure.

Modelling mechanics

Initial conditions

PopMod describes population evolution conditional on initial conditions that define the state of the system at some initial time. These initial conditions consist of the population distribution in non-overlapping terms. If potentially overlapping populations (i.e. descriptions from the left hand side of Table 1) are considered, when the total prevalences p_X and p_C are known the non-overlapping population distribution can be fully determined by determining the prevalence of the joint group. Methods for this are discussed in the section *Disease interactions*.

Runge-Kutta method

The differential equation system is determined by its initial conditions and its parameters. An algebraic description of PopMod differential equation system – using notation from the right-hand side of Table 1 – is:

$$dS/dt = -(h_{S\to X} + h_{S\to C} + h_{S\to D})\cdot S + (h_{X\to S})\cdot X + (h_{C\to S})\cdot C \quad (46a)$$

$$dX/dt = -(h_{X\to S} + h_{X\to XC} + h_{X\to D})\cdot X + (h_{S\to X})\cdot S + (h_{XC\to X})\cdot XC \quad (46b)$$

$$dC/dt = -(h_{C\to X} + h_{C\to XC} + h_{C\to D})\cdot C + (h_{S\to C})\cdot S + (h_{XC\to C})\cdot XC \quad (46c) \quad (46)$$

$$dXC/dt = -(h_{XC\to X} + h_{XC\to C} + h_{XC\to D})\cdot XC + (h_{X\to XC})\cdot X + (h_{C\to XC})\cdot C \quad (46d)$$

$$dD/dt = (h_{S\to D})\cdot S+(h_{X\to D})\cdot X + (h_{C\to D})\cdot C + (h_{XC\to D})\cdot XC \quad (46e)$$

where S, X, C and XC are unknown. Under specified conditions, which apply here, such a differential equation system has a unique, exact and analytical solution, and the solution can be expressed in terms of the eigenvalues and eigenvectors of the 5 x 5 coefficient matrix.[10]

Since finding the required eigenvalues and eigenvectors is here equivalent to solving a fifth-degree polynomial equation, specialized solution algorithms – and access to a substantial amount of processor time – will generally be required. An attractive alternative is therefore the use of numerical techniques, since they yield solutions more cheaply, and without requiring custom routines.

In PopMod, the evolution of the population in time is approximated by a 4th-order Runge-Kutta method, or, optionally, by a 5th-order Runge-Kutta method.[10, please consult reference for further information] The relevant time step is defined as a fraction of the standard reporting interval (the number of divisions of the basic reporting interval must in principle be divisible by 3, but to allow for the possibility of starting with mid-year values in the first year, the number of divisions must be divisible by 6 and the minimum number of divisions is fixed at 12). Note that an n^{th}-order numerical method will in general provide useful results so long as the differentials are smaller than n^{th}-order in the chosen time step.

Each population age- and sex group is modelled as a separate system, and age is updated by taking end-of-year solution values for the "age = α" system as the initial values for the "age = $\alpha + 1$" system in the subsequent model year.

A 4th-order Runge-Kutta method provides solutions to differential equations of the type:

$$\mathrm{d}y_i(x)/\mathrm{d}x = f_i(x, y_i(x)), \tag{47}$$

and is defined by the ansatz (Euler method) that:

$$y_i(x + \Delta x) = y_i(x) + \Delta x \cdot f_i(x, y_i(x)), \tag{48}$$

where:

$$\begin{aligned} & y_i(x + \Delta x) = y_i(x) + (k_{1i} + 2k_{2i} + 2k_{3i} + k_{4i})/6 + O(\Delta x^5), \text{ and} \\ & k_{1i} = \Delta x \cdot f_i(x, y_i) \\ & k_{2i} = \Delta x \cdot f_i(\mathrm{x} + \Delta x/2, y_i + k_{1i}/2) \\ & k_{3i} = \Delta x \cdot f_i(\mathrm{x} + \Delta x/2, y_i + k_{2i}/2) \\ & k_{4i} = \Delta x \cdot f_i(\mathrm{x} + \mathrm{d}x, y_i + k_{3i}). \end{aligned} \tag{49}$$

Note that here $x = t$, y_i = S, X, C, XC and D and that the differential equations (46a-46e) are not explicitly time dependent, that is, $f_i(t, y_i(t)) = f_i(y_i(t))$.[10]

Output interpretation

Standard PopMod output reports $P(t)$ for each population group as end-of-interval (e.g. year-end) values, corresponding to the standard life table quantity l_x. An important derived quantity also included in output is the time at risk experienced by the group during the interval ($\int P(t)\,\mathrm{d}t$), corresponding to the life table quantity L_x (sometimes called "life-years" or "person-years").

For a constant population, population time at risk is calculated $P \cdot \Delta t$. For PopMod populations, population time at risk for the interval $b - a$ is calculated:

$$P_{LY} = \int_a^b P(t)\,\mathrm{d}t \text{ , for example } X_{LY} = \int_a^b X(t)\,\mathrm{d}t\,. \tag{50}$$

When the quantity resulting from eq. (50) with units "person-years" is divided by the length of the time interval with units "years", average population size for the interval ($\hat{P}$, with units "persons") is obtained:

$$\hat{P} = \frac{1}{b-a}\int_a^b P(t)\,\mathrm{d}t\,. \tag{51}$$

$\hat{P}$ thus conforms to the definition of the expected value of the function $P(t)$ on the interval $b–a$. Since $b–a$ is by convention one year (or "chronon" etc.), the normalization to the interval $b–a$ means dividing by 1. Thus, since in this case the numerical quantity is unchanged, substituting different reporting units yields two equally valid interpretations for the same output:

(1) the average population size $\hat{P}$ = E[$P(t)$] during the interval Δt, or
(2) the population time at risk P_{LY} experienced during the interval Δt.

Interpretation (1) also corresponds to average (count) prevalence for the population.

When transition rates are "small" (i.e. the differentials are approximately linear), average population can be interpreted as mid-interval population. Under the same assumptions, mid-year population provides a good estimate of population time at risk.

PopMod numerically evaluates P_{LY} with a standard Newton-Cotes formula for 4-point closed quadrature, sometimes also called Simpson's 3/8-rule.[11] The quadrature formula relies on the values of $P(t)$ determined by the Runge-Kutta method at multiples of the chosen time step. Since these values involve numerical estimation error, there is no simple expression for the order of accuracy of the different output values reported in PopMod.[10]

Discussion

Advantages of the approach

PopMod combines features of existing models (see below) with the possibility to analyse several disease states. It explicitly analyses time evolution and, even more importantly, abandons the constraint of independence of disease states.

A primary advantage of the approach adopted in PopMod is the separate modelling of age and time, and the type of bias inherent in models that do not do so has been previously pointed out.[7] Moreover, it has been independently noted that, without this feature, life-table measures are constrained to adopt – somewhat artificially – either a "period" or a "cohort" perspective.[6] The other chief advantage of PopMod is the ability to deal with heterogeneity of disease and mortality risk by modelling up to four disease states. No previous published generic population model has combined both these features. Note, however, that if disease conditions are independent, and population-dependent effects are not of interest, a multi-state life-table approach should probably be adopted.[20]

A further advantage of PopMod is the introduction of a systematic analytical approach to the modelling of disease interactions. This by itself represents a relatively important advance, as modellers have until now been constrained to model only independent conditions. Furthermore, in spite of the increased

informational demands made by a four-state system, the modelled functional dependency between X-related hazards conditional on C status, and vice versa, reduces the number of exogenous hazards that need to be directly observed. This is of substantial practical importance, since, while direct observation of conditional hazards usually requires a cohort study, it will often be possible to obtain estimates of the required rate ratios from more common case-control studies[21].

Related models

In addition to the multi-state life table family,[3,4] two additional families of mathematical models have some similarity to PopMod. One family comprises the class of models sometimes called incidence, prevalence and mortality (IPM) models.[12-14] Another family (with until now one member) is that of published population models, in particular *Prevent*.[15-18]

IPM models per se have no population or age structure; they can be conceived of as stationary population models (i.e. models of a population in equilibrium, where the numbers of births and deaths in an age group are equal). However, DisMod, probably the IPM model in most common use,[12] has gone through several versions, and the current version allows for hazard trend analysis that relies on modelling a full population structure based on one-year age groups. Notwithstanding, IPM models analyse only a single disease condition in isolation, and, while *Prevent* was explicitly designed to analyse a full population cohort structure, it also analyses only a single disease condition.

Multi-state life tables analyse multiple disease states but published versions have invariably required the assumption of independence across diseases. In addition, multi-state life tables implicitly impose a stationary population assumption by not independently modelling population time and age.

Averaging and its implications

In all compartmental models, of which differential equations models are one type, it is assumed that health and mortality risk are conditional on disease state. In light of the seemingly infinite diversity of real phenomena, this assumption invariably results in "compression", that is, the imposition

of artificial homogeneity. In many cases, compression can be considered a necessary simplifying assumption for the modelling exercise, but in other cases, heterogeneity must be explicitly modelled to avoid the phenomenon of confounding. In a differential equations system, modelling heterogeneity of disease and mortality risk amounts to introducing additional disease states. Thus, PopMod, with four disease states, respresents a substantial increase in complexity over population models modelling only two disease states (e.g. diseased and healthy). PopMod of course includes the two-disease-state model as a special case.

There is also heterogeneity other than of disease and mortality risk. In particular, although real populations change in integer steps at discrete moments in time, a differential equation system represents this process in continuous time. However, this approximation is in general acceptably good when a large number of individuals comprise the population of interest. Moreover, an implication of representing age in a discrete number of statistical bins is modelling a birth-year cohort as though it had a single average age. If births are distributed uniformly throughout the year, the average birthday of the cohort is the mid-year point, and there is no serious objection to this procedure. However, if the cohort average birthday is not be the mid-year point, PopMod's modelled age will differ from the true average age.

It is assumed that conditional hazards are constant within a single reporting interval (e.g. one year), which will in principle be problematic for conditions with high initial case-fatality, for example heart attack (or stroke). This sort of problem can be addressed by defining condition C as "acutely fatal cases" and condition X as "long-term survivors". Similarly, for conditions of determinate duration (e.g. pregnancy), use of a constant hazard rate for "remission" will result in an exponential distribution of waiting time for transition out of the state, whereas a uniform distribution of waiting time is what would be wanted.

All compartmental population models are fundamentally simplifications of reality by means of a system of reduced dimensionality. The mathematical concept of "projection" is useful: the simplified system can be thought of as a "least-squares approximation" to the higher-order real system.[19] The validity

of input parameter values and the accuracy of the solution method determine the actual goodness of fit realized in a particular model. Nevertheless, compression applies to every modelled variable in a differential equations model. Other modelling approaches, such as microsimulation, require much less compression, so the user who wishes to avoid compression systematically should consider adopting the microsimulation approach.

Types of error in PopMod

Sources of error in PopMod can be divided into three types:

(1) Model (or "projection") error due to analysing a simplified system instead of the full one. Model error includes the characterization of scenarios for disease interaction.
(2) Numerical error due to obtaining approximate solution values with numerical techniques.
(3) Parameter error due to uncertainty about observed or derived parameter values.

The 5^{th}-order Runge-Kutta method provides an estimate of the local truncation error inherent in the 4^{th}-order numerical technique. Monte-Carlo analysis of distributions around transition rates can be used to examine parameter uncertainty. However, comparison with a more complex model would be necessary for quantification of model error. A way of investigating the impact of model error would be to construct progressively more realistic and complex models. A spectrum of models, from least to most complex, can thus be imagined, where the "most complex" and necessarily imaginary model has a one-to-one relation to real system it represents. The difference between the results of two adjacent models in such a series would be an expression of model error analogous to the estimate of numerical truncation error afforded by the next-higher-order numerical method.

Although intuitively natural and mathematically valid, in most situations it would be impractical to quantify model error in this laborious way. Nevertheless, model error may, in certain data-rich cases, be estimated by "predicting" outcomes for which numerical data are available for comparison but which are not used as inputs.

Limiting assumptions

Although any state transitions are in principle possible, PopMod assumes that transitions S→XC and X→C do not occur. This is because such transitions can be thought of as the simultaneous occurrence of two transitions (for example, S→XC equals S→X plus X→XC). Note that this does not imply events S→XC and X→C cannot occur within a single reporting interval; rather, it just means the mathematics of PopMod do not represent simultaneous events. A similar feature is the absence of a modelled cause of death "XC".

However, the non-modelled transition S→XC can be imagined if someone in state S simultaneously acquires X and C as a result of, say, very high levels of common risk factors (i.e. someone who suffers a simultaneous heart attack and stroke because of high blood pressure and cholesterol). If such a "simultaneous event" results in mortality, one could potentially speak of a cause of death "XC". Similarly, the non-modelled transition X→C could occur if there were "perfect interference" between two diseases such that acquiring C caused immediate remission from X . If either of these cases is important, PopMod can miss important dynamics.

Authors' contributions

JL devised the methodology, implemented the conceptual and technical development of PopMod, including coordination of co-authors' contributions, and drafted and revised the manuscript. KR and HW contributed to the development of the methodology, drafted certain sections and revised the manuscript. CC and SG contributed to the development of the methodology, revised mathematical formulae throughout, and revised the maunscript. CM provided the initial idea for the model and also contributed technical modifications throughout development of the main ideas presented in this paper. All authors approved the final manuscript.

Conflict of interest

None declared.

Acknowledgements

Reviewers Nico Nagelkerke of Leiden University and Louis Niessen of Erasmus University Rotterdam are gratefully acknowledged. Jan Barendregt and Sake de Vlas of Erasmus University Rotterdam also offered valuable comments and suggestions for revision. David Evans, Raymond Hutubessy, Stephen Lim, Colin Mathers, Sumi Mehta, Josh Salomon and Tessa Tan Torres of the World Health Organization, Geneva, are also gratefully acknowledged for their comments and contributions.

References

1. Shryock HS, Siegel JS: **The life table.** In: *The methods and materials of demography* (Edited by *Stockwell EG*) *Orlando, FL*, *Academic Press* 1976, 250.

2. Hacking I: **The emergence of probability.** *Cambridge, Cambridge University Press* 1975.

3. Barendregt JJ, Bonneux L: **Degenerative disease in an aging population: models and conjectures.** *Enschede, Netherlands, Febodruke* 1998.

4. Manton KG, Stallard E: **Chronic disease modelling.** *London, Charles Griffin* 1988.

5. Schoen R: **Modelling multigroup populations.** *New York and London, Plenum Press* 1987.

6. Barendregt JJ: **Incidence- and prevalence-based summary measures of population health (SMPH): making the twain meet.** In: *Summary measures of population health: concepts, ethics, measurement and applications* (Edited by *Murray CJL, Lopez AD, Salomon JA, Mathers CD*) *Geneva*, *World Health Organization* 2002, 221-231.

7. Preston SH: **Health indices as a guide to health sector planning: a demographic critique.** In: *The epidemiological transition, policy and planning implications for developing countries* (Edited by *Gribble JN, Preston SH*) *Washington DC*, *National Academy Press* 1993.

8. Murray CJL: **Rethinking DALYs.** In: *The Global Burden of Disease: A comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020* (Edited by *Murray CJ, Lopez AD*) *Cambridge, MA*, *Harvard University Press* 1996, 1-98.

9. Anderson RM, May RM: **Infectious diseases of humans: dynamics and control.** *Oxford, Oxford University Press* 1991.

10. Lambert JD: **Numerical methods for ordinary differential systems: the initial value problem.** *Chichester, Wiley* 1991.

11. Ralston A, Rabinowitz P: **A first course in numerical analysis.** 2nd edn. *Mineola, NY, Dover* 1965.

12. Kruijshaar ME, Barendregt JJ, Hoeymans N: **The use of models in the estimation of disease epidemiology.** *Bull World Health Organ* 2002, **80:** 622-628.

13. Murray CJ, Lopez AD: **Quantifying disability: data, methods and results.** *Bull World Health Organ* 1994, **72:** 481-494.

14. Barendregt JJ, Baan CA, Bonneux L: **An indirect estimate of the incidence of non-insulin-dependent diabetes mellitus.** *Epidemiology* 2000, **11:** 274-279.

15. Mooy JM, Gunning-Schepers LJ: **Computer-assisted health impact assessment for intersectoral health policy.** *Health Policy* 2001, **57:** 169-177.

16. Bronnum-Hansen H: **How good is the Prevent model for estimating the health benefits of prevention?** *J Epidemiol Community Health* 1999, **53:** 300-305.

17. Gunning-Schepers LJ: **Models: instruments for evidence based policy.** *J Epidemiol Community Health* 1999, **53:** 263.

18. Bronnum-Hansen H, Sjol A: **[Prediction of ischemic heart disease mortality in Denmark 1982-1991 using the simulation model Prevent].** *Ugeskr Laeger* 1996, **158:** 4898-4904.

19. Strang G: **Linear algebra and its applications.** 3rd edn. *San Diego, CA, Harcourt Brace Jovanovich* 1988.

20. Barendregt JJ, van Oortmarssen GJ, van Hout BA, Van Den Bosch JM, Bonneux L: **Coping with multiple morbidity in a life table.** *Math Popul Stud* 1998, **7:** 29-49, 109.

21. Rothman KJ, Greenland S: **Modern epidemiology.** 2nd edn. *Philadelphia, PA, Lippincott Williams and Wilkins* 1998.

Chapter 3

Effectiveness and costs of interventions to lower systolic blood pressure and cholesterol: a global and regional analysis on reduction of cardiovascular-disease risk

Based on: Murray CJL, Lauer JA, Hutubessy RCW, Niessen LW, Tomijima N, Rodgers A, Lawes CMM, Evans DB. Effectiveness and costs of interventions to lower systolic bloodpressure and cholesterol: a global and regional analysis on reduction of cardiovascular-disease risk *Lancet,* 2003 361: 717-25.

Introduction

Cardiovascular disease (CVD) is a major cause of global disease burden. It accounts for 20.3% of DALYs lost in developed countries, and already causes 8.1% of DALYs lost in developing countries. The World Health Report 2002 quantified the major contribution of tobacco, alcohol, high blood pressure, high cholesterol, low fruit and vegetable intake, physical inactivity and high body mass index to the global burden of disease and the burden of CVD in particular [1,2].

Improved data on the levels of exposure and reassessments of the magnitude of the hazards has led to the recognition that high blood pressure and high cholesterol have much higher impacts on population health than previously thought [3]. Approximately two thirds of stroke and almost half of IHD can be attributed to blood pressure levels above 115 mm Hg. Total cholesterol over 3.8mmol/l accounted for approximately 18% of stroke and 55% of IHD. The joint effects of blood pressure and cholesterol would, of course, be less than additive because of the multi-causality of cardiovascular disease and the joint action of these two risk factors [1]. This regional analysis has also shown that high blood pressure and high cholesterol are major risks to health in all regions of the world, not just high-income countries.

Given the burden of disease caused by high blood pressure and high cholesterol, it is important to evaluate the costs and effects of the available intervention strategies to reduce these risks. These strategies should, however, be seen in the context of more comprehensive approaches to the control of CVD that focus on a number of inter-related risks to health including levels of blood pressure and cholesterol, tobacco consumption, body mass index, physical activity, diet and diabetes [4,5]. In this chapter, we take advantage of the development of standardized methods and companion tools for the evaluation of costs, effects and cost-effectiveness of different interventions within and across regions [6-13]. These methods and tools mean that results of intervention analyses can be compared more meaningfully across interventions and across locations.

Evaluation of the costs and effects of the major intervention strategies for reducing the burden attributable to blood pressure and cholesterol must

address two key debates. First, what are the relative roles of non-personal health services such as mass media messages to change diet or legislation to reduce the salt content of processed foods, and personal health services such as the pharmacological management of cholesterol and hypertension [4,5,14]? Second, should management of blood pressure and cholesterol be based on thresholds for each risk factor seen in isolation, such as treating for a systolic blood pressure over 160 mm Hg, or should management be based on the absolute risk of cardiovascular disease for a given individual taking into account all their known determinants of risk [15]? In this chapter we analyse the population health effects and costs of non-personal health measures, treatment of individual risk factors and treatment based on various levels of absolute risk [16-18].

Methods

Interventions

Seventeen non-personal and personal health service interventions or combinations have been included in this analysis – summary details are provided in *Table 1*. Non-personal health interventions included health education through the mass media focusing on blood pressure, cholesterol and body mass, and either legislation or voluntary agreements on salt content to ensure appropriate labelling and stepwise reductions of the salt content of processed foods. Personal health service interventions included detecting and treating people with elevated levels of cholesterol for two thresholds; treating individuals with elevated systolic blood pressure again using two thresholds; treating individuals for both elevated cholesterol and elevated systolic blood pressure; and treating individuals based on their absolute risk of a cardiovascular event in the next 10 years (called the "absolute risk" approach [20]) using four different absolute risk thresholds. Risk levels are defined by fitting mean risk factor values to observed baseline risk levels. Estimates of the relative risk of modelled risk factors on cardiovascular events are used to predict the absolute risk of individuals with elevated levels of risk factors. Individuals with an absolute risk of cardiovascular disease greater than the threshold all receive a beta-blocker, diuretic, statin, and aspirin.

Table 1. Interventions evaluated

	Intervention	Description
Non-personal interventions		
N1	Salt reduction through voluntary agreements with industry	Cooperation between government and the food industry for stepwise decrease in salt content of processed foods and for labelling
N2	Population-wide reduction in salt intake legislation	Legislation to decrease salt content in processed foods and appropriate labelling
N3	Health education through mass media	Health education through broadcast and print media focusing on body-mass index and cholesterol concentrations
N4	Combined intervention of N2 and N3	Combination of N2 and N3
Personal interventions		
P1 and P2	Individual-based hypertension treatment and education	Treatment of SBP above 160 mm Hg (P1) or above 140 mm Hg (P2) with a standard regimen of β-blocker and diuretic
P3 and P4	Individual treatment for high cholesterol concentrations and education	Treatment with statins for total cholesterol concentrations above 6·2 mmol/L (P3) and above 5·7 mmol/L (P4)
P5	Individual treatment and health education for SBP and cholesterol concentration	Combination of P2 and P3, with treatment thresholds of 140 mm Hg SBP and 6·2 mmol/L for total cholesterol concentration
P6 to P9	Absolute risk approach	People with an estimated combined risk of a cardiovascular event* over the next decade above a given threshold treated for multiple risk factors (with statin, diuretic, β blocker, and aspirin) whatever the values for individual risk actors; four different thresholds were evaluated: 35% (P6), 25% (P7), 15% (P8), and 5% (P9)
Combined personal and non-personal interventions		
(C1 to C4)	Addition of the absolute risk approach at the four thresholds to the combined non-personal health intervention	Combination of N4 with P6 to P9

SBP=systolic blood pressure. *Acute myocardial infarction; angina pectoris; congestive heart failure; first-ever fatal stroke; long-term stroke survivors. The definition of a cardiovascular event differs across studies, so the results reported here may not be strictly comparable with those of similar studies.[19]

Regions

It would be desirable to evaluate all possible combinations of interventions in every country of the world – and for some of the larger countries, to evaluate these combinations at a sub-national level. No country has yet been able to do this, and many countries do not have the technical capacity to evaluate even a few of them. At the other extreme, global estimates are of little use to any specific country. The World Health Organization, through its CHOICE project, provides information on costs and health effects at a subregional level, with the different parts of the world divided by geographic proximity and epidemiology. This allows interventions to be placed into broad categories, such as very cost-effective, cost-effective and cost-ineffective, revealing the extent to which strategies to reduce risks to health should differ across different settings.

In addition, CHOICE provides results in such a way that analysts from countries within a region can adapt them to their settings if they wish. The costs, effects and cost-effectiveness of each of the 17 interventions have been evaluated for 14 epidemiological subregions of the world. These subregions are defined and described in *Annex 1*. The results from three of them are discussed in detail - SearD (in south-east Asia with high rates of adult and child mortality); AmrB (in Latin America with low adult and child mortality); and EurA (in Europe with very low adult and child mortality).

Intervention effects

Effect sizes used in the analysis are based on systematic reviews of randomized trials where possible, or meta-analysis as shown in *Annexes 2-5*. From the evidence of large cohort studies in diverse populations the joint effects between interventions were assumed to be multiplicative, drawing from the evidence of large cohort studies in diverse populations [21-23].

Population health effects due to the interventions are modelled by stochastically simulating age-, sex-, and sub region-specific populations with the observed baseline levels of cardiovascular risk and the observed distribution of risk factors (systolic blood pressure, cholesterol, body mass index, and prevalence of long-term smokers) in those regions [24-25]. Interventions cover either the entire population (non-personal interventions) or sub populations meeting specific characteristics (personal interventions), such as having 10-year risk of a cardiovascular event, or having systolic blood pressure, over a certain threshold. Population level cardiovascular risk (incidence) is recalculated after applying the effectiveness of the intervention and the effect size of the implied change in risk factor levels to the population receiving the intervention.

To translate age and sex specific changes in the risk of cardiovascular disease events into changes in population health quantified using DALYs, a standard multi-state modelling tool, PopMod, was used [8]. In this model, health effects are estimated by tracing what would happen to each age/sex cohort of a given population over 100 years, with and without each intervention. PopMod is a four-state population model simulating the evolution of a population partitioned into four distinct health states – people who have the condition under study, have some other comorbid condition, have both conditions, and have none of the above (but are susceptible). Births and deaths are also included. The states can be considered either as simple states or as aggregates of other substates. Transition rates, such as incidence, remission, and mortality, govern movements between states.

The model is described by a system of ordinary differential equations with each population age and sex group modelled individually as a separate differential system [26]. The model shows the time evolution of the size of the population

age and sex groups, through the four health states and death, in yearly steps. With the appropriate health state valuations, standard life-table measures as well as a variety of summary measures of population health can be derived. Monte-Carlo analysis of uncertainty in transition rates is possible. The side effect relating to the consequences of bleeding associated with the use of aspirin was included. The entire population is subjected to background mortality and morbidity, which is assumed to be independent of the cardiovascular disease states explicitly modelled.

In some cases, mostly in more developed countries, information on intervention effects was available for only one or two settings. The association of blood pressure and cholesterol, however, are remarkably similar in size and shape across Asia, Europe and North America [22,27,28]. We could not obtain evidence about how adherence vary across settings, so no variation was included. Nevertheless, policy-makers must still make decisions about how to use their scarce resources. One approach would be to simply to say that there is no evidence. The approach taken here is to provide the best available evidence, even if this means extrapolating from one setting to another. This approach carries additional uncertainty, especially in the case of behavioural interventions, and this should be considered when interpreting the subsequent results.

Costs

Costs include *programme-level costs* associated with running the intervention (such as administration, training and media), and *patient-level costs* (such as primary care visits, diagnostic tests and medicines). For this analysis, potential cost-savings related to the prevention of CVD events have not been incorporated because the major interest is in identifying the costs of improving population health by preventing CVD events. Costs have been based on a standardized ingredients approach that has been developed by the World Health Organization to facilitate costing of interventions [6-9]. The units of physical inputs required were assessed and multiplied by the unit price for each input. For programme costs the quantities of the required inputs (such as labour, vehicles, office space) were identified from the literature with additional details provided by programme staff in various parts of the world. The quantity of patient-level resource inputs required for a given

health intervention e.g. hospital inpatient days, outpatient visits, medications, laboratory tests etc., were identified in a similar manner. Reporting costs using the ingredients approach is an important part of making the results transparent to policy-makers as well as providing a way for analysts to adapt the results to their own settings if they wish.

Unit costs of programme-level and patient-level resource inputs, such as the salaries of central administrators, the capital costs of vehicles, offices and furniture, or the cost per outpatient visit, were obtained from a review of the literature and supplemented by primary data from programme staff in several countries. Costs of drugs were based on the price of off-patent drugs from the lowest cost vendor of high-quality drugs.

Information on the costs and effectiveness of interventions that are undertaken inefficiently is of little value to decision makers. For that reason we assume capacity utilization of 80% in most settings – e.g. that health personnel are fully occupied for 80% of their time. The results identify, therefore, the set of interventions that, if done relatively efficiently, would be cost-effective in the different settings.

Costs are reported in International Dollars to facilitate more meaningful comparisons across regions. An international dollar has the same purchasing power as the U.S. dollar has in the United States. Costs in local currency units are converted to international dollars using purchasing power parity (PPP) exchange rates rather than official exchange rates. A PPP exchange rate is the number of units of a country's currency required to buy the same amounts of goods and services in the domestic market as a U.S. dollar would buy in the United States. An international dollar is, therefore, a hypothetical currency that is used as a means of translating and comparing costs from one country taking into account differences in purchasing power. The base year is 2000. Details of the assumptions are found in *Annex 2*.

Cost-effectiveness

Average cost-effectiveness ratios are calculated for each intervention by combining the information on the total costs with information on the total health effects in terms of DALYs averted. All costs and effects are discounted

at 3% consistent with the Disease Control Priority Review [29], the first large scale attempt to compare the cost-effectiveness of interventions across diseases, and the recommendations of the US Panel on Cost-Effectiveness in Health and Medicine [30].

Using a standard approach, we have identified the set of interventions a region should purchase to maximize health gain for different budget levels. The order in which interventions would be purchased is called an expansion path and is based on the incremental costs and benefits of each intervention compared to the last intervention purchased.

The Commission on Macroeconomics and Health recently defined interventions that have a cost-effectiveness ratio of less than three times Gross Domestic Product (GDP) per capita as cost-effective [31]. Based on this, three broad categories are defined here. Interventions that gain each year of healthy life (e.g. DALY averted) at a cost less than GDP per capita are defined as very cost-effective. Those averting each DALY at a cost between one and three times GDP per capita are cost-effective, and the remainder are not cost-effective.

The results of cost-effectiveness analysis should not be used formulaically – starting with the intervention with the lowest cost-effectiveness ratio, choosing the next most attractive intervention, and continuing until all resources have been used [10]. The uncertainty surrounding estimates is generally too great to support this approach and, moreover, there are other goals of health policy in addition to improving population health. The tool is most powerful when it is used to classify interventions into broad categories such as those used in this paper. This provides decision-makers with information on which interventions are low cost ways of improving population health, and which improve health at a much higher cost. This information enters the policy debate to be weighed against the impact of the interventions on other goals of health policy.

Sensitivity analysis

Multivariate sensitivity analysis was undertaken to assess the effect of uncertainty in the assumptions on the baseline levels of risks and effect sizes on the costeffectiveness ratios. The first step was to take several samples of hypothetical

individuals from correlated distributions of four risk factors: total cholesterol concentration, systolic blood pressure, smoking, and body-mass index. Then samples were taken from distributions around the population means and standard deviations of the risk factors, as well as around the relative risks and effectiveness estimates from limits developed from the review of relevant publications (*Annex 2 and 3*), producing upper and lower confidence limits on mean incidence. This procedure also includes the effects on costs because different numbers of people will be covered by an intervention under the different scenarios. At the same time, the price of medicines—the most important driver of costs—was allowed to vary from half to double the base estimate.

Results

Table 2 provides the total annualized costs, total annual health effect in terms of DALYs averted and the average cost-effectiveness ratio for each of the 17 interventions in three subregions with differing levels of adult and child mortality and different patterns of risks to health, EurA, AmrB, and SearD. The health benefits of all interventions follow an approximately bell-shaped curve when plotted against age. Depending on the intervention and the region, the curve reaches its maximum at around 60 years of age, with about half of the total intervention benefit occurring at younger ages, and about half at older ages. This is shown for AmrB in *Figure 1*.

All 17 interventions in all three regions are cost-effective according to the Commission on Macroeconomics and Health criterion. In all regions, the four non-personal interventions have cost-effectiveness ratios that are lower than personal health service interventions.

When considered individually, non-personal health interventions to reduce blood pressure and cholesterol are very cost-effective. Measures to reduce salt intake are potentially very cost-effective, with legislation being more cost-effective than voluntary agreements under the assumption that it would lead to the larger reduction in dietary salt intake. The impact of non-personal health service strategies to reduce cholesterol compared to salt reduction strategies depends on the distribution of risk factors in the region – it has a slightly lower impact on population health than legislation to reduce salt in EurA and AmrB, and a substantially higher impact in SearD.

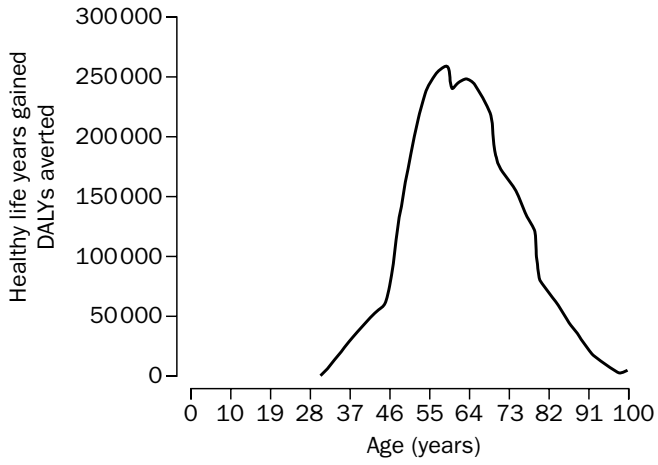
Perhaps surprisingly, personal health service strategies have a much greater potential to reduce the burden of disease even though they are slightly less cost-effective than the population-wide strategies. Treatment of systolic blood pressure above 160 mm Hg falls into the very cost-effective category in all regions. Statins are now available off-patent at very low cost and their use for people with total cholesterol levels above 240 mg/dl (6.2 mmol/L) is also very cost-effective in all regions.

Table 2. Annual costs, effects, and cost-effectiveness of interventions

	AmrB			EurA			SearD		
	Costs ($\$\times10^6$)	DALYs ($\times10^5$)	Cost/DALY ($)	Costs ($\$\times10^6$)	DALYs ($\times10^5$)	Cost/DALY ($)	Costs ($\$\times10^6$)	DALYs ($\times10^5$)	Cost/ DALY ($)
Non-personal									
N1	82	3	24	297	7	44	199	5	37
N2	82	6	13	297	13	23	199	10	19
N3	81	6	14	202	12	16	204	14	14
N4	163	12	14	499	24	21	403	24	17
Personal									
P1	3122	38	81	14777	73	203	1570	44	36
P2	8806	47	186	37385	82	454	6399	71	90
P3	2426	28	87	19187	65	297	2537	54	47
P4	4397	33	133	27142	69	393	4835	69	71
P5	11232	61	183	56572	110	516	8936	107	84
P6	1335	51							
P7	2059	56	37	16015	95	168	2888	88	33
P8	3352	62	54	22226	104	215	5078	106	48
P9	6456	69	93	35750	114	314	10183	133	77
Combined									
C1	1365	54	25	11045	91	121	1829	82	22
C2	2056	58	35	15474	99	157	2994	95	31
C3	3332	64	52	21612	106	203	5149	112	46
C4	6394	71	90	35095	116	302	10173	137	74

*See Table 1 for descriptions of interventions

Figure 1. Total intervention benefit by age, AmrB



However, a comparison of their cost-effectiveness ratios with those of the absolute risk approach shows that treatment based on measured levels of blood pressure or cholesterol alone would not be the preferred option on cost-effectiveness grounds.

The absolute risk approach at a threshold of 35% is always more cost-effective than treatment based on the measured level of either systolic blood pressure or cholesterol. It would avert an additional 65 million DALYs on top of the 21 million DALYs averted by the two non-personal interventions evaluated here.

As the absolute risk threshold is lowered, the health benefits increase but so do the costs – it gets more and more expensive to obtain each additional unit of health benefit. The exact point at which policy makers might choose to set the threshold will vary by setting and will take into account many factors in addition to cost-effectiveness, but it is very cost-effective to reduce the threshold even below 15% in the three regions under consideration, even taking into account the consequences of bleeding associated with the additional use of aspirin.

The cost-effectiveness ratios of the individual interventions do not tell the whole story. *Figures 2(a-c)* plot the annual cost and DALYs averted for each of the 17 interventions in the three regions. The slope of the line connecting the origin to each point is the cost-effectiveness ratio. The steeper the slope the more expensive the intervention is per DALY averted. These figures can also help visualize the incremental cost and incremental health gain of moving from one intervention strategy to another.

From the perspective of how best to maximize population health for the available resources, the optimal overall strategy is a combination of the population-wide and individual-based interventions. The solid lines joining the most cost-effective points in *Figures 2(a-c)* show the optimal choice from a cost-effectiveness perspective. These "expansion paths" join the interventions that would be selected for increasing levels of resource availability. The slopes between them represent the "incremental cost-effectiveness ratio" – or the additional costs required to avert each additional DALY by moving from the

lower to the higher cost intervention. The incremental costs, effects and cost-effectiveness ratios of points on the expansion path are reported in *Table 3*. If resources are extremely scarce, the non-personal interventions would be chosen first.

In all three regions, the expansion path is similar. In settings of extreme resource constraints, one of the non-personal interventions to reduce salt and/or cholesterol would be purchased first. Decision-makers who want to maximize health gain for available resources would next move to a combined strategy of legislated salt reductions in processed foods with mass media campaigns, and then add the absolute risk approach to managing blood pressure and cholesterol. Depending on the resources available, the absolute risk threshold for a cardiovascular event that would trigger intervention with beta blockers, diuretics, statins and aspirin would be lowered. While the total costs, total effects and cost-effectiveness ratios vary considerably across regions, the sequence of intervention strategies that would be purchased is similar.

Table 3. Annual incremental costs, effects, and cost-effectiveness of interventions

	AmrB			EurA			SearD		
	Costs ($\times 10^6$)	DALYs ($\times 10^5$)	Cost/DALY ($)	Costs ($\times 10^6$)	DALYs ($\times 10^5$)	Cost/DALY ($)	Costs ($\times 10^6$)	DALYs ($\times 10^5$)	Cost/DALY ($)
Non-personal									
N2	82	7	13	..	..	..	..	..	..
N3	..	..	..	202	12	16	204	14	14
N2 to N4	81	6	14	..	..	..	..	..	..
N3 to N4	..	..	..	297	12	25	199	10	20
Combined									
N4 to C1	1201	42	29	10546	67	157	1426	58	24
C1 to C2	692	4	160	4428	7	616	1166	13	90
C2 to C3	1276	5	239	6138	8	781	2155	17	129
C3 to C4	7899	7	432	13484	3	1392	5023	6	206

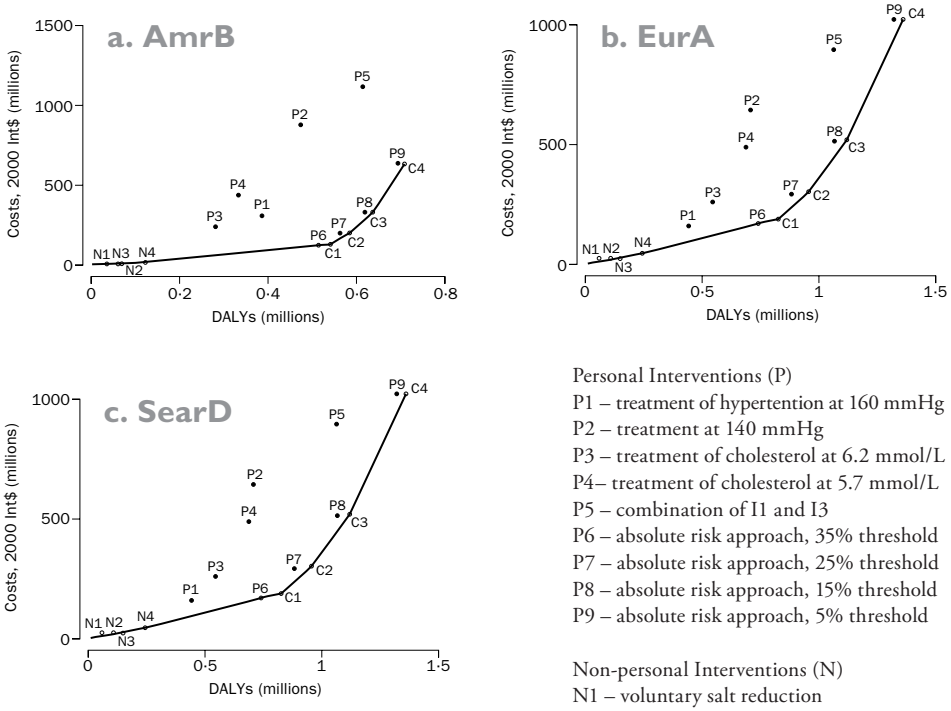
*See Table 7.1 for descriptions of interventions

Figures 2(a-c) also show that the total effect on the burden of disease through the management of absolute risk is substantial. Population-level reduction in cardiovascular disease events of more than 50% is possible. Even in the less developed regions of AmrB and SearD, the absolute magnitude of the changes in the burden of CVD are impressive.

The multivariate uncertainty analysis illustrates that the cost-effectiveness ratios vary – depending on the region, they can be up to 83% higher on average or 53% lower. This variation, though apparently substantial, does

not put any of the interventions classified as very cost-effective into a less desirable category. In addition, the essential features of the expansion paths in the three regions do not change with the changes in assumptions even if the slopes of the segments change somewhat. The order in which the two non-personal health interventions (salt reduction in processed foods and the mass media approach to reducing cholesterol) would be introduced might change, but in all cases one would be chosen as the most desirable option, then the second would be added, before the first personal intervention is considered. In addition, the absolute risk approach is more cost-effective than treating people based on either blood pressure or cholesterol alone in all cases.

Figure 2. Annual Costs and Effectiveness of CVD risk factor interventions



Personal Interventions (P)
P1 – treatment of hypertention at 160 mmHg
P2 – treatment at 140 mmHg
P3 – treatment of cholesterol at 6.2 mmol/L
P4– treatment of cholesterol at 5.7 mmol/L
P5 – combination of I1 and I3
P6 – absolute risk approach, 35% threshold
P7 – absolute risk approach, 25% threshold
P8 – absolute risk approach, 15% threshold
P9 – absolute risk approach, 5% threshold

Non-personal Interventions (N)
N1 – voluntary salt reduction
N2 – legislated salt reduction
N3 – mass media targeting cholesterol
N4 – combination of P2 and P3

Combined Personal and Non-Personal Interventions (C)
C1 – P6 & N4
C2 – P7 & N4
C3 – P8 & N4
C4 – P9 & N4

Figure 3a. Probability of inclusion at 'low' levels of resources availability in EurA

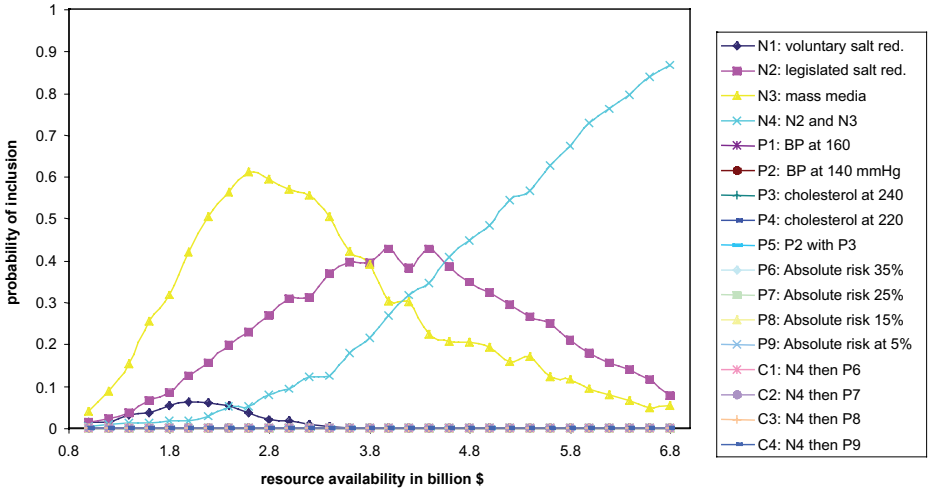


Figure 3b. Probability of inclusion in optimal mix at 'high' levels of resources availability in EurA

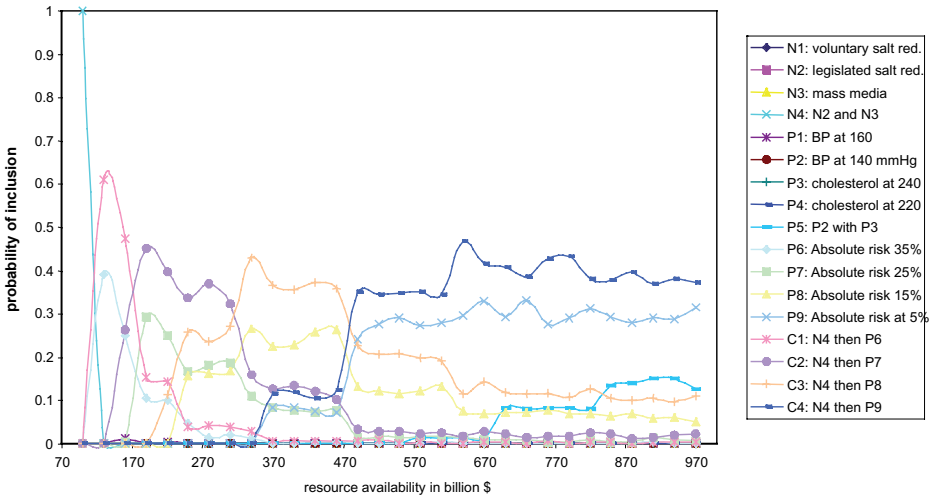


Figure 3 shows the stochastic league diagram for the cardiovascular cluster of interventions for the subregion EurA only as an example when low (*Figure 3a*) and high (*Figure 3b*) levels of resources are available. The vertical axis shows

the probability that an intervention will be included in the optimal package for the level of the resource constraint on the horizontal axis. Estimates of cost and effect are allowed to vary in terms of defined distributions, and, with a Monte Carlo approach, the league table is calculated many times for each postulated level of resource availability. This procedure effectively takes into account uncertainty around the input parameters. The figure shows that the optimal mix will differ depending on the resources available. At relatively low resource levels, population wide interventions are likely the most efficient choice; at higher resource levels, the most efficient choice are combined population wide interventions with the absolute risk interventions. It remains, of course, to be seen if these preventive CVD interventions are cost-effective compared to other ways of using scarce health resources.

Discussion

In all regions, these selected non-personal and personal health interventions to reduce blood pressure and cholesterol are very cost-effective. This result is at odds with the perception that CVD prevention strategies should strictly be the concern of the very wealthy. Implied in these results is a further frameshift in thinking about priorities and public health strategies for developing regions. Even though the benefits documented here are already large, the potential of the non-personal interventions may be even larger. The effect size of these interventions observed in North Karelia, Finland was substantially larger than in the North American demonstration/cluster trials like the Stanford 5-city study [32-34] and the assumptions used here reflect both experiences. With greater understanding of the factors that influence the effectiveness of these non-personal interventions, it could well be possible to develop strategies that increase the population health benefits even further.

The beneficial effects of salt reduction have been subject to debate in the recent literature with one review suggesting that the impact of salt reduction on blood pressure was minimal [35]. That review focused on interventions involving individual dietary advice to reduce salt intake rather than the option to reduce the salt content of processed food as assessed here. Indeed, it concluded that reduction in salt intake population-wide through cutting salt concentrations in processed foods might achieve small reductions in blood pressure across the whole population for sustained periods of time, which

would then have substantial health effects at the populaton level. This was the motivation for the intervention analysed here. In additon, there is evidence that small and repeated reductions in salt intake are not discernable on taste, (eg less salt does not necessarily mean less taste [36]) so that people are not likely to resist the new foods for that reason.

The absolute risk approach to managing blood pressure and cholesterol is very cost-effective in all regions, and has the potential to lead to dramatic reductions in ischaemic heart disease and stroke. Many other combinations of medicines are likely to be as cost-effective as those evaluated here. A meta-analysis of 354 trials, involving 56,000 participants, showed the blood pressure reductions produced by the major classes of drug at standard dose are similar, independent and additive, and that using half standard dose reduces efficacy only by 20% while more than halving side-effects [27,37,38]. There are also probable or proven benefits of these interventions on other important outcomes not measured here, such as dementia, renal failure, peripheral vascular disease, congestive heart failure and the need for coronary artery bypass grafting [39-43]. Additionally, while the effects of blood pressure and cholesterol lowering drugs appear to be due largely to the amount of risk factor reduction achieved [24,25], there may be some additional benefits due to specific agents, such as coronary disease risk reduction with ACE inhibitors [43]. Implementation of risk screening can and should be tailored to the resource levels of national health systems. In high-income countries, risk assessment on the basis of age, sex, measured blood pressure, cholesterol, body mass index, diabetes, tobacco use and clinical history of previous CVD events is practical. In low-resource settings, however, adequate risk screening could be based solely on age, sex, measured blood pressure, body mass index, tobacco use and past CVD events. This would require no sophisticated technology or blood sampling. A 'risk pill' of anti-hypertensives, statin and aspirin could also be packaged as a single compound facilitating compliance.

As the absolute risk threshold used to trigger treatment is lowered, larger and larger fractions of the adult population would be on long-term drug treatment and the number of adverse events would increase. The consequences of this medicalization of potentially the majority of the adult population should be carefully considered. Issues of long-term compliance may also limit the

applicability of the approach in certain populations, including younger people. The potential huge benefits and the apparent cost-effectiveness of the absolute risk approach do seem to justify some large-scale population effectiveness studies. States or provinces in countries facing major CVD challenges could be enrolled to see if the expected population benefits can be achieved in the short-time frame implied by the analysis.

It is well established that in developed countries, levels of blood pressure and cholesterol tend to be worse in the poor than in the rich [44]. Unfortunately, knowledge on how to manage these risks is used more effectively by the higher income, more educated population groups. Consequently, the coverage of interventions to reduce blood pressure and cholesterol is probably lower in the poor. Because of the distribution of these risks, there is a potential for both non-personal and the absolute risk approaches to substantially contribute to the reduction in adult health inqualities. It is a challenge for public health to develop innovative strategies to encourage the uptake of the latter in the poor and disadvantaged [45]. 'Out of the box" thinking may be needed. Studies show that intervention uptake can be affected by financial incentives. Perhaps, lottery tickets should be given to those that reduce their absolute risk by a certain amount in a year.

Why is this analysis apparently suggesting a much bigger impact at lower cost for personal health service interventions to manage blood pressure and cholesterol than many have been expected? First, as part of the Comparative Risk Analysis module of the Global Burden of Disease 2000 project, a clearer picture of the burden of these risk factors worldwide has emerged [2]. Second, new ways of using existing drugs such as the absolute risk method have been developed. Third, lovastain is now off-patent and other statins will follow soon, substantially reducing the cost of these regimens. Fourth, developments in the analysis of hazard data to deal with the effect of measurement error and regression dilution bias [21,46-48] have led to a nearly doubling of the estimated impact of reductions in blood pressure and cholesterol on outcomes. These changes remind us why it will always be important to update and reevaluate strategies that address major public health problems.

The non-personal interventions considered here were even more cost-effective than the personal interventions in the three regions despite having a lower overall impact on population health. Care should be taken in interpreting these results

because the estimates of changes resulting from the mass media intervention were based on changes in behaviour observed in a developed country setting, but even with a halving of the assumed effectiveness, this conclusion is not altered. Based on this assumption, non-personal interventions would be be the first to be introduced. Moreover, the non-personal interventions assessed here are only a selection of those that are possible, and their very nature makes reliable assessment of effects challenging. But this challenge must be accepted. For example, assessing strategies to achieve moderate but widespread changes in manufactured food (e.g. in overall fat content) would be very worthwhile, since sub-optimal cholesterol and blood pressure levels have major dietary components to their aetiology.

Three final considerations are pertinent to the policy debate to which this paper contributes. First, the combination of medicines that prevent people at high risk of cardiovascular disease from having an event would cost just less than $14 per person per year if the lowest cost medicines were purchased internationally. This is simply the costs of the medicines and does not include distribution mark-ups. The ability of poor countries to finance this, and all the other possible cost-effective interventions, from their own resources is limited – some countries spend less than $10 per capita on health each year. The availability of low cost, effective ways to improve health in all settings, many of which are not affordable at current levels of health expenditure, is why WHO has argued strongly for massive injections of resources for health from richer countries that could be used to reduce the burden of disease among the poor.

Second, this paper has focused on reducing the health consequences associated with cholesterol and blood pressure. It shows which interventions should be given priority when developing a strategy for the control of cardiovascular disease. This is important information for policy-makers responsible for cardiovascular disease control or health promotion. It does not, however, indicate whether controlling cardiovascular disease should receive priority over reducing the risks associated with unsafe sex, for example. This requires consideration of the costs and effects of all possible alternatives. WHO seeks to provide this information through its CHOICE project, and initial results covering several major risks to health can be found in the World Health Report 2002 [1].

Third, cost-effectiveness is only one of the key inputs to final decision about how to allocate scarce resources. Policy-makers have other concerns as well, such as reducing poverty and inequalities, and questions of human rights and community acceptance also influence policy. Another key concern is how different types of interventions can be incorporated into the health infrastructure of the country, or how the infrastructure could be adapted to accommodate the desired strategies. The information presented here is one, but only one, of the critical inputs required to inform the decision making process about efficient ways to reduce risks to health.

Acknowledgements

We thank Ilja Borysenko, Brodie Ferguson, Margaret Hogan, and Margaret Squadrani for assistance with calculations, figures, and formatting; Taghreed Adam, Julia Lowe, Ken Redekop, Joshua Salomon, Kenji Shibuya, Tessa Tan Torres, Steve Vander Hoorn, Rod Jackson and staff of the Non-communicable Disease and Mental Health Cluster of WHO for input on assumptions and comments of various drafts.

References

1. World Health Organization. *The World Health Report 2002: reducing risks, promoting healthy life*. Geneva: WHO, 2002.

2. Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJL. Comparative Risk Assessment Collaborating Group. Selected major risk factors and global and regional burden of disease. *Lancet* 2002; **360:** 1347–60.

3. Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, 1st edn. Cambridge, MA: Harvard University Press, 1996.

4. Puska P. Development of public policy on the prevention and control of elevated blood cholesterol. *Cardiovasc Risk Factors* 1996; **6:** 203–10.

5. World Health Organization. Innovative care for chronic conditions: building blocks for action. Geneva: WHO, document No. WHO/MNC/CCH/02.01: 2002.

6. Johns B, Baltussen R, Hutubessy RCW. Programme costs in the economic evaluation of health interventions. *Cost Eff Resour Alloc* 2003; **1**(1): 1.

7. Adam T, Evans DB, Koopmanschap MA. Cost-effectiveness analysis: can we reduce variability in costing methods? *Int J Technol Assess Health Care* 2003; **19**(2): 407–20.

8. Baltussen RM, Adam T, Tan Torres T, et al. *Generalized costeffectiveness analysis: a guide*. Geneva: WHO, Global Programme on Evidence for Health Policy: 2002. Available at: http://www.who.int/evidence/cea

9. Adam T, Evans DB, Murray CJL. Econometric estimation of country-specific hospital costs. *Cost Eff Resour Alloc* 2003; **1**(1): 3.

10. Murray CJ, Evans DB, Acharya A, Baltussen RM. Development of WHO guidelines on generalized cost-effectiveness analysis. *Health Econ* 2000; **9:** 235–51.

11. Hutubessy RC, Baltussen RM, Evans DB, Barendregt JJ, Murray CJ. Stochastic league tables: communicating cost-effectiveness results to decision-makers. *Health Econ* 2001; **10:** 473–77.

12. Hutubessy RCW, Baltussen RMPM, Tan Torres-Edejer T, Evans DB. Generalised cost-effectiveness analysis: an aid to decision making in health. *Appl Health Econ Health Policy* 2002; **1:** 89–95.

13. Baltussen RM, Hutubessy RC, Evans DB, Murray CJ. Uncertainty in cost-effectiveness analysis. Probabilistic uncertainty analysis and stochastic league tables. *Int J Technol Assess Health Care* 2002; **18:** 112–19.

14. Martin I. Implementation of WHO/ISH Guidelines: role and activities of WHO. *Clin Exp Hypertens* 1999; **21:** 659–69.

15. Law MR, Wald NJ. Risk factor thresholds: their existence under scrutiny. *BMJ* 2002; **324:** 1570–76.

16. Cooper RS, Rotimi CN, Kaufman JS, Muna WF, Mensah GA. Hypertension treatment and control in sub-Saharan Africa: the epidemiological basis for policy. *BMJ* 1998; **316:** 614–17.

17. Marshall T, Rouse A. Resource implications and health benefits of primary prevention strategies for cardiovascular disease in people aged 30 to 74: mathematical modelling study. *BMJ* 2002; **325:** 197–99.

18. Ebrahim S, Smith GD. Exporting failure? Coronary heart disease and stroke in developing countries. *Int J Epidemiol* 2001; **30:** 201–05.

19. Anderson KM, Odell PM, Wilson PW, Kannel WB. Cardiovascular disease risk profiles. *Am Heart J* 1991; **121:** 293–98.

20. Jackson R, Barham P, Bills J, et al. Management of raised blood pressure in New Zealand: a discussion document. *BMJ* 1993; **307:** 107–10.

21. Neaton JD, Wentworth D. Serum cholesterol, blood pressure, cigarette smoking, and death from coronary heart disease: overall findings and differences by age for 316,099 white men. *Arch Intern Med* 1992; **152:** 56–64.

22. Eastern Stroke and Coronary Heart Disease Collaborative Research Group. Blood pressure, cholesterol, and stroke in eastern Asia. *Lancet* 1998; **352:** 1801–07.

23. Jee SH, Suh I, Kim IS, Appel LJ. Smoking and atherosclerotic cardiovascular disease in men with low levels of serum cholesterol: the Korea Medical Insurance Corporation Study. *JAMA* 1999; **282:** 2149–55.

24. Lawes CMM, Vander Hoorn S, Law MR, Elliott P, Rodgers A. High blood pressure. In: Ezzati M, Lopez AD, Rodgers A, Murray CJL, eds. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*. Geneva: WHO, 2004.

25. Lawes CMM, Vander Hoorn S, Law MR, Rodgers A. High cholesterol. In: Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJL, eds. *Comparative quantification of health risks: global and regional burden of disease attributable to selected major risk factors*. Geneva: WHO, 2004.

26. Lambert JD. *Numerical methods for ordinary differential systems: the initial value problem*. Chichester: Wiley, 1991.

27. Law MR, Wald NJ, Thompson SG. By how much and how quickly does reduction in serum cholesterol concentration lower risk of ischaemic heart disease? *BMJ* 1994; **308:** 367–72.

28. Prospective studies collaboration. Age-specfic relevance of usual blood pressure to vascular mortality: a meta-analysis of individual data for one million adults in 61 prospective studies. *Lancet* 2002; **360:** 1903–13.

29. Jamison DT, Mosley WH, Measham AR, Bobadilla JL. *Disease control priorities in developing countries*. New York: Oxford University Press, 1993.

30. Gold MR, Siegel JE, Russel LB, Weinstein MC. *Cost-effectiveness in health and medicine*. New York: Oxford University Press, 1996.

31. WHO Commission on Macroeconomics and Health. *Macroeconomics and health: investing in health for economic development. Report of the Commission on Macroeconomics and Health*. Geneva: WHO, 2001.

32. Fortmann SP, Varady AN. Effects of a community-wide health education program on cardiovascular disease morbidity and mortality: the Stanford Five-City Project. *Am J Epidemiol* 2000; **152:** 316–23.

33. Vartiainen E, Puska P, Jousilahti P, Korhonen HJ. Cardiovascular diseases and risk factors in Finland. *Prev Med* 1999; **29:** S124–29.

34. Sellers DE, Crawford SL, Bullock K, McKinlay JB. Understanding the variability in the effectiveness of community heart health programs: a meta-analysis. *Soc Sci Med* 1997; **44:** 1325–39.

35. Hooper L, Bartlett C, Davey SG, Ebrahim S. Systematic review of long term effects of advice to reduce dietary salt in adults. *BMJ* 2002; **325:** 628.

36. Rodgers A, Neal B. Less salt does not necessarily mean less taste. *Lancet* 1999; **353:** 1332.

37. Law MR, Wald NJ, Morris JK, Jordan RE. Value of low dose combination with blood pressure lowering drugs: analysis of 354 randomised trials. BMJ 2003; **326**(7404): 1427–34.

38. Sacks FM, Tonkin AM, Shepherd J, et al. Effecct of pravastatin on coronary disease events in subgroups defined by coronary risk factorsthe prospective pravastatin pooling project. *Circulation* 2000; **102:** 1893–1900.

39. Richard F, Fromentin-David I, Ricolfi F, et al. The angiotensin I converting enzyme gene as a susceptibility factor for dementia. *Neurology* 2001; **56:** 1593–95.

40. Collins R, Armitage J, Parish S, Sleight P, Peto R. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002; **360:** 7–22.

41. Tzourio C, Anderson C, Chapman N, Woodward M, Neal B, MacMahon S, Chalmers J; PROGRESS Collaborative Group. Effects of blood pressure lowering with perindopril and indapamide therapy on dementia and cognitive decline in patients with cerebrovascular disease. *Arch Intern Med* 2003; **163**(9):1069–75.

42. Forette F, Seux ML, Staessen JA, et al. Prevention of dementia in randomised double-blind placebo-controlled Systolic Hypertension in Europe (Syst-Eur) trial. *Lancet* 1998; **352:** 1347–51.

43. Neal B, MacMahon S, Chapman N. Effects of ACE inhibitors, calcium antagonists, and other blood-pressure-lowering drugs: results of prospectively designed overviews of randomised trials. *Lancet* 2000; **356:** 1955–64.

44. Marmot M, Shipley M, Brunner E, Hemingway H. Relative contribution of early life and adult socioeconomic factors to adult morbidity in the Whitehall II study. *J Epidemiol Community Health* 2001; **55:** 301–07.

45. Haynes RB, McKibbon KA, Kanani R. Systematic review of randomised trials of interventions to assist patients to follow prescriptions for medications. *Lancet* 1996; **348:** 383–86.

46. MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease: part 1, prolonged differences in blood pressure: prospective observational studies corrected for the regression dilution bias. *Lancet* 1990; **335:** 765–74.

47. Prospective studies collaboration. Cholesterol, diastolic blood pressure, and stroke: 13 000 strokes in 450 000 people in 45 prospective cohorts. *Lancet* 1995; **346:** 1647–53.

48. Lawes CM, Rodgers A, Bennett DA, Parag V, Suh I, Ueshima H, MacMahon S; Asia Pacific Cohort Studies Collaboration. Blood pressure and cardiovascular disease in the Asia Pacific region. *J Hypertens* 2003; **21**(4): 707–16.

Annex

Annex Table 1. Epidemiologic subregions

Region*	Mortality stratum**	Countries
Afr	D	Algeria, Angola, Benin, Burkina Faso, Cameroon, Cape Verde, Chad, Comoros, Equatorial Guinea, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Liberia, Madagascar, Mali, Mauritania, Mauritius, Niger, Nigeria, Sao Tome and Principe, Senegal, Seychelles, Sierra Leone, Togo
Afr	E	Botswana, Burundi, Central African Republic, Congo, Côte d'Ivoire, Democratic Republic Of The Congo, Eritrea, Ethiopia, Kenya, Lesotho, Malawi, Mozambique, Namibia, Rwanda, South Africa, Swaziland, Uganda, United Republic of Tanzania, Zambia, Zimbabwe
Amr	A	Canada, United States of America, Cuba
Amr	B	Antigua and Barbuda, Argentina, Bahamas, Barbados, Belize, Brazil, Chile, Colombia, Costa Rica, Dominica, Dominican Republic, El Salvador, Grenada, Guyana, Honduras, Jamaica, Mexico, Panama, Paraguay, Saint Kitts And Nevis, Saint Lucia, Saint Vincent and the Grenadines, Suriname, Trinidad And Tobago, Uruguay, Venezuela
Amr	D	Bolivia, Ecuador, Guatemala, Haiti, Nicaragua, Peru
Emr	B	Bahrain, Cyprus, Iran (Islamic Republic of), Jordan, Kuwait, Lebanon, Libyan Arab Jamahiriya, Oman, Qatar, Saudi Arabia, Syrian Arab Republic, Tunisia, United Arab Emirates
Emr	D	Afghanistan, Djibouti, Egypt, Iraq, Morocco, Pakistan, Somalia, Sudan, Yemen
Eur	A	Andorra, Austria, Belgium, Croatia, Czech Republic, Denmark, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Luxembourg, Malta, Monaco, Netherlands, Norway, Portugal, San Marino, Slovenia, Spain, Sweden, Switzerland, United Kingdom
Eur	B	Albania, Armenia, Azerbaijan, Bosnia and Herzegovina, Bulgaria, Georgia, Kyrgyzstan, Poland, Romania, Slovakia, Tajikistan, The Former Yugoslav Republic of Macedonia, Turkey, Turkmenistan, Uzbekistan, Yugoslavia
Eur	C	Belarus, Estonia, Hungary, Kazakhstan, Latvia, Lithuania, Republic of Moldova, Russian Federation, Ukraine
Sear	B	Indonesia, Sri Lanka, Thailand
Sear	D	Bangladesh, Bhutan, Democratic People's Republic of Korea, India, Maldives, Myanmar, Nepal
Wpr	A	Australia, Japan, Brunei Darussalam, New Zealand, Singapore
Wpr	B	Cambodia, China, Lao People's Democratic Republic, Malaysia, Mongolia, Philippines, Republic of Korea, Viet Nam, Cook Islands, Fiji, Kiribati, Marshall Islands, Micronesia (Federated States of), Nauru, Niue, Palau, Papua New Guinea, Samoa, Solomon Islands, Tonga, Tuvalu, Vanuatu

* Afr = African Region; Amr = Region of the Americas; Emr = Eastern Mediterranean Region; Eur = European Region; Sear = South East Asian Region; Wpr = Western Pacific Region

** A subregions have very low rates of adult and child mortality; B = low adult, low child; C = high adult, low child; D = high adult, high child; E = very high adult, high child mortality

Annex Table 2. Assumptions on effectiveness and costs for non-personal interventions

Model variables		Assumptions	Sources
Basic epidemiology	Ischaemic heart disease	Acute myocardial infarction; angina pectoris; congestive heart failure.	Murray and Lopez, 1996 [1]
	Cerebovascular disease	First-ever fatal stroke cases; long-term stroke survivors.	Murray and Lopez, 1996 [1]
Risk factor and epidemiology	Blood pressure	Region-, age- , and sex-specific systolic blood pressure levels; age-specific relative risks of CVD event for 1mm Hg change in systolic blood pressure (see Annex 4).	World Health Organization, 2002 [2]
	Cholesterol	Region-, age- , and sex-specific cholesterol levels; age-specific relative risks of CVD event for 1mmol/l change in total blood cholesterol (see Annex 4).	
	Body mass index	Region-, age- , and sex-specific body mass index levels age-specific relative risks of CVD event for 1 unit change in body mass index (see Annex 4).	
	Smoking	Region-, age- , and sex-specific prevalence of long-term smokers; age-specific relative risks of CVD event for unit change in prevalence of long-term smokers (see Annex 4).	
Programme level costs	Intervention N1	Central administration and planning costs at 95% coverage.	Johns,2002 [3]; Adam et al., 2002 [4]
	Intervention N2	Central administration, planning and enforcement costs at 95% coverage.	
	Intervention N3	Central administration costs, planning, media costs and printed materials at 80% coverage.	
Effectiveness	Intervention N1	Region- age- and sex-specific blood pressure changes associated with a 15% reduction in total dietary salt intake (see Annex 5).	Law et al., 1991 [5]
	Intervention N2	Region- age- and sex-specific blood pressure changes associated with a 30% reduction in total dietary salt intake (see Annex 5).	Frost et al., 1991 [6] Law et al., 1991 [7]
	Intervention N3	2% reduction in total blood cholesterol levels.	Lawes et al., 2002 [8]
	Intervention N4	Combined effect of interventions N2 and N3.	Tosteson et al., 1997 [9]

1. Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, 1st edn. Cambridge, MA: Harvard University Press, 1996.
2. World Health Organization. The World Health Report 2002: reducing risks, promoting healthy life. Geneva: WHO, 2002.
3. Johns B, Baltussen R, Hutubessy RCW. Programme costs in the economic evaluation of health interventions. Cost Eff Resour Alloc 2003; 1(1): 1
4. Adam T, Evans DB, Murray CJL. Econometric estimation of country-specific hospital costs. Cost Eff Resour Alloc 2003; 1(1): 3.
5. Law MR, Frost CD, Wald NJ. By how much does dietary salt reduction lower blood pressure? I Analysis of observational data among populations. BMJ 1991; 302: 811–15.
6. Frost CD, Law MR, Wald NJ. By how much does dietary salt reduction lower blood pressure? II Analysis of observational data within populations. BMJ 1991; 302: 815–18.
7. Law MR, Frost CD, Wald NJ. By how much does dietary salt reduction lower blood pressure? III Analysis of data from trials of salt reduction. BMJ 1991; 302: 819–24.
8. Lawes C, Feigin V, Rodgers A. Estimating reductions in blood pressure following reductions in salt intake by age, sex and WHO region. Auckland: Clinical Trials Research Unit, University of Auckland, 2002.
9. Tosteson AN, Weinstein MC, Hunink MG, et al. Cost-effectiveness of populationwide educational approaches to reduce serum cholesterol levels. Circulation 1997; 95: 24–30.

Annex Table 3. Assumption for personal and combined interventions

Model variables		Assumption	Sources
Basic epidemiology	Ischaemic heart disease	Acute myocardial infarction; angina pectoris; congestive heart failure.	Murray, Lopez, 1996 [1]
	Cerebrovascular disease	First-ever fatal stroke cases; long-term stroke survivors.	
	Upper gastrointestinal bleeding	Adverse effect of anti-platelet therapy with low dose aspirin.	Hernandez-Diaz, Rodriguez, 2002 [2]
Risk factor and epidemiology	Blood pressure	Region-, age- , and sex-specific systolic blood pressure levels; global age-specific relative risks of CVD event for 1mm Hg change in systolic blood pressure (Annex 7.3).	World Health Organization, 2002 [3]
	Cholesterol	Region-, age- , and sex-specific cholesterol levels; global age-specific relative risks of CVD event for 1mmol/l change in total blood cholesterol (Annex 4).	
	Body mass index	Region-, age- , and sex-specific body mass index levels; global age-specific relative risks of CVD event for 1 unit change in body mass index (Annex 4).	
	Smoking	Region-, age- , and sex-specific prevalence of long-term smokers; global age-specific relative risks of CVD event for unit change in prevalence of long-term smokers (Annex 4).	
	Intervention coverage	Coverage of antihypertensive drug treatment among respondents aware of high blood pressure. Coverage of cholesterol-lowering drug treatment among respondents aware of high cholesterol.	Molarius, Tuomilehto, et al, 1998; Tolonen, Kullasmaa et al, 1999; Tolonen, Ferrario et al, 1999 [4-6]
Patient level costs	Intervention P1 & P2	Drug treatment: 50mg/day atenolol (beta-blocker) ; 25mg/day hydrochlorothiazide (diuretic). Provider visits: 4 health-care provider visits/year; 1.5 outpatient visits/year for health education. Laboratory tests: annual renal function, lipid profile and blood sugar tests.	Unit prices of health facilities are based on Adam et al. [7]
	Intervention P3 & P4	Drug treatment: 30mg/day lovastatin. Provider visits: 4 health-care provider visits/year; 1.5 outpatient visits/year for health education. Laboratory tests: total cholesterol and hepatic function.	Unit price drugs are based on the International Drug Price indicator by Management Science for Health (www.erc.msh.org)
	Intervention P5	Drug treatment: 50mg/day atenolol (beta-blocker) ; 25mg/day hydrochlorothiazide (diuretic) and 30mg/day lovastatin. Provider visits: 4 health-care provider visits/year; 1.5 outpatient visits/year for health education. Laboratory tests: annual renal function, lipid profile and blood sugar tests; total cholesterol and hepatic function.	Unit prices laboratory tests are based on Arizona Health Care Cost Containment System (AHCCCS) Fee Schedule (www.state.me.us/wcb/departments/omrs/1999%20MFS/radiology.htm)
	Intervention P6-P9	Drug treatment: 50mg/day atenolol (beta-blocker) ; 25mg/day hydrochlorothiazide (diuretic), 30mg/day lovastatin and 100 mg/day acetylsalicylic acid (aspirin). Provider visits: 4 health-care provider visits/year; 1.5 outpatient visits/year for health education. Laboratory tests: annual renal function, lipid profile and blood sugar tests; total cholesterol and hepatic function.	Hay, Lyubashevsky, et al, 1996 [8]
	Upper gastrointestinal bleeding	Secondary level hospital stay of 2.7 days for GBD non-A sub regions and tertiary level hospital stay of 4.8 days for GBD A sub regions.	
Effectiveness	Intervention P1 & P2	33% reduction of the difference between actual systolic blood pressure and 115 mm Hg.	References [9-24]
	Intervention P3 & P4	20% reduction in total blood cholesterol. Combined effect of P2 and P3.	Collins, Armitage et al, 2002 [25] Anon [26]
	ntervention P5	Combined effect of P2 and P3 with additional 20% reduction of absolute	
	Intervention P6-P9	risk for antiplatelet therapy. Effects of N4 evaluated first. Then the reduced number of people at risk of a	
	Intervention C1-C4	Cardiovascular event are subject to the same costs and effects as with P6-9	

Only for GBD A sub regions

1. Murray CJL, Lopez AD. *The global burden of disease: a comprehensive assessment of mortality and disability from diseases, injuries, and risk factors in 1990 and projected to 2020*, 1st edn. Cambridge, MA: Harvard University Press, 1996.
2. Hernandez-Diaz S, Rodriguez LA. Incidence of serious upper gastrointestinal bleeding/perforation in the general population: review of epidemiologic studies. *J Clin Epidemiol* 2002; 55: 157–63.
3. World Health Organization. *The World Health Report 2002: reducing risks, promoting healthy life*. Geneva: WHO, 2002.
4. Molarius A., Tuomilehto J, Kuulasmaa K, for the WHO MONICA Project. *Quality assessment of data on hypertension control in the WHO MONICA Project*. Geneva: WHO, 1998 (available at www.ktl.fi/publications/monica/hyperten/hbpdrug.htm).
5. Tolonen H, Kuulasmaa K, for the WHO MONICA Project. *Quality assessment of data on use of aspirin in the WHO MONICA Project*. Geneva: WHO, 1999 (available at www.ktl.fi/publications/monica/aspirin/aspirinqa.htm).
6. Tolonen H, Ferrario M, Minoja M, for the WHO MONICA Project. *Quality assessment of data on awareness and treatment of high cholesterol in the WHO MONICA Project*. Geneva: WHO, 1999 (available at www.ktl.fi/publications/monica/hich/hchdrug.htm).
7. Adam T, Evans DB, Murray CJL. Econometric estimation of country-specific hospital costs. *Cost Eff Resour Alloc* 2003; 1(1): 3.
8. Hay JA, Lyubashevsky E, Elashoff J, Maldonado L, Weingarten SR, Ellrodt AG. Upper gastrointestinal hemorrhage clinical—guideline determining the optimal hospital length of stay. *Am J Med* 1996; 100: 313–22.
9. Coope J, Warrender TS. Randomised trial of treatment of hypertension in elderly patients in primary care. *BMJ* 1986; 293: 1145–51.
10. Amery A, Birkenhager W, Brixko P, et al. Mortality and morbidity results from the European Working Party on high blood pressure in the elderly trial. *Lancet* 1985; 1: 1349–54.
11. Hypertension-Stroke Cooperative Study Group. Effect of antihypertensive treatment on stroke recurrence. *JAMA* 1974; 229: 409–18.
12. MRC Working Party. Medical Research Council trial of treatment of hypertension in older adults: principal results. *BMJ* 1992; 304: 405–12.
13. Medical Research Council Working Party. MRC trial of treatment of mild hypertension: principal results. *BMJ* 1985; 291: 97–104.
14. Helgeland A. Treatment of mild hypertension: a five year controlled drug trial: the Oslo study. *Am J Med* 1980; 69: 725–32.
15. Leren P, Helgeland A. Oslo hypertension study. *Drugs* 1986; 31 (suppl 1): 41–45.
16. SHEP Cooperative Research Group. Prevention of stroke by antihypertensive drug treatment in older persons with isolated systolic hypertension: final results of the Systolic Hypertension in the Elderly Program (SHEP). *JAMA* 1991; 265: 3255–64.
17. Anon. Effects of treatment on morbidity in hypertension: results in patients with diastolic blood pressures averaging 115 through 129 mm Hg. *JAMA* 1967; 202: 1028–34.
18. Anon. Effects of treatment on morbidity in hypertension: II, results in patients with diastolic blood pressure averaging 90 through 114 mm Hg. *JAMA* 1970; 213: 1143–52.
19. Yusuf S, Sleight P, Pogue J, Bosch J, Davies R, Dagenais G. Effects of an angiotensin-converting-enzyme inhibitor, ramipril, on cardiovascular events in high-risk patients. *N Engl J Med* 2000; 342: 145–53.
20. MacMahon S, Sharpe N, Gamble G, et al. Randomized, placebo-controlled trial of the angiotensin-converting enzyme inhibitor, ramipril, in patients with coronary or other occlusive arterial disease. *J Am Coll Cardiol* 2000; 36: 438–43.
21. Progress Collaborative Group. Randomised trial of a perindopril-based bloodpressure-lowering regimen among 6,105 individuals with previous stroke or transient ischaemic attack. *Lancet* 2001; 358: 1033–41.
22. Teo KK, Burton JR, Buller CE, et al. Long-term effects of cholesterol lowering and angiotensin-converting enzyme inhibition on coronary atherosclerosis: the Simvastatin/Enalapril Coronary Atherosclerosis Trial (SCAT). *Circulation* 2000; 102: 1748–54.
23. Pitt B, Byington RP, Furberg CD, et al. Effect of amlodipine on the progression of atherosclerosis and the occurrence of clinical events. PREVENT Investigators. *Circulation* 2000; 102: 1503–10.
24. Staessen JA, Fagard R, Thijs L, et al. Randomised double-blind comparison of placebo and active treatment for older patients with isolated systolic hypertension. *Lancet* 1997; 350: 757–64.
25. Collins R, Armitage J, Parish S, Sleight P, Peto R. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20 536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002; 360: 7–22.
26. Antithrombotic Trialists' Collaboration. Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients. *BMJ* 2002; 324: 71–86.

Annex Table 4. Relative risks of CVD events for unit changes in systolic blood pressure, total blood cholesterol, body mass index levels and prevalence of long-term smokers.

	Relative risks for ischaemic heart disease				
	Age groups				
	30-44	45-59	60-69	70-79	80+
Systolic blood pressure (mm Hg)	1.07	1.05	1.03	1.02	1.01
Total blood cholesterol (mmol/l)	3.65	2.08	1.55	1.42	1.42
Body mass index (kg/m^2)	1.11	1.09	1.05	1.04	1.03
	Relative risks for stroke				
	Age groups				
	30-44	45-59	60-69	70-79	80+
Systolic blood pressure(mm Hg)	1.09	1.07	1.05	1.03	1.02
Total blood cholesterol(mmol/l)	1.48	1.35	1.25	1.17	1.09
Body mass index (kg/m^2)	1.19	1.09	1.06	1.06	1.02
	Relative risks for smoking and cardiovascular death				
	Age groups				
	30-44	45-59	60-69	70-79	80+
Male	2.43	2.43	1.84	1.70	1.38
Female	2.18	2.18	2.12	1.70	1.31

Annex Table 5. Percentage reduction in systolic blood pressure for 15% and 30% salt reduction intake in three regions.

	AmrB				EurA				SearD			
	15% reduction		30% reduction		15% reduction		30% reduction		15% reduction		30% reduction	
	Male	female	male	Female	male	Female	Male	female	male	female	male	female
30-44	-1.1%	-0.8%	-2.2%	-1.6%	-1.3%	-0.9%	-2.6%	-1.9%	-1.1%	-1.0%	-2.1%	-1.9%
45-59	-1.3%	-1.1%	-2.7%	-2.2%	-1.6%	-1.3%	-3.1%	-2.6%	-1.4%	-1.2%	-2.8%	-2.5%
60-69	-1.7%	-1.5%	-3.4%	-2.9%	-1.9%	-1.6%	-3.8%	-3.3%	-1.9%	-1.6%	-3.7%	-3.3%
70-79	-2.0%	-1.7%	-4.0%	-3.4%	-2.2%	-1.9%	-4.4%	-3.7%	-2.2%	-2.0%	-4.5%	-3.9%
80+	-2.4%	-2.0%	-4.8%	-3.9%	-2.6%	-2.1%	-5.2%	-4.3%	-2.7%	-2.4%	-5.4%	-4.8%

These estimates were made by applying the relationship between sodium intake and blood pressure as estimated by Law et al [1-3] to the WHO age, sex and region subgroups [4], there have been no re-analyses of the strength of this association in this paper.

1. Law MR, Frost CD, Wald NJ. By how much does dietary salt reduction lower blood pressure? I Analysis of observational data among populations. *BMJ* 1991; 302: 811–15.
2. Frost CD, Law MR, Wald NJ. By how much does dietary salt reduction lower blood pressure? II Analysis of observational data within populations. *BMJ* 1991; 302: 815–18.
3. Law MR, Frost CD, Wald NJ. By how much does dietary salt reduction lower blood pressure? III Analysis of data from trials of salt reduction. *BMJ* 1991; 302: 819–24.
4. Lawes C, Feigin V, Rodgers A. Estimating reductions in blood pressure following reductions in salt intake by age, sex and WHO region. Auckland: Clinical Trials Research Unit, University of Auckland, 2002.

Chapter 4

Achieving the millennium development goals for health: cost effectiveness analysis of strategies to combat malaria in developing countries

Based on: Morel CM, Lauer JA, Evans DB. Achieving The millennium development goals for health: Cost effectiveness analysis of strategies to combat malaria in developing countries. *BMJ*, 2005, 331:1299–1302.

Abstract

Objective To determine the cost effectiveness of selected malaria control interventions in the context of reaching the millennium development goals for malaria.

Design Generalised cost effectiveness analysis.

Data sources Efficacy data came from the literature and authors' calculations supported by expert opinion. Quantities for resource inputs came from the literature and from expert opinion; prices came from the WHO-CHOICE database.

Methods Costs were assessed in year 2000 international dollars and effects were assessed as disability adjusted life years averted by a 10 year implementation programme. Analysis was restricted to sub-Saharan regions where the most deadly form of malaria, due to resistant *Plasmodium falciparum*, is most prevalent. The impact on population health for various interventions and their combinations was evaluated at various coverage levels by using a state-transition model. Sensitivity analysis was done for age weights and discounting.

Results High coverage with artemisinin based combination treatments was found to be the most cost effective strategy for control of malaria in most countries in sub-Saharan Africa.

Conclusions A much larger infusion of resources than those currently available is needed to make headway in the fight to roll back malaria. On cost effectiveness grounds, in most areas in sub-Saharan Africa greater coverage with highly effective combination treatments should be the cornerstone of malaria control. However, treatment alone can achieve less than half the total benefit obtainable through a combination of interventions, and scaling up the use of impregnated mosquito nets or indoor spraying with insecticides is also critical. Intermittent presumptive treatment of pregnant women can bring a small but important additional health gain at relatively low cost.

Introduction

Each year, more than one million people, mostly children and pregnant women, die from malaria. The human toll is tragic, and the economic cost is enormous.[1] [2] Most of these deaths could be avoided, however, as effective and affordable ways to prevent and treat malaria exist. In recognition of the scope of the problem, malaria control is embedded in one of the millennium development goals of the United Nations: to "combat HIV/AIDS, malaria and other diseases."[3]

Although insufficient data are available to fully assess global experience since 2000, malaria related mortality seems to have increased since 1990, probably owing to a combination of factors, including increasing exposure to the disease,[4] increasing resistance to antimalarial drugs,[5] and stagnant levels of coverage with interventions (R W Snow, personal communication, 2005). Complex emergencies and resistance to insecticides have also contributed. [6] Achieving the millennium development goals clearly requires a massive scaling up of interventions against malaria.

However, it is important to ask whether current interventions are used appropriately and what is the most cost effective way to scale up activities to the levels needed. In particular, which prevention or treatment strategies, and what combination, are most effective and where? We use a generalised cost effectiveness analysis (See reference 10, Chapter 3) to examine the costs and effects of scaling up seven interventions against malaria and their most promising combinations. This paper deviates from others in this series[7] by focusing only on sub-Saharan Africa, where 90% of deaths from malaria occur.[8]

Whereas most economic studies have compared the relative cost effectiveness of implementing interventions for prevention or treatment individually—that is, considering the best use of small amounts of additional resources—this study used a generalised framework allowing for interactions, as well as for consideration of whether current practice is optimal and what the implications are of massively scaling up.

Methods

Geographical focus

We focused on two sub-Saharan African regions: Afr-E (predominantly Southern and Eastern Africa), defined as African countries with high child mortality (all causes) and very high adult mortality (all causes), and Afr-D (predominantly Western Africa), African countries with high child mortality and high adult mortality. Annex table A gives a list of the countries by region.

Both regions are predominantly areas with endemic high transmission of malaria due to *Plasmodium falciparum*, although burden of disease differs somewhat. According to the World Health Report 2000, incidence of symptomatic malaria in children aged under 5 years was 1436 per thousand in Afr-D, whereas in Afr-E it was 1184 per thousand; these differences are due to patterns of urbanisation and the elevation of populated sites. In Afr-E, cause specific child mortality is slightly higher at 8 per thousand as opposed to 7 per thousand in Afr-D.

Interventions

A limited number of means are available to fight malaria. Preventive interventions, based on vector control, include insecticide treated nets and indoor residual spraying. For treatment of malaria, several drugs exist and a few are relatively inexpensive. However, resistance to most drugs is growing rapidly. Recently, combination treatments with and without artemisinin derivatives have been tested and found not only to be effective but also to slow the growth of resistance.[9] Intermittent treatment of pregnant women—aimed largely at reducing neonatal mortality—is also an option. We evaluated seven individual interventions and combinations thereof (box 1).

Box 1: Interventions evaluated

- Insecticide treated bed nets (ITN)
- Indoor residual spraying (IRS)
- Case management with chloroquine (CQ)
- Case management with sulfadoxine-pyrimethamine (SP)
- Case management with non-artemisinin based (CQ-SP) combination treatment (Comb)
- Case management with artemisinin based combination treatment (ACT)
- Intermittent presumptive treatment with SP in pregnancy (IPTp) (See Annex for details)

Some countries in sub-Saharan Africa still officially recommend chloroquine as first line treatment for malaria despite increasing resistance and declining cure rates. Although others have moved to sulfadoxine-pyrimethamine, resistance has also compromised its effectiveness. As a result, awareness is growing of the need to increase the use of artemisinin derivatives (especially in combination treatment) as resistance to these compounds is still extremely low (even non-existent) in sub-Saharan Africa.[10] In this study, we evaluated the cost effectiveness of chloroquine, sulfadoxine-pyrimethamine, non-artemisinin based combinations, and artemisinin based combinations as first line treatment (we did not consider complicated malaria needing admission to hospital).

Population at risk and coverage

We evaluated interventions at 50%, 80%, and 95% target coverage to allow for unit costs and effectiveness that may vary with coverage. We estimated effective coverage as target coverage multiplied by population at risk.[8] We based region-wide estimates of population at risk (the proportion living in a malaria endemic area: 98% for Afr-D and 69% for Afr-E) on country specific figures published in 2003.[8] Table 1 shows estimates of current coverage,[8] used for calculating the null scenario.[7]

Table 1 Current coverage* with selected malaria control interventions

Intervention	Afr-D (%)	Afr-E (%)
Case management with chloroquine	23	27
Case management with sulfadoxine-pyrimethamine (SP)	3	3
Case management with non-artemisinin based combination treatment	0	0
Case management with artemisinin based combination treatment	0	0
Insecticide treated bed nets	2	2
Indoor residual spraying	0	3
Intermittent presumptive treatment with SP in pregnancy	0	0

*Percentage of people at risk who are given the drug at outpatient clinic or antenatal clinic, sleep under an insecticide treated net, or live in a house that has been sprayed.

Estimating the net effectiveness of interventions

We expressed the efficacy of nets and indoor spraying as a reduction in incidence and, thereby, a reduction in mortality, modelled here through case fatality (table 2). We estimated the net effectiveness of treatment, taking into account patients' behaviour (adherence to the regimen), pharmacokinetics (probability of success when the regimen is not followed), and biogenetics (resistance of the parasite to the drug). These factors (table 3) determine the number of expected treatment failures,[11] which we subtracted from a common baseline of 98% efficacy. We reduced the net effectiveness of bed nets, but not spraying, to account for imperfect adherence. Table 4 shows estimates of net effectiveness. Annex table B reports the detailed assumptions on effectiveness.

Table 2 Baseline efficacy (both Afr-D and Afr-E)

Intervention	Baseline reduction in incidence (%)	Baseline reduction in case fatality* (%)
Insecticide treated bed nets	50	20
Indoor residual spraying	50	20
Case management with chloroquine	0	98
Case management with sulfadoxine-pyrimethamine (SP)	0	98
Case management with non-artemisinin based combination treatment	0	98
Case management with artemisinin based combination treatment	0	98
Intermittent presumptive treatment with SP in pregnancy	0	5

*Before patients' adherence and parasites' drug resistance are taken into account.

Table 3 Parameters used for the calculation of net effectiveness

Intervention	Resistance		Behavioural/drug characteristics	
	R_0	r	Adherence (%)	Probability of success when not fully compliant (%)
Insecticide treated bed nets	–	–	65	0
Indoor residual spraying	–	–	100	0
Case management with chloroquine	0.35	0.1	40	20
Case management with sulfadoxine-pyrimethamine (SP)	0.1	0.4	90	0
Case management with non-artemisinin based combination treatment	0.1	0.1	35	35
Case management with artemisinin based combination treatment	0.001	0.05	35	45
Intermittent presumptive treatment with SP in pregnancy	0.1	0.1	80	0

Table 4 Net effectiveness of the interventions

Intervention	Reduction in incidence (%)	Reduction in case fatality (%)
Insecticide treated bed nets	50	20
Indoor residual spraying	50	20
Case management with chloroquine	–	27
Case management with sulfadoxine-pyrimethamine (SP)	–	44
Case management with non-artemisinin based combination treatment	–	47
Case management with artemisinin based combination treatment	–	63
Intermittent presumptive treatment with SP in pregnancy	–	3.3

A population model[12] combined estimates of incidence, prevalence, and mortality (table 5)[13] with estimates of prevalence and severity from the burden of disease study to project the population impact of intervention scenarios in terms of healthy years of life lived.[7] Differences in total population healthy years under the intervention and baseline scenarios are expressed as disability adjusted life years (DALYs) averted.

Table 5 WHO estimates of the burden of malaria

	Incidence	Deaths	DALYs
Afr-D	176 547 940	515 528	19 088 877
Afr-E	163 442 874	470 845	17 680 061
Worldwide	381 987 866	1 120 697	41 998 576

DALY=disability adjusted life year.
Source: WHO programme on global burden of disease.

Costs

Estimated costs measure the value of resources needed to provide the intervention[7] and are expressed in international dollars ($int, a hypothetical unit of currency that has the same purchasing power that the US$ has in the United States at a given point in time, thus showing the average value of local currency units within each region's borders). We calculated costs in light of experience from effectiveness trials, using data from the WHO-CHOICE database, existing literature, and expert opinion.[7] As we explicitly assumed training of human resources to be a substantial part of malaria interventions, training costs are reported separately. We used the CostIt model (WHO, 2002) to aggregate cost components and total costs for the 10 year implementation horizon. Details of the approach are discussed by Baltussen et al.[14]

Unit costs—We obtained unit costs of inputs, such as salaries, capital equipment, drugs, storage, buildings, office supplies, and furniture from a review of the literature supplemented by primary data from several countries

(a full list of estimated unit costs is available at www.who.int/evidence/cea). Additional details can be found in Johns et al, Johns and Baltussen, and Adam et al.[15][16][17]

Distribution costs—We assumed distribution costs to be most sensitive to changes in coverage levels and calculated them with a standard mark-up based on the average of free on board; cost, insurance, and freight; and additional trade related distributional costs.[15] [16]

Media costs—A substantial component of malaria control is creating public awareness of and demand for health services. We accordingly included media costs for all interventions and estimated them according to whether they consisted of a centrally determined policy change (for example, case management guidelines) or were intended to change population behaviour (for example, insecticide treated bed nets). We included both public campaigns (all interventions) and targeted social marketing (bed nets only, through an extensive level of outreach). We obtained benchmarks from cost analyses of existing malaria control programmes in sub-Saharan Africa.[18]

Labour costs—We estimated labour costs according to the educational level of the worker—for example, administrative staff or medical staff—and the number of full time equivalents needed for administration, training, or delivery of the intervention.

Cost profiles— Figures 1 and 2 summarise cost profiles for 95% coverage (generally the most efficient coverage level) by patient, programme, and training costs. Note that for some interventions patient costs are the smallest proportion of total costs.

Figure 1 Cost profile of interventions at 95% coverage, Afr-D. See box 1 for abbreviations

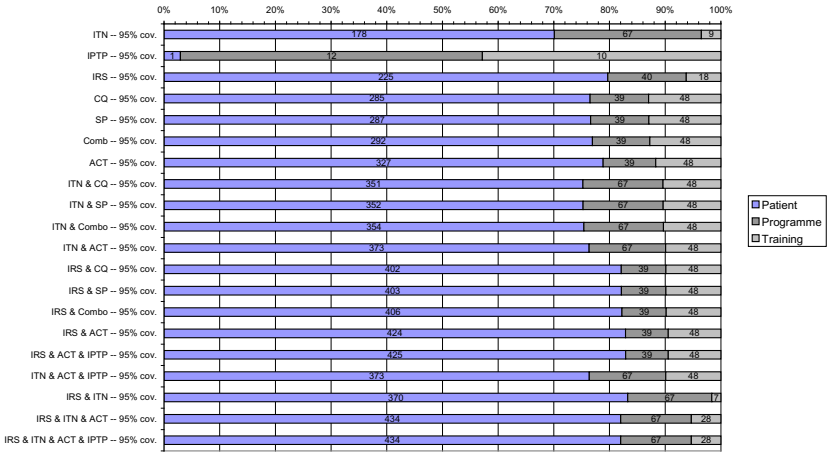
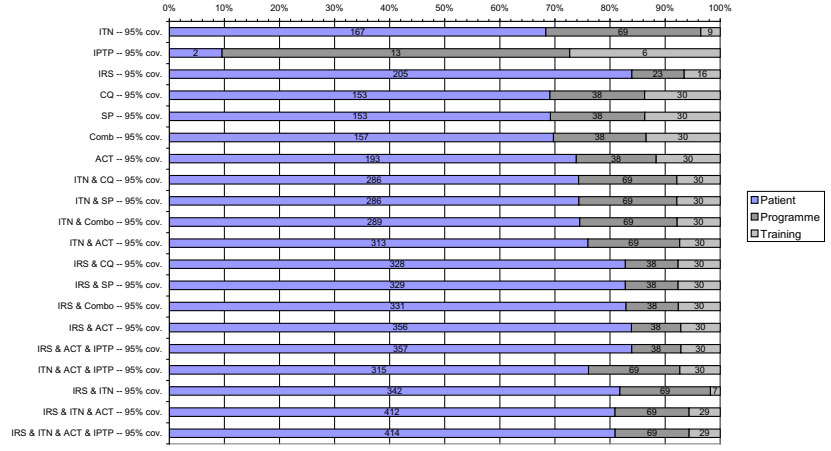


Figure 2 Cost profile of interventions at 95% coverage, Afr-E. See box 1 for abbreviations



Results

Population level cost effectiveness estimates for individual and combined interventions are shown in table 6 (dominant interventions only) and in figures 3 and 4 (all interventions).

Table 6 Costs, effectiveness, and cost effectiveness of the health maximising set of interventions (see Web Table 3 for detailed results for all interventions)

Region and intervention	Average yearly costs ($int)	Average yearly effectiveness (DALYs averted)	Average cost effectiveness ($int/ DALY averted)	Incremental cost effectiveness ($int/ DALY averted)
Afr-D				
Insecticide treated bed nets (95% coverage)	178,216,076	8,872,378	29	29
Insecticide treated bed nets plus indoor residual spraying plus case management with artemisinin based combination therapy plus intermittent presumptive treatment in pregnancy (95% coverage)	434,229,098	14,561,792	36	48
Afr-E				
Case management with artemisinin based combination therapy (80% coverage)	194,055,200	5,055,437	38	38
Case management with artemisinin based combination therapy plus insecticide treated bed nets (80% coverage)	343,906,545	8,136,888	42	49
Case management with artemisinin based combination therapy plus insecticide treated bed nets (95% coverage)	412,334,271	9,338,292	44	57
Case management with artemisinin based combination therapy plus insecticide treated bed nets plus indoor residual spraying (95% coverage)	509,794,378	10,956,811	46.5	60
Case management with artemisinin based combination therapy plus insecticide treated bed nets plus indoor residual spraying plus intermittent presumptive treatment in pregnancy (95% coverage)	510,921,742	10,964,451	46.6	148

DALY=disability adjusted life year; $int=international dollar.

Figure 3 Cost effectiveness plane showing 60 analysed interventions (20 individual and combination interventions at three assumed coverage levels) and expansion path (see text), Afr-D. DALY=disability adjusted life year; see box 1 for other abbreviations

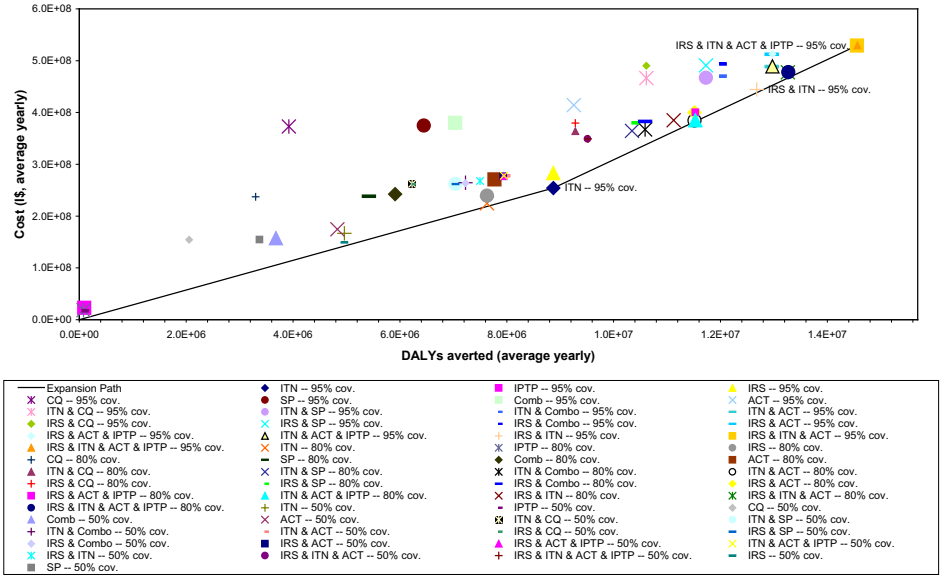
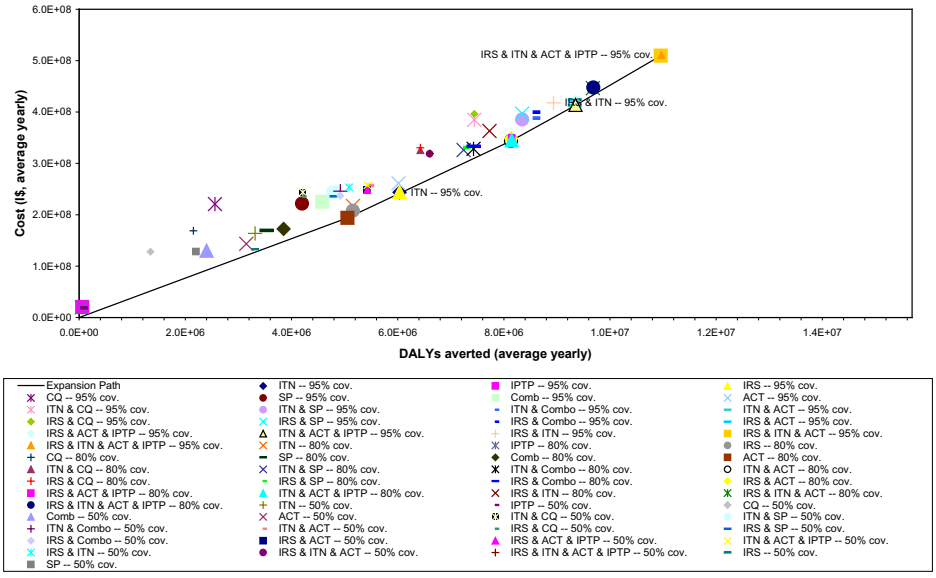


Figure 4 Cost effectiveness plane showing 60 interventions (20 individual and combination interventions at three assumed coverage levels) and expansion path (see text), Afr-E. DALY=disability adjusted life year; see box 1 for other abbreviations



The "expansion paths" in figures 3 and 4, described in the methods paper for this series,[7] show the order in which interventions would be selected at different levels of resource availability. Notable differences exist between the regions. In the Afr-D region, insecticide treated bed nets at 95% target coverage are the most cost effective intervention overall, and would be the first chosen where resources are very limited, whereas in Afr-E, case management with artemisinin based combination treatments at 80% target coverage is the most cost effective intervention overall. In Afr-D, the second intervention on the path represents the combination of insecticide treated bed nets, indoor residual spraying, case management with artemisinin based combination therapy, and intermittent presumptive treatment with sulfadoxine-pyrimethamine in pregnancy, also at 95% target coverage.

In Afr-E, case management with artemisinin based combination treatments at 80% target coverage is followed by the combination of case management with artemisinin based combination therapy and insecticide treated bed nets (80%

coverage); then the combination of case management with artemisinin based combination therapy and insecticide treated bed nets (95% coverage); then the combination of case management with artemisinin based combination therapy, insecticide treated bed nets, and indoor residual spraying (95% coverage); and, finally, the combination of case management with artemisinin based combination therapy, insecticide treated bed nets, indoor residual spraying, and intermittent presumptive treatment with sulfadoxine-pyrimethamine in pregnancy (95% coverage). In both regions, the ultimate stage involves the use of case management with artemisinin based combination treatment, insecticide treated bed nets as well as indoor residual spraying, and intermittent presumptive treatment in pregnancy, all at 95% coverage.

All malaria interventions are highly cost effective, with average cost effectiveness ratios (except intermittent presumptive treatment with sulfadoxine-pyrimethamine in pregnancy) in the order of 10-100 $int /DALY averted. Nevertheless, the size of potential health gains, as well as incremental cost effectiveness ratios, are more favourable in Afr-D than in Afr-E, as a higher proportion of the population is at risk in Afr-D. That allows more people to be covered, thereby reducing the costs per person covered (fig 5, fig 6, table 6).

Figure 5 Incremental and average cost effectiveness ratios for the health maximising interventions, Afr-D. DALY=disability adjusted life year; see box 1 for other abbreviations

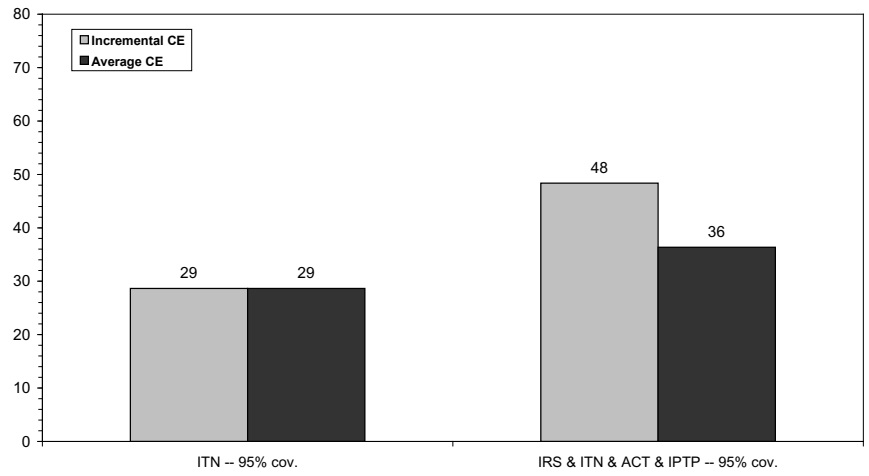
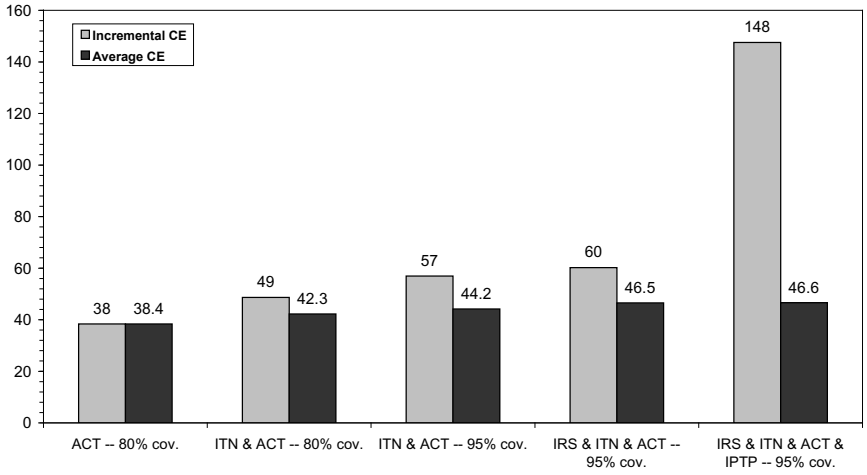


Figure 6 Incremental and average cost effectiveness ratios for the health maximising interventions, Afr-E. DALY=disability adjusted life year; see box 1 for other abbreviations



Discussion

Principal findings

Progress towards the millennium development goals has been disappointing and major action is needed if sub-Saharan Africa is to achieve the malaria specific target. Although external assistance is critical, countries also need to re-evaluate current strategies (box 2). A distinguishing characteristic of this analysis is the health system focus, integrating evidence about costs and effectiveness at a level relevant for policy makers.

Box 2: Main messages—malaria and the millennium development goals

- Five years on from the declaration of the millennium development goals, progress in controlling malaria remains uncertain, particularly in sub-Saharan Africa
- Health system decision makers in most countries in sub-Saharan Africa (see text for details) should consider switching treatment strategies to artemisinin based combinations as the foundation of effective malaria control
- Reinvigorating efforts to scale up use of impregnated mosquito nets and indoor spraying with insecticides is critical
- Where these interventions are being successfully implemented, intermittent presumptive treatment of pregnant women can bring a small but important additional health gain
- A much larger infusion of resources than those currently available is needed to make headway in the fight to roll back malaria

Recent developments such as the Global Fund to Fight AIDS, Tuberculosis and Malaria, and the interest of various foundations, promise increased funding for malaria control. To date, however, attention has been focused on the drugs, other expendables, and, to a limited degree, additional staffing needed to scale up health interventions. This narrow focus ignores the additional system resources needed to implement and sustain a massive scale up of activity, as well as the relatively fixed constraints on existing staff time and health facilities. Our results suggest that the costs of reaching agreed targets are considerably greater than those costs that receive most current attention.

Artemisinin based combination treatments are more expensive than other drugs and, in the short term, non-artemisinin based combinations such as amodiaquine-sulfadoxine-pyrimethamine might be used as an interim measure.[19] However, as resistance to these drugs is increasing, recommendations should be guided by local resistance patterns,[20] and a full switch to artemisinin based combinations should be seriously considered in all areas with high transmission of drug resistant *P falciparum* malaria.

Differences from other studies

This analysis differs from previous studies for malaria in that we have specifically assessed combinations of interventions, rather than assuming that costs and effects sum up when interventions are used concurrently. Such studies can be misleading at worst, and, at best, decision makers lack necessary information about relevant combinations.

In the only previous study analysing sub-Saharan Africa, Goodman et al found the cost/DALY averted of insecticide treated bed nets to be $19-85 (US$1995) compared with $16-29 for indoor residual spraying.[21] The estimates presented here (~30 $int /DALY for either intervention at 95% coverage in Afr-D and ~40 $int/DALY at 95% coverage in Afr-E) are roughly comparable. However, what does not emerge from typical cost effectiveness trials is the conclusion that implementing spraying and nets together has a cost effectiveness ratio roughly similar to that of either one alone (~35 $int/DALY at 95% coverage in Afr-D and ~47 $int/DALY at 95% coverage in Afr-E), from the realisation of substantial efficiencies due to shared costs (reducing the unit costs of both) counterbalanced by the reduced health benefits resulting from adding spraying to widespread use of bed nets.

Limitations of the analysis

The relative effectiveness of spraying and impregnated bed nets has been debated.[22] [23] Although we have assumed that they are equally effective in reducing morbidity and mortality, their respective advantages in a given setting will depend on whether mosquitoes are endophilic, the length of time they rest indoors,[24] and the time children go to bed.

People's willingness to use bed nets effectively is often reduced by user charges and the need to reimpregnate nets if they are not long lasting. Moreover, nets must be kept in good condition and used consistently despite their at times uncomfortable insulating properties.[25]

Indoor residual spraying, on the other hand, is generally paid for and carried out by public authorities once or twice a year, which in theory makes adherence less burdensome; however, house owners may replaster or paint over recently

sprayed walls because of the smell of insecticide or visible deposits,[23] and residents must be willing to let sprayers into their homes. Furthermore, the relative effectiveness of these interventions will depend on local transmission patterns. For example, in epidemic-prone areas, spraying is likely to be an effective emergency measure as it can be implemented relatively quickly and needs less logistic effort than a bed net campaign.[25]

Our assumptions about the effectiveness of intermittent presumptive treatment with sulfadoxine-pyrimethamine in pregnancy were very conservative. We did not include health benefits for the mother and the effect on the infant was limited to a reduction in case fatality due to malaria in the first year of life. Costing of intermittent presumptive treatment was as part of a routine antenatal package of care with only the incremental costs of dedicated activities considered.

Owing to lack of data on reductions in all cause mortality, we assumed that when drugs are taken under ideal conditions they are 98% effective in preventing cause specific mortality. While this study does try to account for parasite resistance, imperfect patient compliance, and pharmacokinetic properties, actual effectiveness is likely to include factors omitted in the analysis. Modelling implementation over a 10 year period, moreover, may not fully capture the contrast between drugs with high versus low growth rates of resistance (for example, sulfadoxine-pyrimethamine *v* artemisinin based combination treatment). Consequently, our estimate of the cost effectiveness of artemisinin based combination treatments may be conservative compared with when a longer term perspective is taken.[26] Finally, the study allows no conclusions to be drawn about financing methods, which can be of utmost importance.[27]

Implications of the study

In conclusion, with the availability of increased international funding, a re-evaluation of existing and potential strategies is appropriate. As an excessive disease specific focus is undesirable, an adequate portion of new funds should be allocated to strengthening the health system, which determines the long term viability of disease control activities.

On purely cost effectiveness grounds, this study suggests that countries in sub-Saharan Africa should be moving to combination therapy with new drugs as rapidly as possible or that efforts should be regenerated to ensure that the use of nets is scaled up appropriately. Even where the health system is weak, prevention will never be a stand alone strategy, as informal distribution mechanisms ensure the availability of antimalarials—both effective and highly ineffective—even in the most isolated areas. All attempts should be made to ensure that the antimalarials that are accessed are indeed those that work. These are relatively simple messages, but ones that remain robust even when underlying assumptions are varied significantly.

Acknowledgements

We thank Anne Mills, Jo Lines, Chris Curtis, Brian Greenwood, and Tessa Tan-Torres for their assistance over the course of this project. We also thank Eline Korenromp, Kamini Mendis, Pascal Ringwald, Andrea Bosman, Wilson Were, and Peter Olumese for their helpful comments.

Contributors

CMM and JAL contributed equally to the planning, conduct, interpretation, and writing up of the work. CMM was responsible for assessing and estimating all malaria specific data and methods; JAL was responsible for the final estimation and modelling of costs and effects at population level. DBE revised the analysis plan and contributed critical interpretive content. All authors revised and approved the final draft. JAL is the guarantor.

Funding

CMM received consulting fees from the World Health Organization.

The views expressed are solely those of the authors and do not necessarily reflect the decisions or stated policy of the London School of Hygiene and Tropical Medicine or the World Health Organization.

References

Competing interests: None declared.

1. Sachs J, Malaney P. The economic and social burden of malaria. *Nature* 2002;415:680-5.
2. Chima RI, Goodman CA, Mills A. The economic impact of malaria in Africa: a critical review of the evidence. *Health Policy* 2003;63:17-36.
3. United Nations. Millennium development goals. World Bank Group, 2005.
4. Hay SI, Guerra CA, Tatem AJ, Noor AM, Snow RW. The global distribution and population at risk of malaria: past, present, and future. *Lancet Infect Dis* 2004;4:327-36.
5. White NJ. The assessment of antimalarial drug efficacy. *Trends Parasitol* 2002;18:458-64.
6. Roll Back Malaria, World Health Organization, UNICEF. *World malaria report 2005*. Geneva: WHO, 2005.
7. Evans DB, Adam T, Tan-Torres Edejer T, Lim SS, Cassels A, Evans T, et al. Achieving the millennium development goals for health: time to reassess strategies? *BMJ* 2005;331:1133-1136.
8. World Health Organization, UNICEF. *The Africa malaria report 2003*. Geneva: WHO, UNICEF, 2003:120.
9. White N. Delaying antimalarial drug resistance with combination chemotherapy. *Parassitologia* 1999;41:301-8.
10. White NJ. Averting a malaria disaster. *Lancet* 1999;353:1965-7.
11. Goodman C, Coleman PG, Mills AJ. Changing the first line drug for malaria treatment—cost-effectiveness analysis with highly uncertain inter-temporal trade-offs. *Health Econ* 2001;10:731-49.

12. Lauer JA, Roehrich K, Wirth H, Charette C, Gribble S, Murray CJL. PopMod: a longitudinal population model with two interacting disease states. *Cost Eff Resour Alloc* 2003;1:6.

13. World Health Organization. Global burden of disease estimates 2002. www3.who.int/whosis/menu.cfm?path=evidence,burden (accessed 16 September 2004)

14. Baltussen R, Adam T, Tan-Torres Edejer T. *Methods for generalised cost-effectiveness analysis: a guide. WHO guidelines on cost-effectiveness analysis*. Geneva: WHO, 2003.

15. Johns B, Baltussen R, Adam T, Hutubessy R. Programme costs in the economic evaluation of health interventions. *Cost Eff Resour Alloc* 2003;1:1.

16. Johns B, Baltussen R. Accounting for the costs of scaling up health interventions. *Health Econ* 2004;13:1117-24.

17. Adam T, Evans D, Murray C. Econometric estimation of country-specific hospital costs. *Cost Eff Resour Alloc* 2003;1:3.

18. Stevens W, Wiseman V, Ortiz J, Chavasse D. Cost-effectiveness of a nationwide insecticide treated net programme: the case of Malawi. *Malar J* 2005;4:22.

19. Staedke SG, Mpimbaza A, Kamya MR, Nzarubara B, Dorsey G, Rosenthal PJ. Combination treatments for uncomplicated falciparum malaria in Kampala, Uganda: randomised clinical trial. *Lancet* 2004;364:1922-3.

20. Bakyaita N, Dorsey G, Yeka A, Banek K, Staedke SG, Kamya MR, et al. Sulfadoxine-pyrimethamine plus chloroquine or amodiaquine for uncomplicated falciparum malaria: a randomised, multisite trial to guide national policy in Uganda. *Am J Trop Med Hyg* 2005;72:573-80.

21. Goodman CA, Coleman PG, Mills AJ. Cost-effectiveness of malaria control in sub-Saharan Africa. *Lancet* 1999;354:378-84.

22. Guyatt HL, Kinnear J, Burini M, Snow RW. A comparative analysis of insecticide-treated nets and indoor residual spraying in highland Kenya. *Health Policy Plan* 2002;17:144-53.

23. Goodman CA, Minzava AEP, Diamini SS, Sharp BL, Mthembu DJ, Gumede JK. Comparison of the cost and cost-effectiveness of insecticide-treated bednets and residual house-spraying in KwaZulu-Natal, South Africa. *Trop Med Int Health* 2001;6:280-95.

24. Sharp BL, LeSueur D, Bekker P. Effect of DDT on survival and blood feeding success of Anopheles arabiensis in northern Kwazulu, Republic of South Africa. *J Am Mosq Control Assoc* 1990;6:197-202.

25. Curtis CF, Mnzava AEP. Comparison of house spraying and insecticide-treated nets for malaria control. *Bull World Health Org* 2000;78:1389-400.

26. Coleman PG, Morel CM, Shillcutt SD, Goodman CA, Mills AJ. A threshold analysis of the cost-effectiveness of artemisinin-based combination therapies in sub-Saharan Africa. *Am J Trop Med Hyg* (in press).

27. Hanson K, Goodman C, Lines J, Meek S, Bradley D, Mills A. *The economics of malaria control interventions*. Geneva: WHO Global Forum for Health Research, 2004.

Annex

Individual and combination interventions

In addition to the seven individual interventions discussed in the main text (reproduced here for convenience):

1) Insecticide-treated bed nets (ITN);
2) Indoor residual spraying (IRS);
3) Case management with chloroquine (CQ);
4) Case management with sulfadoxine-pyrimethamine (SP);
5) Case management with non-artemisinin-based (CQ-SP) combination therapy (Comb);
6) Case management with artemisinin-based combination therapy (ACT);
7) Intermittent presumptive treatment with SP in pregnancy (IPTp);

the following combinations were analysed:

8) ITN & CQ;
9) ITN & SP;
10) ITN & Comb;
11) ITN & ACT;
12) IRS & CQ;
13) IRS & SP;
14) IRS & Comb;
15) IRS & ACT;
16) IRS & ACT & IPTp;
17) ITN & ACT & IPTp;
18) IRS & ITN;
19) IRS & ITN & ACT;
20) IRS & ITN & ACT & IPTp.

Each of these 20 interventions was analysed at each of the three coverage levels (50%, 80% and 95%) and in each of the two regions (Afr-D and Afr-E).

Intervention-specific assumptions

Drugs

The current subregional coverage of each case-management intervention was estimated as a population-weighted average on the basis of the country population and country-level estimates of the proportion of children who receive CQ or other antimalarials.[1] If no data were available on the proportion of sick children receiving drugs, rather than assume zero coverage for the country, the country was excluded from both the numerator and denominator of the coverage calculation. Estimates for the use of 'any antimalarial' that were greater than those estimates quoted specifically for CQ were assumed to refer to that country's first-line drug (if other than CQ, otherwise that country's second-line drug). If specific coverage estimates for CQ were not available, it was assumed that 80% of 'any antimalarial' consisted of that country's first-line drug, with the remaining 20% consisting of the second-line drug.

Estimating drug resistance

Resistance was estimated by means of a standard logistic growth function (eq. 1). Projected effectiveness estimates for case management interventions and IPTp were thus adjusted for changing resistance profiles with increasing duration of use. Projected drug resistance was estimated as a function of the initial level of resistance prior to intervention implementation, the growth rate of resistance, and the maximum level of resistance (carrying capacity). As the estimated growth in resistance over the 10-year implementation horizon was in all cases approximately linear, annual numbers of treatment failures were averaged over the 10-year implementation period to obtain an average failure rate.

Resistance was measured as:[1]

$$R_{i,t} = k_i \left[\frac{R_{i,0}}{R_{i,0} + (k_i - R_{i,0}) \exp(-r_i t)} \right],$$

where $R_{i,t}$ is the proportion of parasites that resistant to drug i at time t, R_0 is the initial level of resistance, r is the growth rate in resistance, k is the maximum level of resistance and t is time (see Table 1, main text). The growth rate of resistance was estimated in conjunction with biogeneticists on the basis of drug-specific pharmacokinetic properties and experience in the field.[2]

Drug treatment failure

Treatment failure was calculated as:[1]

$$F_{i,t} = 1 - \{(1 - R_{i,t}) \cdot m_i + (1 - R_{i,t}) \cdot (1 - m_i) \cdot p_i\},$$

where F is treatment failure, m is the probability of full adherence, and p is the probability of treatment success without full compliance (see Table 2, main text).

Net drug effectiveness

Net effectiveness, E, was calculated as:

$$\mathrm{E} = B - F_{i,t},$$

where B is baseline efficacy, and $F_{i,t}$ is treatment failure.

Preventive interventions

No vector resistance to insecticide was accounted for. Potential benefits of prevention on unborn children (i.e. in terms of higher birth weight) were not included.

IRS and ITN were assumed to benefit from a 9% "knock-on" effect (i.e. transmission externality) at modelled coverage levels: that is, the costs of a given level of effective coverage were reduced by about 8% (= 1 − 1/1.09), while keeping the effects unchanged. This is likely a conservative estimate.[3]

For ITNs, coverage estimates refer to individuals, while, for IRS, coverage estimates refer to the household (e.g. a 50% coverage level for IRS indicates that 91% of houses within 50% of transmission areas are assumed to be effectively sprayed.

ITN

ITNs were assumed to be long lasting and thus not requiring reimpregnation. ITNs were assumed to last on average 5 years.[4] One ITN was attributed to each 1.5 persons.[5] Estimates of ITN effectiveness assumed 65% adherence, defined as sleeping under an untorn net.

IRS

Each spray team was assumed to cover 1170 houses per year.

Case Management

Case management refers to the first-line antimalarial drug administered to symptomatic patients presenting at outpatient clinics. One health center visit per case was assumed, as well as that a patient given an antimalarial drug does indeed have malaria (and therefore receives the drug benefit).

IPTp

Women of childbearing age were assumed to be those aged 15 to 49. Sixty percent of first-time mothers (estimated to account for 30% of births) were assumed to receive two doses of IPTp within the context of routine antenatal care (ANC) services. Sixty percent of pregnancies considered "at risk" due to HIV (estimated to account for an additional 20% of births) were assumed to receive three doses. A total of 15% of pregnant mothers (split evenly between first and multiparous high-risk pregnancies) treated with IPTp were assumed to have babies that would not survive the neonatal period. A potential reduction in incidence of placental malaria from IPTp was excluded from consideration, as its associated morbidity is difficult to quantify due to the asymptomatic nature of the disease in endemic areas.[6]

Resistance to SP within the IPTp context was based on estimates of resistance to SP when used as a first-line treatment.[7] However, it was assumed that when SP is used for IPTp it is no longer used in standard case management, since if SP is concurrently used as the first line treatment for malaria, it's usefulness as a preventive treatment will be diminished. Resistance to SP was assumed to grow more slowly when used only for IPTp.

Annex references

1. Goodman C, Coleman PG, Mills AJ. Changing the first line drug for malaria treatment - cost-effectiveness analysis with highly uncertain inter-temporal trade-offs. *Health Economics* 2001;10:731-749.
2. Roper C. personal communication, 2004.
3. Curtis CF, Maxwell CA, Finch RJ, Njunwa KJ. A comparison of use of a pyrethroid either for house spraying or for bednet treatment against malaria vectors. *Tropical Medicine and International Health* 1998;3(8):619-631.
4. WHO, RBM, UNICEF, PSI, MSH. Sources and Prices of Selected Products for the Prevention, Diagnosis and Treatment of Malaria, 2004.
5. Lines J. personal communication, 2005.
6. WHO. Global burden of disease estimates 2002. Geneva: http://www3.who.int/whosis/menu.cfm?path=evidence,burden, 2002.
7. CostIt [program]. 1 version. Geneva: WHO, 2002.
8. WHO. *The Africa Malaria Report 2003*. Geneva: UNICEF and World Health Organization, 2003:120.
9. Lengeler C. *Insecticide-treated bednets and curtains for preventing malaria*. Cochrane Library, Issue 1. Oxford, 2001.
10. Curtis CF, Mnzava AEP. Comparison of house spraying and insecticide-treated nets for malaria control. *Bull World Health Organ* 2000;78(12):1389-1400.
11. Lengeler C. Comparison of malaria control interventions. *Bull World Health Organ* 2001;79(1).
12. Goodman CA, Coleman PG, Mills AJ. *Economic analysis of malaria control in sub-Saharan Africa*. Geneva: Global Forum for Health Research, 2000:185.

Annex table A: Countries analysed in each African region

Afr-D

Algeria
Angola
Benin
Burkina Faso
Cameroon
Cape Verde
Chad
Comoros
Equatorial Guinea
Gabon
Gambia
Ghana
Guinea
Guinea Bissau
Liberia
Madagascar
Mali
Mauritania
Mauritius
Niger
Nigeria
Sao Tome and Principe
Senegal
Seychelles
Sierra Leone
Togo

Afr-E

Botswana
Burundi
Central African Republic
Congo
Cote d'Ivoire
Democratic Republic of the Congo
Eritrea
Ethiopia
Kenya
Lesotho
Malawi
Mozambique
Namibia
Rwanda
South Africa
Swaziland
Uganda
United Republic of Tanzania
Zambia
Zimbabwe

Annex table B: Estimation of effectiveness

Intervention/Parameter	Value	Source/Notes
ITN		
Adherence	65%	Estimate within range suggested to achieve impact[8]
Probability of success when not fully compliant	0%	Estimate based on binary response of whether or not an individual has slept under the net
— resulting reduction in incidence	50%	[8]
— resulting reduction in case fatality	20%	[9]
IRS		
Adherence	100%	Estimate based on household level to achieve impact
Probability of success when not fully compliant	0%	Estimate based on dosing schedule and pharmacokinetic properties
— resulting reduction in incidence	50%	Assumed to be similar to ITNs[3,10,11]
— resulting reduction in case fatality	20%	Assumed to be similar to ITNs[3,10,11]
CQ		
Adherence	40%	Estimate based on relative length and complexity of the regimen
Probability of success when not fully compliant	20%	Estimate based on dosing schedule and pharmacokinetic properties

Initial resistance	0.35	[2]
Growth rate of resistance	0.1	[2]
— resulting reduction in case fatality	27%	See Drugs for details of calculation
SP		
Adherence	90%	Estimate based on relative length and complexity of the regimen
Probability of success when not fully compliant	0%	Estimate based on dosing schedule and pharmacokinetic properties
Initial resistance	0.1	[2]
Growth rate of resistance	0.4	[2]
— resulting reduction in case fatality	44%	See Drugs for details of calculation
CQ-SP		Assumed to be co-administered, not co-formulated
Adherence	35%	Estimate based on relative length and complexity of the regimen
Probability of success when not fully compliant	35%	Estimate based on dosing schedule and pharmacokinetic properties
Initial resistance	0.1	[2]
Growth rate of resistance	0.1	[2]
— resulting reduction in case fatality	47%	
ACT		Assumed to be blister-packed and/or co-formulated
Adherence	35%	Estimate based on relative length and complexity of the regimen

Probability of success when not fully compliant	45%	Estimate based on dosing schedule and pharmacokinetic properties
Initial resistance	0.001	[2]
Growth rate of resistance	0.05	[2]
— reduction in case fatality	63%	See Drugs for details of calculation
IPTp (SP)		
Adherence	80%	Estimate based on relative length and complexity of the regimen
Probability of success when not fully compliant	10%	Estimate based on dosing schedule and pharmacokinetic properties
Initial resistance	0.1	[2]
Growth rate of resistance	0.1	[2]
— reduction in case fatality	3.3%	Assumes baseline reduction of 5% based on model developed by Goodman et al 2000[12]; See methods section for details of calculation of net effect

Chapter 5

Deriving logistic distributions from basic processes

Based on: Lauer JA, Greenland S. Deriving logistic distributions from basic processes (in submission).

Summary

The logistic distribution is one of the most widely used models in statistics. While its popularity may be attributed to its simplicity, it can often be derived from more basic contextual considerations, for example by subtracting a quadratic term from a linear differential equation and integrating. The quadratic term is a convenient choice when approximating some unknown density dependence in the outcome. We show that the logistic form is the limiting distribution of a particular Markov process on a lattice, and that the apparent density dependence emerges as a consequence of the lattice embedding. For applications in which time or age is a conditioning variable, the underlying discrete process lends substantive meaning to the use of its limiting description in modeling or estimation. We describe these properties for a logistic analysis of classical models of infection. We also derive a "square-logistic" model from basic considerations.

Introduction

The logistic transform from the real line to the unit interval is defined by

$$\text{Eq. (1)} \qquad \text{expit}(x) \equiv e^x/(1 + e^x) = 1/(1 + e^{-x}).$$

This function is common in statistical and biological research, and is ubiquitous in some fields, such as epidemiology. In most statistical applications it is merely a range-limiting function for regressions confined to the unit interval. Specifically, the logistic regression model corresponds to use of its inverse,

$$\text{Eq. (2)} \qquad \text{logit}(u) = \ln[u/(1 - u)],$$

as the link function in a generalized-linear model (McCullagh and Nelder, 1989). In this usage it provides easily interpreted coefficients and a closed form, unlike the probit (inverse-normal) transform (Hosmer and Lemeshow, 2000).

Our primary concern here is with the also-common usage of the logistic transform as a model for the distribution of a random variable (e.g., Greenland and Lash, 2008). We begin by describing the distribution and its history. We will describe how this distribution (like the Gaussian) arises as the limit of a natural diffusion process, providing a deeper rationale for the logistic distribution as a starting model in several applications. Along the way we note some natural connections between logistic regression and the logistic distribution, that is, between the logistic transform used as a link function versus used as a distribution.

The logistic distribution

Let $P(y) \equiv \Pr(Y \leq y)$ and $Q(y) \equiv \Pr(Y > y)$ be the cumulative ("risk") and survival distributions for a random variable Y. The standard logistic distribution then has $P(y) = \text{expit}(y)$ and

$$\text{Eq. (3)} \qquad Q(y) = \text{expit}(-y) = 1/(1 + e^y),$$

with density

$$\text{Eq. (4)} \qquad p(y) = P(y)Q(y) = \text{expit}(y)\text{expit}(-y) = e^y/(1 + e^y)^2$$

and hazard $p(y)/Q(y) = P(y)$; in fact the standard logistic is definable as the distribution with hazard equal to risk (much as the standard exponential is definable as the distribution with density equal to survival). Like the normal distribution, the logistic is symmetric, unimodal, and unbounded, and generalizes to include location and scale parameters m and s by substituting $z = (y-m)/s$ for y in the above formulas. The logistic is heavier-tailed than the normal however, with standard deviation $\sigma = s(\pi/3^{1/2}) \approx 1.81s$, whereas the normal has scale s equal to its standard deviation σ.

Like the normal, the logistic has several simple mathematical derivations. Suppose U is uniform on the unit interval; then $F = U/(1-U)$ is an F-variate with 2-numerator and 2-denominator degrees of freedom, and $Y = \ln(F) = \text{logit}(U)$ is logistic. The logistic distribution is thus a special case of the log-F distribution (Jones, 2004). There is a related Bayesian derivation. Because a uniform variate is also a beta variate, the logistic is also a logit-beta distribution, and is thus a conjugate prior for logistic regression (Greenland, 2003). In particular, the logistic distribution generalizes Laplace's law of succession for a binomial parameter (Good, 1983) to logistic-regression coefficients: The coefficient β of a regressor X can be given a logistic prior by adding two records to the data, one with 1 success and one with 1 failure, with all regressors (including the constant) set to zero except X, which is set to 1 in each added record. Ordinary maximum likelihood applied to the resulting augmented data then gives the corresponding posterior mode, and location and scale parameters can be added by slight modification of the added record (Greenland, 2007).

Historical origins

Cramer (2005) recounts that Verhulst (1838) developed the logistic distribution as a result of questions put to him by his teacher, the "astronomer turned statistician", Alphonse Quetelet. Verhulst apparently arrived at the logistic model as one of several modifications to exponential growth, modifications undertaken to avoid "impossible values" of unbounded growth. The exponential growth model is

$$\text{Eq. (5)} \qquad \frac{dP(t)}{dt} = \alpha P(t),$$

where α is a constant. With the boundary condition $P(0) = P_0$ the solution is

Eq. (6) $$P(t) = P_0 \exp(\alpha t) .$$

For positive α, $P(t)$ increases faster than any power of t (Rudin, 1976).

Cramer states the logistic form was but one alternative explored by Verhulst. Let the exponential term of the differential equation for growth be denoted by the function $f(P(t)) \equiv \alpha P(t)$; Verhulst investigated various forms for a second term, $g(P(t))$

Eq. (7) $$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = f(P(t)) - g(P(t)) .$$

The logistic distribution results when g is quadratic. The form usually seen in the demographic and ecologic literature is

Eq. (8) $$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = r\left(1 - \frac{P(t)}{K}\right)P(t) ,$$

where r represents the "intrinsic growth rate" and K the "carrying capacity". We call this form the "Verhulst equation". By letting $p = P/K$ and $\tau = rt$, the Verhulst equation can be written

Eq. (9) $$\frac{\mathrm{d}p(\tau)}{\mathrm{d}\tau} = p(\tau)(1 - p(\tau)) = p(\tau) - p^2(\tau)$$

(Cantrell and Cosner, 2003). In the following we typically assume this rescaling has been done. Although the other forms investigated by Verhulst are not mentioned in Cramer, non-quadratic forms for g were later proposed by Gompertz and Bertalanffy in the eponymous growth models (Fletcher, 1974).

The choice of g quadratic might appear arbitrary. For example, Cardy (1996) writes:

> Consider a population of individuals whose local density is described by the fluctuating variable $n(r, t)$. When n is small, there is a birth rate, λ, but, when the local population gets too large, it is limited by a shortage of food. In the absence of fluctuations, such a system may be modelled by a rate equation of the form
>
> $$n' = \lambda n - \mu n^2 . \qquad (10.53)$$

> In this approximation, there is a dynamic phase transition at $\lambda = 0$. For $\lambda < 0$, the population always eventually dies out, while, for $\lambda > 0$, it achieves a steady state with density λ/μ.

Cardy does not explain why "shortage of food" should be quadratic in population size, and the squared term appears to be only a convenient choice. By "approximation", Cardy evidently means that the differential equation (10.53) is local and non-stochastic. He continues:

> To describe the system more completely, the effects of diffusion and noise must be added. Thus (10.53) gets replaced by
>
> $$n' = D\nabla^2 n + \lambda n - \mu n^2 + \zeta\,, \qquad (10.54)$$
>
> where D is the diffusion coefficient.

Other than the added "noise" term ς, Cardy's (10.54) is equivalent to Fisher's equation for gene diffusion (Fisher, 1937).

Cantrell and Cosner (2003), citing Enright (1976), derive logistic growth by expressing the birth (respectively mortality) rate as the sum of a constant, density-independent, term and a negative (resp. positive) term that varies linearly with population

$$\text{Eq. (10)} \qquad \frac{\mathrm{d}p(t)}{\mathrm{d}t} = \big[(a - bp(t)) - (c + dp(t))\big]p(t)\,.$$

Collecting density-independent ($a - c \equiv r$) and density-dependent terms ($b + d \equiv q$), the resulting expression

$$\text{Eq. (11)} \qquad \frac{\mathrm{d}p(t)}{\mathrm{d}t} = \big[r - qp(t)\big]p(t)$$

is equivalent to the Verhulst equation when K is set equal to r/q. Thus, quadratic density dependence in net reproduction can be derived from a linear dependence of the birth (resp. mortality) rate on density.

Other standard presentations introduce the logistic distribution as a statistical convenience, for example, in non-linear models for censored response (Aldrich and Nelson, 1984), or as a variance-reduction device (Armitage, Berry and Matthews, 2002). In the following, we describe contexts where the logistic

distribution is neither arbitrary nor approximate, and thus its rationale need not rely solely on statistical convenience or mathematical simplicity. In those contexts, the logistic equation does not enforce damping by means of vague constraints such as "shortage of food," "resistance to further growth," or alternative heuristics. Instead, the logistic equation is the limiting form of a particular random walk on a lattice subject to local interaction. Furthermore, the logistic form is a "hydrodynamic limit" in the sense that smooth macroscopic dynamics are deduced from random microscopic interactions (Kipnis and Landim, 1999).

We show, moreover, that the growth ("reaction") term of reaction-diffusion equations such as Fisher's law for gene advance is, under this derivation at least, itself a diffusion (i.e., a Markov process with a continuous time path), and thus non-local in the sense of Cardy. In other words, we develop a context in which most of the motivating claims for logistic growth and logistic reaction-diffusion equations do not apply: we show that logistic growth yields the logistic distribution and a lattice embedding imposes an intrinsic constraint. The derivation is relevant, at least as an idealization, to a variety of estimation and modeling situations found in population dynamics, epidemiology, ecology and related disciplines. Finally, we develop a logistic analysis of classical infection models.

Contact process on the lattice

To begin our development, we show how the logistic distribution arises from a Markov process subject to local interaction (Spitzer, 1970). Related fields in probability and statistical physics such as contact processes (Harris, 1974), percolation theory (Grimmet, 1999) and interacting particle systems (Liggett, 2005), provide background.

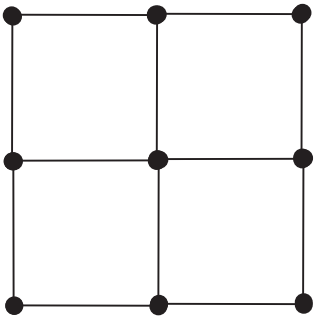
A contact process is a type of Markov process whose domain is a lattice of points (Harris, 1974). It is commonly understood as an idealization of the spread of infection or of a population. Interaction between sites is local: the state of a site depends only on the state of sites in its immediate neighborhood (i.e., those "in contact"). In a typical specification, sites take the value of 0 (uninfected, unoccupied) or 1 (infected, occupied) according to a rule specifying a non-zero probability of occupation only when a neighboring site is occupied.

If adjacent sites are joined by a bond (edge) when occupied, contact processes form a branch of percolation theory on graphs. Originally, percolation refers to the movement or filtering of fluids through porous materials. The mathematical study of percolation was presumably motivated by questions like "What is the probability that the centre of a large porous stone, when immersed in water, is wetted?" (Grimmet, 1999), or, more formally, "What is the probability of a long connected path on a lattice where sites are connected with probability p and unconnected with probability $1 - p$?". An object is said to be percolated if traversed by a connected path; "infinite" paths arise when the scale of the object's structure is negligible compared with its size (Grimmet, 1999). If the "object" is a population of individuals (e.g., represented by a lattice of points) and the contact process is an infection, percolation is equivalent to the emergence of an epidemic.

Our contact process derives from directed (oriented) percolation, in other words where edges form in only a given direction (Durrett, 1984). Directed percolation is frequently a model for dynamic, non-equilibrium, processes (Cardy, 1996; Dickman, 1997). We restrict attention to a simple directed process on the square lattice.

Formally, a lattice is a set whose elements correspond one-to-one with the set of integers **Z**; a square lattice has two-dimensional elements and can thus be mapped onto $\mathbf{Z}^2$. A small subset of the square lattice is shown in Figure 1.

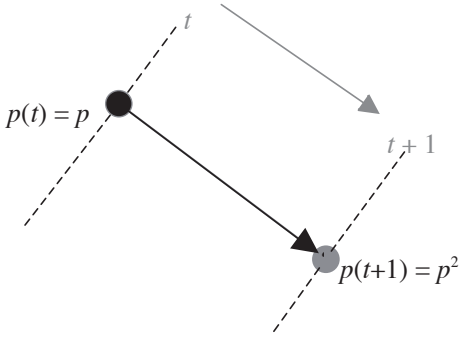
Figure 1 Small subset of the square lattice



In Figure 2, a lattice site is shown to be "occupied" by means of a heavy black dot (upwards and to the left); it is shown as "unoccupied" with a lighter grey dot (downwards and to the right).

Occupied sites may propagate to adjacent unoccupied sites in the positive time direction. Whether an occupied site will successfully propagate is determined according to a probability model. In Figure 2, the uncertain outcome is represented by a question mark. When two adjacent sites are occupied, they are joined by an arrow (directed edge), indicating the connection (an open directed path). Sites are occupied with probability p.

Figure 2 The contact process as an example of directed percolation



Let a random site transected by the line t be occupied. The adjacent site at time $t + 1$ can be occupied, independently with probability p. If the site at time t is not occupied, the probability of occupation for the site at time $t + 1$ is zero. Thus, unconditionally, the probability of a site at $t + 1$ being occupied is p^2, the probability that the parent site is occupied times the conditional probability that the offspring site is occupied given the parent is occupied (see Figure 2).

The change in the probability of site occupation with respect to time is equal to the probability of site occupation at time $t + 1$ minus the probability of site occupation at time t:

Eq. (12) $\Delta p = p(t+1) - p(t) = p^2 - p.$

If the change in time Δt is 1, the ratio of changes $\Delta P/\Delta t$ is also $p^2 - p$. Since the time interval in the lattice is of unit size, the difference ratio $\Delta P/\Delta t$ can be written as the discrete-time derivative

Eq. (13) $$\frac{\mathrm{d}p}{\mathrm{d}t} = p^2 - p\,,$$

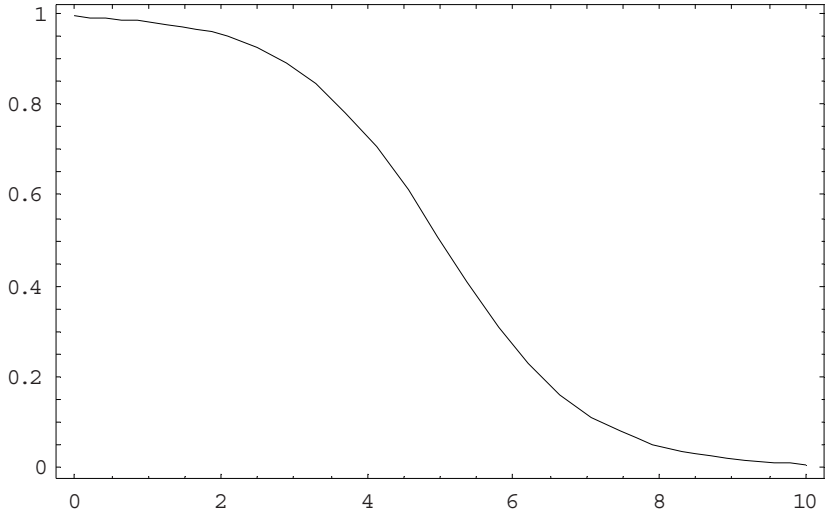
which is a differential equation defining the logistic cumulative distribution as the distance between adjacent sites (Δt) goes to zero.

The cumulative logistic distribution is usually defined as the function with derivative $p(1 - p) = p - p^2$, in other words, the negative integral of our expression (cf. Lloyd, 1967 for a review of standard derivations). However, the two forms are complementary and define the same distribution. First, with the addition of arbitrary constants (α, β), the solution of the differential equation $p^2 - p$ (for positive values of the constants) is a monotonically decreasing logistic function

Eq. (14) $$p(t) = \frac{\alpha}{\beta + \exp(\alpha t + \alpha C)}\,.$$

(C is a constant of integration.) See Figure 3.

Figure 3 The solution of $p^2 - p$

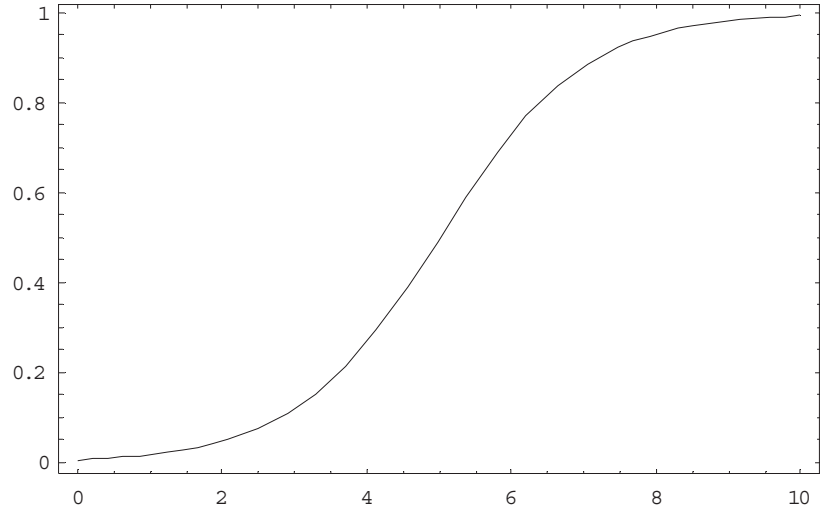


For the same constants the equation

$$\text{Eq. (15)} \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = p - p^2,$$

defines a monotonically increasing logistic function (Figure 4) that is the complement (in the sense that the abscissas for a given ordinate sum to 1) of the function shown in Figure 3.

Figure 4 The solution of $p^2 - p$

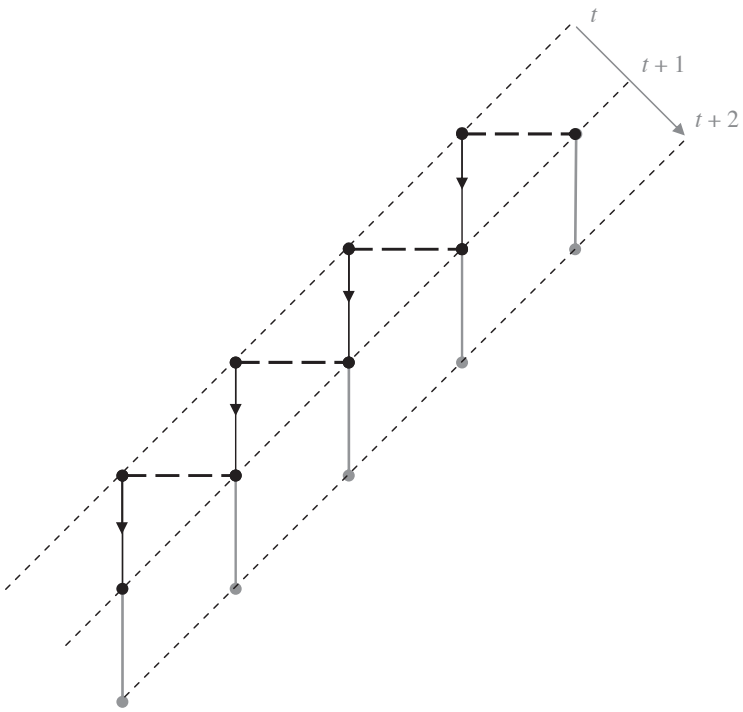


Indeed, this follows from the fact that $P = \text{expit}(y)$ and $Q = \text{expit}(-y)$ for the logistic distribution. Thus, the complement of the change in the probability of site occupation is the change in the probability that a site is not occupied.

We now extend the description for a single pair of sites so that the contact process constitutes a percolation model. First, we consider the necessary conditions for the process to be logistic in form; then we ask under what conditions the process can percolate the lattice. Suppose at time t there are N sites, of which a proportion $p = n/N$ are occupied (Figure 5; for convenience occupied sites are shown as adjacent). Each site at time t connects to two sites in the positive-time direction (diagonally downwards to the right). Suppose further for the moment that site occupation can proceed only vertically downwards, in the direction of the solid-black arrows (see Figure 5). Then induction on t shows that the probability of site occupation evolves according to

Eq. (16) $$\frac{\mathrm{d}p}{\mathrm{d}t} = p^2 - p\,,$$

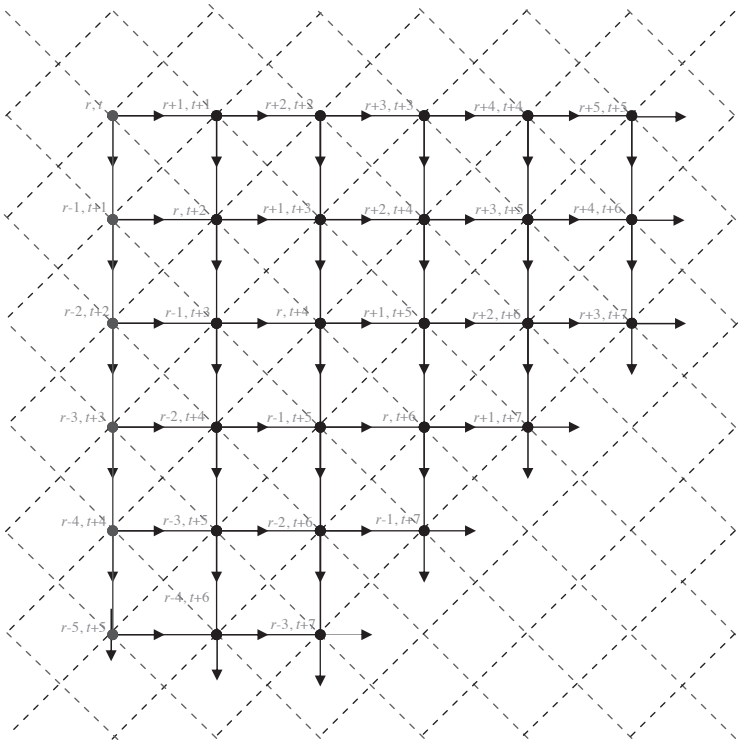
Figure 5 Vertical propagation cannot fill the square lattice



as previously. Restricting propagation to the vertical (resp. horizontal) direction, at any point in time each of the sites occupied with probability p can propagate to one other site only, which is unconditionally occupied with probability p^2. Thus, the logistic form arises by constraining propagation from an occupied site to one other site only.

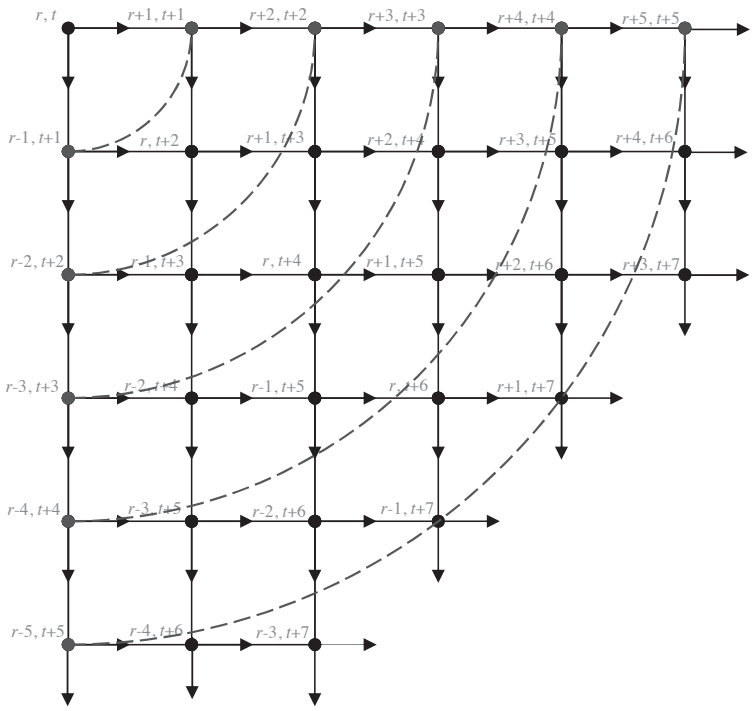
However, this process can percolate the lattice only in a single direction. Although occupied adjacent sites are connected (in Figure 5, horizontal dashed lines are used for these passive connections), propagation in the vertical (resp. horizontal) direction only does not percolate the lattice in both directions. Since the restriction is necessary to preserve the logistic form, it would appear that a logistic contact process cannot percolate the lattice bidirectionally. To emphasize this point, suppose the process initiates at a single site at location r

Figure 6 Square lattice and superimposed epidemic lattice. Without bidirectional (orthogonal) propagation only the part of the lattice shown in grey can be occupied.



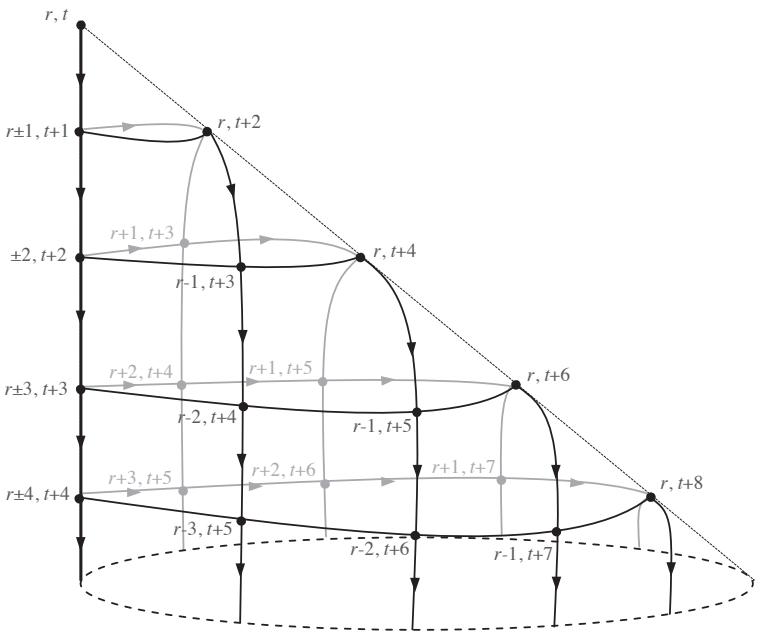
chosen at random at time t (Figure 6). For concreteness, suppose it represents an infectious disease with a high probability of transmission ($1 - p << 1$). Unless infection can proceed both horizontally and vertically, the potential epidemic will be restricted to the single grey vertical (resp. horizontal) line defined by the points $(r - n, t + n)$. On the other hand, if each site can infect both downstream sites, that is $(r, t) \rightarrow (r - 1, t + 1)$ and $(r, t) \rightarrow (r + 1, t + 1)$, the probability of infection is no longer logistic (cf. Cardy, Table 10.1, p. 203). Notice, with reference to Figure 6, that an epidemic defines a superimposed square lattice, rotated by 90 degrees with respect to the original one.

Figure 7 Epidemic lattice with cyclic boundary condition



Thus, while a logistic differential equation is a valid representation of the epidemic when each site potentially infects only one other, a modification is required if infection probability is to evolve logistically and the epidemic is to be capable of percolating in both directions. If the points connected by the grey dashed arcs in Figure 7 are superimposed, bidirectional transmission can proceed freely. This modification is equivalent to applying a cyclic boundary condition. In other words, the lattice is folded upon itself on the line defined

Figure 8 Epidemic lattice after folding



by the points $(r + n, t + n)$ and $(r - n, t + n)$ (Figure 8). With the folding, logistic evolution is maintained; moreover, although each site can infect only one other, the entire lattice can now be percolated. Finally, only space is made cyclical by the boundary condition, so the folding respects temporal irreversibility (Figure 9).

With the exception of three points whose connectivity is changed by the folding, the epidemic domain (i.e., the conical part of the lattice) remains a square lattice, albeit non-planar. The affected points, whose post-folding connectivity is 1, 3 and 3, respectively, are (r, t), $(r, t + 2)$ and $(r \pm 1, t + 1)$. Thus, with the exception of these points, another planar square lattice can be obtained by projecting the epidemic domain from the apex of the cone onto a flat surface. In the projected lattice, the positive time direction is radially outwards from the image of the apex. The projection shows that, despite the changes imposed by the folding, the epidemic lattice can cover the plane (Figure 10). It also suggests that the logistic contact process can serve as a model for spatial diffusion from a point source.

Figure 9 Silhouettes of the *r*−*t* lattice following the folding (top view and side view; time shown in light gray dashed lines, location in dark gray)

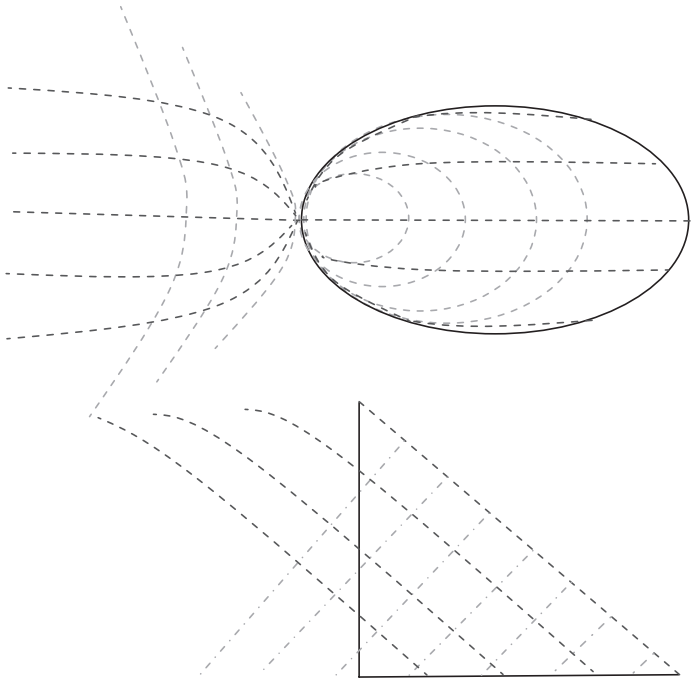
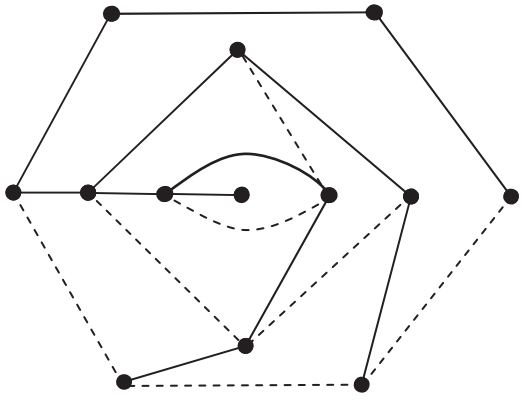


Figure 10 Projection of the intial part of the epidemic lattice onto a plane



The logistic equation as a hyperbolic conservation law

An important class of random-walk processes with zero-mean flux on the lattice converges to the Gaussian probability distribution (Kipnis and Landim, 1999). Brownian motion in $\mathbf{R}^n$ is another such class (Gardiner, 2004). Unlike Gaussian processes, since the particles flow in a characteristic direction, logistic diffusion is not zero-mean flux and is therefore described in the limit by a first-order hyperbolic conservation law rather than by variants of the second-order heat equation (Kipnis and Landim, 1999).

Although folding the lattice changes the geometry of the epidemic domain it leaves lattice connectivity largely unchanged. On the folded lattice, at any point in time the change in the probability of site occupation is

$$\text{Eq. (17)} \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = p^2 - p \,,$$

as before. Similarly, the change in the probability of a site not being occupied is

$$\text{Eq. (18)} \qquad \frac{\mathrm{d}p}{\mathrm{d}t} = p - p^2 \,,$$

as before. Site occupation can percolate the entire lattice. The standard logistic distribution is thus the limiting distribution of the contact process as the number of points goes to infinity and as the distance between points (Δt) goes to zero. It follows that the "damping" imposed by logistic growth can be explained solely as a consequence of the lattice embedding, without the need to invoke density-dependent arguments.

The logistic description of the contact process is hydrodynamic in the sense that, in the above limits, the parameter p evolves smoothly according to a differential equation (Kipnis and Landim, 1999). The hydrodynamic limit provides the justification for associating the discrete random walk with the continuous distribution. As previously suggested, moreover, with the lattice embedding, the Verhulst ordinary differential equation (expressing change with respect to time only) defines a partial differential equation (expressing change with respect to time and position). That is, on the lattice, the logistic

equation is a model of spatial as well as temporal dynamics. Specifically, the Verhulst equation can now be characterized as a first-order partial differential equation of the form

Eq. (19) $$\partial_t p + \partial_r f(p) = 0$$

(Kipnis and Landim, 1999), where $f(p)$ is the (mean) flux of the quantity p across sites, and r is location. Applying the chain rule, the equation can be rewritten in "quasilinear" form

Eq. (20) $$\partial_t p + a(p)\partial_r p = 0$$

(Bressan, 2000), where $a(p) = f'(p)$ is the derivative of flux. A system of such equations where $A(p) \equiv Df(p)$ has all real eigenvalues defines a "hyperbolic system of conservation laws" for p, and the eigenvalues of A are called the wave speeds of p (LeFloch, 2002).

In general, solutions to first-order quasilinear hyperbolic equations are not well behaved, and develop discontinuities with the characteristics of "shock waves" (Bressan, 2000). Nevertheless, if $a(p)$ is a constant a, the relation is an example of a linear advection equation

Eq. (21) $$\partial_t p + a\partial_r p = 0$$

(LeFloch, 2002), a subclass of hyperbolic conservation laws with explicit solutions. The logistic differential equation is a hyperbolic conservation law of this type since, in the lattice embedding, $\partial_r p$ is equal in the limit to $p^2 - p$. As noted previously, the intrinsic growth rate (here called the wave speed, a) can be set to 1 by a change of units, now seen to be the natural coordinate system defined by the eigenvalues of A. Thus, the Verhulst equation is a linear advection equation with an explicit analytical solution in terms of the logistic distribution.

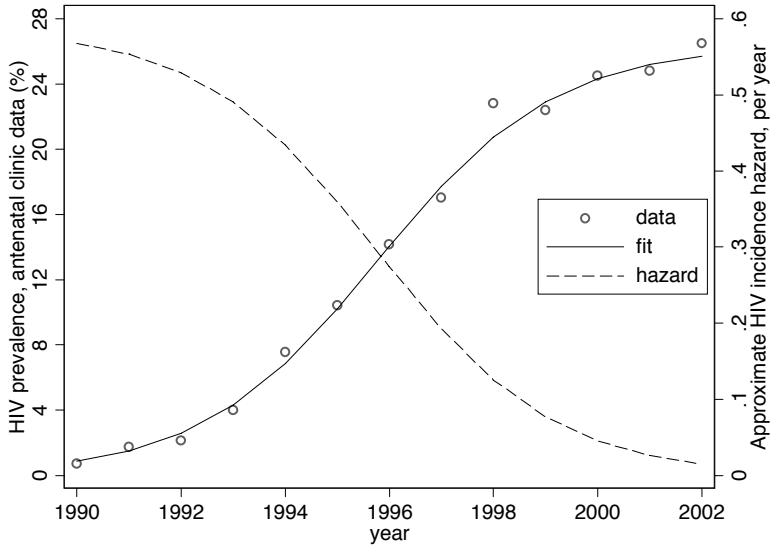
To summarize: Constant flux implies the possibility of a "natural" description of flows or transmission in terms of the logistic rather than the Gaussian distribution. Thus formal analogies between the distributions (Cramer, 2004; Banks, 1994; Cox DR and Snell, 1989) are complemented by parallel derivations as limiting behaviors in particle systems.

Logistic distributions in practice

Banks (1994) cites numerous applications of the logistic distribution to growth problems that might be more naturally modeled as diffusions according to the contact process described here. We mention two additional examples relevant to epidemiology.

National data on HIV prevalence have been collected by the South African Department of Health, Pretoria, at antenatal clinics in South Africa since 1990. The pooled data, fitted model and hazard are shown in Figure 11. Overall coherence of the data with the logistic distribution suggests that the basic features of logistic diffusion may be relevant to understanding the infection process. For example, HIV infection in women of childbearing age appears to proceed according to a simple random walk.

Figure 11 Time series of HIV prevalence and infection hazard in South African antenatal clinics



Some non-transmissible outcomes are also well modeled by logistic diffusion. For example, suppose in a time period t an individual has a probability p of not experiencing transition to an absorbing state. The state of the individual

in each period is represented by a lattice site adjacent to his or her site in the previous period. Not experiencing the transition by t means the lattice sites representing that individual's state are occupied in that and in all previous periods.

Now suppose that, for an individual not having experienced the event, the probability of not experiencing the event is p in every period. If the event does not occur by t, the probability of not experiencing the event by the subsequent period is p^2. Conversely, if transition to the absorbing state has already occurred by t, the probability of not experiencing the event by $t + 1$ is then zero. In this model, the individual can be said to "interact" only with his or her state and only "locally" (i.e., across adjacent time periods). Under these assumptions the probability of an individual's experiencing the event follows a logistic diffusion in the limit of a small time interval.

Provided each individual shares the common parameter p, the model can be extended to a population by adjoining the corresponding number of sites in the lattice. Such a model comprises a series of collections of Bernoulli trials with parameter p, which is the standard rationale for logistic regression. In the limit of many individuals, not only the evolution of the cumulative probability but also the series of realized outcomes will follow a logistic distribution. An example is time-to-failure data for breastfeeding by month of age, where non-breastfeeding is the absorbing state, and the prevalence of non-breastfeeding by month of age is well approximated by a logistic survival distribution (Lauer et al., 2004; see the reference for a detailed set of figures).

The logistic distribution in models of infection

SI model

The logistic form arises from the standard susceptible–infective (SI) model of disease transmission, called by Daley and Gani (1999) "the simple epidemic in continuous time." We show that it is also the limiting form of a related discrete process. In the SI model, the equation of transmission (i.e., for the number of infectives, y, in a population of size N) is given by

$$\text{Eq. (22)} \qquad \frac{\mathrm{d}y}{\mathrm{d}t} = \beta xy = \beta y(N - y),$$

where x is the number of susceptibles. The solution can be written

Eq. (23) $$y(t) = \frac{y_0 N}{y_0 + (N - y_0)\exp(-\beta N t)},$$

which is equivalent to logistic survival. The relation giving rise to this form is the "law of mass action", i.e., that the rate of change in infectives is proportional to the product of infectives and susceptibles (where β is per-capita probability of infection).

The classical SI model thus relies on an assumption that susceptible and infective populations are "well mixed". As opposed to logistic diffusion, the law of mass action proceeds from a macroscopic law about the ensemble of particles (in particular, their equiprobability of interaction) and "averages down" to the level of the individual particle. This is conceptually very different than an individual-level microscopic law that is aggregated up to obtain the law for the ensemble. Nevertheless, the law of mass-action is also an exact derivation of the logistic distribution.

SIR model

The SIR model (Kermack and McKendrick, 1927) also has a fundamental relationship to the logistic distribution.

Although the SIR model does not have an exact solution, an approximate solution is obtained by means of a series expansion, yielding the result (Daley and Gani, 1999)

Eq. (24) $$z(t) \approx \frac{\rho^2}{x_0}\left(\frac{x_0}{\rho} - 1\right) + \frac{\alpha\rho^2}{x_0}\tanh\left(\frac{1}{2}\gamma\alpha t - \varphi\right),$$

where $z(t)$ is the number of removals at time t, $\rho = \gamma/\beta$ is the relative removal rate, x_0 is the initial number of susceptibles and α and φ are combinations of other constants.

This result can be written in logistic form by means of the identity

Eq. (25) $$\frac{c_1 - c_2 + (c_1 + c_2)\exp(2(c_3 t + c_4))}{1 + \exp(2(c_3 t + c_4))} = c_1 + c_2 \tanh(c_3 t + c_4),$$

where

$$\text{Eq. (26)} \qquad c_1 = \frac{\rho^2}{x_0}\left(\frac{x_0}{\rho} - 1\right),\ c_2 = \frac{\alpha\rho^2}{x_0},\ c_3 = \frac{1}{2}\gamma\alpha,\ \text{and}\ c_4 = -\varphi\,.$$

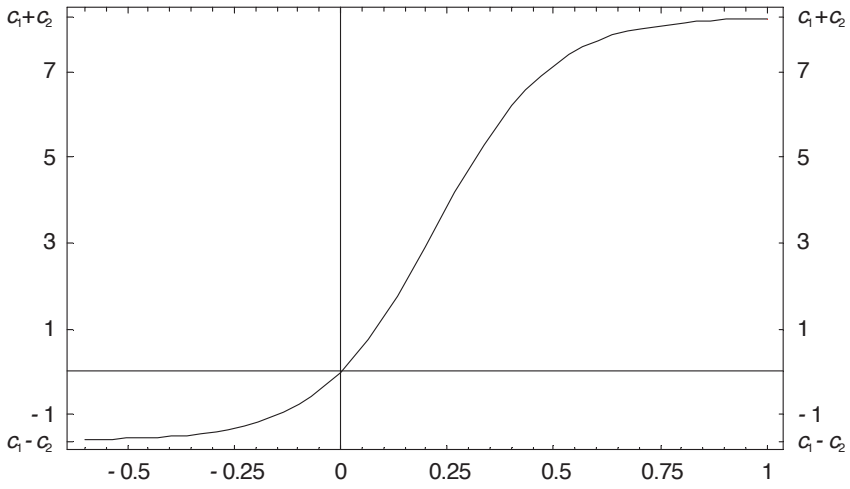
The lower asymptote of the logistic curve is set by $c_1 - c_2$, the upper asymptote by $c_1 + c_2$, and the rate and the level (i.e. the $z(t) = 0$ intercept) of logistic growth by c_3 and c_4, respectively (Figure 12). As shown by Kendall (1956), this approximate solution is exact if per-capita force of infection obeys the following density dependence

$$\text{Eq. (27)} \qquad \beta(z) = \frac{2\beta_0}{(1 - \frac{z}{\rho}) + (1 - \frac{z}{\rho})^{-1}},$$

where $\beta 0$ is the original (constant) per-capita force of infection (Daley and Gani, 1999).

The approximate solution from series expansion is a description of the SIR epidemic even when there is no density dependence in the force of infection.

Figure 12 Removals are a logistic function of time in the approximate solution to the SIR model



In that solution, the evolution of susceptibles $x(t)$ is a quadratic polynomial ($a + bu + cu^2$) in terms of the logistic argument u

Eq. (28) $$x(t) = \frac{x_0^2 + \rho^2}{2x_0} + c_2 \tanh(c_3 t + c_4) + \frac{1}{2}\alpha c_2 \tanh^2(c_3 t + c_4),$$

and infectives also evolve according to a quadratic polynomial ($a' + c'u^2$) in the logistic term

Eq. (29) $$y(t) = -2x_0 N + x_0^2 + 2x_0\rho - \rho^2 + \frac{1}{2}\alpha c_2 \tanh^2(c_3 t + c_4).$$

Thus, up to constant terms, the infective series is the square of a simple logistic series. Furthermore, the quadratic logistic $a' + c'u^2$ has a valid interpretation as an independent differential equation model, that is, the solution of the evolution equation

Eq. (30) $$\frac{dp}{dt} = p(1-\sqrt{p}) = \sqrt{p}(\sqrt{p} - p) = p - p^{3/2}.$$

It follows that model parameters corresponding to a "square-logistic" process (such as an infectives time series in an SIR epidemic) can be estimated by transforming the time series in terms of a "logit-square-root" model:

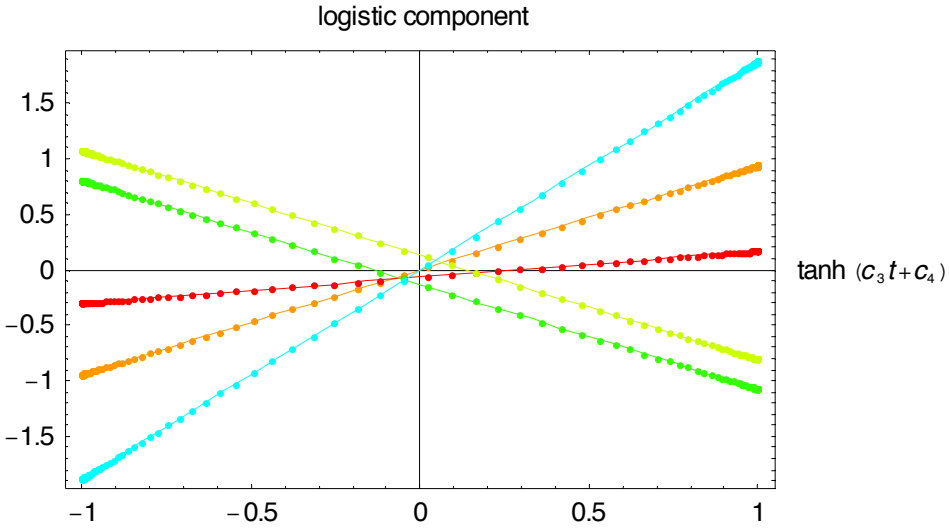
Eq. (31) $$\log\left(\frac{\sqrt{p}}{\sqrt{a} - \sqrt{p}}\right) = c_3 + c_4 t,$$

where a is the upper asymptote of the series. In microscopic terms, square-logistic diffusion requires that the probability of site occupation at $t + 1$ be inflated by a factor $1/\sqrt{p}$ in comparison with simple logistic diffusion. In continuous time, this is a hyperbolic conservation law with non-constant wave speed $a(p) = \exp[\log(p)\cdot 2^{-t}]$. The approximate SIR equations for removals and infectives therefore constitute a system of hyperbolic conservation laws.

While it would be natural to extend that system to include the logistic form for susceptibles ($a + bu + cu^2$), there appears to be no obvious way to write the wave-speed equation for this quantity, at least in terms of a contact process on a lattice. Nonetheless, since susceptibles are merely the complement of infectives and removals, we conjecture that there exists a lattice or other graph for which the general logistic quadratic form can be represented as a diffusion with non-constant wave speed.

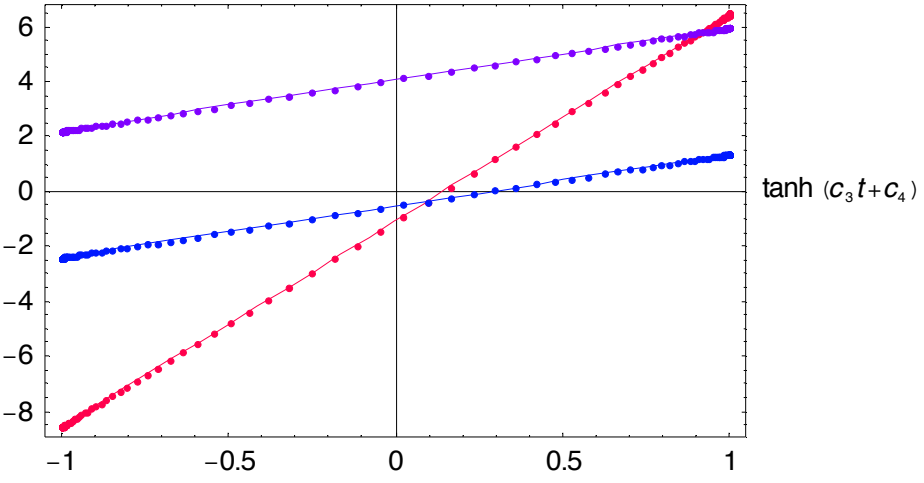
In any case, the approximate SIR model can be factored as five separate logistic curves, where each population element $z(t)$, $x(t)$ and $y(t)$ is expressed as either a simple logistic term or as the product of two logistic terms (Figure 13).

Figure 13 Kendall's SIR model can be factored into five simple logistic curves: *z(t)*, cyan line, *x(t)*, the product of the red and light-orange lines plus a constant, *y(t)*, the product of the bright-green and lime-green lines plus a constant



Finally, in Kendall's exact model, the function $\beta(z)$ can be written as the ratio of a linear and a quadratic logistic (Figure 14).

Figure 14 Per-capita infectiousness in Kendall's exact SIR model is the ratio of a logistic curve (red line) over the product of two logistic curves (blue and lavender lines)



Generalized linear modeling

In a generalized linear framework with a quadratic loss function (McCullagh and Nelder, 1989), estimation minimizes the quantity:

Eq. (32) $$(y-m)'V^{-1}(y-m)$$

where V is the covariance matrix of the data y and m is the vector of sample means. The normal equations for least-squares estimation are thus

Eq. (33) $$\frac{\partial m'}{\partial \beta}V^{-1}(y-m)=0.$$

For the exponential family of distributions, the normal equations are the same as the score equations from the loglikelihood (Cox C, 1996). Thus, within this family, in either a least-squares or a likelihood framework, estimation employs an identical set of differential equations for the regression.

The generalized-linear framework requires specification of a link function, h, such that:

Eq. (34) $$\eta = h(\mu),$$

where μ is the conditional mean and $\eta = x'\beta$ is the linear predictor; x may include a constant (McCullagh and Nelder, 1989). For various distributions, there is a "canonical" link for which there is a sufficient statistic of dimension equal to that of the model manifold (i.e., the span of the linear predictor in the regression space). For Bernoulli outcomes with probability parameter p, $\mu = p$ and the canonical link is the logit (McCullagh and Nelder, 1989), so that the inverse link is logistic. Taking the derivative of the mean yields

Eq. (35) $$\frac{\partial \mu}{\partial \eta} = \mu - \mu^2 = p(1-p)$$

Thus, for the set of applications described here, the canonical link for Bernoulli regression can be viewed as arising from the dynamics of logistic diffusion.
In a quasi-likelihood framework, the distribution of the data is not characterized beyond a variance relation (Wedderburn, 1974), and the canonical link is defined as the function satisfying

Eq. (36) $$\frac{\partial \mu}{\partial \eta} = \nu(\mu)$$

where $\nu(\mu)$ specifies the dependence of the variance on the mean (McCullagh and Nelder, 1989; Cox C, 1996). Although the logistic distribution is not a member of the exponential family (Cox DR and Snell, 1989), because the Bernoulli variance is defined by the logistic link it still has identical quasi-likelihood, likelihood and least-squares estimators.

Even without a restriction on distributional family, the variance function imposes strong restrictions on the data. In a generalized-linear framework these restrictions enter either implicitly (through specification of the dependence of μ on x) or explicitly (through specification of the dependence of the variance of y on μ). In either case, the estimation procedure produces a differential equation for the regression. When η is a linear function of time, choice of the link/variance function is equivalent to selection of a dynamic model for the data-generating process.

The logistic regression model states that regression is linear after the logit transformation. For transmissible outcomes (such as infection, drug addiction, product preferences, and some forms of time-to-failure data), the results we have described suggest that, as a consequence of underlying diffusion-like phenomena, the individual Bernoulli outcomes may sometimes follow a logistic regression on time (or, in birth-cohort models, age). When considering additional covariates, logit-additive approximations to varying temporal effects may sometimes be expected based on mechanical considerations. Furthermore, if the covariate levels define separate diffusion compartments within which we have logit-linearity in time (i.e., if the covariates define a system of hyperbolic conservation laws), the overall model will be logit linear with time-covariate product terms ("interactions").

Conclusion

We have described how the logistic distribution can be derived from a description of the microscopic dynamics of diffusion on the square lattice, and that with this embedding it has both a temporal and a spatial interpretation. Its appearance in various applications to population dynamics, epidemiology and ecology could thus be predicted from its status as the limiting distribution of a stochastic process that is a plausible model for generating the data. In particular, a family of polynomial representations of the logistic distribution plays a fundamental role in classical models of infection, at least as an approximation. In special cases of importance these relations induce a correspondence between dynamics of the data-generating process and classical estimation processes.

Acknowledgements

Brian G. Williams, Edward A. Frongillo, Jr., Philippe G. LeFloch, Claudio Landim, Carel Pretorius, Nico Nagelkerke, J. S. Cramer and Ekkehard Kopp are gratefully acknowledged for their helpful comments.

References

1. Aldrich JH, Nelson FD. Linear probability, logit, and probit-models**.** Beverly Hills, California, USA, Sage Publications, 1984 [Lewis-Beck MS (Series Editor) *Quantitative Applications in the Social Sciences, No. 07-045*.]

2. Armitage P, Berry G, Matthews JNS. Modelling categorical data. In *Statistical methods in medical research*. 4th Edition. Oxford, England, Blackwell, 2002:485-502.

3. Banks RB. *Growth and diffusion phenomena: mathematical frameworks and applications*. Berlin, Springer, 1994.

4. Bressan A. *Hyperbolic systems of conservation laws: the one-dimensional Cauchy problem*, Oxford University Press, 2000.

5. Cantrell RS, Cosner C. *Spatial ecology via reaction-diffusion equations*. Chichester, England, John Wiley and Sons Ltd, 2003.

6. Cardy J. *Scaling and renormalization in statistical physics*. Cambridge, Cambridge University Press, 1996.

7. Cox DR, Snell EJ. *Analysis of binary data*, 2nd ed. London, Chapman and Hall/CRC, 1989 [*Monographs on statistics and applied probability, 32*].

8. Cox C. Nonlinear quasi-likelihood models: applications to continuous proportions**.** *Computational Statistics and Data Analysis* 1996, 21:449-461.

9. Cramer JS. The early origins of the logit model. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* (*Stud Hist Philos Biol Biomed Sci*). 35(4): 613-626, 2004.

10. Daley DJ, Gani J. *Epidemic modelling: an introduction*. Cambridge, Cambridge University Press, 1999.

11. Dickman R. Critical phenomena at absorbing states. In *Nonequilibrium statistical mechanics in one dimension*, Edited by Privman V. Cambridge University Press, 1997.

12. Durrett R. Oriented percolation in two dimensions, *Annals of probability*, 1984, 12(4):999-1040.

13. Enright JT. Climate and population regulation: the biogeographer's dilemma. *Oecologia*, 1976, 24:295-310.

14. Fisher RA. The wave advance of advantageous genes, *Annals of eugenics*, 1937, 7:355-369.

15. Fletcher RI. The quadric law of damped exponential growth, *Biometrics*, 1974, 30:111-124.

16. Greenland S, Lash TL. Bias analysis. Ch. 19 in *Modern Epidemiology*. 3rd Edition. Edited by Rothman KJ, Greenland S, Lash TL. Philadephia, Pennsylvania, USA: Lippincott Williams and Wilkins; 2008, 345-380.

17. Gardiner CW. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, 3rd ed. Berlin, Springer, 2004.

18. Greenland S. Generalized conjugate priors for Bayesian analysis of risk and survival regressions. *Biometrics*, 2003, 59:92-99.

19. Greenland S. Prior data for non-normal priors. *Statistics in Medicine*, 2007, 26:3578-3590.

20. Grimmet G. *Percolation*, 2nd ed. Berlin, Springer, 1999 [Volume 321, A series of comprehensive studies in mathematics].

21. Harris TE. Contact interactions on a lattice. *Annals of probability*, 2(6): 969-988, 1974.

22. Hosmer D, Lemeshow S. *Applied logistic regression*, 2nd ed. John Wiley and Sons, New York, 2000.

23. Jones MC. Families of distributions arising from distributions of order statistics. *Test*, 13:1-43, 2004.

24. Kendall DG. Deterministic and stochastic epidemics in closed populations. *Proceedings of the Third Berkeley Symposium in Mathematical Statistics and Probability*, 4:149-165, 1956.

25. Kermack WO, McKendrick AG. A contribution to the mathematical theory of epidemics, *Proceedings of the Royal Society of London, Series A*, 115:700-721, 1927.

26. Kipnis C, Landim C. *Scaling limits of interacting particle systems*. Berlin, Springer, 1999.

27. LeFloch PG. *Hyperbolic systems of conservation laws: the theory of classical and nonclassical shock waves*. Basel, Birkhauser, 2002.

28. Liggett TM. *Interacting particle systems*. Berlin, Springer, 2005.

29. Lauer JA, Betrán AP, Victora CG, de Onís M, Barros AJD. Breastfeeding patterns and exposure to suboptimal breastfeeding among children in developing countries: review and analysis of nationally representative surveys. *BMC Medicine*, 2004, 2:26.

30. Lloyd PJ. American, German and British antecedents to Pearl and Reed's logistic curve. *Population studies*, 21:99-108, 1967.

31. McCullagh P, Nelder JA. *Generalized linear models*, 2nd Ed. Cambridge, England: Chapman and Hall, 1989:1-19.

32. Rudin W. *Principles of Mathematical Analysis*, 3rd ed. New York, McGraw-Hill, 1976.

33. Spitzer F. Interaction of Markov processes. *Advances in mathematics*, 5: 246-290, 1970.

34. Verhulst P-F. Notice sur la loi que la population suit dans son accroissement. *Correspondance mathématique et Physique, publiée par A. Quetelet*, 10:113-120, 1838.

35. Wedderburn RWM. Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* 1974, 61:439-447.

Chapter 6

Breastfeeding patterns and exposure to suboptimal breastfeeding among children in developing countries: review and analysis of nationally representative surveys

Based on: Lauer JA, Betrán AP, Victora CG, de Onís M, Barros AJD. Breastfeeding patterns and exposure among children in developing countries: review and analysis of nationally representative surveys. *BMC Medicine*, 2004, 2:26.

Abstract

Background: Suboptimal breastfeeding is associated with higher mortality among infants and young children in the developing world. We describe patterns in 'exclusive breastfeeding' and 'any breastfeeding' rates and quantify exposure to suboptimal breastfeeding among children aged two years or younger in developing countries.

Methods: We reviewed nationally representative surveys that collected data on breastfeeding rates in 94 developing countries. Surveys were categorized by completeness and comprehensiveness of data. Complete and comprehensive data were analysed with minimum chi-square regression. With a fitting procedure, estimated parameters were used to impute missing observations for incomplete or non-comprehensive surveys. Breastfeeding indicators were calculated and are reported for 135 developing countries by UN region.

Results: Amongst infants aged six months or younger in the developing world, the prevalence of exclusive breastfeeding is 39% and the prevalence of no breastfeeding is 5.6%. The prevalence of continued breastfeeding is 86% and 68% for infants and children aged 6-11 and 12-23 months, respectively, in the developing world. Imputation expands population coverage of indicators, especially for infants. Breastfeeding trends are highly linear and estimated parameters defining the age-specific attrition hazard are robust. Survey-reported rates, particularly for exclusive breastfeeding, appear to have systematic upward bias, and exposure estimates must be considered conservative.

Conclusions: Compliance with breastfeeding recommendations in developing countries is low, and more attention should be given to increasing breastfeeding - especially exclusive breastfeeding - and to monitoring trends. Although the introduction of more standardized and better validated survey instruments is desirable, since data coverage, completeness and comprehensiveness are extensive, global exposure assessment is relatively robust. Moreover, the regularity of breastfeeding patterns show existing survey data capture real biological and social phenomena. Our method for the analysis of breastfeeding rates provides a potent tool for summarizing trends, validating observations, translating and extrapolating indicators (as well as projecting and imputing estimates when necessary) and should support more effective child health monitoring.

Background

Objectives

For almost all infants, breastfeeding remains the simplest, healthiest and least expensive feeding method that fulfils the infant's nutritional needs. The prevalence and duration of breastfeeding are therefore recognized as important health indicators, and their impact on infant and child health has been frequently studied [1-3]. The aim of this study is to describe patterns in 'exclusive breastfeeding' and 'any breastfeeding' rates reported in complete, comprehensive and nationally representative surveys and to quantify exposure to suboptimal breastfeeding in 135 developing countries among children aged two years or younger. 'Suboptimal breastfeeding' is used as a generic term to denote exposure to any increased risk relative to age-specific minimum risk.

Policy background

Recent policy debate has focused on the optimal duration of exclusive breastfeeding in infancy, and WHO commissioned a systematic review [1] to elucidate the issue. In 2001, the World Health Assembly passed a resolution recommending exclusive breastfeeding for the first six months of life as a global public health recommendation [1,2]. International consensus is that optimal breastfeeding practice for infants and young children consists of exclusive breastfeeding for the first six months of life, with continued breastfeeding up to two years of age and beyond [2,3].

Exposure categories

The term 'category' as used here refers primarily to individual status. For infants ≤6 months of age, exclusive breastfeeding represents 'theoretical minimum' [4] exposure. 'Exclusive breastfeeding' means the infant receives only breast milk from the breast, or expressed breast milk, and receives no other liquids or solids with the exception of drops or syrups consisting of vitamins, mineral supplements or medicines [5].

Ideally, exposure to suboptimal breastfeeding for infants ≤6 months of age would be measured as a continuous variable. Such a variable might express,

for example, the proportion of energy, water and nutrients in the diet derived from breast milk. However, since all survey-reported data on breastfeeding are categorical, and in view of the fact that there is heterogeneity in the categories employed, we define two categories: 'partial breastfeeding' and 'no breastfeeding' for the purpose of measuring exposure to increased risk among infants ≤6 months of age. The defining characteristic of partial breastfeeding is that the infant ≤6 months of age receives some breast milk, but not exclusively.

For infants >6 months and children ≤2 years of age, theoretical minimum exposure is defined as 'continued breastfeeding', and 'no breastfeeding' is the sole exposure category of increased risk. The defining characteristic of continued breastfeeding is that the infant >6 months or child ≤2 years of age receives at least some breast milk regardless of the quantity or the presence of other foods or liquids in the diet. For all infant and child age groups, the defining characteristic of 'no breastfeeding' is that the infant or child receives no breast milk.

For analysis purposes we further define the category 'any breastfeeding', which refers to infants or children receiving breast milk regardless of quantity or the presence of other foods or liquids in the diet. For infants ≤6 months of age, 'any breastfeeding' is equivalent to the category 'exclusive or partial breastfeeding', and for infants >6 months and children ≤2 years of age is identical to the category 'continued breastfeeding'. For children >2 years of age analysed in subregional estimation models, 'any breastfeeding' is used merely as a descriptive category without regard to risk status.

For infants >6 months of age, the failure to provide safe and appropriate complementary foods has been identified as a risk factor. We do not investigate this exposure in the present study.

Breastfeeding indicators

For population-level assessment, exposure categories are aggregated across individuals and expressed as indicators. Therefore, breastfeeding indicators, in the sense used here, are primarily summary measures of population-level exposure with direct relevance for health outcomes. Although many types

of indicator are in common use, the breastfeeding indicators reported here are measures of cross-sectional prevalence for the defined exposure categories and age ranges. Cross-sectional prevalence is also called 'point prevalence' or 'period prevalence'.

Since surveys include prevalence estimates for states that may only indirectly be a measure of exposure according to our definitions, some of our indicators (complex indicators) represent sums of other indicators. For example, partial breastfeeding represents an aggregation of survey feeding categories such as 'breast milk plus formula', 'breast milk plus solid foods', 'predominant breastfeeding', and so on. Continued breastfeeding likewise represents an aggregation of various indicators. However, exclusive breastfeeding and no breastfeeding are simple, not complex, indicators.

Breastfeeding and HIV

Breastfeeding for more than one year has been estimated to pose a 10-20% risk of HIV transmission in children of infected mothers [6]. The risk of HIV transmission needs to be balanced, however, against competing risks for infant mortality and morbidity. WHO states that 'when replacement feeding is acceptable, feasible, affordable, sustainable and safe, avoidance of all breastfeeding by HIV-positive women is recommended; otherwise, exclusive breastfeeding is recommended during the first months of life; and that those who choose other options should be encouraged to use them free from commercial influences' [6,7].

Since recommended practice depends on mothers' HIV status and other individual factors, only an individual-level assessment can explicitly account for the offsetting risks posed by mother-to-child transmission of HIV through breastfeeding. The population-level indicators reported here cannot measure this risk in a manner consistent with current international infant-and-child-feeding recommendations. In a related study, we use an outcomes-based assessment to estimate the potential magnitude of the risks of mother-to-child transmission of HIV through breastfeeding (results not shown).

Methods

Data sources

Data sources are nationally representative surveys - published and unpublished - that we identified as of June 2002 with data on breastfeeding in 94 developing countries. Most surveys were undertaken by the Demographic and Health Surveys (DHS) programme of Macro International, Calverton, MD, USA [8], or by UNICEF under its initiative on Multiple Indicators Cluster Surveys (MICS) [9]. About a dozen surveys were undertaken by national institutions.

For developing countries with more than one survey, the most recent was used for calculating breastfeeding indicators, but all available complete, comprehensive and nationally representative surveys that we identified were used for estimating patterns in exclusive and any breastfeeding rates. The Annex to this article presents a list of the developing countries for which data were identified, the data source used for the indicators we report, and the corresponding survey year and data category (see below).

DHS surveys report prevalence estimates for various breastfeeding indicators for children in 18 two-month age groups through the end of the third year of life (0-1, …, 34-35 months of age). DHS estimates are based on 24-hour recall, in which mothers or caregivers are asked about the infant or child's food and fluid intake during the 24-hour period preceding the interview. Because of the wide age range (comprehensiveness, or coverage in breadth), fine age stratification (completeness, or coverage in depth) and use of comparable feeding categories, DHS survey data on breastfeeding are the most complete and comprehensive, and we consider them the most reliable. They are also nationally representative. Data from DHS surveys are termed 'category A'. About 78% of the surveys used for calculating breastfeeding indicators are category A.

MICS surveys report prevalence estimates only for selected indicators, such as 'exclusive breastfeeding, 0-3 months of age', and are not as consistent as DHS surveys in covering either a given set of age groups or a given age range.

Consequently, MICS data required imputation or extrapolation (see below) of missing observations of two-month prevalence for analysis on a comparable basis with those of DHS. Like DHS, MICS surveys employ 24-hour recall, but use a questionnaire superior in some respects to those of DHS surveys (see Discussion, 24-hour recall). However, we found some evidence of low internal validity in MICS surveys, although this impression could be due merely to otherwise benign typographical errors in reports. Data from MICS surveys are considered 'category B'.

Other, non-standard, surveys report various breastfeeding indicators estimated with diverse methods for a variety of age groups. When such data were nationally representative they were included, but required imputation of missing observations of two-month prevalence for analysis on a comparable basis with those of DHS. A few surveys with data reported as cohort rather than period measures additionally required translation (see below) prior to imputation. Although it is possible that non-standard surveys are less reliable than MICS, since there was no obvious reason to conclude this, and since the sort of analysis they required for comparability with DHS surveys was similar to that required for MICS, non-standard surveys are also considered 'category B'.

Table 1. Prevalence estimates for breastfeeding indicators, by subregion and age group

Region	Breastfeeding indicator estimates (%)						
Subregion	Infants <6 months of age			Infants 6–11 months of age		Children 12–23 months of age	
	Exclusive	Partial	No	Continued	No	Continued	No
Africa	**24.9**	**71.2**	**3.9**	**91.8**	**8.2**	**69.9**	**30.1**
Eastern	41.4	56.1	2.4	95.1	4.9	75.6	24.4
Middle	19.4	79.6	1.0	96.6	3.4	76.8	23.2
Northern	36.5	53.9	9.6	77.7	22.3	49.6	50.4
Southern	8.2	75.7	16.0	70.4	29.6	46.7	53.3
Western	6.1	92.1	1.8	96.8	3.2	74.9	25.1
Asia (excluding Japan)	**44.9**	**50.7**	**4.5**	**87.5**	**12.5**	**72.4**	**27.6**
Eastern	58.6	36.3	5.1	85.7	14.3	ND[a]	ND
South-Central	42.1	55.3	2.6	93.3	6.7	78.8	21.2
South-Eastern	37.5	55.0	7.5	76.7	23.3	61.7	38.3
Western	17.7	72.0	10.3	71.3	28.7	37.3	62.7
Latin America and the Caribbean	**30.8**	**51.2**	**18.0**	**59.9**	**40.1**	**36.5**	**63.5**
Caribbean	25.8	63.8	10.4	64.6	35.4	34.2	65.8
Central America	23.4	55.0	21.6	60.2	39.8	37.0	63.0
South America	35.1	48.0	16.9	59.3	40.7	36.4	63.6
Developing countries[b]	**38.7**	**55.7**	**5.6**	**85.8**	**14.2**	**68.3**	**31.7**

[a] ND = no data. [b] The region Oceania (including developing countries Guam, Fiji, French Polynesia, New Caledonia, Papua New Guinea, Samoa, Solomon Islands and Vanuatu) was excluded, as nationally representative breastfeeding data were not available for any developing countries in Oceania.

Surveys with potential category-B data were not included unless they contained at least one observation of either exclusive, any or no breastfeeding. About 22% of the surveys used for calculating breastfeeding indicators are category B, with about half non-standard, and half MICS. The term for the implicit data category for countries with no nationally representative data on breastfeeding is 'category C'. As shorthand, we refer to countries with category-A, B or C data as category-A, B or C countries, respectively.

Available data for infants were category A in 75 countries; for children 12-23 months of age, data were category A in 73 countries. Available data for infants were category B in 19 countries; for children 12-23 months of age, data were category B in 11 countries. A total of 41 countries were category C for infants; for children 12-23 months of age, 51 countries were category C.

Reporting regions

The 135 developing countries were grouped according to the UN regional and subregional classification (see Table 1) [10]. Japan was excluded from regional and subregional estimates for Asia. The region Oceania (including developing countries Guam, Fiji, French Polynesia, New Caledonia, Papua New Guinea, Samoa, Solomon Islands and Vanuatu) was excluded, as no nationally representative breastfeeding data were obtained for any developing countries in Oceania.

Analysis regions

Although UN regions are geographical designations only, WHO has developed a subregional classification on the basis of both geographical and epidemiologic criteria. The 14 subregions represent a classification of countries in the six WHO regions according to four possible patterns of child and adult mortality [11]. Since no WHO region presents more than three of the defined patterns, there are substantially fewer than the theoretically possible 24 subregions.

The mortality patterns used in the classification are defined so as to locate a country with regard to the epidemiologic transition, and therefore represent a macro-level classification of conditions affecting demography, development

and proportional causes of death [12,13]. For estimating age trends in exclusive and any breastfeeding rates, we used a stratification defined by the nine WHO subregions found in the developing world.

Basic calculation of indicators

Basic calculation methods described in this section were used to calculate indicators for category-A countries. Other countries first required either imputation of missing observations or extrapolation or, occasionally, translation of observations prior to calculation of indicators.

Country indicators for no breastfeeding were calculated as weighted averages of reported estimates of the cross-sectional prevalence of no breastfeeding for two-month age groups by summing over the relevant age ranges (0-5, 6-11 and 12-23 months of age) with population weights. Population weights for aggregation across age ranges (aggregation within category-A countries) were calculated using the numbers of infants or children surveyed in each two-month age group. The country indicator for exclusive breastfeeding for infants ≤6 months of age was similarly calculated as a weighted average of the cross-sectional country prevalence of exclusive breastfeeding for two-month age groups over the range 0-5 months of age.

The country indicator for partial breastfeeding for infants ≤6 months of age was calculated as 100% – (%Exclusive breastfeeding + %No breastfeeding), where '%Exclusive breastfeeding' and '%No breastfeeding' are the country indicators (expressed as percentages) for exclusive and no breastfeeding, respectively, for the age range 0-5 months. For infants >6 months and ≤2 years of age, the country indicator for continued breastfeeding for each age range (6-11 and 12-23 months of age) was calculated as 100% – %No breastfeeding, where '%No breastfeeding' is the country indicator for no breastfeeding for the relevant age range.

Regional breastfeeding indicators were calculated as infant-population weighted averages of country indicators (aggregation between category-A countries). Estimates of infant population are based on figures published by the United Nations [13]. Weights calculated with country infant populations were also used to calculate regional indicators for children 12-23 months of age.

Subregional estimation models

We used subregional regression models to estimate age trends in exclusive and any breastfeeding rates for category-A countries. The models are stratified by the nine WHO subregions found in the developing world. Subregion-specific results were used to impute missing observations of breastfeeding rates for category-B countries in the corresponding subregions. We refer to any estimation, imputation, extrapolation, translation or projection as 'analysis'.

Breastfeeding rates were transformed with the logit function prior to analysis [14]. The logit of breastfeeding rate p is $\text{Log}(p/1-p)$, where p is a proportion (that is, p = %Rate/100). For data where, as here, country-specific observations of p are available only for defined age groups, and where the dependent variable represents a dichotomous response, minimum chi-square regression, also called Berkson-Theil weighted least squares, yields unbiased, minimum-variance estimates of regression betas [14].

The regression equation is:

$$\text{Log}(p/1-p) = a + bx + \varepsilon, \qquad \text{Equation 1}$$

where p is the breastfeeding rate, x is month of age, a and b are the parameters to estimate (regression betas) and ε is the error term. Regressions were run separately for exclusive and any breastfeeding rates in each of the nine subregions. Parameter estimates were used to predict breastfeeding rates by solving the regression equation for p:

$$\hat{p} = \frac{\text{Exp}(\hat{b} + \hat{a}x)}{1 + \text{Exp}(\hat{b} + \hat{a}x)}. \qquad \text{Equation 2}$$

The regression equation implies p is a logistic function of age.

A linear model for the logit of prevalence is equivalent to a log-linear model for prevalence odds. The log-linear odds model is one of the most common in epidemiology [15], but its use is often justified on merely pragmatic grounds [16]. However, on the assumption that breastfeeding attrition rates continually

decrease (that is, become more negative) with age, the logarithm of breastfeeding prevalence odds is constrained to be linear, implying the logit model is correctly specified for estimation of regression betas (results not shown).

Imputation

The parameters (*a*, *b*) estimated for category-A countries were the starting point for imputing missing observations of two-month prevalence for category-B countries in the same subregion. The first step was to calibrate (by changing the intercept, *b*, or by changing the slope, *a*, and the intercept, *b*) the subregional category-A trend line so as to fit available observations of the breastfeeding rate in the desired category-B country in a least-squares sense. Secondly, missing observations of breastfeeding prevalence in the desired category-B country were predicted with the calibrated trend line. Finally, indicators were calculated for the desired category-B country using the basic calculation methods described above, with the difference that population weights were calculated on the basis of the country's infant mortality rate.

For example, suppose country X in subregion Y has category-B data consisting of an observation of cross-sectional prevalence of exclusive breastfeeding of 16% for infants 0-3 months of age. The observed prevalence can be considered as a weighted average of two unobserved prevalences, for infants 0-1 and 2-3 months of age, respectively. Now suppose the subregional trend for category-A countries in subregion Y predicts (after back-transformation with the logistic function) a cross-sectional prevalence of 44% exclusive breastfeeding for infants 0-1 months of age and 34% for infants 2-3 months of age. Applying population weights calculated with country X's infant mortality rate (aggregation within category-B countries), the two predictions imply a predicted cross-sectional prevalence of exclusive breastfeeding of 40% for infants 0-3 months of age.

Keeping the slope of the estimated trend constant, the intercept is changed until the predicted prevalence of exclusive breastfeeding for infants 0-3 months of age in country X is 16%, as observed. Following calibration of the trend line to country X's observation, it was possible to predict the prevalence of exclusive breastfeeding for infants 4-5 months of age using the new intercept, b'. Again applying population weights calculated with country X's infant

mortality rate, the three predicted prevalences of exclusive breastfeeding are summed, yielding an imputed estimate for the exclusive breastfeeding indicator for country X.

Now suppose country X has additional category-B data consisting, for example, of an observation of the cross-sectional prevalence of no breastfeeding of 18% for infants ≤6 months of age. Taking 100% – 18% yields an estimate of any breastfeeding of 82% for infants ≤6 months of age. Now suppose the trend for category-A countries in subregion Y predicts a cross-sectional prevalence of 76% any breastfeeding for infants 0-1, 70% for infants 2-3, and 64% for infants 4-5 months of age. Applying population weights calculated with country X's infant mortality rate, the three predictions imply a cross-sectional prevalence of 70% any breastfeeding for infants ≤6 months of age, which is the same as an estimated prevalence of 30% no breastfeeding for the age group.

Keeping the slope of the estimated trend constant, the intercept is changed until the predicted prevalence of any breastfeeding for infants ≤6 months of age in country X is 82% (corresponding to a predicted prevalence of no breastfeeding of 18%, as observed). With the new intercept, b', and population weights calculated for the corresponding two-month age groups, it is possible to impute the prevalence of no breastfeeding for infants 6-11 months of age, as well as for children 12-23 months of age.

When, as in the above example, available data for country X contain only one observation of exclusive and no breastfeeding, it is possible to fit the observed prevalence exactly by changing the intercept of the estimated subregional trend. However, if country X has two or more observations of prevalence for a breastfeeding category, in general it will not be possible to fit all observations exactly. In such a case, it was necessary to change both the slope and the intercept to fit available observations. Here, again, although with two observations it is possible to fit them exactly by changing two parameters, if more observations are available, generally it will not be possible to fit them all exactly. When such cases arose, a least-squares approach was used, whereby the parameter values that best fit, in the least-squares sense, all available observations were chosen.

Note that the slope parameter estimates the rate of change of the logit of prevalence, while the intercept estimates the logit of prevalence at 0 months of age. We made the arbitrary decision that it was preferable to change the intercept. Therefore, as a rule of thumb, when it was necessary to change both parameters, the intercept was changed first to fit the observations as nearly as possible before changing the slope. If necessary, this two-step (intercept then slope) fitting procedure was repeated until a defined tolerance level (that is, sum of squared deviations less than a given threshold) was satisfied.

Once all available observations for country X were fit - either exactly or in the least squares sense - with the trend line thus calibrated it was possible to predict a complete and comprehensive series of cross-sectional prevalence for exclusive and any breastfeeding. However, if, in a given country, observations were available for only one rate (exclusive or any), the estimated subregional trend was used without adjustment to predict the other. In any case, once predictions were available for both exclusive and any breastfeeding, it was possible to predict rates of partial and continued breastfeeding by relying on the fact that, for infants ≤6 months of age, the percentages of exclusive, partial and no breastfeeding must add to 100%, and, for infants 6-11 and children 12-23 months of age, percentages of continued and no breastfeeding must add to 100%.

Country estimates of indicators calculated with predictions deriving from subregional estimation, whether or not a subsequent fitting procedure was performed, are called imputed indicators. To impute indicator estimates for category-C countries, the subregional averages calculated on the basis of both category-A and category-B countries were used without adjustment.

Extrapolation

If enough observations were available for a category-B country, it was possible to estimate a country-specific trend line. In such cases, observed rates were extrapolated by means of a regression on age for that country alone, and the resulting series of predicted rates was averaged with population weights to obtain estimates of indicators. Because it relies on a country's own data, this procedure is called extrapolation in order to distinguish it from imputation (applying estimates based on other countries' data). Depending on available

data, a country might have extrapolated estimates for one rate (exclusive or any) and imputed estimates for the other.

Translation between cohort and period indicators

DHS and similar surveys report estimates of cross-sectional (period) prevalence for two-month age groups. However, in a few countries, estimates were reported in longitudinal (cohort) terms, that is, as the proportion of a birth cohort remaining in a category at a specific age. Prior to analysis, longitudinal estimates were translated into cross-sectional ones by assuming the implicit cohort attrition rate applied to the other infant and child cohorts in the population, and averaging the resulting series with population weights calculated on the basis of the country's infant mortality rate. Since translation yields a complete but not comprehensive series of predictions, extrapolation or imputation was performed to obtain the series of predictions required for calculation of indicators.

Uncertainty

A random-effects model (that is, with country-specific random effects) can be used to obtain estimates for the standard error of regional estimates of breastfeeding indicators based on category-A country data. However, the kind of error analysed with a random-effects (or similar) model is that arising from observing only part of the entire population, and is termed statistical error. Depending on the validity of the survey methods and responses, category-A data may also involve systematic measurement error.

Indicators based on category-B or C data involve a further source of uncertainty deriving from use of the subregional estimation model, and also, in the case of category B, the validity and statistical error of available data. Reported indicators therefore potentially involve statistical, model and measurement error, which interact non-linearly. Model error and measurement error are not captured by random-effects estimates, and there is no general analytical method capable of taking account of these sources of error.

Results

Indicator estimates

Table 1 shows cross-sectional prevalence estimates for breastfeeding indicators by UN region. Exclusive breastfeeding rates are low, at about 25% in Africa, 45% in Asia and 31% in Latin America and the Caribbean, or 39% of infants ≤6 months of age in developing countries. However, most of the infant population in developing countries is receiving some breast milk, as the prevalence of no breastfeeding is only 5.6% for infants, ranging from about 4% in Africa to 18% in Latin America and the Caribbean. The prevalence of continued breastfeeding is about 86% for infants 6-11 months of age in the developing world, and ranges from 92% and 88% in Africa and Asia, respectively, to 60% in Latin America and the Caribbean. For children 12-23 months of age, the prevalence of continued breastfeeding drops to about 70% and 72% in Africa and Asia, respectively, and to 37% in Latin America and the Caribbean.

Uncertainty analysis performed with the random-effects model yielded confidence intervals on the order of 1-2 percentage points for each indicator (results not shown).

Population coverage of indicators

The distribution of age-specific population by data category in each UN region is shown in Table 2. In Africa and Latin America/the Caribbean, about 80% and 91% of the infant population, respectively, was covered with category-A data. In Asia 64% of the infant population was covered with category-A data.

For children 12-23 months of age, 86% of the population in Latin America and the Caribbean was covered with category-A data, as was 79% in Africa and 59% in Asia. For developing countries, 70% of infants and 67% of children 12-23 months of age were covered with category-A data.

When category-B data are considered, population coverage in developing countries increases to 95% for infants (representing about 106 million infants worldwide) and 73% for children 12-23 months of age. Inclusion of category-B data increases population coverage most in Africa (for both infants

and children 12-23 months of age) and Asia (for infants only). Lack of data for China on children 12-23 months of age substantially limits population coverage in Asia for that age group.

Subregional estimation models

Selected results for the subregional estimation models are presented in Tables 3 and 4 and Figures 1-9. These figures show estimated age trends for exclusive breastfeeding for the age groups 0-1 to 8-9 months of age (upper panels; reference line, 1% prevalence), and for any breastfeeding for the age groups 0-1 to 22-23 months of age (lower panels; reference line, 15% prevalence). Regressions use minimum chi-square estimation with dependent variable the logarithm of prevalence odds (that is, $\text{Log}(p/1 - p)$, where p is prevalence of the breastfeeding indicator), and independent variable the oldest month of age by two-month age group; exclusive breastfeeding was estimated for five age groups 0-1 to 8-9 months of age, any breastfeeding for 18 age groups 0-1 to 34-35 months of age (shown only for children ≤2 years of age); n is the number of pooled observations; inverse-variance weights used in estimation are shown by the relative size of data markers; regressions for category-A countries are stratified by geographical and epidemiologic criteria as noted in the figure legends. Data sources are listed in the Annex.

For almost all countries, the logit transformation reduced a markedly non-linear age trend to one well described by the estimated slope and constant (Figure 10). Some countries are systematically over- or under-predicted by the subregional trend, but still show highly linear patterns (Figure 11). Although for a small group of countries systematic non-linearity remained (Figure 12), at the subregional level the linear component, as measured by the F statistic (results not shown), was always strongly significant.

Some countries showed highly variable observations month by month, which is suggestive of the limitations of 24-hour recall or the presence of systematic measurement error (Figure 13). These anomalies were particularly evident in observations of exclusive breastfeeding at older ages (Figures 14-17).

Analysis of Equation 2, the logistic expression for prevalence, shows that the regression betas (a, b) fully define the attrition hazard for the corresponding

Table 2. Population coverage by subregion, age group and data category

Region	Infants aged 0–11 months				Children aged 12–23 months			
Subregion	Population with data (%)[a]	Category			Population with data (%)[a]	Category		
		A	B	C		A	B	C
Africa	**99.1**	**79.5**	**19.6**	**0.9**	**97.6**	**79.2**	**18.4**	**2.4**
Eastern	99.4	91.8	7.6	0.6	99.4	91.8	7.6	0.6
Middle	96.5	25.9	70.6	3.5	96.5	25.9	70.6	3.5
Northern	99.8	52.3	47.5	0.2	95.1	57.2	37.9	4.9
Southern	95.1	91.5	3.6	4.9	95.1	91.5	3.6	4.9
Western	100	100	0	0	97.9	96.7	1.2	2.1
Asia (excluding Japan)	**93.5**	**63.5**	**30.0**	**6.5**	**62.0**	**59.4**	**2.5**	**38.0**
Eastern	94.1	0	94.1	5.9	0	0	0	100
South-Central	97.3	97.3	0	2.7	90.9	90.9	0	9.1
South-Eastern	84.6	74.5	10.1	15.4	84.7	74.5	10.1	15.4
Western	83.6	65.1	18.5	16.4	65.2	50.3	14.9	34.8
Latin America and the Caribbean	**92.6**	**91.2**	**1.4**	**7.4**	**86.2**	**86.2**	**0**	**13.8**
Caribbean	84.9	77.3	7.6	15.1	84.9	84.9	0	15.1
Central America	98.2	95.6	2.6	1.8	98.2	98.2	0	1.8
South America	90.8	90.6	0.2	9.2	80.6	80.6	0	19.4
Developing countries[b]	**94.8**	**70.1**	**24.7**	**5.2**	**72.9**	**66.8**	**6.1**	**27.1**

Percentage of infant and child population in each region for which there were complete and comprehensive (category A), incomplete or non-comprehensive (category B) or no (category C) data on breastfeeding. [a] Includes categories A and B. [b] The region Oceania (including developing countries Guam, Fiji, French Polynesia, New Caledonia, Papua New Guinea, Samoa, Solomon Islands and Vanuatu) was excluded, as nationally representative breastfeeding data were not available for any developing countries in Oceania.

Table 3. Summary regression results for age trends for exclusive breastfeeding

Stratum name	Stratum characteristics	Intercept (95% confidence interval)	Slope (95% confidence interval)	Standard error of the residual
AfrD	Africa: high adult, high child mortality	-0.98 (-1.33, -0.63)	-0.21 (-0.29, -0.14)	1.66
AfrE	Africa: very high adult, high child mortality	0.68 (0.29, 1.07)	-0.38 (-0.47, -0.29)	1.79
AmrB	Americas: low child, low adult mortality	0.23 (-0.20, 0.66)	-0.39 (-0.50, -0.29)	1.14
AmrD	Americas: high adult, high child mortality	0.91 (0.47, 1.34)	-0.39 (-0.49, -0.28)	1.10
EmrB	Eastern Mediterranean: low child, low adult mortality	-0.84 (-1.75, 0.06)	-0.47 (-0.80, -0.14)	0.57
EmrD	Eastern Mediterranean: high adult, high child mortality	0.26 (-0.25, 0.76)	-0.29 (-0.39, -0.18)	0.90
SearB	South-east Asia: low child, low adult mortality	1.01 (0.55, 1.48)	-0.47 (-0.58, -0.35)	0.92
SearD	South-east Asia: high adult, high child mortality	1.03 (0.83, 1.23)	-0.38 (-0.42, -0.33)	0.63
WprB	Western Pacific: low child, low adult mortality	0.58 (-0.15, 1.31)	-0.45 (-0.68, -0.22)	1.49

Regressions use minimum chi-square estimation with dependent variable logarithm of prevalence odds (that is, Log(p/1 - p), where p is prevalence of exclusive breastfeeding), and independent variable oldest month of age by two-month age groups; regressions for category-A countries stratified by geographical and epidemiologic criteria as noted in the table; exclusive breastfeeding estimated for five age groups 0–1 to 8–9 months of age. Data sources listed in the Annex (Additional file 1).

Table 4. Summary regression results for age trends for any breastfeeding

Stratum name	Stratum characteristics	Intercept (95% confidence interval)	Slope (95% confidence interval)	Standard error of the residual
AfrD	Africa: high adult, high child mortality	4.28 (4.05, 4.51)	-0.20 (-0.21, -0.19)	1.30
AfrE	Africa: very high adult, high child mortality	4.71 (4.43, 5.00)	-0.20 (-0.21, -0.18)	1.12
AmrB	Americas: low child, low adult mortality	1.34 (1.16, 1.52)	-0.12 (-0.13, -0.11)	0.79
AmrD	Americas: high adult, high child mortality	3.57 (3.29, 3.86)	-0.18 (-0.20, -0.17)	0.83
EmrB	Eastern Mediterranean: low child, low adult mortality	2.85 (2.58, 3.12)	-0.22 (-0.23, -0.20)	0.36
EmrD	Eastern Mediterranean: high adult, high child mortality	3.40 (3.16, 3.64)	-0.17 (-0.18, -0.16)	0.80
SearB	South-east Asia: low child, low adult mortality	3.58 (3.36, 3.79)	-0.14 (-0.15, -0.13)	0.49
SearD	South-east Asia: high adult, high child mortality	4.03 (3.72, 4.34)	-0.14 (-0.15, -0.12)	0.89
WprB	Western Pacific: low child, low adult mortality	2.18 (1.77, 2.60)	-0.14 (-0.16, -0.12)	2.77

Regressions use minimum chi-square estimation with dependent variable logarithm of prevalence odds (that is, Log(p/1 - p), where p is prevalence of any breastfeeding), and independent variable oldest month of age by two-month age groups; regressions for category-A countries stratified by geographical and epidemiologic criteria as noted in the table; any breastfeeding estimated for 18 age groups 0–1 to 34–35 months of age. Data sources listed in the Annex (Additional file 1).

breastfeeding rate as a function of age (results not shown). The average hazards for exclusive (upper panel) and any (lower panel) breastfeeding are displayed in Figure 18 versus month of age.

Discussion

24-hour recall

An assessment based on 24-hour recall will not always represent true exposure status, since infants may change feeding practices when ill or during the absence of the mother. In particular, 24-hour recall will generate a large number of false positives for exclusive breastfeeding [17]. This conclusion is supported by results from community-based studies [18,19]. Other studies have found that infants classified as exclusively breastfeeding with 24-hour recall receive other liquids and even solid foods on a less-than-daily basis [17,20].

In a prospective analysis published in 2000, the difference between data from 24-hour recall and those from mothers' daily records was compared, and it was found that 24-hour recall overestimated exclusive breastfeeding rates by an absolute magnitude of about 40% percentage points at both two and four months of age (92% versus 51% at two months, and 73% versus 30% at four months of age) [21]. The most common reason for misclassification was the

consumption of water or water-based drinks. Infants receiving water or water-based drinks in addition to breast milk are referred to as 'predominantly breastfed'.

The study shows infants whose true status is 'predominant' are frequently classified as 'exclusive' with 24-hour recall. Consistent with international recommendations, only infants not receiving water or water-based drinks should be classified as exclusive, and studies have shown that the introduction of any fluids other than breast milk is associated with increased risk of morbidity and mortality [22-24]. Since data from the same study show that about 30% of infants classified as 'exclusive' or 'predominant' receive some solid-food supplements, it appears there is a significant tendency to under-report small-quantity items administered in addition to breast milk.

An additional cause of overestimation is the use of filter questions. In many surveys, including DHS, the mother is asked 'For how long have you given your child only the breast?'. Although instructions following the question specify '(not even water)', since the instructions are not part of the question itself, information on the consumption of water may not be systematically elicited. Moreover, since the 24-hour-recall module is implemented only for infants who are not reported to be given 'breast only', there is no mechanism for the identification of false positives. A more valid approach is used in UNICEF's MICS surveys, in which 24-hour recall is performed for all infants, regardless of whether or not they are reported to be given 'breast only'. Our own experience is that unless mothers are specifically asked about water and herbal teas, they will report that they gave only breast milk because water and teas are not perceived as foods (CGV, personal communication).

Systematic measurement error

We highlight the presence of anomalous observations of exclusive breastfeeding among older children in category-A data (see Figures 14-17), and interpret these results as providing evidence of measurement error resulting in systematic over-reporting in DHS surveys for exclusive and possibly other breastfeeding categories. In our view, while these findings have clear implications for the validity of observations of exclusive breastfeeding in DHS data at all age groups, the interpretation must take account of differences between regions

and countries (see Figures 14-17), perhaps due to differing perceptions about infant-and-child feeding practices. However, the anomalies detected become obvious only at older ages, and there are valid reasons why a small percentage of older children may be exclusively breastfed on a given day (for example, illness of child or no food in the house). However, exclusive breastfeeding rates in some countries (particularly in the regions AfrD, EmrD and WprB) display patterns that probably could not be completely explained in this manner. While the true explanation may remain unknown in the absence of further validation studies, it seems plausible that misunderstanding on the part of survey respondents or interviewers, or validity problems with the survey instrument beyond the limitations of 24-hour recall, are important contributing factors.

Implications for exposure assessment

Validity problems in DHS surveys have implications for the measurement of the whole distribution of exposure to suboptimal breastfeeding, especially for infants ≤6 months of age. If the exposure baseline (that is, the group at no increased risk) is not validly measured, it is likely that classifications of low or intermediate risk (for example, the category predominant breastfeeding) are similarly underestimates of true risk. Conversely, it is unlikely that infants classified as at low, intermediate or high-risk are in fact at no, low or intermediate risk.

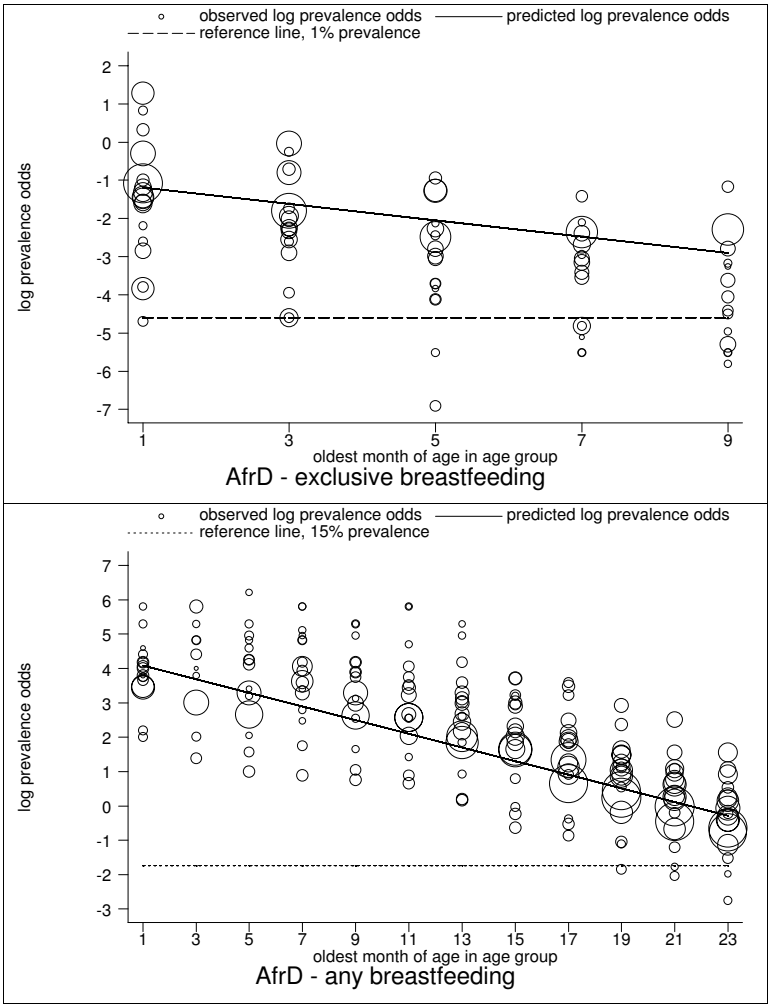


Figure 1
Estimated age trends for exclusive breastfeeding and for any breastfeeding in AfrD (Africa: high-adult, high-child-mortality stratum). Exclusive breastfeeding (upper panel, $n = 81$), any breastfeeding (lower panel, $n = 301$).

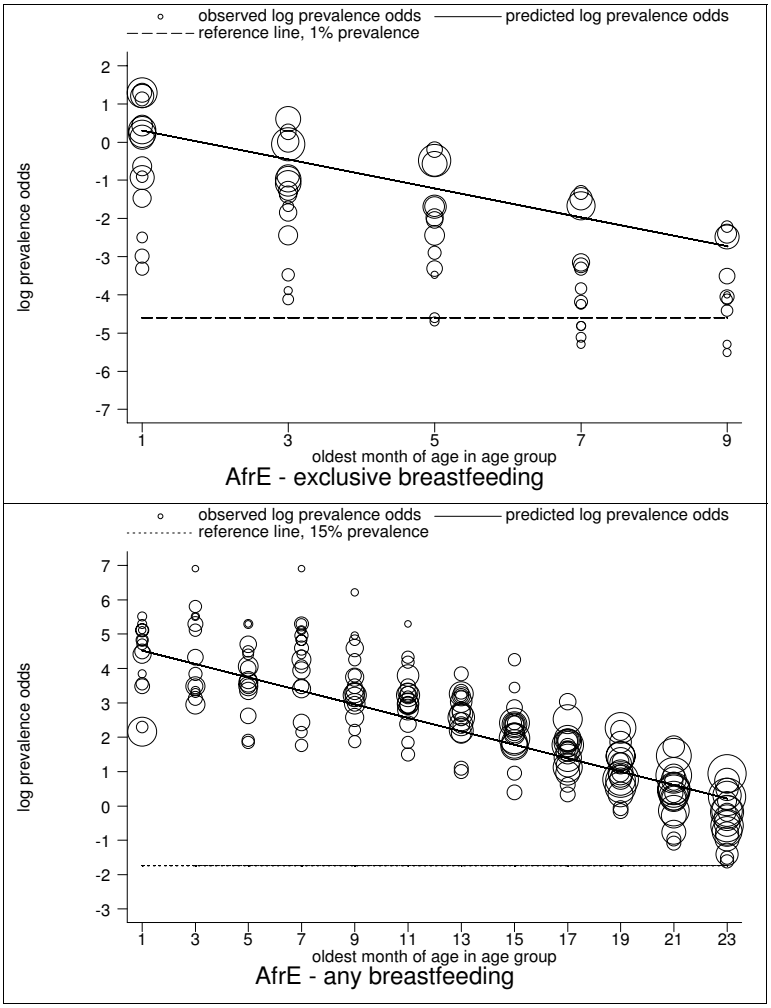


Figure 2
Estimated age trends for exclusive breastfeeding and for any breastfeeding in AfrE (Africa: very-high-adult, high-child-mortality stratum). Exclusive breastfeeding (upper panel, *n* = 66), any breastfeeding (lower panel, *n* = 304).

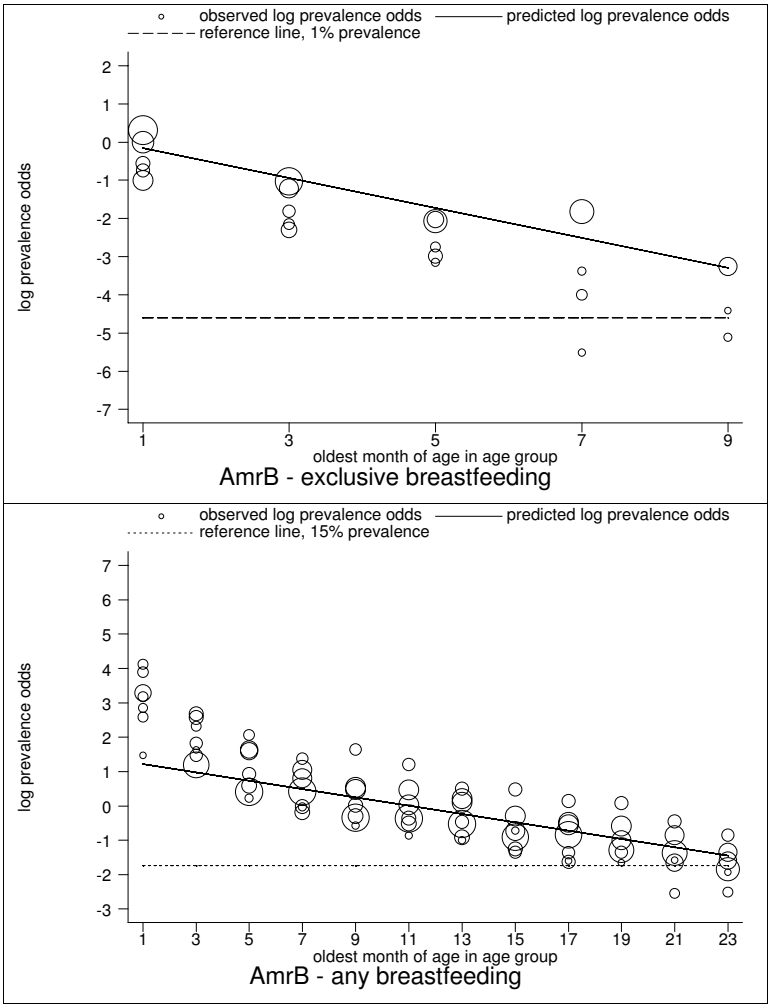


Figure 3
Estimated age trends for exclusive breastfeeding and for any breastfeeding in AmrB (Americas: low-adult, low-child mortality-stratum). Exclusive breastfeeding (upper panel, *n* = 22), any breastfeeding (lower panel, *n* = 117).

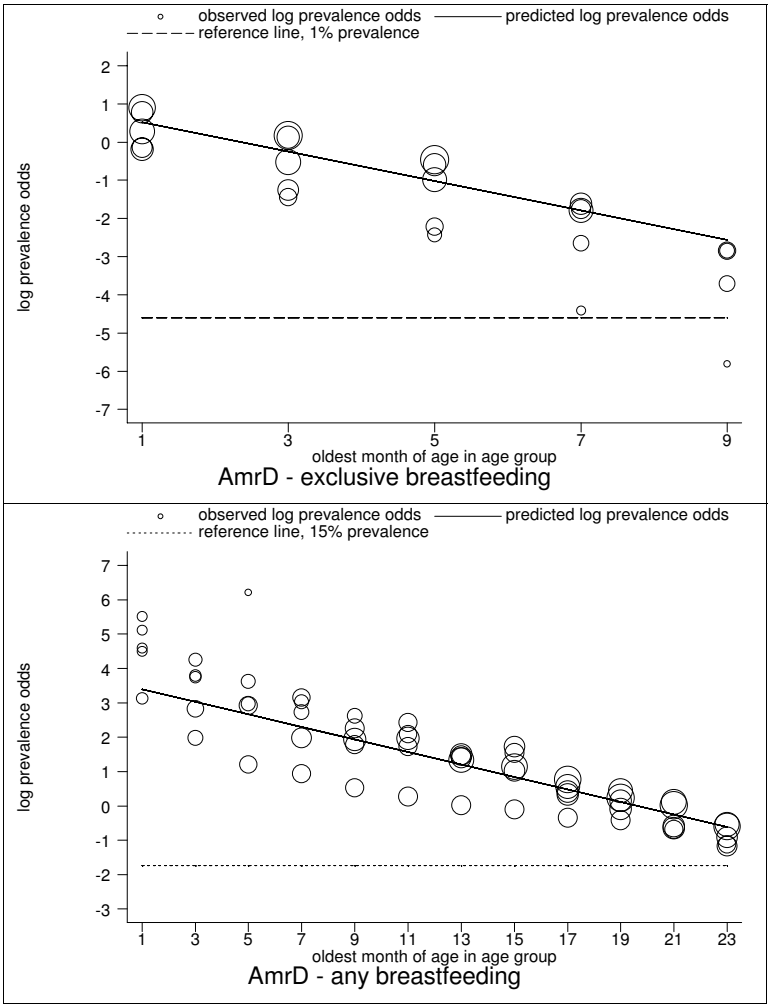


Figure 4
Estimated age trends for exclusive breastfeeding and for any breastfeeding in AmrD (Americas: high-adult, high-child mortality stratum). Exclusive breastfeeding (upper panel, *n* = 24), any breastfeeding (lower panel, *n* = 89).

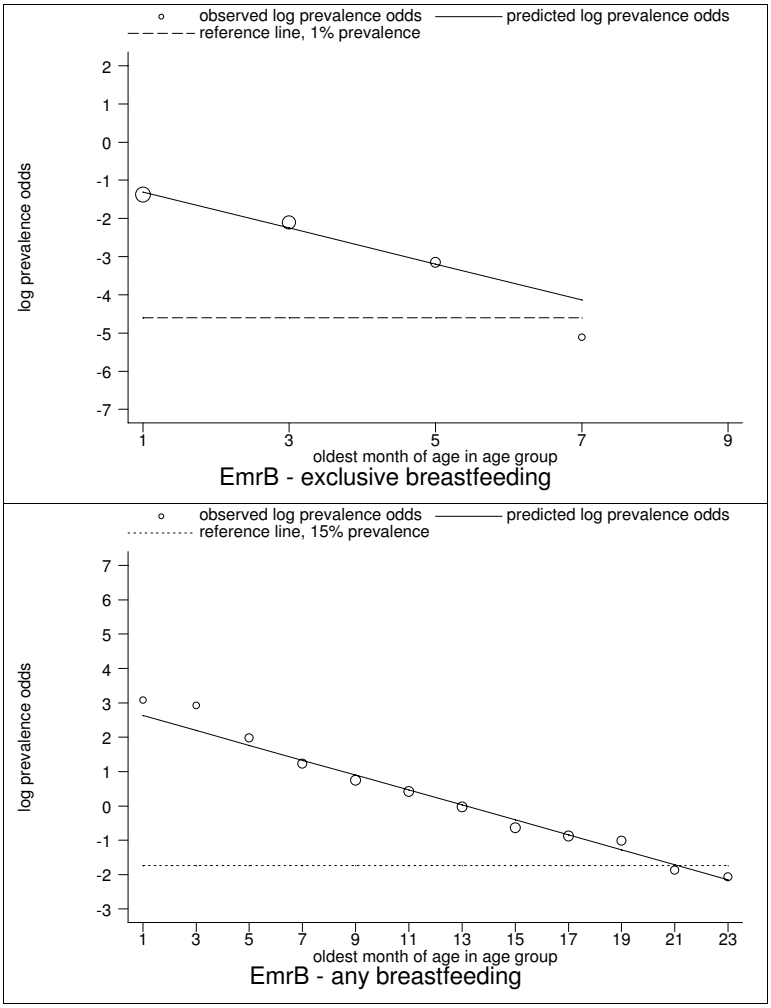


Figure 5
Estimated age trends for exclusive breastfeeding and for any breastfeeding in EmrB (Eastern Mediterranean: low-adult, low-child-mortality stratum). Exclusive breastfeeding (upper panel, *n* = 4), any breastfeeding (lower panel, *n* = 18).

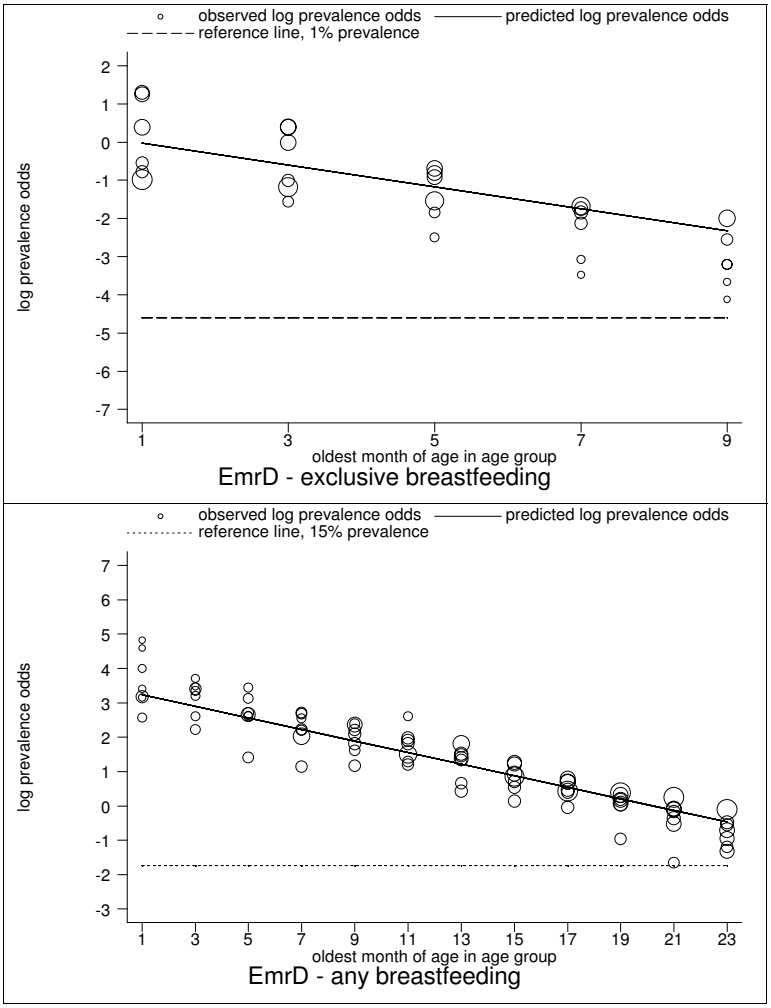


Figure 6
Estimated age trends for exclusive breastfeeding and for any breastfeeding in EmrD (Eastern Mediterranean: high-adult, high-child-mortality stratum). Exclusive breastfeeding (upper panel, *n* = 30), any breastfeeding (lower panel, *n* = 125).

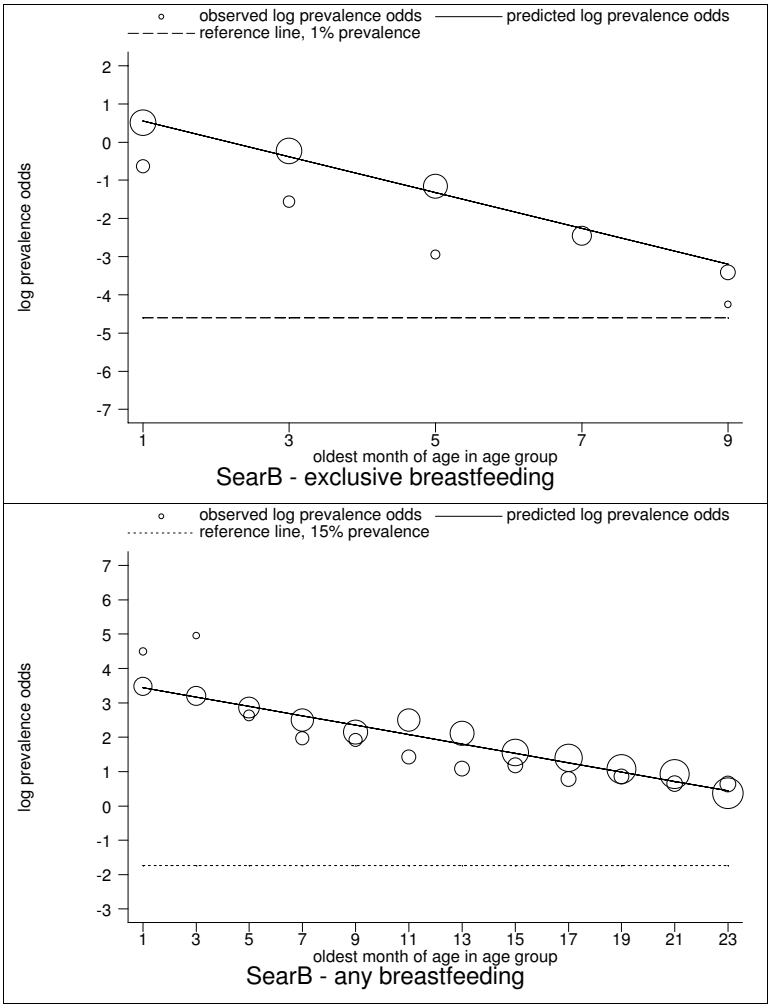


Figure 7
Estimated age trends for exclusive breastfeeding and for any breastfeeding in SearB (South-east Asia: low-adult, low-child-mortality stratum). Exclusive breastfeeding (upper panel, *n* = 9), any breastfeeding (lower panel, *n* = 36).

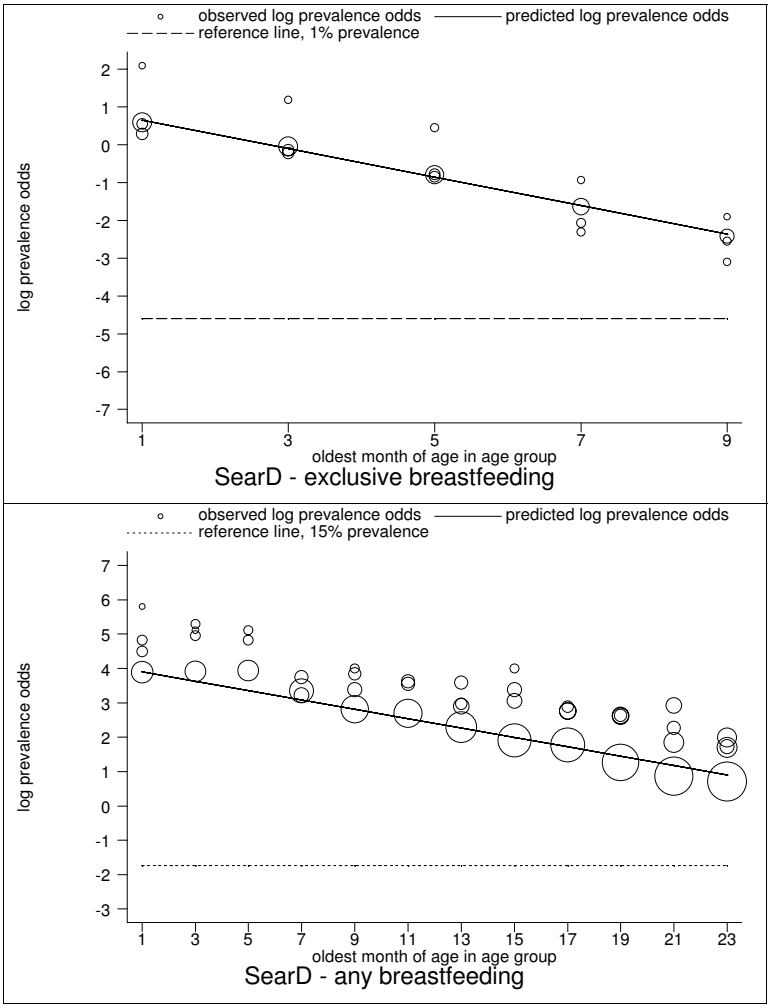


Figure 8
Estimated age trends for exclusive breastfeeding and for any breastfeeding in SearD (South-east Asia: high-adult, high-child-mortality stratum). Exclusive breastfeeding (upper panel, *n* = 20), any breastfeeding (lower panel, *n* = 69).

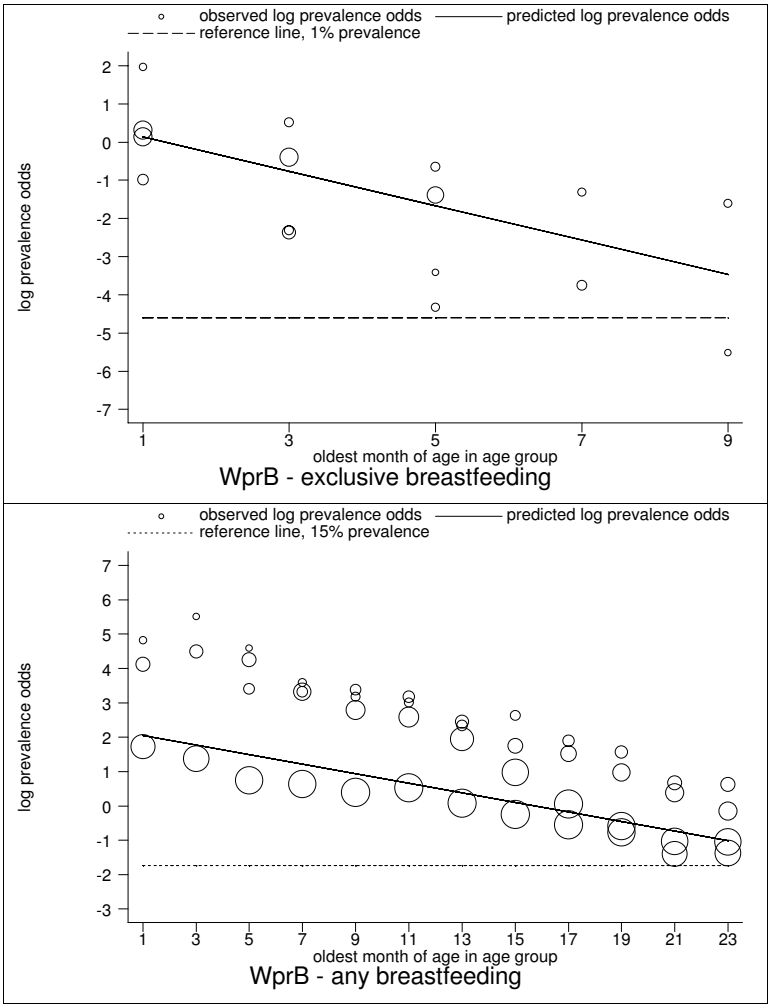


Figure 9
Estimated age trends for exclusive breastfeeding and for any breastfeeding in WprB (Western Pacific: low-adult, low-child-mortality stratum). Exclusive breastfeeding (upper panel, *n* = 16), any breastfeeding (lower panel, *n* = 70).

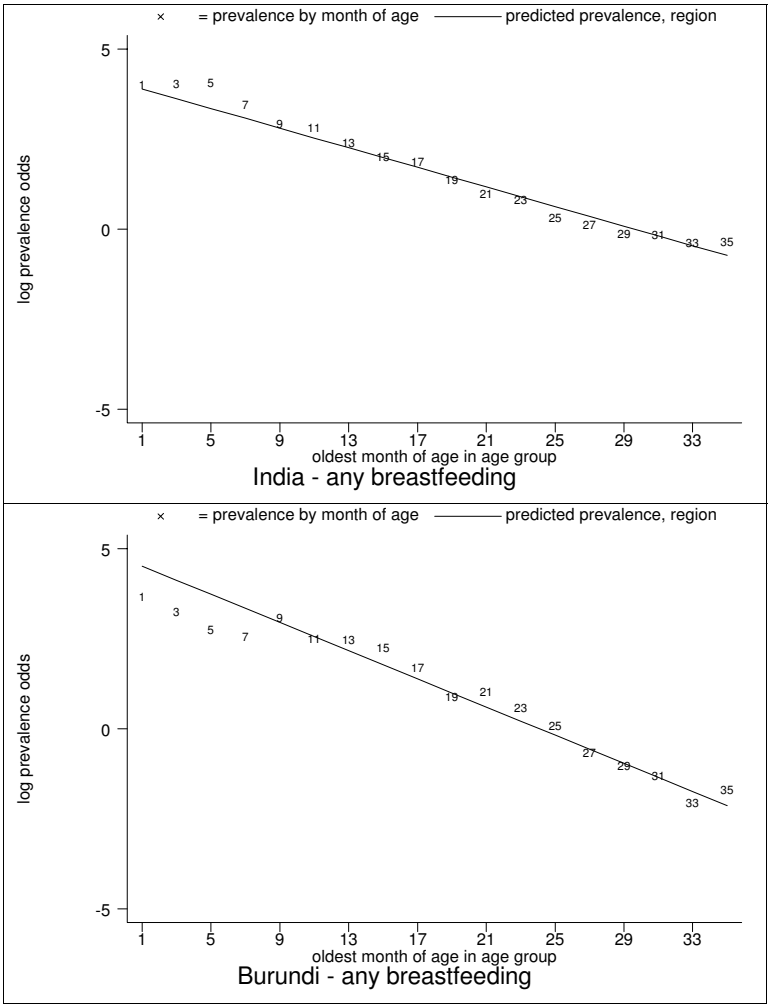


Figure 10
Examples of good predictive performance for India (upper panel) and Burundi (lower panel). Data markers show oldest month of age for the observation. The patterns are typical for many developing countries.

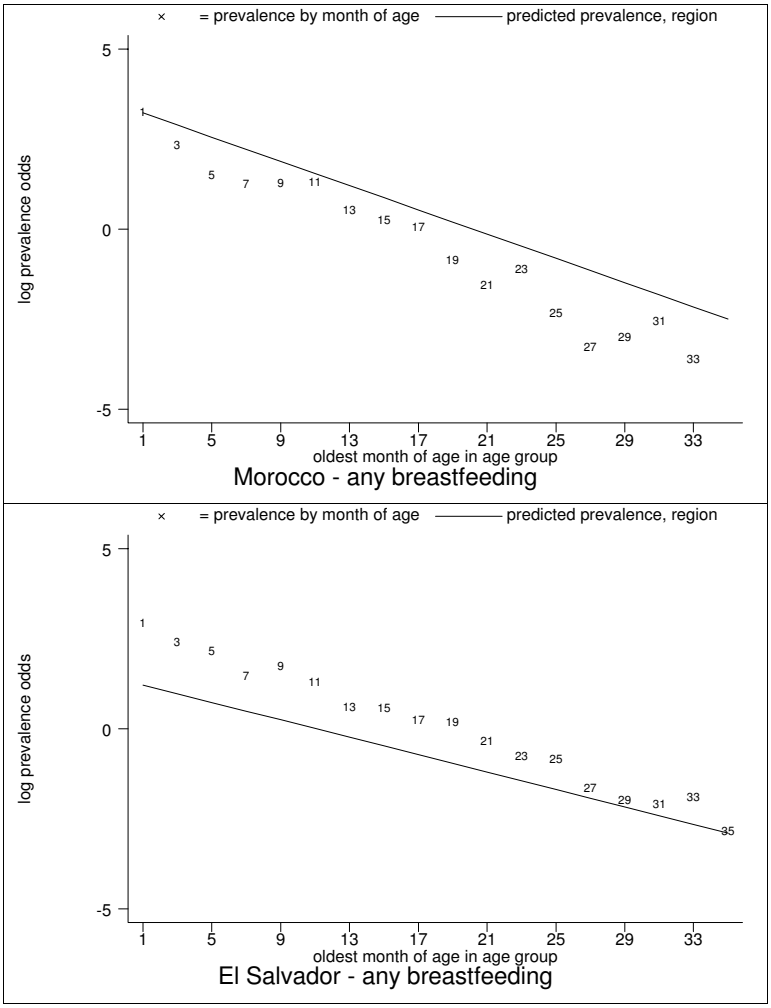


Figure 11
Examples of systematic over- (Morocco, upper panel) and under- (El Salvador, lower panel) prediction. Data markers show oldest month of age for the observation. The patterns are typical for many developing countries.

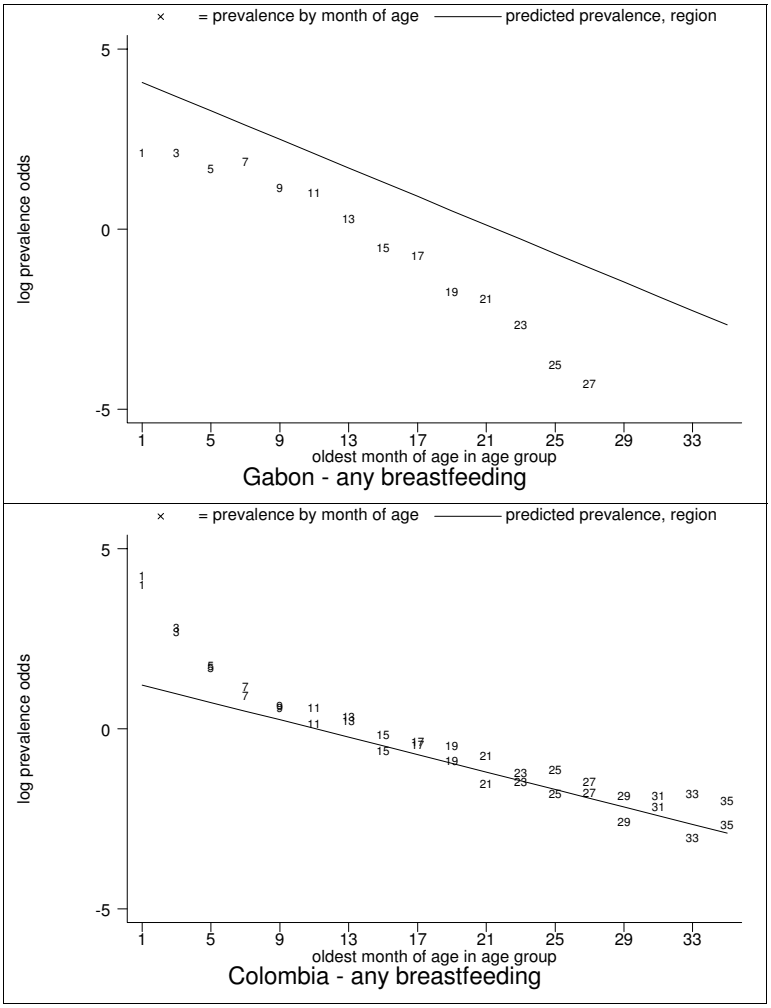


Figure 12
Examples of systematic non-linearity: concave (Gabon, upper panel); convex (Colombia, younger age groups, lower panel). Data markers show oldest month of age for the observation. Only a few developing countries display these patterns.

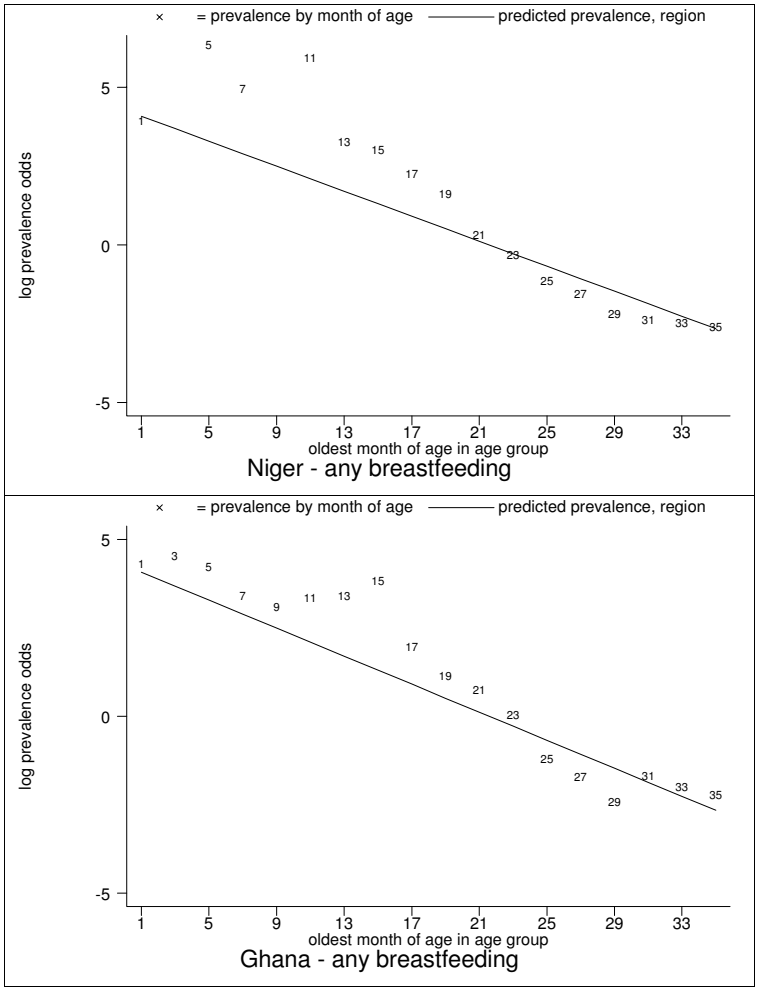


Figure 13
Examples of probable measurement error: especially months of age five and 11 (Niger, upper panel) and 11, 13 and 15 (Ghana, lower panel). Data markers show oldest month of age for the observation. High variability could be due to limitations of 24-hour recall, but consistently high age-specific values suggest systematic measurement error.

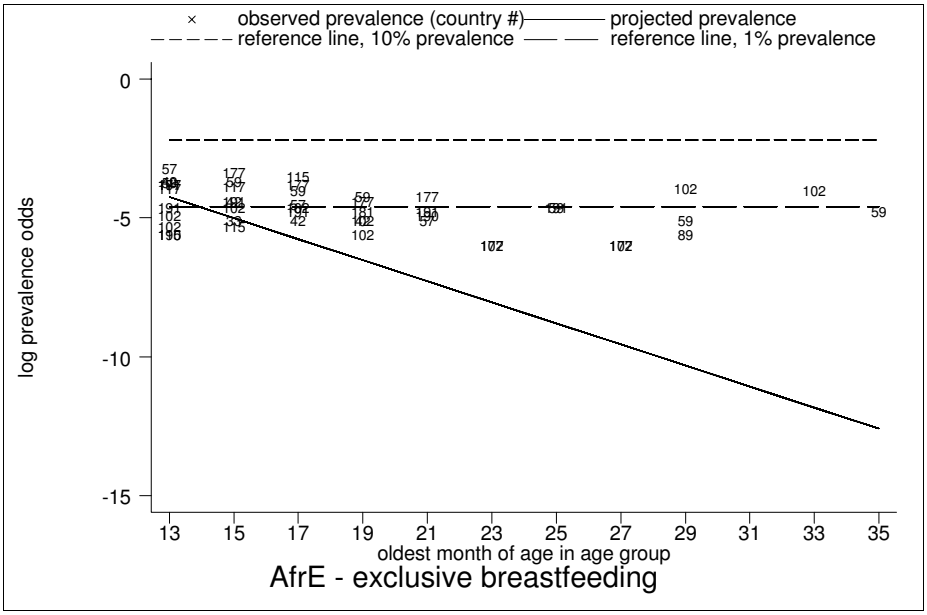


Figure 14
Projected and observed exclusive breastfeeding for age groups 12–13 to 34–35 months of age, with reference lines of 10% (short dashes) and 1% (long dashes) prevalence in AfrE. Low but non-zero observed rates of exclusive breastfeeding throughout second and third year of childhood in AfrE, with persistent rates of about 1%, show limitations of 24-hour recall. This pattern is also typical for AmrB, AmrD, SearB and SearD (EmrB had no observations of exclusive breastfeeding during the second and third year of childhood). Data markers show country identifiers.

Conclusions

The size of the gap between breastfeeding practice and recommendations in developing countries is striking. More attention should be given to increasing breastfeeding, especially exclusive breastfeeding, and to monitoring trends.

The introduction of more standardized and better validated survey instruments would be a valuable addition to child health monitoring. However, while existing data on breastfeeding are not perfect, in view of their extensive coverage, completeness and comprehensiveness, global exposure assessment is relatively robust. Although most data used here are category A, inclusion of category-B data allows for a more complete assessment in developing countries, especially among infants and in Africa.

Category-A data present anomalies suggesting the limitations of current survey methods and the presence of systematic measurement error. Nevertheless, the regularity and consistency of observed patterns of breastfeeding (see Figures 1-11) support the view that existing data capture the effects of real biological and social processes. More studies of the sort done by Aarts and colleagues [21] are probably necessary for a full understanding to emerge of measurement error and validity problems in breastfeeding surveys.

While the random-effects model yields a valid expression of statistical error, measurement and model error - essentially unquantifiable here - are larger by probably an order of magnitude or more. On balance, we believe our estimates must be interpreted as conservative (that is, lower-bound) estimates of exposure to suboptimal breastfeeding, especially non-exclusive breastfeeding, among children in developing countries. To our knowledge, these are the first published global estimates reporting exclusive breastfeeding rates for the infant population ≤6 months of age [25-27].

Our method for the analysis of breastfeeding rates provides a potent tool for summarizing trends, validating observations, translating and extrapolating indicators (as well as projecting and imputing estimates when necessary) and should support more effective child health monitoring.

Competing interests

None declared.

Abbreviations

AfrD (Africa: high-adult, high-child-mortality stratum)
AfrE (Africa: very-high-adult, high-child-mortality stratum)
AmrB (Americas: low-adult, low-child mortality-stratum)
AmrD (Americas: high-adult, high-child mortality stratum)
DHS (Demographic and Health Surveys)
EmrB (Eastern Mediterranean: low-adult, low-child-mortality stratum)
EmrD (Eastern Mediterranean: high-adult, high-child-mortality stratum)
HIV (human immunodeficiency virus)
MICS (multiple indicators cluster surveys)
SearB (South-east Asia: low-adult, low-child-mortality stratum)
SearD (South-east Asia: high-adult, high-child-mortality stratum)
WHO (World Health Organization)
WprB (Western Pacific: low-adult, low-child-mortality stratum)
UN (United Nations)
UNICEF (United Nations Children's Fund)

Authors' contributions

JAL devised the analysis methods and drafted the manuscript. APB conceived of the study, collected and collated survey data, participated in interpretation of results and drafted key sections of the manuscript. CGV drafted key sections of the manuscript, participated in interpretation of results and revised the manuscript for essential intellectual content. MdO assisted with collection and interpretation of survey data and revised the manuscript for essential intellectual content. AJDB participated in development and interpretation of statistical methods. All authors revised and approved the final manuscript.

Acknowledgements

The authors wish to thank reviewers Laura Caulfield, Edward A. Frongillo, Jr., and Alan S. Ryan for helpful comments. The views expressed are solely those of the authors and do not necessarily reflect the decisions or stated policy of the Federal University of Pelotas or the World Health Organization.

References

1. Kramer MS, Kakuma R: *The optimal duration of exclusive breastfeeding: a systematic review* (WHO/NHD/01.08). Geneva, Switzerland: World Health Organization; 2002.

2. Fifty-fourth World Health Assembly (Geneva, 14-17 May 2001): *Resolution WHA54.2: Infant and young child nutrition* (WHA54.2). Geneva, Switzerland: World Health Organization; 2001. [http://www.who.int/gb/EB_WHA/PDF/WHA54/ea54r2.pdf]

3. WHO: *The optimal duration of exclusive breastfeeding: report of an expert consultation, Geneva, Switzerland, 28-30 March 2001* (WHO/NHD/01.09). Geneva, Switzerland: World Health Organization; 2002.

4. Murray CJ, Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S: **Comparative quantification of health risks: conceptual framework and methodological issues.** *Popul Health Metr* 2003, **1:**1.

5. WHO: *Indicators for assessing breast-feeding practices* (WHO/CDD/SER/91.14). Geneva, Switzerland: World Health Organization; 1991.

6. WHO, UNICEF: *Global strategy for infant and young child feeding.* Geneva, Switzerland: World Health Organization; 2003.

7. Newell ML: **Prevention of mother-to-child transmission of HIV: challenges for the current decade.** *Bul World Health Organ* 2001, **79:**1138-1144.

8. Demographic and health surveys: Providing information for informed decisions in population, health and nutrition. [www.measuredhs.org]

9. UNICEF: Monitoring the situation of children and women. [www.childinfo.org]

10. United Nations: *World Population Prospects: the 2000 Revision*. New York: United Nations Department of Economic and Social Affairs, Population Division; 2001.

11. WHO: *The world health report 2002: reducing risks, promoting healthy life*. Geneva, Switzerland: World Health Organization; 2002.

12. Salomon JA, Murray CJL: **The epidemiologic transition revisited: compositional models for causes of death by age and sex.** *Populand Dev Rev* 2004, **28**:205-228.

13. Murray CJ, Ferguson BD, Lopez AD, Guillot M, Salomon JA, Ahmad OB: **Modified logit life table system: principles, empirical validation, and application.** *Population Studies* 2003, **57:**165-182.

14. Aldrich JH, Nelson FD: **Linear probability, logit, and probit-models.** In *Quantitative Applications in the Social Sciences, No. 07-045*. Edited by Lewis-Beck MS. Beverly Hills, California, USA: Sage Publications; 1984.

15. Greenland S: **Introduction to regression models.** In *Modern Epidemiology*. 2nd Edition. Edited by Rothman KJ, Greenland S. Philadephia, Pennsylvania, USA: Lippincott Williams and Wilkins; 1998:359-399.

16. Armitage P, Berry G, Matthews JNS: **Modelling categorical data.** In *Statistical methods in medical research*. 4th Edition. Oxford, England: Blackwell; 2002:485-502.

17. Piwoz EG, Creed de Kanashiro H, Lopez de Romaña G, Black RE, Brown KH: **Potential for misclassification of infants' usual feeding practices using 24-hour dietary assessment methods.** *J Nutr* 1995, **125:**57-65.

18. Morrow AL, Guerrero ML, Shults J, Calva JJ, Lutter C, Bravo J, Ruiz-Palacios G, Morrow RC, Butterfoss FD: **Efficacy of home-based peer counselling to promote exclusive breastfeeding: a randomised controlled trial.** *Lancet* 1999, **353:**1226-1231.

19. Haider R, Ashworth A, Kabir I, Huttly SR: **Effect of community-based peer counsellors on exclusive breastfeeding practices in Dhaka, Bangladesh: a randomised controlled trial.** *Lancet* 2000, **356:**1643-1647.

20. Boerma JT, Sommerfelt AE: *An assessment of the quality of breast-feeding data in DHS-I surveys*. Calverton, Maryland, USA: Macro International; 1993

21. Aarts C, Kylberg E, Hornell A, Hofvander Y, Gebre-Medhin M, Greiner T: **How exclusive is exclusive breastfeeding? A comparison of data since birth with current status data.** *Int J Epidemiol* 2000, **29:**1041-1046.

22. Victora CG, Smith PG, Vaughan JP, Nobre LC, Lombardi C, Teixeira AM, Fuchs SM, Moreira LB, Gigante LP, Barros FC: **Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil.** *Lancet* 1987, **2:**319-322.

23. Brown KH, Black RE, Lopez de Romaña G, Creed de Kanashiro H: **Infant feeding practices and their relationship with diarrhoeal and other diseases in Huascar (Lima), Peru.** *Pediatrics* 1989, **83:**31-40.

24. Arifeen SE, Black RE, Antelman G, Baqui AH, Caulfield LE, Becker S: **Exclusive breastfeeding reduces acute respiratory infection and diarrhea deaths among infants in Dhaka slums.** *Pediatrics* 2001, **108:**E67.

25. WHO: *Global data bank on breast-feeding. Breast-feeding: the best start in life* (WHO/NUT/96.1). Geneva, Switzerland: World Health Organization; 1996.

26. WHO: *Complementary feeding of young children in developing countries: a review of current scientific knowledge* (WHO/NUT/98.1). Geneva, Switzerland: World Health Organization; 1998:137.

27. Betrán AP, de Onís M, Lauer JA, Villar J: **Ecological study of effect of breast feeding on infant mortality in Latin America.** *BMJ* 2001, **323:**303-306.

Annex: Data sources used in the calculation of breastfeeding indicators

Country	Year	Data Category	Source
Algeria	2000	B	UNICEF. Enquête nationale sur les objectifs de la fin décennie: Santé mères et enfants Algérie 2000. République Algérienne Démocratique et Populaire, Ministère de la Santé et de la Population, and Institut National de Santé Publique. 2001. Internet communication of 11 January 2002 from http://www.childinfo.org/MICS2/natlMICrepz/MICSnatrep.htm.
Angola	1996	B	UNICEF. Angola Multiple Indicator Cluster Survey, 1996. Internet communication of 19 December 2001 from http://www.childinfo.org/mics/micsrept.html
Argentina	2000	B	Encuestas LACMAT 3.3. Procesado por la Unidad Coordinadora de Programas Materno-Infantiles y Nutricionales. Programas Materno-Infantiles Provinciales. Ministerio de Salud de la Nación. Enero 2001. (Surveys carried out during Breastfeeding Week).
Armenia	2000	A	Demographic and Health Survey (DHS). Armenia Demographic and Health Survey 2000. Measure DHS+ and ORC Macro, Calverton, Maryland USA, 2001.
Azerbaijan	2000	A	UNICEF. Azerbaijan Multiple Indicator Cluster Survey, 2000. (Internet communication of 11 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm)
Bangladesh	2000	A	National Institute of Population Research and Training (NIPORT), Mitra and Associates and ORC Macro. Bangladesh Demographic and Health Survey 1999-2000. National Institute of Population Research and Training (NIPORT), Mitra and Associates and ORC Macro, Dhaka, Bangladesh and Calverton, 2001.
Belize	1991	A	Ministry of Finance/Belize Family Life Association/Ministry of Health/CDC. 1991 Belize Family Health Survey. Final Report. US Department of Health and Human Services, Atlanta, Georgia 30333, USA. 1992.
Benin	1996	A	Demographic and Health Surveys (DHS). Bénin. Enquête Démographique et de Santé 1996. Demographic and Health Surveys. Macro International Inc. Calverton, Maryland, USA. 1997.
Bolivia	1998	A	Demographic Health and Surveys (DHS). Bolivia Encuesta Nacional de Demografia y Salud 1998. Macro International Inc. Calverton, Maryland, USA, 1998.
Botswana	2000	B	UNICEF and Central Statistics Office Republic of Botswana. Botswana Multiple Indicator Survey 2000. UNICEF. 2001. (Internet communication of 8 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm).
Brazil	1996	A	Demographic Health and Surveys (DHS), Soceidade Civil Bem-Estar Familiar no Brasil B, Fundação Instituo Brasileiro de Geografiia e Estatistica I, Ministério da Saúde, USAID, UNICEF. Pesquisa Nacional Sobre Demografia e Saúde 1996 Brazil. Macro International Inc. Calverton, Maryland, USA, 1997.
Burkina Faso	1999	A	Demographic and Health Survey (DHS). Enquête Démographique et de Santé Burkina Faso 1998-1999. Macro International Inc. Calverton, Maryland, USA, 2000.
Burundi	2000	B	UNICEF, Ministere de la planification du developpement et de la reconstruction, Institut de statistics et d'etudes econimques du Burundi. Enquête Nationale d'Evaluation des Conditions de vie de l'Enfant et de la Femme au Burundi. ENECEF-BURUNDI 2000: Rapport Préliminaire. (Internet communication of 8 January 2002 from http://www.childinfo.org/ MICS2/natlMICSrepz/MICSnatrep.htm).

Country	Year	Data Category	Source
Cambodia	2000	A	UNICEF. Cambodia Demographic and Health Survey 2000. UNICEF, 2001. Internet communication of 8 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSrep.htm.
Cameroon	1998	A	Demographic and Health Surveys (DHS). Enquête Démographique et de Santé 1998 - Cameroun. Macro International Inc. Calverton, Maryland, USA, 1999.
Cape Verde	1998	A	Instituto Nacional de Estatística. Inquérito Demográfico e de Saúde Reprodutiva Cape Verde 1998. Instituto Nacional de Estatística, Avenida Amílcar Cabral, CP 116 Praia Cabo Verde, 1999. (e-mail: inecv@cvtelecom.cv)
Central African Republic	1995	A	Ndamobissi, Robert, Gora Mboup et Edwige Opportune Nguélébé. 1995 Enquête Démographique et de Santé, République Centrafricaine 1994-95. Calverton, Maryland, USA; Direction des Statistiques Démographiques et Sociales et Macro International Inc.
Chad	1996-97	A	Ouagadjio, Bandoumal, Kostelngar Nodjimadji, Joël Nodjimbatem Ngoniri, Ningam Ngakoutou, Keumaye Ignégongba, Joël S. Tokindang, Oumdague Kouo, Bernard Barrère, et Monique Barrèere. 1998 Enquête Démographique et de Santé, Tchad 1996-97. Calverton, Maryland, USA; Bureau Central du Recensement et Macro International Inc.
Chile	1993	A	Castillo C, Atalah E, Riumallo J, Castro R. Breast-feeding and the nutritional status of nursing children in Chile. Bulletin of PAHO, 1996;30:125-132.
China	1998	B	Zhenying F, Suying C, Wu H, Gang F. Exclusive breastfeeding and growth of infants under 4 months in China. Journal of Hygiene Research, 2000; 29(5):275-278.
Colombia	2000	A	Ojeda G, Ordónez M, Ochoa LH. Salud Sexual y Reproductiva en Colombia. Encuesta Nacional de Demografía y Salud 2000. Measure DHS+, Macro International Inc., Calverton, Maryland, USA, 2000.
Comoros	1996	A	Demographic and Health Surveys (DHS). Enquête Démographique et de Santé aux Comores 1996. Macro International Inc. Calverton, Maryland USA, 1997.
Costa Rica	1993	B	Caja Costarricense del Seguro Social/CDC. Fecundidad y Formación de la Familia. Encuesta Nacional de Salud Reproductive de 1993. Programa Salud Reproductiva, Apartado 1434-1011-Griega, San Jose, Costa Rica. 1994.
Cote d'Ivoire	1999	A	Demographic Health and Surveys (DHS). Enquête Démographique et de Santé en Côte d'Ivoire EDSCI-II 1998-1999. Macro International Inc. Calverton, Maryland, USA, 1999.
Cuba	2000	A	UNICEF. Cuba Multiple Indicator Cluster Survey, 2000. Internet communication of 19 December 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm.
Democratic Republic of the Congo	1995	B	Ministére du Plan et Reconstruction Nationale. Enquête national sur la situation des enfants et des femmes au Zaire en 1995. Rapport final. UNICEF, PNUD, OMS, Kinshasa, Democratic Republic of the Congo, 1996.
Dominican Republic	1996	A	Demographic Health and Surveys (DHS). Encuesta Demográfica y de Salud 1996 República Dominicana 1996. Macro International Inc. Calverton, Maryland, USA, 1997.
Ecuador	1999	A	USAID/CDC/Naciones Unidas. Fondo de Población. ENDEMAIN-III Ecuador. Informe General. Centro de Estudios de Poblacion y Desarrollo Social. Quito, Ecuador. 2000.
Egypt	2000	A	El-Zanaty F, Way AA. Egypt Demographic and Health Survey 2000. Ministry of Health and Population [Egypt], National Population Council and ORC Macro, Calverton, Maryland, USA, 2001.
El Salvador	1998	A	Asociación Demográfica Salvadoreña, CDC, USAID. República de El Salvador, CA. Encuesta Nacional de Salud Familiar. Informe final. FESAL-98. Atlanta, Georgia USA: Centers for Disease Control and Prevention (CDC) USA, 2000.

Country	Year	Data Category	Source
Eritrea	1995	A	Demographic and Health Surveys (DHS). Eritrea Demographic and Health Survey 1995. Macro International Inc.Calverton, Maryland USA, 1997.
Ethiopia	2000	A	Demographic and Health Survey (DHS). Ethiopia Demographic and Health Survey 2000. Calverton, Maryland, USA: ORC Macro, 2001.
Gabon	2000	A	Direction Générale de la Statistique et des Économiques (DGSEE), Gabon, ORC Macro. Gabon Enquête Dèmographique et de Santé 2000. Direction Générale de la Statistique et des Économiques, Direction Génerale de la Statistique et des Etudes Economiques, United Nations Fund for Population, and ORC Macro, Calverton, Maryland, USA, 2001.
Gambia	2000	A	UNICEF. The Gambia. Draft Report. Multiple Indicator Cluster Survey, 2000. Internet communication of 19 December 2001 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm
Georgia	2000	B	Servanescu F, Morris L, Nutsubidze N, Imnadze P, Shaknazarova M. Women's Reproductive Health Survey Georgia, 1999-2000. National Center for Disease Control, Georgia, Centers for Disease Control and Prevention, Atlanta, Georgia, USA, 2001.
Ghana	1998	A	Demographic and Health Survey (DHS). Ghana Demographic and Health Survey 1998. Macro International Inc. Calverton, Maryland, USA, 1999.
Guatemala	1999	A	Demographic Health and Survey (DHS). Guatemala Encuesta Nacional de Salud Materno Infantil 1998-1999. Macro International Inc. Calverton, Maryland, USA, 1999.
Guinea	1999	A	Direction Nationale de la Statistique (Guineé) et Macro International Inc. 2000. Enquête Démographique et de Santé, Guineé 1999. Calverton, Maryland USA: Direction Nationale de la Statistique et Macro International Int. (DHS)
Guinea-Bissau	2000	A	UNICEF. Multiple Indicator Cluster Survey, Guinea-Bissau. Internet communication of 19 December 2001 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm
Haiti	2000	A	Demographic Health and Surveys (DHS). Enquête Mortalité, Morbidité et Utilisation des Services EMMUS-III Haïti 2000. Calverton, Maryland USA: ORC Macro, 2001.
Honduras	1996	A	Demographic Health and Surveys (DHS), Ministerio de Salud, Asociación Hondureña de Planificacion de Familia (ASONPLAFA), USAID/Honduras, and CDC. Encuesta Nacional de Epidemiologia y Salud Familiar 1996 Honduras. Atlanta, Georgia, USA: Centers for Disease Control and Prevention USA, 1997.
India	1999	A	Demographic Health Survey (DHS). National Family Health Survey (NFHS-2) 1998-99 India. ORC Macro, Calverton, Maryland, USA, 2000.
Indonesia	1997	A	Demographic Health and Surveys (DHS). Indonesia Demographic and Health Survey 1997. Macro International Inc. Calverton, Maryland, USA, 1998.
Iran, Islamic Republic of	1995	B	Ministry of Health and Medical Education under Secretary for Public Health of the Islamic Republic of Iran and UNICEF. The Multiple Health Indicator Cluster Survey of Iran 23-27 September 1995. Ministry of Health and Medical Education under Secretary for Public Health of the Islamic Republic of Iran and UNICEF Tehran, 1996.
Iraq	1997	A	Ministry of Health (Iraq) and UNICEF. Situation Analysis of children and women in Iraq. Internet communication of 19 December 2001 from www.childinfo.org/Other/Iraq_sa.pdf
Jamaica	1997	B	National Family Planning Board. Reproductive health Survey. Jamaica 1997. CDC, Atlanta, Georgia 30333. USA. 1999.

Country	Year	Data Category	Source
Jordan	1997	A	Demographic Health and Survey (DHS). Jordan Population and Family Health Survey 1997. Macro International Inc. Calverton, Maryland, USA, 1998.
Kazakhstan	1999	A	Demographic Health and Survey (DHS). Kazakhstan Demographic and Health Survey 1999. Macro International Inc. Calverton, Maryland, USA, 2000.
Kenya	1998	A	Demographic and Health and Survey (DHS). Kenya Demographic and Health Survey 1998. Macro International Inc. Calverton, Maryland, USA, 1999.
Kyrgyzstan	1997	A	Demographic and Health Survey (DHS). Kyrgyztan Demographic and Health Survey 1997. Calverton, Maryland, USA: Macro International Inc., 1998.
Lebanon	2000	B	UNICEF. Multiple Cluster Survey on the Situation of Children In Lebanon 2001. UNICEF. 2001. Internet communication of 14 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm.
Liberia	1999-00	A	UNICEF. Liberia National Nutrition Survey 1999-2000. Ministry of Health and Social Welfare, Christian Health Association of Liberia and United Nations Children's Fund. Monrovia, Liberia, 2001.
Libyan Arab Jamahiriya	1996	B	Pan Arab Project for Child Development (PAPCHILD). Arab Libyan Maternal and Child Health Survey 1996. The Great Socialist People's Libyan Arab Jamahiria The General People's Committee for Health and Social Insurance. EL-Faresssia st Cairo, League of Arab States, 1998
Madagascar	1997	A	Demographic and Health and Surveys (DHS). Enquête Démographique et de Santé 1997, Madagascar. Macro International Inc. Calverton, Maryland, USA, 1998.
Malawi	2000	A	National Statistical Office, Measure DHS+. Malawi Demographic and Health Survey 2000. ORC Macro, Calverton, Maryland, USA, 2001.
Mali	1996	A	Demographic and Health Surveys (DHS). République du Mali Enquête Démographique et de Santé 1995-1996. Cellule de Planification et de Statistique Ministère de la Santé, Direction Nationale de la Statistique et de l'Informatique Bamako M, and Macro International Inc., Calverton, Maryland USA, 1996.
Mauritania	2000-01	A	Office National de la Statistique (ONS) [Mauritanie] et ORC Macro. 2001. Enquête Démographique et de Santé mauritanie 2000-2001. Calverton, Maryland, USA: ONS et ORC Macro.
Mexico	1999	A	González-Cossío T, Moreno-Macías H, Rivera Dommarco J, Villalpando Hernández S, Shamah Levy T, Monterrubio Flores E, Hernández Garduño A. Breast-feeding Practices in Mexico: Results from the Second National Nutrition Survey (NNS-II) 1999 (in press).
Mongolia	2000	B	UNICEF. Multiple Indicator Cluster Survey Mongolia 2000. UNICEF. 2002. Internet communication of 8 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm, UNICEF.
Morocco	1995	A	Demographic and Health Surveys (DHS). Enquête de Panel sur la Population et la Santé (EPPS) 1995, Maroc. Calverton, Maryland USA; Macro International Inc., 1996.
Mozambique	1997	A	Demographic and Health Surveys (DHS). Moçambique Inquérito Demográfico e de Saúde 1997. Calverton, Maryland, USA: Macro International Inc., 1998.
Myanmar	1995	B	Ministry of Health, Myanmar and UNICEF. Monitoring Progress Towards Goals of The World Summit for Children through Multiple Indicator Cluster Survey. Department of Planning and Statistics, Ministry of Health, Myanmar and UNICEF, 1995.

Country	Year	Data Category	Source
Namibia	1992	A	Demographic and Health Surveys (DHS). Namibia Demographic and Health Survey 1992. Calverton, Maryland USA: Macro International Inc., 1993.
Nepal	1996	A	Demographic and Health Survey (DHS). Nepal Family Health Survey 1996. Macro International Inc. Calverton, Maryland, USA, 1997.
Nicaragua	1998	A	Demographic and Health Surveys (DHS), Instituto Nacional de Estadísticas y Censos, Ministerio de Salud, Nicaragua. Encuesta Nicaragüense de Demografía y Salud 1998. Macro International Inc. Calverton, Maryland, USA, 1999.
Niger	2000	A	UNICEF. Enquête à Indicateurs Multiples de la fin de la Décennie (MICS2) Niger 2000. UNICEF, 2001. Internet communication of 8 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSrep.htm, UNICEF.
Nigeria	1999	A	Demographic and Health Surveys (DHS), United Nations Population Fund, USAID. Nigeria Demographic and Health Survey 1999. National Population Commission and ORC/Macro, Calverton, Maryland USA, 2002.
Oman	1995	B	Gulf Family Health Survey. Oman Family Health Survey 1995. Preliminary Report. Edited by Sulaiman AJM, Al-Riyami A, Farid S. Ministry of Health, Muscat and Council of Health Ministers of GCC States, Riyadh, 1996.
Pakistan	1991	A	Demographic and Health Surveys (DHS). Pakistan Demographic Health Survey 1990/1991. Columbia, Maryland USA: IRD/Macro International Inc., 1992.
Paraguay	1996	A	Centro Paraguayo de Estudios de Población (CEPEP) and Centers for Disease Control and Prevention (CDC). Encuesta Nacional de Demografía y Salud Reproductiva 1995-96 (ENDSR-95/96). Centro Paraguayo de Estudios de Población, Asunción, Paraguay, and CDC, Atlanta, Georgia USA, 1997.
Peru	1996	A	Demographic and Health Survey (DHS), USAID, UNFPA, UNICEF, Ministerio de Salud [Peru]. Encuesta Demográfica y de Salud Familiar 1996 Peru. Macro International Inc. Calverton, Maryland, USA, 1997.
Philippines	1998	A	Demographic and Health Surveys (DHS). Philippines National Demographic and Health Survey 1998. Macro International Inc. Calverton, Maryland, USA, 1999.
Rwanda	2000	A	Demographic and Health Surveys (DHS). Enquête Démographique et de Santé EDSR-II Rwanda 2000. Calverton, Maryland USA; ORC Macro, 2001.
Saudi Arabia	1991	B	Al-Mazrou Y, Al-Shehri S, Al-Jeffri M, Farag MK, Baldo MH. Saudi Maternal and Child Health Survey 1991. General Directorate of Health Centres, Ministry of health, Riyadh, Saudi Arabia, 1992.
Senegal	1997	A	Demographic and Health Surveys (DHS) and Ministère de l'Economie, des Finances et du Plan. Enquête Démographique et de Santé au Sénégal (EDS-III) 1997. Direction de la Prévision et de la Statistique, Point E, B.P. 116, and Macro International Inc., Calverton, Maryland, USA, 1997.
Sierra Leone	2000	A	Government of Sierra Leone/UNICEF/UNFPA/UNHCR/ACTION AID/CCF/CHASL. The status of women and children in Sierra Leone. A household survey report (MICS-2). Internet communication of 19 December 2001 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm
Somalia	2000	B	UNICEF. Multiple Indicator Cluster Survey 2000 Somalia. UNICEF. 2000. Internet communication of 11 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm, UNICEF.
South Africa	1998	A	Demographic and Health Surveys (DHS). South Africa Demographic and Health Survey 1998. Full report. Department of Health, Republic of South Africa and Demographic and Health Surveys. Macro International Inc, Calverton, Maryland, USA.
Sri Lanka	1993	A	Demographic and Health Surveys (DHS). Sri Lanka Demographic and Health Survey 1993. Colombo, Sri Lanka, 1995.

Country	Year	Data Category	Source
Sudan	1993	B	Pan Arab for Child Development (PAPCHILD). Sudan Maternal and Child Health Survey 1992/93. Federal Ministry of Health National Directorate of Motherhood, Childhood & Family Planning National Centre of Health Information. 1995. Republic of Sudan, League of Arab States,1993
Suriname	2000	B	UNICEF. Suriname Multiple Indicator Cluster Survey 2000. UNICEF. 2001. Internet communication of 8 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSrep.htm.
Swaziland	2000	A	UNICEF and Central Statistical Office of Swaziland. Multiple Indicator Cluster Survey 2000. UNICEF. 2002. Internet communication of 10 December 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm.
Tajikistan	2000	A	UNICEF. Multiple Indicator Cluster Survey, Tajikistan. Internet communication of 19 December 2001 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSnatrep.htm
Togo	1998	A	Demographic and Health Surveys (DHS). Enquête Démographique et de Santé Togo 1998. Macro International Inc. Calverton, Maryland, USA, 1999.
Tunisia	1995	B	Ministère de la Santé Publique. Office National de la Famille et de la Population. L'Enquête Tunisienne sur la Santé de la Mère et de l'Enfant. Rapport Principal 1996. Project Pan Arab pour la Promotion de l'Enfance. Ministère de la Santé Publique, Tunisia, 1996.
Turkey	1998	A	Demographic Health and Survey (DHS). Turkey Demographic and Health Survey 1998. Macro International Inc. Calverton, Maryland, USA, 2000.
Turkmenistan	2000	A	Demographic and Health Surveys (DHS). Turkmenistan Demographic and Health Survey 2000. Measure DHS+ and ORC Macro. Calverton, Maryland, USA, 2001.
Uganda	1995	A	Statistics Department Ministry of Health and Economic Planning, Demographic and Health Surveys. Uganda Demographic and Health Survey 1995. Macro International Inc. Calverton, Maryland, USA, 1996.
United Republic of Tanzania	1999	A	Demographic and Health Surveys (DHS). Tanzania Reproductive and Child Health Survey 1999. Macro International Inc. Calverton, Maryland, USA, 2000.
Uzbekistan	2000	A	UNICEF. Multiple Indicator Cluster Survey Republic of Uzbekistan 2000. UNICEF. 2000. (Internet communication of 8 January 2002 from http://www.childinfo.org/MICS2/natlMICSrepz/MICSrep.htm).
Viet Nam	2000	A	National Institute of Nutrition, General Statistical Office. 2000 Vietnam: Child and Mother Nutrition Situation. Medical Publishing House. Hà Nôi, Viet Nam. 2001.
Yemen	1997	A	Demographic and Health Surveys (DHS). Yemen Demographic and Maternal and Child Health Survey 1997. Macro International Inc. Calverton, Maryland, USA, 1998.
Zambia	1996	A	Demographic and Health Surveys (DHS). Zambia Demographic and Health Survey 1996. Macro International Inc. Calverton, Maryland, USA, 1997.
Zimbabwe	1999	A	Demographic and Health Survey (DHS). Zimbabwe Demographic and Health Survey 1999. Calverton, Maryland, USA: Macro International Inc., 2000.

Chapter 7

Deaths and years of life lost due to suboptimal breastfeeding among children in the developing world: a global ecological risk assessment

Based on: Lauer JA, Betrán AP, Barros AJD, de Onís M. Deaths and years of life lost due to suboptimal breastfeeding among children in the developing world: a global ecological risk assessment. *Public health nutrition*, 2006, 9(6):673-685.

Abstract

Objective

We estimate attributable fractions, deaths and years of life lost among infants and children ≤2 years of age due to suboptimal breastfeeding in developing countries.

Design

We compare actual practices to a minimum-exposure pattern consisting of exclusive breastfeeding for infants ≤6 months of age and continued breastfeeding for older infants and children ≤2 years of age. For infants, we consider deaths due to diarrhoeal disease and lower respiratory infections, and deaths due to all causes are considered in the second year of life. Outcome measures are attributable fractions, deaths, years of life lost and offsetting deaths potentially caused by mother-to-child transmission of HIV through breastfeeding.

Setting

Developing countries.

Subjects

Infants and children ≤2 years of age.

Results

Attributable fractions for deaths due to diarrhoeal disease and lower respiratory infections, respectively, are 55% and 53% for the first six months of infancy, 20% and 18% for the second six months, and are 20% for all-cause deaths in the second year of life. Globally, as many as 1.45 m lives (117 m years of life) are lost due to suboptimal breastfeeding in developing countries. Offsetting deaths caused by mother-to-child transmission of HIV through breastfeeding could be as high as 242,000 (18.8 m years of life lost) if relevant WHO recommendations are not followed.

Conclusions

The size of the gap between current practice and recommendations is striking when one considers breastfeeding involves no out-of-pocket costs, that there exists universal consensus on best practices, and that implementing current international recommendations could potentially save 1.45 m children's lives each year.

Introduction

Recent work on major global risk factors

In 2002, the World Health Organization (WHO) highlighted the public health importance of major global risk factors for burden of disease.[1] Modifiable childhood nutritional risk factors such as underweight (3.75 m deaths) and zinc (789,000 deaths), vitamin A (778,000 deaths) and iron (841,000 deaths) deficiency were estimated to be responsible for over 6 m childhood deaths per year. Environmental risk factors such as unsafe water, sanitation and hygiene, whose effects fall overwhelmingly on children, were estimated to be responsible for an additional 1.73 m deaths per year.

Although WHO did not analyse suboptimal breastfeeding, other recent analyses of causes of childhood death have qualified it as one of "the most powerful shared risk factors"[2] and estimated that 1.3 m deaths could be prevented in 42 high-mortality countries by increased levels of breastfeeding among infants.[3] That work, by the Bellagio Child Survival Group, suggests that suboptimal breastfeeding is a childhood risk factor of the same magnitude as the other major nutritional and environmental risk factors studied by the WHO Collaborating Group. [1;4;5]

On a number of technical grounds, the present work extends that published by the Bellagio Group: we use published estimates of exposure for the entire developing world (135 countries),[6] we apply estimates of risk consistent with current international infant-feeding recommendations and, consistent with best available evidence,[7] we assess the effect of suboptimal breastfeeding throughout the second year of life. Finally, we present a detailed account of methods, assumptions and data sources.

Policy background

Excluding neonatal disorders, the most important causes of childhood deaths in developing countries are lower respiratory infections and diarrhoea.[2] It has been long understood that breastfeeding reduces not only the incidence of these diseases but also their severity and duration.[8-11]

Recent debate has focused on the optimal duration of exclusive breastfeeding in infancy, and, in 1999, WHO commissioned a systematic review of evidence,[12] following which the World Health Assembly voted in 2001 to recommend exclusive breastfeeding for the first six months of life as a global public health recommendation.[13] On the basis of the best available evidence, therefore, current international consensus holds that optimal breastfeeding practice for infants and young children consists of exclusive breastfeeding for the first six months of life with continued breastfeeding up to two years of age and beyond.[14]

Nevertheless, breastfeeding for more than one year has been estimated to pose a 10–20% risk of HIV transmission in children of infected mothers.[15] Since the risk of HIV transmission clearly needs to be balanced against other risks faced by infants and children, WHO currently states that "when replacement feeding is acceptable, feasible, affordable, sustainable and safe, avoidance of all breastfeeding by HIV-positive women is recommended; otherwise, exclusive breastfeeding is recommended during the first months of life"; WHO emphasizes, however, "that those who choose other options should be encouraged to use them free from commercial influences".[15;16]

Methods

Design

Using the method of population attributable risk,[17] we report estimates of attributable fractions, deaths and years of life lost due to suboptimal breastfeeding among children in the developing world. Population attributable risk measures the proportion of an outcome potentially avertable through reduced exposure to a risk factor. The outcomes considered here are childhood

deaths due to diarrhoeal disease and lower respiratory infections in 12 global subregions; the corresponding risk factor is suboptimal breastfeeding.

Reduced exposure is defined here as a counterfactual distribution of childhood breastfeeding defined with reference to current international recommendations.[13] As recommendations distinguish children 0–5, 6–11 and 12–23 months of age, all our analyses are stratified for these age groups. The counterfactual distribution constitutes the minimum level of exposure that is in principle possible; such a reference distribution has been termed "theoretical minimum exposure"[17]. We use the term "suboptimal breastfeeding" throughout to refer to practices inconsistent with this defined minimum exposure distribution.

Estimates of population attributable risk require information on current exposure status and the relative risk of the studied outcomes in the exposed group. Our exposure estimates were derived from a comprehensive review of breastfeeding patterns obtained from nationally representative data sources in the developing world. Calculation methods and validation procedures have been described elsewhere.[6] Our estimates of relative risk were derived from a re-analysis of data identified through a review of suitable epidemiological studies; the studies, as well as the analysis methods used to derive and validate our risk estimates, are described more fully below.

All outcome estimates refer to 2000, the year of the most recent comprehensive review of the global burden of disease, our source for data on deaths and years of life lost. Estimates of population attributable risk are mutiplied by total numbers of deaths and years of life lost due to diarrhoeal disease and lower respiratory infections to yield estimates of attributable deaths and attributable years of life lost.

Risk of suboptimal breastfeeding

Current knowledge

In 2000, the WHO Collaborative Study Team published a systematic review of the risk of death attributable to suboptimal breastfeeding.[7] For the purposes of an assessment such as the one presented here, a major limitation

of that study was that pooled analysis was possible only in terms of aggregate causes of death (e.g. all infections), and only for the dichotomous categories "any" vs "no" breastfeeding. Moreover, no pooled estimate was published for suboptimal breastfeeding in the second year of life.

Risk categories

For infants ≤6 months of age, theoretical minimum risk is defined as "exclusive breastfeeding". In conformity with available exposure data,[6] we define two increased risk categories, "partial breastfeeding" and "no breastfeeding," resulting in a polytomous minimum risk distribution for this age group. For infants >6 months of age and for children ≤2 years of age, theoretical minimum risk is defined as "continued breastfeeding", with "no breastfeeding" the sole increased risk category.[6]

Risk data

When the WHO Collaborative meta-analysis was published,[7] only one study with data for young infants in terms of polytomous exposure categories had been identified, a case-control study performed a number of years ago in Brazil.[9] A recent cohort study in Bangladesh[11] has since been published, also with polytomous exposure data. Each study presents certain limitations for the present purposes. While the Brazilian study characterized its reference group in terms of "predominant breastfeeding" rather than exclusive breastfeeding, the study performed in Bangladesh enrolled virtually no infants in the no-breastfeeding category. Furthermore, the small number of deaths recorded in the Bangladesh study severely limited its power to estimate mortality risk for infants >6 months of age.

Risk model

For use in the present study, risk data, consisting of observations of exposure and outcomes in infant and child populations, require aggregation in terms of the age groups and outcomes defined here as well as transformation, by means of a mathematical structure representing the relationships in this information, into summary estimates of risk.

For the two infant age groups, re-analysis of the Brazilian data affected the age distribution in the sample and it was necessary to analyse cases and controls irrespective of matching. Although ignoring matching can bias the association

only towards the hypothesis of no effect,[18] exploratory comparisons showed that estimates of risk were reduced only slightly.

We used maximum-likelihood estimation of a logistic regression model implemented with the Stata procedure *logit*. Selection of covariates was facilitated with a hierarchical model embodying a particular realization of a distal–proximal causal web. A model with all variables at the top (most distal) level was fit, and variables with *P*-values greater than 0.20 were removed, one at a time, starting with the highest *P*-value, until all remaining variables had *P*-values less than 0.20. The procedure was repeated for subsequent (more proximal) levels, keeping variables retained in previous levels irrespective of their *P*-values in subsequent regressions.

For the second year of life, five previous studies contained sufficient information to compute the risk of no breastfeeding on all cause mortality,[7] and we used those data to compute a variance-weighted pooled mean of all-cause-mortality risk for children >1 and ≤2 years of age using the Stata procedure *metan*.[19]

External validation

Risk estimates were compared where possible with those from other studies and settings to ensure that they were reasonably consistent with other data and could therefore be applied across the diverse settings studied here. We made detailed comparisons, in polytomous terms, to appropriately re-analysed data from the study in Bangladesh, and in dichotomous terms, to the two other data sets[20;21] with information on suboptimal breastfeeding and fatal outcomes in infants identified by the WHO Collaborative meta-analysis.

Outcomes

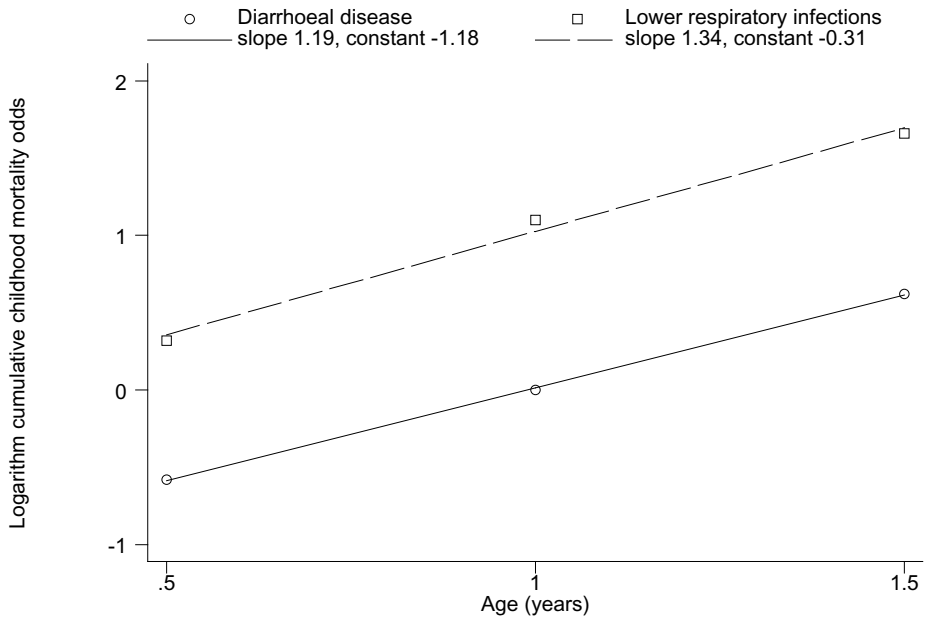
For infants, only deaths due to diarrhoeal disease and lower respiratory infections are considered, but for children ≤2 years of age, deaths due to all causes are included. The corresponding outcomes we derive are attributable fractions, attributable deaths and attributable years of life lost. An upper bound on deaths potentially caused by mother-to-child transmission of HIV through breastfeeding is estimated for 32 high-prevalence countries in the developing world. All estimates are reported for developing countries by UN subregion.[22]

We use WHO figures for cause-specific deaths and years of life lost in 2000, which are estimated subject to a variety of internal and external validity constraints[23;24] for the world and a set of defined epidemiological subregions.[25] WHO cause-specific death estimates are reported, by sex, for infants (≤ 1 year of age) and other children (>1 and ≤ 5 years of age) and six other age groups of varying width; estimates of all-cause deaths are reported, by sex, for infants, other children and successive five-year age groups.[26]

Age distribution of infant outcomes

As breastfeeding recommendations distinguish children 0–5, 6–11 and 12–23 months of age,[13] it was necessary to estimate the proportion of cause-specific or all-cause deaths occurring in various age subgroups. We used published data

Figure 1. Linear relation between the logarithm of cumulative childhood mortality odds and age for selected causes. Circles (diarrhoeal disease) and squares (lower respiratory infections) show pooled data points;[27] solid and dashed lines show the estimated relation. Estimated parameter values reported in the figure.



on the age distribution of deaths due to diarrhoeal disease and lower respiratory infections among children[27] to estimate a linear relation between the logarithm of childhood mortality odds and age for the relevant cause categories (Figure 1). The estimated slope determines the age distribution of cause-specific deaths and the constant, the overall level of risk in the published data.

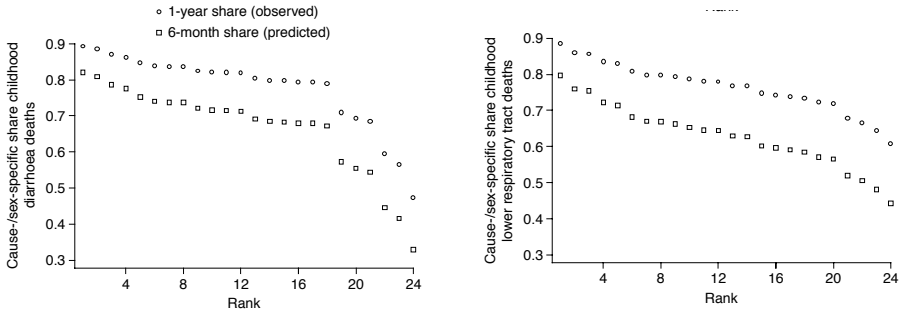
To apply this relation to the subregions defined here, we assumed that the age distribution of deaths remains the same across regions but that the level of risk is context specific. Accordingly, the estimated relations were calibrated, by changing the constant, to fit WHO estimates of cause-and-sex-specific deaths occurring by 1 year of age for each subregion. The calibrated relation was then used to predict subregion-specific deaths by 6 months of age. The resulting estimates (predicted) are shown in Figure 2 by cause and sex, along with the reported WHO estimates (observed) for comparison. We assumed that the subregional age distributions of years of life lost by 6 months of age were the same as those we estimated for deaths.

A similar procedure was used to obtain estimates of the proportion of childhood mortality occurring by the end of the second year of life. Here, though, we used estimated all-cause mortality risk at one and five years of age obtained from each country's life table[26] and simply interpolated, on the odds scale, values for the second year of life.

Offsetting outcomes

Since, for HIV-positive mothers, breastfeeding can result in transmission of HIV infection to the infant, international recommendations take into account the need to assess competing risks. As a shorthand locution, we use the term "offsetting outcomes" throughout to denote the expected number of childhood deaths or years of life lost that would be caused by HIV-positive mothers were they to breastfeed their children at the same counterfactual rate as non-HIV-positive mothers. It is important to distinguish this artificial concept of "offsetting deaths" (based on an assumption of breastfeeding without regard to HIV status and, therefore, in a manner inconsistent with international recommendations) from deaths that might potentially be caused as a result of the vertical transmission of HIV infection through full compliance with international breastfeeding recommendations, which specifically require a judgement concerning the balance of risk. "Offsetting deaths" as used here thus refers to theoretical deaths

Figure 2. Share of childhood deaths (diarrhoeal disease, upper panel; lower respiratory infections, lower panel).
Observed (circles, 1 year) and predicted (squares, 6 month) cause-and-sex-specific shares by subregion, by rank.



that would attributable to breastfeeding by HIV-positive mothers were they to breastfeed fully but without regard to individual- and setting-specific factors. With such "indiscriminate" breastfeeding (i.e. without considering individual- and setting-specific factors), "offsetting outcomes" would reduce the population benefit achievable by breastfeeding because some of the breastfed children will die from AIDS. While a number of offsetting deaths of children of HIV-positive mothers might be unavoidable even taking into account individual- and setting-specific factors, logically, such deaths must always be fewer than those attributable to so-called indiscriminate breastfeeding by HIV-positive mothers. We estimate here the offsetting outcomes of indiscriminate breastfeeding as an upper bound on the potential harms of breastfeeding by HIV-positive mothers.

We analysed offsetting deaths in 32 high-prevalence countries (Footnote, Table 5). The population at risk for mother-to-child transmission through breastfeeding was assumed to be the subregion's infant population, and we assumed independence in the probabilities of a mother's being HIV positive,[28] having a live-born child survive the neonatal period and breastfeeding her child for at least one year. The probability of HIV transmission for children of infected mothers who breastfeed for at least one year was assumed to be 15%,[15] and homogeneous across breastfeeding categories. The probability of mothers' breastfeeding for at least one year was calculated as the complement of the no breastfeeding rate among infants.[6] We assumed that all infants becoming HIV positive through breastfeeding die at exact age 5 years.[29]

Results

Risk of suboptimal breastfeeding

Table 1 shows a comparison, by exposure status and cause of death, of adjusted and crude estimates of risk for infants from the re-analysed Brazilian data. Pooled and individual estimates of the risk of no breastfeeding for deaths due to all causes during the second year of life are shown in the "forest plot"[19] depicted in Figure 3.

External validation

To validate the polytomous estimates for infants, data from the studies in Brazil and Bangladesh were tabulated on a comparable basis and re-analysed; re-analysis for Bangladesh was possible only for infants ≤6 months of age (*personal communication*, Shams Arifeen). After adjusting for a similar set of covariates (Footnote, Table 1), levels of risk in the two studies are similar and show overlapping confidence intervals (Table 2).

Table 1 Crude and adjusted* odds ratios (OR) and 95% confidence intervals (CI) of infant mortality due to suboptimal breast-feeding, by cause of death and exposure category. Estimates obtained from reanalysis of data from the study reported in reference 9

Age group/cause of death	Suboptimal breast-feeding category	OR, crude estimate (95% CI)	OR, adjusted estimate (95% CI)
0–5 months			
Diarrhoeal disease	Partial	3.9 (2.0–7.7)	2.4 (1.1–5.2)
Diarrhoeal disease	No	8.1 (4.6–14.5)	7.3 (3.7–14.4)
Lower respiratory tract infections	Partial	3.2 (1.6–6.4)	2.8 (1.4–5.8)
Lower respiratory tract infections	No	5.6 (3.1–10.0)	4.7 (2.4–9.1)
6–11 months			
Diarrhoeal disease	No	2.2 (1.0–4.5)	2.5 (1.1–5.7)
Lower respiratory tract infections	No	2.3 (0.9–5.7)	2.8 (1.0–7.7)

* Covariates controlled for were father living in the same household, income per capita, infant's age in months, number of previous pregnancies, mother's age in years, birth weight in grams (lower respiratory tract infections, 0–5 months); mother's education in years, mother's ethnicity, infant's age in months, household has flush toilet, birth interval in months, mother's age in years, birth weight in grams, hospital birth, delivery by Caesarian section, number of antenatal visits (diarrhoeal disease, 0–5 months); mother's education in years, household head employed, infant's age in months, birth weight in grams, low Apgar score at birth (lower respiratory infections, 6–11 months); and income per capita by quartile, male infant, infant's age in months, household has piped water, number of persons per sleeping room, number of previous pregnancies, mother's age in years (diarrhoeal disease, 6–11 months).

Figure 3. Risk estimates for non-breastfeeding vs continued breastfeeding during the second year of childhood (all causes of death) in five different settings. Relative size of squares indicates relative sample size (% weight) and the diamond shows the variance-weighted mean of the squares (overall).

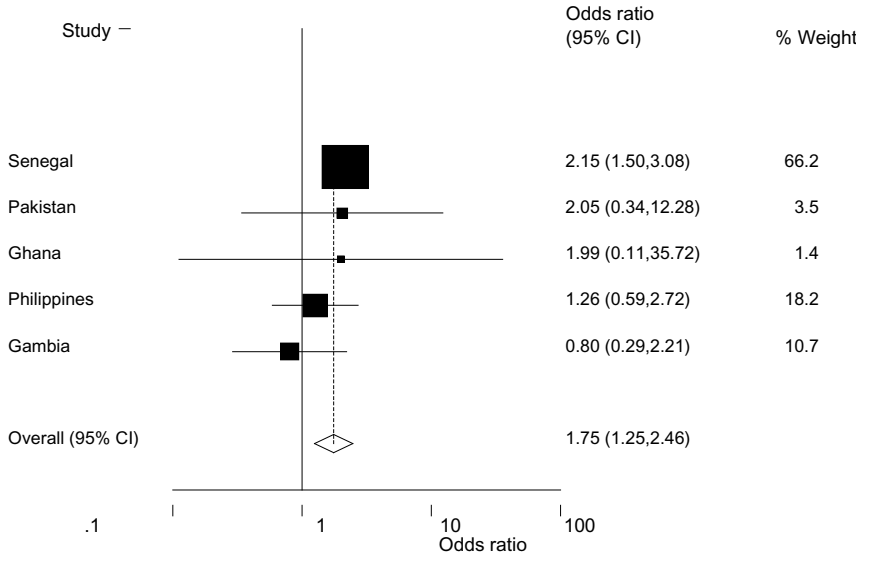


Figure 4 compares age-specific levels of risk for deaths due to infectious diseases found in Brazil with those in each of the other two studies (Philippines[20] and Pakistan[21]) where dichotomous (i.e. "any" vs "no" breastfeeding) comparisons were possible. The plot suggests decreasing age-specific risk, linear in the logarithm of the odds ratio. The estimates from Brazil are consistent with those from the other settings and are generally intermediate in risk.

Attributable fractions

Attributable fractions denote the proportion of an outcome attributable to the avoided exposure and are reported by cause in Table 3. For infants ≤6 months of age, attributable fractions are highest in Latin America and the Caribbean, followed by Africa and Asia. Globally, attributable fractions for younger infants account for more than half of fatal outcomes due to the selected causes in developing countries. For older infants, attributable fractions are highest

in Latin America and the Caribbean, followed by Asia and Africa. Globally, attributable fractions for older infants account for nearly one-fifth of fatal outcomes due to selected causes in developing countries. For children in the second year of life, attributable fractions are highest in Latin America and the Caribbean, followed by Africa and Asia, and globally account for one-fifth of all fatal outcomes in developing countries.

Attributable deaths

Attributable deaths are calculated by multiplying attributable fractions by total deaths and are reported by cause in Table 4. Globally, we estimate that 1.45 m childhood deaths are attributable to suboptimal breastfeeding in developing countries. A total of 96% of all attributable infant deaths (i.e. 1.24 m deaths) occur during the first 6 months of life, with 168,000 attributable deaths occurring during the second year of life (12% of all attributable childhood deaths). Attributable childhood deaths are much higher in Asia (830,000) and Africa (572,000) than in Latin America and the Caribbean (50,000), due to the effects of larger population (Asia) and higher underlying death rates (Africa and parts of Asia). Globally, attributable infant deaths due to lower respiratory infections are higher (685,000) than those due to diarrhoeal disease (600,000), although the pattern is reversed in Latin America and the Caribbean (20,000 and 23,000 deaths, respectively).

Table 2. Comparison of polytomous risk estimates (odds ratios and 95% confidence intervals) for infant mortality due to selected causes, by exposure category and age group. Estimates obtained from reanalysis of data from the studies reported in references 9 Victora et al.) and 11 (Arifeen et al.)

Cause of death, age group	Study	No breast-feeding	Partial breast-feeding	Predominant/exclusive breast-feeding*
Diarrhoeal disease, 0–5 months†	Arifeen *et al.*	4.2 (0.5–34.3)	3.6 (1.6–8.1)	1.0
Diarrhoeal disease, 0–5 months	Victora *et al.*	7.3 (3.7–14.4)	2.4 (1.1–5.2)	1.0
Lower respiratory tract infections, 0–5 months†	Arifeen *et al.*	5.7 (1.2–26.5)	2.7 (1.4–5.4)	1.0
Lower respiratory tract infections, 0–5 months	Victora *et al.*	4.7 (2.4–9.1)	2.8 (1.4–5.8)	1.0
Diarrhoeal disease, 6–11 months‡	Victora *et al.*	2.5 (1.1–5.7)	1.0	–
Lower respiratory tract infections, 6–11 months‡	Victora *et al.*	2.8 (1.0–7.7)	1.0	–

* Reference category with predominant and exclusive breast-feeding groups pooled.
† Reanalysed hazard ratios for infants aged 0–5 months, adjusted for selected confounding variables (S Arifeen, personal communication). Confounding variables included child's age and sex; mother's age, education, height, place of birth, parity and history of previous child death; father's education; and household religion, monthly income and economic status.
‡ Estimates for this age group could not be calculated (S Arifeen, personal communication).

Attributable years of life lost

Attributable years of life lost are calculated by multiplying attributable fractions by total years of life lost and are reported by cause and age group in Table 5. Globally, 117 m years of life lost are attributable to suboptimal breastfeeding among children in developing countries.

Offsetting outcomes

Offsetting deaths and offsetting years of life lost potentially caused by mother-to-child transmission of HIV through indiscriminate breastfeeding are reported in Table 6. Globally, offsetting deaths (242,000) potentially amount to 17% of those attributable to suboptimal breastfeeding in the developing world, but are as high as 36% in Africa as a whole and 48% in Eastern Africa. In Southern Africa, however, deaths potentially caused by mother-to-child transmission of HIV through breastfeeding are 3 times as high as those attributable to suboptimal breastfeeding.

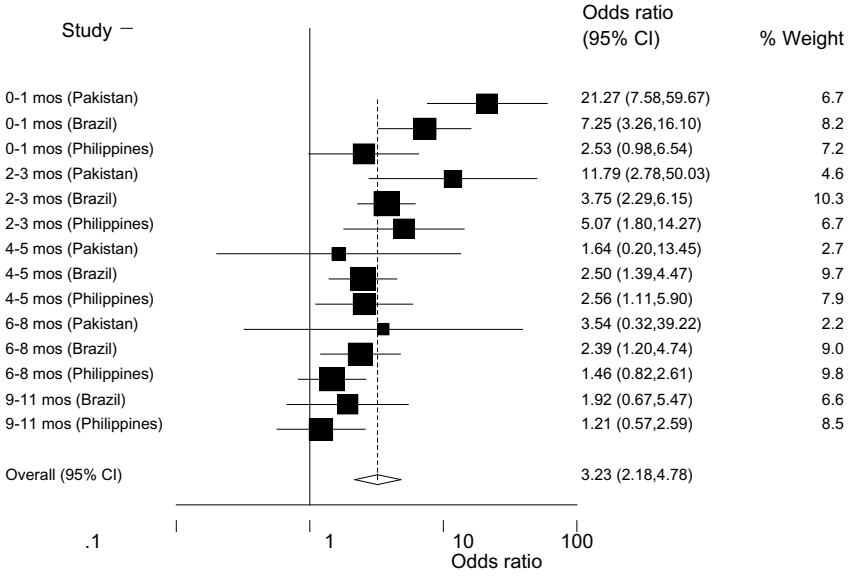
Discussion

Comparisons

Compared with estimates for other childhood risk factors, our estimate of childhood deaths attributable suboptimal breastfeeding (1.45 m) is substantially lower than estimates of deaths attributable to unsafe water, sanitation and hygiene (1.73 m) and less than half (39%) as high as childhood deaths attributable to underweight (3.75 m).[1] Fatal adult risk factors are responsible for many more deaths: for example, high blood pressure (7.14 m) and cholesterol (4.42 m) jointly account for 11.6 m cardiovascular deaths worldwide.[1]

However, when outcomes are measured in terms of attributable years of life lost, suboptimal breastfeeding (117 m years) ranks second only to underweight (127 m years) and well ahead of the next largest childhood or adult risk factor studied by the WHO Collaborating Group (unsafe sex, 77 m years).[1] High blood pressure (56 m years) and cholesterol (35 m years) jointly account for 91 m years of life lost.[1] The change in the ranking of risk factors when compared with different outcome measures is due to the fact that deaths caused by suboptimal breastfeeding occur at earlier ages than those attributable to other risk factors.

Figure 4. Comparison of age specific crude risk estimates for non-breastfeeding vs any breastfeeding in infants (all infectious causes of death). Relative size of squares indicates relative sample size (% weight) and the diamond shows the variance-weighted mean of the squares (overall).



Previously, we estimated that 52,000 infant deaths due to diarrhoeal disease and lower respiratory infections in Latin America and the Caribbean were attributable to suboptimal breastfeeding.[30] Our present estimate for the same region, ages and causes is 43,000 deaths, lower by 18% but within the uncertainty interval of the former estimate (31,000–73,000). The difference can be explained by the fact that the relative risks used here are lower.

Our estimate of childhood attributable deaths (1.45 m) appears to be higher than that published by the Bellagio Child Survival Group (1.3 m). However, as we include deaths during the second year of life and do not restrict the analysis to 42 countries,[3] our total should be higher. Comparing only attributable infant deaths, our estimate (1.28 m), which does not account for deaths potentially caused by mother-to-child transmission of HIV, is lower

Table 3. The fraction of mortality attributable to suboptimal breast-feeding by sub-region, age group and cause of death

	Infants < 6 months of age		Infants 6–11 months of age		Children 12–23 months of age
Region/sub-region	DD	LRTI	DD	LRTI	All causes
Africa	59	55	13	11	19
Eastern	52	48	8	7	16
Middle	59	54	6	5	16
Northern	57	58	29	25	29
Southern	66	67	35	31	30
Western	63	58	5	5	17
Asia (excluding Japan)	52	50	18	16	18
Eastern	46	45	21	18	18*
South-Central	52	48	11	9	14
South-Eastern	56	55	30	26	23
Western	63	62	34	30	33
Latin America and the Caribbean	61	65	42	38	34
Caribbean	61	65	42	38	34
Central America	64	68	42	37	33
South America	60	63	42	38	34
Developing countries	55	53	20	18	20

DD – diarrhoeal disease; LRTI – lower respiratory tract infection.
* Since exposure data for Eastern Asia were not available for the second year of life, the attributable fraction is the regional average attributable fraction for the age group.

by 20,000 deaths than the figure published by the Bellagio Group. However, before accounting for the offsetting effects of HIV/AIDS, their figure is 1.5 m deaths (i.e. 17% higher than our corresponding estimate). It is difficult to identify the reasons for this difference since the data sources and methods used by the Bellagio Group are not publicly available.[3, Ref. 53]

Table 4. Numbers of deaths attributable to suboptimal breast-feeding, by sub-region, age group and cause of death

	Attributable deaths (000 s)				
	Infants < 6 months of age		Infants 6–11 months of age		Children 12–23 months of age
Region/sub-region	DD	LRTI	DD	LRTI	All causes
Africa	197.0	255.7	7.4	7.1	104.5
Eastern*	68.2	113.0	2.7	2.3	32.6
Middle	30.1	35.8	1.1	0.6	18.0
Northern*	17.8	19.6	1.2	1.7	6.3
Southern	7.4	2.4	0.6	0.3	3.9
Western	73.6	84.9	1.8	2.2	43.6
Asia (excluding Japan)	359.3	385.5	12.8	16.4	56.4
Eastern*	35.8	35.2	2.2	2.4	7.7‡
South-Central	265.6	304.3	6.2	8.8	33.7
South-Eastern	38.2	23.1	3.2	3.0	9.0
Western*	19.6	22.9	1.2	2.2	5.9
Latin America and the Caribbean	20.9	18.1	2.1	2.3	7.0
Caribbean*	2.9	2.2	0.2	0.3	1.3
Central America	4.5	5.7	0.5	0.6	2.0
South America	13.4	10.2	1.4	1.3	3.6
Developing countries†	577.2	659.3	22.3	25.8	167.8

DD – diarrhoeal disease; LRI – lower respiratory tract infection.
* A small number of developing countries were omitted as they are not member countries of the World Health Organization and basic demographic and epidemiological data were not available: Reunion (Eastern Africa); Western Sahara (Northern Africa); Hong Kong Special Administrative Region and Macau (Eastern Asia); Gaza Strip (Western Asia); Guadeloupe, Martinique, Netherlands Antilles and Puerto Rico (Caribbean).
† The United Nations region Oceania (including developing countries Guam, Fiji, French Polynesia, New Caledonia, Papua New Guinea, Samoa, Solomon Islands and Vanuatu) was omitted as exposure data were not available for these countries.
‡ Since exposure data for Eastern Asia were not available for the second year of life, the estimate of deaths is based on the regional average attributable fraction for the age group.

Table 5. Years of life lost attributable to suboptimal breast-feeding, by sub-region, age group and cause of death

	Attributable years of life lost (000 s)				
	Infants < 6 months of age		Infants 6–11 months of age		Children 12–23 months of age
Region/sub-region	DD	LRTI	DD	LRTI	All causes
Africa	15 834	20 541	617	587	8477
Eastern*	5478	9107	223	190	2644
Middle	2403	2871	90	51	1462
Northern*	1433	1582	100	142	513
Southern	597	189	51	22	320
Western	5923	6792	153	182	3539
Asia (excluding Japan)	28 980	31 098	1070	1371	4576
Eastern*	2893	2866	188	200	626‡
South-Central	21 416	24 533	520	733	2736
South-Eastern	3085	1853	265	254	730
Western*	1586	1845	97	184	483
Latin America and the Caribbean	1682	1459	179	190	569
Caribbean*	232	178	20	27	108
Central America	366	461	42	52	165
South America	1084	819	117	110	296
Developing countries†	46 496	53 097	1866	2148	13 622

DD – diarrhoeal disease; LRTI – lower respiratory tract infection.
* A small number of developing countries were omitted as they are not member countries of the World Health Organization and basic demographic and epidemiological data were not available: Reunion (Eastern Africa); Western Sahara (Northern Africa); Hong Kong Special Administrative Region and Macau (Eastern Asia); Gaza Strip (Western Asia); Guadeloupe, Martinique, Netherlands Antilles and Puerto Rico (Caribbean).
† The United Nations region Oceania (including the developing countries Guam, Fiji, French Polynesia, New Caledonia, Papua New Guinea, Samoa, Solomon Islands and Vanuatu) was omitted as exposure data were not available for these countries.
‡ Since exposure data for Eastern Asia were not available for the second year of life, the estimate of years of life lost is based on the regional average attributable fraction for the age group.

Table 5. Years of life lost attributable to suboptimal breast-feeding, by sub-region, age group and cause of death

Region/sub-region	Offsetting deaths (000 s)*	Offsetting years of life lost (000 s)*
Africa	208.6	16 240
Eastern	106.1	8255
Middle	17.6	1373
Northern	2.8	218
Southern	43.7	3401
Western	38.5	2994
Asia (excluding Japan)	31.7	2465
Eastern	3.1	239
South-Central	24.4	1898
South-Eastern	4.2	329
Western	–	–
Latin America and the Caribbean	1.8	138
Caribbean	–	–
Central America	–	–
South America	1.8	138
Developing countries	242.1	18 842

* Analysed for 32 selected high-prevalence countries in the developing world: Burundi, Ethiopia, Kenya, Malawi, Mozambique, Rwanda, Uganda,United Republic of Tanzania, Zambia, Zimbabwe (Eastern Africa); Angola,Cameroon, Central African Republic, Democratic Republic of the Congo (Middle Africa); Sudan (Northern Africa); Botswana, Lesotho, Namibia,South Africa, Swaziland (Southern Africa); Burkina Faso, Côte d'Ivoire, Ghana, Guinea, Nigeria (Western Africa); China (Eastern Asia); India (South-Central Asia); Indonesia, Myanmar, Thailand, Vietnam (South-Eastern Asia); Haiti (Caribbean).

Offsetting outcomes

Estimates of offsetting outcomes are sensitive to the estimated probability of HIV transmission in breastfeeding children of infected mothers, which could be as much as 33% higher or lower than the figure used here while still consistent with available evidence.[15] Offsetting outcomes for individual subregions will also be sensitive to estimates of HIV prevalence among women of childbearing age.[28] Furthermore, while pooled data from developed countries show 70% survival at 15-year follow up for seroconverters <5 years of age,[29] corresponding values for developing countries are probably much lower, and we arbitrarily assumed survival to age 5.

On the other hand, assuming an homogeneous risk of vertical transmission by breastfeeding status may overstate offsetting deaths.[31] In addition, although WHO has estimated that 1.6 m babies are born each year to HIV-positive mothers,[15] the independent-risk model we use here predicts that nearly 2 m babies are born annually to HIV-positive mothers in just 32 high-prevalence countries, in other words, possibly a substantial overestimate due to not accounting for the effect of seropositivity on fertility. Moreover, estimating vertical HIV transmission with cross-sectional breastfeeding prevalences such as those used here[6] probably results in overestimates; nevertheless, since infants of infected mothers who breastfeed for less than one year are presumably exposed to some risk, we assumed all such infants face the same risk of HIV transmission as those who breastfeed for one full year.

Substantial inter-regional heterogeneity in the relative magnitude of the benefits and harms of indiscriminate breastfeeding underscores that this important child-health intervention requires the sort of individual and setting-specific assessment recommended by WHO.[13] Moreover, an ecological analysis like that reported here cannot precisely estimate the fatal outcomes that could be offset by HIV infection for the very reason that recommended practice depends on the mother's HIV status and other individual- and setting-specific factors.

Offsetting outcomes, however, establish an upper bound to fatal outcomes attributable to breastfeeding by HIV-positive mothers. Notably, even in the

hypothetical case of breastfeeding without regard to HIV status, the offsetting outcomes would not alter the importance of suboptimal breastfeeding as a global risk factor. Without the possibility of being more precise, the estimate of fatal childhood outcomes attributable to non-compliance with international breastfeeding recommendations lies in the range 1.21–1.45 m deaths (98–117 m years of life lost). The lower bound of this range represents a potential reduction in the global benefit of breastfeeding of nearly one-fifth (19%) compared with widely reported estimates of the number of deaths avoidable by improved breastfeeding prior to the explosion of the AIDS pandemic in Africa.[32;33]

Multi-causality

Recognizing the multiple causality of outcomes, double counting of deaths due to unsafe water, sanitation and hygiene and those due to suboptimal breastfeeding (both contributing to diarrhoeal disease), between suboptimal breastfeeding and underweight (both contributing to diarrhoeal disease and all-cause deaths in the second year of life), or between suboptimal breastfeeding and indoor air quality (both contributing to lower respiratory infections) cannot be excluded.[17] The synergistic effect of selected childhood diseases and risk factors has moreover been highlighted.[2]

Confounding and measurement error

As mothers in developing countries who breastfeed their babies tend to be poorer than those who do not, residual confounding by socio-economic status could result in underestimates of the risk of suboptimal breastfeeding due to clustering of common risk factors with poverty. However, other potential confounding factors could lead to overestimates: since babies born with higher birth weights are more likely to breastfeed, mothers' failure to breastfeed underweight babies (i.e. with higher risk of mortality) may result in "selection bias".

One may further argue that, while our risk estimates are calculated relative to a baseline of "predominant breastfeeding",[9] we measure the prevalence of unexposed in terms of exclusive breastfeeding. However, we have argued in other work,[6] on the basis of both internal and external validity criteria,

that infants reported in surveys as exclusively breastfed typically include a notably high proportion (ca. 15%) who are receiving other liquids or even solid foods. We believe it is consequently unlikely that the nominal exposure data derived from nationally representative surveys, on the one hand, and from epidemiological studies that form the basis of the estimates of risk of suboptimal breastfeeding, on the other, are inherently comparable despite their use of similar terms to describe breastfeeding status. On balance, therefore, assessment of the existence or net effect of bias due to confounding or measurement error is uncertain.

Implications for developed countries

As childhood deaths due to diarrhoeal disease and lower respiratory infections are very rare in developed countries, and second-year childhood deaths even rarer,[23;24;26] the present study is confined to developing countries. However, if outcomes were defined to include breastfeeding-attributable morbidity, developed and developing countries alike would have a substantial attributable burden of disease.[34-36]

Conclusions

The age distribution of attributable outcomes reinforces the importance of emphasizing exclusive breastfeeding for infants ≤6 months of age. The substantial benefits of continued breastfeeding in the second year of life, moreover, highlight the importance of this neglected aspect of infant-and-child feeding recommendations.

The size of the gap between practice and recommendations is striking, especially considering that breastfeeding involves no out-of-pocket costs, that there exists universal consensus on best practices, and that implementing current international recommendations could save between 1.21 and 1.45 m lives (between 98 and 117 m years of life) among children in the developing world each year. Although the HIV/AIDS pandemic has substantially mitigated the benefits of universal (i.e. indiscriminate) breastfeeding, especially in parts of Africa, full compliance with current international breastfeeding recommendations remains one of the most important child health goals.

Moreover, breastfeeding is arguably cheaper and more readily implementable than interventions such as improvements in water, sanitation and hygiene, or in general nutritional status, since unlike so many interventions the principal barriers to improved breastfeeding are not to be found in the direct monetary costs of delivery.

Nevertheless, breastfeeding, especially exclusive breastfeeding, imposes a time and energy cost on the mother, and may also have important indirect costs, such as the inability of breastfeeding mothers to participate in the wage-earning workforce. Notwithstanding, virtually all mothers can breastfeed provided they have accurate information and support within their families and communities and from the health system. Recent findings from two experimental studies[37;38] and one quasi-experimental one[39] demonstrate that breastfeeding counselling given in hospitals and in the community is a cost-effective way of increasing exclusive breastfeeding. A similar conclusion was reached by a systematic review of the effects of breastfeeding support.[40]

However, implementation of large-scale breastfeeding counselling programmes demands substantial resources and political commitment. Ideally, breastfeeding support would be provided as an integral part of routine health services for mothers. This implies the need to train health workers, lay and peer counsellors and certified lactation consultants who can help build mothers' confidence, improve feeding technique and prevent or resolve problems. Similarly, women in paid employment require minimum enabling conditions to prolong the duration of exclusive breastfeeding, such as paid maternity leave, part-time work arrangements, on-site nurseries, facilities for expressing and storing breast milk and breastfeeding breaks.

A remaining challenge is to understand better the causes that determine the practice of exclusive breastfeeding in different settings. These include national breastfeeding policies; socio-economic and cultural factors such as medical attitudes, commercial advertising, family pressures, maternal work demands, women's employment legislation and maternal beliefs; biological factors such as infant's size, sex, growth rate, development, interest/desire and appetite and maternal lactation capacity; and the local epidemiology of HIV/AIDS.

Acknowledgements

The views expressed are solely those of the authors and do not necessarily reflect the decisions or the stated policy of the World Health Organization. We are grateful to Shams Arifeen for re-analysing the data from Bangladesh and to Cesar G. Victora for guidance throughout the course of the research.

Competing interests

None declared.

Authors' contributions

JAL and APB designed the study, analysed the data and results and drafted the manuscript. AJDB designed and interpreted the statistical methods used to re-analyse risk in the data from the Brazilian study. MdO drafted key manuscript sections and participated in the interpretation of data and results. All authors revised and approved the final manuscript.

References

1. The World Health Report 2002. Reducing risks, promoting healthy life. Geneva, Switzerland: World Health Organization, 2002.

2. Black RE, Morris SS, Bryce J. Where and why are 10 million children dying every year? *Lancet* 2003;**361**:2226-34.

3. Jones G, Steketee RW, Black RE, Bhutta ZA, Morris SS. How many child deaths can we prevent this year? *Lancet* 2003;**362**:65-71.

4. Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S, Murray CJ, Comparative Risk Assessment Collaborating Group. Selected major risk factors and global and regional burden of disease. *Lancet* 2002;**360**:1347-60.

5. Ezzati M, Vander Hoorn S, Rodgers A, Lopez AD, Mathers CD, Murray CJ. Estimates of global and regional potential health gains from reducing multiple major risk factors. *Lancet* 2003;**361**:271-80.

6. Lauer JA, Betrán AP, Victora CG, de Onis M, Barros AJ. Breastfeeding patterns and exposure to suboptimal breastfeeding among children in developing countries: review and analysis of nationally representative surveys. *BMC Medicine* 2004;**2**:26.

7. WHO Collaborative Study Team on the Role of breastfeeding on the Prevention of Infant Mortality. Effect of breastfeeding on infant and child mortality due to infectious diseases in less developed countries: a pooled analysis. *Lancet* 2000;**355**:451-5.

8. Brown KH, Black RE, Lopez de Romaña G, Creed de Kanashiro H. Infant feeding practices and their relationship with diarrhoeal and other diseases in Huascar (Lima), Peru. *Pediatrics* 1989;**83**:31-40.

9. Victora CG, Vaughan JP, Lombardi C, Fuchs SMC, Gigante LP, Smith PG *et al*. Evidence for protection by breast-feeding against infant deaths from infectious diseases in Brazil. *Lancet* 1987;**2**:319-22.

10. Hanson LA. Breast-feeding as a protection against gastroenteritis and other infections. *Acta paediatrica scandinavica* 1985;**74**:641-2.

11. Arifeen SE, Black RE, Antelman G, Baqui AH, Caulfield LE, Becker S. Exclusive breastfeeding reduces acute respiratory infection and diarrhea deaths among infants in Dhaka slums. *Pediatrics* 2001;**108**:E67.

12. Kramer, M. S. and Kakuma, R. The optimal duration of exclusive breastfeeding. A systematic review. WHO/NHD/01.08. 2002. Geneva, Switzerland, World Health Organization.

13. World Health Assembly. Infant and young child nutrition. Resolution WHA54.2 (document WHA54.2; available from: http://www.who.int/gb/ebwha/pdf_files/WHA54/ea54r2.pdf). 2001. Geneva, Switzerland, World Health Organization.

14. The optimal duration of exclusive breastfeeding. Report of an expert consultation, Geneva, Switzerland, 28-30 March 2001. WHO/NHD/01.09. 2002. Geneva, Switzerland, World Health Organization.

15. Global strategy for infant and young child feeding. Geneva, Switzerland: World Health Organization, 2003.

16. Newell ML. Prevention of mother-to-child transmission of HIV: challenges for the current decade. *Bulletin of the World Health Organization* 2001;**79**:1138-44.

17. Murray CJ, Ezzati M, Lopez AD, Rodgers A, Vander Hoorn S. Comparative quantification of health risks: Conceptual framework and methodological issues. *Population Health Metrics* 2003;**1**:1.

18. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia, USA: Lippincott Williams and Wilkins, 1998.

19. Bradburn, M. J., Deeks, J. J., and Altman, D. G. *Stata technical bulletin 44* (sbe24). Internet communication of 8 January 2002 at web site http://www.stata.com. 2002.

20. Yoon PW, Black RE, Moulton LH, Becker S. Effect of not breastfeeding on the risk of diarrheal and respiratory mortality in children under 2 years of age in Metro Cebu, The Philippines. *American journal of epidemiology* 1996;**143**:1142-8.

21. Hanson LA, Ashraf R, Karlber J, Lindbland BS, Jalil F. Breast feeding is a natural contraceptive and prevents disease and death in infants, linking infant mortality and birth rates. *Acta paediatrica* 1994;**83**:3-6.

22. World Population Prospects. The 2000 Revision. New York: United Nations. Department of Economic and Social Affairs. Population Division, 2001.

23. Murray, C. J., Lopez, A. D., Mathers, C. D., and Stein, S. The Global Burden of Disease 2000 project: aims, methods and data sources. (GPE Discussion paper No. 36). 2001. Geneva, Switzerland, World Health Organization.

24. Mathers, C. D., Stein, S., Ma Fat, D., Rao, C., Inoue, M., Tomijima, N., Lopez, A. D., and Murray, C. J. Global Burden of Disease 2000: Version 2 methods and results. (GPE Discussion paper No. 50). 2002. Geneva, Switzerland, World Health Organization.

25. The World Health Report 2003. Shaping the future. Geneva, Switzerland: World Health Organization, 2003.

26. Lopez, A. D., Ahmad, O. B., Guillot, M., Inoue, M., Ferguson, B. D., and Salomon, J. A. Life tables for 191 countries for 2000: data, methods, results. GPE Discussion paper No. 40. 2001. Geneva, Switzerland, World Health Organization.

27. Kirkwood BR, Gove S, Rogers S, Lob-Levyt J, Arthur P, Campbell H. Potential interventions for the prevention of childhood pneumonia in developing countries: a systematic review. *Bulletin of the World Health Organization* 1995;**73**:793-8.

28. UNICEF. http://www.unicef.org/infobycountry/. 2003. Accessed December 4, 2003.

29. Collaborative Group on AIDS Incubation and HIV Survival including the CASCADE EU concerted Action. Time from HIV-1 seroconversion to AIDS and death before widespread use of highly-active antiretroviral therapy: a collaborative re-analysis. *Lancet* 2000;**355**:1131-7.

30. Betrán AP, de Onis M, Lauer JA, Villar J. Ecological study of effects of breast feeding on infant mortality in Latin America. *BMJ* 2001;**323**:303-6.

31. Coutsoudis A, Pillay K, Spooner E, Kuhn L, Coovadia HM. Influence of infant-feeding patterns on early mother-to-child transmission of HIV-1 in Durban, South Africa: a prospective cohort study. *Lancet* 1999;**354**:471-6.

32. Infant and young child nutrition (progress and evaluation report; and status of implementation of the International Code of Marketing of Breast-milk Substitutes): Report by the Director-General. Document EB/9317. 2001. Geneva, Switzerland, World Health Organization.

33. Walker ARP, Adam FI. Breast-feeding in sub-Saharan Africa: outlook for 2000. *Public Health Nutrition* 2000;**3**:285-92.

34. Silfverdal SA. Protective effect of breastfeeding on invasive Haemophilus influenzae infection: a case-control study in Swedish preschool children. *International journal of epidemiology* 1997;**26**:443-50.

35. Levine OS, Farley M, Harrison LH, Lefkowitz L, McGeer A, Schwartz B. Risk factors for invasive pneumococcal disease in children: a population-based case-control study in North America. *Pediatrics* 1999;**103**.

36. Nafstad P. Breastfeeding, maternal smoking and lower respiratory tract infections. *European respiratory journal* 1996;**9**:2623-9.

37. Morrow AL, Guerrero ML, Shults J, Calva JJ, Lutter C, Bravo J *et al*. Efficacy of home-based peer counselling to promote exclusive breastfeeding: a randomised controlled trial. *Lancet* 1999;**353**:1226-31.

38. Haider R, Ashworth A, Kabir I, Huttly SR. Effect of community-based peer counsellors on exclusive breastfeeding practices in Dhaka, Bangladesh: a randomised controlled trial. *Lancet* 2000;**356**:1643-7.

39. Lutter CK, Perez Escamilla R, Segall A, Sanghvi T, Teruya K, Wickham C. The effectiveness of a hospital based program to promote exclusive breastfeeding among low income women in Brazil. *American journal of public health* 1997;**87**:659-63.

40. Sikorski, J., Renfrew, M. J., Pindoria, S., and Wade, A. Support for breastfeeding mothers (Cochrane Review). In: The Cochrane Library, Issue 1 . 2002. Oxford: Update Software.

Annex: Supplementary material

Table A1. Prevalence of breastfeeding indicators, by subregion and age groups.[6]

Region/subregion	Breastfeeding prevalence estimates (%)						
	Infants <6 months of age			Infants 6-11 months of age		Children 12-23 months of age	
	Exclusive	Partial	None	Continued	None	Continued	None
Africa	**24.9**	**71.2**	**3.9**	**91.8**	**8.2**	**69.9**	**30.1**
Eastern	41.4	56.1	2.4	95.1	4.9	75.6	24.4
Middle	19.4	79.6	1.0	96.6	3.4	76.8	23.2
Northern	36.5	53.9	9.6	77.7	22.3	49.6	50.4
Southern	8.2	75.7	16.0	70.4	29.6	46.7	53.3
Western	6.1	92.1	1.8	96.8	3.2	74.9	25.1
Asia	**44.9**	**50.7**	**4.5**	**87.5**	**12.5**	**72.4**	**27.6**
Eastern*	58.6	36.3	5.1	85.7	14.3	ND	ND
South-Central	42.1	55.3	2.6	93.3	6.7	78.8	21.2
South-Eastern	37.5	55.0	7.5	76.7	23.3	61.7	38.3
Western	17.7	72.0	10.3	71.3	28.7	37.3	62.7
Latin America and the Caribbean	**30.8**	**51.2**	**18.0**	**59.9**	**40.1**	**36.5**	**63.5**
Caribbean	25.8	63.8	10.4	64.6	35.4	34.2	65.8
Central America	23.4	55.0	21.6	60.2	39.8	37.0	63.0
South America	35.1	48.0	16.9	59.3	40.7	36.4	63.6
Developing countries	**38.7**	**55.7**	**5.6**	**85.8**	**14.2**	**68.3**	**31.7**

Table A2. Odds ratios of suboptimal breastfeeding and selected control variables for infant mortality due to lower respiratory infections among infants 0-5 months of age.

	Variable	Odds ratio	95% Confidence interval	*P*-value
Level 1	*Father not living with child*	2.0	1.1 - 3.6	0.03
	Income per capita			0.02
	1st quartile	—		
	2nd quartile	0.7	0.4 - 1.3	
	3rd quartile	0.9	0.5 - 1.6	
	4th quartile	0.4	0.2 - 0.7	
Level 3	*Infant's age (months)*			0.001
	0 - 1	—		
	2 - 3	3.2	1.7 - 6.1	
	4 - 5	2.4	1.2 - 4.7	
	No. previous pregnancies			0.06
	0	—		
	1	0.7	0.3 - 1.5	
	2-3	1.7	0.8 - 3.5	
	4+	2.2	0.9 - 5.7	
	Mother's age (years)			0.13
	14 - 19	—		
	20 - 24	0.6	0.3 - 1.1	
	25 - 29	0.4	0.2 - 1.0	
	30 - 45	0.4	0.1 - 0.9	
Level 4	*Birth weight (g)*			0.15
	<2500	1.9	0.9 - 4.3	
	2500 - 2999	1.5	0.9 - 2.5	
	3000+	—		
	Breastfeeding mode			<0.001
	Predominant/exclusive	—		
	Partial	2.8	1.4 - 5.8	
	None	4.7	2.4 - 9.1	

Table A3. Odds ratios of suboptimal breastfeeding and selected control variables for infant mortality due to lower respiratory infections among infants 6-11 months of age.

	Variable	Odds ratio	95% Confidence interval	*P*-value
Level 1	*Mother's education (years)*			0.13
	0	3.1	1.1 - 9.1	
	1 - 4	1.4	0.6 - 3.3	
	5+	—		
	Household head employed	0.5	0.2 - 1.4	0.20
	Infant's age (months)			0.04
	6 - 8	—		
	9 - 11	0.4	0.2 - 1.0	
Level 4	*Birth weight (g)*			0.01
	<2500	5.4	1.8 - 15.8	
	2500 - 2999	2.7	1.1 - 6.6	
	3000+	—		
	Low Apgar score at birth	0.2	0.0 - 1.9	0.10
	Non-breastfed	2.8	1.0 - 7.7	0.03

Table A4. Odds ratios of suboptimal breastfeeding and selected control variables for infant mortality due to diarrhoeal disease among infants 0-5 months of age.

	Variable	Odds ratio	95% Confidence interval	*P*-value
Level 1	*Mother's education (years)*			0.06
	0	2.7	0.9 - 8.3	
	1-4	2.7	1.0 - 7.2	
	5-8	1.7	0.6 - 4.6	
	9+	—		
	Mother's ethnicity			0.10
	White	—		
	Mixed race	1.4	0.8 - 2.4	
	Black	1.7	1.0 - 3.0	
	Infant's age (months)			<0.001
	0 - 1	—		
	2 - 3	2.9	1.6 - 5.2	
	4 - 5	2.7	1.4 - 5.1	
L 2	*Household has no flush toilet*	1.7	1.1 - 2.7	0.02
Level 3	*Birth interval (months)*			0.002
	First child	2.8	1.5 - 5.5	
	9 - 20	1.6	0.9 - 3.1	
	21 - 50	0.8	0.3 - 2.0	
	51+	—		
	Mother's age (years)			0.20
	14 - 19	—		
	20 - 24	1.5	0.8 - 2.9	
	25 - 29	0.8	0.4 - 1.7	
	30 - 45	1.0	0.4 - 2.1	
	Birth weight (g)			0.03
	<2500	2.7	1.3 - 5.8	
	2500 - 2999	1.4	0.8 - 2.5	
	3000+			
Level 4	*Birth not in hospital*	3.7	1.3 - 10.2	0.01
	Delivery by Caesarian section	0.6	0.2 - 1.3	0.10
	No. of antenatal care visits			0.02
	0	—		
	1 - 4	0.9	0.5 - 1.8	
	5 - 8	0.5	0.2 - 1.0	
	9+	0.4	0.2 - 0.9	
	Breastfeeding mode			
	Predominant/exclusive	—		<0.001
	Partial	2.4	1.1 - 5.2	
	None	7.3	3.7 - 14.4	

Table A5. Odds ratios of suboptimal breastfeeding and selected control variables for infant mortality due to diarrhoeal disease among infants 6-11 months of age.

	Variable	Odds ratio	95% Confidence interval	*P*-value
Level 1	*Income per capita*			0.11
	1st quartile	—		
	2nd quartile	0.7	0.3 - 1.7	
	3rd quartile	0.8	0.3 - 1.8	
	4th quartile	0.3	0.1 - 0.9	
	Male infant	1.5	0.8 - 2.9	0.19
	Infant's age(months)			<0.001
	6 - 8	—		
	9 - 11	0.2	0.1 - 0.5	
Level 2	*Household has piped water*	0.5	0.2 - 1.0	0.04
	No. persons per sleeping room			02
	1 - 2	—		
	3 - 5	2.5	0.7 - 9.1	
	6+	3.9	0.9 - 17.2	
Level 3	*No. previous pregnancies*			0.10
	0	—		
	1	4.2	1.4 - 12.4	
	2 - 3	1.4	0.4 - 4.3	
	4+	2.2	0.6 - 8.8	
	Mother´s age (years)			0.30
	14 - 19	—		
	20 - 24	0.4	0.2 - 1.2	
	25 - 29	0.5	0.1 - 1.5	
	30 - 45	0.2	0.0 - 0.8	
	Not breastfed	2.5	1.1 - 5.7	0.02

Chapter 8

Trends in caesarean section in developed countries: supply, demand and opportunities for control

Based on: Lauer JA, Betrán AP, Merialdi M, Wojdyla D. Trends in caesarean section in developed countries: supply, demand and opportunities for control (in submission).

Abstract

Objective

To study population-level determinants of caesarean section trends in developed countries.

Data sources/study setting

National-level data (1980-2004) from developed countries on trends and potential determinants of caesarean section utilization.

Study design

The effect of maternal mortality, national income, hospital infrastructure and the health system financing and human resources profile on caesarean section rates was analysed with a dynamic econometric model.

Data collection/extraction methods

Annual data on utilization and potential determinants of caesarean section were obtained from health statistical services and international organizations.

Principal findings

The capacity of the health system to deliver surgical obstetric care, its financing structure, and possibly also its human resources profile, have stronger aggregate-level effects on caesarean section rates than does income.

Conclusions

Health system factors are potentially important aggregate-level determinants of caesarean section utilization which have been overlooked in debates focusing on the impact of women's choices and doctors' preferences.

Introduction

Caesarean section rates are high and continue to rise in developed countries.[1][2][3] , However, the impact of guidelines and recommendations in curbing their growth has been limited: in 1985, representatives of a study group convened by the World Health Organization wrote, "there is no justification for any region to have caesarean section rates higher than 10–15%."[4] Although levels of 10–15% were considered high but acceptable at the time, average caesarean rates in most developed regions (with the exception of eastern Europe) now exceed 20%; the recommendation thus appears to have been largely overtaken by events.[5] Neverthless, little research exists on determinants of caesarean section utilization, at either the aggregate[6] or the individual level,[7][8] and, until recently,[9] the few randomised trials that have been published have found no effect, for the intervention studied, on rates of caesarean delivery.[10][11]

Understandably, in such a context, there is concern that apparently inexorably rising rates of caesarean delivery have the potential to divert human and financial resources from other, arguably higher priority, interventions.[12] Furthermore, the possibility that indiscriminate use of caesarean section can have a negative impact on maternal and neonatal health has been raised[13] and has recently received support from a number of studies.[3][14][15][16] On the other hand, it has been argued that reducing caesarean delivery rates would have a detrimental effect on mothers' and infants' health, and that patients' choices should be considered.[17]

Against this background, we set out to study trends in caesarean delivery rates in developed countries with the aim of identifying potential population-level determinants. Previous ecological research,[6] limited to cross-sectional analysis in Latin America, has suggested that it is primarily income that determines caesarean section rates at aggregate level. Here we use a cross-country dynamic regression model to exploit the additional information available from data on trends to present evidence that caesarean rates in developed countries respond not only to economic incentives such as income but also, and more strongly, to important modifiable health-system factors.

Methods

Our analysis focuses on developed countries as defined in the United Nations' classification.[18] Developed countries include countries in Europe and Australia, Canada, Japan, New Zealand and the United States of America. Trend data were collected for 38 of these countries since 1980. Variables used in the analysis, and their interpretation, are described in the Table.

Potential determinants of CS rates

Determinant	Reason for inclusion
Medical	
• Maternal mortality ratio*	Indicator of medical need
Non-medical	
Income	
• Gross domestic product per capita (GDP per capita)*	Indicator of consumer demand
Health-system infrastructure	
• Number of hospitals per capita#	Relatively fixed indicator of the capacity of the health system to deliver surgical obstetric procedures
• Number of hospital beds per capita#	More flexible indicator of the capacity of the health system to deliver surgical obstetric procedures
Health-system financing and human resources organization	
• Proportion of total health expenditure derived from government sources#	Indicator of the financial organization of the health system
• Number of midwives per capita#	Indicator of the human resources profile of the reproductive health system

*Data available for all 38 countries ("basic" data).
Data available for only 25 countries ("full" data).

For European countries, data were obtained from the European Health for All Database (www.who.dk), maintained by the WHO European Regional Office. For Australia, Canada, Japan, New Zealand, the United Kingdom and the United States, national health statistical services issue regular publications and maintain web pages with information on maternal health indicators.

Previous research on international comparisons of caesarean delivery rates has relied on cross-sectional data from Latin America and simple bivariate correlation models.[6] As a preliminary analysis, and to establish a baseline for comparison with the results from our trend analysis, we performed a simple linear regression of caesarean section rate versus income per head in Latin America and in developed countries, using a previously published cross-sectional data set, in order to confirm previous findings.[3]

Subsequently, trend data on caesarean section rates and their potential determinants (Table) were analysed using a dynamic econometric model

combining a standard time-series analysis with a simple panel-data model for cross-country variation.[19][20] (For full details of analysis methods, see Annex).

A basic set of indicators (maternal mortality ratio and income per head) was available for 38 countries (Table), with an average of over 11 years of observation per country. A larger set of indicators (consisting of the basic indicators plus estimates of hospitals, hospital beds and midwives per head as well as the proportion of total expenditure on health derived from government sources) was available for 25 countries, with an average of over 6 years of observation per country.

Results from both data sets (basic and full) using the dynamic econometric model are reported below. Since the 38 countries for which basic data were available might differ systematically from the 25 countries for which full data were available, a model with basic indicators only was studied for both sets of countries in order to establish that countries with full data were not qualitatively different than those with basic data only.

Results

Cross-sectional analysis

In the same set of developed countries as those studied in the dynamic model, a log-log linear regression of cross-sectional caesarean section rate versus cross-sectional income per head predicts that a doubling in income corresponds to a 33% (95% CI, 18% – 46%) increase in caesarean delivery rate (see Annex for details of the log-log regression). For comparison with previous research,[6] the same model estimated in Latin American countries only suggests that a doubling in income would be associated with a 77% (67% – 87%) increase.

Basic dynamic model

At only 4% (3% – 6%), the estimate of the effect of a doubling in income obtained from the basic dynamic model is substantially lower than the estimate obtained from cross-sectional analysis (Annex, Table 3). A small positive coefficient was associated with calendar year, indicating an average

increase of 0.3% (0.2% – 0.4%) per year after controlling for other variables in the basic model (Annex, Table 1). There was also a negative but insignificant association with maternal mortality (Annex, Table 1).

These results were qualitatively robust across all alternative model specifications using the basic indicators. Standard tests of model validity show that no estimation assumptions were unsatisfied.[21]

Full dynamic model

When the full set of indicators was analysed, a doubling in income per head was found to correspond to an increase in caesarean section rates of 6% (4% – 8%; Annex, Table 3). Notably, however, variables associated with the capacity of the health system to deliver surgical obstetric care were also found to have a significant positive effect on caesarean utilization rates: for example, a doubling in the stock of hospitals per head corresponded to a 15% (4% – 26%) increase in caesarean section rates (Annex, Table 3); a doubling in the number of hospital beds per head, however, was associated with approximately a 26.8% (12.2% – 41.4%) increase in the caesarean section rate (Annex, Table 2).

Moreover, the financial organization of the health system appeared to have an even stronger effect on caesarean utilization: a doubling in the share of health expenditure derived from government sources was found to correspond to a 29.8% (9.6% – 50%) decrease in caesarean rates (Annex, Table 2). Although the coefficient was not quite significant (Annex, Table 3), there was a suggestion that a doubling in the number of midwives per head would result in a 3% (–1% – 6%) increase in caesarean section rates, which is contrary to what might be expected. There was a small, although also not significant, increase in caesarean section rates associated with increased maternal mortality ratio (Annex, Table 2).

When the full model was analysed for long-run relationships (Annex, Table 4), the effect of health system financing was seen to be much larger, with a doubling of the share of health expenditure from public sources implying a 95% (42% – 149%) reduction in the caesarean section delivery rate in the long run. An effect of similar magnitude but in the opposite direction was observed for the number of hospital beds per head, which was associated with an 86% (45% – 126%) increase in the caesarean section rate in the long run.

Although the coefficient was not significant, the effect of doubling the number of hospitals per head was predicted to have only a 12% (–18% – 42%) increase in the caesarean section rate in the long run. Possibly because hospitals are a relatively fixed measure of health system infrastructure, the long-term effect of the hospital stock was found to be of about the same magnitude as that in the short run. Although still not significant, a doubling in the number of midwives per head was found to be associated with a 14% (–4% – 32%) decrease in the caesarean delivery rate in the long run; in other words, the effect of the number of midwives per head appeared to be different in the long run, and moreover consistent with expectations, as compared with the estimated effect in the short term. A switch in the direction of effect also obtained for income per head in the long run, since a doubling in income was found to imply a 5% (–3% – 14%) decrease in caesarean section rates in the long run (although the coefficient was not significant).

See Annex for further details on the estimation model, interpretation and results.

Discussion

The estimate of a 77% increase in the caesarean section rate for a doubling of income in the Latin American region that we obtained from the preliminary cross-sectional analysis is remarkably close to the value of the linear correlation reported for that region previously.[6] For developed countries, however, the cross-sectional relationship, estimated here at 32%, is substantially weaker. In any case, both previous ecological research[6] and our own preliminary cross-sectional analysis could be claimed to support the hypothesis that, at aggregate level, caesarean section rates respond strongly to income, or to factors that are themselves strongly associated with income. Results from cross-sectional analyses would therefore seem to raise the possibility of dramatically increasing caesarean section rates with rising incomes in the future.

Our principal new finding, however, is that the relationship of caesarean section rates with income is in fact substantially weaker when longitudinal rather than cross-sectional data are analysed. Since, in either the basic or the

full dynamic model, the effect of income on caesarean section rates is an order of magnitude weaker than that found in cross section, estimates derived from cross-sectional studies would appear to be biased. Such bias could be attributable to either the failure to control for the dynamic aspects of the relationship, as well as possibly also the absence of relevant control variables in the cross-sectional analyses. Our second important finding is that, when health system variables are included, a much richer picture of the population-level determinants of caesarean section rates emerges than that available from previous research.

The main strength of the study is its application of dynamic econometric models to health-care utilization trends in developed countries so as to explore competing hypotheses about aggregate-level determinants of caesarean section rates. Its main limitations are those inherent in the ecological nature of the data;[22] there are clearly individual-level factors affecting the utilization of caesarean section[9] which this study could only measure in the aggregate. A specific limitation of the econometric model is its assumption that different countries respond similarly to determinants of caesarean section utilization.[23] Finally, since several of the estimates for coefficients of interest were not significant, we cannot be certain that their reported value was not positive or negative due merely to chance.

The conventional model for growth in caesarean section rates implies that caesarean delivery is a conventional economic good, in the sense that the higher one's income the more one is inclined to "purchase" it. We call such a model "demand-driven". A demand-driven model is consistent with the hypothesis that it is primarily women's choices that determine caesarean section rates. Although a demand-driven model receives support from the results presented here, the size of the estimated effect is nevertheless much smaller than that previously reported.[6]

The observed effect of the number of hospitals and of hospital beds per head suggests that, in addition to demand, supply factors are also important. A supply-driven model would imply that, regardless of medical need, the greater the capacity of the health system to deliver surgical obstetric care, the more will be delivered. Such a model suggests that "suppliers" of ceasarean delivery

(e.g. obstetricians) have substantial influence on delivery mode, and contribute importantly to rising caesarean section rates. A "supply-driven" model also receives support from the data analysed here.

Nevertheless, health system factors such as the human-resources and financing profile are seen to have the largest impact on caesarean utilization rates. Health system financing, in both the short and the long run in fact, is the single factor among those studied here with the strongest impact on aggregate levels of caesarean delivery. This finding suggests the importance of a previously under-recognized model for determinants of caeasarean section, one related neither to supply or to demand factors but rather to the health system itself. Health system factors are largely institutional, in other words, related to the legal environment in which health-care decisions are made.

Overall, our results therefore suggest that, in the context of debates about whether patients' choices or doctors' preferences are more responsible for rising caesarean delivery rates,[24][25], health system factors may be an important overlooked population-level determinant. One obvious implication is that caesarean delivery rates might be amenable to control through policy instruments acting at the health system level. While it is acknowledged that such instruments would be likely to affect a broad range of other outcomes as well, these findings nevertheless suggest novel avenues for policy intervention and investigation into determinants of utilization of this important obstetrical procedure.

References

1. Dobson R. Caesarean section rate in England and Wales hits 21%. *BMJ* 2001; 323:951.

2. Hamilton BE, Martin JA, Sutton PD. Births: Preliminary data for 2003. National Vital Statistical Reports 2004; 53:1-18.

3. Betrán AP, Merialdi M, Lauer JA, Bing-shun W, Thomas J, Van Look P, et al. Rates of caesarean section: analysis of global, regional and national estimates. *Paediatr Perinatal Epidemiol* 2007;28:98-113.

4. World Health Organization. Appropriate technology for birth. *Lancet* 1985;2:436-437.

5. Lauer JA, Betrán AP. Decision aids for women with a previous caesarean section: focusing on women's preferences improves decision making. *BMJ* 2007; 334:1281-1282.

6. Belizán JM, Althabe F, Barros FC, Alexander S. Rates and implications of caesarean sections in Latin America: ecological study. *BMJ* 1999; 319:1397-1400.

7. Najmi RS, Rehan N. Prevalence and determinants of caesarean section in a teaching hospital of Pakistan. *J Obstet Gynaecol* 2000;20:479-483.

8. Lei H, Wen SW, Walker M. Determinants of caesarean delivery among women hospitalized for childbirth in a remote population in China. *J Obstet Gynaecol Can* 2003;25:937-943.

9. Montgomery AA, Emmett CL, Fahey T, Jones C, Ricketts I, Patel RR, et al. Two decision aids for mode of delivery among women with previous caesarean section: randomised controlled trial. *BMJ* 2007; 334:1305-1309.

10. Shorten A, Shorten B, Keogh J, West S, Morris J. Making choices for childbirth: a randomized controlled trial of a decision-aid for informed birth after caesarean. *Birth* 2005;32:252-261.

11. Fraser W, Maunsell E, Hodnett E, Moutquin J-M. Randomized controlled trial of a prenatal vaginal birth after cesarean section education and support program. *Am J Obstet Gynecol* 1997;176:419-425.

12. Wagner M. Fish can't see water: the need to humanize birth. *Int J Gynaecol Obstet* 2001; 75 Suppl 1: S25-S37.

13. Wagner M. Choosing caesarean section. *Lancet* 2000; 356:1677-1680.

14. Thomas J, Paranjothy S; Royal College of Obstetricians and Gynaecologists Clinical Effectiveness Support Unit. *The national sentinel caesarean section audit report*. London: RCOG Press, 2001.

15. Villar J, Valladares E, Wojdyla D, Zavaleta N, Carroli G, Velazco A, et al. Caesarean delivery rates and pregnancy outcomes: the 2005 WHO global survey on maternal and perinatal health in Latin America. *Lancet* 2006;367:1819-1829.

16. Hall MH, Bewley S. Maternal mortality and mode of delivery. *Lancet* 1999; 354:776.

17. Sachs BP, Castro MA. The risk of lowering cesarean-delivery rate. *N Engl J Med* 2003; 340:54-57.

18. United Nations. *World Population Prospects: the 2004 revision*. New York: United Nations, 2004.

19. Dielman TE. *Pooled cross-sectional and time series data analysis*. New York, United States: Marcel Dekker, 1989.

20. Bond SR. Dynamic panel data models: a guide to micro data methods and practice. *Portuguese economic journal* 2002; 1:141-162.

21. Arellano M, Bond S. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of economic studies* 1991; 58:277-297.

22. Rothman KJ, Greenland S. *Modern epidemiology*. Philadelphia, PA: Lippincott-Raven Publishers, 1998.

23. Pesaran MH, Smith R. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics* 1995; 68:79-113.

24. Showalter E, Griffin A. Commentary: all women should have a choice. *BMJ* 2001;319:1401.

25. Castro A. Commentary: increase in caesarean section may reflect medical control not women's choice. *BMJ* 2001;319:1401-1402.

Annex: estimation methods, interpretation and detailed results

Characteristics of the estimation model

A time-series regression model using lagged dependent variables is typically used to analyse complex systems whose time-dependent outcome (here, caesarean section rate) is held to depend on numerous intermediate outcomes, some of which may not be directly observed. In such an approach, the effect of the intermediate outcomes is controlled for by using lagged values of the dependent variable as instruments (i.e. variables with many of the same statistical properties as the variables that cannot be observed). The effect of other, directly observed, covariates in the regression equation can then be inferred from their coefficients in the standard way.

Since we wish however to study relationships between caesarean section rates and their determinants across a group of countries with similar characteristics, a standard time-series model is inadequate.[1] We therefore use a so-called "dynamic panel" model, which combines the time series (i.e. "dynamic") approach with a simple model for inter-country (i.e. "panel") variation.[2][3]

A requirement of the particular dynamic panel model used here is that inter-country variation is measurable in terms of a single parameter; the model requires, in other words, the assumption that slope coefficients are homogeneous across countries and that inter-country differences are expressable by means of country-specific intercepts.[4]

Theoretical basis of the model

On the hypothesis that caesarean delivery is a conventional economic good, it is expected that caesarean rates will respond positively (i.e. in direct relationship) to income per capita as a result of increased patient demand. On the alternative hypothesis that doctors' preferences are an important determinant of caesarean section rates (i.e. the hypothesis that, other things being equal, obstetricians prefer patients to have caesarean deliveries), a direct relationship with caesarean rates is posited for measures of the capacity of

the health system to deliver surgical obstetric care (for example, the stock of hospitals, or of hospital beds, per capita). On the still different hypothesis that health system factors, such as the human resources profile of obstetric care and the organization of health system financing, are important determinants of caesarean section rates, it is expected that caesarean utilization would respond negatively (i.e. in inverse relationship) to the number of midwives per capita or to the proportion of total expenditure on health derived from government sources. Finally, since it has been shown to be a covariate of caesarean section rates in cross-sectional analysis,[5] we include the maternal mortality ratio as a control variable representing medical need.

With the exception of calendar year, all variables are entered as the logarithm.

The two model specifications that we report (i.e. "basic" and "full", described below) are robust in the sense that, except as noted immediately following, in any alternative model specification tested, the coefficients of all variables included in the reported model specifications were strongly significant; conversely, the coefficients of none of the variables not included in the reported model specifications were significant in any of the alternative model specifications. The sole exceptions were that the coefficients for maternal mortality ratio and for the number of midwives per capita in the current period, *t*, were not significant in most of the model specifications tested; nevertheless, these variables were retained in the corresponding reported model specifications on account of their interest as potential determinants of caesarean section rates. The inclusion or exclusion of these variables in any case had no important effect on the coefficient values of any of the other variables reported.

The basic model, ignoring error terms, can be written:

$$CS_t = c_1 \cdot CS_{t-1} + c_2 \cdot CS_{t-2} + c_3 \cdot GDP_t + c_4 \cdot GDP_{t-1} + c_5 \cdot MMR_t + c_6 \cdot Year_t.$$

The full model, ignoring error terms, can be written:

$$CS_t = c_1 \cdot CS_{t-1} + c_2 \cdot CS_{t-2} + c_3 \cdot GDP_t + c_4 \cdot GDP_{t-1} + c_5 \cdot MMR_t + c_6 \cdot Year_t + c_7 \cdot Hosp_t + c_8 \cdot Hosp_{t-1} + c_9 \cdot HospBed_t + c_{10} \cdot Midw_t + c_{11} \cdot Midw_{t-1} + c_{12} \cdot PubHExp_t.$$

A list of variables and their abbreviations follow:

Caesarean section rate in the current year (t)	CS_t
Caesarean section rate in the previous year ($t-1$)	CS_{t-1}
Caesarean section rate two years previoius to the current year ($t-2$)	CS_{t-2}
Income per head in the current year (t)	GDP_t
Income per head in the previous year ($t-1$)	GDP_{t-1}
Maternal mortality ratio in the current year (t)	MMR_t
Number of hospitals per head in the current year (t)	$Hosp_t$
Number of hospitals per head in the previous year ($t-1$)	$Hosp_{t-1}$
Number of hospital beds per head in the current year (t)	$HospBed_t$
Number of midwives per head in the current year (t)	$Midw_t$
Number of midwives per head in the previous year ($t-1$)	$Midw_{t-1}$
Proportion of total expenditure on health from government sources in the current year (t)	$PubHExp_t$
Calendar year (t)	Year.

Raw coefficient estimates for the variables are reported below in Annex Table 1 (basic model) and Annex Table 2 (full model).

The dynamic model is estimated using the Stata (version 9) procedure xtabond.

Model and parameter interpretation

Log-log regression

Regression with logarithmically transformed dependent (i.e. left-hand side) and independent (i.e. right-hand side) variables yield coefficients that can be interpreted as *elasticities*, which means that the coefficient value gives the proportional change in the dependent variable (i.e. caesarean section rate) associated with a doubling in the independent variable (i.e. potential determinant).

Changes rather than levels

In the *Results* section of the main text, it is claimed that "changes" in income or other variables are a potential determinant of caesarean section rates. This is a common interpretation applied to dynamic econometric models of the type used here, and it relies on the following argument.

Taking income as an example, the relevant part of the regression equation can be written as follows:

$$CS_t = \ldots + c_3 \cdot GDP_t + c_4 \cdot GDP_{t-1} + \ldots.$$

If the estimated value of the coefficient c_3 is approximately equal in absolute value to the estimated value of the coefficient c_4, and if, in addition, c_4 is negative in sign (i.e. $c_3 > 0$ and $c_3 = -c_4$), we can then write the regression equation as, with approximate equality:

$$CS_t \approx \ldots + c_3 \cdot GDP_t - c_3 \cdot GDP_{t-1} + \ldots.$$

The following rearrangement and relabelling is then possible:

$$CS_t \approx \ldots + c_3 \cdot (GDP_t - GDP_{t-1}) + \ldots \Rightarrow$$

$$CS_t \approx \ldots + c_3 \cdot \Delta GDP + \ldots,$$

where Δx is used as an abbreviation for $x_t - x_{t-1}$ (i.e. "changes in x"). When the above conditions hold, this argument gives rise to the interpretation that it is changes in, rather than levels of a variable (e.g. income) that determine caesarean section rates.[6]

This interpretation can in fact be legitimately applied whenever the confidence intervals of the coefficients, for example c_3 and c_4, are substantially overlapping (i.e. the coefficients are equal within statistical error). Since, in practice, the coefficients will rarely be exactly equal, their absolute values can be averaged to produce a statistically more robust estimate of the impact of changes in the independent variable on the dependent variable.

Averaging the coefficients is equivalent to defining a new coefficient based on a combination of the original ones:

$$c_\Delta \equiv (c_3 - c_4)/2.$$

The transformed coefficient c_Δ is then used to infer the effect of changes in the independent variable.

This is the procedure used to interpret the effect of income, the number of hospitals per capita and the number of midwives per capita on caesarean section rates. The transformed coefficients, and their approximate confidence intervals, corresponding to these variables are reported below in Annex Table 3.

Long run

In *Results*, so-called "long run" relationships are reported for certain variables. This is another standard interpretation for econometric models of this type.

This interpretation is motivated by the argument that, in the long run, the values of variables at different time periods (i.e. at t, $t-1$ and so forth) will become equal. In other words, the dynamic system estimated by the model is assumed to reach equilibrium in the long run in the sense that there are no further changes in the values of any of the variables.

However, if the use of the term "long run" seems objectionable for any reason, one can alternatively think of the long run relationships as equilibrium relationships. They are, in other words, the relationships that would obtain if and when the system estimated by the model reaches a steady state. The long run coefficients thus show the intrinsic response (equilibrium elasticity) of the estimated dynamic system to exogeneous shocks in any of the variables.

If we are interested in the long-run relationship between caesarean section and income, for example, we suppress the corresponding subscripts for period (since the variables have reached steady state and are therefore equal in all periods). The relevant part of the regression equation can then be written:

$$CS_t = c_1 \cdot CS_{t-1} + c_2 \cdot CS_{t-2} + c_3 \cdot GDP_t + c_4 \cdot GDP_{t-1} + \ldots \Rightarrow$$

$$CS = c_1 \cdot CS + c_2 \cdot CS + c_3 \cdot GDP + c_4 \cdot GDP + \ldots.$$

The common terms for caesarean section and income are then collected together and their coefficients added:

$$(1 - c_1 - c_2) \cdot CS = (c_3 + c_4) \cdot GDP + \ldots.$$

The long-run relationship of caesarean section and income can then be calculated by solving the above equation for CS:[6]

$$CS = (c_3 + c_4)/(1 - c_1 - c_2)\cdot GDP + \ldots.$$

This defines a new, long run ("LR") coefficient for the independent variable in terms of the estimated coefficients; it expresses the proportional response of the caesarean section rate to exogeneous shocks in income at steady state:

$$c_{LR} \equiv (c_3 + c_4)/(1 - c_1 - c_2)\ .$$

We have calculated long run elasticities of caesarean section rates for income per head, hospitals per head, hospital beds per head and the number of midwives per head. The transformed coefficients, and their approximate confidence intervals, for the long run elasticities for these variables are reported below in Annex Table 4.

Uncertainty estimates

Approximate 95% confidence intervals are reported in the annex tables for both raw and transformed coefficient estimates. The estimates for the raw coefficients are derived directly from the standard errors of the regression coefficients in the conventional manner. The confidence intervals reported for the transformed coefficients, however, are derived using a statistical technique called the "delta method" that relies on a linear-order series expansion of the equation for the transformed coefficient. The variance of the linear-order expansion of the coefficient equation can then be calculated by using standard statistical identities. This approach takes into account the main (i.e. first-order) effect of the variance and covariance of the raw coefficients (Annex Table 5).

Annex tables

Annex Table 1: Raw coefficients estimates for the basic model[a]

Variable	Coefficient	*P* value	95% Confidence Interval	
CS[b] (*t*−1)	0.646	0.000	0.557	0.735
CS (*t*−2)	0.178	0.000	0.091	0.265
GDP[c]	0.040	0.000	0.020	0.060
GDP (*t*−1)	−0.045	0.000	−0.062	−0.029
MMR[d]	−0.013	0.373	−0.043	0.016
Year	0.003	0.000	0.002	0.004

[a] Sargan test of over-identifying restrictions: $\chi^2(479) = 404.7$; Probability > $\chi^2 = 0.994$. Arellano-Bond test that average autocorrelation in residuals of order 1 is 0: H_0: no autocorrelation; $z = -14.00$, Probability > $z = 0.000$. Arellano-Bond test that average autocorrelation in residuals of order 2 is 0: H_0: no autocorrelation; $z = 0.580$, Probability > $z = 0.562$.
[b] CS = caesarean section rate.
[c] GDP = gross domestic product per capita (income per head).
[d] MMR = maternal mortality ratio.

Annex Table 2: Raw coefficients estimates for the full model[a]

Variable	Coefficient	*P* value	95% Confidence Interval	
CS[b] (*t*−1)	0.430	0.000	0.318	0.542
CS (*t*−2)	0.257	0.000	0.150	0.363
GDP[c]	0.052	0.000	0.022	0.081
GDP (*t*−1)	−0.069	0.000	−0.094	−0.049
MMR[d]	0.011	0.569	−0.027	0.049
Hospitals per capita	0.167	0.007	0.045	0.289
Hospitals per capita (*t*−1)	−0.130	0.024	−0.242	−0.017
Hospital beds per capita	0.268	0.000	0.122	0.414
Midwives per capita	0.006	0.805	−0.045	0.058
Midwives per capita (*t*−1)	−0.051	0.011	−0.091	−0.012
Share of health expenditure from government sources	−0.298	0.004	−0.500	−0.096
Year	0.007	0.000	0.005	0.010

[a] Sargan test of over-identifying restrictions: $\chi^2(479) = 198.43$; Probability > $\chi^2 = 1.000$. Arellano-Bond test that average autocorrelation in residuals of order 1 is 0: H_0: no autocorrelation $z = -4.72$, Probability > $z = 0.0000$. Arellano-Bond test that average autocorrelation in residuals of order 2 is 0: H_0: no autocorrelation $z = 0.390$, Probability > $z = 0.698$.
[b] CS = caesarean section rate.
[c] GDP = gross domestic product per capita (income per head).
[d] MMR = maternal mortality ratio.

Annex Table 3: Transformed regression coefficients for changes in independent variables

ΔVariable (model)	Equation for transformed coefficient	Numerical value	Approximate 95% Confidence Interval	
ΔGDP[a] (basic)	$(c_3 - c_4)/2$	0.04	0.03	0.06
ΔGDP[a] (full)	$(c_3 - c_4)/2$	0.06	0.04	0.08
ΔHospitals per capita (full)	$(c_7 - c_8)/2$	0.15	0.04	0.26
ΔMidwives per capita (full)	$(c_{10} - c_{11})/2$	0.03	−0.01	0.06

[a] GDP = gross domestic product per capita (income per head).

Annex Table 4: Transformed regression coefficients for the long-run elasticities of the independent variables

Variable (model)	Equation for transformed coefficient	Numerical value	Approximate 95% Confidence Interval	
GDP[a] (basic)	$(c_3 + c_4)/(1 - c_1 - c_2)$	−0.02	−0.11	0.07
GDP[a] (full)	$(c_3 + c_4)/(1 - c_1 - c_2)$	−0.05	−0.14	0.03
Hospitals per capita (full)	$(c_7 + c_8)/(1 - c_1 - c_2)$	0.12	−0.18	0.42
Hospital beds per capita (full)	$c_9/(1 - c_1 - c_2)$	0.86	0.45	1.26
Midwives per capita (full)	$(c_{10} + c_{11})/(1 - c_1 - c_2)$	−0.14	−0.32	0.04
Share of health expenditure from government sources (full)	$c_{12}/(1 - c_1 - c_2)$	−0.95	−1.49	−0.42

[a] GDP = gross domestic product per capita (income per head).

Annex Table 5: Approximate variance–covariance estimates for the raw coefficients in the full model

Variable	CS_{t-1}	CS_{t-2}	GDP_t	GDP_{t-1}	MMR_t	$Hosp_t$	$Hosp_{t-1}$	$HospBed_t$	$Midw_t$	$Midw_{t-1}$	$PubHExp_t$	Year
CS_{t-1}	0.003											
CS_{t-2}	−0.002	0.003										
GDP_t	−0.0001	0.00002	0.0002									
GDP_{t-1}	0.00002	0.0001	0.0001	0.0002								
MMR_t	0.00009	0.0004	−0.00003	0.00002	0.0004							
$Hosp_t$	0.001	0.001	0.0001	0.0001	0.00002	0.004						
$Hosp_{t-1}$	−0.001	−0.0002	0.0001	−0.00006	0.00005	−0.002	0.003					
$HospBed_t$	−0.001	−0.001	0.0002	−0.00001	0.00006	−0.002	0.0003	0.006				
$Midw_t$	0.0002	0.0002	−0.00003	−0.00003	−0.0002	0.0001	−0.0003	−0.001	0.001			
$Midw_{t-1}$	0.0001	0.0001	0.00001	−0.00003	−0.0006	0.0002	−0.0003	−0.0003	−0.00005	0.0004		
$PubHExp_t$	0.002	0.002	0.00008	0.00008	−0.0001	0.001	0.00003	−0.003	0.00001	0.00009	0.011	
Year	−0.00003	−0.00002	~0	~0	~0	−0.00002	0.00002	0.0001	−0.00002	0.00001	0.00003	~0

Annex references

1. Hsiao C. *Analysis of panel data*. Cambridge, England: Cambridge University Press, 1986.

2. Dielman TE. *Pooled cross-sectional and time series data analysis*. New York, United States: Marcel Dekker, 1989.

3. Bond SR. Dynamic panel data models: a guide to micro data methods and practice. *Portuguese economic journal* 2002; 1:141-162.

4. Pesaran MH, Smith R. Estimating long-run relationships from dynamic heterogeneous panels. *Journal of econometrics* 1995; 68:79-113.

5. Betrán AP, Merialdi M, Lauer JA, Bing-shun W, Thomas J, Van Look P, et al. Rates of caesarean section: analysis of global, regional and national estimates. Paediatr Perinatal Epidemiol 2007;28:98-113.

6. Arellano M, Bond S. Some tests of specification for panel data: Monte Carlo evidence and an application to employment equations. *Review of economic studies* 1991; 58:277-297.

Chapter 9

Discussion and conclusions

Discussion

Restatement of the general research questions

The general questions developed in the Introduction are:

— *Should we care about the longitudinal (i.e. time) dimension of population health?* In order to answer this question, we attempted to show why longitudinal population dynamics are an important concern for the measurement of population health.

— *Do longitudinal measures allow us to answer questions about population health better than other measures?* There, we focused on how we can apply longitudinal models to measure population health benefits and risks.

— *Are standard models from other scientific fields relevant to questions of population health measurement?* For the last question, we applied common models from epidemiology and econometrics to questions in population health measurement and related areas.

We proposed to illustrate the general questions through several more specific questions. Based on our research, we now propose answers to those questions.

Is it possible to develop an easy-to-use, reasonably generic, longitudinal model that can be used for population health measurement with generally available epidemiologic and demographic data?

As shown in Chapter 2, it is possible to develop a generic longitudinal model that can be used for population health measurement. The model discussed is a multi-state dynamic population life table ("PopMod"). Although it might be claimed that it is easy to use, with a few exceptions, it has been used only by trained specialists employed by institutions with an established interest in longitudinal population health measurement (e.g. The World Health Organization, Erasmus University, Harvard University). While it is possible to apply the model using generally available epidemiological and demographic data, these data typically require transformation, and in some cases preliminary modelling, before the required input parameters can be estimated.

It appears that there is a tension in generic models between generality and realism: it will come as no surprise to researchers in the field that the more realistic the model, the less generic it is. Here, we propose as a working definition of a realistic model one that has an acceptable face validity for a specialist familiar with the modelling question. Realism is a principal aim motivating the modelling of independent age and time dimensions. Generality is important for the comparability of results, as well as to promote ease of use by those who may not be specialists in a given subject area. We outline following how these divergent aims are addressed in PopMod.

Generality and realism

The model presented in Chapter 2 allows for four health states to be represented. Notionally these states are: (1) a healthy, or susceptible, state, (2) an index disease state, (3) a second disease state and (4) the state of having both diseases. Each state is subject to a specific profile of mortality and other rates (see Figure 7, Chapter 1). However, since the labels on the states are arbitrary, they can be interpreted in any other way consistent with this disease process (including interpretations in which some transition rates are set to zero). The states have been used in various applications to represent either stages in a disease process or risk factors predisposing to disease.

However, those wishing to model a multi-stage disease, such as for some cancers or chronic diseases, have found that four states are occasionally insufficient to satisfy the criterion of realism as defined above. This problem arises because such diseases can involve an arbitrary number of clinically and epidemiologically distinct stages accessed in a progressive, usually non-reversible, fashion. Representation of such diseases in PopMod can only be done by the compression of the desired number of states into four or fewer. Similarly, analysts interested in modelling multiple interacting risk factors and diseases (such as for physical activity, diet, obesity and related health outcomes), have found a limiting factor in the number of possible interactions between modelled health states. Moreover, although it has not been a typical complaint, the problem of modelling continuous risk-factor distributions is completely unaddressed by a model like PopMod. We return in more detail to the problem of modelling interacting disease and risk-factor states in the appendix *Modelling complex interactions in longitudinal population models*.

On the other hand, a number of analysts have found that it was sufficient to represent the disease process of interest with only a single state, resulting in frequent use of a two-state version of the model. Examples have been road traffic accidents, heavy alcohol use, blindness, deafness and mental health conditions like schizophrenia and depression. Although this approach has not been based on a claim that the conditions are not complex, progressive or multifactorial in causality, it highlights the fact that many analysts have in practice been able to capture what they considered to be the essential features of their problem without the need for comorbidity, disease staging or interactions.

Based on this experience, it appears there is no reason why a partition of the population into four health states should have any special claim to be considered the right level of complexity to adopt for a generic disease model. Although four states may be suited to the original conception of an index, comorbid and joint condition, our experience suggests that the questions that can be addressed with such a model may not be truly generic concerns but rather specific to the context of two causes of death that are strongly interacting, such as ischaemic heart disease and stroke.

Interest from the public

Interest in longitudinal population health modelling has come in the form of regular requests for download of the PopMod model. Most requests are generated by the WHO-CHOICE web page (www.who.int/choice), where the availability of the program is advertised. Since we have maintained continuous records, we have recorded about 75 download requests per year. However, we have received relatively few follow-up questions, at a rate only slightly higher than 1 every 4 weeks. In addition, most such requests have concerned administrative aspects of the download process. As only rarely have we received technical queries, the most likely conclusion is that potential users of PopMod are not becoming actual users.

Although this is in some sense a failure, it is doubtful that it represents a failure for the cause of longitudinal modelling per se. Likewise, although life expectancy is a widely cited concept, the ordinary period life table has never been widely used by non-specialists. Similarly, although the concept of "burden of disease" has thoroughly penetrated current discourse on public

health, only a fraction of the community has ever been engaged in actual burden of disease estimation. Thus, although ease of use remains a barrier to the diffusion of longitudinal modelling, it would not seem to be an absolute one.

Data needs

On the other hand, the need to transform observed data to estimate the input parameters required for generic longitudinal population models may pose a more critical barrier. Since data come in many forms, only a few of which are amenable to being used directly as input parameters, statistical and other problem-solving skills are usually required in order to estimate the parameters values required. Moreover, the cumulative impact of the assumptions required for parameter estimation can undermine the objective of realism sought by use of a multi-state longitudinal population model. It can also be argued that the use of ad-hoc assumptions in estimating parameter values limits a generic model's claim to be generic.

Is it feasible to apply a generic, longitudinal model to quantify real population health risks and what sort of insights can be obtained through its application?

As demonstrated in Chapter 3 and Chapter 4, quantifying population health risks and benefits with a generic population model is feasible. In Chapter 3, we applied the model to an analysis of interventions for cardiovascular disease, a major cause of non-communicable disease burden; in Chapter 4 we analysed interventions for malaria, a major cause of infectious disease burden. The corresponding research reports have been prominently published; the paper on interventions for cardiovascular disease has in addition been widely cited. This suggests that the application of generic longitudinal population models to real public-health problems can be interesting, plausible and of potential policy relevance.

Both of these research examples concern studies of cost effectiveness, rather than population health modelling per se. As noted earlier, our choice of method of population health measurement was guided in part by the need to find answers to questions about the impact on health of interventions. In other words, an important motivation for the measurement approach was to

quantify the impact of changes from one, usually currently observed, state to another, usually hypothetical, state. This is called *counterfactual analysis* (Pearl, 2000).

Cost-effectiveness analysis

Cost-effectiveness analysis provides a rational basis for priority setting, and requires measuring effectiveness. The advantage of measures of effectiveness obtained from longitudinal population models is that the age, period and cohort position of health benefits are accounted for and can be tabulated when required. Failure to account for these classification variables can lead to inconsistencies in health measurement.

One example is that statements such as "*intervention A averted* x-*thousand deaths*" is, without further information, imprecise. To make the statement somewhat more precise, an explicit time period must be specified. Often, the intended time horizon is one year, i.e. "*intervention A, if implemented now, would avert* x-*thousand deaths in the following year*". This interpretation is reasonable because a one-year time period is held to be meaningful in the case of a stationary population. Thus, the implicit use of a one-year time period in health measurement (as, for example, with ordinary life-table analysis) is equivalent to an implicit assumption that the population is stationary (Preston, 1993).

For a non-stationary population, a one-year time horizon is of limited relevance. To adequately distinguish between interventions in such a population, the age and the time position of the averted deaths must be precisely specified. Moreover, since interventions with intertemporal impact induce non-stationary population effects, to distinguish between such actions it can also necessary to identify the cohort position of the events. For these reasons, a longitudinal model allows for a more meaningful measure of intervention benefit than that obtained with ordinary life table analysis.

Prediction versus projection

The time dimension of a longitudinal population model implicitly raises the question of whether it should be considered to represent historical time. In other words, does a longitudinal population simulation represent a *prediction* about the future course of the population? While there is no logical problem with interpreting a population simulation as a prediction, in most cases it will

be better to avoid this view. The strongest reason is perhaps simply the fact that predictions are always wrong, and they are usually wrong even according to approximate criteria. Although making statements about the future of a population is logically possible, doing so therefore tends to make population modelling non-scientific. Population models nevertheless arguably display an interest in the future, or more properly stated, in a hypothetical (often counterfactual) state of population health. The hypothetical state typically represents either a desired or a deprecated future state of population health.

For example, a simulation based on current parameter values could be said to represent the future the population would face if it continued to experience currently estimated parameters, all other factors being held constant. If it is not a prediction, a simulation based on the current scenario is still of interest if current parameter values represent factors relevant for the question at hand. Current values might be relevant because, for example, they represent the parameters that medical theory claims are relevant for health. They might also be of interest because they are the parameters that most easily modifiable through intervention. Although there are many reasons a given set of parameters can be relevant, when applied to a modelling question, these reasons take the form of an hypothesis: *if these parameters are relevant, then the simulation is of interest*. Notice that the hypothesis does not necessarily claim that the current parameters are solely those that determine the future population health.

To distinguish simulations from predictions, they are occasionally referred to as *projections* (Keyfitz and Caswell, 2005), the usual term employed in the field of matrix population models (Caswell, 2001). A projection is distinguished from a prediction because its time dimension is mainly a mathematical construct used to represent the desired object of analysis. A longitudinal population model, in other words, uses a time dimension for purposes of measurement, and not for the sake of making statements about the future.

Comparability and generality

Cost-effectiveness analysis using longitudinal population models and using standard life tables can yield the same ordinal ranking of interventions. This is not surprising because measurements made using the same model have the same implicit time horizon. Moreover, exceptions to invariance in ordinal ranking can often be explained by factors unrelated to the choice of model. One might therefore

decide that answers to questions about cost-effectiveness analysis are insensitive to the choice of model. Is there then any advantage to generic longitudinal population modelling for the purpose of cost-effectiveness analysis?

Although different models may yield similar rankings, comparisons across models nevertheless remain problematic. One advantage of cost-effectiveness analysis using generic longitudinal population models may therefore be unrelated, strictly speaking, to the longitudinal aspects of the model but rather to the comparability resulting from use of a generic model. This kind of advantage can be referred to as that of standardization: since analyses performed with a similar set of methods can be mutually compared, standardization facilitates, for example, sector-wide cost-effectiveness analysis, where interventions with many different kinds of effect (and costs) must be mutually compared.

Relevance for policy

The more explicit are the assumptions on which they are based, and the more comparable the health measurements, the more relevant they are for policy making. Although relevance for policy making is different than use in policy making, discussing the reasons for this distinction goes beyond the scope of the questions we have investigated here. Work in the context of WHO-CHOICE has nevertheless resulted in a few experiences where it is possible to trace a direct connection between cost-effectiveness analysis and subsequent policy making.

Most prominent among these examples is probably our experience with the Estonian Ministry of Social Affairs. The ministry sought our assistance for the use of cost-effectiveness analysis of interventions designed to decrease alcohol and tobacco use. A revised policy on excise taxes was subsequently implemented. The ministry continues to work with WHO on applications of cost-effectiveness analysis to national health policy.

Does the logistic family of models offer an approach relevant to longitudinal population health measurement?

We have shown that models of the logistic family can be interpreted longitudinally in two types of case: first, when the model is an explicitly

longitudinal model describing exponential or logistic population growth or disease transmission, or when the model describes an age-dependent set of outcomes that can be interpreted as time dependent under the assumption of a stationary population.

In the context of population health measurement, the first case is mainly relevant to phenomena of contagion or growth. In Chapter 5, we show how individual-level random processes, which are an idealization of infection transmission or the spread of a population, are relevant to the population-level description of growth or of an epidemic. We also argue there that the same mechanism makes the second case (implicit logistic models) relevant to the estimation of population processes measured in cross section.

In both cases, the applications concern those where the term "logistic model" refers to a distribution function for a random variable with a logistic cumulative probability distribution. Since this distribution arises naturally as the limit of a discrete random process, for applications where time or age is a conditioning variable, the mechanics of the discrete process lend substantive meaning to the use of its limiting description in modelling or estimation.

In particular, contagious outcomes (such as infection, drug addiction and product preferences), may follow a logistic regression on the time predictor as a consequence of underlying diffusion-like phenomena. When considering additional covariates, logit-additive approximations to varying temporal effects may sometimes be expected due to similar mechanisms. Furthermore, if the covariate levels define separate diffusion compartments within which there is logit-linearity in time, the overall model will be logit linear with time-covariate product terms ("interactions").

Is it feasible to apply logistic models to quantify real population health risks and what sort of insights can be obtained through their application?

Application of logistic models to population health measurement

In Chapter 6 and Chapter 7, we present an application of logistic modelling to an important child-health risk factor, suboptimal breastfeeding. Although

suboptimal breastfeeding remains a significant global concern (Black RE et al., 2008), exposure measurement has been hampered by the poor availability of population-representative breastfeeding data. In the work presented in Chapter 6, we show that the logit-linear age structure observed in breastfeeding data from countries with comprehensive nationally representative surveys can facilitate the use of partial data from other countries.

Logistic regression of breastfeeding prevalence on age yields a simple two-parameter (slope and intercept) logit function for breastfeeding time-to-failure rates. The slope is interpretable as a constant per-capita risk of stopping breastfeeding and the intercept as the initial breastfeeding odds at age zero. With this model it is possible to emply a simple fitting procedure to impute missing data in a target country. The procedure consists of fixing the slope as found in a country with similar characteristics to the target country, but where there is a comprehensive survey, and then selecting the intercept so as to fit the limited data available from the target country. Once values for both parameters are obtained in this way, the entire age-specific breastfeeding schedule can be solved for in the target country by using the inverse logit transformation, thereby extending the geographic range of coverage with breastfeeding indicators.

In Chapter 7, we apply the exposure estimates from Chapter 6, together with results from a re-analysis of a previously published case-control study on the mortality risk of suboptimal breastfeeding, to develop estimates of the population attributable fraction of fatal outcomes in children. In other words, relying on the stationary population assumption, estimates of exposure and risk, respectively from two cross-sectional studies, are used to estimate mortality burden in a longitudinal framework.

A counter example

The application of cross-sectional methods to fundamentally longitudinal problems in population health measurement is nevertheless not advisable in the absence of supporting evidence, or when evidence in favour of the stationary population interpretation is weak. We describe an example of such a case in Chapter 8, where an econometric longitudinal model, called the "dynamic panel" model, was applied to the analysis of trends in caesarean section utilization rates at population level.

The study was motivated by the fact that rates of caesarean section utilization continue to rise amid concerns that they represent a diversion of resources from activities with greater health benefit, and that unnecessary caesarean sections can have a negative impact on mothers' and children's health.

One of the principal aims was to measure the relationship between caesarean section rates and income, which was motivated in part by our previous work (Betrán et al., 2007) in which income was suggested to have a large impact on caesarean section rates. When the trend data were analysed dynamically, however, the relationship of caesarean section rates with income was found to be substantially weaker, implying that the estimates we had derived with cross-sectional methods were biased.

Another finding was that, when health system infrastructure, human resources and financing profile are controlled for, a richer picture of potential population-level determinants of caesarean section rates emerges. Although debate has tended to center around either a "women's choice" or a "doctors' preference" model to explain determination of delivery method (Castro, 2001; Showalter and Griffin, 2001), we found that the largest effect on caesarean section utilization at population level was associated with aspects of the health system infrastructure and its human resources and financing profile. The use of longitudinal methods was necessary for making these distinctions.

Priorities for further research

Uncertainty analysis in longitudinal population models

The fact that indirect modelling and estimation methods are often required to estimate the parameters used by multi-state longitudinal population life tables implies that such models use estimated inputs to produce yet further estimates as outputs. The outputs therefore convolve input uncertainty with the effect of model-specific processes. Characterization of the impact of uncertainty in model inputs on model outputs is therefore important for understanding population health measurement with multi-state longitudinal models. Despite the long history of use of life-table methods, we have nevertheless been unable to find published work analysing this problem. In Appendix I, *Uncertainty analysis in longitudinal population models*, we propose an approach where we

focus on the uncertainty of input rates only; uncertainty in the estimates of the size and relative size of input populations is considered secondary, as is uncertainty in the estimates of the input prevalence of the diseases, risk factors or conditions analsyed as well as uncertainty in the estimates of input health state valuations. The main thrust of the approach is to analyse both demographic and epidemiological transition rates as random variables. The analysis is applied first to the case of a single population exposed to a single risk of transition to an absorbing state (i.e. the ordinary life-table model), then extended to the multi-state dynamic life-table framework. The main theme that emerges is that continuous-time dynamic models imply a simple "additive-in-the-exponent" rule for rates; this rule carries over, with minor modifications, to the case where rates are random variables, where additivity needs also to consider the statistical properties of the input rates.

Complex interactions in longitudinal population models

As noted, the goal of generality in model formulation can be illusory, since the need for realism according to subject matter criteria inevitably introduces an ad-hoc component to the model. In particular, one of the problems identified for the would-be user of PopMod is the need to model more than four population states. Since, as the number of states x increases, the number of possible transitions between states increases approximately as x^2, estimating the corresponding input parameters using only generally available data rapidly becomes a complex problem. For example, a typical modelling situation might require determining the joint distribution of the prevalence of $m \times k$ population states when only $m + k$ marginal distributions are observed. There is the related problem of partitioning a total disease rate r across the $m \times k$ joint population states. A formal statement of the problem is provided as Appendix II, Modelling complex interactions in longitudinal population models, where both the prevalence and the rate aspects of the problem are characterized in terms of a system of linear equations. Simple matrix algebra theory permits the development of general insights regarding the amount of information required for the solutions to these problems.

The two appendices represent preliminary steps that we have taken to address these important areas for future research. In order to complete the theory of the time dimension in the measurement of population health, this preliminary work should be expanded upon and complemented by additional research.

Conclusions

1. Longitudinal population models are superior to others because they permit the tabulation, when required, of the age, period and cohort position of health events.
2. Generic multi-state longitudinal population models can be developed, but require trade offs between generality and realism.
3. Generic multi-state longitudinal models promote consistent standards of population health measurement.
4. Generic multi-state longitudinal population models can be applied to real problems using available epidemiological and demographic data, but they often require estimates of some epidemiological and demographic parameters that are not typically available.
5. Cost-effectiveness analysis using generic multi-state longitudinal population models can contribute to rational priority setting and policy making.
6. Standard statistical and epidemiological models can be used to obtain information about required input parameters that are not directly observed, but their use for multi-state longitudinal population modelling may require assumptions that are potential sources of bias.

References

1. Betrán AP, Merialdi M, Lauer JA, Bing-shun W, Thomas J, Van Look P, Wagner M. Rates of caesarean section: analysis of global, regional and national estimates, *Paediatric and perinatal epidemiology*, 2007, 21:98–113.

2. Black RE, Allen LH, Bhutta ZA, Caulfield LE, de Onis M, Ezzati M, Mathers C, Rivera J, for the Maternal and Child Undernutrition Study Group. Maternal and child undernutrition: global and regional exposures and health consequences. *Lancet*, 371(9608):243-260, 2008.

3. Caswell H. *Matrix population models: construction, analysis, and interpretation.* Sunderland, Massachusetts, Sinauer Associates, 2001.

4. Chisholm D, Rehm J, Van Ommeren M, Monteiro M. Reducing the global burden of hazardous alcohol use: a comparative cost-effectiveness analysis. *Journal of studies on alcohol*, 65(6):782-793, 2004.

5. Castro A. Commentary: increase in caesarean section may reflect medical control not women's choice. *BMJ*, 319:1401-1402, 2001.

6. Keyfitz N, Caswell H. *Applied mathematical demography*, 3rd ed. Berlin, Springer, 2005.

7. Pearl J. *Causality: models, reasoning and inference*. Cambridge University Press, 2000.

8. Preston SH. Health indices as a guide to health sector planning: a demographic critique. In: The epidemiological transition, policy and planning implications for developing countries (Edited by: Gribble JN, Preston SH) Washington DC, National Academy Press, 1993.

9. Singer B, Spilerman S. The representation of social processes by Markov models. *The American journal of sociology*, 82(1):1-54, 1976.

10. Showalter E, Griffin A. Commentary: all women should have a choice. *BMJ*, 319:1401, 2001.

Appendix I

Uncertainty analysis in longitudinal population models

Introduction[1]

Multi-state longitudinal life tables are used to measure population health as well as to study the effect of specific interventions implemented for the sake of improving population health (Lauer et al. 2003). How does uncertainty in parameter values used as inputs affect the projections of multi-state longitudinal life tables? It seems that the uncertainty in the results of such models will depend on the individual contributions of multiple uncertain parameters, as well as on the dynamic characteristics of the model itself.

Despite their obvious interest, these and related questions seem to have been ignored in the relevant literature on demography, ecology (e.g. matrix population models; Caswell, 2001) and epidemiology. In fact, to identify methods relevant to the analysis of these questions, one has to turn to the mathematical, physical and statistical literature on stochastic processes and Markov models.

We shall discuss uncertainty in the rate parameters of the population model only. This restriction is partly for the sake of convenience and partly a reflection of the fact that rates of mortality and disease are greater determinants of population health than other parameters (e.g. the sizes, or relative sizes, of initial populations). Moreover, we shall not develop detailed results for the distributions of individual life table quantities but shall rather discuss the methods and principles that allow for a general approach to such calculations.

1 Harald Wirth, Klaus Röhrich, Jan Barendregt and Dik Habbema are gratefully acknowledged for their helpful discussions on this appendix.

Formal statement of the problem

For the purposes of population health measurement, we consider a differential equation system of the following form:

$$\text{Eq. (1)} \qquad \boldsymbol{p}'(t) = \boldsymbol{C} \cdot \boldsymbol{p}(0),$$

where $\boldsymbol{p}' \equiv d\boldsymbol{p}/dt$ denotes the derivative of $\boldsymbol{p}$ with respect to time, $\boldsymbol{C}$ is a matrix of non-time-varying rates and $\boldsymbol{p}(0)$ is the vector of initial populations. For the case of the population model, it is natural to consider $\boldsymbol{p}(0)$ as the population of interest partitioned in the set of defined age groups.

It is well known that the general solution to the system of differential equations can be written as (see, for example, Strang, 1988):

$$\text{Eq. (2)} \qquad \boldsymbol{p}(t) = e^{\boldsymbol{C} \cdot t} \cdot \boldsymbol{p}(0).$$

This result means that initial populations evolve exponentially with time; Eq. (2) is merely a concise way of representing the population model.

In terms of Eq. (2), we are interested in the uncertainty of $\boldsymbol{p}(t)$ when there is uncertainty in the elements of $\boldsymbol{C}$. We are also interested in the uncertainty of various life-table functions of the populations $\boldsymbol{p}(t)$, which we can write as $f(\boldsymbol{p}(t))$. Important examples of such functions are the average population size in an interval (population residence time) and the sums of populations or of average populations over a given horizon (Lauer et al., 2003).

Scope of the analysis

Uncertainty in $\boldsymbol{p}(t)$ can in general be attributed to the following sources:
— model uncertainty,
— stochastic uncertainty (i.e. random variation) in the modelled process, and
— uncertainty in model parameters.
We briefly describe each of these sources below.

Model uncertainty. Model uncertainty refers to uncertainty in the populations $\boldsymbol{p}(t)$ when the specific mathematical method used to model the population (e.g. Eq. (2)), or its related assumptions, are incorrect or represent a simplification or idealization of the true underlying population process. For example, exponential evolution in continuous time may be used to approximate an underlying process that is in fact discontinuous. Alternatively, rates may not be constant but time varying. Typically, however, model uncertainty arises when the matrix $\boldsymbol{C}$ is missing elements that are important for the question under study. This can arise due to the fact only a small number of known transitions are typically represented in $\boldsymbol{C}$, as well as to the fact that there are many important transitions that are unknown (e.g. exposure to unknown risk factors).

Stochastic uncertainty. Stochastic uncertainty arises when the population process is inherently stochastic. It is useful to distinguish the uncertainty arising from modelling a discrete process by means of a continuous one, or from modelling a random process by means of a deterministic one (both of which would be causes of model uncertainty in terms of the definition above), from the intrinsic variability of an inherently random process.

Parameter uncertainty. Parameter uncertainty refers to the fact that parameters in the population model may be unknown or are measured imperfectly. Measurement uncertainty can either be statistical or systematic. Statistical measurement uncertainty refers to random variation in either the measurement tools or in the measurers (which implies that measurement is a random process, i.e. subject to inherent random variation). Since statistical parameter uncertainty can either be model error (for example, if the measurement tool is deterministic but the process measured is random) or

simple stochastic uncertainty, it is therefore best considered under one or the other of those respective headings. On the other hand, systematic parameter uncertainty can arise due to the presence in the measurement tool of non-random dependence on important parameters whose values are unmeasured or unknown or to systematic bias in the measurement method (cf. Hubble telescope). Notice therefore that systematic parameter uncertainty can therefore be considered as a form of model uncertainty.

In terms of these definitions, the analysis here applies to uncertainty in model outcomes due to stochastic variability in the elements of the matrix of rates $\boldsymbol{C}$. However, the approach is also well suited to the characterization of statistical parameter uncertainty arising due to random errors in the measurement tool or in the measurers). Finally, inherently random processes modelled by deterministic processes will display model error that is assimilable to the sort of uncertainty characterized here. We do not address other forms of model uncertainty or systematic parameter uncertainty.

In the notation of (Lauer et al., 2003), the matrix of rates $\boldsymbol{C}$ can be written as follows:

$$\begin{pmatrix} h_{S\to X}+h_{S\to C}+m & -h_{X\to S} & -h_{C\to S} & 0 \\ -h_{S\to X} & h_{X\to XC}+h_{X\to S}+m+f_X & 0 & -h_{XC\to X} \\ -h_{S\to C} & 0 & h_{C\to XC}+h_{C\to S}+m+f_C & -h_{XC\to C} \\ 0 & -h_{X\to XC} & -h_{C\to XC} & h_{XC\to C}+h_{XC\to X}+m+f_X+f_C \end{pmatrix}$$

Note that, other than the stochastic component, all matrix entries are non-time varying. Transitions between health states are denoted by h, and distinguished by subscripts, and background mortality and case fatality rates are labelled m and f, respectively, the latter also distinguished by subscripts. In terms of that notation, $\boldsymbol{p}(0) = (S(0)\ X(0)\ C(0)\ XC(0))^{\mathrm{T}}$, where the population of each possible health state is defined at $t = 0$.

Analysis of specific cases

It is convenient to consider initially the case of a single population exposed to a single transition, in other words, the case of a standard cohort life table, where $\boldsymbol{C} = \mu$ and $\boldsymbol{p}(0) = p(0)$.

Cohort life table under uncertainty

We refer to a continuous-time life table comprising a single state exposed to constant mortality rate μ_y, where the subscript y indexes the rate prevailing during period y (notionally, one year). With subscripts now denoting time, within any period the population at time t is given by:

Eq. (3) $$p_t = p_{t-1} e^{-\mu_y \cdot t}.$$

Suppose $p_{t-1} = p_0$ and $p_0 = 1$ (the latter condition can always be enforced without loss of generality by dividing both p_t and p_0 by p_0). Suppose, moreover, that $t = 1$ year and that μ_y has units "year^{-1}". Then we can write:

Eq. (4) $$p_1 = e^{-\mu_1}$$

for the population surviving at the end of year 1.

Now suppose that, starting from the beginning of year 2, the population surviving at the end of year 1 is exposed to μ_2. Substituting the value obtained for p_1, it follows that the population surviving at the end of year 2 can be written as:

Eq. (5) $$p_2 = e^{-\mu_1} \cdot e^{-\mu_2} = e^{-(\mu_1+\mu_2)}.$$

Similarly, if the population surviving at the end of year 2 is exposed to μ_3 from the beginning of year 3, the surviving population at the end of year 3 is given by:

Eq. (6) $$p_3 = e^{-(\mu_1+\mu_2)} \cdot e^{-\mu_3} = e^{-(\mu_1+\mu_2+\mu_3)}.$$

Under these assumptions, it is clear that in year t,

Eq. (7) $$p_t = e^{-\sum_t \mu_y}.$$

Notice, however, that these assumptions are merely those of the cohort version of the standard life table in continuous time where age-specific mortality is assumed constant within an age group but differs by year of age.

Independent rates

Now suppose that μ_y no longer represents a constant mortality rate but is instead the mean value of a normally distributed random rate h_y, with variance σ_y^2, and that the h_y are independent. We refer to this concept of random variability as the "null model". The null model provides a useful reference case for comparison with more complicated types of variability.

It follows from the definition of the log-normal random variable that P_1 is not a constant with value $e^{-\mu_1}$ but rather a log-normally distributed random variable with mean $e^{-\mu_1+\frac{1}{2}\sigma_1^2}$ and variance $e^{-2\mu_1+2\sigma_1^2}-e^{-2\mu_1+\sigma_1^2}$, the latter equal to $e^{-2\mu_1+\sigma_1^2}(-1+e^{\sigma_1^2})$ after factoring. Notice that, when compared with the case of constant rates, mean mortality is reduced by the quantity $\frac{1}{2}\sigma_y^2$. Regardless of the precise distributional form, when rates are random the mean population at year's end will show correction factors whose values are subtracted from mean mortality. Thus, the mean number of survivors when rates are random is greater than the number of survivors when rates are constant.

The sum of t independent normal random variables with means μ_y and variances σ_y^2 is itself a normally distributed random variable with mean $\sum_t \mu_y$ and $\sum_t \sigma_y^2$ (Gardiner, 2004). It follows that p_t is a log-normally distributed random variable with mean $e^{\sum_t \mu_y+\frac{1}{2}\sum_t \sigma_y^2}$ and variance $e^{-2(\sum_t \mu_y+\frac{1}{2}\sum_t \sigma_y^2)+\sum_t \sigma_y^2}+e^{-2(\sum_t \mu_y+\frac{1}{2}\sum_t \sigma_y^2)+2\sum_t \sigma_y^2})$, the latter expression equal to $e^{-2\sum_t \mu_y}(-1+e^{\sum_t \sigma_y^2})$ after cancellations and factoring.

Since the distributional family of populations in the cohort null model is invariant by year, it is natural to think of the quantities $\sum_t \mu_y$ as "pseudorates" (i.e. "cumulative rates") with variance $\sum_t \sigma_y^2$. Since pseudorates are in this respect indistinguishable from actual (annual) rates, uncertainty analysis of the cohort null model can be greatly simplified. For example, if all annual rates have mean μ and variance σ^2, the mean population after t years is simply $e^{t\mu+\frac{1}{2}t\sigma^2}$, and the corresponding expression for population variance is $e^{-2t\mu}(-1+e^{t\sigma^2})$.

Population correlations when rates are independent

Although rates are independent in the cohort null model, age-specific populations follow a specific correlation law. Write $h_y = \mu_y + dh_y$, where dh_y is a normally distributed random variable with a mean of zero and variance σ_y^2. We know from the population model that:

$$\text{Eq. (8)} \qquad p_t = p_{t-1} e^{-h_y}.$$

Rewriting h_y in terms of deviations around its mean, we can moreover write:

$$\text{Eq. (9)} \qquad e^{-h_y} = e^{-\mu_y - dh_y}.$$

By both adding and subtracting the quantity $\frac{1}{2}\sigma_y^2$ in the exponent and factoring we then write:

$$\text{Eq. (10)} \qquad e^{-h_y} = e^{-\mu_y + \frac{1}{2}\sigma_y^2} e^{-dh_y - \frac{1}{2}\sigma_y^2}.$$

It follows that the population model can be written as:

$$\text{Eq. (11)} \qquad p_t = p_{t-1} e^{-\mu_y + \frac{1}{2}\sigma_y^2} e^{-dh_y - \frac{1}{2}\sigma_y^2}.$$

But Eq. (11) says that p_t is equal to the product of p_{t-1}, a positive constant $e^{-\mu_y + \frac{1}{2}\sigma_y^2}$ (whose value is less than 1), and a random variable $e^{-dh_y - \frac{1}{2}\sigma_y^2}$ with a mean of 1 (since $-dh - \frac{1}{2}\sigma_y^2$ is normally distributed with mean $-\frac{1}{2}\sigma_y^2$ and variance σ_y^2, its exponential is a log-normal random variable with mean $e^{-\frac{1}{2}\sigma_y^2 + \frac{1}{2}\sigma_y^2}$, which is equal to e^0).

A closely related model from the time-series literature is the first-order autoregressive (AR-1) process (Goldberger, 1991):

$$\text{Eq. (12)} \qquad p_t = \alpha p_{t-1} + \varepsilon_t,$$

where ε_t is a random variable with zero mean and variance σ^2 and α is a constant whose value is less than 1. In a similar notation, the cohort null model could be written:

$$\text{Eq. (13)} \qquad p_t = \alpha_y p_{t-1} \xi_t,$$

where α_y is equal to $e^{-\mu_y + \frac{1}{2}\sigma_y^2}$ and ξ_t is equal to $e^{-dh_y - \frac{1}{2}\sigma_y^2}$.

With identically distributed rates, by induction on t it can be seen that the covariance of two populations t years different in age is given by $\alpha^t\sigma^2$; this is clearly the same correlation law as for the additive AR-1 process (Goldberger, 1991), justifying the application of the term "first-order autoregressive" to the null cohort model. Moreover, when rates differ by year, it is not hard to see that the covariance of two populations separated by t years is given, by induction on t, by the variance of the first population σ_1^2 times the product of the t coefficients α_y, in other words, $\sigma_1^2 \cdot \Pi_t \alpha_y$. The error process of the cohort null model is thus a multiplicative analog of the standard AR-1 process, with the additional possibility that α_y is a period-specific constant.

Such a model is also closely related to geometric Brownian motion and to the geometric random walk (Franke et al., 2004). Taking the logarithm of both sides of Eq. (11), we get:

$$\text{Eq. (14)} \qquad \log p_t = \log p_{t-1} - \mu_y - \mathrm{d}h_y.$$

But, since $\mathrm{d}h_y$ has a mean of zero, we can also write:

$$\text{Eq. (15)} \qquad \log p_t = -\mu_y + \log p_{t-1} + \mathrm{d}h_y.$$

Thus, the logarithm of population in the cohort null model follows a random walk with a period-specific drift term $-\mu_y$. This differs from the standard geometric random walk in that the drift term can be period specific; it differs from geometric Brownian motion in that $\mathrm{d}h_y$ is not continuously random.

In sum, the correlation law for populations in the cohort null model is well defined. The null model correlation law can also be related to well known stochastic-process and time-series models.

Distributional characteristics of residence time

So far we have focused on the uncertainty analysis of population quantities. The population residence time (i.e. aggregate population life years, p_{LY}) in an interval is another quantity of interest.

Population residence time is equal to the time integral of the population within the interval. For example, in the cohort null model for the first year, we have (Lauer et al., 2003):

$$\text{Eq. (16)} \qquad p_{\mathrm{LY}(1)} = \int_0^1 p_0 e^{-h_1 t}\,\mathrm{d}t = \frac{1-e^{-h_1}}{h_1} \cdot p_0 .$$

For h_1 random, the right-hand side of Eq. (16) defines a distribution for $p_{\mathrm{LY}(1)}$. Although p_0 is non-stochastic by assumption, in general p_{t-1} is a random variable. Nevertheless, we focus first on the distributional characteristics of the expression $(1 - e^{-h})/h$. In a continuous-time life table, this expression gives the general formula for the time average of survivors of a population exposed to mortality rate h during a one-year interval. Since the time average of survivors is nothing other than the population residence time (life years) accumulated during the interval, the distribution defined by Eq. (16) is of general interest for population health measurement.

Since the distributional characteristics of residence time cannot in general be given in terms of closed-form analytical expressions (exceptions are when h is gamma or uniform distributed), we seek approximate expressions for its mean and variance. For any standard distributions for h, approximations can be found by means of the delta method (Goldberger, 1991). For the mean, the delta method to linear order implies taking the function of the mean $f(\mu)$ as an approximation to the mean of the function $\mathrm{E}[f(h)]$, in other words we take the quantity $(1 - e^{-\mu})/\mu$ as an approximation to the mean of $(1 - e^{-h})/h$.

To obtain the delta-method approximate variance, the Taylor-series expansion of the function around the mean of h is truncated at linear order, and the theorem for variance of a linear function is used. This yields the expression $(e^{-2\mu}(1-e^{\mu}+\mu)^2)\mu^{-4}$ as the linear-order delta-method approximation to the variance of $(1 - e^{-h})/h$. Thus, linear-order approximations for both the mean and the variance of residence time are expressable in terms of the mean only of the rate h.

If it is relevant to consider the value of p_{t-1} as a random variable, residence time in the interval is therefore the product of two random variables. Calculating the expectation and variance of the product of random variables requires the covariance of the random variables. By using the delta-method result for the mean (i.e. $1 - e^{-\mu})/\mu$), the required covariance can be given as approximately $-(e^{-\mu_y+\frac{1}{2}\sigma_y^2})\mu^{-2}\sigma^2$.

Correlated rates

We maintain the assumption that rates h_y are normally distributed with mean μ_y and variance σ_y^2 but relax the assumption of independence by interval. Successive rates therefore have a non-zero covariance. Since the expected value of a sum is equal to the sum of expected values, the mean value of any period-specific pseudorate is still $\sum_t \mu_y$. However, the variance of the pseudorate is no longer equal to $\sum_t \sigma_y^2$. Moreover, the pseudorates are no longer normally distributed, and the distribution of year-end populations will not be log-normal.

We claim that the distribution of year-end populations exposed to random rates in a single-state population model will be given by the distribution of the exponential of some function *f*(.), where the arguments of *f*(.) are random rates not necessarily independently, normally or identically distributed. However, because of the nature of the population model, *f*(.) is in fact restricted to a simple sum, i.e. $f(h_y) = \Sigma h_y$. In the null model, attention was restricted to the case where arguments of *f*(.) are independent normal random variables. In the correlated case we drop the restriction of the independence of the h_y and consider non-independent normally distributed random variables. The sums of such random variables are not normally distributed, which has implications for caculating the moments of $e^{f(h)}$. We shall see that, without much additional complication, we can subsequently drop the assumption of normality. In all of these cases, our aim is to find expressions for the distribution of $e^{f(h)}$, or at least for its mean and variance. In the null model, we can determine the distribution of $e^{f(h)}$ exactly; in other cases we can do so only approximately.

A natural approach is to characterize the distribution of $e^{f(h)}$ in terms of the mean and variance of *f*(.). Such an approach is attractive since *f*(.) is always additive and its mean and variance can always be easily obtained. Although the mean will always be given by the sum of the means of its arguments, to obtain the variance of *f*(.), the covariance between rates needs to be taken into account. For example, the variance of the sum of two random rates, h_1 and h_2, is given by:

$$\text{Eq. (17)} \qquad \mathrm{Var}(f) = \mathrm{Var}(h_1 + h_2) = \mathrm{Var}(h_1) + \mathrm{Var}(h_2) + 2\mathrm{Cov}(h_1, h_2)\,.$$

For a pseudorate in year 3, three possibly non-zero covariance terms need to be considered:

$$\text{Eq. (18)} \qquad \begin{aligned} \mathrm{Var}(h_1 + h_2 + h_3) = \mathrm{Var}(h_1) + \mathrm{Var}(h_2) + \mathrm{Var}(h_3) + \\ 2\mathrm{Cov}(h_1, h_2) + 2\mathrm{Cov}(h_1, h_3) + 2\mathrm{Cov}(h_2, h_3) \end{aligned}$$

It follows that, for the variance of the pseudorate in year t, there are $\binom{t}{2}$ potentially non-zero covariance terms.

Despite the growing number of terms in the variance of $f(.)$, provided that the means, variances and covariances of the h_y are known, the mean and variance of $f(.)$ can be exactly determined. A simple model for non-independent rates that satisfies this condition is one where rates follow an AR-1 process. Under such or similar conditions, therefore, determining the exact mean and variance of $f(.)$ is feasible.

Series expansion methods

Approximating the mean and variance of $e^{f(h)}$ can be done with methods of series expansion. If it is sufficent to estimate the mean and variance of $e^{f(h)}$, one approach is to express $e^{f(h)}$ as a series expansion in terms of powers of a dummy constant and to take the required moments of the terms of a truncated version of the series (Weigel, 1986). If, on the other hand, one needs to characterize more fully the distribution of $e^{f(h)}$, the methods of Edgeworth expansion, whereby the density of $e^{f(h)}$ is approximated in terms of a modified normal density, are available (Kolassa, 2006).

Taking the first approach, we write $e^{\lambda f}$ as a series expansion in the powers of λ and take the mean of both sides (equivalent to writing f in terms of its characteristic function):

$$\text{Eq. (19)} \qquad \langle e^{\lambda f} \rangle = 1 + \lambda \langle f \rangle + \tfrac{1}{2} \lambda^2 \langle f^2 \rangle + \ldots.$$

The right-hand side can then be rewritten (Weigel, 1986):

$$\text{Eq. (20)} \qquad \langle e^{\lambda f} \rangle = \exp\{\langle \lambda f \rangle + \tfrac{1}{2} \lambda^2 (\langle f^2 \rangle - \langle f^2 \rangle) + O(\lambda^3 f^3)\}.$$

Setting $\lambda = 1$, the mean of $e^{f(h)}$ can be written in terms of the mean, μ, and variance, σ^2, of $f(.)$ and a term that is $O(f^3)$:

Eq. (21) $\langle e^f \rangle = \exp\{\mu + \frac{1}{2}\sigma^2 + O(f^3)\}.$

Note that when f is normally distributed, $O(f^3)$ must be identically equal to zero.

Truncating Eq. (21) at second order in the exponential, the mean of $e^{f(h)}$ can thus be written as approximately:

Eq. (22) $\langle e^f \rangle \approx \exp\{\langle f \rangle - \frac{1}{2}(\langle f^2 \rangle - \langle f \rangle^2)\}.$

To obtain the variance of $e^{f(h)}$, we proceed as follows. Since:

Eq. (23) $\mathrm{Var}(X) = \langle X^2 \rangle - \langle X \rangle^2,$

it follows that:

Eq. (24) $\mathrm{Var}(e^f) = \langle e^{2f} \rangle - \langle e^f \rangle^2.$

But since, by Eq. (21):

Eq. (25) $\langle e^{2f} \rangle = \exp\{\langle 2f \rangle - \frac{1}{2}(\langle 4f^2 \rangle - \langle 2f \rangle^2) + O(f^3)\},$

and:

Eq. (26) $\langle e^f \rangle^2 = \exp\{2\langle f \rangle + \langle f^2 \rangle - \langle f \rangle^2 + O(f^3)\},$

it follows that:

Eq. (27) $\mathrm{Var}(e^f) \approx \exp\{2\langle f \rangle - 2\langle f^2 \rangle + \langle f \rangle^2\} - \exp\{2\langle f \rangle + \langle f^2 \rangle - \langle f \rangle^2\},$

with exact equality when f is normal. So approximate expressions for the mean and the variance of $e^{f(h)}$ can be written in terms of the mean and variance of f.

The above approach is equivalent to approximating the exponential of an arbitrary random variable by the exponential of a normally distributed random variable. This is evidently analogous to approximating an arbitrary distribution by a normal distribution with the same mean and variance. While this may seem to be a somewhat arbitrary procedure, it is equivalent to truncating the Edgeworth expansion of f at the first term (Kolassa, 2006);

an expression for the error of truncation is moreover obtained. When f is even approximately normal, this is a reasonable procedure and is in fact commonly used (Weigel, 1986).

However, it is straightforward to obtain a better approximation if necessary by writing the distribution of $e^{f(h)}$ directly in terms of its Edgeworth expansion (also referred to as the cumulant expansion):

Eq. (28) $$e^{f(h)} \sim \phi(e^f) \times [1 + \kappa_3 h_3(e^f)/6 + (\kappa_4 h_4(e^f)/24 + 10\kappa_3^2 h_6(e^f)/720) + \ldots],$$

where $\phi(e^f)$ is the normal density function with the same mean and variance as e^f, the κ_i are the cumulants of e^f of order i and $h_i(e^f)$ are the Hermite polynomials (McCullagh, 1987). The higher-order cumulants of e^f ($i > 2$) will in general have to be estimated from a sample (possibly obtained by simulation) of e^f. The sample-dependent nature of the estimates of the higher-order cumulants makes it relevant to express the error in truncation in terms of a measure of sample size. However, even for theoretical calculations, it is almost never necessary to include more than the first or second non-constant term in square brackets (McCullagh, 1987). Truncation error can be expressed in terms of the sample size, n, used for estimation of the cumulants. Truncating the series after the first non-constant term in square brackets yields error $O(n^{-1})$; truncating after the second term (i.e. not including the dots) yields error $O(n^{-3/2})$ (McCullagh, 1987).

Recall that the covariance of populations depends only on the mean value of year-end populations (e.g. in the independent case $\alpha_y = e^{-\mu_y + \frac{1}{2}\sigma_y^2}$). In the dependent case, therefore, α_y will depend only on the mean of $e^{f(h)}$. Thus dependence in rates affects the correlation of populations only indirectly, in other words, through the means of the pseudorates. In the dependent case as well, the pseudorates therefore remain the key to the uncertainty analysis of the population model.

Implications of various random models for h

It is possible to make several qualitative statements about the behaviour of $f(.)$ for different random distributions h:

— When h is uniformly distributed, Σh rapidly converges to an approximately normal distribution as the number of terms h increases.

— When h is gamma distributed with a common parameter θ, Σh will be gamma distributed, so when h is gamma distributed with parameters θ_y approximately equal, Σh will be approximately gamma distributed.
— When h is log-normally distributed, Σh will be approximately log-normally distributed (Wu et al., 2005).

Population model under uncertainty

The conclusions of the cohort model generalize relatively easily to the full population model. Considering other age-group cohorts, the conclusions stated above remain valid: under the null model, year-end populations are log-normally distributed random variables whose means and variances follow a simple additive-in-the-exponent rule and null-model populations can be characterized in terms of stochastic processes whose covariance properties are well established.

One additional complication is possible, however. Rates may be either correlated or uncorrelated across age groups. However, whatever the correlation pattern across age-specific rates, age-specific populations across the respective cohorts will be correlated through their respective population means.

Multi-state life table under uncertainty

Introduction

When there is more than one possible state in the population model, as in the longitudinal population model described in (Lauer et al., 2003), we have the exponential of the matrix of transition rates (Eq. (2)) rather than the exponential of a single transition rate. The exponential of a matrix is defined in terms of the infinite series of the powers of the matrix (cf. Singer and Spilerman, 1976). However, for diagonalizable matrices (i.e. for square matrices of full rank) the matrix exponential has a closed-form expression in terms of the eigenvalues and eigenvectors. A diagonalizable matrix $\boldsymbol{C}$ can be written

$$\text{Eq. (29)} \qquad \boldsymbol{C} = \mathrm{N}\cdot\Lambda\cdot\mathrm{N}^{-1},$$

where N is the matrix whose columns are the eigenvectors and Λ is a diagonal matrix of the eigenvalues. It is then easy to show that

Eq. (30) $\exp(\boldsymbol{C}) = \mathbf{N}\cdot\exp(\Lambda)\cdot\mathbf{N}^{-1}$,

and that $\exp(\Lambda)$ is equal to the diagonal matrix whose elements are the exponentials of the individual eigenvalues (Strang, 1988). All $\boldsymbol{C}$ matrices for a population model will be square matrices of full rank, so these results apply to our case. In the population-model context, the eigenvalues are analogous to the rate in the scalar case; the eigenvectors represent the initial population values, and it is seen in the equation above, as in the equation for the scalar case, that they multiply the exponentiated eigenvalues in the matrix exponential (Caswell, 2001). Each distinct eigenvalue of $\boldsymbol{C}$ is associated with a unique state in the population model, and expresses the "long-term growth potential" of that state (Caswell, 2001).

When rates have units "year^{-1}", the matrix exponential gives a matrix of probabilities whose i,j-th element expresses the one-year probability of transition from state i to state j in the population model (Singer and Spilerman, 1976). As such, each element in the matrix exponential represents the sum of the probabilities of making the $i\rightarrow j$-transition considering all possible paths from state i to state j, including paths requiring an infinite number of steps (Van Kampen, 1992). As it affords what is called a "spectral decomposition" of the transition matrix $\boldsymbol{C}$, the matrix exponential provides a very useful formalism for the multi-state population model. The matrix exponential representation allows an analogous representation to that given for the single-cohort case in Eq. (3):

Eq. (31) $\boldsymbol{p}(t) = \mathbf{A}(\boldsymbol{C}) \cdot \boldsymbol{p}(t-1)$.

Here $\mathbf{A}(\boldsymbol{C})$ is the notation adopted for the matrix of survival probabilities (written as a function of the transition operator $\boldsymbol{C}$ to emphasize its dependence on the rates).

When rates are random, $\mathbf{A}(\boldsymbol{C})$ can be partitioned, as in Eq. (11), into a matrix of the constant means and a matrix of multiplicative random deviations. The multi-state longitudinal population model can then be written:

Eq. (32) $\boldsymbol{p}(t) = \langle\mathbf{A}(\boldsymbol{C})\rangle \otimes \mathbf{d}\langle\mathbf{A}(\boldsymbol{C})\rangle \cdot \boldsymbol{p}(t-1)$,

where $\langle \mathbf{A}(\boldsymbol{C}) \rangle$ refers to the matrix of means of the elements of $\mathbf{A}(\boldsymbol{C})$, $\mathbf{d}\langle \mathbf{A}(\boldsymbol{C}) \rangle$ refers to the matrix of multiplicative random deviations around the means and "⊗" indicates the Hadamard, or element-by-element, product.

In analogy, moreover, with the constant factors α_y of Eq. (13), the elements of $\langle \mathbf{A}(\boldsymbol{C}) \rangle$ are expressable in terms of the series expansion of the elements of $\mathbf{A}(\boldsymbol{C})$. For example, if an element of $\mathbf{A}(\boldsymbol{C})$, which we can notionally represent as e^f (the notion being that f is an eigenvalue of $\boldsymbol{C}$), happens to be log-normally distributed, its mean as given by Eq. (21) will not contain a term of $O(f^3)$ but will be exactly equal to $e^{\mu+\frac{1}{2}\sigma^2}$, where μ and σ^2 are the mean and variance, respectively, of f (Gardiner, 2004). In this case, the corresponding element in the matrix of deviations, $\mathbf{d}\langle \mathbf{A}(\boldsymbol{C}) \rangle$, is a log-normally distributed random variable with a mean of 1 and variance equal to $e^{-2\mu_1+\sigma_1^2}(-1+e^{\sigma_1^2})$.

Distribution of elements of A(C)

Since, however, each element of $\mathbf{A}(\boldsymbol{C})$ is in fact of the form $\Sigma i g_i \cdot e^f g_i^{-1}$, where g_i represents eigenvector i of $\boldsymbol{C}$ and f represents an eigenvalue of $\boldsymbol{C}$, elements of $\mathbf{A}(\boldsymbol{C})$ will in general not be log-normal in distribution. However, algebraic expressions for the eigenvalues and eigenvectors of the matrix $\boldsymbol{C}$ can be calculated for any arbitrary population model with four or fewer states. In addition, for many reasonable population models containing five or more states, the eigenvectors and eigenvalues can still be calculated as algebraic expressions. Entries in $\mathbf{A}(\boldsymbol{C})$ can therefore often be expressed in terms of analytic functions of algebraic expressions of random variables.

For example, for a population model with the 2 x 2 matrix $\boldsymbol{C}$:

$$\begin{pmatrix} a+c & -b \\ -c & a+b \end{pmatrix},$$

the eigenvalues are:

$$\begin{pmatrix} a \\ a+b+c \end{pmatrix},$$

the eigenvectors are the columns of:

$$\begin{pmatrix} b/c & -1 \\ 1 & 1 \end{pmatrix},$$

and the corresponding matrix $\mathbf{A}(\boldsymbol{C})$ is:

$$\begin{pmatrix} \dfrac{e^a(b+ce^{b+c})}{b+c} & -\dfrac{be^a(-1+e^{b+c})}{b+c} \\ -\dfrac{ce^a(-1+e^{b+c})}{b+c} & \dfrac{e^a(c+be^{b+c})}{b+c} \end{pmatrix}.$$

What can we say in this case about the distribution of the elements of $\mathbf{A}(\boldsymbol{C})$ when the elements of $\boldsymbol{C}$ are random? If one assumes that rates are normally distributed and independent, their sums and differences will also be normally distributed. In addition, the exponentials of normally distributed random variables are log-normally distributed. So the numerators of the individual elements can be reduced to a sum of the products of normal and log-normal distributions. Since the denominators in this case are themselves normal distributions, it is reasonable to expect that the ratio distribution will be approximately log-normally distributed (as an heuristic, one can imagine that the normal densities "cancel" in both the numerator and denominator).

However, for a population model such as that in (Lauer et al., 2003), with a 4 x 4 matrix $\boldsymbol{C}$, these expressions contain hundreds of terms, including polynomial combinations of elements of $\boldsymbol{C}$ and roots of polynomial combinations of elements of $\boldsymbol{C}$. In such a case, no obvious intuition seems available concerning the distribution of the elements of $\mathbf{A}(\boldsymbol{C})$; calculating the distributional form analytically is moreover a virtual impossibility.

We suggest the following conjecture. Since, when many random terms combine additively a normal distribution is the result, and when many such terms combine multiplicatively a log-normal distribution results, when many random terms combine both additively and multiplicatively, depending on the predominance of one or the other operation, as well as on the relative magnitudes of the terms, one can expect either the normal or the log-normal

form to dominate, at least as an approximation. We support this conjecture below with results from simulations.

Cumulant expansion of elements of A(C)

As both the eigenvalues (f) and eigenvector elements (g) of $\boldsymbol{C}$ are functions of random variables, the cumulant expansion of the elements will depend on the distributions of g and of f, as well as on their correlations. Moreover, as there is no guarantee that elements of $\mathbf{A}(\boldsymbol{C})$ are normal in distribution, their cumulant expansion will in general have a potentially infinite number of terms. Consequently explicit expressions for their means and variances will necessarily be approximate, and the elements of $\mathbf{d}\langle\mathbf{A}(\boldsymbol{C})\rangle$ can likewise be written as only approximately equal to 1 in expectation. Although this implies that we are limited to writing Eq. (26) as an approximate equality only, any arbitrary level of precision can be obtained by means of additional terms in the cumulant expansion if necessary.

Simulation approach

This difficulty of calculating an explicit distributional form for the elements of $\mathbf{A}(\boldsymbol{C})$ when its elements are not normally distributed suggests an efficient use of the simulation method for purposes of studying the uncertainty of a given multi-state longitudinal population model:

1. One defines distributions for the random rates that make up the elements of $\boldsymbol{C}$.
2. Multiple random draws are made from each of these distributions, so as to produce a distribution of random matrices $\boldsymbol{C}$.
3. For each draw of $\boldsymbol{C}$, the matrix of one-year probabilities $\mathbf{A}_y(\boldsymbol{C}_y)$, which is in general understood to be period specific, is computed numerically.
4. The resulting distribution of random matrices $\mathbf{A}_y(\boldsymbol{C}_y)$ is used to compile distributions for each of its elements.
5. The distributions of each of the elements of $\mathbf{A}_y(\boldsymbol{C}_y)$ is used to calculate numerically the desired number of cumulants for each of the element distributions.
6. Using the results from step (5), the matrices $\langle\mathbf{A}(\boldsymbol{C})\rangle$ and $\mathbf{d}\langle\mathbf{A}(\boldsymbol{C})\rangle$ are constructed.
7. The population model is then approximately represented in the form of Eq. (26).

8. When the uncertainty of multiple-year life-table quantities are required, the matrix products $\Pi_t\langle \mathbf{A}_y(\boldsymbol{C}_y)\rangle$ and $\Pi_t \mathbf{d}\langle \mathbf{A}_y(\boldsymbol{C}_y)\rangle$ are calculated.
9. The product $\Pi_t\langle \mathbf{A}_y(\boldsymbol{C}_y)\rangle$ gives the correlation constant for the population model and the product $\Pi_t \mathbf{d}\langle \mathbf{A}_y(\boldsymbol{C}_y)\rangle$ gives the random component.
10. The delta method can be used to calculate the mean and variance of derived life-table quantities like population residence time.

It can thus be seen that the distribution of the elements of the matrix $\mathbf{A}_y(\boldsymbol{C}_y)$ is sufficient for defining the uncertainty characteristics of any life-table quantity in the multi-state population model. Just as for the pseudorates, the exponential component of elements of $\Pi_t\mathbf{A}_y(\boldsymbol{C}_y)$ will show additivity in the eigenvalues, so it may be possible to achieve some simplifications in the above methods (e.g. calculating only the distributions of the eigenvalues of $\boldsymbol{C}$). However, for the purposes of simulation it will usually be just as convenient to calculate the full expression for the components of $\Pi_t\mathbf{A}_y(\boldsymbol{C}_y)$.

Simulation results

Regarding our conjecture concerning the distributional form of the elements of $\mathbf{A}_y(\boldsymbol{C}_y)$, we stated above that it should be either approximately normal or approximately log-normal. This is indeed what is found in simulations based on population models such as the one presented in (Lauer et al., 2003). In fact, these simulations suggest that elements of $\mathbf{A}(\boldsymbol{C})$ show a strong tendency overall to converge towards an approximately normal distribution (Figure). The histograms of the elements of $\mathbf{A}_y(\boldsymbol{C}_y)$ display a bell-shaped curve well fit by the normal density, that shows either negligible skewness or a slight log-normal skewness).

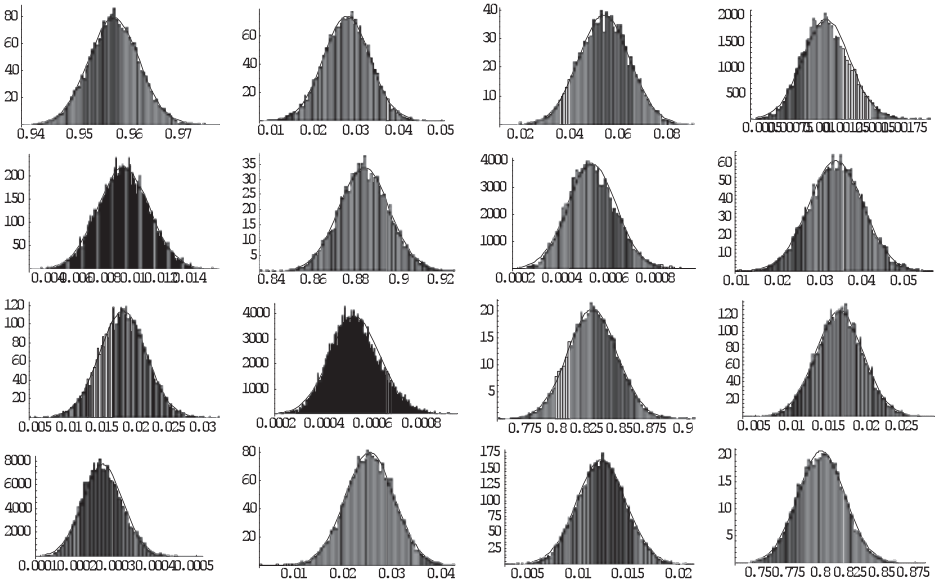


Figure: Histograms of elements of **A**(***C***) for the population model in [Lauer et al., 2003], showing the corresponding line of normal density. The distributions were obtained from 10,000 simulations where the elements of ***C*** are independent and normally distributed.

Notice, however, that the distribution of anti-diagonal elements {(*i*, *j*): (1, 4), (2, 3), (3, 2), (4, 1)} in the array are noticeably more skew (i.e. more log-normal) than the distribution of other elements, which are only negligibly if at all skew. This pattern has a meaning in terms of the population model in (Lauer et al., 2003), since the anti-diagonal elements represent transitions that cannot be made in a single step, that is {($j \rightarrow i$): ($4 \rightarrow 1$), ($3 \rightarrow 2$), ($2 \rightarrow 3$), ($1 \rightarrow 4$)}, but only as a result of a series of two or more steps.

Thus, the shortest path that implements the transition {($4 \rightarrow 1$) = (XC $\rightarrow$ S)} is either the path {(XC $\rightarrow$ C)} $\cup$ {(C $\rightarrow$ S)} or the path {(XC $\rightarrow$ X)} $\cup$ {(X $\rightarrow$ S)}. Similarly, the shortest path implementing the transition {($3 \rightarrow 2$) = (X $\rightarrow$ C)} is either the path {(X $\rightarrow$ XC)} $\cup$ {(XC $\rightarrow$ C)} or the path {(X $\rightarrow$ S)} $\cup$ {(S $\rightarrow$ C)}. Since the probabilities of making a double or higher-order transistion always involve the product of at least two probabilities, it is reasonable to expect that the distributions of the probabilities of these transitions should be more log-normal.

The same is true of the diagonal probabilities, which represent the probability of starting in a state and ending in that same state in the same interval. The relevant paths also contain at least two steps. Taking the logarithm of the distributions from the simulations shows that, for all distributions other than the diagonal and anti-diagonal distributions, the logarithmically transformed distributions are distorted whereas the transformed diagonal and anti-diagonal distributions appear symmetric. This suggests that the diagonal and anti-diagonal distributions have a more pronounced tendency to converge to log-normal distributions than the other distributions.

These findings suggest the possibility of making the following emendation to the simulation procedure, above. After step (4), insert:

4a. If the distributions of any of the elements of $\mathbf{A}_y(\boldsymbol{C}_y)$ are noticeably skew, take the logarithm of the values and calculate the cumulants of this transformed distribution (where a low-order truncation of the cumulant expansion will be more justified). Exponentiate to recover the original distribution.

Conclusions

The problem of determining the distributional features of $\boldsymbol{p}(t)$ in a multi-state longitudinal population model can be seen to reduce to the problem of determining the distribution of elements of $\mathbf{A}(\boldsymbol{C})$ when $\boldsymbol{C}$ is a random matrix. The demographic or statistical literature on Markov models does not appear to address the question of random elements in the operator matrix $\boldsymbol{C}$. The indicative results we present here are nevertheless a straightforward extension of the linear algebra theory that underlies most multi-state modelling theory.

References

1. Caswell H. *Matrix population models: construction, analysis, and interpretation*. Sunderland, Massachusetts, Sinauer Associates, 2001.
2. Franke J, Härdle W, Hafner C. *Statistics of Financial Markets*. Berlin, Springer, 2004.
3. Gardiner CW. *Handbook of stochastic methods for physics, chemistry and the natural sciences*, 3rd ed. Berlin, Springer, 2004.
4. Goldberger AS. A course in econometrics. Cambridge, Massachusetts, Harvard University Press, 1991.
5. Kolassa JE. *Series Approximation Methods in Statistics*, 3rd ed. Springer, 2006.
6. Lauer JA, Roehrich K, Wirth H, Charette C, Gribble S, Murray CJL. PopMod: a longitudinal population model with two interacting disease states. *Cost effectiveness and resource allocation*, 2003, 1:6.
7. McCullagh P. *Tensor methods in statistics*, London, Chapman and Hall, 1987.
8. Singer B, Spilerman S. The representation of social processes by Markov models. *The American journal of sociology*, 82(1):1-54, 1976.
9. Gilbert Strang. *Linear algebra and its applications*, 3rd ed. Harcourt Brace Jovanovich, 1988.
10. Van Kampen NG. *Stochastic processes in physics and chemistry*. Amsterdam Elsevier, 1992.
11. Weigel FW. *Introduction to Path-Integral Methods in Physics and Polymer Science*, Singapore, World Scientific Publishing Company, 1986.
12. Wu J, Mehta NB, Zhang J. Flexible Lognormal Sum Approximation Method, IEEE Global Telecommunications Conference (GLOBECOM), Vol. 6, pp. 3413-3417, November 2005 (IEEE Xplore).

Appendix II

Modelling complex interactions in longitudinal population models

Introduction[1]

Generic population health models [1] offer the possibility of easy-to-use longitudinal models that can be used for population health measurement. In many cases, generic models can be implemented with generally available epidemiologic and demographic data and can be used to quantify real population health risks. In spite of these possibilities, analysts who are faced with the need to model complex disease processes, such as those presented by many chronic diseases, for example, may find that four health states are not sufficient for a decent level of realism. Typically chronic disease processes can involve an arbitrary number of disease stages accessed in a progressive, non-reversible fashion. Or, if an analyst is interested in modelling multiple risk factors and disease states, the factor limiting the realism of the model may not be the number of states available in the model but rather the number of interactions between risk factors, in other words, the need to model combinations of risk factors requiring the definition of many more than four states. For such situations, four different health states is far too few and the use of generic models [1] will therefore ultimately be unsatisfying for some population health problems and a greater level of heterogeneity will be sought. Although different approaches are possible, and many of them will share a similar set methods for addressing the problem we describe, microsimulation is one natural approach. With microsimulation methods, one can readily investigate questions of multiple interacting risk factors and causes of disease (i.e. "causal web"). However, the technology of microsimulation does not change the fundamental need to characterize a joing distribution from a set of observations of marginal distributionsw. Moreover, the question of characterizing in quantitative terms the interactions in the causal web becomes important. Thus, the kinds of data transformation that are required to define the initial state of the population simulation can be more complex. Several generic insights are nevertheless possible.

When one wishes to model several population states in a longitudinal

1 I am grateful to Claude Charette and Steve Gribble of Statistics Canada for their assistance in helping me to start thinking about the problem discussed in this appendix and for their having worked through a complete example.

population model, lack of information about the joint distribution of risk (either longitudinal, in terms of rates, or cross sectional, in terms of prevalences) is a common problem. Even for the simple case of estimating the effect of a risk factor within a given strata of the population, in the absence of knowledge about the joint distribution of risk factors within the stratum, the estimation of effects is – in most cases – hopelessly confounded [2]. Of course, if risk distributions were statistically independent, the problem would be trivial since the joint distribution would merely be calculated as the product of the marginal distributions. Yet departures from independence are often precisely what make the joint distribution of risk of interest [2]. So determining the joint distributions of non-independent risk factors is not only an ubiquitous but also an important problem.

These questions can be related more generally to concepts of "interaction" as discussed in the epidemiological literature [3-22]. However, in the mainstream literature, interaction is typically defined with respect to rates, relative rates or incidence proportions [3]. In any case, since rates, relative rates and relative incidence proportions are measured longitudinally, one can conclude that the kind of interaction typically discussed in the epidemiological literature is what could be called "longitudinal interaction" (i.e. related to interaction in rates). Despite its similar importance, the problem of determining the joint distribution of risk factors and stratification variables in cross section typically goes underrecognized, although it has been mentioned as a source of unavoidable estimation bias [19,21].

In either the longitudinal or the cross-sectional case, however, it is important to be modest in trying to relate statistical concepts of dependence and independence to biological processes of interaction [6]. It is important to keep in mind that, among other problems in moving from a statistical to a biological concept, that a variety of different biological interactions can produce the same statistically observed outcomes[13]. Recognizing the inherent gulf between biological and statistical concepts, we develop here a pragmatic approach to "interaction" that is defined primarily in statistical terms. Our concept will therefore be mostly agnostic about the true underlying biological mechanisms, although it will occasionally be possible to rule out certain biological mechanisms as inconsistent with the observed data.

Moreover, in the approach presented here, headway can be made on both the problem of longitudinal and cross-sectional interaction through the application of a single conceptual framework and approach. Indeed, since the current cross-sectional distribution of exposure is merely the cumulative result of past events, it should in principle be possible to approach the problem of the joint prevalence of risk factors (i.e. static, or cross-sectional, interaction) as well as the decomposition of aggregate rates (i.e. dynamic, or longitudinal, interaction) with a common set of methods.

For both conceptual clarity and measurement purposes, however, it is useful to distinguish dynamic and static concepts of interaction in the following exposition, where we define a minimal set of necessary and sufficient information to determine the joint distribution in the static case as well as solve the decomposition of aggregate rates. We specify that the "joint distribution" can be understood to refer either to the joint distribution of persons or of person time and that the methods employed do not depend on one or the other distributional concept, although it is of course always useful to use a consistent set of definitions in the course of a particular problem. It is also useful to recognize, at least as an approximation, that the concepts of prevalence, rates and odds can be referred to together, as necessary, by means of the common term "risk" [3] without damage to either common sense or the main conclusions of the argument.

Static problem

To fix ideas, we refer to a case of stratification variables and risk factors, although the actual boundaries between the two concepts are fluid. Typical stratification variables might be age, sex, income or educational status, whereas typical risk factors are usually the exposures of primary interest in the question at hand. However, these designations are mainly arbitrary since risk factors will typically become stratification variables once they are well known. In any case, suppose there are K strata for the stratification variable $\underline{M} = \{M_k\}$, $k = 1, \ldots, K$, and J levels of risk factor $\underline{X} = \{X_j\}$, $j = 1, \ldots, J$. It should be clear that we refer here to the partition of the population into a 2-way discrete classification.

We may start by noticing that, in such a case, it will always be possible to define the joint distribution of risk within strata as the solution of a system of $K \times J$ independent equations. For the problem to be well posed, these equations, whatever their actual positive content, must be written in terms of a vector of the unknown joint prevalences $\boldsymbol{p}$, a matrix of known coefficients $\boldsymbol{A}$ and a vector of observable constants $\boldsymbol{b}$. The observable constants $\boldsymbol{b}$ will typically be either the marginal prevalences or zeros (discussed below), and the coefficients $\boldsymbol{A}$ will typically be either known relative risks or other known constants (discussed below). We return to this in more detail subsequently. In any case, based on a suitable identification of the different elements of the system, we then write $\boldsymbol{A} \cdot \boldsymbol{p} = \boldsymbol{b}$. It is well known that whenever the matrix $\boldsymbol{A}$ is of full rank, the system has a unique solution for $\boldsymbol{p}$.

To define solution conditions that are consistent with the empirical problem, it is reasonable to assume that the marginal distributions $\boldsymbol{b}$ of both X and M are known. The marginal distributions can be thought of here as the population-level risk of (i.e. the probability of finding) the members of the different risk-factor and stratification-variable levels based on uniform random sampling from the entire population. Actual observations of $\boldsymbol{b}$ will typically not strictly fulfill these conditions but rather approximate them.

These marginal prevalences, which we assume are observed, define $J + K - 1$ independent summing-up equations. That is, for each marginal prevalence, we can always write K equations of the form $m_i = \sum_{j=1}^{J} p_{i,j}$, as well as for each risk factor J equations $x_j = \sum_{i=1}^{K} p_{i,j}$. Note that $p_{i,j}$ are unknown in both cases When written in matrix form, these $J + K - 1$ equations define a submatrix of $\boldsymbol{A}$ whose coefficients are all ones (1s). Note however that although we could in principle write $J + K$ equations, only $J + K - 1$ of the equations are actually independent, since one of the equations can always be expressed in terms of a linear combination of the others (i.e. by summing across either risk-factor or stratification-variable levels, the total population is recovered: $1 = \sum_{j=1}^{J} p_{i,j} = \sum_{i=1}^{K} p_{i,j}$).

Since the observed marginal prevalences yield only $K + J - 1$ independent equations in terms of the joint prevalences, $(K \times J) - (K + J - 1) = (J - 1) \times (K - 1)$ additional equations are required in order for the system to be solvable. This number $(J - 1) \times (K - 1)$ is evidently equal to zero only when K or $J = 1$.

We propose that, in general, it will be reasonable to characterize these $(J-1) \times (K-1)$ additional equations in terms of relative risk relations. The advantage of relative risk measures for this purpose are several. First, and most importantly, relative risk provides a concise way to define interaction in statistically observable terms. Second, estimates of relative risk as defined below are typically available from case-control studies. Third, it will often be possible to make reasonable guesses about the magnitude of a given relative risk even in the absence of data. Finally, using relative risk measures such as those defined below to construct the required additional equations to complete the system $\boldsymbol{A}\cdot\boldsymbol{p} = \boldsymbol{b}$ does not for the moment require any assumption about the functional form of relative risk (e.g. additive, multiplicative etc.). We return to this topic when we discuss dynamic interaction, below.

As is typical, we define relative risk as a ratio of conditional risk

$$\text{Eq. (1)} \qquad \wp_{k,1} \equiv \frac{p_k(X = x_j)}{p_1(X = x_j)},$$

where p_k represents the probability of finding an exposure unit (person or person-time) in stratum level k conditional upon a specific exposure status (e.g. $X = x_j$). We say that when there is interaction between levels of the risk-factor and the stratification-variables $\wp_k \neq 1$. Subsequently, we define more precisely what kind of conditional risk measures are required to solve $\boldsymbol{A}\cdot\boldsymbol{p} = \boldsymbol{b}$.

Suppose, for example, that, for each risk-factor level j, an equation of the form

$$\text{Eq. (2)} \qquad \wp_k = \frac{p_k(X = x_j)}{p_1(X = x_j)}$$

can be written, where $X = x_j$ signifies that there is exposure to risk factor level j. This represents the typical sort of measure that is available from case-control studies. Suppose for the moment that there are two stratification-variable levels ($k = 1, 2$), one risk factor level, one baseline, or non-exposed, level, and that the interaction of risk-factor level j with stratum 2 is written $\wp_{2,j}$

$$\text{Eq. (3)} \qquad \wp_{2,j} = \frac{p_{2,j}/m_2}{p_{1,j}/m_1},$$

where the marginal prevalence of measurement units in stratum 1 is given by m_1, the marginal prevalence of measurement units in stratum 2 is given by m_2 and we write $p_{2,j}$ and $p_{1,j}$ for the unknown joint prevalence of risk-factor level j in stratum-level 1 and 2, respectively. Rearranging the latter equation for the relative risk, we can then evidently write

Eq. (4) $$p_{1,j} m_2 \wp_{2,j} - p_{2,j} m_1 = 0 \,,$$

which is a linear equation in terms of the unknown joint prevalences $p_{2,j}$ and $p_{1,j}$. On the basis of the observation of the value $\wp_{2,j}$, such an equation can be added to the $(K + J - 1)$ equations that were previously written in terms of the summing-up constraints for the marginal prevalences.

Since, when the marginal prevalences are known (which is our assumption), it is always possible to write the summing-up constraints, and since each relative risk $\wp_{i,j}$ that we know defines one additional equation in terms of a subset of the joint prevalences $p_{i,j}$ and some of the marginal prevalences m_i and x_i, it will always be possible to solve for the full set of joint prevalences whenever we know the value of a set of $(J-1) \times (K-1)$ complete and independent relative risks. The equations are shown in matrix form for the $J = K = 3$ case in the Figure. The solution can be obtained by matrix algebra or by Gaussian elimination (i.e. repeat substitution).

Figure: The matrix equations for the J = K = 3 case.

$$\begin{pmatrix} m_1 & -m_3\rho_{3,b} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -m_2\rho_{2,b} & m_1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & m_1 & -m_3\rho_{3,c} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -m_2\rho_{2,c} & m_1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} p_{3,b} \\ p_{1,b} \\ p_{2,b} \\ p_{3,c} \\ p_{1,c} \\ p_{2,c} \\ p_{1,a} \\ p_{2,a} \\ p_{3,a} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ x_b \\ m_1 \\ m_2 \\ m_3 \\ x_c \end{pmatrix}$$

What is an "independent" or "complete"relative risk in this context? Each relative risk relates the conditional prevalence of a single stratum-risk pair with that of another stratum-risk pair. Independence simply requires that there are no redundant stratum-risk pairs among the available set of observed relative risks. Completeness, which is in fact satisfied whenever there are $(J-1) \times (K-1)$ independent relative risks, requires that all existing strata and risk levels can be mutually related, for example, for four strata, three stratum-risk pairs are required to mutually relate all four strata. For example, if the strata are labeled i, j, k, and l, any of the sets of strata pairs $\{i \rightarrow j,\ j \rightarrow k, k \rightarrow l\}$, $\{i \rightarrow j,\ j \rightarrow k, j \rightarrow l\}$ and $\{i \rightarrow j,\ j \rightarrow k, i \rightarrow l\}$ are complete. A complete set of strata pairs is required for each risk level except for "baseline" risk since the relative risk of the "last" risk factor level can always be expressed as in terms of the other relative risks.

So far we have discussed the case of a single risk factor with J levels $\underline{X} = \{X_j\}$, $j = 1, \ldots, J$ and a single stratification variable with K levels $\underline{M} = \{M_k\}$, $i = 1, \ldots, K$. However, when there are additional risk factors or stratification variables that are also of interest, provided there are enough additional relative risks estimates available, the approach described above can be iteratively applied to find the solution to the joint prevalence. To begin, we notice first that additional risk factors (or stratification variables) must be either interacting or non-interacting with the initial pair of risk factor and stratification variable. If the additional variables are non-interacting, a new joint distribution can easily be obtained by multiplying each element of the joint distribution solved for in the first step by each element of the marginal distribution of the new risk factor (or stratification variable).

However, if the new risk factor or stratification variable is interacting, it must interact either with the existing risk factor, with the existing stratification variable or with both. If it interacts with only one of them, the process of writing summing-up constraints and relative-risk equations can be repeated with that one. If, on the other hand, it interacts with both of them, then one of the interactions can be chosen first. The new variable must interact either with levels, or with aggregates of levels, of the marginal distribution of the chosen variable. On the basis of that interaction (i.e. using the relative risk observations that are available), one then write a

new system of the form $\boldsymbol{A}\cdot\boldsymbol{p} = \boldsymbol{b}$, which will again be solvable whenever the matrix $\boldsymbol{A}$ can be written so as to be of full rank. The same process can then be repeated for the other interaction.

Provided the conditions for the existence and uniqueness of a solution are satisfied for each step (i.e. the matrix $\boldsymbol{A}$ is of full rank), this process can be repeated an arbitrary number of times. This is true even when none of the joint prevalences that are sought are actually known to start with. Moreover, because of considerations of symmetry, regardless of which risk-factor/stratification-variable pair is chosen as the starting point (assuming there are several alternatives), the end-point of the iterated solution process must be the same (i.e. in practice, the solution is constrained by the available data on marginal distributions and relative risk).

One may futher observe that, due to measurement error or bias in measurement, the supposedly known quantities can in fact be uncertain. As a result, the mapping from the space of the known quantities (marginal prevalences) to the space of the joint prevalences may be subject to uncertainty. In this case, a notional uncertainty analysis can proceed as follows. Suppose error in the known quantities is normally distributed. Since the solution vector of joint prevalence is, according to matrix theory, a linear combination of the known quantities, its elements will also tend to have normal distributions. This tendency will be more pronounced the more independent known quantities are represented in the solution expressions.

Dynamic problem

As noted earlier, the epidemiological literature has typically been concerned what we have termed dynamic interaction (interaction in rates).

Interaction hypotheses

In the static case, we have discussed interaction without regard to its functional form (cf. Eq. 3, which depends on the magnitude of interaction but not on its specification). This sort of measure is non-parametric in the sense that whatever kind of interaction is present, it is captured in the data that are observed. This type of assumption is consistent moreover with the sort of observations of static interaction that are typically available

from case-control studies. In the case of dynamic interaction, however, considerable attention has been devoted to a number of different canonical forms of interaction and their implications [3,11,14-16].

Additive models

For example, a simple additive model for rates has been defined as follows [3]

$$\text{Eq. (5)} \qquad r_k(\underline{x}) = \alpha_k + \underline{x\beta} \,,$$

where $\underline{x}$ is the vector of exposures, k indexes the stratum and β is a vector of effect measures expressing the change in risk per unit increase in the exposure $\underline{x}$. Since $r_k(\underline{0}) \equiv \alpha_k$, the additive model implies (or can be derived from) additivity of the rate difference

$$\text{Eq. (6)} \qquad D_k(x) = r_k(\underline{x}) - r_k(\underline{0}) = \underline{x\beta} \,.$$

Since the model of additive rate-different is synonymous with the hypothesis of "no causal interaction" under Rothman's sufficient-component cause model [3], the lack of causal interaction can be taken as a definition of the simple model of additive rate difference, and a finding of additive rate difference can be taken as evidence if favour of the absence of causal interaction.

Another model typically referred to as "additive" is the additive relative rate model. The rate model that corresponds to the additive relative-rate model can be written as follows [14,16]

$$\text{Eq. (7)} \qquad r_k(\underline{x}) = \alpha_k(1 + \underline{x\beta^*}) \,,$$

where β^* is a vector of constant effect measures expressing the change in relative risk per unit increase in $\underline{x}$ (note that this is a different measure of effect than that which is relevant to the model of addtive rate difference, above). The additive relative rate model implies that excess relative risk, defined as

$$\text{Eq. (8)} \qquad E_k(\underline{x}) \equiv \frac{r_k(\underline{x})}{r_k(\underline{0})} - 1 = \underline{x\beta^*}$$

is also additive. The additive relative rate model is also a proportional risk model [3], since, as can be seen from the definition of the model, stratum-specific exposure effects differ only by a constant of proportionality α_k. The

additive relative rate model is therefore is not a model of non-interaction (as has sometimes been claimed) but rather a restriction on interaction to a particular multiplicative form (i.e. constant of proportionality α_k).

A less restricted additive model is the following [3]

Eq. (9) $$r_k(\underline{x}) = \alpha_k + \underline{x\beta_k}\ .$$

Here, the effect of exposure is stratum specific, and interaction is not limited to a particular mathematical specification. Such stratum-specific interaction would therefore need to be observed. This less restricted additive model can also be written in terms of an additive relative (or excess) risk model

Eq. (10) $$r_k(\underline{x}) = \alpha_k + \underline{x\beta_k} = \alpha_k(1 + \frac{x\beta}{\alpha_k}) \equiv \alpha_k(1 + \underline{x\hat{\beta}_k})\ .$$

Here, each element β_k is divided by the corresponding α_k and renamed $\hat{\beta}_k$, yielding an additive relative risk model with stratum specific coefficients. This is in fact the specification in which additivity of risk is usually tested for in studies [3].

In every additive model but the simple additive model risk-difference model of no interaction, therefore, the assumed additivity relation in underlying rates, risk or odds can be represented in terms of a corresponding additivity relation in relative risk, relative rates or relative odds. Moreover, the different additive models represent a spectrum of (additive) specifications for interaction: the least restricted model of Eq. 9 allows stratification variables to interact with risk factors in a non-specified manner; the model of additive relative risk allows risk factors to interact with stratification variables but limits the interaction to constant factors α_k; finally, the additive risk model implies no interaction between stratification variables and exposures [3].

Equivalence between additive models

The simple additive model for rates (Eq. 5) is equivalent to an additive model for relative rates (Eq. 7) only if [3]:

— there is only one stratum, or
— the baseline rate $r_k(\underline{0})$ is constant across strata (i.e. $r_k(\underline{0}) = r(\underline{0})$).

However, we have shown that is always possible to solve for unique values for $r_k(\underline{x})$ when $J + K$ marginal prevalences are known and there are $(J-1) \times (K-1)$ independent known relative risks. We now show that, in general, the existence of such a solution will almost always require baseline risk to differ across strata, implying that the additive model for rates is almost never equivalent to the additive model for relative rates.

The stratum-specific (i.e. joint) risk, rate or odds that are the solution of the matrix equation correspond to a basis vector in the space of the joint prevalences. Moreover, precisely because the solution vector for the joint prevalences is a basis in the space of joint prevalence, each solution element is necessarily algebraically different (since if they were the same, one element would be redundant and, contrary to the construction of the problem, the basis would then not span the space of joint prevalence). We can nevertheless ask under what conditions these algebraically different solution elements may be numerically equal.

The conditions for this can be determined by setting the different solution expressions for baseline risk equal, $r_{i,0} = r_{j,0}$ for all possible i,j pairs, where i and j index the stratum levels. For K levels, the number of possible i,j pairs where $i \neq j$ can be written compactly as the binomial coefficient $\binom{K}{2}$. When the relative risks are independent and complete, solutions for baseline risk will moreover necessarily be expressed in terms of all $J \times K$ known quantities for marginal risks and relative risks. The baseline risks cannot be expressed in terms of fewer than $J \times K$ known quantities because solvability of the system requires each of the known quantities and no fewer. Each of the $\binom{K}{2}$ equalities that need to be written to enforce equality of baseline risk across strata is therefore an expression written in terms of each of the $J \times K$ known quantities. Since each equality can be solved in terms of one of the $J \times K$ known quantities, each of the $\binom{K}{2}$ equality constraints induces a corresponding equality constraint on each of the $J \times K$ other known quantities it contains.

Since there are $\binom{K}{2}$ equalities that can be solved for a given known quantity, these in turn induce $\left\{\binom{K}{2}\right\}$ pairwise equalities in terms of the remaining $J \times K - 1$ known quantities. From this group of pairwise constraints, another

one of the $J \times K - 2$ remaining known quantities can then be solved for. However, this solved-for quantity now has to satisfy $\left\{\left\{\left(\begin{smallmatrix} K \\ 2 \end{smallmatrix}\right)\right\}\right\}_{2}$ constraints expressed in terms of the remaining $J \times K - 3$ known quantities. Proceeding in this fashion, it can be seen that the binomial coefficient $\binom{K}{2}$ must be nested $J \times K$ times, in other words, until all the known quantities except for the last one have been expressed in terms of all the others. For the case of 4 strata and 4 known quantities, the first known quantity therefore has to satisfy 6 constraints, the second 15, the third 105 and the last 5460. However, since the order of choosing the known quantities is arbitrary, symmetry requires that each of the known quantities should satisfy 5460 additional constraints.

However, since each such additional constraint represents a linear combination of the solution values for the joint risk, rate or odds, even one such additional constraint is too many for the known quantities to satisfy. This is because the only combinations of the solution values for joint risk, rate or odds that also are combinations of the known quantities are those specified in the system of equations $\boldsymbol{A} \cdot \boldsymbol{p} = \boldsymbol{b}$, which was solved to obtain the joint prevalences. In fact, the matrix $\boldsymbol{A}$ from that system of equations is guaranteed by construction (i.e. by the full rank condition) to represent a unique transformation that maps elements from the space of the joint prevalences to the space of the known quantities.

If any single additional linear combination was in fact possible, it would be equivalent to adding an extra row to the original matrix equation. However, since the original matrix equation is the unique transformation that maps the joint prevalences into the known quantities, two possibilities arise: either the additional row is linearly dependent on the other, original, rows, in which case the matrix is singular (and the "additional" constraint is therefore implicit in the original system and is consequently not an additional constraint), or the additional row is independent (and the matrix is therefore inconsistent). In either case, the full rank condition is violated and a solution to the matrix equation is impossible. Therefore, any additional constraints on the known quantities in terms of the joint solved-for prevalences are not possible.

There is nevertheless one case where the condition of equality of baseline risk does not impose any additional constraints on the known quantities. That is the case of two strata only (regardless of the number of risk levels). Since $\binom{2}{2} = 1$, equality of baseline risk across the two strata can induce only a single equality constraint on the joint solution quantities, but, since $\binom{\binom{2}{2}}{2} = 0$, no equality constraints are in turn induced on the marginal prevalences).

It has already been noted in [3] that the additive relative risk model is equivalent to the additive risk model only when there is only one stratum or when baseline risk is equal across strata. The arguments presented here provide an important restriction on the second possibility, namely, that it is only possible in principle for baseline risk to be equal across strata when there are only two strata.

The additive relative risk model can therefore only be equivalent to the simple additive risk model of no interaction when there are at most two strata (and it is by no means guaranteed to hold in the two-strata case). In any other case, the additive risk model cannot be equivalent to the additive relative risk model.

Multiplicative model

A multiplicative model for risk is typically written [21]:

$$\text{Eq. (11)} \qquad r_k(\underline{x}) = \exp\beta(\underline{x}) \cdot r_k(\underline{0}),$$

from which it can be seen that relative risk can be written:

$$\text{Eq. (12)} \qquad R_k(\underline{x}) \equiv \frac{r_k(\underline{x})}{r_k(\underline{0})} = \frac{\exp\beta(\underline{x}) \cdot r_k(0)}{\exp\beta(\underline{0}) \cdot r_k(0)} = \frac{\exp\beta(\underline{x})}{\exp\beta(\underline{0})}.$$

Since by Eq. 11, $\exp\beta(\underline{0})$ must be equal to 1, it follows that

$$\text{Eq. (13)} \qquad R_k(\underline{x}) = \exp\beta(\underline{x}),$$

which is a multiplicative model for relative risk. The multiplicative relative risk model is of considerable interest in practical work since estimates of relative risk obtained using such a model are interpretable as standardized parameters, i.e. parameters standardized to the risk distribution in the study population from which they were obtained [21].

Multiplicative versus additive models of interaction

While it has been claimed that additivity of risk, rate and odds remains the "most basic manifestation of independent action" [6], empirical additivity of rates does not by itself imply the absence of interaction [3] and it is therefore a necessary but not sufficient condition for a definite claim about the absence of interaction. Interaction, moreover, as we have discussed above, can have different specifications, including additive and multiplicative relative rate specifications.

In any case, for certain simple biological models of interaction, it is in general possible to say whether an additive or multiplicative model is preferred. For example, in simple multi-stage models model derived from studies of carcinogenesis, it can be shown that two risk factors acting on the same stage of pathogenesis interact in an additive fashion, whereas risk factors that act on different stages interact multiplicatively [10, 12].

For example, a simple carcinogenesis model posits that a single cell, by undergoing a series of transformations, results in a cell capable of malignant replication; such a model is consistent with the well known power-law relationship for cancer incidence [10]. Since noxious exposures increase the probability of transformations, when they facilitate separate stages of the transformation to malignancy their joint effect combines multiplicatively, but when they facilitate the same stage, the joint effect combines additively. A multi-stage multiplicative model has solid empirical support for non-hormonally dependent cancers of epithelial origin [10].

Therefore, when pathogenesis is multi-factorial and exposures affect different stages of disease, a multiplicative model appears to be the canonical model, whereas when exposures affect the same stage of disease occurrence, an additive model may be canonical. Whether the additive model is an additive rate or risk model will depend on the formulation of the underlying disease model, i.e. in terms of either rates or relative rates and, as we have shown, an addtive rate model will almost never be an additive relative rate model. In any case, the additive relative rate model is of some interest because of its use as a common model to test for additivity.

Finally, additive and multiplicative relative rate models cannot be equal: each specification has different quantitative implications for outcomes. In particular, a multiplicative model of interaction will in general result in a higher rate of disease than any additive model.

Since "it is difficult to imagine situations in which typical stratifying factors … would be unassociated with any causal cofactors" of the exposures $\underline{x}$, we propose that the additive relative rate model and the multiplicative relative rate model may be representing two different possible forms of interaction that respond to different possible models of biological interaction in multi-stage disease causation. As we shall see in the following section, this characterization offers the additional advantage that both cases can be characterized in terms of relative risk only.

Rate decomposition

Now we consider transition events, in particular from disease-free to diseased status. Transition events are modelled by rates, and are therefore fundamentally longitudinal. Here, to fix ideas, we focus on the case of hazards, i.e. instantaneous rates.

The total population rate of transition to disease y is equal to the per-capita rate of transition, r^y, multiplied by the prevalence of the group without disease y, p^0 (i.e. the population prevalence of the group capable of experiencing the transition), in other words, the quantity $r^y \cdot p^0$. We write p with superscript 0 to emphasize the prevalence of the group without disease y, and we write r with superscript y to emphasize the rate of transition to y, respectively. Suppose that the disease-free population is furthermore divided into subgroups by risk-factor and stratification-variable status according to the methods of the previous section regarding the static problem. The total population transition rate $r^y \cdot p^0$ must then necessarily be equal to the sum of the products of per-capita rates in the different subgroups and the prevalences of the respective subgroups. Suppose that the subgroups are denominated by subscripts denoting stratum/risk-factor levels and that we write superscripts to emphasize status with respect to disease y. When there are two strata (1, 2) and two risk-factor levels (a, b), we can then write

$$\text{Eq. (14)} \qquad r^y \cdot p^0 = r^y_{1,a} \cdot p^0_{1,a} + r^y_{2,a} \cdot p^0_{2,a} + r^y_{1,b} \cdot p^0_{1,b} + r^y_{2,b} \cdot p^0_{2,b} .$$

By the methods of the static problem, above, the prevalence values in an equation of this type are assumed to be known. Since it is the sort of quantity which is in principle observable, we further suppose here that the total population rate, $r^{y} \cdot p^{0}$, is also known. However, since there are $K \times J$ unknown subgroup rates, for the moment, we have only one equation with $K \times J$ unknowns. In the absence of $(K \times J) - 1$ additional independent equations, it is clearly impossible to solve for the unknown subgroup rates. This is a substantially worse situation than that which was faced in the static problem, where we required only $(K - 1) \times (J - 1)$ additional equations, which we supposed were available in the form of relative risks. In other words, here we need $K + J - 2$ more equations than we needed to solve the static case. What can we do to make the information demands in the dynamic case in practice no greater than they were in the static case?

One way is to suppose that we know only a relatively small number of relative rates but that we know that relative rates follow a particular mathematical form, such as additive or multiplicative. In such a case, we can then solve for the different subgroup rates. In other words, if we know only $(K - 1) + (J - 1)$ independent relative rates *and* we know a rule $f(.)$ for combining relative rates, we can easily derive the additional relative rates required to make the system solvable. To continue the dynamic example we have started, above, if we know the following two relative rates

Eq. (15) $$\wp_{1\to 2,a} = \frac{r_{2,a}}{r_{1,a}}, \wp_{1,a\to b} = \frac{r_{1,b}}{r_{1,a}},$$

then we can easily derive the third required relative rate

Eq. (16) $$\wp_{1\to 2,a\to b} = \frac{r_{2,b}}{r_{1,a}} = f(\wp_{1\to 2,a}, \wp_{1,a\to b}) \cdot$$

If relative rates are multiplicative, the derived relative rate is calculated

Eq. (17) $$\wp_{1\to 2,a\to b} = \wp_{1\to 2,a} \cdot \wp_{1,a\to b},$$

but if relative rates are additive, the derived relative rate is calculated

Eq. (18) $$\wp_{1\to 2,a\to b} = (\wp_{1\to 2,a} - 1) + (\wp_{1,a\to b} - 1) + 1.$$

Following the approach for the static case, with these three relative rates it is possible to write enough equations in the four unknown subgroup-specific rates to solve for all the subgroup rates.

While in general it is impossible to say with any confidence what is the correct relative risk law to apply, broad statements of the following type appear to be not unreasonable:

— The additive rate model is unlikely to apply since there will in general be interaction (i.e. the statistical association of stratification and exposure variables).
— Interaction can be characterized according to two canonical forms in terms of relative rates (additive and multiplicative).
— Each form has specific implications for outcomes, and each can be related with profit to a specific part of the epidemiological literature.
— The multiplicative form for interaction of relative risk implies a larger subgroup risk than the additive form of interaction.
— The multiplicative form of interaction has the interpretation that exposures operate at different stages of pathogenesis, or, in other words, through different causal pathways.

References

1. Lauer JA, Roehrich K, Wirth H, Charette C, Gribble S, Murray CJL. PopMod: a longitudinal population model with two interacting disease states. *Cost effectiveness and resource allocation*, 2003, 1:6.

2. Greenland S. Ecologic versus individual-level sources of bias in ecologic estimates of contextual health effects. *Int J Epidemiol* 2001;**30**:1343-50.

3. Greenland S. Additive risk versus additive relative risk models. *Epidemiology*. 1993; 4:32-36.

4. Walker AM Rothman KJ (1982) Models of varying parmatric form in case-referent studies. *Am J Epidemiol* 115, 129-137.

5. Greenland S. Basic problems in interaction assessment. *Environmental health perspectives supplements*, 101, supplement 4, 59-66.

6. Greenland S, Poole C. Invariants and non-invariants in the concept of interdependent effects. *Scan J Work Environ Health*, 1988;14:125-129.

7. Rothman KJ, Greenland S, Walker AM. Concepts of interaction. *Am J Epidemiol*, 1980; 112:467-470.

8. Miettinen OS. Causal and preventive interdependence: elementary principles. *Scan J Work Environ Health*, 1982;8:159-168.

9. Weinberg CR. Applicability of the simple independent action model to epidemiologic studies involving two factors and a dichotomous outcome. *Am J Epidemiol*, 1986; 123: 162-173.

10. Breslow NE, Day NE. *Statistical methods in cancer research, vol. I. The analysis of case-control studies*. Lyon, International Agency for Research on Cancer, 1980.

11. Breslow NE, Day NE. *Statistical methods in cancer research, vol. II. The design and analysis of cohort studies*. Lyon, International Agency for Research on Cancer, 1987.

12. Doll R (1971) The age distribution of cancer: implications for models of carcinogenesis. *J R Stat Soc Lond series A*, 134, 133-156.

13. Siemiatycki J Thomas DC (1981) Biological models and statistical interactions: an example from multi-stage carcinogenesis. *Int J Epidemiol*, 10, 383-387.

14. Moolgavkar SH, Venzon DJ. General relative risk regression models for epidemiologic studies. *Am J Epidemiol*, 1987; 126: 949-961.

15. Thomas DC. General relative-risk models for survival time and matched case-control analysis. *Biometrics*, 1981; 37:673-686.

16. Breslow NE, Storer BE. General relative-risk functions for case-control studies. *Am J Epidemiol*, 1985; 122:149-162.

17. Breslow NE (1976) Regression analysis of the log odds ratio: a method for retrospective studies. *Biometrics*, 32, 409-416.

18. Holford TR (1980) The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36, 229-306.

19. Maldonado G, Greenland S. A comparison of the performance of model-based confidence intervals when the correct model form is unknown: coverage of asymptotic means. *Epidemiology*, 1994; 5: 171-182.

20. Maldonado G, Greenland S. Interpreting model coefficients when the true model form is unknown. *Epidemiology*, 1993; 4: 310-318.

21. Greenland S, Maldonado G. The interpretation of multiplicative model parameters as standardized parameters. *Statistics in Medicine*, 1994; 13: 989-999.

22. Maldonado G, Greenland S. Simulation study of confounder-selection strategies. *Am J Epidemiol*, 1993; 138: 923-936.

Summary

The health of populations is a concern for both governments and civil society. Nevertheless, progress in the technical dimensions of health measurement has not been matched by progress in political and popular concepts. A lack of consensus regarding key methods of population health measurement has limited the diffusion of modern concepts. In particular, disagreements about how to represent age and time reflect divergent views about how to represent population health. We propose an approach based on the use of generic multi-state dynamic population models and we examine whether the approach is feasible, whether it can be applied to real problems, and what sort of existing tools can be used to meet any additional data needs it entails.

In Chapter 2, we show that generic longitudinal multi-state modelling methods are feasible. However, their use appears to involve an increasing commitment to the use of mathematical tools as well as an increasing reliance on the ready availability of cheap computing power. Moreover, since such models, especially multi-state ones, are more complex than standard life tables, it is typically more difficult to estimate the required model parameters on the basis of observed epidemiological and demographic data.

The use of generic longitudinal population models as a tool for measuring population health benefit in an extended project of cost-effectiveness analysis at WHO seems notable, since it represents a significant institutional commitment to a form of population health measurement that is subtantially more sophisticated -- in terms of both the technical and conceptual aspects -- than either life table methods or burden of disease ("DALY") models. These cost effectiveness applications, two of which are presented in Chapter 3 and Chapter 4, demonstrate that generic longitudinal modelling methods allow important policy questions to be addressed in a way that is more satisfactory than with non-longitudinal methods.

The use of generic models also facilitates comparisons across interventions and across disease areas in the health sector, factors that are important for ensuring policy relevance of population health measurements. Although the link between the results of the analysis of health measurement questions and related health policy is perhaps not always direct, the publications corresponding to these chapters have attracted a high level of interest. WHO also continues to work directly with a range of countries on projects of cost effectiveness and related policy development in the health sector using these and similar tools.

However, generic longitudinal methods of population health measurement must acknowledge several serious challenges to their claims of realism and generality. In addition, the more widespread adoption of generic longitudinal models has arguably been limited by the fact that the corresponding modelling work is laborious and requires specialized skills. As alluded to above, supplying such models with input parameters can require the use of auxiliary modelling assumptions that can potentially militate against the desired realism and generality of the methods. Finally, multi-state generic longitudinal models are subject to model-specific sources of uncertainty that are difficult to quantify.

To address these limitations and difficulties, a variety of methods from econometrics, statistics and epidemiology are in principle at the disposal of the modeller. The logistic family of models is one example discussed in detail in Chapters 5, 6 and 7. Chapter 5 presents a novel account of the logistic model that illuminates aspects of its wide use in epidemiology and related fields. When such models are used as estimation models, with age or time as an independent variable, they arguably constitute dynamic models. Those that involve age can be termed implicit dynamic methods under the assumption of population stationarity in the relevant outcome. Those that involve time, such as models of epidemic progression, are typically explicitly longitudinal. The use of logistic models as an auxiliary method can therefore help satisfy the parameter needs of generic longitudinal population models.

However, the validity of the use of logistic and related models for this purpose will in general depend on the truth of a variety of assumptions. In some cases, the relevant assumptions will rely on the applicability to the estimation situation of an implicit mechanism of logistic diffusion that is described in Chapter 5. The work presented in Chapters 6 and 7 demonstrates in detail the application of these assumptions to a particular case. In addition to the assumption about logistic diffusion, that work relies explicitly on an assumption of population stationarity in the phenomenon of interest.

These chapters show that, when used judiciously, the stationary population assumption, an ubiquitous and nearly unavoidable feature in the use of standard life table methods, can, especially when paired with standard models from epidemiology and related fields, provide a useful device for the exploitation of limited observed data for the estimation of parameters in longitudinal population models. Likewise, a number of common epidemiological measures, such as the population attributable fraction or odds ratios obtained from case-control studies, usually also implicitly embody a population stationarity assumption.

The work presented in Chapter 7 relies on the assumption of population stationarity in all of these ways: stationarity in a logistic outcome, the interpretation of odds ratios as rate ratios and the use of population attributable fraction as measure of health impact. On the other hand, Chapter 8 demonstrates the potential dangers of indiscriminately using the stationarity assumption by providing an example where the implications of the stationarity assumption can be directly compared with those from an explicitly dynamic (i.e. non-stationary) model using the same data set.

In order to complete the theory of the time dimension in the measurement of population health, the work presented here should be expanded upon and complemented by additional research focusing on, among other things, the measurement of uncertainty in longitudinal population models and on the implications, as model complexity increases, of the need to estimate a rapidly growing number of model parameters from a relatively fixed number of observable quantities.

Longitudinal population models clearly represent a greater level of complexity than standard life table analysis. The concepts they involve, however, were all either implicitly or explicitly present at the birth of modern graphical methods of population health measurement in the later part of the 19th century. So long as the legitimate limitations of generic longitudinal models and their auxiliary methods are recognized, the availability of cheap computing power can provide a means of quantifying and operationalizing the original conceptual richness of population health measurement.

Samenvatting

Volksgezondheid is een zaak die zowel de overheid als de burgers aangaat. De ontwikkeling in de technische aspecten van volksgezondheidsmeting in de beleidsmatige en publieke sfeer heeft echter geen gelijke tred gehouden met die in de wetenschap. Een gebrek aan consensus betreffende basismethoden voor het meten van gezondheid op populatieniveau heeft ruime verspreiding van moderne begrippen in de weg gestaan. Met name uiteenlopende meningen over hoe rekening gehouden moet worden met leeftijd en tijd verraden verschillen van inzicht over hoe volksgezondheid het beste gemeten kan worden. Dit proefschrift stelt een benadering voor waarbij gebruik wordt gemaakt van zogenaamde generieke meerdimensionale dynamische bevolkingsmodellen, waarbij ook onderzocht wordt of de benadering uitvoerbaar is, of ze toegepast kan worden op problemen in de werkelijkheid, en welk type meetinstrument voldoet aan de datavereisten die dit met zich meebrengt.

In Hoofdstuk 2 tonen wij de haalbaarheid aan van een aanpak die gebaseerd is op generieke longitudinale meerdimensionale modellen. Deze modellen lijken een intensieve toepassing van wiskundige methoden te vereisen, en zijn afhankelijk van de beschikbaarheid van goedkope rekencapaciteit. Omdat dergelijke modellen, met name de meerdimensionale modellen, complexer zijn dan standaard overlevingstafels, is het als regel moeilijker om de vereiste modelparameters op basis van waargenomen epidemiologische en demografische gegevens te schatten.

Generieke longitudinale populatiemodellen worden onder meer voor het meten van gezondheidseffecten gebruikt bij kosteneffectiviteitanalyses bij de Wereld Gezondheids Organisatie (WHO). Dit is opvallend, omdat deze vorm van meten van volksgezondheid, die technisch en conceptueel veel verfijnder is dan methodes van overlevingstafels of ziektelast modellen

(„DALY"), een belangrijke inzet van de WHO vereist. Deze toepassingen van kosteneffectiviteitanalyse, waarvan er twee worden gepresenteerd in Hoofdstuk 3 en Hoofdstuk 4, tonen aan dat bij het gebruik van generieke longitudinale modelleringsmethoden, beleidsvragen op een bevredigender wijze beantwoord kunnen worden dan met niet-longitudinale methodes.

Het gebruik van generieke longitudinale populatiemodellen vergemakkelijkt tevens de vergelijking van verschillende interventies en van verschillende ziektegroepen in de gezondheidssector, wat de beleidsrelevantie vergroot. Ofschoon er geen direct verband is tussen de resultaten van de analyse van vraagstukken van gezondheidsmeting en het daarmee verbonden gezondheidsbeleid, hebben de publicaties waar deze hoofdstukken op zijn gebaseerd veel aandacht getrokken. De WHO zal met een reeks landen blijven samenwerken aan kosteneffectiviteitanalyses en verwante beleidsontwikkelingen in de gezondheidssector met gebruik van de hier beschreven technieken.

Het realisme en de generaliseerbaarheid die worden geclaimd door de generieke longitudinale methodes voor de meting van de volksgezondheid, zijn serieus in twijfel getrokken. Bovendien zal een gebruik van dit type modellen vermoedelijk beperkt blijven omdat het benodigde modelleringswerk arbeidsintensief is en speciale vaardigheden vereist. Zoals hierboven al aangestipt kan het bepalen van de inputparameters van dergelijke modellen het gebruik vereisen van aanvullende veronderstellingen die strijdig kunnen zijn met gewenst realisme en generaliseerbaarheid. Tot slot zijn meerdimensionale generieke longitudinale modellen behept met moeilijk te kwantificeren oorzaken van onzekerheid.

Om deze tekortkomingen en problemen te verhelpen, heeft de modelleur methodes uit de econometrie, statistiek en de epidemiologie tot zijn beschikking. De logistische familie van modellen is een voorbeeld dat in detail in hoofdstukken 5, 6 en 7 wordt besproken. Hoofdstuk 5 bespreekt een interpretatie van het logistische model die een nieuw licht werpt op het wijdverspreid gebruik in de epidemiologie en verwante aandachtsgebieden. Wanneer dergelijke modellen als schattingsmodellen worden gebruikt, met leeftijd of tijd als onafhankelijke variabele, kunnen zij als dynamisch worden beschouwd. De modellen die rekening houden met leeftijd kunnen

aangeduid worden als impliciete dynamische rekenmethodes onder een stationariteitsaanname van de onderliggende populatie. De modellen die rekening houden met tijd, zoals modellen van epidemieën, zijn als regel expliciet longitudinaal. Logistische modellen kunnen daarom helpen te voldoen aan de parameterbehoeften van generieke longitudinale populatiemodellen.

Het geldig gebruik van logistische en aanverwante modellen zal afhangen van de geldigheid van de onderliggende veronderstellingen. In sommige gevallen zullen de relevante veronderstellingen berusten op de toepasbaarheid van een mechanisme van logistische diffusie, zoals beschreven in Hoofdstuk 5. Hoofdstukken 6 en 7 beschrijven de toepassing van deze veronderstellingen op een specifiek voorbeeld. Naast de veronderstelling van logistische diffusie berust dit voorbeeld op de aanname van stationariteit van het beschreven fenomeen.

Deze hoofdstukken tonen aan dat de stationaire populatie veronderstelling, een bijna onvermijdelijke eigenschap bij het gebruik van de standaard overlevingstafels, in combinatie met standaardmodellen van epidemiologie en verwante gebieden, van nut kan zijn voor het volledig benutten van empirische data voor het schatten van parameters in longitudinale populatie modellen. Zo wordt ook bij een aantal veel gebruikte epidemiologische maten, zoals de populatie attributieve fractie en de odds ratio verkregen uit case-control studies, meestal impliciet van een stationaire populatie uitgegaan.

Het werk dat in Hoofdstuk 7 wordt gepresenteerd berust op de aanname van verschillende soorten stationariteit : stationariteit in een logistisch resultaat, de interpretatie van odds ratio's als rate ratio's en het gebruik van de populatie attributieve fractie als een maat van gezondheidseffect. Anderzijds toont Hoofdstuk 8 met een voorbeeld de potentiële gevaren van de stationaire populatie veronderstelling aan wanneer deze zonder onderscheid gebruikt wordt. In dit voorbeeld worden de implicaties van de stationariteitsveronderstelling duidelijk door directe vergelijking met resultaten van een expliciet dynamisch (d.w.z. niet-stationair) model dat dezelfde gegevens gebruikt.

Om de theorie van de tijdsdimensie bij het meten van gezondheid op populatie niveau te voltooien, moet het gepresenteerde werk in dit proefschrift

aangevuld worden met onderzoek dat gericht is op het meten van onzekerheid in longitudinale populatie modellen en, met gebruik van ingewikkelder modellen, op de implicaties voor het schatten van een toenemend aantal model parameters op basis van een relatief vast aantal waarneembare feiten.

De longitudinale populatie modellen zijn duidelijk complexer dan standaard overlevingstafels. Echter de gebruikte begrippen waren impliciet dan wel expliciet aanwezig bij de geboorte van de moderne grafische methoden van volksgezondheidsmeting in de late 19de eeuw. Zolang er goed inzicht is in de bestaande tekortkomingen van generieke longitudinale modellen en hun hulpmethodes, kan de steeds goedkopere rekencapaciteit een middel zijn om de conceptuele rijkdom van het meten van de volksgezondheid te kwantificeren en te operationaliseren.

Acknowledgements

I would like to thank current and former co-workers at WHO, in particular, Chris Murray, David Evans, Tessa Tan-Torres, Raymond Hutubessy, Louis Niessen, Rob Baltussen, Colin Mathers, Stephen Lim, Dan Chisholm and Taghreed Adam.

I would also like to thank my friends, mentors, and sometimes also valued colleagues, Klaus Röhrich, Harald Wirth, Steve Gribble, Claude Charette, Carel Pretorius, Brian Williams and Sander Greenland, all of whom have contributed to this thesis in important ways. I also mention the many co-authors from whose contributions and expertise I have significantly benefited in the course of performing the research that forms the basis of this thesis, only some of whom are named here.

I thank also my advisors, Dik Habbema and Jan Barendregt, both of whom have been exceedingly generous in guiding me in the course of preparing this thesis and whose example of mentorship will therefore remain a permanent inspiration for me.

I would also like to thank my parents, Ken and Marjorie, and the other family members who have encouraged me in this project. Finally, I would like to thank my wife, Ana Betrán, and my daughters Andrea and Susana, for their loving support.

Curriculum vitae

Jeremy A. Lauer has worked as an Economist at the World Health Organization, in the department of Health Systems Financing, since 1998. Prior to that, he worked for three years in the Office of Publications of the World Health Organization. Jeremy received training in economics at the University of Wisconsin, Madison, from which he graduated in 1991 with a Master of Science in Economics and a Master of Arts in Agricultural and Applied Economics.

He first worked in the health field as a researcher with Dr Benjamin S. Carson in the Department of Neurosurgery at the Johns Hopkins School of Medicine from 1992 to 1994, where, with Dr Carson and other workers, he published several papers on pediatric neurosurgical topics and procedures. In 1982, Jeremy had enrolled at St John's College in Annapolis, Maryland, where he received a Bachelor of Arts degree, with a double major in mathematics and philosophy and a minor in physics, in 1986. Following graduation, Jeremy worked for two years in the Office of Admissions at St John's before matriculating at UW Madison.

In the department on Health Systems Financing at WHO (formerly, Global Programme on Evidence for Health Policy), Jeremy has been involved in a variety of fields of research in international public health under the direction of Dr Christopher J. L. Murray, Dr David B. Evans and Dr Tessa Tan-Torres Edejer. He was part of the core team that prepared the global country assessments ("rankings") on health system performance assessment in 2000. Jeremy was also part of the core team that assessed the cost effectiveness of preventive interventions (including novel medical and public health therapies) for cardiovascular disease, that formed a part of the *World Health Report 2002*. More recently, he was part of the core team analysing the cost effectiveness

of interventions targeted to achieving the Millennium Development Goals, a project whose methods and results were published as a series of papers by the *BMJ* in late 2005. Jeremy has published in a range of public health and medical journals and books, on topics from maternal and child health to health systems research and cost-effectiveness analysis.

Jeremy has been particularly involved in the modelling of intervention effectiveness at population level, a field which provides ample scope for his interests in mathematics, statistics, econometrics, demography and epidemiology. He has been responsible for developing, in collaboration with other workers, the core methods for population effectiveness assessment used by WHO-CHOICE, the WHO programme on cost effectiveness assessment. Since 2003, Jeremy has, in addition to his duties at WHO, worked part time towards a doctorate degree under the supervision of Professor J. D. F. (Dik) Habbema at Erasmus University Rotterdam and Dr Jan Barendregt of the University of Queensland in Australia.