Clinical prediction models in reproductive medicine:

Applications in untreated subfertility and in IVF treatment

Claudine C. Hunault

ACKNOWLEDGEMENTS

The work presented in this thesis was conducted at the Department of Public Health, with additional financial support of the Department of Reproductive Medicine of the Erasmus MC-University Medical Centre Rotterdam, and the Department of Reproductive Medicine of the University Medical Center Utrecht, The Netherlands.

The studies described in this thesis were supported by the Netherlands Organization for Scientific Research (ZON/MW).

The Department of Public Health of the Erasmus Medical Center, Rotterdam and the National Institute of Public Health and the Environment (RIVM) provided financial support for the publication of this thesis.

Cover: *F0082542·580953*, © HALLMARK CARDS · Foto: Zefa.
The firms Hallmark Cards Nederland B.V. and Fotostock B.V. have been contacted by the author of this thesis for copyrights.
For any question, please contact the author of this thesis (chunault@hetnet.nl).

Printed by: Optima Grafische Communicatie, Rotterdam, The Netherlands.
ISBN: 90-8559-233-X

Clinical prediction models in reproductive medicine:
Applications in untreated subfertility and in IVF treatment

Klinische predictiemodellen in de voortplantingsgeneeskunde:
toepassingen in onbehandelde sub-fertiliteit en in IVF behandeling

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
17 November 2006
at 9 hrs

by

Claudine Colette Hunault
born in Angers, France

Doctoral Committee

Promotors:	Prof.dr. J.D.F. Habbema Prof.dr. E.R. te Velde
Other members:	Prof.dr. Th.J.M. Helmerhorst Prof.dr. N.S. Macklon Prof.dr. T. Stijnen
Copromotor:	Dr. M.J.C. Eijkemans

“As you sow, so shall you reap”

To the memory of my late father

Contents

1

Introduction

For many couples the advent of a baby is the most beautiful gift of Life. However, the motivations to wish a child are very different. For example it may be personal (to realize yourself as an individual), relational (to establish a more complete relationship with your partner) or social (to reach a status which is highly valued by society).

When in earlier days the wish to have a child could not be fulfilled, people had to resign to childlessness or had to turn to adoption. Currently, medical techniques are able to resolve problems for many albeit not all couples. Involuntary childlessness by itself does not threaten physical health but can have a strong impact on the psychological and social well being of couples (van Balen and Trimbos-Kemper, 1995). The impact of experiencing difficulties in conceiving is often underestimated by the general public (Bertarelli, 2000). Many infertile couples find their problem stressful and may experience deterioration in their sexual relationship, which exacerbates the problem.

The probability of conception among couples who are involuntary childless varies according to the causes and determinants of the sub/infertility. Knowledge of the chance to conceive, with and without treatment, is important in the process of counselling couple. In the past, clinical experience was the only available "tool" to estimate these chances. However, to quantify and integrate the effects of the many prognostic factors involved is difficult. In recent years, the increased demand for expensive reproductive treatments and the limited health care resources have enhanced the need for rational decision making in reproductive medicine. Clinical prediction rules based on statistical modelling, may help the clinician in the management of sub/infertility by making reliable prognostic statements.

This thesis is devoted to two prediction problems in reproductive medicine: (1) prediction of the chance to conceive among untreated infertile couples and (2) prediction of the chance to conceive with IVF treatment. The aim is to construct and validate statistical prediction models that may be used to estimate the individual chance of a particular couple.

We will briefly discuss prevalence and causes of infertility, the basic infertility work up, prognosis and entry into sub/infertility, prognosis in IVF and development and validation of clinical prediction models. Four research questions are formulated and we indicate which chapters address these four questions.

SUB/INFERTILITY

Infertility is usually defined as a failure of conception within one year of unprotected intercourse. Since 85-92 percent of couples who wish to have a child, realize a pregnancy within one year after stopping birth control, the prevalence of infertility is 10-15 percent according to this definition. However, the percentage of couples who are really infertile in the sense of sterile, and thus not able to ever have a child, is at most 2-4% (Greenhall and Vessey 1990). The often used

definition of infertility – a failure to conceive within one year - is, therefore, misleading, because among the couples who have not conceived within one year, many still have a reasonable albeit reduced chance to conceive spontaneously while only a minority is sterile (te Velde *et al*., 2000) (Habbema *et al*., 2004). We therefore prefer to distinguish subfertility or reduced fertility – the chance to become pregnant is reduced but not zero - from infertility: there is no chance at all.

Causes of subfertility and their prevalence

The most common causes of sub/infertility and their estimated prevalence are summarised in **Table 1**.

Table 1: Main causes of sub/ infertility (Hull *et al.*, 1985).

Diagnostic category	Frequency
Unexplained infertility	20%
Mild Sperm defect	35%
Failure of ovulation	21%
Tubal damage	14%
Severe Sperm defect	5%
Cervical hostility	5%

If a couple has not conceived within one year of unprotected intercourse, a basic infertility work-up is warranted for which the WHO provided guidelines (Rowe *et al.*, 1993). Traditionally, the work-up begins with an investigation of medical history and lifestyle issues and a physical examination. Many tests are available to investigate the cause of infertility (**Table 2**). According to the diagnosis established and the couple characteristics, the physician may decide to initiate a treatment or not. This decision should be based on the rational balance of the pregnancy chances with and without treatment, with the costs and the burden for the couple. The same holds for decisions during treatment on which next step to take. In both situations, prediction models may be of help.

Prognosis in untreated sub/infertility

The population average of the monthly chance to become pregnant is between 20% and 30% but there is a large variation between couples (Bongaarts, 1975). Because of this heterogeneity, the couples who have not conceived after one year of unprotected coitus form a selection with a relatively low monthly chance.

However, failure of conception after one year does not mean that the probability of conception is zero. In **Figure 1**, 69% of the couples conceive within half a year and 84% within one year. Of the 16% of the couples that has not become pregnant within 12 months, only about 25%, is really sterile.

Therefore, a rather high percentage (here: 75%) of these couples will eventually still conceive spontaneously, 66% of whom during the second year. In **Figure 1**, the chance of conception is only related to duration of infertility. In reality, it is affected by multiple other factors, such as the woman's age, previous pregnancy, and the presence of pathology (Collins *et al.*, 1995) (Snick *et al.*, 1997) (Eimers *et al.*, 1994) that may be detected during the diagnostic work-up.

Table 2: Tests often used in the diagnostic work-up of the subfertile couple.

Evaluation	Test
WOMAN	
Ovulation testing	Basal Body Temperature charts (BBT) LH measurements in blood or urine Ultrasound (Follicular growth)
Ovarian function tests	Hormone levels (FSH, Estradiol, Inhibin B) Antral follicle count by ultrasound
Luteal Phase testing	Serum progesterone level Endometrial biopsy
Cervical mucus and sperm function	Postcoital test (PCT)
Uterus and tubal patency	Chlamydia antibody test Hysterosalpingogram Hysteroscopy Laparoscopy
MAN	
Semen	Semen analysis: - Sperm count, - Motility, - Forward progression, - Morfology - Total semen volume, Sperm antibodies
Hormone tests	FSH levels Testosterone Prolactin

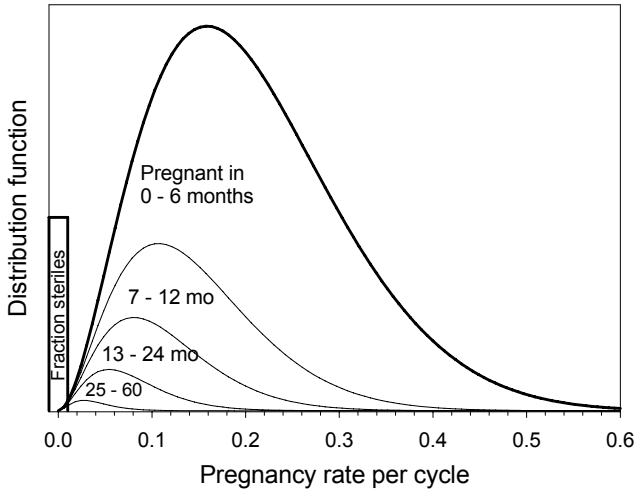

Figure 1, extracted from a demographic study (Bongaarts, 1975) shows the distribution of the monthly chance of spontaneous pregnancy in a natural population wishing to procreate. A small group (4%) is sterile and has zero chance to conceive. For the other couples, the monthly chances vary between 0 and 0.6, with most of the couples having a chance between 0.1 and 0.3. Because of selection, patients who do not succeed in conceiving after several months have a distribution shifting more and more towards the left, with an increasingly lower monthly chance.

Prognosis in *in vitro* fertilization (IVF)

In *in vitro* fertilization, the ovum is fertilized outside the mother's body. This 'empirical' assisted reproductive technology (ART) does not aim to remove the cause of infertility but to achieve a pregnancy despite the sub/ infertility. In 1978 the first baby after IVF was born in the UK. Other countries followed soon, with steadily increasing production. From 1996 to 2000, 63,414 IVF or ICSI cycles were started in the Netherlands (Kremer *et al.*, 2002).

The first indication for IVF was tubal pathology or severe endometriosis but nowadays IVF is also indicated in the other diagnostic categories mentioned in **table 1**. Several technologically advanced variants of IVF have been developed in the last decades of which intra cytoplasma sperm injection (ICSI) used for severe sperm disorders, is the most important one. We will focus on the classical IVF procedure in this thesis.

Pregnancy chances in IVF are limited, but can be enhanced by increasing the number of embryos transferred. Such an increase will also increase the number of multiple pregnancies (Dickey, 2003) (ESHRE 2000). **Table 3** shows the rate of multiple deliveries in ART in several countries. A recent review reported that twins conceived with ART had a 5 percent higher rate of intensive care admissions, a 21 percent higher rate of premature and Caesarian delivery rates, and a 27 percent increased risk of being small for gestational age, compared to singleton pregnancies (Helmerhorst *et al.*, 2004). Figures for triplets will be worse. Multiple pregnancies are also associated with higher rates of neonatal mortality, handicaps, and malformation (Olivennes, 2000) (Hansen *et al.*, 2002) (Stromberg *et al.*, 2002) (Moll *et al.*, 2003). Moreover, health economic studies have indicated that much of the financial burden levied by IVF treatment is caused by multiple pregnancies (Wølner-Hanssen and Rydhstroem, 1998) (ESHRE, 2000).

Table 3: Deliveries in relation to multiple births after IVF (1997) (Oliva and Arnau, 2001).

Country	Singleton (%)	Twins (%)	Triplets (%)	Quadruplets (%)
Denmark	75.4	24.2	0.4	0.0
Finland	72.6	26.3	1.1	0.0
France	74.4	23.7	1.9	0.0
Germany	72.2	22.3	5.3	0.2
Greece	63.1	35.6	2.7	0.0
Italy	73.6	20.3	5.1	0.9
Norway	70.6	28.2	1.2	0.0
Portugal	77.6	16.4	6.0	0.0
Sweden	74.2	25.4	0.4	0.0
Switzerland	76.2	20.6	3.2	0.0
United Kingdom	70.7	25.9	3.3	0.0
Spain	54.5	32.7	11.0	0.9

In view of these risks associated with multiple pregnancies, there is an important ongoing debate about whether and when more than one embryo should be transferred in IVF.

CLINICAL PREDICTION MODELS

The past decades have witnessed the methodological development and clinical application of prediction models that aim to predict the prognosis, or course of disease, in a particular patient. Prognosis refers to the possible outcomes of a disease and the frequency with which they can be

expected to occur (Laupacis *et al.*, 1994). In internal medicine, for example oncology, outcome refers to death within a certain time period, to survival time or to recurrence of disease. The emphasis is usually on unfavourable outcome. The opposite occurs in reproductive medicine, where the favourable outcome, pregnancy, is the focus of attention. The reason for this difference is that the event that ends the period of follow up dictates if we are prognosticating the favourable or the unfavourable outcome. In oncology, cure from cancer is not observable, but death or recurrence is. In reproductive medicine, impossibility to conceive/sterility is only observable in exceptional cases; usually only the favourable outcome, pregnancy, is observable; therefore, the models concern the prediction of the chance of pregnancy without treatment ('spontaneous' pregnancy) (Eimers *et al.*, 1994) (Collins *et al.*, 1995) (Snick *et al.*, 1997) or with therapeutic interventions, in particular IVF treatment (Stolwijk *et al.*, 1996) (Wheeler *et al.*, 1998) (Martin and Welch, 1998). The main reason for using clinical prediction models is to assist patients and clinicians in making difficult choices. During the counselling process of their patients, clinicians have to weight the expected benefit of a treatment against the possible side effects and costs of this treatment. The outcome of a disease process is usually related to multiple factors and the chances of the different possible outcomes are therefore difficult to estimate intuitively (Wiegerinck, 1999). For instance, in case of unexplained subfertility, the decision to wait longer or to initiate a treatment is not straightforward, because the chance of pregnancy can also be high without treatment and treatment can have possible side effects such as complications from multiple pregnancy or the ovarian hyper stimulation syndrome (OHSS).

Clinical prediction models are tools developed to assist clinicians in making these choices. They combine patient information (coming from the patient history, physical examination, laboratory testing, and so on) to background information obtained from previous studies. Prediction models realise the synthesis of the effects of multiple factors on the prognosis of a disease and can predict the occurrence of an event (e.g., ongoing pregnancy) for an individual patient. However, they may perform disappointingly in clinical practice. Therefore, prognosis evidence should be studied adequately. The characteristics of good prognostic evidence for clinical use, focusing on study design, statistical analyses, and evaluation and presentation of results have recently been described by Eijkemans (Eijkemans, 2004) and are reproduced in the next four pages.

Study Design

The essence of prognosis is to relate patient and disease characteristics known at present to outcome in the future. Therefore, longitudinal data from a cohort study are required to construct such a prognostic model. The cohort study may be either prospective or retrospective. From a methodological point of view the prospective design is to be preferred, as it allows investigators to standardize treatment, guarantees that variable of interest are registered, minimizes missing data, and maximizes completeness of follow up. In the design phase of the study decisions have to be made on the outcome measure(s) and on the potential predictive variables, based on prior information from literature and pathophysiological knowledge.

Data Considerations

The outcome that is to be predicted is often dichotomous (e.g. pregnancy [yes/no] in an *in vitro* fertilization (IVF) cycle (Templeton *et al*., 1996)), or a continuous “time to event” (e.g. with time to spontaneous pregnancy or live birth (Eimers *et al*., 1994) (Collins *et al*., 1995) (Snick *et al.,* 1997)), and cumulative pregnancy rates in IVF (Stolwijk *et al.,* 1996). An observation on time to pregnancy is called “censored” if follow-up ended before pregnancy occurred.

Coding of predictor variables is preferably as detailed as possible (i.e., continuous of the underlying phenomenon is measured on a continuous scale). Categorizing of variables leads to loss of information and should be performed, if at all, in the analysis phase only.

Choice of Model

Several types of statistical techniques may be applied to relate predictor variables to an outcome. Among others, we mention regression models, neural networks (Cross *et al.,* 1995) (Baxt, 1995) and classification and regression trees (Breiman *et al.,* 1984). We will focus on regression models, because they are most frequently used in medical applications.

A regression model relates an outcome variable (Y) to one (univariable) or to the weighted sum of several (multivariable) predictor variables ($X = \{X_1 \ldots X_p\}$). The weights by which the variables are multiplied are called "regression coefficients" ($\beta_i, i = 1 \ldots p$) and present the strength of the association between a predictor and the outcome. The weighted sum is called the prognostic index (PI). In formula:

$$PI = \beta_1 X_1 + \ldots + \beta_p X_p$$

Three main types of regression models can be distinguished, dependent on the type of outcome data (**Table 4**): (1) ordinary linear regression for continuous outcomes, (2) logistic regression for dichotomous (or binary) outcomes, and (3) Cox proportional hazards regression for censored time to event data (also known as survival data).

Informative Censoring

Cox regression makes the assumption that censoring is uninformative. In a time to pregnancy analysis this means that the fact that the follow-up of the patient ended before the patient had become pregnant (censoring) is not related to her pregnancy chances. Reasons for censoring could be dropout because the patient moved to another residence or because the couple had marital problems. In general, censoring because of study design or protocol is uninformative, because no prognosis-related selection of patients occurs. In the analysis of spontaneous pregnancy chances, censoring often occurs because a treatment is started. Informative censoring may occur when patients are selectively treated only when their spontaneous pregnancy chances are being regarded as low.

Model Development

Determining which predictor variables will be part of the model is a major challenge. Already in the design phase of the study choices have been made on which variables to register. Usually a wide scope of variables is included so as to not miss factors that could later appear to be important. For prognostic modeling this is usually not sensible. Including all registered variables as predictors in the prognostic model may lead to serious over-fitting and also to redundancy in case of mutually strongly correlated predictors. Furthermore, it is impractical to use a model with many predictors. Often, a reduction in the number of predictor variables is required.

Preferably, a conservative approach should be used in model reduction: start with a limited set of variables whose prognostic value has been established in the literature. It is generally advised to use no more than one potential predictor on every 10 cases (i.e., patients in ordinary linear regression or events in logistic and Cox regression) (Harrell *et al*., 1985). Experience in large simulation studies shows that a model with this approach will perform well in a new, independent data set (Steyerberg *et al.,* 2000). Often, further reduction of the number of predictors is achieved by applying statistical criteria such as backward or forward stepwise selection. These methods

delete or include variables in a stepwise fashion according to repeated statistical significance testing. In this way non-informative variables are excluded from the model, but the danger is that some informative variables are also excluded. Furthermore, it involves statistical testing and fitting on the same data, known to lead to biased estimates of the coefficients for the variables that are included (Miller, 1990) (Hurvitch and Tsai, 1990). Simulation studies have shown that stepwise selection with the usual significance level of $P = 0.05$ may lead to poor model performance in new data. It is better to apply a less strict P value: P values of 0.10, 0.20, and even 0.50 have been proposed (Steyerberg *et al.,* 1999).

Table 4: Three types of outcome data, with corresponding regression technique and interpretation of regression coefficients

Type of outcome Y Regression technique	Relationship between outcome (*Y*) and prognostic index (*PI*)	Interpretation of regression coefficient β_i
Continuous Linear	$Y\|X = \beta_0 + PI$ [a,b]	β_i: change in average outcome per unit change of the variable
Dichotomous Logistic	$\Pr\{Y = 1 \mid X\} = \frac{1}{1 + Exp[-(\beta_0 + PI)]}$ [a,b,c]	$Exp(\beta_i)$ = Odds Ratio (OR): change in Odds on outcome per unit change of the variable
Censored time to event Cox proportional hazards	$\Pr\{\text{free of event at } Y = T \mid X\} =$ $S(Y = T \mid X) = S(Y = T \mid 0)^{Exp(PI)}$ [b,d]	$Exp(\beta_i)$ = Hazard ratio (HR): change in hazard on outcome per unit change of the variable

[a] β_0 is the intercept of the regression formula
[b] $PI = \beta_1 X_1 + \ldots + \beta_p X_p$
[c] $\Pr\{Y=1|X\}$ is the probability that the dichotomous outcome Y will take the value 1, for a patient with predictor variables X.
[d] $\Pr\{\text{free of event at } Y=T|X\} = S(Y = T|X)$ is the probability that the event has not yet occurred at follow-up time $Y = T$, for a patient with predictor variables X (this is also known as the Survival probability, referring to the origins of this method that lie in analysis of mortality). $S(Y = T|0)$ is the 'baseline' survival probability, corresponding to a patient with all predictor variables equal to 0. Note that for each follow-up time T the survival probability can be calculated, giving a survival curve. When the event of interest is pregnancy instead of mortality the 'one minus survival' curve is usually preferred.

Missing Data

Missing data may form a serious problem in multivariable modeling for several reasons. When the outcome variable is missing, bias may occur if values are missing selectively. For instance, women who become pregnant may be more willing to respond to a questionnaire on the outcome of treatment than women who don't get pregnant. As a result, the chance of pregnancy will be overestimated when only the women who responded are used for analysis. When a predictor variable is missing, usually no bias is introduced when the couple is dropped from the analysis. However, it is a waste of data to drop a couple in which the outcome and many predictor variables have been measured, just because one of the predictor variables is missing. Imputation techniques may be used to keep these cases in the analysis (Little, 1992). Although imputation doesn't produce new independent information, it prevents existing information from being dropped.

Validation

The purpose of developing a prognostic model is to provide valid predictions in future patients. Validity refers to a number of concepts that all relate to whether the model predictions can be trusted:

- Random or prediction error: How precise are the predictions?
- Systematic error or reliability: Do the predictions agree with observations? Has to do with model misspecification, or omitted predictor variables.
- Discrimination: To what extent are the model predictions able to separate between outcome categories?

When validity is assessed on exactly the same patients that were used to develop the model, we speak of "apparent validity". Validity in new but similar patients from the same setting is called "internal validity", whereas "external validity" refers to patients from another time or place. For assessment of external validity a new study should be set up. Preferably, this should be a prospective longitudinal cohort study, just like the study that was used to develop the model.

Usually, we can only assess apparent and internal validity when developing a model. For apparent validity we simply apply the model to the same data that were used to develop the model. For internal validity we use the same data, but model development and assessment of validity are performed on different parts of the data. Cross-validation is a method in which the data set is randomly split in a number of parts of equal size. In turn, each part is used to evaluate a model that has been developed on the complementary parts. Well-known examples are the split quarter and split-half methods. In the most extreme variant, a single patient is left out to evaluate a model build on all other patients. This is the repeated for each patient (leave-one-out method). The major drawback of cross-validation is that it is inefficient because only part of the data is used for model development or evaluation. A method that uses all available data is the bootstrap method. In this method the model-building process (selection of variables in the model and parameter estimation) is repeated a pre-specified number of times (e.g., 200 times). Each repetition consists of creating a new data set (bootstrap sample) by drawing cases with replacement from the original data. The resulting model from each bootstrap sample is evaluated on the original data. This procedure mimics what would happen if new data were collected repeatedly in the same setting (Efron and Tibshirani, 1993) and therefore assesses internal validity.

External validity is assessed on patient data from another time or another place. The real test of the model, whether it can be used in another setting, is here at stake, and for practice this is the most important test.

Statistical Prediction Error

Uncertainty about the predictions of the model will always exist because of the statistical imprecision of the regression coefficients. For a given patient profile, the model produces a predicted outcome, with an associated 95% confidence interval. The prediction is to be interpreted as a mean value for this patient profile, and the confidence interval refers to the uncertainty about this mean value.

Systematic Error, Reliability or Calibration

Predictions of the model may also show systematic error when compared with actual observations. Reliability (or calibration) refers to how well predictions agree with observations. For example, if the model predicts that a patient has a 30% probability of getting pregnant, is the probability really 30%? Of course the real probability cannot be observed in an individual patient but one could imagine a group of identical patients and determine their frequency of pregnancy.

Assessing apparent reliability is equivalent to verifying that the model fits well to the data, by performing a goodness of fit test such as the Hosmer-Lemeshow for logistic regression (Hosmer and Lemeshow, 1989). Lack of fit may occur with improper modeling of the "dose-response" relationship between a continuous predictor and the outcome (e.g., the usual linear form is chosen where the data follow a nonlinear "bathtub" kind of curve), or because important interactions between predictors are omitted from the model. To resolve lack of fit, subject knowledge from pathophysiology may be of greater help than statistical "data dredging", because of the risk of chance findings and overfitting.
When there is no (more) lack of fit, the agreement between predictions and observations will be good in the data that were used to develop the model. However, statistical theory (Copas, 1983) (Van Houwelingen and Le Cessie, 1990) and recent simulation studies (Steyerberg *et al.,* 2000) have shown that in new patients the model predictions will be overoptimistic: higher than average predictions are too high and lower than average predictions are too low, a phenomenon related to "regression to the mean". It becomes worse when the number of candidate predictor variables is higher relative to the number of patients with the outcome. It has been shown that reliability in new patients will improve if all regression coefficients are corrected by a shrinkage factor (Van Houwelingen and Le Cessie, 1990). The bootstrap method is well suited to estimate the shrinkage factor that is required.

Discriminative Ability
Discrimination refers to the ability of the model to discriminate between good and poor prognosis patients, and it is a quantification of the degree to which predicted probabilities are lower for patients with poorer outcome. For logistic models it may be measured by the area under the ROC curve (AUC), for Cox models by the c-statistic (Harrell *et al.,* 1984). They have a similar interpretation: give two randomly selected patients with different outcomes (e.g., one getting pregnant, the other not or the time to conception of one patient is shorter than for the other one), AUC or c-statistic gives the probability that the model prediction is worse in the patient with the worst outcome. For sensible models it should be higher than 0.5, which is the AUC of a non-informative "flip of a coin" model.
"Apparent" discriminative ability is always optimistic. Internal validation techniques such as the bootstrap method will result in lower and more appropriate values for AUC or c-statistics than the apparent value.

Presentation
Finally, the results of model development and (internal) validation have to be presented. This may be done in the form of the regression formula, with the estimated coefficients, corrected by a shrinkage factor, obtained from the statistical software. Although the calculation of the prognostic index may be done with pen and paper, for logistic regression and the Cox model the transformation from prognostic index to predicted outcome variable (the "link function") is too complicated to perform by hand (**Table 4**), and users will have to implement the formula in computer software such as a spreadsheet. Alternatively, a score chart may be constructed, consisting of a score table in which, for each value of the predictors, a score is assigned corresponding to the regression coefficient from the model. For this purpose, continuous predictors are divided into categories. The clinician has to look up the scores corresponding to the values of the predictor variables and add them, giving a sum score. Finally, the probability corresponding to the sum score can be read from a graph or table that also has to be provided.
For an example of a score chart for predicting treatment-independent pregnancy, see **Table 5** and **Figure 2**.

Alternatively, a fully graphical presentation form, such as a nomogram, may be chosen. For example, to obtain an overall prediction of pregnancy chances, prior to start of Clomiphene Citrate (CC) medication in anovulatory infertility (and thus prior to knowledge about the chance of ovulation with CC or the chance of pregnancy in case of ovulation), a nomogram was constructed (Imani *et al.,* 2002).

Use of prediction rules

An important question is whether a published prediction rule can be used in another setting or whether first some adaptations should be made, reflecting the difference between the new setting and the one in which the prediction rule was developed. Another question arises in spontaneous pregnancy prediction where we are in the luxurious situation that several carefully developed prediction models have been published. Which one should be used? The one developed in a setting that most resembles the setting where the rule will be applied? Or the best validated model? Or is it possible to combine the prediction rules into a "synthesis model" that is even more useful? These questions have to be considered from case to case.

RESEARCH QUESTIONS

This thesis describes the development, synthesis and validation of clinical prediction models in two clinical problems of reproductive medicine: the prediction of the chance of pregnancy in untreated subfertility and in infertility treated by IVF. The following specific research questions are addressed:

1. **Does the combination of existing models for predicting the chance of pregnancy among untreated subfertile couples result into improved predictions**?

2. **Are the synthesis model and the Eimers model for predicting the chance of pregnancy among untreated subfertile couples externally valid?**

3. **Can a valid model be developed for assisting in the choice between single and double embryo transfer?**

OUTLINE OF THIS THESIS

In *Chapter 2,* a clinical prediction model is developed to estimate the chance of pregnancy leading to live birth among untreated subfertile couples. This model is based on the synthesis of three previous studies (Research question 1).
Chapters 3 and 4 describe the external validation of two models predicting the chance of pregnancy leading to live birth among untreated subfertile couples (Research question 2).
In *Chapter 5,* the development of a clinical prediction model estimating the chance of ongoing pregnancy and twin pregnancy after dual (dET) and single embryo-transfer (sET) is described and in *Chapter 6*, the external validity of the model developed in chapter 5 is assessed (Research question 3).
Chapter 7 concludes the thesis with a general discussion in which the research questions are answered and conclusions are drawn.

Table 5: Pocket chart of the Eimers model for calculating the chance pregnancy within 1 year for untreated subfertile couples.

Predictor		Infertility score[a]
Woman's age	21 to 25 years	0
	26 to 30 years	2
	31 to 35 years	4
	36 to 40 years	6
	41 to 45 years	8
Duration of subfertility		
	1 year	0
	2 years	2
	3 to 4 years	4
	5 to 6 years	8
	≥ 7 years	12
Female subfertility		
	Secondary	0
	Primary	7
Fertility problems in male's family		
	No	0
	Yes	5
PCT		
	Progressive	0
	Non progressive	10
	Negative	20
Motility		
	≥ 60%	0
	40%-60%	3
	20%-40%	7
	0%-20%	10
Prognostic Index		---

[a] Circle the infertility scores for each of the predictors and add them to obtain the prognostic index

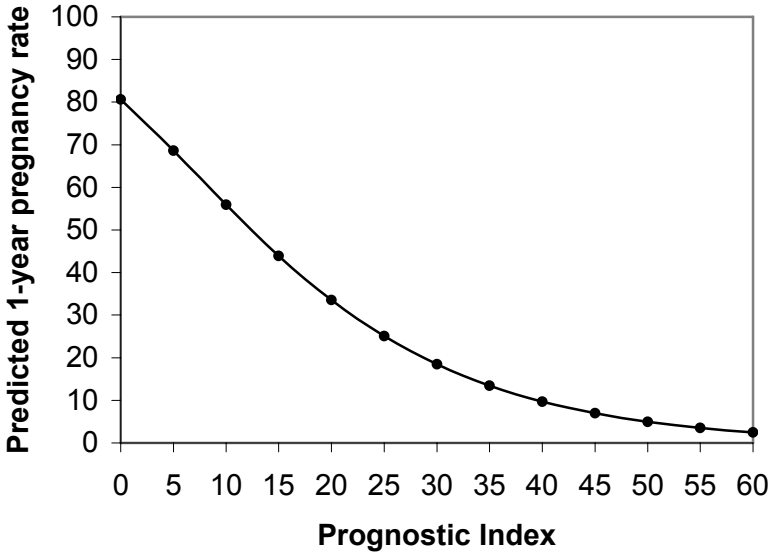

Figure 2: Relation between prognostic index (see **Table 5**) and chance of pregnancy within 1 year.

References

Baxt WG. (1995) Application of artificial neural networks to clinical medicine. *Lancet*. 346, 1135-8.

the Bertarelli Foundation Scientific Board. (2000) Public perception on infertility and its treatment: an international survey. *Hum. Reprod*. 15, 330-4.

Bongaarts J. (1975) A method for the estimation of fecundability. Demography 12, 645-60.

Breiman L, Friedman JH, Olshen RA, Stone CJ. (1984) *Classification and Regression Trees*, Wadsworth, Belmont, CA.

Collins JA, Burrows EA, Willan AR. (1995) The prognosis for live birth among untreated infertile couples. *Fertil Steril.* 64, 22-8.

Copas JB. (1983) Regression, prediction and shrinkage. *JR Stat Soc B.* 45, 311-54.

Cross SS, Harrison RF, Kennedy RL. (1995) Introduction to neural networks. *Lancet.* 346, 1075-9.

Dickey RP. (2003) A year of inaction on high-order multiple pregnancies due to ovulation induction. *Fertil Steril*. 79, 14-6.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD. (1994) The prediction of the chance to conceive in subfertile couples. *Fertil Steril*. 61, 44-52.

Efron B, Tibshirani RJ. (1993) *An introduction to the bootstrap.* Chapman & Hall Inc, London, UK.

Eijkemans MJC. (2004) *Fertility in populations and in patients: Population studies on natural fertility and prediction of treatment outcome in anovulatory infertile patients.* Thesis Erasmus University Rotterdam. Rotterdam, The Netherlands.

Eshre Capri Workgroup Group. (2000) Multiple Gestation Pregnancy. *Hum Reprod.* 15, 1856-64.

Greenhall E and Vessey M. (1990). The prevalence of subfertility: a review of the current confusion and a report of two new studies. *Fertil Steril.* 54, 978-83.

Habbema JDF, Collins J, Leridon H, Evers JLH, Lunenfeld B, te Velde ER. (2004) Towards less confusing terminology in reproductive medicine: a proposal. *Fertil Steril.* 82, 36-40.

Hansen M, Kurinczuk JJ, Bower C, Webb S. (2002) The risk of major birth defects after intracytoplasmic sperm injection and in vitro fertilization. *N Eng J Med.* 346, 725-30.

Harrell FE, Lee KL, Califf RM, Pryor DB, Rosati RA. (1984) Regression modelling strategies for improved prognostic prediction. *Stat Med.* 3, 143-52.

Harrell FE, Lee KL, Matchar DB, Reichert TA. (1985) Regression models for prognostic predictions: advantages, problems, and suggested solutions. *Cancer Treat Rep*. 69, 1071-7.

Helmerhorst FM, Perquin DA, Donker D, Keirse MJ. (2004) Perinatal outcome of singletons and twins after assisted conception: a systematic review of controlled studies. *Bmj*. 328, 261.

Hosmer DW, Lemeshow S. Applied logistic regression. New York, NY: John Wiley & Sons Inc, 1989:140-145.

Hull MG, Glazener CM, Kelly NJ, Conway DI, Foster PA, Hinton RA, Coulson C, Lambert PA, Watt EM, Desai KM. (1985) Population study of causes, treatment, and outcome of infertility. *Bmj.* 291, 1693-7.

Hurvitch CM, Tsai CL. (1990) The impact of model selection on inference in linear regression. *Am Statist.* 44, 214-217.

Imani B, Eijkemans MJ, te Velde ER, Habbema JD, Fauser BC. (2002) A nomogram to predict the probability of live birth after clomiphene citrate induction of ovulation in normogonadotropic oligoamenorrheic infertility. *Fertil Steril,* 77, 91-7.

Kremer JAM, Beekhuizen W, Bots RSGM, Braat DDM, van Dop PA, Jansen CA, *et al.* (2002). Resultaten van in-vitro fertilisatie in Nederland, 1996-2000. *Ned Tijdschr Geneeskd.* 146, 2358-63.

Laupacis A, Wells G, Richardson WS, Tugwell P. (1994) Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. *JAMA*. 272, 234-7.

Little RJA. (1992) Regression with missing X's: a review. *J Am Stat Assoc.* 87, 1227-1237.

Martin PM, Welch HG. (1998) Probabilities for singleton and multiple pregnancies after in vitro fertilization. *Fertil Steril.* 70, 478-81.

Miller AJ. (1990) *Subset selection in regression,* Chapman & Hall, London, UK.

Moll AC, Imhof SM, Cruysberg JRM, Schouten-van Meeteren AYN, Boers M, van Leeuwen FE. (2003) Incidence of retinoblastoma in children born after in-vitro fertilisation. *Lancet.* 361, 309-10.

Oliva G, Arnau J. (2001) Assisted reproductive technology: the national and European situation. *CAHTA's newsletter*, Barcelona, Spain.

Olivennes F. (2000) Avoiding multiple pregnancies in ART. Double trouble: yes a twin pregnancy is an adverse outcome. *Hum Reprod*. 15, 1663-5.

Rowe PJ, Comhaire FH, Hargreave TB and Mellows HJ. (1993) *WHO manual for the standardized investigation and diagnosis of the infertile couple.*, Cambridge University Press, Cambridge, UK.

Snick HK, Snick TS, Evers JL, Collins JA. (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod*. 12, 1582-8.

Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. (1999) Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. *J Clin Epidemiol*. 52, 935-42.

Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JD. (2000) Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. *Stat Med.* 19, 1059-79.

Stolwijk AM, Zielhuis GA, Hamilton CJ, Straatman H, Hollanders JM, Goverde HJ, van Dop PA, Verbeek AL. (1996) Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. *Hum Reprod.* 11, 2298-303.

Stromberg B, Dahlquist G, Ericson A, Finnstrom O, Koster M, Stjernqvist K. (2002) Neurological sequelae in children born after in-vitro fertilisation: a population-based study. *Lancet*. 359, 461-5.

te Velde ER, Eijkemans R, Habbema HD. (2000) Variation in couple fecundity and time to pregnancy, an essential concept in human reproduction. *Lancet*. 355, 1928-9.

Templeton A, Morris JK, Parslow W. (1996) Factors that affect outcome of in-vitro fertilisation treatment. *Lancet.* 348, 1402-6.

van Balen F and Trimbos-Kemper TC. (1995) Involuntarily childless couples: their desire to have children and their motives. *J Psychosom Obstet Gynaecol*. 16, 137-44.

van Houwelingen HC, Le Cessie S. (1990) Predictive value of statistical models. *Stat Med.* 9, 1303-25.

Wheeler CA, Cole BF, Frishman GN, Seifer DB, Lovegreen SB, Hackett RJ. (1998) Predicting probabilities of pregnancy and multiple gestation from in vitro fertilization: a new model. *Obstet Gynecol.* 91, 696-700.

Wiegerinck MA, Bongers MY, Mol BW, Heineman MJ. (1999) How concordant are the estimated rates of natural conception and in-vitro fertilization/embryo transfer success? *Hum Reprod*. 14, 689-93.

Wølner-Hanssen P, Rydhstroem H. (1998) Cost-effectiveness analysis of in-vitro fertilization: estimated costs per successful pregnancy after transfer of one or two embryos. *Hum Reprod.* 13, 88-94.

2

Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three previous models

ABSTRACT

Background
Several models have been published for the prediction of spontaneous pregnancy among subfertile patients. The aim of this study was to broaden the empirical basis for these predictions by making a synthesis of three previously published models.

Methods
We used the original data from the studies of Eimers *et al.* (1994), Collins *et al.* (1995) and Snick *et al.* (1997) on couples consulting for various forms of subfertility. We developed a so-called three-sample synthesis model for predicting spontaneous conception leading to live birth within 1 year after intake based on the three data sets. The predictors used are duration of subfertility, women's age, primary or secondary infertility, percentage of motile sperm, and whether the couple was referred by a general practitioner or by a gynaecologist (referral status). The performance of this model was assessed according to a 'jack-knife' analysis. Because the post-coital test (PCT) was not assessed in one of the samples, a synthesis model including the PCT was based on two samples only.

Results
The ability of the synthesis models to distinguish between women who became pregnant and those who did not was comparable to the ability of the one-sample models when applied in the other samples. The reliability of the predictions by the three-sample synthesis model was somewhat better. Predictions improved considerably by including the PCT.

Conclusions
The synthesis models performed better and had a broader empirical basis than the original models. They are therefore better suitable for application in other centres.

INTRODUCTION

In clinical practice infertility is often defined as a failure to become pregnant during a 12-month period of regular, unprotected intercourse. If no pregnancy ensues, not only patients but also some doctors are inclined to think that such couples require immediate treatment because they can be considered as (almost) sterile. However, the chance of becoming pregnant, after 1 year of trying, is highly variable and depends on many female and male factors (te Velde *et al*., 2000*)*. If the spontaneous pregnancy prospects are low, immediate treatment is justified, but if they are high treatment should be withheld and the couple should be encouraged still to go for a spontaneous pregnancy. A reliable estimate of the chance of spontaneous pregnancy is therefore important for being able to make the appropriate decisions for the couple during the counselling process.

Several models for predicting spontaneous pregnancy have been developed (Comhaire, 1987) (Eimers *et al*., 1994) (Wichmann *et al*., 1994) (Collins *et al*., 1995) (Snick *et al*., 1997). We selected the three studies in which data from both partners were collected prospectively and in which the dependency of the predictors was corrected for by multivariable analysis (Eimers *et a*l., 1994) (Collins *et al*., 1995) (Snick *et al*., 1997). In the present study the data of the three selected studies were pooled to form a new data set.

Because the spontaneous pregnancy rate was found to be higher for subfertile couples referred by a general practitioner to a secondary centre than for couples referred by a gynaecologist to a tertiary centre (Wouts *et al.*, 1987) (Snick *et al.,* 1997), we assessed the importance of the care setting as a potential independent predictor for spontaneous pregnancy.

The aim of this study was to develop one or more prediction models, which more reliably predict the individual chance of pregnancy in subfertile couples and have a broader empirical base than the three individual models.

MATERIALS AND METHODS

Patients

The three studies contain couples from a Dutch University hospital (Eimers *et al*., 1994), eleven Canadian University hospitals (Collins *et al*., 1995), and a Dutch general hospital (Snick *et al*., 1997). In the following, the three studies are referred to as ‘Eimers’, ‘Collins’ and ‘Snick’, according to the name of the first author. Patients with an ovulation disorder, tubal pathology or azoospermia were excluded, because the choice of treatment is usually obvious in such patients and waiting for a spontaneous pregnancy is not a realistic option.

After these exclusions, the data set contained 996 couples from the Eimers study, 1061 couples from the Collins study and 402 couples from the Snick study, resulting in a total of 2459 couples. The pooled data set was used for the analysis of the present study. The study was approved by each institution’s research ethics review board.

Definitions and modifications of the predictive and outcome variables of the original models, and the construction of two synthesis models

The definitions of the predictors as used in the three models have been described in the original publications (Eimers *et a*l., 1994) (Collins *et al*., 1995) (Snick *et al*., 1997) and are summarized in **Table 1**. In all three studies, semen samples were collected and analysed according to the available standards of the World Health Organization (World Health Organization, 1980, 1987).

The duration of subfertility was defined by the time interval from discontinuation of contraceptive activities until registration at the fertility centre. Primary and secondary subfertility were defined as subfertility without and with a previous pregnancy respectively.

Different definitions were used for the following variables: sperm factor in the Eimers and Collins models, Post-Coital Test (PCT) in the Eimers and Snick models and outcome of success in the Eimers model as compared to the other two (**Table 1**).

The pooled individual data from the three samples were used to construct two synthesis models (a three-sample and a two-sample model) after modifying some of the predictors in order to make the data sets of the three original studies compatible. In the three-sample synthesis model, only variables were included which were available in all three samples. They include the duration and type of subfertility (primary or secondary), the woman's age and the percentage of motile sperm. In the Snick sample the percentage of motile sperm was not present in the database and therefore the percentage of progressive motile sperm had to be converted into the percentage of motile sperm. A linear model linking these two semen parameters derived from the Eimers data, was used for this conversion. The effect of the woman's age was modelled as a continuous declining fertility function, with a more rapid decline after 31 years (Van Noord-Zaadstra *et al.*, 1991). Four patients had to be excluded because two or more of the predictors were missing. For 104 patients (4%) with only one missing predictor, the missing values were imputed ('filled in'), based on the correlation with other predictors (Little, 1992). Imputation is a better method for handling missing data than simply excluding them, provided that certain conditions are met (Harrell, 2001). We used the so-called Expectation Maximisation method to estimate missing values by an iterative process (SPSS Inc., Chicago, IL). In all three data sets, the period of time (in months) couples were observed, either until a conception leading to live birth or treatment was started or until the end of the follow-up period, was available. Live birth was defined as a child still living 1 week after birth. A pregnancy leading to live birth within 1 year after intake, was taken as the outcome variable for both synthesis models.

In addition, we studied whether the referral status of the couple had a significant relation with the outcome after correcting for the other included predictors. Referral status indicates whether the couple was referred by a general practitioner (secondary-care couple) or by a gynaecologist (tertiary-care couple). All patients from the Snick study were secondary-care couples and all couples from the Collins study were considered as tertiary-care couples. Only the Eimers study included both secondary-care and tertiary-care couples and was therefore used to estimate the effect of the referral status.

The result of the PCT was not available in the Collins sample and therefore could not be included in the three-sample synthesis model. Therefore, we also developed a two-sample synthesis model including the result of the PCT based on the Snick and Eimers data sets. The PCT was scored in three categories in the Eimers model and in two categories in the Snick model (**Table 1**). Therefore, we transformed the three categories of the Eimers patients into two categories by combining the categories 'positive non-progressive' and 'negative' (abnormal) and contrasted them with 'positive progressive' (normal) according to the Snick qualifications (see **Table 1**). All other predictors in the two-sample synthesis model were the same as in the three-sample synthesis model.

For the sake of validation we also constructed three one-sample models (both with and without PCT) and three two-sample models (Snick-Eimers, Snick-Collins, Eimers-Collins both with and without referral status; the Snick-Eimers model also with PCT) to be able to perform the jack-knife analysis (see later) using the modified predictors from the three different data sets. We also analyzed whether the modifications necessary to make the data sets compatible, changed the discriminative ability of the three original models. Two score charts were constructed for easy application of the two models.

Table 1: Predictors and outcome of the three published models for spontaneous pregnancy in couples without tubal defect or ovarian disorder (Eimers *et al.*, 1994; Collins *et al.*, 1995; Snick *et al.*, 1997). Each model bears the name of the first author of the study.

	Eimers[a]	Collins	Snick
Duration of subfertility	> 1 year, continuous (y)	< 36 months, yes/no	<24 months, yes/no
Woman's age	Continuous (y)	< 30 years, yes/no	---[c]
Pregnancy history	Primary/secondary subfertility	Primary/secondary Subfertility	---[c]
Sperm factor	Percentage of motile sperm	Seminal defect[b], yes/no	---[c]
Post-Coital-Test	positive progressive (progressively motile sperm), positive non-progressive (locally motile or nonmotile sperm), negative (no sperm)	n.a.[d]	Normal (progressively motile sperm), abnormal (locally motile, nonmotile or no sperm)
Outcome	Spontaneous pregnancy within one year after intake[e]	Spontaneous pregnancy leading to live birth	Spontaneous pregnancy leading to live birth

[a]The model also contains 'presence of fertility problem in male's family' but this predictor was not considered in this study.
[b]Seminal defect was defined as one or more of sperm density <20 × 10^6/mL, <40% of sperm with progressive motility, or <40% morphologically normal sperm.
[c]Not included in the model, but available in the database.
[d]Not available in the database.
[e]Spontaneous pregnancy leading to live birth is also available in the Eimers database.

Any model based on a sample containing its own patients, will tend to give too sharp predictions when applied to other patients (Steyerberg *et al.*, 2001a). We corrected for overoptimism in the newly developed synthesis models by applying a shrinkage factor to the coefficients (Harrell *et al*., 1996).

Descriptive analyses

Differences in couple characteristics between the three samples were tested by Kruskal-Wallis test or chi square test (Altman, 1997). The effects of the predictors were compared between the three samples and expressed as fecundity ratios, which are equivalent to the hazard ratios in survival analysis. Differences in fecundity ratios and in spontaneous pregnancy chances between the samples were tested using multivariable Cox analyses (Altman, 1997).

Performance measurement

How good are the probabilistic predictions of 1 year pregnancy prospects of the two synthesis models? To obtain an unbiased estimate of their performance they should be validated in samples which were not used for their construction. We therefore applied the jack-knife principle to

estimate the performance of the three-sample synthesis model, as follows. We developed a two-sample model from two of the three samples and assessed its performance in the third truly independent sample. We repeated this procedure three times for each of the two-sample combinations. For the two-sample synthesis model with PCT, this procedure was not possible. Instead, we cross-validated the Eimers PCT model on the Snick sample and vice-versa, and compared the performance of these models with the performance of the models without PCT.

Performance was measured by assessing the ability of the model to distinguish between women who became pregnant and those who did not (discrimination) and by assessing the agreement between the observed and the predicted probabilities of pregnancy (reliability). We applied three performance measures. The c-statistic or area under the receiver operating characteristic curve (AUC) was used for assessing discrimination (Harrell *et al*., 1996). The c-statistic is the probability that from a random pair of women, the woman who first becomes pregnant had a higher predicted probability of spontaneous pregnancy.

The reliability ratio (or calibration slope) assesses reliability (Steyerberg *et al*., 2001b). A ratio of 1 indicates a perfect calibration of the joint effect of the predictors included in the model. With a ratio smaller than 1, high probability predictions are too high and low probability predictions are too low, and for a ratio greater than 1 the bias is the other way round.

The third measure assesses overall reliability of the predictions. It measures the difference in overall predicted spontaneous pregnancy rate (SPR) between the tested model and a reference model. Ideally, there is no difference (0%). In our study, the reference is always the one-sample model on its own sample.

Calculations were performed with SPSS (SPSS Inc., Chicago, IL) and S-plus (MathSoft Inc., Seattle, WA) programs.

RESULTS

The couple characteristics of the three samples differed in a number of respects. The couples from the Collins sample were older, had a longer duration of subfertility, and had also a higher percentage of motile sperm ($P < 0.001$). There were significantly fewer patients with secondary subfertility in the Eimers sample ($P < 0.001$). More patients started a treatment within the first year in the Collins sample (38%) compared to the Eimers sample (8%) and the Snick sample (15%).

The cumulative rate of spontaneous pregnancy leading to live birth within 1 year also differed significantly between the three samples (37%, 24% and 18%, respectively for Snick, Eimers and Collins, $P < 0.001$). The referral status of the couple appeared to be an independent predictor for spontaneous pregnancy after adjusting for the other characteristics ($P = 0.001$), and was therefore included in the synthesis models.

In the pooled data of the three-sample synthesis model, the age of the woman and the duration of subfertility had an adverse effect on fecundity (adjusted fecundity ratios/year = 0.95, 95% CI 0.93-0.98 and 0.83, 95% CI 0.78-0.88 respectively). Secondary female subfertility (as compared to primary) increased the chance of spontaneous pregnancy (adjusted fecundity ratio = 1.79, 95% CI 1.46-2.19). Sperm motility increased pregnancy chances by 8% for every 10% motility increase (adjusted fecundity ratio = 1.08, 95% CI 1.04-1.13). The estimates for the two-sample Snick-Eimers model of the above-mentioned variables which are used for the two-sample synthesis model with PCT, are comparable. According to the two-sample synthesis model with PCT, couples with a normal PCT had a two to three times higher chance of spontaneous pregnancy leading to live birth than couples with an abnormal PCT (adjusted fecundity ratio = 2.6, 95% CI: 2.0-3.4).

It appeared that the modifications in the predictors of the one-sample models as required for pooling did not change the discriminative ability of the three original models very much: the c-statistics of the one-sample models were almost identical for the Eimers and Collins models (c=0.69 and 0.66 respectively) and slightly improved for the Snick model (c=0.64 instead of 0.62).

Table 2 summarises the predictive performance of the various one- and two-sample models without PCT and of the final three-sample synthesis model. For the jack-knife evaluation of the synthesis model, one-sample models and all possible two-sample combinations were applied to the third truly independent sample and compared to the performance of the one-sample model when applied to its own sample, the performance of which can be considered as the reference (the highest possible performance to be expected). As expected, the performance of the one-sample models in the truly independent samples considerably decreased compared to the reference measures. However, when comparing the performance of the two-sample models in the independent sample to that of the one-sample models, the two reliability measures improved, while discrimination remained about the same.

Table 2: Performance of the one-, two- and three-sample (synthesis) models in the three samples: Snick (n=402), Eimers (n=992) and Collins (n=1061). All models use the woman's age, duration of subfertility, secondary subfertility and sperm motility. Some models also use referral status ('+ referral').

Sample	Model	Discrimination c statistic[b]	Reliability Reliability ratio[c]	Reliability Difference in SPR[d]
Snick	Snick (Ref.)[a]	0.59	1	0%
	Eimers	0.59	0.6[e]	-6%
	Collins	0.58	0.6	-11%
	Eimers-Collins	0.59	0.7	- 10%
	Eimers-Collins (+ referral)	0.59	0.7	- 6%
	Snick-Eimers-Collins (+ referral)	0.59	0.8	- 3%
Eimers	Eimers (Ref.)[a]	0.66	1	0%
	Snick	0.64	1.3	7%
	Collins	0.62	0.7[e]	-5%
	Snick-Collins	0.62	0.6[e]	- 4%
	Snick-Collins (+ referral)	0.62	0.6[e]	- 3%
	Snick-Eimers-Collins (+ referral)	0.64	0.9	-1%
Collins	Collins (Ref.)[a]	0.66[c]	1	0%
	Snick	0.65	1.5[e]	14%
	Eimers	0.62	0.8	11%
	Snick-Eimers	0.64	1	11%
	Snick-Eimers (+ referral)	0.64	1	8%
	Snick-Eimers-Collins (+ referral)	0.65	1.2	3%

[a]This model is the reference for the considered sample. For the reference model, the reliability ratio and the difference in SPR are perfect (1 and 0% respectively)
[b]c statistic (range 0.5-1): 1=optimal, higher=better discrimination.
[c]Reliability ratio: 1=optimal, closer to 1=better; lower than 1: high model predictions are too high and low model predictions are too low ; higher than 1: the other way around.
[d]Difference in Spontaneous Pregnancy Rate (SPR): 0 is optimal, closer to 0 = better. Positive/negative sign: the predictions of the tested model are higher/lower than those of the reference model.
[e]Reliability Ratio significantly different from 1 ($p<0.05$).

When referral status was included in the two-sample models the difference in SPR again improved. The performance of the three-sample synthesis model was, as expected, better than two-sample models because the individual samples were used to construct the prediction rule and to assess its performance. Both the discriminative ability and the reliability improved in the three-sample model and approached the reference measures.

Next, we developed the two-sample synthesis model with PCT, using the samples from Eimers and Snick only, because the Collins data do not include PCT. We compared their performance with and without adding the PCT. The results are given in **Table 3**. It appears that when the Snick-models are cross-validated on the Eimers sample and vice-versa, the models with PCT perform considerably better than those without PCT, on all three performance measures. In particular, the discrimination statistic is much better, indicating that the PCT contains considerable independent prognostic information.

The formulas of the three- and two-sample synthesis models are given in the Appendix. Score-charts are presented in **Figure 1**.

Table 3: Performance of models with and without PCT in the Snick and Eimers samples.

Sample	Model	Discrimination c statistic[b]	Reliability Reliability ratio[c]	Reliability Difference in SPR[d]
Snick	Snick	0.59	1	0%
	Snick (+ PCT)[a]	0.64	1	0%
	Eimers	0.59	0.6[e]	-6%
	Eimers (+ PCT)	0.64	0.8	-2%
	Snick-Eimers (+ PCT & referral)	0.64	0.9	1%
Eimers	Eimers	0.66	1	0%
	Eimers (+ PCT)[a]	0.69	1	0%
	Snick	0.64	1.3	7%
	Snick (+ PCT)	0.67	1	4%
	Snick-Eimers (+ PCT & referral)	0.69	1	0%

[a b c d e]: see Table 2.

In order to get more insight into how well the three- and two-sample synthesis models perform in a clinical situation, their performance for different prognostic categories was assessed (**Table 4a**, **b**). For purposes of comparison, the three-sample model was also considered in the Snick and Eimers samples only (**Table 4c**). The results indicate that many more patients can be classified in the extreme categories (very poor and very good prognosis) when using the two-sample synthesis model including the PCT as compared to the three-sample synthesis model without PCT.

							Score 3-sample	Score 2-sample
Woman's age (y)	21-25	26-31	32-35	36-37	38-39	40-41		
3-sample model	**0**	**3**	**7**	**10**	**13**	**15**	….	
2-sample model	**0**	**2**	**6**	**9**	**11**	**12**		….
Duration of subfertility (y)	1	2	3-4	5-6	7-8			
3-sample model	**0**	**3**	**7**	**12**	**18**		….	
2-sample model	**0**	**2**	**5**	**9**	**13**			….
Type of subfertility		Secondary	Primary					
3-sample model		**0**	**8**				….	
2-sample model		**0**	**6**					….
Motility (%)	≥ 60	40-59	20-39	0-19				
3-sample model	**0**	**2**	**4**	**6**			….	
2-sample model	**0**	**2**	**4**	**6**				….
Referral status	Secondary-care couple		Tertiary-care couple					
3-sample model		**0**	**4**				….	
2-sample model		**0**	**4**					….
Post-Coital-Test		Normal	Abnormal					
2-sample model		**0**	**14**					….
			Prognostic Index Score (Sum)				….	….

Procedure: circle the score for each of the variables, transfer to rightmost column and add to get the prognostic index score. Insert the score in the appropriate figure below in order to read off the chance of spontaneous pregnancy within 1 year resulting in live birth. (Example: according to the 3-sample synthesis model, a couple with a 28-year-old woman, with primary subfertility of 2 years duration, with 30% motile sperm, referred by a gynaecologist has a prognostic index score equivalent to 3 + 3 + 8 + 4 + 4 = 22. This score corresponds to a cumulative 12-months spontaneous pregnancy rate of 21%.)

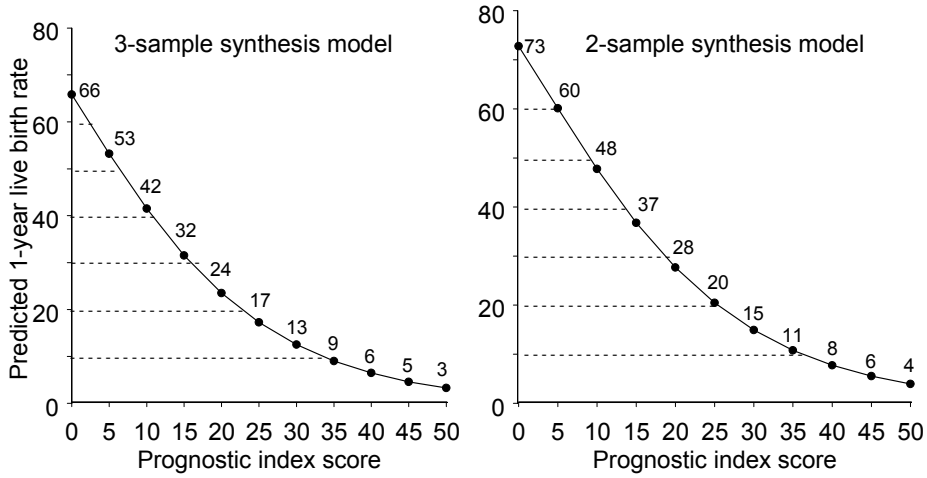

Figure 1 Score chart of the 3-sample and 2-sample synthesis models to estimate the chance of spontaneous pregnancy within 1 year after intake resulting in live birth. Upper part: calculating the score; lower part: predicting 1-year pregnancy rate.

Table 4: Predicted probabilities and observed proportions of 1 year pregnancy leading to live birth, when applying the synthesis models.

(a). Three-sample synthesis model, applied to all three samples (n=2455).

Predicted probability	N	(%)	Observed	95% CI
< 20%	991	(40%)	13%	11-16%
20-40%	1256	(51%)	29%	26-31%
> 40%	208	(9%)	50%	42-58%
All	2455	(100%)	24%	22-26%

(b). Two-sample synthesis model with PCT, applied to the Snick and Eimers samples (n=1394).

Predicted probability	N	(%)	Observed	95% CI
< 20%	616	(44%)	12%	9-15%
20-40%	441	(32%)	34%	29-39%
> 40%	337	(24%)	47%	41-52%
All	1394	(100%)	28%	25-30%

(c). Three-sample synthesis model, applied to the Snick and Eimers sample (n=1394).

Predicted probability	N	(%)	Observed	95% CI
< 20%	510	(37%)	15%	12-18%
20-40%	732	(53%)	32%	28-36%
> 40%	152	(11%)	49%	40-57%
All	1394	(100%)	28%	25-30%

DISCUSSION

In the present study we attempted to combine and improve the predictive performance of three previously published models for predicting spontaneous pregnancy, by combining their data and constructing two synthesis models. One was based on the information of each of these studies using five predictive variables: the three-sample synthesis model. The other was based on the information of two studies using six predictive variables: the two-sample synthesis model. The three original studies had so much in common, that synthesis, in our opinion, was justified. They shared a similar cohort design, had the same aim (to develop a predictive model for subfertile couples to estimate the probability of spontaneous pregnancy), were based on similar data from both partners collected during a comparable period in the seventies and eighties and the data were analyzed with the same statistical techniques (multivariable Cox regression). That the data were collected 20-30 years ago is an advantage because many more patients could be followed until a spontaneous pregnancy did (or did not) occur, simply because fewer treatment modalities were available in those days. Moreover, in contrast to the rapid development in therapeutic possibilities, there has not been a break-through in diagnostic methodology. In fact, the same information derived from the previous history and some simple diagnostic tests already performed 20–30 years ago, are still used nowadays for a first screening-type of diagnostic work-up. Therefore, these data collected in the past are at present the best available data to determine the chance of spontaneous pregnancy within a certain time period. In fact, it is questionable whether it would be possible to develop new predictive models to estimate the spontaneous chance to conceive in our Western societies, where various effective treatment modalities are available which often are applied shortly after intake.

Before we were able to construct both synthesis models, we had to make the datasets compatible. First, we excluded the patients with an ovulation disorder, tubal pathology and azoospermia in the Snick and Collins samples, because such patients were not included in the Eimers sample. The prognostic variables selected in the original models slightly differed (**Table 1**). Because the senior authors of the three originally published models took part in the present study, we could make use of the original data sets and made them compatible with one another by slight modifications. In this way the pooled data could be used now as one new data set.

The care setting between the three samples differed considerably and we found that the referral status appeared to be an independent predictor: after correcting for all other variables, a couple referred by a family doctor to a gynaecologist in a secondary centre appeared to have a better chance to conceive spontaneously than a couple referred by a gynaecologist to a tertiary centre. Apparently, apart from all known variables, there must be some concealed selection, which is reflected by the type of referral status. Apparently, some patients referred by the general practitioner to a gynaecologist of a secondary centre became pregnant (either spontaneously or after treatment) before they could be referred to a tertiary centre. Therefore, we included referral status in the newly developed synthesis models.

How to decide whether or not our synthesis models performed better or worse than the original ones? Ideally, such validation should be performed prospectively in a different population. For reasons already mentioned, to perform such a study at the present time in a large population with a sufficiently long follow-up, is almost impossible. We tried several times, so far without success. Another way to assess the performance of the synthesis models in the future, is to apply the jack-knife principle (see **Table 2**). We reasoned that if the two-sample models performed systematically better than the one-sample models, it is reasonable to assume that the three-sample synthesis model would still perform better than the two-sample models. Indeed there was a trend of better performance when comparing the two-sample models with the one-sample models. Especially the two reliability measures improved. Moreover, when adding the referral status, the performance of the two-sample models further improved the predictive performance, especially for the second reliability measure. Apart from the results of the jack-knife analysis, there is another argument in favour of the three-sample synthesis models. It is based on data derived from three different settings in two different parts of the world and collected under different circumstances. Therefore, it has a broader empirical base than the three original models.

We were not able to validate the synthesis model with PCT by the jack-knife method, because the result of the PCT was only available in two of the three databases. This is unfortunate, because the data of **Table 3** clearly demonstrate that performance, especially the discriminative power, greatly improves when adding the PCT. However, the external validation of Eimers in Snick and vice-versa is quite good, also when the PCT is added. Since, apart from the PCT, the same variables as in the three-sample synthesis model were used in the two-sample model, it is reasonable to assume that the same arguments used for the three-sample synthesis model also apply to the two-sample synthesis model. However, the argument of the wider empirical base, applies to a lesser degree to the two-sample synthesis model since it is based on two Dutch populations only.

How can the results of the predictions obtained help the clinician to counsel the individual couple? Most couples have tried for more than 1 year –often much longer– and demand immediate treatment. In their judgement, further waiting is senseless because they consider themselves as infertile. Moreover, the psychological pressure caused by feelings of uncertainty and frustration increase their desire for immediate action. In addition, most couples overestimate the success of ART and grossly underestimate the related risks. (Elster, 2000) (Olivennes, 2000) (Ericson & Kallen, 2001) (Grobman *et al.*, 2001) (Schieve *et al.*, 2002) (Stromberg *et al.*, 2002) (Land & Evers, 2003) (Moll *et al.*, 2003). The estimations of spontaneous pregnancy leading to

live birth can be a tool in advising the couple in the following manner (see **Table 4**). If the chances are low e.g. below 20%, there is no point in further waiting, and advising the couple to quickly undergo treatment is realistic. In contrast, if the chances are favourable e.g. above 40%, the couple should be strongly encouraged to wait for another year, because there is an ~ 50% chance of success. The couple should be advised that there is no ART with an equal chance of success without any risk. Further waiting is certainly worthwhile in such cases. In the middle group (above 20% and below 40%) predictions approximate the overall probability of 30-25% and the advice given depends on the balance between the probability of success, the degree of frustration and the risks of ART. These examples demonstrate that the sharp predictions –the low and high ones– are clinically useful. Predictions in the middle group hardly provide additional information for the individual couple.

The data of **Table 4** show that the two-sample synthesis model (PCT included) performs better in this respect than the three-sample synthesis model. In the latter, sharp predictions are only possible in less than half of the couples, whereas in the former this proportion is almost 70%. The superiority of the favourable predictions is noteworthy: about one quarter of all couples could be advised to wait for another year because their chances of spontaneous pregnancy leading to live birth, are almost 50%. When using the three-sample synthesis model, this advice can only be given to 10% of the couples.

We conclude that both synthesis models perform better than the originally published ones and have a broader empirical basis. They can be used both by family doctors and by gynaecologists when considering to refer couples for (further) treatment. Although far from being perfect, they contain the best prognostic information predicting spontaneous pregnancy, so far available.

APPENDIX

The general formula of a Cox model is:

$$S(t,x) = S_0(t)^{\exp(\beta_1 x_1 + \ldots + \beta_2 x_2)} = S_0(t)^{\exp(PI)}$$

where S(t,x) is the function expressing the probability that no pregnancy has occurred at time t.

The predicted probability (P) of spontaneous pregnancy within 1 year after intake leading to live birth according to the three-sample synthesis model including the referral status of the couple is:

$P = 100 \times (1 - 0.181^{\exp(PI)})$

Where the prognostic index (PI) = -0.03 × AGE1 –0.08 × AGE2 – 0.19 × duration of subfertility - 0.58 × primary subfertility + 0.008 × percentage of motile sperm - 0.25 × tertiary-care couple

AGE1 is the woman's age if the age is lower or equal to 31 years and 31 years if the age is > 31 years ; AGE2 is the difference (woman's age – 31 years) if the woman's age > 31 years and zero otherwise ; a tertiary couple is a couple referred by a gynaecologist.

The synthesis model with PCT is based on the Snick and Eimers samples. The formula of the two-sample synthesis model with PCT becomes:

$P = 100 \times (1 - 0.17^{\exp(PI)})$

and the prognostic index (PI) = -0.03 × AGE1 –0.06 × AGE2 – 0.13 × duration of subfertility - 0.44 × primary subfertility + 0.008 × percentage of motile sperm - 0.24 × tertiary-care couple – 0.95 × abnormal PCT

The result of the PCT in the initial cycle was coded as abnormal when no forward-moving sperm cell were found in the whole mucus sample.

References

Altman DG. (1997) *Practical statistics for medical research.* Chapman & Hall ed. Padstow, London, UK.

Collins JA, Burrows EA, Willan AR. (1995) The prognosis for live birth among untreated infertile couples. *Fertil Steril*. 64, 22-8.

Comhaire FH. (1987) Simple model and empirical method for the estimation of spontaneous pregnancies in couples consulting for infertility. *Int J Androl*. 10, 671-80.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD. (1994) The prediction of the chance to conceive in subfertile couples. *Fertil Steril*. 61, 44-52.

Elster N. (2000) Less is more: the risks of multiple births. The Institute for Science, Law, and Technology Working Group on Reproductive Technology. *Fertil Steril*. 74, 617-23.

Ericson A and Kallen B. (2001) Congenital malformations in infants born after IVF: a population-based study. *Hum Reprod*. 16, 504-9.

Grobman WA, Milad MP, Stout J, Klock SC. (2001) Patient perceptions of multiple gestations: an assessment of knowledge and risk aversion. *Am J Obstet Gynecol*. 185, 920-4.

Harrell FE Jr., Lee KL, Mark DB. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med.* 15, 361-87.

Harrell FE Jr. (2001) *Regression modelling Strategies: with applications to linear models, logistic regression, and survival analysis.* Springer-Verlag, New-York, USA.

Land JA and Evers JL. (2003) Risks and complications in assisted reproduction techniques: Report of an ESHRE consensus meeting. *Hum Reprod.* 18, 455-7.

Little RJA. (1992) Regression with missing X's: a review. *Journal of the American Statistical Association.* 80, 1198-202.

Moll AC, Imhof SM, Cruysberg JRM, Schouten-van Meeteren AYN, Boers M, van Leeuwen FE. (2003) Incidence of retinoblastoma in children born after in-vitro fertilisation. *Lancet.* 361, 309-10.

Olivennes F. (2000) Avoiding multiple pregnancies in ART. Double trouble: yes a twin pregnancy is an adverse outcome. *Hum Reprod.* 15, 1663-5.

Schieve LA, Meikle SF, Ferre C, Peterson HB, Jeng G, Wilcox LS. (2002) Low and very low birth weight in infants conceived with use of assisted reproductive technology. *N Eng J Med* 346, 731-7.

Snick HK, Snick TS, Evers JL, Collins JA. (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod.* 12, 1582-8.

Steyerberg EW, Eijkemans MJ, Harrell FE Jr., Habbema JD. (2001a) Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making*. 21, 45-56.

Steyerberg EW, Harrell FE Jr., Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. (2001b) Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol.* 54**,** 774-81.

Stromberg B, Dahlquist G, Ericson A, Finnstrom O, Koster M, Stjernqvist K. (2002) Neurological sequelae in children born after in-vitro fertilisation: a population-based study. *Lancet.* 359, 461-5.

te Velde ER, Eijkemans R, Habbema HD. (2000) Variation in couple fecundity and time to pregnancy, an essential concept in human reproduction. *Lancet.* 355, 1928-9.

Van Noord-Zaadstra BM, Looman CW, Alsbach H**,** Habbema JD, te Velde ER, Karbaat J. (1991) Delaying childbearing: effect of age on fecundity and outcome of pregnancy. *Bmj.* 302, 1361-5.

Wichmann L, Isola J, Tuohimaa P. (1994) Prognostic variables in predicting pregnancy. A prospective follow up study of 907 couples with an infertility problem. *Hum Reprod*. 9, 1102-8.

World Health Organization. (1980) *Laboratory Manual For The Examination of Human Semen and Semen-Cervical Mucus Interaction*. Press Concern, Singapore.

World Health Organization (1987) *Laboratory Manual For The Examination of Human Semen and Semen-Cervical Mucus Interaction*. Cambridge University Press, Cambridge.

Wouts MH, Duisterhout JS, Kuik DJ, Schoemaker J. (1987) The chance of spontaneous conception for the infertile couple referred to an academic clinic for reproductive endocrinology and fertility in The Netherlands. *Eur J Obstet Gynecol Reprod Biol.* 26, 243-50.

3

Validation of a model predicting spontaneous pregnancy among subfertile untreated couples

ABSTRACT

Objective
To provide an external validation of the Eimers model which predicts spontaneous pregnancy among subfertile couples, within the first year after the definitive establishment of the diagnostic category.

Design
Live birth rates predicted by an adapted version of the Eimers model were tested against observed live birth rates in a Canadian cohort study.

Setting
Fertility clinics in University medical Centres
Patients: 1061 couples consulting for subfertility due to cervical hostility, male subfertility, or unexplained subfertility.

Interventions
None

Main outcome measure(s)
The discriminative ability and reliability of the predictions from the model

Results
The live birth rate was lower in the Canadian population than in the Eimers population (p=0.007). Overall, the prognostic effect of the predictors did not differ significantly in both populations (p=0.91). The model showed moderate predictive power in the Canadian population (c index = 0.62). With adjustment of the average live birth rate, the reliability of the model was satisfactory.

Conclusions
The Eimers model gave reliable spontaneous pregnancy predictions in the Canadian validation population after adjustment of the average live birth rate.

INTRODUCTION

In clinical medicine, studies focusing on the generalisability of prognostic models are infrequent. Most of the time, the original publication about the development of a predictive model is not followed by an external validity study (Smeenk, 2000). This can be due to the difficulty to find a suitable validation population, or to an absence of consensus on what the relevant prognostic factors are. However, the necessity to test prognostic models in other populations has been often emphasised (Harrell, 1996) (Justice, 1999) (Altman, 2000). Untested prognostic models can show disappointing performance when used among patients from a different but plausibly related population (Stolwijk, 1996).

In this perspective, we aimed to externally validate the model developed by Eimers *et al.* (Eimers, 1994) (the 'Eimers model') for predicting the chance of subfertile couples to conceive spontaneously during the first year after intake i.e after the diagnostic category has been definitively established. This model was developed in a cohort of couples consulting the department of gynaecology of the University Hospital Utrecht for subfertility. The internal validity of this model was evaluated and found to be adequate.

To assess how this prognostic model performs in an other population, we used a previously published independent data set from the Canadian Infertility Therapy Evaluation Study (CITES) (Collins, 1995). This data set represents an interesting opportunity because of the setting of the Canadian population (large cohort of subfertile couples, referred to University Hospitals) and because the period of data collection was approximately the same as in the study done to develop the model (1984 to 1987 for the Canadian study, 1974 to 1984 for the Dutch study). We applied the model to these patients to test its transportability. Emphasis is on the ability to discriminate between couples getting and not getting a live child, and on the agreement between the predicted and observed probabilities of spontaneous pregnancy leading to live birth.

MATERIALS AND METHODS

The Eimers model

The Eimers model (Eimers, 1994) was developed in 1994 to estimate the chance of spontaneous pregnancy among subfertile couples. This model is only appropriate for couples with cervical hostility, male subfertility or unexplained subfertility, but not for couples with azoospermia, ovulatory problem or tubal disease. The predicted probability (P) of spontaneous pregnancy within one year after intake according to the Eimers model is:

$$P = 100 \text{ X } (1 - 0.81^{\exp(PI)})$$

Where the prognostic index (PI) = - 0.029*woman's age - 0.12*duration of subfertility + 0.55*secondary subfertility - 0.37*presence of subfertility problem in the male's family + 0.75* positive non-progressive PCT + 1.46*positive PCT + 0.013*percentage of motile sperm

PCT is the best result of the Post-Coital Test (PCT) during the initial cycle.

The validation population

The validation was performed on patients from the Canadian Infertility Therapy Evaluation Study (CITES). Details on data collection and description of patient characteristics have been published previously (Collins, 1995). Couples with azoospermia, ovulatory problems or tubal disease were excluded because the Eimers model does not cover these causes. After these exclusions, the validation population was composed of 1061 couples. The registration date was the start of the follow-up.

The study was approved by each institution's research ethics review board.

Use of an alternate model

Because the CITES study did not have a PCT and information about fertility problems in male's family, we had to use an alternate model with four predictors: age of the woman, duration of subfertility, type of subfertility and percentage motile sperm. To test whether the absence of the PCT could be compensated by one more sperm parameter, we tested two extensions of the alternate model: with the percentage of sperm cells with normal morphology and with the sperm density (number of sperm cells per ml). Assessment of cell morphology and sperm density was according to the recommendations of the World Health Organization at the time of the study (World Health Organization, 1980).

We compared the predicted probabilities of the alternate models with those of a model including the six variables of the original model, by calculating Pearson's correlation coefficient. Live birth was taken as the outcome variable because the outcome of the original model, "pregnancy", was not available in the validation database. "Live birth" was defined as a living child at the time of hospital discharge after delivery.

Statistical analysis

The Eimers model was based on a Cox analysis. The model was developed in 1994 and was originally validated using the 'split-half' method. Since then, techniques to assess the validity of a model have evolved and it is considered appropriate to correct models for over-optimism (*i.e.* the phenomenon that coefficients in regression models are fitted too extremely during the development phase for predictive purposes) (Steyerberg, 2001). We carried out this correction by applying a shrinkage factor (Van Houwelingen, 1990) of 0.77 to the coefficients of the alternate models. We used this corrected version of the model in all subsequent analyses. In the appendix, we show a score chart of the original Eimers model, corrected for over-optimism (shrinkage factor of 0.79).

In order to compare the effects of the woman's age, duration of subfertility, type of subfertility, and percentage of motile sperm in the development and the validation populations, we fitted a Cox regression model with these 4 predictors to both populations. We analysed the significancy of any differences in the hazard ratios between the two populations using t-tests.

We evaluated the predictive performance of the model by two measures: the ability to discriminate between women getting and not getting a child (discrimination), and the agreement between the predicted and observed probabilities of pregnancy leading to live birth (reliability). The discrimination of the model was evaluated by the c index. The c index is identical to the area under a receiver operating characteristic (ROC) curve for binary outcomes. It can be interpreted as the probability, for a random pair of women, that the woman with the poorest chance of pregnancy leading to live birth is also the woman with the longest time to pregnancy leading to live birth. The c index ranges from 0.5 (no discrimination) to 1.0 (perfect discrimination).

We examined to what degree the predicted chance of spontaneous pregnancy leading to live birth agreed with the observed births, at one year. First, we grouped patients by predicted probability categories of 20%. Then, we plotted the mean of the predicted probability against the observed live birth rate obtained by Kaplan-Meier analysis, for each category. In addition, we tested whether the joint effect of the predictors was miscalibrated, by studying the slope of the linear predictor ('calibration slope') (Steyerberg, 2001). A slope of 1 means a perfect agreement between the predicted and observed probabilities. A slope significantly different from 1 indicates that the joint effect of the predictors is not correctly specified.

The average live birth rate (or baseline probability) in Cox regression gives the predicted live birth rate for a referent patient profile after correcting for the influence of the predictors in the model. It may differ between the development and the validation populations, even if the effects of the predictors are correctly specified.
We hypothesised that the baseline live birth probability in the development population did not differ from the baseline probability in the validation population. To test this hypothesis, we applied a Cox model with the predictors of the alternate model and an extra covariate indicating the origin of the population (development or validation population) to the pooled data from both populations (996 + 1017 couples). Statistical significance of the extra covariate indicates that there is a difference of baseline probabilities between the two populations. In that case, it is sensible to adjust the model by replacing the original baseline probability by the baseline probability of the validation population. Otherwise, reliable predictions in the validation population will not be possible

RESULTS

Among the 1061 Canadian couples, 1017 (96%) had values for all four predictors. 44 couples had missing values for the sperm motility and were excluded from the analyses.

Correlation between the predictions of the original and the alternate model
In **Table 1**, the predicted probabilities according to the alternate model with four variables are compared with the predicted probabilities done by the model with the six original variables. Pearson's correlation coefficient between the predictions of the two models equalled 0.80 in the development population (R^2=0.63). A difference of 10% or more between the predictions of the 4-variable and the 6-variable model was observed for 12.5% of the couples. Extending the alternate model with sperm density or with the percentage of sperm cells with normal morphology as a compensation for the PCT did not improve the correlation. For the analyses, we therefore used the four-variable model.

Table 1: Percentages of couples by category of predicted Live Birth Rates (LBR) at one year, in the development population (n=996), provided by the 4-variable alternate model and by the 6-variable original model. The grey cells indicate interval-correspondence between both predictions.

Predicted LBR with the 6-variable model

>50%				0.1%	0.1%	0.1%
40-50%			0.3%	2.4%	0.2%	
30-40%		0.2%	8.2%	0.8%	0.5%	0.1%
20-30%		8.1%	8.3%	2.7%	0.1%	
10-20%	2.7%	19%	11.5%	1.5%		
0-10%	21.4%	11.4%	0.3%			
	0-10%	10-20%	20-30%	30-40%	40-50%	>50%

Predicted LBR with the 4-variable model

Comparison of the couples characteristics in the development and the validation populations

The couples characteristics are shown in **Table 2**. The woman's age, the duration of subfertility and the number of secondary subfertility in the validation and development population were quite similar. However, the percentage of motile sperm was on average higher in the validation population. Treatment in the first year was much more often initiated in the validation population (38% in the validation population versus 8% in the development sample), which explains why few spontaneous live births were conceived in one year.

Table 2: Couples characteristics in the development and in the validation population. The development population is from a Dutch University Hospital, 1974-1984 (n=996), (Eimers, 1994). The validation population is from the CITES study, and concerns 11 Canadian referral hospitals, 1984-1987 (n=1061), (Collins, 1995).

		Median	Lower Quartile	Upper Quartile	SD	Missing (%)
Woman's age (years)	Development	29	26	31	3.7	0
	Validation	30	27	33	4.1	0
Duration of subfertility (years)	Development	3	2	5	2.4	2.1
	Validation	3	2	5	2.3	0
Motility in the 1st sperm analysis (%)	Development	40	20	60	20	0
	Validation	55	38	70	24	4.1
Follow-up (months)	Development	17	7	48	33	3.6
	Validation	8	3	16	19	0

		N	%	Missing (%)
Secondary subfertility	Development	170	17	0
	Validation	236	22	0
Spontaneous live birth during total follow-up	Development	355	36	0
	Validation	175	16	0
Spontaneous live birth within one year	Development	207	21[a]	0
	Validation	133	13[a]	0
Treatment started within 12 months	Development	80	8	3.1
	Validation	406	38	1.8

[a] Estimated by Kaplan-Meier analyses

Comparison of the predictors

The effects of the four predictors (expressed as hazard ratios for live birth) differed somewhat (**Table 3**). The effects of the age of the woman, of the duration of subfertility and of the type of subfertility were somewhat stronger in the validation population than in the development population ($p>0.05$). In contrast, motility had a statistically significantly smaller effect ($p=0.005$) on the chance of live birth in the validation population than in the development population: 10% more motility increased the chance of live birth by 15% in the development population and by 4% in the validation population.

Table 3: Hazard ratios of live birth (HR) among the development and the validation populations.

	Development population		Validation population	
	HR	95% CI	HR	95% CI
Woman's age (year)	0.98	0.95 - 1.01	0.93	0.89 – 0.97
Duration of subfertility (year)	0.90	0.84 - 0.95	0.78	0.70 - 0.88
Female subfertility (secondary)	1.54	1.20 - 1.97	1.82	1.25 - 2.66
Sperm motility (per 10%)	1.15	1.09 - 1.22	1.04	0.97 - 1.12

Predictive performance

For the four-variable model we found a c index of 0.62 within the validation population compared to a c index of 0.66 in the development population.

The average live birth rate was significantly lower in the validation population than in the development population (p=0.007). **Figure 1A**, shows the reliability of the model without adjustment of the average live birth rate i.e. using the average live birth rate of the development population. The curve was lower than the ideal diagonal line, showing that the predicted live birth rates were higher than the observed live birth rates in the validation population. The joint effect of the predictors was well specified, since the slope did not differ statistically from 1 (calibration slope 0.98 , p=0.91).

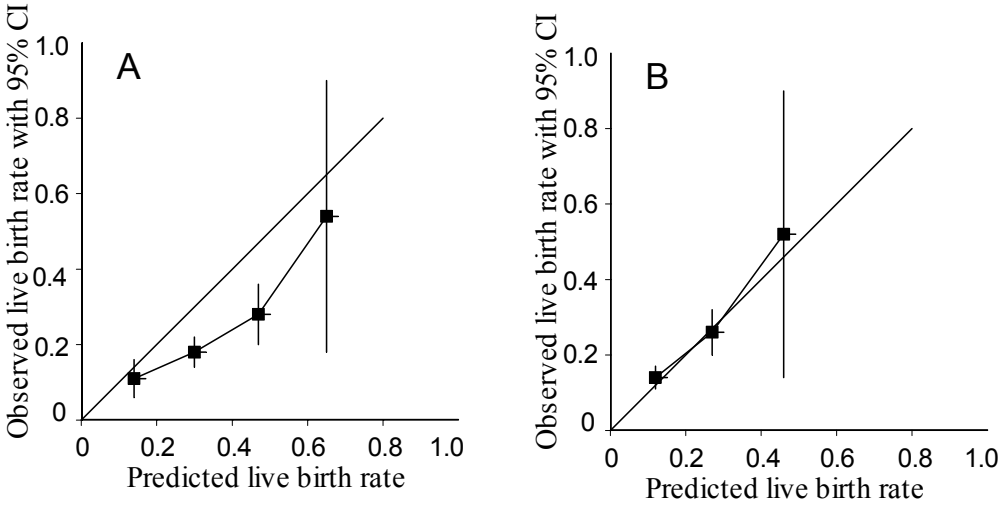

Figure 1 Reliability of the model (A) without and (B) with adjustment on the live birth rate of the validation population. The squares correspond to the groups formed by pooling according to the predicted probabilities. The vertical lines are the 95% confidence intervals of the observed live birth probabilities, estimated by Kaplan-Meier analysis.

Figure 1B shows the improvement when an adjustment of the average live birth rate was performed (using the average live birth rate of the validation population). The curve was closer to the ideal diagonal line. Moreover, the range of predicted live birth rate was narrower with adjustment (0.12 to 0.46) than without adjustment (0.14 to 0.65). This difference in average live birth rate can be illustrated by an example. A couple formed by a 29-year-old woman with primary subfertility for 3 years and a man with 50% motile sperm had a predicted live birth rate of 24% without adjustment of the average live birth rate and 15% with adjustment, at 12 months.

DISCUSSION

The external validation of the Eimers model on a Canadian patient population turned out to be a far from trivial exercise. The original model included six predictors of which four were available in the Canadian validation population. The joint effect of these four predictors did not differ statistically between the development and the validation populations (p=0.91) whereas the effect of the individual predictors differed. The stronger effect of three clinical predictors -woman's age, duration of subfertility and type of subfertility (primary or secondary)- in the validation population was balanced by the weaker effect of the percentage of motile sperm.

PCT results and information about fertility problems in male's family were not available in the validation data set. The correlation between the predictions from the 6-variable model and the 4-variable model was reasonable (0.80) and could not be improved by using an extra predictor (the percentage of sperm cells with normal morphology or the sperm density) as compensation for the absent PCT. We are currently planning a study in which data are collected on all six predictors with the explicit purpose of prospective validation of the prediction model.

The outcome variable "live birth" was available in the validation population but the original outcome variable "ongoing pregnancy" not. In the development data set however, both outcomes were available and analyses show that 89% (191/215) of the ongoing spontaneous pregnancies within the first year after intake ended in live births. The difference in average live birth rate between the two populations may be explained by differences in background variables such as different referral patterns between the two populations. For example, because of long distances, probably more patients were directly referred by a specialist in the Canadian population. In the Dutch population, 34% of the patients were referred by a family doctor. We speculate that these differences might have an (unknown) impact on the degree and seriousness of the fertility problem.

Difference in tobacco use between the two populations could have also partly explained the lower average live birth rate in the validation population, but we observed on the contrary that the percentage of smoking women was higher among the women of the Eimers study (48%) than the percentage of smoking women in the general Canadian population for the same period (36%) (Rootman, 1988).

Absence of PCT results in the alternate model resulted in a reduced discriminative ability in the development population (c index = 0.66 vs 0.71 in the 6-variable model). We expect therefore that the discriminative ability of the model would also have been better in the validation population than the current c index of 0.62 when all six predictors could have been used. PCT was not routinely done in Canada. Currently, although the effectiveness of the PCT is disputed (Oei, 1998) (Glazener, 2000) (Hull, 1999) (Cohlen, 1999), clinicians often use it in the evaluation of a subfertile couple. The selection of the PCT in the Eimers and Snick models (Snick, 1997) supports its importance.

In recent years, few studies focused on the validation of prognostic models in the field of infertility. Snick *et. al* (Snick, 1997) also encountered the problem of PCT unavailable in the validation dataset. They used three variables ("semen defect", "woman's age <30 years" and the type of infertility) to compensate for the absence of PCT results. As in our study, their alternate model had a lower discriminative ability than the original model.

Models usually loose discriminative power when they are applied in an external population. The Stolwijk model (Stolwijk, 1996), predicting pregnancy after in-vitro fertilization, had a c index of 0.67 in the development population, and of 0.61 in a validation population. The c index score 0.62 of our alternate model applied to the validation population is quite comparable to other

reported c-values for the same prediction problem: 0.65 for Collins and 0.67 for Snick (Snick, 1997). This suggests an inherent difficulty in prediction of spontaneous pregnancy.

In a review on generalizability of prediction rules, Justice et al. distinguishes five types of external validation: methodological, spectrum, geographic, follow-up period and historical (Justice, 1999). In our case, the variables had similar definitions. There was a difference in spread of the values of the motility in the first semen analysis between the two populations, although both studies used the same WHO protocol. This laboratory variable is less objective than predictors such as the woman's age and duration of subfertility. Differences in motility scores may be explained by differences in operationalization of the motility concept. The geographic component has already been addressed in connection with spectrum differences as a possible explanation for the difference in baseline live birth rates.

There was also an age difference between the two populations. The validation population was older, with 37% of the women older than 31 years, versus 22% in the development population. 31 years is the critical age above which female fecundity has been found to start decreasing (Van Noord-Zaadstra, 1991). In the model, the effect of woman's age is considered as a linear variable. The fact that a higher proportion of women was older than 31 years in the validation population could partially explain that the model performed not so well in the validation population. An adaptation of the Eimers model has been previously proposed to better express the effect of the age for women older than 35-year (Habbema, 1997).

Also, a much larger percentage of couples started a treatment within the first year in the validation population (38% versus 8%). These patients were considered as censored observations. As emphasised by Graf *et al.* (Graf, 1999), censoring is a problem for the assessment of the accuracy of a predictive model. Concerning the historical transportability, the periods of data collection were quite similar for the development population (1974 to 1984) and the validation population (1984 to 1987). However, we do not know whether the Eimers model will also be valid for current patients. Age of the woman, percentage of motile sperm and duration of subfertility are predictors included in the model, which permit to take into account recent trends, such as postponed childbearing, decreased semen quality in some countries, or the fact that procreation is more and more often planned within a short period.

In conclusion, the Eimers prediction model had to be adapted in several aspects before it resulted in satisfactory application in an external population. As a by-product of this validation study, we came up with an adapted version of the original model, which is presented in the appendix. This updated model uses the shrunken coefficients and takes account of the rapidly declining female fecundity after 31 years. This updated model is therefore in our opinion more suitable for prediction in new patients than the original version.

APPENDIX

Score chart of the internally validated Eimers model to estimate the probability of spontaneous pregnancy in subfertile couples within one year.

						Infertility Score [a]
Woman' s age (years)	21-25	26-31	32-35	36-39	40-41	
Score	0	2	5	7	10	--
Duration of infertility (years)	1	2	3-4	5-6	7-8	
Score	0	1	3	7	10	--
Female infertility	Secondary		Primary			
Score	0		6			--
Fertility problems in male's family	No		Yes			
Score	0		4			--
Post-Coitum Test	Progressive		Non progressive		Negative	
Score	0		8		17	--
Motility (%)	≥ 60	40-59	20-39	0-19		
Score	0	2	5	8		--
					Prognostic Index (Sum score)	--

[a] Circle the infertility score for each of the parameters and add them to the prognostic index. Use the curve in **Figure 2** to estimate the PR within 1 year.

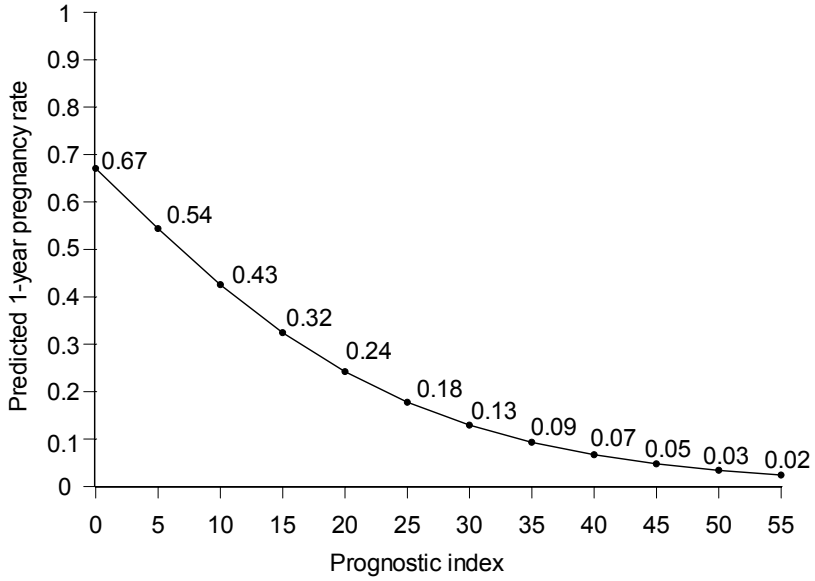

Figure 2 Relation between prognostic index and pregnancy within 1 year.

References

Altman DG, Royston P. (2000) What do we mean by validating a prognostic model? *Stat Med*. 19, 453-73.

Cohlen BJ, te Velde ER, Habbema JD. (1999) Postcoital testing. Postcoital test should be performed as routine infertility test. *Bmj*. 318, 1007; discussion 1008-9.

Collins JA, Burrows EA, Willan AR. (1995) The prognosis for live birth among untreated infertile couples. *Fertil Steril*. 64, 22-8.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD. (1994) The prediction of the chance to conceive in subfertile couples. *Fertil Steril*. 61, 44-52.

Glazener CM, Ford WC, Hull MG. (2000) The prognostic power of the postcoital test for natural conception depends on duration of infertility. *Hum Reprod*. 15, 1953-7.

Graf E, Schmoor C, Sauerbrei W, Schumacher M. (1999) Assessment and comparison of prognostic classification schemes for survival data. *Stat Med*. 18, 2.

Habbema JDF, Steyerberg EW. *Predictive tools for clinical decision support.* In: van Bemmel JH, Musen MA eds. Handbook of Medical Informatics. Chapter 18. Houten/Diegem, 1997.

Harrell FE, Jr., Lee KL, Mark DB. (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 15, 361-87.

Hull MG, Evers JL. (1999) Postcoital testing. Criterion for positive test was not given. *Bmj*. 318, 1008-9.

Justice AC, Covinsky KE, Berlin JA. (1999) Assessing the generalizability of prognostic information. *Ann Intern Med.* 130, 515-24.

Oei SG, Helmerhorst FM, Bloemenkamp KW, Hollants FA, Meerpoel DE, Keirse MJ. (1998) Effectiveness of the postcoital test: randomised controlled trial. Bmj. 317, 502-5.

Rootman I, Warren R, Stephens T, Peters L. *Canada's Health Promotion Survey*. Technical Report. Eds. Minister of Supply and Services Canada, Otawa, 1988.

Smeenk JM, Stolwijk AM, Kremer JA, Braat DD. (2000) External validation of the Templeton model for predicting success after IVF. *Hum Reprod.* 15, 1065-8.

Snick HK, Snick TS, Evers JL, Collins JA. (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod.* 12, 1582-8.

Steyerberg EW, Eijkemans MJ, Harrell FE Jr, Habbema JD. (2001) Prognostic modelling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making.* 21, 45-56.

Steyerberg EW, Harrell FE Jr, Borsboom GJJM, Eijkemans MJC, Vergouwe Y, Habbema JDF. (2001) Internal validation of predictive models: Efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol*. 54, 774-81.

Stolwijk AM, Zielhuis GA, Hamilton CJ, Straatman H, Hollanders JM, Goverde HJ, van Dop PA, Verbeek AL. (1996) Prognostic models for the probability of achieving an ongoing pregnancy after in-vitro fertilization and the importance of testing their predictive value. *Hum Reprod.* 11, 2298-303.

Van Houwelingen JC, Le Cessie S. (1990) Predictive value of statistical models. *Stat Med.* 9, 1303-25.

Van Noord-Zaadstra BM, Looman CW, Alsbach H, Habbema JD, te Velde ER, Karbaat J. (1991) Delaying childbearing; effect of age on fecundity and outcome of pregnancy. *Bmj.* 302, 1361-5.

World Health Organization. *Laboratory manual for the examination of human semen and semen-cervical mucus interaction.* In: Belsey MA, Eliasson R, Gallegos AJ, Moghissi KS, Paulson CA, Prasa MRN, editors. Singapore: Press Concern, 1980.

4

Prospective validation of two models predicting pregnancy leading to live birth among untreated subfertile couples

ABSTRACT

Objective
Models predicting clinical outcome need external validation before they can be applied safely into daily practice. This study aimed to validate two models for the prediction of the chance of treatment independent pregnancy leading to live birth among subfertile couples.

Methods
The first model uses the woman's age, duration and type of subfertility, percentage of progressive motility and referral status. The second model in addition uses the result of the post-coital test. For validation, these characteristics were collected prospectively in two University hospitals for 302 couples consulting for subfertility. The models' ability to distinguish between women who became pregnant and women who did not (discrimination) and the agreement between predicted and observed probabilities of treatment independent pregnancy (calibration) were assessed.

Results
The discrimination of both models was slightly lower in the validation sample than in the original sample which provided the model. Calibration was good: the observed and predicted probabilities of treatment independent pregnancy leading to live birth did not differ for both models.

Conclusions
The chance of pregnancy leading to live birth was reliably estimated in the validation sample by both models. The use of PCT improved the discrimination of the models. These models can be useful in counselling subfertile couples.

INTRODUCTION

When counselling a subfertile couple, the decision to treat should be based on the pregnancy prospects without treatment of this specific couple and not on a uniform criterion i.e. not having conceived within ≥ 12 months of unprotected intercourse. Treatments such as intrauterine insemination or IVF should be proposed only to couples with a sufficiently low probability of treatment independent pregnancy in order to avoid unnecessary medication and subsequent complications such as twin pregnancies, which are in it-self associated with higher peri-natal mortality rates and more long term health and psycho-social sequaelae (ESHRE, 2000; Jones, 2003; Hansen, 2002; Stromberg, 2002; Moll, 2003). Treatment independent pregnancy rates within 1 year after intake i.e after the diagnostic category has been definitively established, from 0 to 50% and more have been reported among couples with subfertility due to unexplained subfertility, mild male factor or cervical hostility (Eimers, 1994; Collins, 1995; Snick, 1997; Hunault, 2004). An accurate estimation of the chance of treatment independent pregnancy for an individual couple is hence important, and may be provided by a prediction model.

We have previously developed two models to improve the prediction of treatment independent pregnancy (Hunault, 2004). These models were based on three previous studies and therefore called "synthesis models". The population in which the two models were developed included couples consulting for various forms of subfertility (unexplained subfertility, subfertility due to cervical hostility or to a mild male factor), and referred by a general practitioner or by a gynecologist. The first model includes the following predictors: the woman's age, duration of subfertility, type of subfertility (primary or secondary), percentage of motile sperm cells and referral status of the couple. The second model includes the same predictors, plus the result of the best post-coital-test (PCT). In clinical practice, such a model could be used to categorize a couple as having a poor, intermediate or good chance of conceiving without treatment. If the chance is poor, the couple should be advised to undergo treatment. If the chance is high, the couple should be encouraged to wait with treatment. If the chance is intermediate, the advice could be driven by the preferences of the couple concerning effectiveness, costs and risks of treatment.

The internal validity of the models has been found to be satisfactory but an internally validated model can easily produce poor predictions in future patients or in patients from other centers (Justice, 1999). The aim of the present study was to externally validate the two treatment independent pregnancy prediction models i.e. to assess whether these models predict well in a sample of subfertile patients different from the sample of patients used to develop the models.

METHODS

Patients

This study was approved by the local institutional medical and ethical review boards and written informed consent was obtained from all participants.

The standardized initial screening included the clinical examination of both partners, a (i.e. the first) semen sample analysed according to WHO criteria (WHO, 1999), recording of a basal body temperature (BBT) chart, a mid-luteal progesteron determination, a post-coital test (PCT), a transvaginal ultrasound and serum *Chlamydia* antibody testing. A hysterosalpingography (HSG), or a laparoscopy with tubal patency testing was performed if *Chlamydia* antibodies were present or in case of risk factors for tubal pathology (ectopic pregnancy or abdominal surgery history).

Three hundred and two couples from the Rotterdam and Utrecht University hospitals were prospectively enrolled in the study between January 1998 and August 2002. Inclusion criteria were: (a) woman's age below 40 years, (b) duration of subfertility of ≥ 1 year, (c) cycle duration >21 and < 35 days, (d) normal physical examination (no body shape and stature suggesting Turner's syndrome, BMI < 30, normal secondary sexual characteristics, no abnormal findings on pelvic and gynecological examination) and ultrasonography (no uterus abnormalities) (e) serum FSH concentrations within normal limits (1-10 IU/l) (f) normal midluteal serum progesteron (≥ 28 nmol/L) (g) subfertility due to mild male-, cervical- or unexplained-subfertility. Mild male factor was defined as a total motile count (TM) of at least 7 x 10^6. Semen analysis was considered normal if sperm concentration was > 14×10^6/ml, if grade A progressive motility was > 18% and if the percentage of normal morphology was > 8% (strict Kruger criteria, Ombelet, 1997). The PCT was considered as positive if on average one progressively moving spermatozoa was found in at least 6 High Power Field (WHO, 1999). In case of a negative result, timing of post coïtal testing was done using transvaginal ultrasound. Subfertility was attributed to cervical hostility if a correctly timed PCT revealed no progressive motile spermatozoa in optimal cervical mucus in combination with normal semen samples, or if PCT was repeatedly negative regardless of the condition of the cervical mucus (WHO, 1999). The diagnosis of unexplained subfertility was made when all investigations were normal. Couples with uni and/or bilateral tubal disease, ovulatory disorder (abnormal serum progesteron in the mid-luteal phase) or endocrine disorders (abnormal prolactin or thyroid malfunction) or males with azoospermia were excluded. In summary, the inclusion and exclusion criteria of the population in which the models were validated were the same as of the population in which the models were developed, except the semen criteria, which were stricter in the validation sample: in the development sample, only men with azoospermia were excluded, whereas men with severe male factor were also excluded in the validation sample.

All patient characteristics were collected prospectively: the woman's age, duration of subfertility, type of subfertility (primary or secondary), percentage of motile sperm in the first semen analysis, result of the best PCT during the initial screening and referral status (whether the couple was referred by a general practitioner or by another gynecologist). The following definitions were used. Duration of subfertility: the interval in years from discontinuation of contraceptive activities until registration at the fertility center; primary subfertility: women who never conceived; secondary subfertility: subfertility after prior conceiving for the women; live birth: living child at the time of hospital discharge after parturition. The number of observation months of couples was counted until either conception leading to live birth, or treatment was started, or because the study stopped before the end of their follow-up.

Analysis

Differences in couple characteristics between the validation sample and the original sample that provided the model, were tested by Kruskal-Wallis test for continuous variables and chi-square test for categorical variables. The prognostic effects of the patient characteristics included in the model were studied in the validation sample and expressed as hazard ratios for live birth, using a multivariable model.

The synthesis models we aimed to validate are Cox models predicting the chance of treatment independent pregnancy leading to live birth within 1 year after inclusion (Hunault, 2004). The model without PCT has been developed using data on 2459 couples obtained by pooling the data of three studies, the Eimers, Collins and Snick studies (Eimers, 1994; Collins, 1995; Snick, 1997). The model with PCT is based on the data of two studies (the Eimers and Snick studies) since the PCT was not investigated in the third study (the Collins study). The formula's of the

models are given in the appendix. The probability of live birth was calculated for each couple of the validation sample, according to both models.

The calibration and the discrimination of the models were assessed to test the validity of the model in the validation sample. Calibration refers to the agreement between predicted and observed probabilities of treatment independent pregnancies whereas discrimination is the model's ability to distinguish between the women who became pregnant and those who did not.

Calibration was assessed graphically by plotting the observed 1-year live birth rate against the predicted 1-year live birth probability in a calibration plot (Miller, 1993). We statistically tested whether the mean predicted and observed probabilities of pregnancy leading to live birth were different. Furthermore, we tested whether the predictions were too extreme (too low estimates for low probabilities and too high estimates for high probabilities), and whether the observed and predicted ongoing pregnancy rates were systematically different (Harrell, 1996). The discriminative ability of the model was quantified by the c statistic, which is equivalent to an area under the ROC curve. A c statistic ranges from 0.5 (no discriminative power) to 1 (perfect discrimination). The c statistic is the probability that from a random pair of women, the one with the highest predicted probability of treatment independent pregnancy leading to live birth will be the first to succeed.

In order to assess and compare the clinical usefulness of the two models, the patients of the validation sample were grouped in three categories of predicted chances of treatment independent pregnancy leading to live birth within 1 year, <20%, 20-40%, 40% and more. Clinical usefulness of a model was expressed as the percentage of patients assigned by the model to the two extreme categories.

Calculations were performed using commercially available software packages (SPSS Inc., Chicago, IL, USA, 1999 and S-plus 2000, MathSoft Inc., Seattle, WA, USA, version 2000). A p value < 0.05 was considered to indicate statistical significance.

RESULTS

Three hundred and two couples were included (213 patients from Utrecht and 89 couples from Rotterdam). The chance of pregnancy leading to live birth did not differ significantly between the Utrecht and Rotterdam clinics (p=0.15). We pooled the two data sets into the 'validation sample' to assess the validity of the synthesis models. The couple characteristics of the development and validation samples are summarized in **Table 1**. Women from the validation sample were older but their duration of subfertility was shorter compared to the women from the development sample. Secondary subfertility, normal PCT and referral by a general practitioner were more frequent in the validation sample. The time until treatment was much shorter in the validation sample, in which 71% of couples started a treatment within the first year after intake, compared to only 23% in the development sample.

The live birth rate estimate at 12 months did not differ significantly between the validation sample and the development sample (24% and 31%, p=0.12). The effects of the predictors were in the same direction in the development and validation samples. In the validation sample, couples with a normal PCT had a nearly four times higher chance of treatment independent pregnancy leading to live birth than couples with an abnormal PCT after adjusting for the woman's age, primary subfertility, duration of subfertility, motility and referral status (Hazard Ratio equal to 3.7, 95% CI: 1.09-12.7).

The c statistic was 0.59 (95% CI: 0.46-0.73) and 0.63 (95% CI: 0.51-0.75) for the synthesis models without and with PCT respectively, when used in the validation sample. The two c-statistics differed statistically (p=0.04). **Figure 1** shows that both models were well calibrated.

On average, the observed probabilities were closest to the ideal diagonal line for the model with PCT. The mean predicted and observed probabilities of live birth did not differ significantly for the models without and with PCT (P=0.3 and 0.6 respectively). The predictions were not statistically too extreme (neither too low estimates for low probability-patients, nor too high estimates for high probability-patients), and no systematic difference was observed between observed and predicted pregnancy rates (P=0.13 for model without PCT and P=0.6 for model with PCT).

Table 1. Couple characteristics in the development sample (N=2459 for model without PCT and N=1398 for model with PCT) and in the validation sample (N=302).

Variable	Median (upper and lower quartile)		P value
	Development sample	Validation sample	
Woman's age (y)	29 (27-32)	32 (29-35)	<.001
Duration of subfertility (y)	2.5 (1.6-4)	2 (1.5-3)	<.001
Follow-up (months)	11 (5-27)	4 (2-7)	<.001
Total No. of sperm cells (10^{-6}/ml)	42 (16-85)	55 (33-111)	<.001
Motile sperm cells (%)	47 (30-60)	54 (43-62)	<.001
Sperm cells with normal morphology (%)	55 (42-74)	12 (7-17)	<.001
	% (number of patients)		
	Development sample	Validation sample	
Secondary subfertility	22 (530)	37 (112)	<.001
Referral status [a]			
GP	30 (743)	61 (184)	<.001
Gynecologist	68 (1675)	38 (114)	
PCT			<.001
Normal	55 (769)	79 (238)	
Abnormal	45 (624)	17 (52)	
Missing	0 (5)	4 (12)	
Treatment started within 12 months	23 (574)	71 (215)	<.001

[a] 41 and 4 missing values for referral status respectively in the development and validation samples

Table 2 shows that the model with PCT was clinically more useful than the model without PCT since the low and high prediction categories applied to 52% (18% + 34%) of the patients when using the model with PCT versus 36% (25% + 11%) when using the model without PCT. The two models tended to overestimate the probability of live birth in the category of predicted chances above 40% because the estimate is only 36% (**Table 2**). This is consistent with **Figure 1**.

Table 2: Predicted probabilities and observed proportions of 1-year treatment-independent-pregnancy leading to live birth, using the synthesis models with and without PCT.

Predicted probability	N	(%)	Observed	95% CI
	Model without PCT			
< 20%	70	(25%)	7%	2-22%
20-40%	184	(65%)	33%	23-46%
> 40 %	30	(11%)	36%	18-64%
Average prediction: 26.9% [a]	284	(100%)	32%	23-42%
	Model with PCT			
< 20%	52	(18%)	8%	3-22%
20-40%	135	(48%)	34%	20-53%
> 40 %	97	(34%)	36%	24-52%
Average prediction: 33.6% [a]	284	(100%)	32%	23-42%

[a] 12 patients with missing value for PCT, 4 patients with missing value for referral status and 2 patients with missing value for outcome variable.

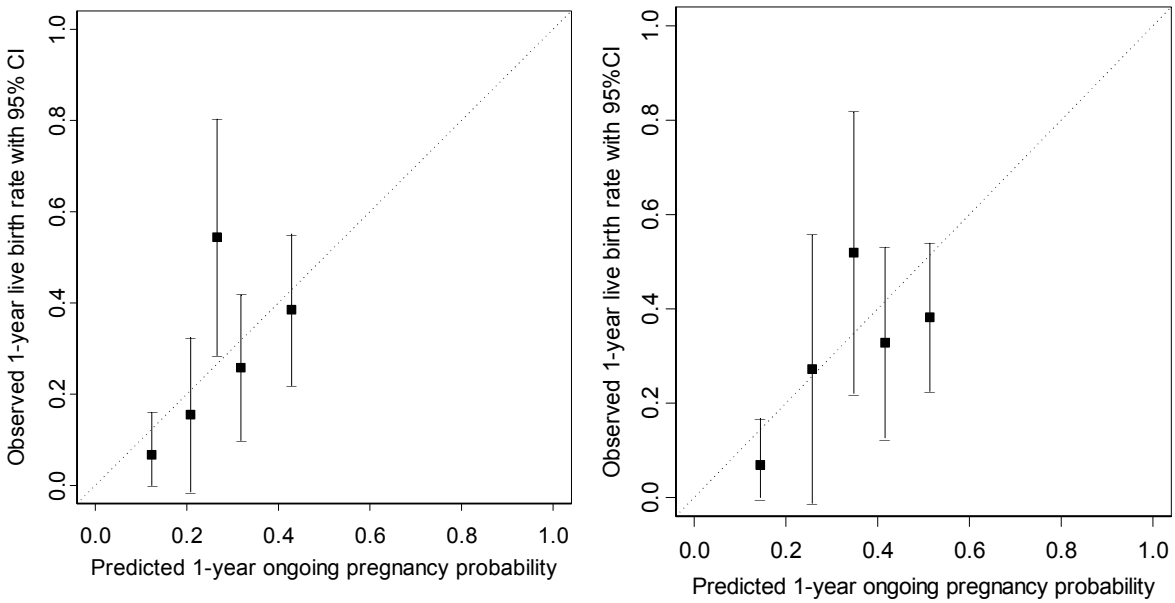

Figure 1 Calibration plots of the models: (**A**) without PCT and (**B**) with PCT. The squares correspond to the groups formed by pooling according to the predicted probabilities. The vertical lines are the 95% confidence intervals of the observed 1-year live birth probabilities, estimated by Kaplan-Meier analyses.

DISCUSSION

We assessed the validity of two models predicting the chance of pregnancy leading to live birth in untreated subfertile couples in a different population than the sample of patients used to develop the models. This study shows that the models were well calibrated i.e. the predicted probabilities did not differ significantly from the observed probabilities. The model including the result of the PCT discriminated better between women who became pregnant and women who did not than the model without PCT (c statistic equal to 0.63 and 0.59 respectively).

The discriminative ability was slightly lower in the validation sample than in the data of the three studies used to develop the models. In the latter, the c statistic varied between 0.59 and 0.64 for the model without PCT and between 0.64 and 0.67 for the model with PCT after internal validation (Hunault, 2004). The lower c statistics observed in the validation sample could be due to the fact that the validation sample is a more homogeneous group with patients having less extreme chances of pregnancy without treatment (predicted chance of treatment independent pregnancy ranging between 5% and 68%, SD=13 in the validation sample compared to predicted chance ranging between 1% and 75%, SD=14 in the development sample).

PCT is an important predictor of treatment independent pregnancy in this sample of patients. This result is interesting since the way in which the PCT is performed in one of the two study centers has changed in the last years. The effect of the result of the PCT in our model has been estimated using data from the Eimers study (Eimers, 1994) and from the Snick study (Snick, 1997). In the Eimers study, the PCT was performed in the fertility laboratory whereas it is currently performed by the clinicians (senior or junior residents). In the Snick study, the PCT was performed by one of the four experienced gynecologists of the peripheral hospital. The prognostic power of the PCT has previously been established for couples with duration of subfertility shorter than 3 years (Glazener, 2000), that is 80% of our validation sample. The

repeated finding that the PCT is an important predictor suggests that the level of experience of the person performing the PCT does not count.

Currently, various effective treatment modalities are available. In our validation sample, treatment was often started early, also for patients who still had a good chance of treatment independent pregnancy, even in the center with a long standing history of use of clinical prediction models (the Utrecht clinic). Among the 27 patients with a predicted probability of 50% or more according to the model with PCT, 52% started a treatment within 6 months after intake (79% in the Utrecht clinic and 21% in the Rotterdam clinic). These 77 couples had a median duration of subfertility of 1.6 years, a median woman's age of 29 years and a median sperm motility of 60%. Eighty five percent of them were referred by a general practitioner and had a secondary subfertility. The PCT was normal in all cases. Because of the high percentage of treatment initiated within the first year, few treatment independent live births were conceived in one year. The statistical power of Cox analysis is related to the number of events (45 treatment independent pregnancies leading to live birth in this study) so the fact that no significant "lack of fit" (calibration) of the model was detected does not mean that calibration was perfect. The calibration of the model should be confirmed in a study with a larger number of couples.

Could the use of the models improve the counselling of couples in comparison with the actual IUI and IVF guidelines of the Dutch Society of Obstetrics and Gynaecology (Dutch acronym: NVOG; www.nvog.nl). According to these guidelines, IUI -and eventually IVF- treatments are offered to patients with unexplained subfertility according to the woman's age and the duration of infertility. We categorized the patients from the validation sample without missing value for the predictors of the models in two groups, patients who should be treated immediately and patients who should have an expectant management, according to the criteria of the Dutch IUI and IVF guidelines, see **Table 3**. Within the group who should be treated immediately, 10% of the patients had a predicted probability of treatment independent live birth above 40% according to the model including PCT. In the group who should have expectant management, 11% of the patients had a predicted probability of treatment independent live birth below 20%. Moreover, about half of the patients fall in the intermediate class, in which patient preferences and counselling are particularly important. These findings suggest that use of the models may be valuable in clinical practice in addition to a guideline like the Dutch one. The patients with a predicted probability of <20% had a median duration of subfertility of 3 years, a median woman's age of 33 years and a median sperm motility of 35%. Fourty eight percent of them were referred by a general practitioner and 19% had a secondary subfertility. The PCT was normal in 25% of the cases. The patients with a predicted probability of >40% had a median duration of subfertility of 1.7 years, a median woman's age of 30 years and a median sperm motility of 54%. Eighty one percent of them were referred by a general practitioner and 62% had a secondary subfertility. The PCT was normal in all cases.

When deciding whether a couple should be offered IUI or IVF treatment or not depends not only of the probability of treatment independent pregnancy. The probability of pregnancy with treatment is also important. If the latter is low as well, starting treatment is without any sense.

If the models are used as a tool in counselling, the model with PCT is more useful than the model without PCT since the poor (below 20%) and good (above 40%) prognosis categories applied to more patients (52% versus 36%). The study has several implications for clinical patient practice. Only six readily available patients characteristics are necessary to use the model with PCT (woman's age, duration of subfertility, type of subfertility (primary or secondary), referral status of the couple, progressive motility from the first semen analysis, and result of the first correctly timed PCT). The models apply to couples with subfertility due to unexplained subfertility, cervical hostility and mild male factor. They have a broad basis of underlying patients populations and provide reliable predictions. Using these models would be useful for

identifying those couples in which the treatment independent chance of live birth is higher than 40%. These couples should be strongly encouraged to restrain from any ART program in the near future. These models might furthermore facilitate a more balanced choice of ART in those couples with lower chances of treatment independent live birth.

APPENDIX

The general formula of a Cox model is:

$$S(t,x) = S_0(t)^{\exp(\beta_1 x_1 + \ldots + \beta_2 x_2)} = S_0(t)^{\exp(PI)}$$

The predicted probability (P) of treatment independent pregnancy within one year after intake leading to live birth according to the synthesis model including the PCT result is:

$P = 100 \times (1 - 0.18^{\exp(PI)})$

Where the prognostic index (PI) = -0.03*AGE1 –0.08*AGE2 – 0.19*duration of subfertility - 0.58*primary subfertility + 0.008*percentage of motile sperm - 0.25*tertiary-care couple

The formula of the synthesis model with PCT is:

$P = 100 \times (1 - 0.17^{\exp(PI)})$

and the prognostic index (PI) = -0.03*AGE1 –0.06*AGE2 – 0.13*duration of subfertility - 0.44*primary subfertility + 0.008*percentage of motile sperm - 0.24*tertiary-care couple –0.95* abnormal PCT

AGE1 is the woman's age if the age is lower or equal to 31 years and 31 years if the age is > 31 years ; AGE2 is the difference (woman's age – 31 years) if the woman's age > 31 years and zero otherwise; a tertiary couple is a couple referred by a gynecologist. Duration of subfertility is measured in years. For primary subfertility, tertiary couple and abnormal PCT, the value is 1 if true, 0 if not true.

The result of the PCT in the initial cycle was coded as abnormal when no forward-moving sperm cell was found in the whole mucus sample.

References

Collins JA, Burrows EA, Willan AR (1995) The prognosis for live birth among untreated infertile couples. *Fertil Steril*. 64, 22-8.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD (1994) The prediction of the chance to conceive in subfertile couples. *Fertil Steril.* 61, 44-52.

Eshre Capri Workgroup Group (2000) Multiple Gestation Pregnancy. Hum Reprod. 15, 1856-64.

Glazener CM, Ford WC, Hull MG (2000) The prognostic power of the post-coital test for natural conception depends on duration of infertility. Hum Reprod. 15, 1953-7.

Hansen M, Kurinczuk JJ, Bower C, Webb S (2002) The risk of major birth defects after intracytoplasmic sperm injection and in vitro fertilization. N Eng J Med. 346, 725-30.

Harrell FE Jr, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med. 15, 361-87.

Hunault CC, Eijkemans MJC, te Velde ER, Collins JA, Evers JLH, Habbema JDF (2004) Two new prediction rules for spontaneous pregnancy leading to live birth among subfertile couples, based on the synthesis of three models. Hum Reprod. 19, 2019-26.

Jones HW (2003) Multiple births: how are we doing? Fertil Steril. 79, 17-21

Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. Ann Intern Med. 130, 515-24.

Miller ME, Langefeld CD, Tierney WM, Hui Sl, McDonald CJ (1993) Validation of probabilistic predictions. Med Decis Making. 13, 49-58.

Moll AC, Imhof SM, Cruysberg JRM, Schouten-van Meeteren AYN, Boers M, van Leeuwen FE (2003) Incidence of retinoblastoma in children born after in-vitro fertilisation. Lancet. 361, 309-10.

Ombelet W, Bosmans E, Janssen M, Cox A, Vlasselaer J, Gyselaers W, Vandeput H, Gielen J, Pollet H, Maes M, Steeno O, Kruger T (1997) Semen parameters in a fertile versus subfertile population: a need for change in the interpretation of semen testing. Hum Reprod. 12, 987-93.

Snick HK, Snick TS, Evers JL, Collins JA (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. Hum Reprod. 12, 1582-8.

Stromberg B, Dahlquist G, Ericson A, Finnstrom O, Koster M, Stjernqvist K (2002) Neurological sequelae in children born after in-vitro fertilisation: a population-based study. Lancet. 359, 461-5.

World Health Organization (1999) WHO laboratory manual for the examination of human semen and sperm-cervical mucus interaction. Fourth edition, Cambridge University Press, Cambridge, UK.

5

A Prediction Model for Selecting Patients for Elective Single Embryo Transfer in IVF

ABSTRACT

Objective
Construction of a prediction model to enable the selection of patients for elective single embryo transfer.

Design
Retrospective cohort study.

Setting
Fertility Center in a Tertiary Referral University Hospital.

Patient(s)
642 women undergoing their first IVF treatment cycle, where no more than two embryos were transferred.

Intervention(s)
Database analysis.

Main Outcome Measure(s)
Ongoing pregnancy and multiple pregnancy.

Result(s)
In multivariate analysis**,** the best predictors for ongoing pregnancy were female age, the number of retrieved oocytes, the developmental stage score and the morphology score of the two best embryos available for transfer, and the day of transfer. Younger age and high quality of transferred embryos were the best predictors for increased risk of multiple pregnancy. The resulting model enables the calculation of probabilities of pregnancy and twin pregnancy. Depending on embryo quality, there is a threshold age under which the chance of singleton pregnancy is higher if one embryo is transferred compared to two embryos.

Conclusion(s)
Application of this model may enable a reduction in the chance of twin pregnancy without compromising singleton pregnancy rates in a subgroup of patients undergoing IVF.

INTRODUCTION

The high incidence of multiple births remains a major concern in relation to IVF treatment. In Europe, the overall rate in 1997 was 30%, with twins at 26% and triplets 4% (EIM, 1997). In the same year, data for the USA revealed that 39% of deliveries were multiple births, with twins accounting for 32% and triplets or higher for 7% of deliveries (SART and ASRM, 2000). These disturbing figures do not include those additional multiple pregnancies that end in spontaneous abortion or are subject to fetal reduction.

Increasing concern over the high incidence of adverse outcomes associated with multiple pregnancy (Rufat,1994) (Fauser, 1999) (Olivennes, 2000) (Jones, 2001) has recently led to the development in the USA of guidelines recommending the number of embryos to be transferred in certain patients (ASRM, 1999). In Western Europe and in particular the Nordic countries transfer of a maximum of two embryos has become the norm (EIM, 1997). However, the chances of twin pregnancy remain high, and twin deliveries are also associated with higher neonatal mortality, handicap and malformation rates (Olivennes, 2000). Moreover, cost studies indicate that much of the financial burden levied by IVF treatment is caused by twin pregnancies (Wølnner-Hansen, 1998) (ESHRE, 2000).

Single embryo transfer for women at particular risk of multiple pregnancy may offer a feasible approach to reducing multiple pregnancy without impacting greatly on overall pregnancy rates (Vilska, 1999). A recent prospective study suggested that acceptable pregnancy rates could be achieved when elective embryo transfer was carried out in women selected on basis of age and embryo quality (Gerris, 1999). The ability to identify those treatment cycles at particular risk of leading to multiple pregnancy and for which single embryo transfer would not reduce the chance of achieving a singleton pregnancy may encourage the adoption of single embryo transfer into clinical practice.

The aims of the present study were to identify clinical and laboratory predictors of pregnancy and twin pregnancy following the transfer of two embryos, and to construct a prediction model that could facilitate the selection of patients for elective single embryo transfer.

MATERIALS AND METHODS

Patient Characteristics and Treatment Protocol

A database containing clinical and laboratory information on all IVF treatment cycles carried out at the University Hospital Rotterdam between December 1993 and December 1998 was analysed. This prospectively designed database is required of all centres practising IVF by Dutch regulatory agencies. The local medical ethics committee and national regulations ensuring anonymity and privacy with respect to patient data were followed.

The ovarian stimulation and embryo culture procedures employed in our IVF program have previously been published (Huisman, 2000). In summary, all patients were down regulated with gonadotropin-releasing hormone (GnRH) agonist and subsequently underwent controlled ovarian hyperstimulation with gonadotropins. When at least one follicle had reached a diameter of 18 mm and more than 3 a diameter >15 mm, a single injection of human chorionic gonadotropin (hCG) 10,000 IU was administered subcutaneously. Transvaginal ultrasound guided follicular aspiration was performed 34-36 h later and luteal support was provided by daily vaginal administration of 600 mg micronized progesterone. Oocytes were inseminated with motile spermatozoa 4-5 h after follicular aspiration.

The resulting best quality embryo and second best quality embryo were selected for embryo transfer. Embryo transfer was carried out on day 3, 4 or 5 after oocyte retrieval, depending on the day of the week on which retrieval had taken place, as described previously (Huisman, 2000). Data from first IVF cycles were subject to analysis. Cycles arising from single embryo transfer, oocyte donation, cryo-thaw embryos cycles and intra-cytoplasmic injection of sperm (ICSI) together with those cycles not resulting in embryo transfer were excluded. In total, 642 consecutive first cycles with 2-embryos transfers were subject to analysis.

Assessment of Embryo Quality

Embryo quality was assessed on the same day as transfer and expressed by two scores: the developmental stage score (Cummins, 1986) and the morphology score (Shulman, 1993).

The developmental stage score included three categories: 'advanced' 'appropriate' and 'retarded' (recorded as 3, 2 and 1 respectively) and was defined differently according to the day of transfer. Day three embryos which had reached the 8-cell stage of embryonic development were considered to be 'advanced', whereas those which had not yet reached the 6-cell stage were considered to be 'retarded'. Day three embryos at the 6 or 7-cell stage were considered to be 'appropriate'. Four days after ovum pick-up, cavitating morulae, morulae, and 4 to 12 cell stage embryos were observed, and respectively classified as advanced, appropriate, and retarded embryos. After 5 days of culture, blastocysts and all earlier stages could be observed. The stages advanced, appropriate and retarded were allocated to expanded blastocysts, blastocysts and morulae or non-cavitating embryos, respectively.

The morphology score was treated as a continuous variable. Day three embryos were assessed for the number of blastomeres, their appearance and degree of fragmentation. Day four embryos did not allow for a number count of blastomeres since the collection of individual cells appeared as a solid mass (morulae) with indistinguishable membranes. Instead, the degree of embryo compaction and the presence of separated cells or fragments were taken into account when assessing quality. For day five embryos, cavitation, expansion and compaction formation were assessed (Gardner, 2000).

For each day, the percentage of fragments in relation to the total volume of the embryo was also assessed (Bolton, 1989). When no fragmentation was evident and the development stage was appropriate for their age, they were described as "high grade" embryos and scored '4'. Embryos showing developmental delay and more than 50% fragmentation were described as "low grade" and scored '1'. Embryos of better and worse intermediate quality were scored as grade "intermediate high" and "intermediate low" respectively.

A pregnancy was defined as a ongoing when fetal heart activity was observed 12 weeks following embryo transfer (ET). Twin-pregnancy was defined as two embryo sacs, each containing an embryo with fetal heart activity at 12 weeks of gestation.

Data analysis

In order to predict individual chances of pregnancy and twin pregnancy, two prediction models were constructed; the 'pregnancy model' (for both singleton and twin) and the 'twin-pregnancy model'. Predictors significantly associated with these outcomes in univariate analyses were analysed by multivariable logistic regression, using a backward stepwise elimination procedure. We chose a cut-off P-value < 0.10 for inclusion of predictors in the final pregnancy and twin-pregnancy models. In addition, age of the woman and day of transfer, both variables shown to importantly influence IVF outcomes in previous studies (Gerris, 1999) (Templeton, 1996) were included in the models, irrespective of their significance.

Although the database was restricted to cases of two-embryo transfer, we were also interested to assess what the outcome would have been in each cycle if just 1 embryo had been transferred. To

do this we were unable to use the logistic models because they assumed that the implantation chances of two embryos transferred into the same uterus are independent. This is not the case however, since the receptivity of the uterus is common to both embryos. Specific statistical methods are required to predict the outcome following single embryo transfer from data relating to two embryo transfer cycles. We used a method developed by Zhou (Zhou, 1998) based on an earlier concept of Speirs (Speirs, 1983), in which survival of the embryos transferred to the uterus depend on their own inherent viability's (E_1 and E_2) and on the receptivity of the uterus (U) that they share.

In such an 'embryo-uterus model', variables specific for the woman (such as her age) can be included as predictors of the uterine receptivity and other variables specific for the transferred embryo (such as embryo quality) or for the woman can be entered as predictors of the embryo viability. We considered several embryo-uterus models consisting of some or all the predictors selected in the final logistic models. The best fitting embryo-uterus model to the pregnancy and multiple pregnancy prediction models was selected. This selected embryo-uterus model was then corrected for over-optimism because estimated coefficients are too extreme for predictive purposes (Van Houwelingen, 1990). We therefore applied a shrinkage factor equal to 0.85 as determined by a bootstrap re-sampling procedure (Efron, 1993). This model could predict both pregnancy following single embryo transfer and pregnancy and twin-pregnancy following two-embryo transfer.

In order to assess the degree of agreement between the predictions produced by the selected embryo-uterus model and by the logistic models for pregnancy and twin-pregnancy, we computed intraclass correlation coefficients. The reliability of the predictions produced by the 3 models was statistically tested by the Hosmer-Lemeshow goodness-of-fit test (Hosmer, 1989). The predictive ability of the three models was assessed by determining the area under the receiver-operating characteristics (ROC) curves.

Graphs were constructed to illustrate how the embryo-uterus model might be applied in practice. Since the representation of all the predictors would make the graphs complex and difficult to read, we elected to construct the graphs using a simplified embryo-uterus model restricted to the most important and objective predictors of pregnancy and twin-pregnancy.

RESULTS

The median age of the women undergoing treatment was 32 years (range 21-43) and the median duration of fertility was 3.9 years (range 0.2-17). 39% had secondary infertility. The indications and outcomes from IVF for the 642 patients are given in **Table 1**.

Table 1: IVF indications and outcomes (642 first IVF cycles).

	%	N
Indication for IVF		
Tubal	29	189
Male factor	28	182
Idiopathic infertility	30	190
Others	13	81
Characteristics of the best embryo		
Developmental stage score		
Retarded	14	91
Appropriate	20	127
Advanced	66	424
Morphology score "high grade"	78	503
Characteristics of the 2nd best embryo		
Developmental stage score		
Retarded	30	193
Appropriate	26	168
Advanced	44	281
Morphology score "high grade"	65	419
Day of transfer		
3	31	201
4	27	170
5	42	271
Ongoing Pregnancies	26	170
Ongoing Multiple-Pregnancies	9	60

In univariate analyses, significant predictors for pregnancy were the woman's age, the total sperm count, the number of pre-ovulatory follicles, the number of retrieved oocytes, the number of embryos suitable for transfer, the developmental stage and the morphology scores of the best two embryos (**Table 2**). The significant predictors of twin-pregnancy in univariate analyses were the woman's age, the number of retrieved oocytes, the number of embryos suitable for transfer and the developmental stage and the morphology scores of the best two embryos. An odds-ratio (OR) > 1 indicates a favourable effect on fecundity and an OR< 1 a negative effect.
Multivariate analysis, allowing correction for other variables, resulted in the predictors shown in **Table 3** being included in the prediction models.

Table 2: Association (univariate analysis) between predictors and chances of pregnancy given by odds ratios (OR) with 95% confidence intervals (CI).

Predictors	Pregnancy[a]		Multiple Pregnancy	
	OR	95% CI	OR	95% CI
Woman's age (per year)	0.95	0.91-0.99	0.93	0.87-0.99
Duration of infertility (per year)	0.96	0.90-1.03	0.96	0.86-1.07
Type of infertility				
Secondary	1.03	0.72-1.48	0.98	0.57-1.69
Indication for IVF				
Tubal	1.00[b]	---	1.00[b]	---
Male factor	0.89	0.56-1.40	1.67	0.84-3.31
Idiopathic infertility	0.83	0.52-1.32	1.23	0.60-2.52
Others	1.08	0.61-1.92	0.60	0.19-1.88
Total No. sperm cells (per 10^{-7}/ml)	1.05	1.01-1.10	1.02	0.95-1.08
Progressive motile sperm cells (per %)	1.01	0.92-1.11	0.98	0.85-1.13
Oestrogen level (per 10^{-3} pmol/l)	1.05	0.99-1.11	1.03	0.93-1.13
No. of pre-ovulatory follicles (per follicle)	1.02	1.00-1.04	1.01	0.98-1.03
No. of retrieved oocytes (per oocyte)	1.05	1.02-1.07	1.04	1.00-1.07
Proportion of oocytes fertilized (per 10%)	1.04	0.96-1.13	1.08	0.96-1.22
Day of embryo transfer				
Day 3	1.31	0.87-1.99	1.26	0.69-2.33
Day 4	1.28	0.83-1.98	0.93	0.47-1.85
Day 5	1.00[b]	---	1.00[b]	---
No. of embryos suitable for transfer (per embryo)	1.09	1.05-1.13	1.07	1.02-1.14
Stage development of the best embryo				
Retarded	0.14	0.06-0.34	0.07	0.01-0.54
Appropriate	0.49	0.31-0.80	0.22	0.08-0.61
Advanced	1.00[b]	---	1.00[b]	---
Stage development of the 2nd best embryo				
Retarded	0.24	0.15-0.39	0.12	0.04-0.33
Appropriate	0.50	0.32-0.77	0.47	0.24-0.89
Advanced	1.00[b]	---	1.00[b]	---
Morphology score of the best embryo (range 1-4)	0.54	0.41-0.71	0.46	0.25-0.82
Morphology score of the 2nd best embryo (range 1-4)	0.67	0.57-0.80	0.58	0.41-0.81

[a] 'Pregnancy' includes both singleton and twin-pregnancy.
[b] reference category.

Pregnancy prediction model

170 of the 642 women undergoing their first IVF cycle achieved either a singleton or multiple ongoing pregnancy (26%). The best predictors for ongoing pregnancy were the number of retrieved oocytes, the development stage of the second best embryo transferred and the morphology score of the best embryo transferred (**Table 3**). The statistical significance of the number of retrieved oocytes after multivariate analysis was p=0.07, allowing the inclusion of this predictor in the pregnancy model since the cut-off p value for inclusion was 0.1. The fact that the 95% confidence intervals in **Table 3** were computed for a level of significance of p=0.05 explains why the number of retrieved oocytes do not appear significant (confidence interval including 1).

A significant interaction was observed between the developmental stage score of the embryo and the day of transfer. This interaction factor was quantified and entered into the model. In contrast, including an interaction factor between the morphology score and the day of transfer gave no significant improvement to the power of the model. The predictive ability of the model measured by the area under the ROC curve was 0.68.

Table 3: Association (multivariate analysis) between predictors and pregnancy and twin pregnancy respectively after correction for other variables given by odds ratios (OR) with 95% confidence intervals (CI).

	OR	95% CI
Pregnancy model[a]		
Woman's age	0.98	0.94-1.02
Number of retrieved oocytes	1.03	0.99-1.05
Morphology score of the best embryo	1.56	1.16-2.11
Stage development score for the 2nd best embryo		
Score = 1 on day 3	1[b]	---
day 4	0.68	0.24-1.89
day 5	0.40	0.15-1.04
Score = 2 on day 3	1.20	0.76-1.90
day 4	1.13	0.48-2.65
day 5	0.75	0.32-1.75
Score = 3 on day 3	1.44	0.57-3.62
day 4	1.86	0.70-4.97
day 5	1.42	0.58-3.43
Twin pregnancy model[c]		
Woman's age	0.95	0.89-1.01
Morphology score of the best embryo	1.56	0.87-2.79
Stage development score for the 2nd best embryo		
Score = 1 on day 3	1[b]	---
day 4	0.58	0.10-3.53
day 5	0.23	0.03-1.58
Score = 2 on day 3	1.44	0.69-3.04
day 4	1.11	0.26-4.76
day 5	0.74	0.16-3.33
score = 3 on day 3	2.08	0.47-9.22
day 4	2.13	0.45-10.2
day 5	2.37	0.57-9.85

[a] Intercept = -2.66
[b] reference category
[c] Intercept = -4.61

Twin-pregnancy prediction model

Fifty-eight (9%) out of the 642 women achieved a twin-pregnancy (34% of all pregnancies) and two women a triplet-pregnancy. A higher probability of achieving a twin-pregnancy was associated with younger age, increasing morphology score of the second best embryo, and increasing development stage score of the best embryo. The number of retrieved oocytes was not a significant predictor of multiple pregnancy. The predictive ability of the model measured by the area under the ROC curve was 0.71.

Embryo-uterus prediction model

We identified five prognostic factors for the chance of survival of each embryo: the woman's age, the number of retrieved oocytes, the developmental stage score and the morphology score of the transferred embryo, and the day of embryo transfer. In contrast, uterine receptivity was not statistically significantly influenced by woman specific covariates. The predictive ability of the embryo-uterus model measured by the area under the ROC curve was 0.67 for pregnancy and 0.72 for twin-pregnancy. The intraclass correlation coefficients between the predictions produced by the logistic models for pregnancy and twin-pregnancy, and the predictions produced by the embryo-uterus model were 0.90 and 0.94 respectively. All three models fitted well to the data as the Hosmer-Lemeshow test was not significant for pregnancy or for twin-pregnancy predictions.

Graphs showing singleton and twin pregnancy chances were constructed from a simplified embryo-uterus model restricted to three predictors: the woman's age, the development stage of the two best embryos available and the day of transfer (**Figure 1**). The embryo developmental stage score was preferred to the morphology score because the former is less liable to subjective interpretation.

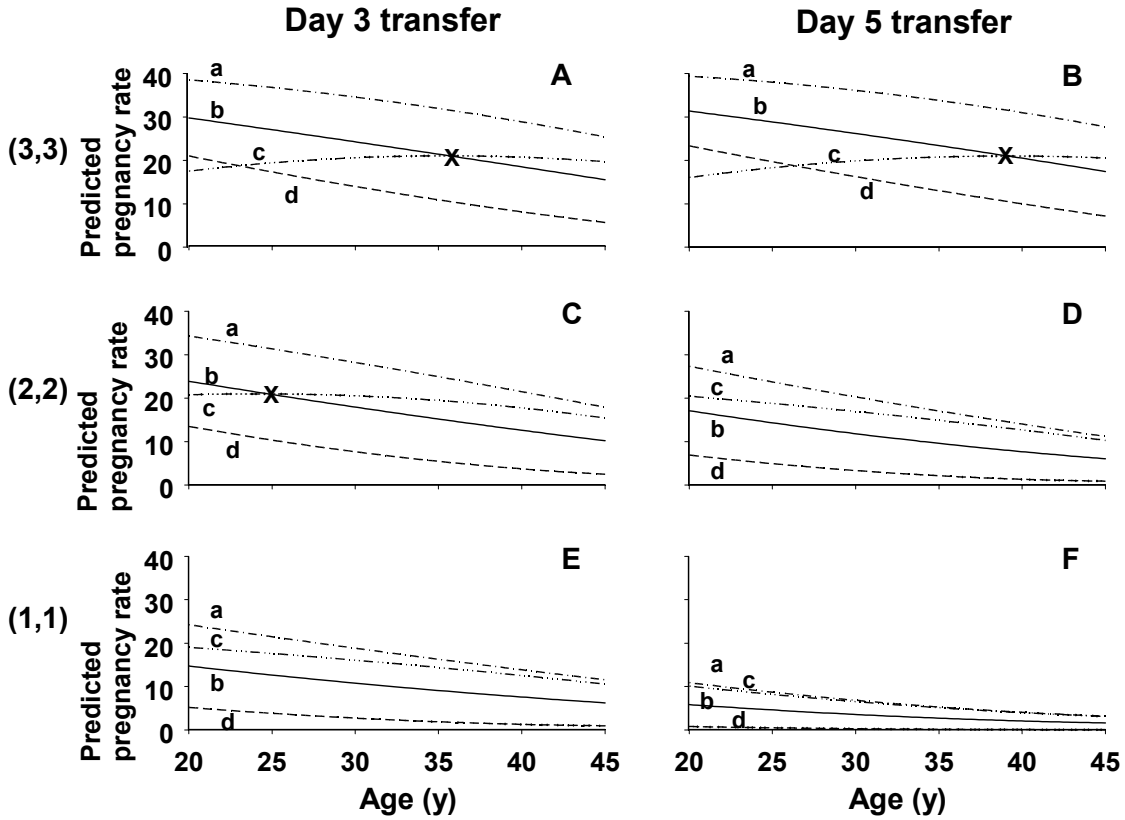

Figure 1: Probabilities of pregnancy and twin-pregnancy in relation to female age and development stage of the best two embryos available for transfer. Graphs are given for embryo transfer on the 3rd and 5th day after oocyte pick-up, and for three combinations of embryos with identical developmental stage scores. Development stage is scored such that a score of 3 denotes top quality and a score of 1 denotes poorest quality still suitable for transfer. (3,3) means the combination of two embryos both scored 'advanced' whereas (2,2) and (1,1) are the combinations of two embryos both scored 'appropriate' and 'retarded' respectively. Examples of how the graphs may be used to estimate relative chances of singleton and twin pregnancy are given in the text.

a = Probability of pregnancy if 2 embryos are transferred [- · - · - · - · - · - · -]
b = Probability of pregnancy if 1 embryo is transferred [———————]
c = Probability of singleton pregnancy if 2 embryos are transferred [— · · — · · — · · — · · —]
d = Probability of twin pregnancy if 2 embryos are transferred [- - - - - - - - - - - - - -]

X = patient age below which Single Embryo Transfer will result in a higher chance of singleton pregnancy than transfer of 2 embryos

Graph A shows that a 30 year old woman with two top quality embryos transferred (score 3,3), has a 35% chance of pregnancy (line a), of which 21% chance of singleton and 14% chance of twin-pregnancy (lines c and d). The chance of pregnancy in case of single embryo transfer (SET) is 24% (line b). The benefit of single embryo transfer is thus avoiding the 14% chance of a twin-pregnancy (line d). However, this benefit is balanced against the drawback of single embryo transfer, namely the difference between the chance of pregnancy after two-embryo transfer as compared with single embro transfer. This difference is represented as that between lines a and b, i.e. 35-24=11%. If we consider IVF success only in terms of achieving a singleton pregnancy, single embryo transfer which in this case offers a 24% chance of singleton pregnancy, would be preferable to two-embryo transfer where the chance of singleton pregnancy is only 21%.

When embryo quality is poor however, the transfer of two embryos may increase the chance of singleton pregnancy as compared with single embryo transfer. In graph E, a 30 year old woman whose 2 best embryos are of poor quality, i.e. scored (1,1), is seen to have a chance of singleton pregnancy after two-embryo transfer of 17% (line c) compared with just 12% should SET be carried out (line b). It can be seen from the graphs in **Figure 1** that for given embryo scores and day of transfer, there is a threshold age 'X' where lines b and c cross, below which SET will result in a higher chance of singleton pregnancy than transfer of two embryos. When high quality embryos are available for transfer to a woman whose age is below the threshold, the transfer of one embryo may be preferable to two, as twin pregnancy will be avoided without reducing the chance of singleton pregnancy.

DISCUSSION

By analysing data from IVF cycles where two embryos were transferred, we have constructed models that predict the chances of singleton and twin pregnancy when one or two embryos are transferred. We employed multiple logistic regression using a backward stepwise selection of the variables as a strategy to identify the clinical and laboratory predictors of pregnancy following IVF. The most important predictors identified were the development stage and the morphology score of the two best embryos available for transfer, and the age of the patient. Where just one embryo was available for transfer, this was considered to indicate advanced ovarian ageing and an *a priori* poor prognosis for IVF outcome (Templeton, 1998). Since single embryo transfer was the only option in these cases, data related to these cycles were excluded from the present analysis.

Certain predictors found to be important in previous studies did not appear in our final pregnancy and twin-pregnancy models. For instance, after adjusting for the number of retrieved oocytes, the number of embryos available for transfer was no longer a significant predictor (Templeton, 1998) (Coetsier, 1998). Although the type of infertility (secondary or primary) was highly correlated with the duration of infertility and with the woman's age, neither the type nor the duration of infertility appeared in the final models. This is consistent with the results of previous large studies of the factors affecting outcome from IVF (Templeton, 1996). Other predictors that appear in the medical literature, such as the endometrial thickness (Check, 1991) (Smeenk, 2000), caffeine or smoking consumption and psychological factors (Smeenk, 2000), were not recorded in the database used.

A number of studies have provided models to compute the probabilities of singleton and multiple IVF pregnancies. In a previous retrospective study, predicted singleton and twin-

pregnancies rates were lower compared to the present study (Wheeler, 1998). This may be explained by laboratory differences, the earlier inclusion of patients and by the fact that all cycles in the previous study were considered. Moreover, the quality of the embryos in the previous study was expressed by the mean of the score of the embryos transferred.

As in our study, Martin et al. (Martin, 1998) used the concept of Spiers and focused on first cycles. They provided a simple model assuming a constant implantation rate of the embryos, which is not the case with embryos of different quality. In a more recent study, both ICSI cycles and second and third IVF cycles were included in the analysis (Strandell, 2000). The resulting model contained, in contrast to our prediction model, tubal infertility and the number of good quality embryos as variables, but not embryo quality. Moreover, the dependency between the chances of implantation of both embryos was not taken into account in the construction of the model (Strandell, 2000).

The importance of addressing this aspect is illustrated by considering observed and expected twin pregnancy rates. In our study, the incidence of twin-pregnancy was 9.3 %. If the chance of implantation of each embryo was independent from the other, then the chance of implantation per embryo would be 30%. With a 30% chance of implantation per embryo, the proportion of pregnancies resulting in twins would be expected to be 18% as opposed to the 35% observed.

The graphs shown in **Figure 1** illustrate the importance of the quality of the best embryo in predicting outcome and the positive impact of extended culture (5 versus 3 days) on the ability to select high quality embryos (**Figure 1**, graphs E and F). The prediction model (given in the appendix) can also be used to construct graphs for day 4 embryo transfers and for transfer of other combinations of embryo developmental stage scores (not shown). The graphs also illustrate how the model may aid selection of patients for single embryo transfer when the aim is to maximize singleton pregnancy rates.

However, singleton pregnancy may not always be perceived as the only desired outcome from IVF. **Table 4** illustrates how the patient age and embryo quality under which single embryo transfer might be considered depends on the extent to which twin pregnancy is perceived by the physician and his patients as a ‘positive’ or ‘negative’ outcome. If twin pregnancy is considered to be less desirable than singleton pregnancy but more desirable than no pregnancy, then the threshold ages for single embryo transfer decrease. In contrast, these threshold ages increase when twin pregnancy is deemed as less desirable than no pregnancy (‘negative’ column). If the embryo-uterus model had been applied in our centre, 4 twin-pregnancies instead of 60, and 130 singleton pregnancies instead of 110 would have been obtained, valuing twin-pregnancy as a negative outcome. Were twin-pregnancy deemed as as desirable as no pregnancy or as a more positive outcome, 8 and 48 twin-pregnancies, and 131 and 120 singleton pregnancies would have been obtained respectively.

In conclusion, we have developed a prediction model, which indicates the relative chances of singleton and multiple pregnancy following single and two-embryo transfer. The model can be used to identify those women and treatment cycles where single embryo transfer will improve the singleton pregnancy rate at the expense of twin pregnancies. The mounting evidence for the high costs of twin pregnancies in terms of morbidity, mortality and financial burden supports our contention that singleton prgnancy should be the aim of IVF treatment. If infertile couples are properly informed of the adverse aspects of twin pregnancy they may be happy to undergo single embryo transfer. The presented model may aid in this decision process since for certain couples, single embryo transfer may prevent twin pregnancies without reducing the chance of a singleton pregnancy. Prospective studies are now required to confirm the clinical validity of the model.

Table 4: The effect of attributing different subjective values to twin-pregnancy on a decision as whether 1 or 2 embryos be transferred is shown. Threshold ages below which single embryo transfer should be considered are given for three different perceptions of a twin-pregnancy: twin-pregnancy valued as no pregnancy, as a positive outcome or as a negative outcome.

Day of transfer	Scores of the best & the second best embryos[a]	Value for twin-pregnancy		
		Positive[b]	0	Negative[c]
3	(3,3)	25 years	36 years	42 years
	(3,1)	25 years	36 years	42 years
	(2,2)	always 2[d]	25 years	31 years
	(1,1)	always 2[d]	always 2 [d]	20 years
4	(3,3)	27.5 years	38 years	44 years
	(3,1)	27.5 years	38 years	44 years
	(2,2)	always 2[d]	24 years	30 years
	(1,1)	always 2[d]	always 2[d]	always 2[d]
5	(3,3)	28 years	39 years	45 years
	(3,1)	28 years	39 years	45 years
	(2,2)	always 2[d]	always 2[d]	always 2[d]
	(1,1)	always 2[d]	always 2[d]	always 2[d]

[a] Embryo developmental stage is scored such that a score of 3 denotes top quality and a score of 1 denotes poorest quality still suitable for transfer. (3,3) is the combination of two 'advanced' embryos, whereas (3,1) is the combination of a best embryo scored 'advanced' with a second best embryo scored 'retarded'.
[b] twin-pregnancy valued as +0.5, compared to no pregnancy = 0 and singleton pregnancy = 1
[c] twin-pregnancy valued as -0.5, compared to no pregnancy = 0 and singleton pregnancy = 1
[d] The transfer of 2 embryos is always preferable

APPENDIX

The embryo-uterus model assumes that the survival of an embryo depends on its own inherent viability and the receptivity of the uterus that is shares with the other transferred embryo. For both uterus receptivity (U) and embryo viability (E), a logistic submodel is fitted:

$$U = \frac{1}{1+\exp(-0.26)} = 0.56$$

$$E_i = \frac{1}{1+\exp(-0.6-0.05*\text{age}+0.02*\text{ooc}+0.26*\text{DSS}_i+0.31*\text{MS}_i+(-0.63+0.39*\text{DSS}_i)*\text{day4}+(-1.67+0.9*\text{DSS}_i)*\text{day5})}$$

Where $_i$ is the embryo number (1 or 2), ooc is the number of retrieved oocytes, DSS is the developmental stage score, and MS is the morphology score.
E_1 is the probability that the best embryo is viable, and E_2 that the second best embryo is viable.
Probability of pregnancy when 2 embryos are transferred: $U*[E_1*(1-E_2) + E_2*(1-E_1) + E_1E_2]$
Probability of singleton pregnancy when 2 embryos are transferred: $U*[E_1*(1-E_2) + E_2*(1-E_1)]$
Probability of pregnancy when 1 (the best one) embryo is transferred: $U * E_1$
Probability of twins when 2 embryos are transferred: $U * E_1 * E_2$

References

American Society for Reproductive Medicine (1999). *Guidelines on number of embryo's transferred.* Birmingham, AL: American Society for Reproductive Medicine.

Bolton VN, Hawes SM, Taylor CT, Parsons JH (1989). Development of spare human preimplantation embryos in vitro: an analysis of the correlations among gross morphology, cleavage rates, and development to the blastocyst. *J In Vitro Fert Embryo Transf.* 6, 30-5.

Check JH, Nowroozi K, Choe J, Dietterich C (1991) Influence of endometrial thickness and echo patterns on pregnancy rates during in vitro fertilization. *Fertil Steril.* 56, 1173-5.

Coetsier T, Dhont M. (1998) Avoiding multiple pregnancies in in-vitro fertilization: who's afraid of single embryo transfer? *Hum Reprod.* 13, 2663-4.

Cummins JM, Breen TM, Harrison KL, Shaw JM, Wilson LM, Hennessey JF (1986) A formula for scoring human embryo growth rates in in vitro fertilization: its value in predicting pregnancy and in comparison with visual estimates of embryo quality. *J In Vitro Fert Embryo Transf.* 3, 284-95.

Efron B, Tibshirani RJ (1993) *An introduction to the bootstrap.* Chapman & Hall Inc ; London.

Eshre Capri Workgroup Group (2000). Multiple Gestation Pregnancy. *Hum Reprod.* 15, 1856-64.

European IVF- Monitoring Programme (EIM) for ESHRE (2001) Assisted reproductive technology in Europe, 1997. Results generated from European registers by ESHRE. *Hum Reprod.* 16, 384-91.

Fauser BC, Devroey P, Yen SS, Goesden WF Jr, Baird DT et al. (1999) Minimal ovarian stimulation for IVF: appraisal of potential benefits and drawbacks. *Hum Reprod.* 14, 2681-6.

Gardner DK, Lane M, Stevens J, Schlenker T, Schoolcraft WB (2000) Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer. *Fertil Steril.* 73, 1155-8.

Gerris J, De Neubourg D, Mangelschots K, Van Royen E, Van de Meerssche M, Valkenburg M (1999) Prevention of twin pregnancy after in-vitro fertilization or intracytoplasmic sperm injection based on strict embryo criteria: a prospective randomized clinical trial. *Hum Reprod.* 14, 2581-7.

Gonen Y, Casper RF, Jacobson W, Blankier J (1989) Endometrial thickness and growth during ovarian stimulation: a possible predictor of implantation in in vitro fertilization. *Fertil Steril.* 52, 446-50.

Hosmer DW, Lemeshow S (1989) *Assessing the fit of the model. Applied logistic regression*. John Wiley & Sons Inc, New York, NY.

Huisman GJ, Fauser BC, Eijkemans MJ, Pieters MH (2000) Implantation rates after in vitro fertilization and transfer of a maximum of two embryos that have undergone three to five days of culture. *Fertil Steril.* 73, 117-22.

Jones HW, Schnorr JA (2001) Multiple Pregnancies: a call for action. *Fertil Steril.* 75, 11-3.

Martin PM, Welch HG (1998) Probabilities for singleton and multiple pregnancies after in vitro fertilization. *Fertil Steril.* 70, 478-81.

Olivennes F (2000) Avoiding multiple pregnancies in ART. Double trouble: yes a twin pregnancy is an adverse outcome. *Hum Reprod.* 15, 1663-5.

Rufat P, Olivennes F, de Mouzon J, Dehan M, Frydman R (1994) Task force report on the outcome of pregnancies and children conceived by in-vitro fertilisation (France:1987-89). *Fertil Steril.* 61, 324-30.

Shulman A, Ben-Nun I, Ghetler Y, Kaneti H, Shilon M, Beyth Y (1993) Relationship between embryo morphology and implantation rate after in vitro fertilization treatment in conception cycles. *Fertil Steril.* 60, 123-6.

Smeenk JM, Stolwijk AM, Kremer JA, Braat DD(2000) External validation of the Templeton model for predicting success after IVF. *Hum Reprod.* 15, 1065-8.

Society for Assisted Reproductive Technology (SART) and American Society for Reproductive Medicine (ASRM) (2000) Assisted reproductive technology in the United States: 1997 results generated from the ASRM/SART registry. *Fertil Steril.* 74, 4-11

Speirs AL, Lopata A, Gronow MJ, Kellow GN, Johnston WI (1983) Analysis of the benefits and risks of multiple embryo transfer. *Fertil Steril.* 39, 468-71.

Strandell A, Bergh C, Lundin K (2000) Selection of patients suitable for one-embryo transfer may reduce the rate of multiple births by half without impairment of overall birth rates. *Hum Reprod.* 15, 2520-5.

Templeton A, Morris JK, Parslow W (1996) Factors that affect outcome of in-vitro fertilisation treatment. *Lancet* 348, 1402-6.

Templeton A, Morris JK (1998) Reducing the risk of multiple births by transfer of two embryos after in vitro fertilization. *N Engl J Med.* 339, 573-7.

Van Houwelingen JC, Le Cessie S (1990) Predictive value of statistical models. *Stat Med.* 9, 1303-25.

Vilska S, Tiitinen A, Hyden-Granskog C, Hovatta O (1999) Elective transfer of one embryo results in an acceptable pregnancy rate and eliminates the risk of multiple birth. *Hum Reprod.* 14, 2392-5.

Wheeler CA, Cole BF, Frishman GN, Seifer DB, Lovegreen SB, Hackett RJ (1998) Predicting probabilities of pregnancy and multiple gestation from in vitro fertilization: a new model. *Obstet Gynecol.* 91, 696-700.

Wølner-Hanssen P, Rydhstroem H (1998) Cost-effectiveness analysis of in-vitro fertilization: estimated costs per successful pregnancy after transfer of one or two embryos. *Hum Reprod.* 13, 88-94.

Zhou H, Weinberg CR (1998) Evaluating effects of exposures on embryo viability and uterine receptivity in in vitro fertilization. *Stat Med.* 17, 1601-12.

6

A case study of the applicability of a prediction model for the selection of in vitro fertilization patients for single embryo transfer in another center

ABSTRACT

Study objective: To evaluate the application in a different fertility clinic of a prediction model for selecting IVF patients for elective single embryo transfer.

Design: Retrospective analysis of a large database obtained from a tertiary Infertility Center.

Setting: University Medical Center

Patients: The model, derived at the 'development center' was applied in 494 consecutive first IVF cycles carried out at the 'application center'.

Interventions: Following adjustment of embryo scoring system to be compatible with that employed by the prediction model, it was applied to the development center data. A score chart for predicting the probability of singleton or twin pregnancy was constructed.

Main outcome measure(s) The area under the receiver-operator curve (ROC) was determined to measure the discriminative ability of the model for both ongoing pregnancy and twin pregnancy. Calibration plots were made to assess agreement between predicted and observed pregnancy rates.

Results: The areas under the ROC for predicting ongoing pregnancy and twin pregnancy were 0.66 and 0.70 respectively. Insertion of a correction factor equivalent to the difference in odds ratios for ongoing pregnancy rates between the two centers was required to improve the calibration of the model.

Conclusions. After adaptation, the model performed well in the application center.

Key words
External validation / IVF / prediction model / elective single embryo transfer

INTRODUCTION

In order to be able to decide on the appropriate management of the patient, prognosis plays a central role in clinical medicine. Likewise, in reproductive medicine it is of paramount importance to have knowledge on the probability either of ongoing pregnancy occurring spontaneously or after therapy. Prognostic assessment in reproductive medicine should be based on the multivariate analysis of prospectively collected data in a cohort of subfertile couples (Laupacis, 1994). The results can be applied to future patients either in the same or in other centers. However, the application of such so-called prognostic models in clinical medicine, including reproductive medicine, is complicated and full of potential pitfalls (Stolwijk,1996) (Justice, 1999) (Altman, 2000). The paramount question to be answered before generally applying a prognostic model is: how well does the model predict, and is it applicable in other centers?

The answer to these questions should be based on internal and external validation of the model. First, the accuracy of the model should be internally validated on the patient group from which it was derived. However, internal validation systematically gives a too optimistic impression about the quality of the predictions (Harrel, 1996). External validation by testing the model in other patients and/or in other centers is the gold standard for measuring the quality of prognostic models. Unfortunately, the validation of most prognostic models in reproductive medicine has not gone beyond the stage of internal validation and so the question whether or not they are more generally applicable, has not been answered.

Recently, we developed a prediction model for selecting patients undergoing IVF for elective single embryo transfer (Hunault, 2002). The high incidence of twin pregnancies after IVF poses a major problem: many more complications occur in twin babies as compared to singletons (Olivennes, 2000) (Hazekamp, 2000) (Ozturk, 2001). This problem could be solved by single embryo transfer. However, the chance of pregnancy – singleton or twin – decreases when transferring one instead of two embryos. Therefore, the aim of the prediction model was to select patients who have a high probability of a twin pregnancy after transfer of two embryos, and a reasonable chance of conceiving a singleton after the transfer of one embryo.

In that study, two embryos were transferred in all women. The age of the woman and some quality measures of the transferred embryos appeared to be important predictors both of pregnancy and of twin pregnancy (Hunault, 2002). The most important conclusion of the study was that below a certain age, depending on the quality of the embryos transferred, pregnancy rates are hardly compromised when transferring only one embryo. For example in most women under 30 single embryo transfer is completely justified and strongly recommended. In contrast, in women of 38 or older transfer of 2 embryos or even more, is still preferable. The model appeared to perform reasonably well after internal validation but has not yet been externally validated.

The aim of this study was to evaluate the feasibility of applying this model in another fertility clinic: the problems encountered, the solutions found and the quality of its performance in the application center. We also developed an easy-to-use score chart for daily practice that may assist in the decision to transfer one or two embryos. Although this model relates to a specific clinical problem, it may be of general relevance to clinicians who wish to use prognostic models derived in clinics elsewhere for their own patients.

MATERIAL AND METHODS

The prediction model

The prediction model (Hunault, 2002) was developed at the Erasmus Medical Center in Rotterdam (the 'development center') and was based on data collected during the first IVF cycle of 642 consecutive patients in whom 2 embryos were transferred. It consists of 'Embryo viability' and an 'Uterine receptivity' components (EU model) (Zhou, 1998); and predicts the chance of ongoing pregnancy (both singles and twins) and ongoing twin pregnancy after transfer of one or two embryos.

The model includes five predictors. Three of them - number of retrieved oocytes, developmental stage ('advanced', 'appropriate' and 'retarded' according to the number of cells) and morphology grade (extent of fragmentation and equality of blastomeres) of the best and second best embryo - were selected after multivariate regression analysis with a stepwise selection procedure (for details see Hunault, 2002). Two other predictors, woman's age and day of transfer, were added on clinical grounds. A pregnancy was defined as ongoing when at least one sac with fetal heart activity was observed 12 weeks following ET and as an ongoing twin pregnancy when two embryo sacs with fetal heart activity, were present at 12 weeks of gestation.

The patients of the application center

This model was applied in patients of the University Medical Center Utrecht (the 'application center'). As in the development center, first cycle data from patients in whom two embryos were transferred were included, and patients were excluded when intracytoplasmic sperm injection (ICSI) or oocyte donation had been performed or when cryopreserved embryos were transferred. Patients underwent their IVF treatment between October 1997 and January 2000. Local medical ethics committee and national regulations ensuring anonymity and privacy with respect to patient data, were followed.

Matching the embryo quality score systems of each center

All predictors included in the model were available in the application dataset. Definitions of the predictors were the same for both centers except for the scoring system for embryo quality. Although in both centers embryos were classified according to their developmental stage (number of cells) and morphology (extent of fragmentation and equality of blastomeres), the results had been documented categorically in the application center, and both categorically and continuously in the development center (for details see Hunault, 2002).

Further, the developmental stage classification in the original model consisted of three categories: 'advanced', 'appropriate' and 'retarded'. This division was meaningful because it discriminated well for embryo transfers on day 5, the most common day for embryo transfer in the development center. In contrast, for day 3 and day 4 transfers, the distinction between 'appropriate' and 'retarded' was not useful, since the 'retarded' category hardly occurred on those days. In the application center, embryos were only transferred on day 3 and 4. Therefore, the developmental stage was reclassified into two categories: excellent (equivalent to the category advanced in the original model) and retarded (equivalent to the categories appropriate and retarded in the original model). The modified model using this revised category system was applied in both the development and application centers.

The score chart for calculating pregnancy chances for the modified model is given in graphical form. The treatment protocols of the development center and the application center were otherwise similar (Huisman, 2000) (Van Kooij, 1996).

Methods of analysis

Differences in the characteristics of couples between both samples were tested by Kruskal-Wallis test for the continuous variables and chi-square tests for the categorical variables. The performance of the prediction model was assessed for its discrimination (i.e. the model's ability to distinguish between women who achieved and who did not achieve ongoing pregnancy) and its calibration (i.e. the agreement between predicted and observed probabilities) (Harrell, 1996). Discrimination was assessed by the area under the ROC curve (AUC), which is equivalent to the c-statistic (Harrell, 1996) and lack of calibration was tested by a chi-squared test.

When a prediction model is applied to a sample different from that used to develop the model, the predictions in the new sample of patients are often too extreme: relatively low predictions are too low, while relatively high predictions are too high. To avoid this phenomenon (called "over optimism"), a bootstrapping procedure was used to estimate the degree of over optimism; from this, corrected regression coefficients and c statistics were calculated (Harrell, 1996) (Van Houwelingen, 1990), and used when applying the model in the application sample. The bootstrapping procedure is a method to assess internal validation and the amount of overfitting of the modified model. The amount of overfitting is expressed by the amount of shrinkage by which the regression coefficients should be reduced in absolute size. The standard error of the coefficients is not affected, therefore, the confidence intervals will not change in width, but their location is shifted in the direction of no effect.

Calibration plots were constructed to show the degree of agreement between predicted and observed ongoing pregnancy probabilities.
Calculations were performed using commercially available software packages (SPSS Inc., Chicago, IL, USA, 1999 and S-plus 2000, MathSoft Inc., Seattle, WA, USA, version 2000). 95% confidence intervals were calculated for all important estimates.

RESULTS

The application sample consisted of data of 494 consecutive first IVF cycles in which two embryos were transferred.

In **Table 1**, the relevant patients characteristics, the predictors used and both outcomes (ongoing pregnancy and ongoing twin pregnancy) in the development and application centers were compared. The patients in the application center were on average one year older and more often had primary infertility. The number of retrieved oocytes was higher in the development sample but the percentage of oocytes fertilized was higher in the application sample. The day of ET considerably differed between both centers. While in the application center all transfers occurred either on day 3 or 4, in the development center only about 60% of the transfers were performed on these days.

Furthermore, there was a difference in the outcome categories between both centers in favor of the application center (39% versus 26% for ongoing pregnancy) and 14% versus 9% for ongoing twin pregnancy). The same trend was present in the patients who only had transfers on day 3 or day 4 (39% versus 29 % for ongoing pregnancy, and 14% versus 9% for ongoing twin pregnancy). Twin pregnancies represented about one third of all ongoing pregnancies both in the development sample and in the application sample.

Table 1: Characteristics and outcomes of the patients in the development sample (n=642) and application sample (n=494).

	Development		Application		
	Median or %	(range or n)	Median or %	(range or n)	P value
Woman's age (y)	32	(21-43)	33	(21-41)	0.03 [b]
Duration of infertility (y)	3.9	(0.2-17)	3.8	(0.1-14) [a]	0.06 [b]
Secondary infertility	39%	(250)	31%	(153)	0.005 [c]
Indication for IVF					0.11 [c]
Tubal	29%	(189)	27%	(131)	
Male factor	28%	(182)	25%	(121)	
Idiopathic factor	30%	(190)	36%	(178)	
Others	13%	(81)	13%	(64)	
Day of transfer					<0.001 [c]
Day 3	31%	(201)	74%	(366)	
Day 4	27%	(170)	26%	(128)	
Day 5	42%	(271)	0%	(0)	
No. of retrieved oocytes	10	(2-40)	9	(2-35)	0.01 [b]
No. of embryos suitable for transfer	6	(2-24)	5	(2-29)	0.3 [b]
Percentage of oocytes fertilized	61	(0-100)	67	(0-100)	0.001 [b]
Outcome of all patients					
Ongoing pregnancy (per ET)	26%	(170)	39%	(185) [a]	<0.001 [c]
Ongoing twin Pregnancy (per ET)	9%	(60)	14%	(64) [a]	0.02 [c]
Outcome of patients with day 3 and 4 ET only					
Ongoing pregnancy	29%	(106)	39%	(185) [a]	<0.001
Ongoing twin Pregnancy	9%	(36)	14%	(64) [a]	0.03 [c]

[a] Missing values occurred in the validation sample for duration (8%) and outcome (4%).
[b] Kruskal Wallis test.
[c] Chi square test.

In **Table 2** the embryo characteristics on day 3 and 4 are compared. The quality of most embryos was high and the occurrence of a morphology score below 2 was relatively rare. The difference in ongoing pregnancy rates between both centers cannot be explained by differences in embryo characteristics. The embryos seemed to be of the same quality with regard to the developmental stage, whereas the morphology grade was significantly better in the development center.

Table 2: Embryo characteristics on day 3 and 4 transfers comparing the development (n=371) and the application (n=494) centers. The two categories of the stage development score are defined according to the definition used in the application center.

			Development		Application		P value [a]
			%	(n)	%	(n)	
ET day 3	Morphology grade [b]	0	5%	(11)	9%	(32)	0.04
	of the best embryo	1	12%	(25)	19%	(68)	
		2	82%	(165)	73%	(266)	
	Stage development	retarded	19%	(39)	24%	(88)	0.2
	of the best embryo	excellent	81%	(162)	76%	(278)	
	Morphology grade [b]	0	7%	(15)	4%	(14)	<0.001
	of the 2nd best embryo	1	15%	(31)	37%	(137)	
		2	77%	(155)	59%	(215)	
	Stage development	retarded	43%	(86)	45%	(164)	0.6
	of the 2nd best embryo	excellent	57%	(115)	55%	(202)	
ET day 4	Morphology grade [b]	0	3%	(5)	23%	(30)	<0.001
	of the best embryo	1	11%	(19)	39%	(50)	
		2	86%	(146)	38%	(48)	
	Stage development	retarded	22%	(37)	23%	(29)	0.8
	of the best embryo	excellent	78%	(133)	77%	(99)	
	Morphology grade [b]	0	8%	(13)	34%	(44)	<0.001
	of the 2nd best embryo	1	14%	(23)	45%	(57)	
		2	79%	(134)	21%	(27)	
	Stage development	retarded	41%	(70)	45%	(58)	0.5
	of the 2nd best embryo	excellent	59%	(100)	55%	(70)	

[a] Chi square test
[b] Highest score is best

As 2 embryos were always transferred several combinations were possible. For the developmental stage score, the combination of 2 excellent embryos most frequently occurred (~ 55%). A similar pattern was present for combinations of the morphology score. Moreover, the correlation between stage development score and the morphology grade, both for the best and the second best embryos, was substantial (r = 0.45 and 0.47 respectively).

After correction for over-optimism, the performance of the original and the modified model in the development sample is almost the same (AUC for ongoing pregnancy 0.65 and 0.66 respectively and for ongoing twin pregnancy 0.69 and 0.70 respectively). Therefore, the modified model can be applied to the data of the application center as if it was the original model. With regard to the discriminative ability of the model – the first quality measure -, the AUC in the application center was 0.63 for ongoing pregnancy (a decrease of 0.03) and 0.66 for ongoing twin pregnancy (a decrease of 0.04). With regard to the second quality measure – agreement between predicted and observed pregnancy rates - **Figure 1** shows the calibration plots of the model for ongoing

pregnancy (A) and for ongoing twin pregnancy (B). The calibration curves were above the perfect calibration diagonal over the whole range of probabilities indicating that the predictions were too low, which is in agreement with the better ongoing pregnancy rates in the application center shown in **Table 1**. This systematic difference between predicted and observed probabilities was significant for both ongoing and ongoing twin pregnancy ($p<0.001$).

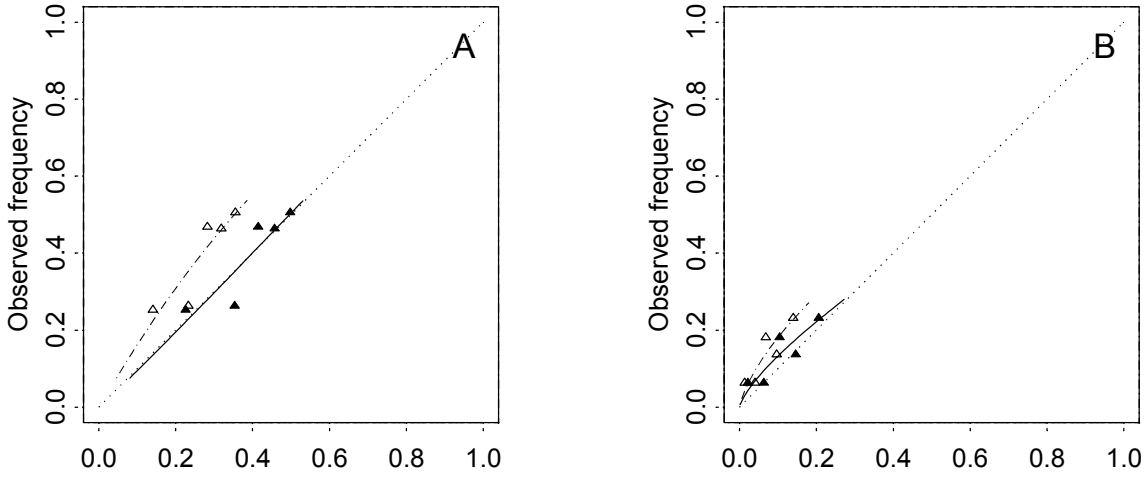

Figure 1: Calibration plots of the modified model applied to the application sample (n=494 couples) (A) for ongoing pregnancy (n=185) and (B) for the twin pregnancy (n=64) ⋯: diagonal indicating perfect calibration; --: calibration curve without correction; —: calibration curve with correction; Δ: patients grouped by quintiles of predicted probability (without correction); ▲: patients grouped by quintiles of predicted probability (with correction).

Theoretically, this difference should be corrected by adapting the baselines of the embryo viability and of the uterine receptivity in the model. However, since both are unobservable in clinical practice, the only way to correct the model is to apply a correction factor equal to the odds ratio between the 2 centers for the ongoing pregnancy rates (odds ratio 1.8) and the ongoing twin pregnancy rates (odds ratio 1.6) over the period of data collection. With this simple correction, the model was well calibrated for ongoing pregnancy ($p=0.45$). For ongoing twin pregnancy, the correction considerably improved the calibration ($p<0.001$) but could not completely overcome the systematic underestimation of the model ($p=0.02$). In **Figure 1**, the improvement by the corrections is demonstrated. The curves are almost the same (ongoing pregnancy) or much closer (ongoing twin pregnancy) in comparison to the ideal diagonal line.

In **Figure 2** the relative weights of 4 of the 5 variables are expressed as scores: the higher the total score the higher the chances of ongoing (twin) pregnancy. The impact of the day of transfer becomes apparent in the scores for developmental stage: a retarded embryo is 'penalized' more strongly on day 4 than on day 3. According to the guidelines given in the legends of **Figure 2**, the final sum scores of both embryos can be translated into the probability of ongoing pregnancy (graph A), ongoing twin pregnancy (graph B), ongoing singleton pregnancy (graph C) and ongoing singleton pregnancy if only the best embryo would be transferred (graph D). The data

and graphs of this figure may assist the clinician in the decision to transfer 1 or 2 embryos in an individual patient.

DISCUSSION

The main aim of this study was to evaluate whether it is possible to adapt a prognostic model developed in one center in such a manner that it is suitable for clinical use in another one. Uncritical use of the prognostic variables of the development center on the application sample, will almost certainly be a disappointing experience. Circumstances usually differ considerably, even if both centers are located within a distance of 60 kilometres in the same country, work in a similar academic setting and have regular clinical interactions, as was the case for the development and application centers of this study. The choice of another center as an application center was further motivated by the fact that external validation in another center is the strongest type of validation of a prediction model. By applying the model to another center, the degree of generalizability of the model may be shown.

Differences between centers are often not apparent at first sight. In our study for example, there were no obvious differences between the patient populations and the ovarian stimulation protocols. The embryo transfer technique and the laboratory procedures also appeared to be similar, although the results of the embryo quality seemed slightly better in the development center (**Table 2**).

However, when the patient populations were further analysed it was apparent that the proportion of ICSI cycles carried out in the development center (18%) was significantly lower than that in the application center (30%). Poor semen quality is known to be detrimental to IVF outcome (Macklon, 2004), but this effect can be corrected by applying ICSI. Since ICSI cycles were excluded from analysis, this may have contributed to the difference in results between the two centers. However, seemingly standardized laboratory procedures may also have untraceable, subjective elements: it is usually impossible to find out why the embryos of one laboratory lead to more pregnancies than the ones of another one. Socio-economic difference may have also contributed to the difference in pregnancy rate between the two centers, but we had no data to test for such a difference.

The first step in the adaptation process is to make sure that the variables in the data sets of both centers mean the same. In our case, this was possible by categorizing one of the embryo quality measures – the developmental stage - according to the system used in the application center and reclassifying the same variable into 2 instead of the original 3 categories. Hereafter, it was possible to construct a modified model, which could be used in the application center. The modified model performed equally well in the developmental sample as the original model did. Also, it discriminated almost equally well in the application center as in the development center.

With regard to calibration, the modified model estimated the occurrence of pregnancy always too low. This phenomenon was a direct consequence of the better ongoing pregnancy rates in the application center. After correction for this systematic difference, ongoing pregnancy rates were predicted quite well. The prediction of ongoing twin pregnancies considerably improved after a similar correction, but still demonstrated some degree of underestimation implying that the proportion of twin pregnancy was slightly higher than predicted. This example demonstrates that it is possible to adapt the original model and make it suitable for clinical use in the application center.

With regards to discrimination, the discrimination associated with areas under the curve was 0.63 to 0.70. For two couples, with survival analysis methodology, this means that there is only a 63% to 70% chance of saying which one will conceive first. These values are not very high and are typical for prediction of pregnancy in fertility studies. Thus, the model will not be very useful in identifying individual couples. The range in predicted probabilities, however, is wide enough to make a clinically relevant distinction between groups of patients below and above a certain threshold probability. Such a threshold probability may be used to decide for single versus double embryo transfer in a clinical protocol.

It is mandatory to monitor the validity of both models in future patients of the development center (original and modified model) and of the application center (modified model). Possibly some variables have to be adapted, because circumstances change over time. For example, stimulation protocols have become less aggressive and the classification used for the variable 'number of oocytes' may no longer be appropriate. The occurrence of 20-30 oocytes retrieved is rare, the optimal number being now somewhere between 5 and 15. (Van der Gaast, 2006).

The score charts in **Figure 2** give some insight in the complex relationships of the 5 variables predicting success or failure after transfer. The two embryo characteristics appear to have the largest impact on the probability of ongoing pregnancy. If, for example, 2 embryos of poor quality are transferred, the chances of ongoing pregnancy are usually less than 10%. In the frequently occurring case of favorable embryo quality combinations, the impact of older age as a differentiating factor becomes large, also because it exerts its influence twice, once in sum score 1 and once in sum score 2. Moreover, the presence of 2 top embryos is relatively rare in elderly patients. For example, of the 220 patients with 2 (almost) perfect embryos (sum score 1 + sum score 2 is at least 25) only 8% were 38 years or older.

Graph A shows that in the most favorable situation possible of a woman below 25 (score 6), who has more than 20 oocytes retrieved (score 3), with 2 perfect embryos with score 8 on morphology grade and score 13 on developmental stage, the probability of ongoing pregnancy is about 45%. The information in graphs B and C shows that the chances on twin and singleton ongoing pregnancy for such a woman, are about 35% and 10% respectively. If only one embryo would be transferred, her chance on a singleton would be almost 40% (graph D). Most physicians with this information would probably prefer to transfer 1 embryo in such a patient. This example demonstrates that the information in **Figure 2** enables the clinician to estimate the probability of ongoing twin and ongoing singleton pregnancy for all individual patients.

Many centers are adopting an SET policy for all patients below the age of 35 or 38 (Schieve, 2006) whereas in older patients dual ET is still the generally accepted rule. Based on our analysis, we propose to consider a more individualized approach. Patients of 30 or younger should usually

Female factors					
Woman's age (years)	20-25	26-30	31-35	36-40	
Score	6	4	2	0	…
No. of retrieved oocytes	20-30	10-20	2-10		
Score	3	1	0		…
			Score woman:		**…**

Best embryo: quality factors					**Second best embryo: quality factors**				
Morphology grade	2	1	0		Morphology grade	2	1	0	
Score	8	4	0	…		8	4	0	…
Developmental stage	Excellent		Retarded		Developmental stage	Excellent		Retarded	
Score									
if Day 3	13		9	…	If Day 3	13		9	…
if Day 4	13		0	…	If Day 4	13		0	…
	Score embryo 1:			**…**		Score embryo 2:			**…**

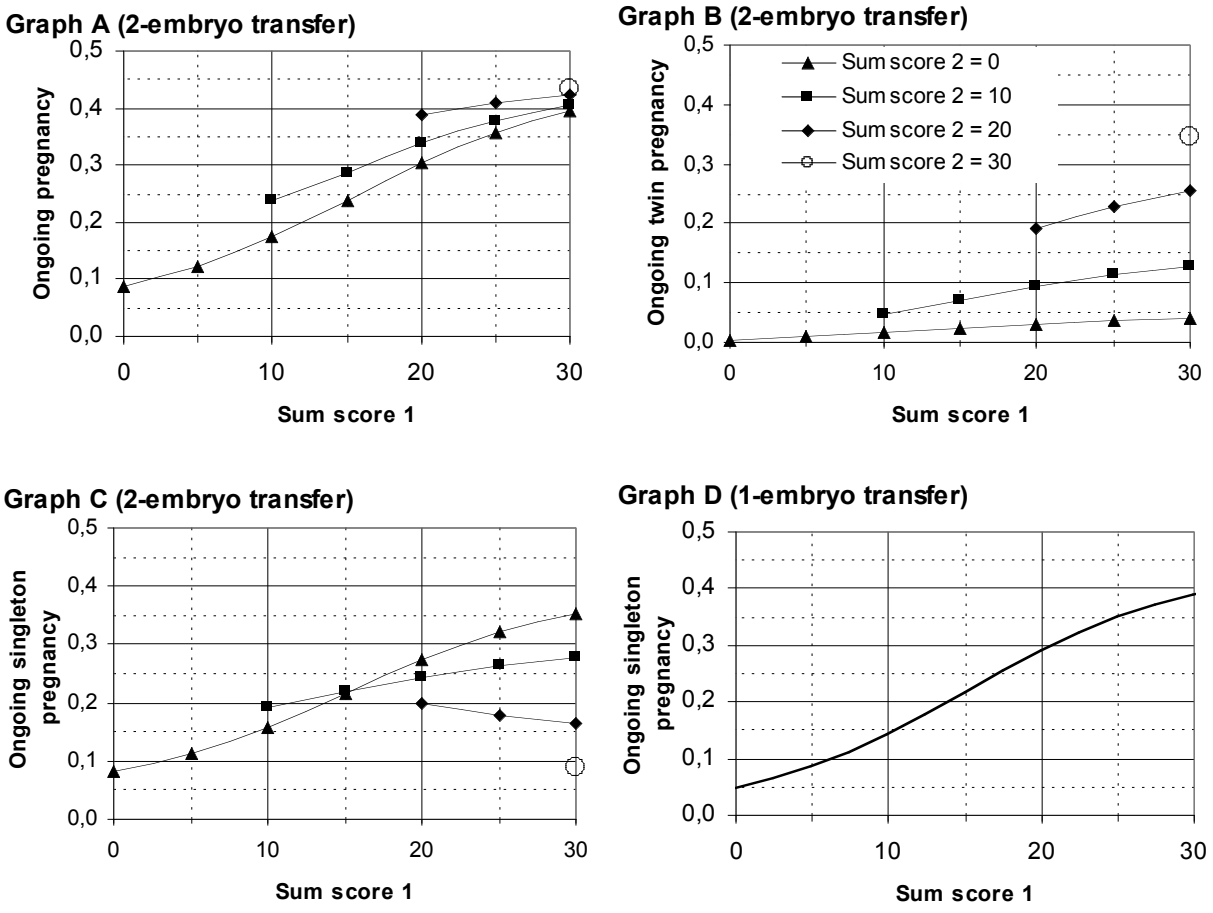

Figure 2: Legend on the following page.

Figure 2: Score chart for predicting the probability of ongoing (twin and/or singleton) pregnancy from the female and embryo factors of the modified EU-model.

The scores can be used to calculate the probability of an ongoing, an ongoing singleton or an ongoing twin pregnancy if 2 embryos are transferred and of an ongoing singleton pregnancy if one embryo is transferred. The procedure runs in 4 steps:

First step: add the scores for woman's age to number of retrieved oocytes to obtain "*Score Woman"*.

Second step: Combine the scores of the morphological grade and developmental stage of the best embryo to obtain "*Score Embryo 1"* and of the second best embryo to obtain "*Score Embryo 2".*

Third step: add the "*Score Woman"* to *the "Score Embryo 1"* to obtain Sum score 1 and the *"Score Woman"* to the *"Score Embryo 2"* to obtain Sum score 2:

Sum score 1 = *Score Woman + Score Embryo 1*
Sum score 2 = *Score Woman + Score Embryo 2*

Fourth step: read the probabilities for an ongoing pregnancy in the corresponding graphs below as follows:
-Graph A: probability of ongoing pregnancy in case of dual ET: use Sum score 1 and Sum score 2
-Graph B: probability of twin pregnancy in case of dual ET: use Sum score 1 and Sum score 2
-Graph C: probability of singleton pregnancy in case of dual ET: use Sum score 1 and Sum score 2
-Graph D: probability of ongoing pregnancy in case of single ET: use Sum score 1

Example: A woman of 32 years had 12 oocytes retrieved at ovum pickup. The two embryos selected for transfer at day 3 had morphology grades of 2 and 1 respectively and both were excellent with regard to the stage development.
Step 1: The prognostic index of the woman is calculated as *Score Woman* = 2+1 = 3.
Step 2: The *Score Embryo 1* is 8+13=21. The *Score Embryo 2* is 4+13 = 17.
Step 3: The sum score1 is 3+21=24 and the sum score2 is 3+17=20.
Step 4: From graph A we infer that her chance of ongoing pregnancy when both embryos are transferred is about 42%. From graphs B and C we read that her chances of a twin and singleton pregnancy are ~ 23% and ~ 19% respectively. Graph D tells us that her chance of a singleton pregnancy is about 34% when the best embryo is transferred.

have one embryo replaced except in the rare cases when few oocytes are retrieved and 2 oocytes of moderate or poor quality are present. In such patients, dual ET might considerably improve the ongoing pregnancy prospects with an acceptable low chance of twin pregnancy. In patients of 38 years or above, dual ET remains the preferred management unless the quality of both embryos is (almost) perfect (sum score 1 + sum score 2 is at least 25).

In such cases single ET is to be preferred. In patients between 30 and 38 single ET is always the right choice if 2 embryos of (almost) perfect quality are present. In all other cases the alternative single versus dual ET, should be decided in dialogue with the couple, according to the guidelines given in the figure.

We conclude that after a careful adaptation, the model shows a good performance in the application center. Monitoring of model-validity in future patients is required. This type of model may assist clinicians in making a rational choice between SET and DET.

References

Altman DG, Royston P (2000) What do we mean by validating a prognostic model? *Stat Med.* 19, 453-73.

Harrell FE, Lee KL, Mark DB (1996) Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med*. 15, 361-87.

Hazekamp J, Bergh C, Wennerholm UB, Hovatta O, Karlstrom PO, Selbing A (2000) Avoiding multiple pregnancies in ART: consideration of new strategies. *Hum Reprod*. 15, 1217-9.

Huisman GJ, Fauser BC, Eijkemans MJ, Pieters MH (2000) Implantation rates after in vitro fertilization and transfer of a maximum of two embryos that have undergone three to five days of culture. *Fertil Steril.* 73, 117-22.

Hunault CC, Eijkemans MJ, Pieters MH, te Velde ER, Habbema JD, Fauser BC, Macklon NS (2002) A prediction model for selecting patients undergoing in vitro fertilization for elective single embryo transfer. *Fertil Steril*. 77, 725-32.

Justice AC, Covinsky KE, Berlin JA (1999) Assessing the generalizability of prognostic information. *Ann Intern Med.* 130, 515-24.

Laupacis A, Wells G, Richardson WS, Tugwell P (1994) Users' guides to the medical literature. V. How to use an article about prognosis. Evidence-Based Medicine Working Group. *JAMA.*272, 234-7.

Olivennes F (2000) Avoiding multiple pregnancies in ART. Double trouble: yes a twin pregnancy is an adverse outcome. *Hum Reprod*. 15, 1663-5.

Ozturk O, Bhattacharya S, Templeton A (2001) Avoiding multiple pregnancies in ART: evaluation and implementation of new strategies. *Hum Reprod.* 16, 1319-21.

Macklon NS, Pieters M, Fauser BCJM (2004) *Textbook of Assisted Reproductive Techniques: Laboratory and Clinical Perspectives*, Eds : Gardner DK, Weissman A, Howles CM, Shoham Z, Taylor and Francis.

Schieve LA. The promise of single-embryo transfer (2006) *N Engl J Med* 354, 1190-3.

Stolwijk AM, Straatman H, Zielhuis GA, Jansen CA, Braat DD, van Dop PA, Veerbeek AL (1998) External validation of prognostic models for ongoing pregnancy after in-vitro fertilization. *Hum Reprod.* 13, 3542-9.

Van der Gaast MH, Eijkemans MJC, van der Net JB, de Boer EJ, Burger CW, van Leeuwen FE, Fauser BCJM, Macklon NS (2006) The optimum number of oocytes for a successful first IVF treatment cycle. *RBM Online* (in press).

Van Houwelingen JC and Le Cessie S (1990) Predictive value of statistical models. *Stat Med*. 9, 1303-25.

Van Kooij RJ, Looman CW, Habbema JD, Dorland M, te Velde ER (1996) Age-dependent decrease in embryo implantation rate after in vitro fertilization. *Fertil Steril*. 66, 769-75.

Zhou H, Weinberg CR (1998) Evaluating effects of exposures on embryo viability and uterine receptivity in in vitro fertilization. *Stat Med*. 17, 1601-12.

7

General discussion

In this final chapter, the research questions will be answered, limitations and implications of the findings in this thesis will be discussed, and conclusions are drawn.

ANSWERS TO THE RESEARCH QUESTIONS

Question 1: **Does the combination of existing models for predicting the chance of pregnancy among untreated subfertile couples result into improved predictions?**
Yes, the predictions of the synthesis model were better than or at least comparable to those of the individual existing models. Moreover the predictions of the synthesis model have a broader empirical basis.

Two models were developed to predict the chance of spontaneous pregnancy leading to live birth among subfertile couples, based on a synthesis of three previously published models (Eimers *et al.*, 1994) (Collins *et al.*, 1995) (Snick *et al.*, 1997). Two models include the PCT as a predictor, the other not (Collins *et al.*, 1995). Whether a patient was referred by her General Practitioner instead of by a gynaecologist, was found to be an independent predictor of pregnancy among untreated subfertile patients. This predictor in the synthesis models was not included in the three original models. This might explain to some extent why the three original models gave less consistent predictions than the synthesis model.

Question 2: **Are models for predicting the chance of ongoing pregnancy among untreated subfertile couples externally valid?**
Not immediately. But after a correction reflecting the overall pregnancy rate in the external population, the models yield reliable predictions.

The Eimers model had to be calibrated by correcting for the difference in overall pregnancy rate between the Eimers population and Collins population in order to provide reliable predictions in the external sample constituted by the Collins population (data collected in the eighties). The synthesis models were valid in the external sample of recent patients (data collected in 2000). In this sample of patients, the PCT was found to remain an important predictor of ongoing spontaneous pregnancy.

Question 3: **Can a valid model be developed for assisting in the choice between single and double embryo transfer?**
An internally valid model for predicting pregnancy and twin-pregnancy chances could be developed, containing the woman's age, number of retrieved oocytes, day of transfer and embryo quality scores as predictors of ongoing pregnancy after embryo-transfer in the first IVF cycle.

The resulting model enables the calculation of probabilities of pregnancy and twin pregnancy, which constitutes important information for the choice between one- and two-embryo transfers.
During external validation, the embryo-quality data of the external centre had to be transformed in order to make them compatible to those of the development centre. Further, the insertion of a correction factor was required to improve the calibration of the model. The model performed well in the application centre only after a careful adaptation. Monitoring the model validity in future patients is required.

LIMITATIONS

Correction of a model for application in another centre

Predictive models can generate unreliable predictions when applied in a centre different from the centre where the model was developed, even after careful internal validation. In chapter 6, when applying the Embryo-Uterus (EU) model to the validation centre, we observed a systematic difference in pregnancy rate between the development and the validation samples. We had to re-calibrate the model by using a correction factor equal to the Odds ratio (OR) for ongoing pregnancy and ongoing twin pregnancy between both samples to obtain more reliable predictions. The question arises, how efficient this correction is in different situations. To deal with this question, we performed an exploratory simulation analysis for a large number of situations of differences in uterine receptivity and embryo viability between development and validation centre (results not shown). The conclusion is that correction performs well in many situations, but not always. Unfortunately, it is not possible to know in advance whether the correction will be efficient.

How useful are the models for future use?

The usefulness of a prediction model in the future depends on the "time-robustness" of the predictors included in the model. Concerning the prediction of spontaneous pregnancy, the variable 'referral status' was found to be an important predictor (chapter 3). This variable is probably related to the –unknown- proportion of couples treated before being referred in the Eimers and Collins samples. Nowadays, treatment is often started earlier, also for patients who still have a good chance of spontaneous pregnancy. Therefore, the effect of the variable "referral status" could have changed over time. In the data collected prospectively to validate the synthesis models, the hazard ratio of the variable "referral status" was increased compared to the pooled Eimers, Snick and Collins data (HR=2.4; 95% CI (1.0-5.8) and HR=1.4; 95% CI (1.1-1.7) respectively). Some other predictors -female's age, duration of infertility and motility- are robust predictors whose effect was estimated at a period where treatment was not started as early as nowadays. The PCT is a debated test. Some studies have expressed doubts about the average quality of the PCT test (Oei *et al.*, 1998). A large study is currently in progress (OFO project) to assess this issue.
Concerning the prediction of the chance of pregnancy in a first IVF cycle, predictors such as the woman's age or number of retrieved oocytes are robust. However, the most important predictor was the embryo quality of the two transferred embryos. Since the way to assess embryo quality is not standardised, as already observed between the development and validation samples, it is likely that "new" predictors of embryo quality will appear in the future, which will make adaptations of our model necessary.

IMPLICATIONS FOR CLINICAL PRACTICE

Pro's of the synthesis approach.

Clinicians have the choice between several published models predicting the chance of pregnancy without treatment. The three original models (Eimers, Snick, Collins), and the two synthesis models use roughly the same diagnostic information and have all been externally validated, to some extent. So the question arises: which model should be used for a new couple?

Table 1 presents four case histories representing a wide spectrum of prognostic possibilities. Case 1 has a very unfavourable prognosis with a relatively old female partner and a long duration of infertility. In contrast, case 2 has a very favourable prognosis with a relatively young female, short duration of infertility and optimal PCT. Case 3 and 4 represent couples with an intermediate prognosis.

Table 1: The four cases histories.

	Case 1	Case 2	Case 3	Case 4
Woman's age	39	21	28	30
Type of subfertility	Primary	Secondary	Primary	Secondary
Duration of subfertility (months)	60	14	18	24
Referral status[a]	Tertiary	Secondary	Secondary	Secondary
Ovulatory disturbance	None	None	None	None
Tubal pathology	None	None	None	None
Post Coital test	No progressive motility	Progressive motility	2 immotile sperm cells	No sperm cells
Sperm analysis				
Volume (ml)	2.5	> 2	2.5	3.5
Concentration (x 10^6/ml)	15	20	8	60
Motility (%)	20	> 40	60	45
Morphology (%)	20	> 40	25	60

[a] Tertiary patients are referred by a gynaecologist, secondary patients by a General Practitioner

Table 2 shows the chance of spontaneous pregnancy leading to live birth at one year after intake, as predicted by the different models. The original six-variable Snick model provided on average higher chances and the Collins model lower chances than the other models. The original Snick model with 4 variables includes the PCT result whereas the original Snick model with 6 variables –developed as an alternative in case the PCT is not available- does not include it. This explains why the Snick model with 4 variables gives much lower chances to cases 3 and 4, which both have an abnormal PCT result, than the 6-variable Snick model.

Table 2: Estimated 1-year live birth chance according to the previously published models and the synthesis models.

	Case 1	Case 2	Case 3	Case 4
Eimers (1994) (ongoing pregnancy)	10	76	25	18
Updated Eimers (2002) (ongoing pregnancy)	9	56	22	15
Collins (1995) (live birth)	4	39	26	20
Snick (four variables) (1997) (live birth)	10	45	15	15
Snick (six variables) (1997) (live birth)	15	54	40	42
Synthesis model without PCT (2004) (live birth)	7	75	39	40
Synthesis model with PCT (2004) (live birth)	5	79	22	27

A major difference between the samples used to develop the three original models was the care setting: the Snick models were developed using secondary patients, the Collins model using tertiary patients and the Eimers model a mix of these two. The synthesis models have a broader empirical base. Interestingly, the synthesis predictions are not the average of the three other models, as one (naively) would expect. This confirms that the synthesis model should really be regarded as a new model, rather than a kind of average of the three original models.

Clinical usefulness of the synthesis models

The clinical usefulness of the two synthesis models was assessed and compared in the patients of the pooled Snick and Eimers samples, because in these patients both models could be evaluated (remember that the PCT was not performed in the Collins study). According to the chance given by the respective model, the patients were grouped in three categories of spontaneous pregnancy leading to live birth within 1 year: poor chance, intermediate and high, with different cut-offs (**Table 3**). The clinical usefulness of a model can be expressed as the percentage of patients assigned by the model in the two extreme categories (low and high chances), because in these categories a definite choice for treatment or for expectant management can be made. This measure of clinical usefulness gives a clinical interpretation to the concept of discriminative ability of a model in a given patient group. The clinical usefulness of the models varies with the cut-offs used, but the model with PCT had always a higher clinical usefulness than the model without PCT. Of course, the percentage predicted in the extreme categories corresponds to clinical usefulness only if the predicted chances are reliable i.e. if the prediction model is well calibrated. On these data used here, this was the case.

Table 3: Clinical usefulness of the synthesis models in the pooled Snick-Eimers samples. The table gives the percentage of patients that have either a low or a high chance of pregnancy according to the synthesis models.

	<20 or >40%	<20 or >35%	<30 or >45%
Model with PCT	68%	77%	76%
Model without PCT	47%	56%	73%

Implementation of prediction models in practice

Most of the time, physicians trust their own intuitive judgment more than prediction models (Liao and Mark, 2003). One of the reasons underlying the physicians' non-use of prediction models may be the belief that prediction models cannot take into account all relevant facts in clinical consultation. Only a few predictors are included in a prediction model whereas the physician has much more information concerning the patient at his disposition. Physicians may also not be convinced that prediction models based and validated on historical data, sometimes even collected 10 to 20 years ago, are appropriate for current practice.

User-unfriendliness of prediction models can also be an obstacle for their use. If the model allows it, like the synthesis models predicting spontaneous pregnancy chances, a score chart can facilitate the implementation of the model. In case of a more complex model, the score chart or nomogram will also become more complex. An example is the embryo-uterus model. This model is complex because intermediate chances (the chances that the best embryo and the second best embryo are viable) are combined to predict the final probability of ongoing pregnancy and ongoing twin pregnancy. Nevertheless, we were able to develop a score chart. In this case, an

electronic version of the prediction model on an Internet site could further facilitate the use of the model.

Topics for further research

Re-calibration of the model for predicting pregnancy after embryo-transfer: The pregnancy rates have improved in the centre where this model has been developed. Therefore, the model should be re-calibrated in order to make predictions for current patients in this centre.

Validation of the model predicting pregnancy after embryo-transfer on single embryo transfer data: Recently, several studies have been published in which elective transfer of only one embryo was performed to reduce the incidence of twin pregnancies in IVF treatment, without compromising the overall ongoing pregnancy rate (Gerris *et al.*, 2002) (De Neubourg and Gerris, 2003) (De Sutter *et al.*, 2003) (Tiitinen *et al.*, 2003) (Van Montfoort *et al.*, 2004). Thus large datasets (more than thousand ET's) have been collected about patients in whom only one embryo was transferred because only one embryo was available or because of the couple's preference. In chapter 6, the model developed to select patients for elective single embryo-transfer in the first IVF cycle has been validated in a sample of patients in whom two embryos were transferred. It would be interesting to validate the model also in a sample of patients in whom single embryo-transfer has been performed.

CONCLUSIONS

- A synthesis of three existing models for the prediction of treatment independent pregnancy chances proved superior to the original models
- The value of the PCT in predicting treatment-independent pregnancy chances has been corroborated in the present study.
- Referral status, i.e. referral by GP or by gynaecologist, is of considerable prognostic importance in predicting treatment-independent pregnancy chances.
- The IVF prediction model developed in this thesis is useful in predicting singleton and twin pregnancy chances after single or double embryo transfer.
-Prediction models in fertility medicine are usually not externally valid.
-External validity of models can often be attained by correcting for the difference in overall pregnancy rate between the development centre and the external centre.

References

Collins JA, Burrows EA, Wilan AR. (1995) The prognosis for live birth among untreated infertile couples. *Fertil Steril*. 64, 22-8.

De Neubourg D, Gerris J. (2003) Single embryo transfer - state of the art. *Reprod Biomed Online*. 7, 615-22.

De Sutter P, Van der Elst J, Coetsier T, Dhont M. (2003) Single embryo transfer and multiple pregnancy rate reduction in IVF/ICSI: a 5-year appraisal. *Reprod Biomed Online*. 6, 464-9.

Eimers JM, te Velde ER, Gerritse R, Vogelzang ET, Looman CW, Habbema JD. (1994) The prediction of the chance to conceive in subfertile couples. *Fertil Steril*. 61, 44-52.

Gerris J, De Neubourg D, Mangelschots K, Van Royen E, Vercruyssen M, Barudy-Vasquez J, Valkenburg M, Ryckaert G. (2002) Elective single day 3 embryo transfer halves the twinning rate without decrease in the ongoing pregnancy rate of an IVF/ICSI programme. *Hum Reprod*. 17, 2626-31.

Liao L, Mark DB. (2003) Clinical prediction models: are we building better mousetraps? *J Am Coll Cardiol*. 42, 851-3.

Oei SG, Helmerhorst FM, Bloemenkamp KW, Hollants FA, Meerpoel DE, Keirse MJ. (1998) Effectiveness of the postcoital test: randomised controlled trial. *Bmj*. 317, 502-5.

Snick HK, Snick TS, Evers JL, Collins JA. (1997) The spontaneous pregnancy prognosis in untreated subfertile couples: the Walcheren primary care study. *Hum Reprod*. 12, 1582-8.

Tiitinen A, Unkila-Kallio L, Halttunen M, Hyden-Granskog C. (2003) Impact of elective single embryo transfer on the twin pregnancy rate. *Hum Reprod*. 18, 1449-53.

Van Montfoort AP, Dumoulin JC, Kester AD, Evers JL. (2004) Early cleavage is a valuable addition to existing embryo selection parameters: a study using single embryo transfers. *Hum Reprod*. 19, 2103-8.

Summary / Samenvatting / Résumé

Summary

This thesis deals with two prediction problems in reproductive medicine. The first is the prediction in infertile couples of the chance to conceive without treatment. The second deals with the prediction of the chance of conception in couples treated with *in vitro* fertilization (IVF).

Chapter 1 introduces the notion of infertility. The difference between the terms infertile and subfertile is explained and the prevalence and the main causes of infertility are summarised. Emphasis is on prognosis in untreated sub/infertility and prognosis in IVF. Next, the methodology is discussed, with emphasis on the development of prediction models and their validation.

Chapter 2 investigates the following question: does combining the information of existing models for predicting the chance of pregnancy among untreated subfertile couples result in improved predictions? The original data of three previously published models are used to develop a synthesis model for predicting pregnancy leading to live birth without treatment. The predictors used are duration of subfertility, women's age, primary or secondary infertility, percentage of motile sperm, and whether the couple was referred by a general practitioner or by a gynaecologist (referral status). Because the post-coital test (PCT) was not assessed in one of the studies, a second synthesis model including the PCT is developed, based on the remaining two studies. The ability of the synthesis models to distinguish between women who did and who did not become pregnant (also called 'discriminative ability') is compatible to the ability to the original models. The reliability of the predictions by the three-sample synthesis model (also called 'calibration' of the model) is somewhat better. Predictions improve considerably by including the PCT in the synthesis model.

Chapter 3 is dedicated to external validation of one of the three previous published models, the so-called Eimers model (according to the first author's name of the study). Indeed, before a prognostic model can reliably be used in clinical practice, it has to be validated in other clinical settings. The Eimers model was developed in 1994 to predict spontaneous pregnancy among subfertile couples. Live birth rates as predicted by the Eimers model are tested against observed live birth rates in a Canadian cohort of 1061 couples consulting for subfertility due to cervical hostility, male subfertility, or unexplained subfertility. Overall, the prognostic effect of the predictors does not differ significantly in both populations. The model shows a moderate discriminative ability in the Canadian population (c index = 0.62). When the Eimers model is adjusted for the difference in the average live birth rate between the Eimers study and the Canadian study, the calibration of the model is satisfactory.

Chapter 4 describes the external validation of the synthesis models that were described in chapter 2. The models are applied to a cohort of 302 couples consulting for subfertility, whose characteristics were collected prospectively on this purpose. As to be expected, both models provide a slightly lower discrimination in the validation sample than in the original sample. Calibration is good: the observed and predicted probabilities of pregnancy without treatment leading to live birth do not differ for both models. The use of PCT improves the discrimination of the models. These models can be useful in counselling subfertile couples.

The next two chapters deal with the prediction of the chance to conceive with IVF treatment. In **Chapter 5**, a prognostic model is developed to help in the selection of patients for elective single

embryo transfer, based on data from 642 women undergoing their first IVF treatment cycle, where two embryos were transferred. The best predictors for ongoing pregnancy are female age, the number of retrieved oocytes, the developmental stage score and the morphology score of the two best embryos available for transfer, and the day of transfer. Younger age and high quality of transferred embryos are the best predictors for increased risk of multiple pregnancy. The resulting model enables the calculation of probabilities of overall pregnancy, single pregnancy and twin pregnancy. Depending on embryo quality, there is a threshold age under which the chance of singleton pregnancy is higher if one embryo is transferred compared to two embryos. Application of this model may enable a reduction in the chance of twin pregnancy without compromising singleton pregnancy rates in a subgroup of patients undergoing IVF.

Chapter 6 evaluates the application of the model developed in Chapter 5 in a different centre. It concerned 494 consecutive first IVF cycles. A first obstacle is the use of different embryo scoring systems in the two centres. It is thus necessary to adapt the embryo scoring system to be compatible with that employed by the prediction model. Secondly, the success rate differs considerably between the two centres and insertion of a correction factor equivalent to the difference in odds ratios for ongoing pregnancy rates is necessary to improve the calibration of the model. A score chart for predicting the probability of singleton or twin pregnancy is constructed. The areas under the ROC (or equivalently the c-index) for ongoing pregnancy and twin pregnancy are 0.66 and 0.70 respectively. After these adaptations, the model performs well in the application centre.

This thesis ends in **Chapter 7** with a general discussion of the findings of the presented studies. The conclusions are as follows: 1) A synthesis model of three existing models for the prediction of pregnancy chances without treatment proved to be superior to the original models. 2) The value of the PCT in predicting treatment-independent pregnancy chances has been corroborated. 3) Referral status, i.e. referral by GP or by gynaecologist, is of considerable prognostic importance in predicting treatment-independent pregnancy chances. 4) The IVF prediction model developed in this thesis is useful in predicting singleton and twin pregnancy chances after single or double embryo transfer. 5) Prediction models in fertility medicine are usually not directly externally valid. 6) External validity of models can often be much improved by simply correcting for the difference in overall pregnancy rate between the development centre and the external centre.

Samenvatting

Dit proefschrift gaat over twee predictieproblemen in de voortplantingsgeneeskunde. Ten eerste is dit de predictie van de kans dat onvruchtbare paren zonder behandeling zwanger worden. Het tweede gaat over de predictie van de kans om zwanger te worden voor paren die behandeld worden met *in vitro* fertilisatie (IVF).

Hoofdstuk 1 licht het begrip onvruchtbaarheid toe. Het verschil tussen de termen onvruchtbaar en subfertiel wordt uitgelegd en de prevalentie en de belangrijkste oorzaken van onvruchtbaarheid worden samengevat. Het accent ligt op voorspelling bij onbehandelde onvruchtbaarheid en voorspelling bij IVF. Vervolgens wordt de methodologie besproken, met de nadruk op de ontwikkeling van predictiemodellen en hun validatie.

In **Hoofdstuk 2** wordt de volgende onderzoeksvraag onderzocht: geeft het combineren van de informatie uit reeds bestaande modellen een betere voorspelling van de kans op zwangerschap voor onbehandelde onvruchtbare paren? De originele data van drie reeds gepubliceerde modellen zijn gebruikt om een synthese model te ontwikkelen, dat de kans voorspelt op een zwangerschap leidend tot een levendgeboren kind, zonder behandeling. De gebruikte predictoren zijn de duur van onvruchtbaarheid, de leeftijd van de vrouw, of het paar al eerder zwanger was geweest, het percentage beweeglijke zaadcellen en of het paar was doorgestuurd door een huisarts of door een gynaecoloog (verwijzingsstatus). Omdat de post-coïtum test (PCT) in een van de studies niet bepaald was, is een tweede synthese model, met de PCT, ontwikkeld op de twee overblijvende studies. Het vermogen van de synthese modellen om vrouwen die zwanger werden te onderscheiden van vrouwen die niet zwanger werden (i.e. de discriminatie van het model) is vergelijkbaar met het onderscheidingsvermogen van de originele modellen. De betrouwbaarheid van de kansschattingen van het model gebaseerd op drie groepen (i.e. de calibratie van het model) is wat beter. De voorspellingen worden aanzienlijk beter met het toevoegen van de PCT in het synthese model.

Hoofdstuk 3 is gewijd aan de externe validatie van een van de drie eerder gepubliceerde modellen, het zogenaamde Eimers model (naar de naam van de eerste auteur van de studie). Voordat een prognostisch model betrouwbaar kan worden gebruikt in de klinische praktijk, moet het gevalideerd worden in andere klinieken. Het Eimers model werd in 1994 ontwikkeld om zwangerschap zonder behandeling te schatten voor subfertiele paren. De voorspelde kansen op een levendgeboorte zijn vergeleken met de frequentie van deze uitkomst in een Canadees cohort van 1061 subfertiele paren. Over het geheel genomen is er geen significant verschil in prognostisch effect van de predictoren tussen de twee populaties. Het model vertoont een matige discriminatie in de Canadese populatie (c index = 0.62). De calibratie van het model is voldoende als het Eimers model wordt aangepast voor het verschil in gemiddelde levendgeboorte percentage tussen de Eimers studie en de Canadese studie.

Hoofdstuk 4 beschrijft de externe validatie van de synthese modellen beschreven in hoofdstuk 2. De modellen zijn toegepast in een cohort van 302 paren met onvruchtbaarheid, van wie de karakteristieken daartoe prospectief werden verzameld. Zoals verwacht, gaven beide modellen een wat lagere discriminatie in de validatie groep dan in de originele groep. Calibratie is goed: de geobserveerde en voorspelde kansen van zwangerschap, zonder behandeling, met een levendgeboren kind als uitkomst zijn niet significant verschillend. Het gebruik van de PCT

verbetert de discriminatie. Deze modellen kunnen nuttig zijn bij de counseling van subfertiele paren.

De volgende twee hoofdstukken gaan over de voorspelling van de kans om zwanger te worden met IVF behandeling. In **hoofdstuk 5** wordt een prognostisch model ontwikkeld om het gemakkelijker te maken om patiënten te selecteren waarbij één embryo kan worden teruggeplaatst. Dit model is gebaseerd op de gegevens van 642 vrouwen die hun eerste IVF behandeling met terugplaatsing van twee embryo's ondergingen. De beste voorspellers voor een doorgaande zwangerschap zijn: de leeftijd van de vrouw, het aantal verkregen oocyten, de ontwikkelingsscore en de morfologische score van de twee beste embryo's, en de dag van terugplaatsing. Jonge leeftijd en hoge kwaliteit van de embryo's voorspellen een grotere kans op een meerlingzwangerschap het best. Het resulterende model maakt de berekening van de kans op zwangerschap, eenlingzwangerschap en tweelingzwangerschap mogelijk. Afhankelijk van de kwaliteit van de embryo's, bestaat er een leeftijdsgrens waaronder de kans op een eenlingzwangerschap groter is als één embryo wordt teruggeplaatst in plaats van twee. Toepassing van dit model kan een vermindering van de kans op een tweelingzwangerschap mogelijk maken zonder het percentage eenlingzwangerschappen te verkleinen, in een subgroep van patiënten die IVF behandeling ondergaan.

Hoofdstuk 6 evalueert de toepassing van het model ontwikkeld in hoofdstuk 5 in een ander centrum. Het betreft 494 consecutieve eerste IVF behandelingen. Een eerste obstakel is het gebruik van verschillende embryo score systemen in de twee centra. Aanpassing van het embryo score systeem om het compatibel te maken met het predictie model is dus noodzakelijk. Ten tweede is de kans op zwangerschap na IVF erg verschillend tussen beide centra en moet voor dit verschil gecorrigeerd worden om de calibratie van het model te verbeteren. Een scoringskaart is gemaakt om de kans op een eenling- of tweelingzwangerschap te voorspellen. Het oppervlak onder de ROC (gelijk aan de c-index) voor doorgaande zwangerschap en voor tweelingzwangerschap zijn respectievelijk 0.66 en 0.70. Na de aanpassingen presteerde het model goed in het andere centrum.

Dit proefschrift wordt afgesloten in **hoofdstuk 7** met een algemene discussie over de bevindingen van de gepresenteerde studies. De conclusies zijn als volgt: 1) Een synthese model gebaseerd op drie reeds bestaande modellen bleek beter te voldoen dan de originele modellen voor het voorspellen van de kansen op zwangerschap zonder behandeling. 2) De waarde van de PCT in de voorspelling van zwangerschap zonder behandeling werd bevestigd. 3) De verwijzingsstatus, i.e. verwijzing door een huisarts of door een gynaecoloog, is van behoorlijk prognostisch belang in de voorspelling van de kansen op zwangerschap zonder behandeling. 4) Het IVF predictiemodel ontwikkeld in dit proefschrift is nuttig voor de voorspelling van de kans op eenling- en tweeling zwangerschappen na terugplaatsing van één of twee embryo's. 5) Predictiemodellen in de voortplantingsgeneeskunde zijn meestal niet direct toepasbaar in andere centra. 6) Externe geldigheid van modellen kan vaak aanzienlijk worden verbeterd door te corrigeren voor het verschil in zwangerschapsfrequentie tussen de centra.

Résumé

Cette thèse traite de deux problèmes de prédiction dans le domaine de la médecine de la reproduction. Le premier problème est de prédire la chance de concevoir sans traitement pour des couples inféconds. Le second est de prédire la chance de concevoir chez les couples traités par fécondation *in vitro* (FIV).

Le **chapitre 1** introduit la notion d'infertilité. La différence de sens entre les termes 'stérile' et 'fécondation réduite' est expliquée, et la prévalence et les principales causes d'infertilité sont résumées. L'accent est porté sur le pronostic de la fécondation réduite/stérilité et le pronostic en fertilisation *in vitro*. Ensuite, la méthodologie utilisée pour réaliser des modèles prédictifs est exposée, et une importance particulière est donnée au développement et à la validation des modèles prédictifs.

Dans le **chapitre 2,** la première question de recherche est examinée : le fait de combiner l'information contenue dans des modèles prédictifs déjà existants améliore-t-il la prédiction de la chance de concevoir pour des couples ayant une fécondation réduite et n'étant pas traités ? Les données originales de trois modèles prédictifs antérieurement publiés sont utilisées afin de développer un modèle dit 'de synthèse' pour prédire les grossesses survenant sans traitement et aboutissant à la naissance d'un enfant vivant. Les facteurs de prédiction utilisés sont la durée de la fécondation réduite, l'âge de la femme, le caractère primaire ou secondaire du problème de fécondation réduite, le pourcentage de spermatozoïdes mobiles, et le fait que le couple soit adressé à l'hôpital par un médecin généraliste ou un gynécologue ('statut de référence'). Un second modèle incluant le test post-coïtal (PCT) est développé à partir de deux études seulement, le PCT n'ayant pas été pratiqué dans la troisième étude. La capacité des modèles de synthèse à différencier les femmes qui sont effectivement devenues enceintes de celles qui ne le sont pas devenues (encore appelée 'capacité discriminatoire' des modèles) est comparable à celle des modèles originaux quand ceux-ci sont appliqués aux autres groupes. La fiabilité des prédictions (encore appelée 'calibration') du modèle développé à partir des trois études est un peu meilleure. Les prédictions sont largement améliorées par l'inclusion du PCT dans le modèle de synthèse.

Le **chapitre 3** est dédié à l'évaluation de la validité externe de l'un des trois modèles antérieurement publiés, le modèle dit 'Eimers' (d'après le nom du premier auteur de l'étude). Avant d'utiliser un modèle de prédiction en pratique clinique courante, un modèle de prédiction doit en effet être validé au préalable, dans un centre différent de celui où il a été développé. Le modèle Eimers fut développé en 1994 afin de prédire les chances de conception parmi les couples ayant une fécondation réduite et n'étant pas traités. Les taux de naissance d'enfants vivants prédits par le modèle Eimers sont ici comparés aux taux de naissance d'enfants vivants observés dans une cohorte canadienne de 1061 couples consultant pour fécondation réduite due soit à une hostilité cervicale, soit à une déficience relative à la fécondité de l'homme, soit à une fécondité réduite inexpliquée. Globalement, la valeur pronostique des facteurs de prédiction ne diffère pas significativement entre les deux populations. Le modèle présente une capacité discriminatoire modérée dans cette population canadienne (c index = 0.62). La calibration du modèle Eimers est satisfaisante quand le modèle est corrigé pour la différence en taux moyen de naissances vivantes entre les deux études.

Le **chapitre 4** décrit la validation externe des modèles de synthèse décrits au chapitre 2. Les modèles sont appliqués à une cohorte de 302 couples consultant pour fécondation réduite, et dont les caractéristiques ont été collectées prospectivement. Logiquement, les deux modèles présentent une capacité discriminatoire légèrement inférieure dans le groupe utilisé pour la validation que dans le groupe utilisé pour le développement des modèles. La calibration est bonne : les probabilités observées et prédites, de grossesses sans traitement aboutissant à la naissance d'enfants vivants, ne diffèrent pas pour les deux modèles. L'inclusion du PCT améliore grandement la discrimination des modèles. Ces modèles peuvent s'avérer utiles pour informer et conseiller des couples présentant une fécondation réduite.

Les deux chapitres suivants traitent de la prédiction de la chance de concevoir avec un traitement de fécondation *in vitro* (FIV). Dans le **chapitre 5**, un modèle prédictif est développé afin d'aider la sélection de patients pour le transfert électif d'un seul embryon. Ce modèle est basé sur les données de 642 femmes ayant bénéficié d'un premier traitement FIV au cours duquel deux embryons furent transférés. Les meilleurs facteurs de prédiction des grossesses sont l'âge de la femme, le nombre d'ovocytes ponctionnés, le score de développement et le score morphologique des deux meilleurs embryons disponibles pour le transfert, et le jour de transfert. Un âge jeune et des embryons d'excellente qualité sont les plus grands facteurs de risque de grossesse multiple. Le modèle obtenu permet de calculer les probabilités de grossesse en général, de grossesse unique et de grossesse gémellaire. Dépendant de la qualité des embryons, il existe un âge limite en-dessous duquel la chance de grossesse unique est plus élevée si un seul embryon est transféré que si deux embryons sont transférés. L'utilisation de ce modèle permet la réduction des risques de grossesses gémellaires sans compromettre le taux de réussite des grossesses uniques dans un groupe particulier de patientes bénéficiant d'un traitement FIV.

Le **chapitre 6** évalue l'application du modèle développé au chapitre 5 dans un centre différent. Il concerne 494 cycles consécutifs de premiers traitements FIV. Un premier obstacle est que les systèmes d'évaluation de la qualité des embryons diffèrent entre les deux centres. Il est par conséquent nécessaire d'adapter le système d'évaluation de la qualité des embryons afin qu'il devienne compatible avec celui utilisé dans le modèle prédictif. Ensuite, le taux de réussite diffère considérablement entre les deux centres et l'insertion d'un facteur de correction égal à la différence des odds ratios pour les taux de grossesses est nécessaire pour améliorer la calibration du modèle. Une grille de scores et des graphes sont construits pour prédire la probabilité de grossesse unique et/ou de grossesse gémellaire. Les aires sous la courbe ROC (équivalentes au c index) pour les grossesses uniques et les grossesses gémellaires sont 0,66 et 0,70 respectivement. Après ces adaptations, le modèle fonctionne bien dans le centre d'application.

Cette thèse se termine au **chapitre 7** avec une discussion générale portant sur les résultats des études présentées. Les conclusions sont les suivantes : 1) Un modèle faisant la synthèse de trois modèles préexistants s'est avéré supérieur aux modèles originaux pour prédire les chances de grossesse sans traitement. 2) L'importance du PCT comme facteur de prédiction des chances de grossesse sans traitement a été corroborée. 3) Le statut de référence, c'est à dire le fait que le couple soit adressé à l'hôpital par un médecin généraliste ou un gynécologue, est d'une importance pronostique considérable dans la prédiction des chances de grossesse sans traitement. 4) Le modèle de prédiction FIV développé dans cette thèse est utile pour prédire les chances de grossesse unique et les risques de grossesse gémellaire après transfert d'un ou de deux embryons. 5) En médecine de la reproduction humaine, les modèles prédictifs ne sont généralement pas directement valides dans un centre différent du centre où ils ont été développés. 6) La validité externe des modèles prédictifs peut souvent être améliorée en corrigeant simplement les modèles

vis à vis de la différence de taux de grossesse entre le centre où le modèle a été développé et celui où il est appliqué.

Acknowledgements

Three people have contributed enormously to the composition of this thesis. My promotor Prof. dr. J. Dik F. Habbema, sketched the broad outlines of my articles and my thesis. He taught me a great deal regarding the evaluation of results and the drafting of scientific articles. I am very grateful to him for this and for his ability to bring to fruition every project undertaken. I owe a great deal to Dr. René J. Eijkemans, my co-promotor, for his creativity, his generosity and his patience when explaining to me, in such a clear-cut manner, statistical concepts or reproductive medicine notions and I sincerely thank him for all the help that he has given me. Prof. dr. Egbert R. te Velde was my second promotor. I am indebted to him for his constant endeavour to make my articles as comprehensible as possible for those clinicians reading them, and for his total commitment to submitting articles of the highest quality. I want to express my gratitude to him here for his important and assiduous contribution to the finishing touches of the majority of articles in this thesis.

Prof. dr. Nick S. Macklon has helped me on many occasions. He participated in the composition of this thesis as co-author and member of the inner doctoral committee. He enthusiastically took part in the creation of the Embryo-Uterus model and improved the style of a number of my articles. I wholeheartedly appreciate his personal dedication, his expertise, and his extreme kindness.
I also give thanks to Prof. dr. Theo Stijnen and Prof. dr. Theo J. Helmerhorst for having accepted to become members of the inner doctoral committee and to Prof. dr. Didi D. Braat and Dr. Eric Boersma for becoming part of the plenary doctoral committee. Moreover, I offer my particular gratitude to Prof. Theo Stijnen for his excellent statistics teaching during the classes at the Netherlands Institute for Health Sciences (NIHES).

Dr. Ewout W. Steyerberg was my mentor within the department. I thank him for his readiness to help, his wisdom and his advice regarding the programming of S-Plus software.
I am grateful to Prof. dr. Myriam G. Hunink for having (without my knowing) sent me in the direction of Dr. Ewout W. Steyerberg following a recruitment interview, which in turn led indirectly to me finding employment at the Public Health Department, Erasmus MC Rotterdam (MGZ in Dutch) in 1998. I also thank her for the excellent quality of her teaching during the classes at NIHES.

I am very thankful to Dr. John A Collins for having accepted to provide me with syntaxes and data collected from the CITES study in Canada, as well as having accepted to be the co-author of several of my articles.
Dr. Herman K. Snick likewise helped me a lot, passing on information about the prognostic model that bears his name, voluntarily commenting on my work and giving me encouragement on numerous occasions. I thank him for his interest and his assistance.

Prof. dr. J.L.H. Evers, Prof. dr. Bart C. Fauser, Dr. Joop S. Laven, Dr. Math H. Pieters, Dr. Sjerp M. Weima, Dr. Ilse A. van Rooij, Dr. Ellen R. Klinkert are co-authors of articles used in this thesis. I am grateful to them for their cooperation. My gratitude also goes out to Dr. Arie Verhoeff, Dr. Durk Berks and Lucienne Bax for their help during the

prospective study in Rotterdam. I thank Dr. Bea A.M. Lintsen, Clazien A.M Bouwmans and Leona Hakkart for their collaboration in the IVF cost-effectiveness study.

Dr. Chantal W. Hukkelhoven and Dr. Yvonne Vergouwe were my first colleagues at the MGZ office. I thank them for the moments spent discussing both the trivial and important ups and downs of our lives, and also for having accepted to be by my side the day I will carry out the academic defense of my thesis. Thanks to Chantal, I became adept at working to music and managed to learn a little about SAS software. As for Yvonne, I remember a certain congress in Sicily and how she helped me to forge a greater understanding of S-Plus software. Laetitia M. Verbeek came onto the office scene later on, bringing her smile, her freshness, and her generosity, a pleasure to be around. I would like also to thank Dr. Nino Mushkudiani and Rolf P. Dreier, my last MGZ-roommates, who gave truly an international atmosphere in the office.
I am grateful to Caspar W. Looman for having unearthed obsolete files from the Eimers study and in particular for having remembered that the "live birth" variable had in fact well been collected in 1984. My thanks are also addressed to Prof. Dr. Johan P. Mackenbach, chair of the Public Health Department, for his advices about the drafting of scientific articles, to Peter Faas for his serenity, and also to the other 'computer-men' (IT team members), Roel Faber, Kees Noordsij-Wagenaar and Ton Gerritsen for their efficient support, to the secretaries: Else van den Engel for her inalterable cheerful temper and all our short conversations in French, Sonja Deurloo-van Dam for her dynamism, and Mirela Antonic for her devotedness. Thanks also to Mona Richter for her smile and her humanism.
I would like also to thank Dr. Adrian V. Hernandez for our conversations in Spanish about, for instance, the utility –or not- of learning a dialect, Dr. John P. Puvimanasinghe for his encouragement and advice, and colleagues or ex-colleagues at MGZ: Dr. Ida Korfage, Dr. Rian Rijnsburger, Resi Mangunkusumo, Dr. Hélène Voeten, Suzanne Polinder, Ilse Oonk, Dr. Jacques Fracheboud, Jan Willem van der Steeg and Pieternel Steures, Sita Tan, Dr. Wilma Stolk, Dr. Vivian Bos, Dr. Margriet van Baar, and more particularly those members or ex-members of the Clinical Decision Sciences group (CKB in Dutch): Dr. Pieta Krijnen, Dr. Merel van Dijk, Dr. Cecile Janssens, Dr. Agnes van der Heide, Dr. Astrid Vrakking, Dr. Elsbeth Voogt, Dr. Judith Rietjens and Gerard Borsboom.

Many thanks to Dr. Denis Hémon (INSERM U754, Villejuif) and Dr. Pierre Verger (Institute for Radiation Protection and Nuclear Safety, Fontenay aux Roses) for having helped me to successfully obtain my DEA (Diploma of Advanced Studies) in Public Health in 1998 in Paris, which represented my first step in the world of epidemiological research.

I am extremely grateful to the whole team at the National Poison Information Center (NVIC in Dutch) of the National Institute of Public Health and the Environment (RIVM in Dutch) for having welcomed me so warm-heartedly into their department, and first and foremost this appreciation goes out to Dr. Jan Meulenbelt, Dr. Irma de Vries and Dr. Tjeert T. Mensinga for having shown confidence in me in my new position. Special thanks go to Tjeert for having immersed me in the meticulous world of ´Good Clinical Practice` and for having forced me to convert to using SAS software. Likewise, many thanks both to Maaike Kruidenier and Irma S. Koot for their devoted and kind cooperation during the Cannabis study, to Dr. Marianne E.C. Leenders for her encouragement and advice, to Rene P.M. van den Hoogen, Marieke A. Bednarzyk and Dr. Marieke A. Dijkman, my current roommates, for their kindness.

Zonder Ina van der Tol had ik niet in alle rust kunnen werken en studeren gedurende al die jaren. Ik heb volledig vertrouwen in haar, want ze kon – en kan – zich over mijn kinderen ontfermen met bekwaamheid, beschikbaarheid en vrijgevigheid. Ik ben haar ontzettend dankbaar. Ik wil daarbij ook Hans bedanken voor zijn vriendschap.
Ik bedank mijn huidige buren, Leontine en Jan Nave, voor hun vriendelijkheid, hun vriendschap en hun aanmoedigingen, mijn ex-buurvrouw, Marjolein Molenaar en mijn vrienden Lucy Meijer en René Strik voor de trouw van hen vriendschap, Ria en Dr. Peter Hermans voor hun aanmoedigingen.
Tevens bedank ik Sensei Wim Krijsman die mij vakkundig de kunst van Aïkido leert, en de budokas van de Isshin club Rotterdam die er voor zorgen dat ik lichamelijk en psychologisch in balans blijf.

Ik bedank mijn schoonfamilie –Opa P. Mattijsen, Omi P. H. Mattijsen-Deij, Dideri Mattijsen en haar kinderen, Sterre, Sjeng en Samuel, Coby en Jos Kelbling en hun kinderen, Martijn en Fu Mei, voor hun trouwe steun, hun interesse, hun aanmoedigingen en hun voortdurende inspanningen om ervoor te zorgen dat ik me in Nederland thuis zou voelen.

Je remercie mes ami(e)s Odile et Bart Schulte-Levasseur, Hanane et Renaud Besselièvre, Claudette et Philippe Pelou-Bonneau, Catherine Cailleau, Virginie et Stéphane Grée-Benat, Montse et Andreú Martin-Plans, Doriana et Jean-Laurent Le Carreres, Francoise et Ivan Travais-Colmont, Isabelle et Michel Rattoray-Gandar et Marie Legoulven pour leur amitié et leur fidélité, malgré le temps qui passe et les kilomètres qui nous séparent.

Je remercie les voisins d'Angers, Noël et Marie-Rose Delaunay, Madame et Monsieur Pichot pour leur aide à ma maman au quotidien.
J'ai une pensée très très affectueuse pour mes proches et ma famille : Huguette et Henri Pioger, Marie-Hélène et Jean-Paul Le Gal, Joël et Arlette Fouche, Rachel et Abdel Lahriri-Martin et Ismaël, Guylette et Jacques Raoul-Jourde, Michel Hunault et Bernadette Cabut, Romain et Charlotte pour leur soutien et leur présence qui même de loin me réchauffent toujours le cœur. A vous tous, un énorme merci !
« Vivre la naissance d'un enfant est notre chance la plus accessible de savoir le sens du mot miracle [Paul Carvel] ». Toi ma maman, Colette Hunault-Pineau, et vous mes enfants, Juliette, Hélène et Jean, vous m'avez permis de ressentir ce sentiment au plus profond de moi-même. Vous êtes les ancres affectives qui me relient au passé, au présent et au futur; je vis par et pour vous. Papa n'est plus là physiquement, mais je pense à lui chaque jour, retournant dans ma tête cette phrase « ne pleure pas celui que tu as perdu, mais réjouis toi de l'avoir connu ».
Lieve Paul, het is niet gemakkelijk voor een man om een vrouw te hebben die na haar 40e nog wil studeren en een scriptie wil schrijven, eh ? Il y a plus de 17 ans, le hasard de la Vie a fait que nous nous sommes rencontrés, et les hasards de l'Amour que nous nous sommes aimés. Aujourd'hui, je souhaite que la flamme qui nous lie continue à nous habiter encore longtemps afin de poursuivre côte à côte notre découverte du chemin de la Vie.

Curriculum Vitae

Claudine C. Hunault was born on May 14, 1965 in Angers, France. She obtained her secondary school diploma with special emphasis on mathematics and physics in Le Mans, in 1983 and began her studies of medicine the same year. She obtained her medical degree (M.D.) in 1994 at the Faculty of medicine of the University of Angers. She did part of her residency training at the Department of Infectious Diseases at the Regional Hospital at Angers (Head: Prof. Dr. Achard, supervisor: Dr. Chennebault) and at the Sector 6 of the Psychiatric Hospital at Allonnes, France (Head: Dr. Pennanech).

Between 1994 en 1997, she worked as substitute General Practitioner, first in the west part of France, and later in the Netherlands, mainly in centres for political refugees (COA Hellevoetsluis and Strijen) and in some GP's offices (Maasdam, Puttershoek and Dirksland).

Between 1997 and 1998, she participated in a study about the identification of risk factors for Post Traumatic Stress Disorder after the 1992 flood in Vaucluse (France), at the Institute for Radiation Protection and Nuclear Safety (Fontenay-aux-Roses) (Head: Dr. Denis Bard; supervisor: Dr. Verger) as part of the Diploma of Advanced Studies (DEA) in Public Health that she completed in 1998 at the University of Paris XI.

In November 1998, she started to work at the Center for Clinical Decision Sciences, Department of Public Health, Erasmus MC Rotterdam, where she conducted the research described in this thesis (Promotors: Prof. dr. Dik Habbema and Prof. dr. Egbert te Velde; Co-promotor: Dr. René Eijkemans). This work was performed in close collaboration with the Departments of Reproductive Medicine of the Erasmus MC Rotterdam and of the University Medical Center Utrecht. The same year, she joined the Netherlands Institute for Health Sciences (NIHES) of the Erasmus University in Rotterdam. She completed the M.Sc. degree in clinical epidemiology in 2000.

In 2005, she joined the National Poison Information Center (NVIC) of the National Institute of Public Health and the Environment (RIVM) (Head: Dr. Meulenbelt) where she works currently as researcher. She participated in 2005-2006 in a research project, with healthy volunteers, on the risks related to highly potent cannabis.

The author is married to Paul Mattijsen, project manager in real estate development. They have three children together –Juliette aged 12 years, Hélène aged 9 years, and Jean aged 6 years. The author is a member of the Amnesty International, World Wild Foundation (WWF) and International Plan associations. She is also treasurer of the French association '*Association Statistique et Santé Publique*' (ASSP). Practising Aïkido, reading English language literature, and playing piano are her current favourite hobbies.