Measuring quality of care

methods and applications to acute neurological diseases

This thesis was printed on FSC certified paper

Financial support by the Netherlands Heart Foundation for the publication of this thesis is gratefully acknowledged
Publication of this thesis was financially supported by the Department of Public Health, Erasmus MC and the Erasmus University

Cover design & lay-out: Luci van Engelen (CVIII Ontwerpers)
Printed by: Printforce, Alphen a/d Rijn

ISBN: 978-90-77283-11-0

Measuring quality of care

methods and applications to acute neurological diseases

Meten van kwaliteit van zorg
methoden en toepassingen op acute neurologische aandoeningen

PROEFSCHRIFT

Ter verkrijging van de graad van doctor aan de
Erasmus Universiteit Rottterdam
op gezag van de
rector magnificus

Prof. dr. H.G. Schmidt

en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
Vrijdag 26 november 2010 om 11.30 uur
door

Hester Floor Lingsma

geboren te Amstelveen.

Promotiecommissie

Promotor	Prof. dr. E.W. Steyerberg
Overige leden	Prof. dr. ir. H. Boersma
	Dr. D.W.J. Dippel
	Prof. dr. A.I.R. Maas

Contents

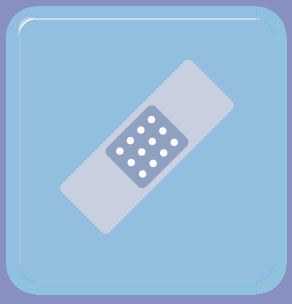


I Introduction









1 General introduction

General introduction

Over the past 20 years, quality of care has become a major topic in health care. Only two decades ago, physicians could be confident that they alone had a social mandate to judge and manage the quality of care.[1] In contrast, in the current era of evidence-based medicine, medical practice is continuously critically evaluated by different stakeholders.

Doctors and hospitals review their own practice with the aim to improve quality of care. They use quality of care information internally, e.g. by internal audits. Or externally, e.g. by comparing different health care providers and learning from best practices. Other stakeholders mostly use quality of care information externally. Governments monitor quality of care to ensure good quality health care. Health care financers try to distinguish good from poor performance to offer good care to their insured. Also patients (or 'consumers') try to compare hospitals and search for the best hospital for their health problem.

All these attempts to evaluate quality of care require that quality of care can actually be measured. This poses numerous difficulties. First there is no uniform definition of quality of care. Experts have struggled for decades to formulate a concise, meaningful, and generally applicable definition of the quality of health care. One of the most widely cited recent definitions, formulated by the Institute of Medicine in 1990[2, 3] holds that quality consists of the 'degree to which health services for individuals and populations increase the likelihood of desired health outcomes and are consistent with current professional knowledge.' The complexity and variability of these and many other definitions of quality is confusing and shows that formulations are dependent on where we are located in the system of care.[4] Different perspectives on and definitions of quality will logically call for different approaches to its measurement and management.

Moreover health care is very complex and many factors determine the outcome of a patient. The care of one specific health care provider is only one of these factors. So far there has been no generally accepted approach or method to measure quality of health care.

Measuring quality of care

Research on quality of care measurement started in the United States of America in the 1970s, followed by Europe, especially the United Kingdom. The current paradigm in quality of care research was established by Donabedian in 1988.[4] He stated that quality of care comprises structure, process and outcome.

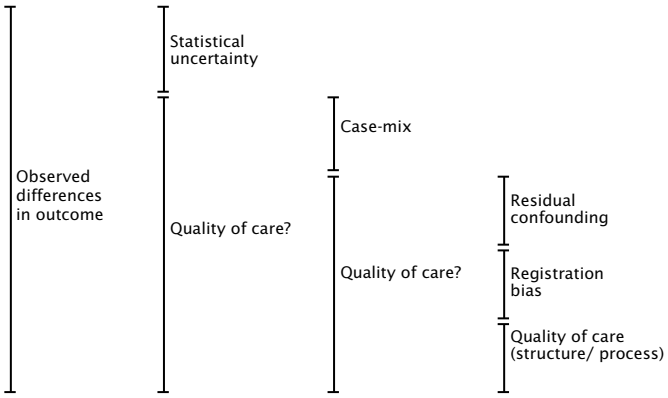
Structure relates to organisation of care, such as number of beds in a hospital. Process relates to actual actions of care, such as whether the patient receives medication within a certain time frame. Outcome includes patient outcome measures such as mortality. Part of the discussion on quality of care measurement has focused on which

of these components of quality to measure. Outcome measures are often used, since they are most relevant to the final aim of measuring quality of care; improvement of patient outcomes, including mortality, morbidity, and poor health. If an outcome measure represents quality of care, it would be expected to be related to relevant process measures.

Outcome measures

Outcome measures for quality of care are surrounded by methodological problems.[5] The two most important challenges are dealing with statistical uncertainty and with differences in case-mix (Figure 1.1). Case-mix is the type or mix of patients treated by a hospital.[6]

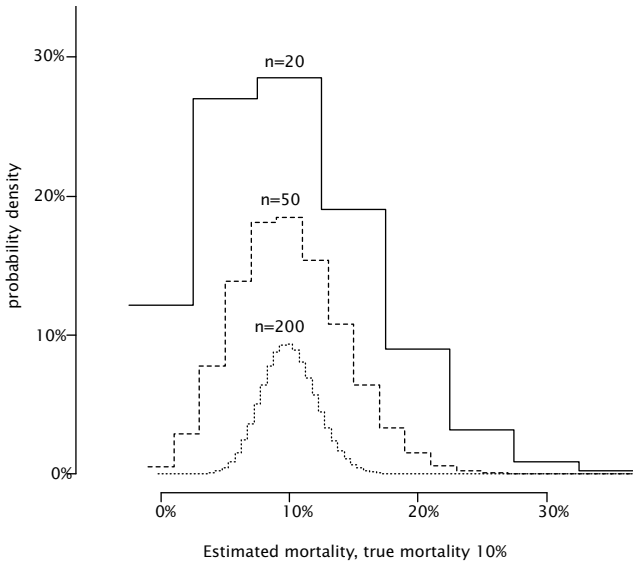
Figure 1.1 Possible sources of observed between-hospital differences in outcome



Statistical uncertainty

When using outcome measures, there will always be some variation in outcome between hospitals, caused just by chance. E.g. when the mortality rate in a hospital with 20 patients is 10%, we expect 2 deaths. Due to pure chance however also 0 or 6 or even more deaths can be observed. The amount of statistical uncertainty is dependent on the number of patients treated in a hospital. When the number of patients increases to 50 or to 200, the role of statistical uncertainty decreases (Figure 1.2).
If statistical uncertainty is ignored, this may lead to overinterpretation of differences in outcome between hospitals.

Figure 1.2 Estimated mortality in relation to number of patients per hospital (n)



A specific form of comparing hospitals is ranking them according to their quality of care. This was already done in 1995 for physician-specific mortality after coronary-artery bypass grafting surgery in New York State.[7] Ranking has the problem that one hospital has to be first and one has to be last, even if the differences are small and the statistical uncertainty is large. Rankings can hence be misleading. Nevertheless rankings are very popular in the press.[8]

Case-mix adjustment

The largest criticism on outcome indicators is that they may more reflect a hospitals' patient population ('case-mix') than quality of care. When hospitals have a different patient population in terms of e.g. age and disease severity, the mortality rates will differ regardless of quality of care.

As an example we compared the case-mix of stroke patients of two Dutch hospitals that participated in the same study (Table 1.1). Since older age, severe stroke and lowered consciousness level are strong predictors of mortality, hospital A is expected to have a higher mortality rate. Thus, ignoring the differences in case-mix will lead to an unfair comparison of the outcomes of the two hospitals.

Table 1.1 Stroke population of two Dutch hospitals

	Hospital A	Hospital B
Mean age (years)	77	65
Severe stroke (%)	28	8
Lowered consciousness at hospital arrival (%)	21	4

To account for patient characteristics that will influence outcome a prognostic model can be used. Prognostic models combine a number of patient characteristics to predict the outcome of interest, most often with regression models.[9] Instead of unadjusted, crude comparisons between hospitals, adjusted outcomes estimated with a prognostic model are preferable.

For appropriate case-mix adjustment it is essential that the prognostic models used for case-mix adjustment consider all relevant prognostic factors that may differ between the hospitals.

Prognostic models have numerous applications besides case-mix adjustment, with direct and indirect applications for quality of care. In this thesis some of these are discussed including the use of models to target treatment, and to adjust for patient characteristics for the estimation of a treatment effect.

The differences between hospitals in outcome that remain after taking into account statistical uncertainty and after case-mix adjustment may be caused by registration bias, residual confounding or real differences in quality of care. The latter can be reduced by quality of care improvement.

Between-hospital differences in outcome are not only of interest from the perspective of quality improvement. Another relevant area is study design. Currently most randomized trials are multi-centre trials, and are conducted in multiple countries. The presence of differences in outcome between the centres may influence the chances of demonstrating a treatment effect in RCTs.

Acute neurological diseases

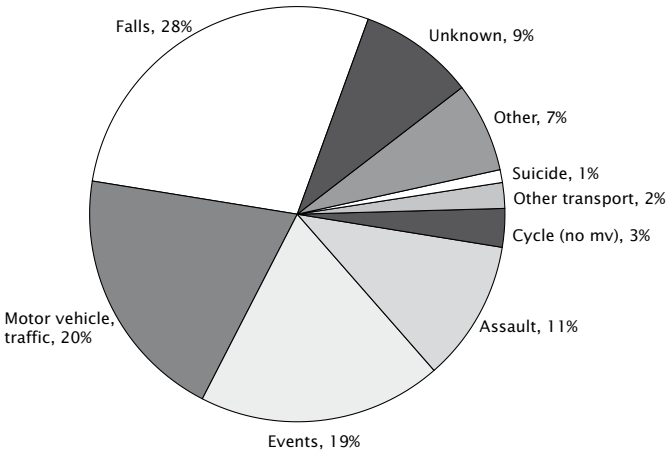
The methods for measuring quality of care that will be studied in this thesis are applied to acute neurological diseases, including traumatic brain injury, stroke, Guilllain-Barré syndrome and subarachnoid haemorrhage.

Traumatic brain injury

Traumatic brain injury (TBI) is an important public health problem worldwide.[10, 11] It is one of the most important causes of death and disability among young adults in the Western world. TBI is generally defined as an injury to the brain induced by external force.

Falls and motor vehicle traffic incidents are the leading cause of TBI. Falls are more prevalent in the Western world, traffic incidents are more prevalent in the developing world. These are followed by events (including sports and recreational injuries) and assaults (Figure 1.3).[12] The age groups at highest risk for TBI are children and young adults. Adults older than 75 years of age have the highest rates of hospitalization and death.

Figure 1.3 Causes of TBI



TBI can be classified according to different aspects, such as injury mechanism, clinical severity, or by assessment of structural abnormalities.[11] Mechanistically, TBI is classically classified as closed, penetrating, crash or blast injury. For classification by clinical severity, the level of consciousness is graded by the Glasgow Coma Scale, using three parameters: eye opening, motor response and verbal response.[13] According to the total GCS, TBI patients are subdivided in three severity classes: mild (GCS 13-15), moderate (GCS 9-12) and severe TBI (GCS < 8). Structural abnormalities can be identified by different imaging techniques, such as CT and MRI scanning. Various classification systems are available for CT abnormalities, of which the Marshall CT classification is most widely used.[14]

The most common outcome measure in TBI is the Glasgow Outcome Scale (GOS), although other scales are available. The GOS is an ordinal 5-point scale, which assesses the overall outcome after TBI (Table 1.1).[15] Often, the scale is dichotomized in favourable versus unfavourable. Final outcome is usually determined at six months after TBI, when clinical recovery is more or less stabilized.

Table 1.1 Glasgow Outcome Scale

Category	Label	Definition
1	Dead	Mortality from any cause
2	Vegetative	Unable to interact with environment, unresponsive
3	Severe Disability	Conscious but dependent
4	Moderate Disability	Independent but disabled
5	Good Recovery	Return to normal occupation and social activities, may have minor residual deficits

Stroke

Stroke ranks second as a cause of death worldwide and is the main cause of disability in high-income countries. In the Netherlands alone, more than 37.000 patients are admitted to hospital for acute stroke each year.[16]

Strokes are either ischemic or hemorrhagic. Ischemic stroke accounts for about 80% of all strokes and results from a transient or permanent reduction of cerebral blood flow caused by occlusion of a cerebral artery or arteriole. The most common causes are atherothrombosis and embolism from the heart.[17]

Different treatment options exist for stroke. Stroke unit care has been proven effective for all stroke patients.[18] In patients with ischemic stroke, treatment with recombinant tissue-plasminogen activator reduces the number of patients with poor outcome at three months by about 9%[19, 20], but the short time window for administration (4.5 hours) and the associated bleeding risk restrict treatment with recombinant tissue-plasminogen activator to a minority of patients. Aspirin, started within 48 hours of symptom onset, is probably effective across the entire range of patients with ischemic stroke, but the benefit is small.[21]

Most phase III stroke trials have used the degree of dependency or death as their main outcome measures. The most commonly used outcome scale for assessing dependency is the modified Rankin Scale[22] This scale quantifies dependency using an ordinal hierarchical grading from 0 (no symptoms) to 5 (severe disability) Sometimes 6 (death) is added to facilitate statistical analysis and interpretation (Table 1.2).

Table 1.2 Modified Rankin Scale

Category	Definition
0	No remaining symptoms
1	No significant disability despite symptoms; able to carry out all usual activities
2	Slight disability; unable to carry out all previous activities, but independent
3	Moderate disability; requiring some help, but able to walk without assistance
4	Moderately severe disability; unable to walk without assistance and unable to attend to own bodily needs without assistance
5	Severe disability; bedridden, incontinent and requiring constant nursing care and attention
6	Dead

Guillain-Barré syndrome

The Guillain-Barré syndrome (GBS) is the most common cause of acute neuro-muscular paralysis in the western world. The incidence is 1.2-2.3 per 100.000 per year.[23, 24] GBS is a post-infectious disorder which occurs in otherwise healthy people. People of all ages can be affected, but incidence increases linearly with age.[23-25]

GBS is characterized by a rapidly progressive bilateral weakness of the extremities, sensory deficits and tendon reflex loss. Cranial nerves and respiratory muscles can also be affected and 20-30% of the patients need mechanical ventilation. GBS is a very heterogeneous disease regarding clinical severity and outcome. Some patients develop mild limb paresis, whereas others develop oculomotor, bulbar, respiratory muscle and limb paralysis and remain bedbound for several months.

Intravenous immunoglobulin and plasma exchange are shown to be effective in patients with GBS.[23, 26, 27] Nowadays, intravenous immunoglobulin (2 g/kg in 2-5 days) has become standard treatment for patients with GBS who are unable to walk unaided and still within the first 2 weeks from onset of weakness [26, 28 29]

The outcome of GBS after 6 or 12 months however has only marginally been improved.[23, 30] Approximately 20% are still disabled after 6 months and a serious long-term impact on the patients' work and private life and that of their partners has been shown.[31]

Subarachnoid haemorrhage

Subarachnoid haemorrhage (SAH) is bleeding into the subarachnoid space – the area between the arachnoid membrane and the pia mater surrounding the brain. SAH originates from arteries localised on the brain's surface. Aneurysmal subarachnoid haemorrhage (aSAH) is an haemorrhage which is caused by rupture of an intracranial aneurysm.[32]

The overall incidence of aSAH is between 5 and 10 per 100,000 person years. The incidence increases with age; and from midlife onwards incidence is higher in women than

in men.[33] The reasons for this higher incidence in women are not clear, but hormonal factors (including hormonal medication) have been suggested as a possible explanation.[34, 35]
Sudden headache is the most characteristic symptom of aSAH; in approximately 75 percent of patients, the onset is within seconds.[36] Often, the headache is accompanied by nausea and vomiting. On admission two-thirds of all patients have depressed consciousness, of whom half are in coma.[37] The patient might regain alertness and orientation or might remain with various degrees of lethargy, confusion, or agitation.
Case-fatality after aSAH has been estimated between 20 and 50 percent.[38, 39] The percentage of persons dying before they reach a hospital has been estimated between 10 and 15 percent.[40, 41] Those who survive the first episode of aSAH, are at risk of rebleeds, delayed ischemia or hydrocephalus. After discharge, approximately 5% of the patients develops epilepsy.[42, 43] Cognitive deficits and psychosocial dysfunction in the first year after SAH are common, even in patients who make a good recovery in terms of self care. [44, 45, 46] Although improvement can be expected up to one and a half year after aSAH, many former patients and their partners experience deficits and reduced quality of life 1 to 2 years after SAH.[47]

Aims and contents

The aim of this thesis is to study methods to measure quality of care with outcome measures, and to apply these methods to different acute neurological diseases.
Specific questions include:

1. What is the role of statistical uncertainty in measuring quality of care with outcome measures?

1a. How large is the effect of statistical uncertainty on between-hospital comparisons?
1b. How should statistical uncertainty be incorporated in outcome measures?

2. What is the role of case-mix variation in measuring quality of care with outcome measures?

2a. How large is the effect of case-mix on between-hospital comparisons?
2b. How can case-mix variation be captured for between-hospital comparisons?

3. How do outcome measures relate to processes of care?

The next parts of this thesis consist of studies on the main methodological issues related to quality of care measurement; statistical uncertainty and case-mix adjustment. In part II, Chapter 2 shows what the effect of statistical uncertainty could be and how it can be incorporated in rankings. Chapter 3 shows how much statistical uncertainty is present in outcome measures that are currently used by the Dutch Healthcare Inspectorate to assess the quality of hospital care. In chapter 4 different statistical software packages that could be used to take into account statistical uncertainty are compared.

Part III is about the development of prognostic models that can be used for case-mix adjustment. Chapter 5 gives an overview of the prognostic models available to predict outcome in TBI patients and some general considerations about prognostic model development. Chapter 6 is about potential improvement of adjustment models in TBI by including extracranial injury as a predictor. Chapter 7 presents a prognostic model that can be used to identify GBS patients that will require artificial ventilation. In chapter 8 a prognostic model to predict outcome in GBS patients is developed. In chapter 9 a prognostic model to predict outcome in aSAH patients is developed.

In the fourth part the methods and models presented in part II and III are applied to TBI and stroke. Chapter 10 studies between-hospital differences in outcome in TBI. Chapter 11 investigates whether these between-hospital differences affect the estimation of the treatment effect in clinical trials.

In chapter 12 between-hospital differences in outcome after stroke are studied and it is assessed whether these differences are more related to patient characteristics or to process measures. Chapter 13 assesses the relation between a process measure in stroke, treatment with statins, and outcome.

The results of the studies in this thesis are further discussed in chapter 14, together with their implications.

References

1. Blumenthal D. Part 1: Quality of care–what is it? *N Engl J Med* 1996; 335(12): 891-4.
2. Lohr KN. Applications of health status assessment measures in clinical practice. Overview of the third conference on advances in health status assessment. *Med Care* 1992; 30(5 Suppl): MS1-14.
3. Lohr KN, Donaldson MS, Harris-Wehling J. Medicare: a strategy for quality assurance, V: Quality of care in a changing health care environment. *QRB Qual Rev Bull* 1992; 18(4): 120-6.
4. Donabedian A. The quality of care. How can it be assessed? *JAMA* 1988; 260(12): 1743-8.
5. Mant J. Process versus outcome indicators in the assessment of quality of health care. *Int J Qual Health Care* 2001; 13(6): 475-80.
6. Iezzoni LI, editor. *Risk Adjustment for Measuring Healthcare Outcomes*. 3th ed. Chigago: AcademyHealth/HAP Book; 2003.
7. Green J, Wintfeld N. Report cards on cardiac surgeons. Assessing New York State's approach. *N Engl J Med* 1995; 332(18): 1229-32.
8. Wang W, Dillon B, Bouamra O. An analysis of hospital trauma care performance evaluation. *J Trauma* 2007; 62(5): 1215-22.
9. Steyerberg EW. *Clinical Prediction Models. A Practical Approach to Developement, Validating and Updating*. New York: Springer; 2009.
10. Ghajar J. Traumatic brain injury. *Lancet* 2000; 356(9233): 923-9.
11. Maas AI, Stocchetti N, Bullock R. Moderate and severe traumatic brain injury in adults. *Lancet Neurol* 2008; 7(8): 728-41.
12. Langlois JA, Rutland-Brown W, Thomas KE. *Traumatic Brain Injury in the United States: emergency department visits, hospitalizations, and deaths*. 2006; .
13. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974; 2(7872): 81-4.
14. Marshall LF, Bowers S, Klauber MR. A new classification of head injury based on computerised tomography. *J Neurosurg* 1991; 75(1): S14-20.
15. Jennett B, Bond M. Assessment of outcome after severe brain damage. *Lancet* 1975; 1(7905): 480-4.
16. Vaartjes I. *Hart- en vaatziekten in Nederland* 2008. Nederlandse Hartstichting; 2008 .
17. Bousser MG, Amarenco P, Chamorro A, Fisher M, Ford I, Fox K, et al. Rationale and design of a randomized, double-blind, parallel-group study of terutroban 30 mg/day versus aspirin 100 mg/day in stroke patients: the prevention of cerebrovascular and cardiovascular events of ischemic origin with terutroban in patients with a history of ischemic stroke or transient ischemic attack (PERFORM) study. *Cerebrovasc Dis* 2009; 27(5): 509-18.
18. Govan L, Weir CJ, Langhorne P, for the Stroke Unit Trialists' Collaboration. Organized Inpatient (Stroke Unit) Care for Stroke. *Stroke* 2008; .
19. Hacke W, Kaste M, Bluhmki E, Brozman M, Davalos A, Guidetti D, et al. Thrombolysis with alteplase 3 to 4.5 hours after acute ischemic stroke. *N Engl J Med* 2008; 359(13): 1317-29.
20. Wardlaw JM, Zoppo G, Yamaguchi T, Berge E. Thrombolysis for acute ischaemic stroke. *Cochrane Database Syst Rev* 2003; (3)(3): CD000213.

21. Sandercock PA, Counsell C, Gubitz GJ, Tseng MC. Antiplatelet therapy for acute ischaemic stroke. *Cochrane Database Syst Rev* 2008; (3)(3): CD000029.
22. van Swieten JC, Koudstaal PJ, Visser MC, Schouten HJ, van Gijn J. Interobserver agreement for the assessment of handicap in stroke patients. *Stroke* 1988; 19(5): 604-7.
23. Hughes RA, Cornblath DR. Guillain-Barre syndrome. Lancet 2005; 366(9497): 1653-66.
24. van Doorn PA, Ruts L, Jacobs BC. Clinical features, pathogenesis, and treatment of Guillain-Barre syndrome. *Lancet Neurol* 2008; 7(10): 939-50.
25. Van Koningsveld R, Van Doorn PA, Schmitz PI, Ang CW, Van der Meche FG. Mild forms of Guillain-Barre syndrome in an epidemiologic survey in The Netherlands. *Neurology* 2000; 54(3): 620-5.
26. van der Meche FG, Schmitz PI. A randomized trial comparing intravenous immune globulin and plasma exchange in Guillain-Barre syndrome. Dutch Guillain-Barre Study Group. *N Engl J Med* 1992; 326(17): 1123-9.
27. Plasmapheresis and acute Guillain-Barre syndrome. The Guillain-Barre syndrome Study Group. *Neurology* 1985; 35(8): 1096-104.
28. Randomised trial of plasma exchange, intravenous immunoglobulin, and combined treatments in Guillain-Barre syndrome. Plasma Exchange/Sandoglobulin Guillain-Barre Syndrome Trial Group. *Lancet* 1997; 349(9047): 225-30.
29. van Koningsveld R, Schmitz PI, Meche FG, Visser LH, Meulstee J, van Doorn PA, et al. Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barre syndrome: randomised trial. *Lancet* 2004; 363(9404): 192-6.
30. Hughes RA, Swan AV, Raphael JC, Annane D, van Koningsveld R, van Doorn PA. Immunotherapy for Guillain-Barre syndrome: a systematic review. *Brain* 2007; 130(Pt 9): 2245-57.
31. Bernsen RA, de Jager AE, Schmitz PI, van der Meche FG. Long-term impact on work and private life after Guillain-Barre syndrome. *J Neurol Sci* 2002; 201(1-2): 13-7.
32. van Gijn J, Kerr RS, Rinkel GJ. Subarachnoid haemorrhage. *Lancet* 2007; 369(9558): 306-18.
33. de Rooij NK, Linn FH, van der Plas JA, Algra A, Rinkel GJ. Incidence of subarachnoid haemorrhage: a systematic review with emphasis on region, age, gender and time trends. *J Neurol Neurosurg Psychiatry* 2007; 78(12): 1365-72.
34. Longstreth WT, Nelson LM, Koepsell TD, van Belle G. Subarachnoid hemorrhage and hormonal factors in women. A population-based case-control study. *Ann Intern Med* 1994; 121(3): 168-73.
35. Mhurchu CN, Anderson C, Jamrozik K, Hankey G, Dunbabin D, Australasian Cooperative Research on Subarachnoid Hemorrhage Study (ACROSS) Group. Hormonal factors and risk of aneurysmal subarachnoid hemorrhage: an international population-based, case-control study. *Stroke* 2001; 32(3): 606-12.
36. Linn FH, Rinkel GJ, Algra A, van Gijn J. Headache characteristics in subarachnoid haemorrhage and benign thunderclap headache. *J Neurol Neurosurg Psychiatry* 1998; 65(5): 791-3.
37. Brilstra EH, Rinkel GJ, Algra A, van Gijn J. Rebleeding, secondary ischemia, and timing of operation in patients with subarachnoid hemorrhage. *Neurology* 2000; 55(11): 1656-60.
38. Benatru I, Rouaud O, Durier J, Contegal F, Couvreur G, Bejot Y, et al. Stable stroke incidence rates but improved case-fatality in Dijon, France, from 1985 to 2004. *Stroke* 2006; 37(7): 1674-9.

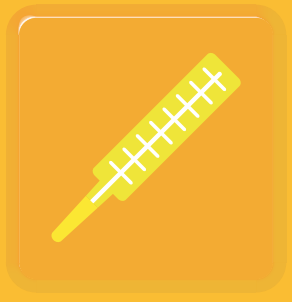
39. Hop JW, Rinkel GJ, Algra A, van Gijn J. Case-fatality rates and functional outcome after subarachnoid hemorrhage: a systematic review. *Stroke* 1997; 28(3): 660-4.
40. Huang J, van Gelder JM. The probability of sudden death from rupture of intracranial aneurysms: a meta-analysis. *Neurosurgery* 2002; 51(5): 1101,5; discussion 1105-7.
41. Koffijberg H, Buskens E, Granath F, Adami J, Ekbom A, Rinkel GJ, et al. Subarachnoid haemorrhage in Sweden 1987-2002: regional incidence and case fatality rates. *J Neurol Neurosurg Psychiatry* 2008; 79(3): 294-9.
42. Buczacki SJ, Kirkpatrick PJ, Seeley HM, Hutchinson PJ. Late epilepsy following open surgery for aneurysmal subarachnoid haemorrhage. *J Neurol Neurosurg Psychiatry* 2004; 75(11): 1620-2.
43. Claassen J, Peery S, Kreiter KT, Hirsch LJ, Du EY, Connolly ES, et al. Predictors and clinical impact of epilepsy after subarachnoid hemorrhage. *Neurology* 2003; 60(2): 208-14.
44. Hackett ML, Anderson CS. Health outcomes 1 year after subarachnoid hemorrhage: An international population-based study. The Australian Cooperative Research on Subarachnoid Hemorrhage Study Group. *Neurology* 2000; 55(5): 658-62.
45. Mayer SA, Kreiter KT, Copeland D, Bernardini GL, Bates JE, Peery S, et al. Global and domain-specific cognitive impairment and outcome after subarachnoid hemorrhage. *Neurology* 2002; 59(11): 1750-8.
46. Powell J, Kitchen N, Heslin J, Greenwood R. Psychosocial outcomes at three and nine months after good neurological recovery from aneurysmal subarachnoid haemorrhage: predictors and prognosis. *J Neurol Neurosurg Psychiatry* 2002; 72(6): 772-81.
47. Powell J, Kitchen N, Heslin J, Greenwood R. Psychosocial outcomes at 18 months after good neurological recovery from aneurysmal subarachnoid haemorrhage. *J Neurol Neurosurg Psychiatry* 2004; 75(8): 1119-24.



II Statistical uncertainty






2 Comparing and ranking hospitals based on outcome after stroke

Lingsma HF, Steyerberg EW, Eijkemans MJC, Dippel DWJ, Scholte op Reimer WJM, van Houwelingen HC, and The Netherlands Stroke Survey investigators. Comparing and ranking hospitals based on outcome: results from The Netherlands Stroke Survey. *Q J Medcine* 2010; 103:99–108

Abstract

Background
Measuring quality of care and ranking hospitals with outcome measures poses two major methodological challenges: case-mix adjustment and variation that exists by chance.

Aim
To compare methods for comparing and ranking hospitals that take these into account.

Methods
The Netherlands Stroke Survey was conducted in 10 hospitals in the Netherlands, between October 2002 and May 2003, with prospective and consecutive enrolment of patients with acute brain ischemia. Poor outcome was defined as death or disability after 1 year (modified Rankin scale 3 or higher). We calculated fixed and random hospital effects on poor outcome, unadjusted and adjusted for patient characteristics. We compared the hospitals using the expected rank, a novel statistical measure incorporating the magnitude and the uncertainty of differences in outcome.

Results
At 1 year after stroke, 268 of the total 505 patients (53%) had a poor outcome. There where substantial differences in outcome between hospitals in unadjusted analysis (χ^2 statistic=48, 9 df, $p<0.0001$). Adjustment for 12 confounders led to halving of the χ^2 ($\chi^2 = 24$). The same pattern was observed in random effects analysis. Estimated performance of individual hospitals changed considerably between unadjusted and adjusted analysis. Further changes were seen with random effect estimation, especially for smaller hospitals. Ordering by Expected Rank led to shrinkage of the original ranks of 1 to 10 towards the median rank of 5.5 and to a different order of the hospitals, compared to ranking based on fixed effects.

Conclusion
In comparing and ranking hospitals, case-mix adjusted random effect estimates and the Expected Ranks are more robust alternatives to traditional fixed effect estimates and simple rankings.

Introduction

Measuring quality of care receives increasing attention. Specifically, ranking of hospitals may be attempted to compare their quality of care. Such ranking on outcome was already done in 1995 for physician-specific mortality after coronary-artery bypass grafting surgery in New York State.[1] Ranking is currently very popular, especially in the lay press.[2]

Measuring quality of care and ranking hospitals has the potential to enable health care financers to identify poor performance. Also patients (or 'consumers') might choose the best hospital for their health problem, and hospitals may learn from best practices. All these applications can have huge consequences for hospitals on for example their budget and reputation, which makes reliability of results extremely important.

Quality of care is often measured with outcomes such as mortality, an approach that is surrounded by many methodological problems.[3] The first major issue is case-mix adjustment.[4] Case-mix adjustment should appropriately capture differences between hospitals in patient characteristics that are outside the influence of actions in the hospital.

The second issue is drawing proper conclusions from the hospital-specific case-mix adjusted outcomes. There will always be some variation in outcome between hospitals, caused just by chance. Disregarding this chance variation may lead to over-interpretation of differences between hospitals since especially smaller hospitals can have an extreme outcome, caused more by chance than by their underlying quality.

Variation between hospitals in binary outcomes is traditionally modelled as fixed effects in a logistic regression model. We can also use a random effect logistic regression model which accounts for variation by chance at the hospital level.[3, 5-10]

Ranking hospitals according to their outcome causes the problem that one hospital has to be first and one has to be last. Simple ranking disregards both the magnitude of the relative differences between the hospitals and the variation that exists by chance, and can hence put hospitals in needless jeopardy.

In this study we use data from the Netherlands Stroke Survey to compare methods for assessment of quality of care that takes into account case-mix and variation by chance.

Methods

The Netherlands Stroke Survey

The Netherlands Stroke survey was conducted in 10 hospitals in The Netherlands: 2 in the north, 4 in the middle, and 4 in the southern regions. The participating hospitals comprised 1 small (<400 beds), 4 intermediate (400 to 800 beds) and 5 large hospitals (>800 beds). Two hospitals were university hospitals.

All patients who were admitted to the neurology department with suspected acute brain ischemia between October 2002 and May 2003 were screened. Patients were enrolled consecutively and prospectively if the initial diagnosis of first or recurrent acute brain ischemia was confirmed by the neurologist's assessment. Trained research assistants collected data from the patients' hospital charts, within 5 days after discharge. At 1 year, survival status was obtained through the Civil Registries. A telephone interview was conducted, based on a structured questionnaire which was sent in advance. Follow up was complete in 96% of the patients. More details on the study population and methods of data collection were reported previously.[11, 12]

Case-mix adjustment

The primary outcome was whether patients were dead or disabled at 1 year after admission, i.e. a score on the modified Rankin scale of 3 or higher. We used a logistic regression model to adjust for case-mix, because we consider case-mix as a confounders since it may be related to the setting and to the outcome and is outside the influence of actions in the hospital. The model we used included 12 patient characteristics: age, sex, stroke subtype (Transient Ischemic Attack (TIA) or ischemic stroke), stroke severity, lowered consciousness level at hospital arrival, Barthel Index 24 hours from admission, previous stroke, atrial fibrillation, ischemic heart disease, diabetes mellitus, hypertension and hyperlipidemia. These variables were selected in previous work on the same dataset with stepwise logistic regression analysis with backward elimination of possible confounders with the Akaike Information Criterion (AIC) for inclusion (equivalent to $P<0.157$ for confounders with 1 degree of freedom).[13] In the first step age, sex and stroke subtype were entered, in the second step the other patient characteristics were added. The model is described in more detail elsewhere.[12]

Hospital effects

We estimated the variation between the hospitals with two different models. The first was a standard fixed effect logistic regression model, with hospital as a categorical variable. We estimated the coefficient for each hospital, compared to the average using an offset variable. We also calculated the χ^2 for the model as the difference in −2 log likelihood for a model with and without hospital, to indicate the total variation between the hospitals. Both the individual coefficients and the variation were calculated with

and without adjustment for case-mix. We refer to the results of the fixed effect models as fixed effect estimates.
Since the fixed effect estimates do not account for variation by chance, we also fitted a random effect logistic regression model. Random effect models account for the fact that part of the variation between hospitals is just chance. They estimate the hospital effects and the total variation 'beyond chance'. This total variation is indicated by the model parameter τ^2. We refer to the results of the random effect models as random effect estimates and these were also fitted with or without adjustment for case mix.

Ranking and rankability
To also account for the variation by chance in rankings, we calculated the expected rank (ER), this is the probability that the performance of a hospital is worse than another randomly selected hospital. The ER incorporates both the magnitude and the uncertainty of the difference of a particular hospital with other hospitals. We can scale the expected ranks ER between 0 and 100% with percentiles based on expected rank (PCER) for easy interpretation and to make the ranks independent of the number of hospitals. The PCER can be interpreted as the probability (as a percentage) that a hospital is worse than a randomly selected hospital, including itself.

To see whether it makes sense to rank the hospitals, we calculated the 'rankability'. The rankability relates the total variation from the random effect models (How large are the differences between the hospitals?) to the uncertainty of the individual hospital differences from the fixed effect model (How certain are the differences?). The rankability can be interpreted as the part of variation between the hospitals that is not due to chance.

More details on the statistical analysis and formulas can be found in appendix 1 and in previous, more detailed work on this topic.[14]

The statistical analysis was performed with R (version 2.5, R foundation for statistical computing, Vienna). The random effect analysis was repeated in SAS (version 9, SAS Inc, Cary, NC) with compatible results. R programming code can be found in appendix 2.

Results

Study population

The study population consisted of 579 patients who were admitted to the hospital because of acute ischemic stroke or TIA. Of these, 505 patients (87%) with complete data on potential confounders and outcome were used in the analysis. The lowest numbers enrolled were 22 and 24 patients in hospitals 5 and 6, and the highest numbers 92 and 99 in hospitals 2 and 7 respectively (Table 2.1).

Table 2.1 Patient characteristics and poor outcome (modified Rankin Scale ≥3), and multivariable odds ratios of patient characteristics in the adjustment model on poor outcome.

Hospital	1	2	3	4	5	6	7	8	9	10	Total	OR (p-value)
N	39	92	31	40	22	24	99	36	50	70	505	
Mean age (years)	77	73	69	65	74	65	68	70	71	72	71	1.5 (<0.001)[4]
Male sex (%)	46	54	61	59	55	67	65	41	56	47	55	0.7 (0.092)
Stroke subtype (% stroke vs TIA)	97	95	97	80	91	63	94	92	88	81	90	1.1 (0.853)
Severe stroke[1] (%)	28	17	16	13	9	8	15	17	14	10	15	3.5 (<0.001)
Lowered consciousness level[2] (%)	21	21	10	15	18	4	2	17	12	11	13	3.5 (0.001)
ADL dependent[2,3] (%)	90	92	100	85	82	54	64	100	84	64	80	2.8 (0.001)
Previous stroke (%)	26	18	26	33	27	33	21	28	26	17	24	1.9 (0.012)
Atrial fibrillation (%)	23	21	16	15	27	8	17	11	22	14	18	1.7 (0.072)
Ischemic heart disease (%)	13	23	29	18	36	21	13	31	26	21	21	2.0 (0.012)
Diabetes mellitus (%)	15	30	23	20	9	21	17	17	16	20	20	2.1 (0.008)
Hypertension (%)	56	47	58	58	59	75	81	50	58	50	59	0.6 (0.018)
Hyperlipedimia (%)	54	46	68	53	73	50	58	44	70	79	59	0.6 (0.040)
Poor outcome (%)	59	72	35	44	73	29	39	78	54	46	53	

1. Paresis of arm, leg and face, homonymous hemianopia and aphasia or other cortical function disorder.
2. At hospital arrival.
3. Barthel Index=20.
4. Odds ratio per decade.

ADL, activities of daily living; TIA, transient ischemic attack

Mean age was 71 (sd=13 years), 278 patients (55%) were male, and the majority (450, 90%) was diagnosed with cerebral infarction (Table 2.1).

At 1 year, 143 patients (28%) had died and 125 of the remaining 362 patients (35%) were disabled (modified Rankin scale score 3, 4 or 5). Thus, the total number of patients with poor outcome at 1 year after stroke was 268 (53%). This percentage varied substantially between hospitals, from 29% poor outcome in hospital 6 to 78% poor outcome in hospital 8 (Table 2.1).

Case-mix adjustment

The strongest predictors of poor outcome were indicators of stroke severity (severe stroke: OR=3.5, p=<0.001; lowered consciousness level: OR=3.5, p=0.001; Activities of Daily Living (ADL) dependency: OR=2.8, p=0.001) and age (OR=1.5 per decade, p=0.001) (Table 2.1).

The Area Under the Curve (AUC) of the total model was 0.804. Sex and stroke subtype were not significant anymore after adding all the confounders in the second step of the model development.

Although the differences in outcome between hospitals were highly significant in unadjusted fixed effects analysis (χ^2 = 48, 9 df, p<0.0001, Table 2.2), they were partly explained by confounders. For example, hospital 2, 5, and 8 had over 70% poor outcome, but mean ages of 73, 74, and 70 years. On the other hand, hospitals with mostly good outcomes had younger patients (e.g. hospital 6, mean age 65 years, 29% poor outcome, Table 2.1). Adjusting the fixed effect analysis for all 12 potential confounders led to halving of the χ^2 seen in unadjusted analysis (χ^2 = 24 instead of 48, Table 2.2). This pattern was also seen in the random effects analysis (τ^2=0.18 versus 0.38, Table 2.2).

Table 2.2 Heterogeneity between hospitals in fixed and random effect logistic regression analysis

	Fixed effect	**Random effect**
Unadjusted	χ^2 = 48, 9 df, p<0.0001	τ^2=0.38, χ^2=24, 1 df, p<0.0001
12 confounders	χ^2 = 24, 9 df, p=0.0042	τ^2=0.18, χ^2=4, 1 df, p=0.0275

χ^2: Difference on −2 log likelihood scale of model with and without hospital.
τ^2: Variance of the random effects on log odds scale

Estimation of differences between hospitals

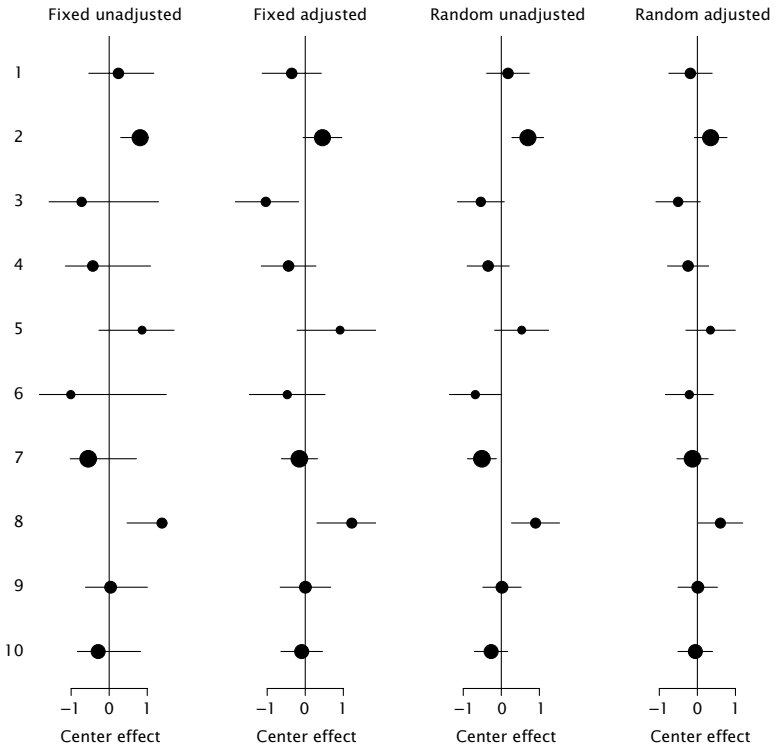
The apparent performance of the individual hospitals changed considerably between unadjusted and adjusted fixed analysis (Table 2.3, Figure 2.1).

Hospital 1 seemed to perform relatively poorly in unadjusted analysis (positive coefficient), while adjusted analysis indicated that the hospital performed relatively well (negative coefficient). This suggests that the positive coefficient was attributable to the unfavourable case-mix of the hospital. Changes for other hospitals were only quantitative, without change of sign, with adjusted differences generally closer to zero.

Table 2.3 Fixed and random effect estimates for differences between hospitals. Values are logistic regression coefficients, compared to the overall average outcome. A positive number means a higher probability on poor outcome.

Hospital	n	Fixed effect unadjusted	Random effect unadjusted	Fixed effect adjusted	Random effect adjusted
1	39	0.24	0.18	−0.36	−0.18
2	92	0.81	0.70	0.45	0.35
3	31	−0.72	−0.54	−1.04	−0.50
4	40	−0.43	−0.35	−0.44	−0.24
5	22	0.86	0.53	0.91	0.34
6	24	−1.01	−0.68	−0.47	−0.21
7	99	−0.55	−0.51	−0.15	−0.13
8	36	1.39	0.90	1.23	0.60
9	50	0.04	0.02	0.00	0.01
10	70	−0.29	−0.27	−0.09	−0.05

Figure 2.1 Differences between centers with unadjusted fixed effect estimates, unadjusted random effect estimates, adjusted fixed effects estimates and adjusted random effect estimates. A positive numbers means a higher probability on poor outcome. Dot size indicates sample size per center.

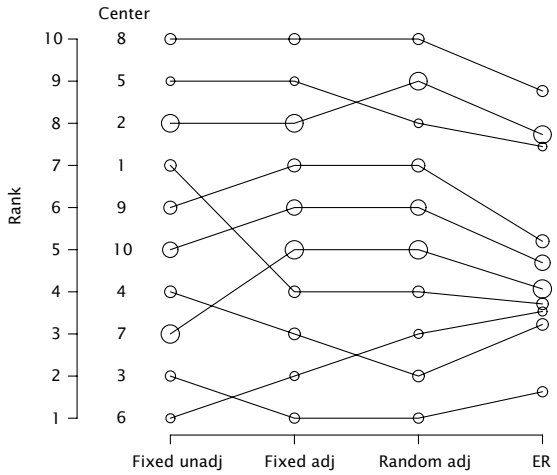


Further changes were seen after accounting for variation by chance with adjusted random effect models (Table 2.3, Figure 2.1). As expected random effect estimation did not affect estimates for the larger hospitals, such as 2 and 7. But for the smaller hospitals, such as hospitals 5, 6, and 8, the point estimates were shrunken considerably. None of the hospitals had a deviation significantly different from the average in the random effect model, but the overall heterogeneity was still statistically significant (Table 2.2).

Ranking and rankability

We first ranked hospitals based on unadjusted and adjusted fixed effect estimates and adjusted random effect estimates. Figure 2.2 shows that some hospitals such as 1, 3 and 4 change rank after adjustment for patient characteristics, and some small hospitals such as 5 and 6 again change rank after accounting for variation by chance with random effect estimation.

Figure 2.2 Ranks (left y axis) of 10 centers in fixed effect unadjusted, fixed effect adjusted, and random effects adjusted analyses, and Expected Rank. Dot size indicates sample size per center.



Subsequently, we calculated the Expected Rank (ER) and Percentile based on Expected Rank (PCER). Figure 2.2 shows that the ER led to shrinkage of the ranks towards the median rank of 5.5, with 6 hospitals having an ER close to this median. Hospital 6 seemed to do best with rank 1 in unadjusted analysis, shifted to rank 2 in adjusted analysis, to rank 3 in random effects analysis, and had an ER around 4, meaning that at most 4 out of 10 hospitals are expected to do better than this hospital.

With the PCER we can express the ERs on a 0 to 100% scale. Hospital 8 had a PCER of 86%, meaning there is a 86% probability that a randomly selected hospital does better than hospital 8. Hospital 3 had the best PCER (17%), meaning that there was only a 17% probability that a randomly selected hospital does better than hospital 3.

The rankability was 55%. This means that of the total variation between hospitals after adjustment, 55% was not due to chance.

Discussion

In this study we found large differences in the proportion of patients with poor outcome after stroke between hospitals. Adjusting for 12 potential confounders led to halving of the chi-square seen in unadjusted analysis, and considerable changes in performance estimates for individual hospitals. Further changes were seen after accounting for uncertainty in the random effect estimation, especially for smaller hospitals. Ordering the hospitals by means of the Expected Rank led to shrinkage of the simple ranks of 1 to 10 towards the median rank of 5.5 and to a different order of the hospitals.

A limitation of our study is we not able to do power calculations because we did not define a formal hypothesis on the difference between the hospitals. Our results should be considered as part of a larger debate on measuring quality of care. Measuring quality of care can have multiple purposes. A first broad distinction can be made between internal and external purposes. The first can for example be an internal quality system, or 'benchmarking', with the initiative at the side of the hospital. The second includes increasing accountability to governments, patients and insurance companies. These purposes are related, since a relatively poor performance might be an incentive for a hospital to stimulate improvements. Such feedback can lead to a continuous quality improvement. The results of this study apply more to external than to internal quality measurement.

If we want to compare hospitals, we can debate what to measure and how to measure it. In this study we focused on outcome (in this case the combination of mortality and disability), but quality of care measures may also include for example patient satisfaction, and organizational issues such as procedures and processes of delivering care.[15, 16] It is known that outcome is not always a valid indicator of quality of care.[12] Therefore some argue that we should concentrate on direct measurement of adherence to clinical and managerial standards.[6] Moreover, measuring adherence to guidelines provides clear directions for improvement of care in all hospitals, not only in those with poor outcome. Examples of such an approach in stroke are the 'Get With The Guidelines' program in the USA and the Scottish Stroke Care Audit.[17, 18] Those in favor of outcome assessment, however, advocate that quality assessment on process level requests a too detailed data collection, and conclusions on quality depend largely on the selection of process measures.[19]

In debates around measuring quality of care based with outcome the issue of case-mix adjustment has received substantial attention.[20, 21] Our study shows that this is indeed very important for stroke outcomes, since half of the differences, in terms of chi-square, between hospitals was explained by differences in case-mix. One hospital even seemed to perform poorly but appeared to perform well after adjustment for their unfavourable case-mix. We used a relatively simple model without any interaction terms for adjustment. In previous work we showed that age, sex and stroke subtype alone have only a moderate predictive strength (AUC: 0.690; AUC of total model: 0.804)[12] The choice for an adjustment model should be a trade-off between the performance of the model and available data. It was surprising that in our model hypertension and hyperlipidemia were protective for poor outcome (OR=0.6). Both were scored if noted in medical history or if diagnosed during hospitalization. Most patients with one of these conditions were already diagnosed before their stroke and thus treated with antihypertensive drugs or statins, this may cause the protective effect.

A second issue in comparing hospitals is variation that exists just by chance. If there are hospitals involved with small samples sizes, using fixed effect models that disregard the variation by chance could lead to exploding estimates of the hospital effects, and over-interpretation of the differences. Random effect models do account for variation by chance; they allow imprecisely estimated outcomes from small hospitals to 'borrow' information from other hospitals, causing their estimates to be shrunk toward the overall mean. Random effect models are thus more robust.[3, 5-10, 20] Our study shows that the random effect estimates are indeed more conservative. In random effect analyses, none of the hospitals had an effect that was significantly different from zero, while some had in the fixed effect analyses. The variation by chance had a large impact on the conclusions drawn about the hospitals. Individual hospitals are often too small to reliably determine whether they are an outlier. Small hospitals are more likely to suffer more from variation just by chance than large hospitals.[23]

We derived the random effect estimates directly from the fitted model since it is easily available now in statistical packages (such as R) and since we were able to reproduce our results with other fitting methods and with other software. The random effect estimates can also be calculated in two steps.[24]

Although random effect analyses are preferable for estimation of differences between hospitals, simple integer ranking based on these random effect estimates disregards uncertainty, and may lead to over-interpretation again. With the expected rank, uncertainty of the hospital effect estimates is also incorporated in the ranking. E.g., we found that 6 of the 10 hospitals were close to the median rank. ERs are a better representation of the random effect estimates. Approaches similar to the ER have been proposed by others.[4, 5, 25-27] For ease of interpretation we calculated the percentile based on expected rank, which is independent from the number of hospitals in the sample and indicates the probability that a hospital is worse than a randomly selected hospital, including itself.

A practical approach to ranking based on risk standardized mortality rates with 95% confidence intervals, estimated with hierarchical (random effects) modelling, was recently been proposed by Krumholz et al.[28] However, interpreting the magnitude and clinical significance of differences between hospitals with overlapping 95% confidence intervals is difficult.[29] The PCER consists of only one number. Although there is an almost universal agreement that a confidence interval is more informative than just an estimate, we believe that for lay people (patients who want to choose between hospitals) it is easier to interpret one number compared to an estimate and its confidence interval. On the other hand, the PCER does not show directly the degree of variation by chance, although it is included in the calculation of the single number. In our perspective the PCER approach could be a useful extension to reporting of provider performance, since it combines the attractiveness of a ranking, provides a single number and is easy to interpret.

Some guidelines have recently been published with respect to statistical methods for public reporting of health outcomes, which suggest 7 preferred attributes of statistical modelling for provider profiling: (1) clear and explicit definition of patient sample, (2) clinical coherence of model variables, (3) sufficiently high-quality and timely data, (4) designation of a reference time before which covariates are derived and after which outcomes are measured, (5) use of an appropriate outcome and a standardized period of outcome assessment, (6) application of an analytical approach that takes into account the multilevel organization of data, and (7) disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples.[30] In this study we have focused mainly on attribute 6. We have also adjusted for case-mix (attribute 7) but the model we used was quite simple and not externally validated. We suggest adding an attribute: consider a measure of rankability to judge what part of the observed differences is not due by chance. It remains however a value judgment when ranking is appropriate. We would suggest that any ranking is meaningless when rankability is low (<50%), that the ER should be used when rankability is moderate (>50% and <75%) and that simple integer ranks are appropriate when rankabilty is high (>75%). ERs and integer ranks will then be very similar.

We label remaining between hospital differences 'unexplained', since there can be many explanations to differences in outcome, including process of care, hospital characteristics, and more (unknown) patient characteristics. We will probably never know how large the 'true' differences are, or be able to completely explain them.[18]

To conclude, this study shows that adjustment for case-mix is crucial in measuring quality of care and ranking hospitals. Case-mix adjusted random effect estimates and the Expected Ranks are more robust alternatives to traditional traditional fixed effect estimates and simple rankings and may assist to prevent over-interpretation.

References

1. Green J, Wintfeld N. Report cards on cardiac surgeons. Assessing new york state's approach. *N Engl J Med.* 1995;332:1229-1232
2. Wang W, Dillon B, Bouamra O. An analysis of hospital trauma care performance evaluation. *J Trauma*. 2007;62:1215-1222
3. Shahian DM, Normand SL, Torchiana DF, et al. Cardiac surgery report cards: Comprehensive review and statistical critique. *Ann Thorac Surg*. 2001;72:2155-2168
4. Goldstein H, Spiegelhalter DJ. League tables and their limitations: Statistical issues in comparisons of institutional performance. *Journal of the Royal Statistical Society* Series A. 1996;159:385-443
5. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: Retrospective analysis of live birth rates. *Bmj.* 1998;316:1701-1704; discussion 1705
6. Lilford R, Mohammed MA, Spiegelhalter D, et al. Use and misuse of process and outcome data in managing performance of acute medical care: Avoiding institutional stigma. *Lancet.* 2004;363:1147-1154
7. Glance LG, Dick A, Osler TM, et al. Impact of changing the statistical methodology on hospital and surgeon ranking: The case of the new york state cardiac surgery report card. *Med Care*. 2006;44:311-319
8. Shahian DM, Torchiana DF, Shemin RJ, et al. Massachusetts cardiac surgery report card: Implications of statistical methodology. *Ann Thorac Surg*. 2005;80:2106-2113
9. Steyerberg EW, Eijkemans MJ, Boersma E, et al. Applicability of clinical prediction models in acute myocardial infarction: A comparison of traditional and empirical bayes adjustment methods. *Am Heart J.* 2005;150:920
10. Smits JM, De Meester J, Deng MC, et al. Mortality rates after heart transplantation: How to compare center-specific outcome data? *Transplantation*. 2003;75:90-96
11. Scholte op Reimer WJ, Dippel DW, Franke CL, et al. Quality of hospital and outpatient care after stroke or transient ischemic attack: Insights from a stroke survey in the netherlands. *Stroke*. 2006;37:1844-1849
12. Lingsma HF, Dippel DW, Hoeks SE, et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: Results from the netherlands stroke survey. *J Neurol Neurosurg Psychiatry*. 2008;79:888-894
13. Akaike H. *Information theory and an extension of the maximum likelihood principle*. 2nd International Symposium on Information Theory. 1973.
14. Houwelingen v, H.C. , Brand R, Louis TA. *Emperical bayes methods for monitoring health care quality.* Available at: http://www.msbi.nl/dnn/People/Houwelingen/Publications/tabid/158/Default.aspx. Accessed June 5, 2009.
15. Jennings BM, Staggers N, Brosch LR. A classification scheme for outcome indicators. *Image J Nurs Sch*. 1999;31:381-388
16. Donabedian A. Methods for deriving criteria for assessing the quality of medical care. *Med Care Rev.* 1980;37:653-698
17. LaBresh KA, Reeves MJ, Frankel MR, et al. Hospital treatment of patients with ischemic stroke or transient ischemic attack using the 'Get with the guidelines' Program. *Arch Intern Med.* 2008;168:411-417

18. Weir NU, Dennis MS. A triumph of hope and expediency over experience and reason? *J Neurol Neurosurg Psychiatry.* 2008;79:852
19. Mant J. Process versus outcome indicators in the assessment of quality of health care. *Int J Qual Health Care.* 2001;13:475-480
20. DeLong ER, Peterson ED, DeLong DM, et al. Comparing risk-adjustment methods for provider profiling. *Stat Med.* 1997;16:2645-2664
21. Krumholz HM, Wang Y, Mattera JA, et al. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation.* 2006;113:1683-1692
22. Timbie JW, Newhouse JP, Rosenthal MB, et al. *Med Decis Making.* 2008;28:419-434
23. Normand SL, Wolf RE, Ayanian JZ, et al. Assessing the accuracy of hospital clinical performance measures. *Med Decis Making.* 2007;27:9-20
24. Thomas N, Longford NT, Rolph JE. Empirical bayes methods for estimating hospital-specific mortality rates. *Stat Med.* 1994;13:889-903
25. Deely JJ, Smith AFM. Quantitative refinements for comparisons of institutional performance. *Journal of the Royal Statistical Society* Series A. 1998;161:5-12
26. Laird N, Louis T. Empirical bayes ranking methods. Journal of Educational Statistics. 1989;14:29-46
27. Shen W, Louis TA. Triple-goal estimates in two-stage hierarchical models. *Journal of the Royal Statistical Society* Series B. 1998;60:455-471
28. Krumholz HM, Normand SL. Public reporting of 30-day mortality for patients hospitalized with acute myocardial infarction and heart failure. *Circulation.* 2008
29. Johnson MA, Normand SL, Krumholz HM. How are our hospitals measuring up?: 'Hospital compare': A resource for hospital quality of care. *Circulation.* 2008;118:e498-500
30. Krumholz HM, Normand SL, Spertus JA, et al. Measuring performance for treating heart attacks and heart failure: The case for outcomes measurement. *Health Aff* (Millwood). 2007;26:75-85

Appendix 1: Formulas

Fixed effect logistic regression:

$\text{Logit}(P(Y_{ij} = 1|X_{ij})) = ßX_{ij} + \theta_i$

with

X_{ij}	the covariates (in this case the confounders) describing the patients characteristics of patient j in hospital i, including the constant term,
ß	the regression coefficients describing the effect of the covariates and the intercept,
θ_i	the effect of hospital i, that is the coefficient with respect to some overall mean.

Random effect logistic regression:

$\text{Logit}(P(Y_{ij} = 1|X_{ij})) = ßX_{ij} + \theta_i$

with

X_{ij}	the covariates (in this case the confounders) describing the patients characteristics of patient j in hospital i, including the constant term,
ß	the regression coefficients describing the effect of the covariates and the intercept,
θ_i	the effect of hospital i, that is the coefficient with respect to some overall mean, drawn from a normal distribution with mean and variance

Expected rank:

$ER_i = 1 = \Sigma_{i \neq k}((F(\theta_i - \theta_k))/\sqrt{(\text{var}(\theta_i) + \text{var}(\theta_k))})$

with

F	the normal distribution function,
$\theta_i - \theta_k$	magnitude of the difference of a particular hospital with other hospitals and,
$\text{var}(\theta_i) + \text{var}(\theta_k)$	the uncertainty in this difference.

Percentiles based on expected ranks:

$PCER_i = 100*(ER_i - 0.5)/N$

with

ER_i	the expected rank of a particular hospital and,
N	the number of hospitals.

Rankability:

$\rho = \tau^2/(\tau^2 + median(s_i^2))$

with:

τ^2	the variance of the random effects,
s_i^2	the variance of the fixed effect individual hospital effect estimates.

Appendix 2: R code

```
# Load required packages
library(Design)
library(lme4) #fits random effect logistic regression models
library(foreign) #can import foreign data files

# Import
cva <- as.data.frame(read.spss('D:/My Documents/.....sav'))

# Test differences between hospitals with fixed and random effects (Table 2.2)

# Hospital in fixed effect analysis('CENTER')for poor outcome('RANKIN6')
unadjusted.ZH <- lrm(RANKIN6~as.factor(CENTER),data=cva)
deviance(unadjusted.ZH)[1] - deviance(unadjusted.ZH)[2]
pchisq(q=deviance(unadjusted.ZH)[1] - deviance(unadjusted.ZH)[2], df=9, 0, F)
# Result: chi2=49.7, df=9, p=1.24e - 7

# Random effects model
unadj.ZH.Laplace <- lmer(RANKIN6~1+(1|CENTER), family=binomial,
method='Laplace', data=cva)
deviance(unadjusted.ZH)[1] - deviance(unadj.ZH.Laplace)
pchisq(q=deviance(unadjusted.ZH)[1] - deviance(unadj.ZH.Laplace), df=1,
lower.tail=F) /2 # divide p-value by 2
# Result: chi2=23, df=1, p= 6.23e - 7

# Full model 12 confounders
full <- lrm(RANKIN6~AGE+SEX+DIAGNOSE+BARTDEP+PRECVA+SEVERESTR+GCSLOW+
AF+IHD+DIAB+HYPTEN+HYPERCH, data=cva, x=T, y=T)
full.ZH <- lrm(RANKIN6~AGE+SEX+DIAGNOSE+BARTDEP+PRECVA+SEVERESTR
+GCSLOW+AF+IHD+DIAB+HYPTEN+HYPERCH+as.factor(CENTER),data=cva)
fullr.ZH.Laplace <- lmer(RANKIN6~AGE+SEX+DIAGNOSE+BARTDEP+PRECVA
+SEVERESTR+GCSLOW+AF+IHD+DIAB+HYPTEN+HYPERCH+(1|CENTER),
family=binomial, method='Laplace', data=cva, x=T,  model=T)
deviance(full)[2] - deviance(full.ZH)[2]
pchisq(q=deviance(full)[2] - deviance(full.ZH)[2], df=9, lower.tail=F)
# Result: chi2=24, df=9, p= 0.00415
deviance(full)[2] - deviance(fullr.ZH.Laplace)
pchisq(q=deviance(full)[2] - deviance(fullr.ZH.Laplace), df=1,lower.tail=F)/2
# Result: chi2=4 df=1 p=0.0275
```

```
# Estimate differences between hospitals (Tables 2.3 and 2.4, figure 2.1)
# make center.effects function for individual hospital effects,
# a matrix with differences against the average

# function for individual hospital effects
center.effects <- function(outcome,center,lp=F) {
Ncenter <- table(center)
ncenters <- length(Ncenter)
resultsR <- matrix(nrow=ncenters,ncol=8)
dimnames(resultsR) <- list(1:ncenters, c('Label', 'GROUP','n', 'p','pmean', 'Coef', 'SE',
'Var'))
# compare to average if no lp is given
if (lp[1]==F) lp <- rep(log(mean(outcome)/
(1-mean(outcome))),length(outcome))#logit function

for (i in 1:ncenters) { # go through all hospitals
f <- lrm.fit(y=outcome[center==i], offset=as.vector(lp[center==i]))
resultsR[i,] <- c(1,i,f$stats[1],sum(outcome[center==i])/f$stats[1],
mean(outcome), f$coef, sqrt(f$var), f$var)
resultsR
} # End loop over hospitals
} # end function for hospital effects

# linear predictors unadjusted and adjusted fixed effects
lpuni <- rep(log(mean(cva$RANKIN6)/(1-mean(cva$RANKIN6))),
length(cva$RANKIN6))+rnorm(length(cva$RANKIN6), mean=0, sd=.001) #logit func-
tion
lp.cva <- full$x  %*% full$coef[2:13] + full$coef[1]

adj.ZH <- center.effects(cva$RANKIN6,center=cva$CENTER,lp=lp.cva)
adj.ZH

unadj.ZH <- center.effects(outcome=cva$RANKIN6, center=cva$CENTER, lp=lpuni)
unadj.ZH

# Adjusted with random effect estimation
rZH <- ranef(fullr.ZH.Laplace, postVar=T) #random effect estimates and variance
RA.ZH <- cbind(as.vector(rZH[[1]]), as.vector(sqrt(rZH[[1]]@postVar)))
names(RA.ZH) <- c('Coef', 'SE')
RA.ZH #Results
```

```
# Rankings
ER <- rep(NA,10)
tau2 <- as.numeric(VarCorr(fullr.ZH.Laplace)[[1]])
for (i in 1:10) {
ER[i] <- 1+ sum(pnorm((RA.ZH [i,1] - RA.ZH [-i,1])/
sqrt(RA.ZH [i,2]^2 + RA.ZH [-i,2]^2)))
} # end loop
PCER <- 100*(ER - 0.5)/10

cbind(rank(unadj.ZH[,'Coef']), rank(adj.ZH[,'Coef']),
rank(RA.ZH[,'Coef']), ER, PCER)

# rankibility rho:
sigma2 <- adj.ZH[,8] # variance of fixed effect estimates
rho <- tau2/(tau2+median(sigma2))
```

3 Rankability of hospitals using outcome indicators

van Dishoeck AM, Lingsma HF, Mackenbach JP, Steyerberg EW. Random variation and rankability of hospitals using performance indicators. Submitted.

Abstract

Objective
There is a growing focus on quality and safety in health care. Outcome indicators are increasingly used to rank hospital performance, but the reliability of ranking is under debate. We aim to quantify the rankability of several outcome indicators of hospital performance currently used by the Dutch government.

Methods
From 52 indicators used by the Netherlands Inspectorate, we selected nine outcome indicators presenting a fraction and absolute numbers. Of these indicators, four were combined into two, resulting in seven indicators for analysis. We used the official data of 97 Dutch hospitals of the year 2007. We estimated uncertainty of the observed outcome within the hospitals (within center variance, σ^2) with fixed effect logistic regression models. We measured heterogeneity (between center variance, τ^2) with random effect logistic regression models. Subsequently, we calculated the rankability by relating heterogeneity to uncertainty within and between centers (τ^2 / (τ^2 +median σ^2)).

Results
Sample sizes varied typically around 200 per center (range of median 90-277) with median 2-21 cases, causing a substantial uncertainty of outcomes per center. Although 4-8 fold differences between hospitals were noted, the uncertainty within the centers caused a poor (< 50%) rankability in 3 indicators and moderate rankability (50-75%) in the other 4 indicators.

Conclusion
The currently used Dutch outcome indicators are not suitable for ranking hospitals. When judging hospital quality the influence of random variation must be accounted for to avoid overinterpretation of the numbers in the quest for more transparency in health care. Adequate sample size is a prerequisite in attempting reliable ranking.

Introduction

There is a growing focus on quality and safety in health care. Increasingly indicators are used to assess hospital performance. In different countries nationwide systems have been set up to monitor the performance of health care institutions using a framework of structure, process and outcome indicators.[1, 2] Public disclosure of the results of hospital performance leads to several inconsistent rankings and there is a growing concern among professionals about the value and reliability of these ranks.[3-10]

Two core components determine the reliability of ranking; within center uncertainty and between center heterogeneity. The amount of uncertainty in the analysis of hospital performance is higher than intuition might have suggested.[11] For low-incidence surgeries and for smaller subgroups in the population uncertainty was large. The smallest hospitals would likely experience five to seven times more uncertainty concerning their ranks.[12] For instance, to assess the influence of uncertainty of revision rates in orthopedic surgery, fixed and random effect models have been used.[13] In contrast to the random-effects model, uncertainty can easily cause overly optimistic or pessimistic outcomes in the fixed-effects model.

Secondly there is heterogeneity between centers.[14] Heterogeneity relates to the true differences beyond chance between centers and can be estimated with random effect models. Both components determine 'rankability' of an outcome indicator.

Lingsma et al used rankability to assess ranking of a small numbers of IVF clinics.[15] They found considerable heterogeneity, while uncertainty per clinic was small because of large numbers. This resulted in a substantial rankability with only 10% of the observed differences between the clinics attributed to chance. There are no minimal sample size requirements for the indicators used by the Dutch government. The numbers may be much smaller, making ranking attempts less reliable. We aim to quantify the rankability of several outcome indicators of hospital performance in the Netherlands.

Methods

Data

We obtained the data from the Netherlands Inspectorate's indicator set. The inspectorate uses this set to assess possible flaws in the quality of care in Dutch hospitals. This obligatory set includes 21 areas with 52 performance indicators (PIs), of which 14 are outcome indicators presenting actual numbers (nominator and denominator) and corresponding percentages. Five indicators were excluded because of clear evidence of registration bias, such as extrapolation of a limited sample in time or patient groups, leaving 9 outcome indicators (Table 3.1). We used the data of 2007, which are publicly available at www.ziekenhuizentransparant.nl. For acute myocardial infarction (AMI), the majority of hospitals reported the in-hospital mortality instead of the 30-day mortality. Several hospitals report both. Using these data, we multiplied 0.74 to the 30-day mortality to include data for the five hospitals that only reported 30-day mortality.

Uncertainty

We used nominator and denominator data for each hospital to create a patient level dataset. We estimated a coefficient for unfavorable outcome for each hospital and compared it to the overall average, using a fixed effect logistic regression model with an offset variable and center as a categorical variable. The standard error of the estimated coefficient, representing the mean outcome (σ^2) indicated the uncertainty within the hospital.

Heterogeneity

We fitted a random effect logistic regression model to estimate unexplained heterogeneity, indicated by τ^2 (the between center variance). Unlike the fixed effect model, the random effect model accounts for the fact that the observed outcomes for smaller hospitals can take on extreme values because of random variation. The variance indicates the differences between hospitals beyond chance.[16]

For the interpretation of τ^2 we calculated a 95% range of odds ratios for the centers compared to the average as $\exp(-1{,}96 * \tau^2);\exp(1{,}96 * \tau^2)$.[17]

Rankability

To estimate rankability, we use the following formula:

$\rho = \tau^2 / (\tau^2 + \text{median } \sigma^2)$

Rankability relates the heterogeneity τ^2 from the random effect logistic regression model (differences between the hospitals) to the standard error σ^2 of the individual hospitals from the fixed effect logistic regression model. Rankability can be interpreted as the part of heterogeneity between hospitals that is due to unexplained differences, and the rest is due to natural variation or chance. Therefore, rankability describes the reliability of ranking.

Table 3.1 Outcome indicators and their description

Indicator	Numerator	Denominator
Nosocomial Pressure Ulcer (PU) prevalence among hospitalized patients	Number of patients with a pressure ulcer grade 2–4	All hospitalized patients who were examined for the presence of PU
Pressure Ulcer (PU) incidence after total hip replacement	Number of patients with a pressure ulcer grade 2–4	All total hip replacement patients
Bile duct leakage within 30 days after cholecystectomy	Number of patients with bile duct leakage within 30 days after cholecystectomy	All patients with a cholecystectomy
Unintended reoperation after colorectal surgery	Number of unintended reoperation after colorectal surgery	All colorectal operations excluding appendix
In hospital mortality after acute myocardial infarction (AMI) for patients younger than 65 years	Number of patients younger than 65 years deceased during hospitalization because of AMI	All patients younger than 65 years hospitalized because of AMI
In hospital mortality after acute myocardial infarction (AMI) for patients of 65 year and older	Number of patients 65 years and older deceased during hospitalization because of AMI	All patients 65 years and older hospitalized because of AMI
Readmission after heart failure for patients younger than 75 year	Number of readmissions after heart failure within 12 weeks after hospital discharge in patients younger than 75 years	All patients younger than 75 years admitted for heart failure.
Readmission after heart failure for patients 75 year and older	Number of readmissions after heart failure within 12 weeks after hospital discharge in patients 75 years and older	All patients younger 75 years and older admitted for heart failure.
Remaining cancer tissue after breast-conserving lumpectomy	Number of patients in whom cancer tissue is left after an initial local excision of a malignant breast tumour	All patients treated with a local excision of a malignant breast tumour

Case-mix adjustment

The data on performance indicators did not include patient characteristics, except for two outcomes; AMI mortality and heart failure re-admission. The original indicators are stratified by age. We combined the indicators AMI <65 years + ≥65 years, and heart failure <75 years + ≥75 years in two datasets and applied a limited age adjustment by putting age group in the fixed part of the random effect model.

The statistical analysis was performed with R statistical software (version 2.7.1, R Foundation for Statistical Computing, Vienna, Austria), using the lme4 library to fit random effect logistic regression models.

Results

We studied nine outcome indicators of which four were combined in two (Table 3.1).

Within center uncertainty

The number of cases with unfavorable outcome, as well as the total number of patients per center varied widely for the different indicators (Table 3.2). For instance, pressure ulcer prevalence varied from 0–39 cases, while the number of patients ranged from 59–548. For cholecystectomy, the number of cases with bile duct leakage was very small (median 2). A considerable number of hospitals reported zero cases (29 out of 97), resulting in a median incidence of leakage of the bile duct of 0,5%. The within center uncertainty was largest among cholecystectomy patients (σ 1.01), and pressure ulcer incidence (σ 0.85), due to small number of cases (Table 3.3).

Table 3.2 Descriptive statistics

Indicator	n center	Median Cases (range)	Median N (range)	Median outcome % (range)
Nosocomial Pressure Ulcer prevalence	93	10 (0–39)	233 (59–548)	3.7 (0–11.1)
Nosocomial Pressure Ulcer incidence total hip replacement	90	2 (0–23)	197 (26–1131)	1.1 (0–8.9)
Leakage of the bile duct within 30 days after cholecystectomy	95	2 (0–7)	255 (109–625)	0.5 (0–3.63)
Unintended reoperation after colorectal surgery	94	15 (0–47)	209 (57–557)	6.9 (0–18.4)
In hospital mortality after AMI age <65 years	88	1 (0–17)	85.5 (4–720)	1.1 (0–6.8)
In hospital mortality after AMI age ≥65 years	88	10 (0–46)	117.5 (28–541)	8.6 (0–20.8)
Readmission after heart failure age <75 years	93	6 (0–30)	77 (13–389)	7.9 (0–22.6)
Readmission after heart failure age ≥75 years	93	10 (0–50)	133 (13–376)	8.0 (0–23.1)
Remaining cancer tissue after breast-saving lumpectomy	94	7 (1–46)	76 (14–300)	10.5 (1.2–35.7)

Table 3.3 Rankability

Indicator	sigma2	tau2	95% range OR		rankability
			–	+	
Nosocomial Pressure Ulcer prevalence	0.19	0.11	0.52	1.91	37%
Nosocomial Pressure Ulcer incidence total hip replacement	0.85	0.16	0.46	2.17	38%
Leakage of the bile duct within 30 days after cholecystectomy	1.01	0.00	1	1	0%
Unintended reoperation after colorectal surgery	0.12	0.29	0.35	2.86	71%
In hospital mortality after AMI age groups combined#	0.19	0.27	0.36	2.76	58%
Readmission after heart failure age groups combined#	0.14	0.15	0.47	2.11	51%
Remaining cancer tissue after breast-saving lumpectomy	0.25	0.28	0.35	2.82	53%

results for the combined age groups are adjusted for age

Between center heterogeneity

Heterogeneity between the centers varied from none (τ^2 0) for cholecystectomy, to τ^2 0.29 for colorectal surgery. The corresponding 95% range of the odds ratios was 0.35 and 2.86 for colorectal surgery, meaning that hospitals at the higher end of the distribution had a 2.86 higher chance of re-operation than in the average hospital. Similar at the lower end of the distribution patients had a 0.35 lower chance of reoperation. This was equivalent to an eight-fold difference between the hospitals for this indicator.

Rankability

Due to the large between center differences, rankability was the highest (71%) for colorectal surgery and the lowest (<50%) for the indicators pressure ulcer prevalence, pressure ulcer incidence, and cholecystectomy. For pressure ulcer the rankability was relatively low despite a σ^2 of 0.19 related to the small between center differences τ^2. Rankability was moderate (50%-75%) for the indicators colorectal surgery, AMI, heart failure readmission, and breast saving lumpectomy.

Adjustment for case-mix revealed that a part of the heterogeneity in the AMI indicator could be explained by age. For heart failure readmission, age was borderline significant. Rankability for the combined indicator AMI was 58% and for heart failure 51%.

Discussion

We tested several outcome indicators on their reliability for ranking hospitals using the concept of rankability. Combining fixed effect logistic regression models and random effect logistic regression models, we could estimate uncertainty within the individual hospitals and the unexplained heterogeneity between hospitals. We found considerable variability due to chance alone within the hospitals. On the other hand, the unexplained differences between the hospitals were small for some indicators. Both lead to low rankability.

The indicators in our research showed substantial uncertainty that influenced rankability. For cholecystectomy, there were no differences other than those by chance alone between the hospitals. Using this indicator for ranking hospitals is useless. This adds to the criticism by de Reuver et al about this indicator.[18] Substantial heterogeneity led to high rankability in the colorectal surgery indicator (71%). Nevertheless, for this indicator it remains unclear how much of these differences are caused by case mix. It is plausible that the indication for surgery such as traumatic injury or colorectal cancer may play a role in reoperation rate. Case mix correction should be performed before using this indicator for ranking hospitals.

The lack of heterogeneity influences the rankability of the pressure ulcer prevalence. For cholecystectomy the between hospital differences did not exceed chance variability. For AMI and heart failure, we were able to perform a simple stratification for two age groups. Combining both age groups resulted in a larger number of cases and total numbers. While rankability of the group of patients younger than 65 was low due to the limited number of cases, the pooled data stratified for age had a moderate rankability (51%).

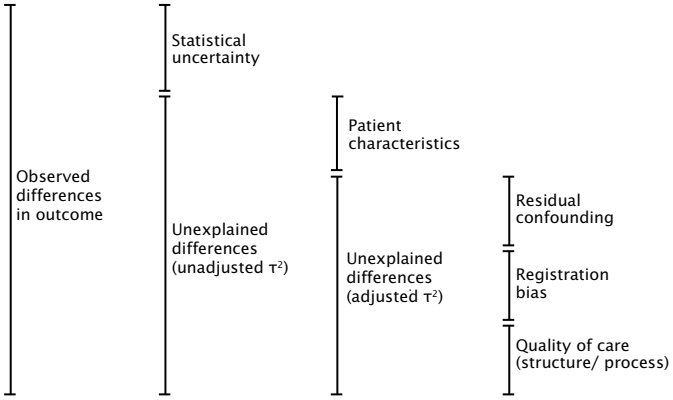
The measurement of rankability provides a way of assessing reliability of ranking. The rankability can be seen the ratio between signal and (statistical) noise. A rankability of 50% then indicates 50% signal on quality of care and 50% statistical noise. A categorization for rankability is yet still arbitrary. Lingsma et al suggested that > 70% rankability should be fair to rank hospitals.[15] Higgins et al assigned adjectives of low, moderate and high to the I^2 values of 25%, 50% and 75%.[19] I^2 is used to measure heterogeneity in meta-analyses[20] and is similar in nature to our rankability measure. I^2 can be interpreted as the percentage of the total variability in a set of effect sizes due to heterogeneity, that is, to between study variability. Adopting this categorization, we found that none of the outcome indicators had a high rankability. It could be argued that in case of moderate rankability, 'expected ranks' should be used that take into account random variability.[12-14] This requires statistical knowledge and access to advanced statistical programs. No ranking attempt should be made with low rankability.

Compared to previous research of IVF patients, rankability in our data was much lower.[15] Only 10% of the differences between IVF centers were explained by chance, which is explained by the large sample size of the IVF centers studied (median 654 cycles). In the Dutch outcome indicators, not only the total numbers were sometimes small (median between 90 and 277) but also the outcome was frequently low. Simple rankings based on fixed effects of hospital performance disregards both the magnitude and the uncertainty of the differences between hospitals.[21] An illustrative example is the cholecystectomy indicator, where the number of cases was too low to detect any differences between hospitals. Small samples and low event rates limit the statistical power of the comparison between hospitals.[22]

This raises questions about minimal power calculations or combining indicators to provide sufficient sample size. Classical power calculation or estimating minimal cases and total numbers might be performed using Cohen's D. D is defined as the difference between two means divided by the standard deviation. Effect sizes are commonly defined as small, d =0.2, medium, d =0.5, and large, d =0.8. We might use a variant of Cohen's D for event rate. The population size for d=0.5 than is at least 200 and at least 800 for d= 0,2 for indicators with sufficient event rates.[23] These numbers can be used as 'a rule of thumb' for the assessment of the reliability of ranking hospitals. Looking at the sample sizes for the pressure ulcer indicator (59-548) in the Dutch hospitals, it is questionable if this indicator will ever be suitable for ranking hospitals. The maximal sample size is limited by the number of beds in a hospital. In case of inadequate numbers the presentation the results of a specific indicator could be done using funnel plots, since this presentation visualizes the differences between hospitals in relation to random variation.[24] Realistic presentation is important to avoid gaming and truly encourage actions to improve the quality of care.[25]

Reliability of ratings depends on sample size and heterogeneity, but also on biases. We can draw a conceptual framework to summarize the elements of between center differences (Figure 3.1).[15] The observed differences can be divided in unexplained differences and chance. By using random effect models chance can be corrected for, leaving patients characteristics, registration bias, quality of care and residual confounding as elements of the unexplained differences. In our research, the uncertainty is accounted for in calculating rankability. Consequently, ranking reflects the total of unexplained differences between hospitals and not true differences in the quality of care. This is a limitation of this study, but the data as publicly reported does not provide any additional information.

Figure 3.1 Conceptual framework of between-center differences. Observed differences can be divided in random variation and unexplained differences, which can be further attributed to patient characteristics that were not adjusted for, residual confounding because of imperfect case-mix correction, registration bias. Differences in quality of care remain as explanation for a final part of between-center differences.



We conclude that none of the currently used Dutch outcome indicators is suitable for ranking hospitals. When judging hospital quality the influence of random variation must be accounted for to avoid overinterpretation of the numbers in the quest for more transparency in health care. Adequate sample size is a prerequisite in attempting reliable ranking.

Acknowledgements

This study was funded by Internal Erasmus MC grant for health care research (Mrace)

References

1. Jencks SF, Cuerdon T, Burwen DR, Fleming B, Houck PM, Kussmaul AE, et al. Quality of medical care delivered to Medicare beneficiaries: A profile at state and national levels. *Jama*. 2000 Oct 4;284(13):1670-6.
2. Berg M, Meijerink Y, Gras M, Goossensen A, Schellekens W, Haeck J, et al. Feasibility first: developing public performance indicators on patient safety and clinical effectiveness for Dutch hospitals. *Health Policy*. 2005 Dec;75(1):59-73.
3. Halasyamani LK, Davis MM. Conflicting measures of hospital quality: ratings from 'Hospital Compare' versus 'Best Hospitals'. *Journal of hospital medicine* (Online). 2007 May;2(3):128-34.
4. Lemmers O, Kremer JA, Borm GF. Incorporating natural variation into IVF clinic league tables. *Human reproduction* (Oxford, England). 2007 May;22(5):1359-62.
5. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet*. 2004 Apr 3;363(9415):1147-54.
6. Mohammed MA, Mant J, Bentham L, Raftery J. Comparing processes of stroke care in high- and low-mortality hospitals in the West Midlands, UK. *Int J Qual Health Care*. 2005 Feb;17(1):31-6.
7. Ranstam J, Wagner P, Robertsson O, Lidgren L. Health-care quality registers: outcome-orientated ranking of hospitals is unreliable. *The Journal of bone and joint surgery*. 2008 Dec;90(12):1558-61.
8. Jacobs R, Goddard M, Smith PC. How robust are hospital ranks based on composite performance measures? *Med Care*. 2005 Dec;43(12):1177-84.
9. Spiegelhalter D. Ranking institutions. *The Journal of thoracic and cardiovascular surgery*. 2003 May;125(5):1171-3; author reply 3.
10. Anderson J, Hackman M, Burnich J, Gurgiolo TR. Determining hospital performance based on rank ordering: is it appropriate? *Am J Med Qual*. 2007 May-Jun;22(3):177-85.
11. Diehr P, Cain K, Connell F, Volinn E. What is too much variation? The null hypothesis in small-area analysis. *Health Serv Res*. 1990 Feb;24(6):741-71.
12. Davidson G, Moscovice I, Remus D. Hospital size, uncertainty, and pay-for-performance. *Health care financing review*. 2007 Fall;29(1):45-57.
13. Robertsson O, Ranstam J, Lidgren L. Variation in outcome and ranking of hospitals: an analysis from the Swedish knee arthroplasty register. *Acta orthopaedica*. 2006 Jun;77(3):487-93.
14. Normand S-LT, Shahian DM. Statistical and clinical aspects of hospital profiling. *Statistical Science*. 2007;22(2):206-26.
15. Lingsma HF, Eijkemans MJ, Steyerberg EW. Incorporating natural variation into IVF clinic league tables: The Expected Rank. *BMC medical research methodology*. 2009 Jul 16;9(1):53.
16. Spiegelhalter DJ. Handling over-dispersion of performance indicators. *Qual Saf Health Care*. 2005 Oct;14(5):347-51.
17. Spiegelhalter D, Abrahams, Myles. *Baysean approaches to clinical trials and health care evaluation* 2004.

18. de Reuver PR, Gouma DJ. [Bile leakage. A performance indicator with markedly different consequences for the patient, specialist and care insurer]. *Nederlands tijdschrift voor geneeskunde*. 2007 Aug 4;151(31):1709-12.
19. Higgins JP, Thompson SG, Deeks JJ, Altman DG. Measuring inconsistency in meta-analyses. *Bmj*. 2003 Sep 6;327(7414):557-60.
20. Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. Assessing heterogeneity in meta-analysis: Q statistic or I2 index? *Psychological methods*. 2006 Jun;11(2):193-206.
21. Lingsma HF, Dippel DW, Hoeks SE, Steyerberg EW, Franke CL, van Oostenbrugge RJ, et al. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. *J Neurol Neurosurg Psychiatry*. 2008 Aug;79(8):888-94.
22. Dimick JB, Welch HG, Birkmeyer JD. Surgical mortality as an indicator of hospital quality: the problem with small sample size. *Jama*. 2004 Aug 18;292(7):847-51.
23. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*. second ed. Philadelphia: Lawrence Erlbaum Associates 1988.
24. Spiegelhalter DJ. Funnel plots for comparing institutional performance. *Stat Med*. 2005 Apr 30;24(8):1185-202.
25. Gibberd R, Hancock S, Howley P, Richards K. Using indicators to quantify the potential to improve the quality of health care. *Int J Qual Health Care*. 2004 Apr;16 Suppl 1:i37-43.

4 Comparing software packages for random effect models

Baoyue L, Lingsma HF, Steyerberg EW, Lesaffre E. Logistic random effects regression models: a comparison of statistical packages. *BMC Medical Research Methodolody.* Provisionally accepted.

Abstract

Background
Logistic random effects models are popular to analyze multilevel data with a binary or ordinal outcome. Here, we aim to compare different statistical software implementations of these models.

Methods
We used individual patient data from 8509 patients in 231 centers with moderate and severe Traumatic Brain Injury (TBI) enrolled in eight Randomized Controlled Trials (RCTs) and three observational studies. We fitted logistic random effects regression models, with the 5-point Glasgow Outcome Scale (GOS) as outcome, both dichotomized as well as ordinal with center as a random effect, and as covariates age, motor score, pupil reactivity and study. We then compared estimates of the fixed and random effects from different statistical packages. Bayesian approaches included MLwiN, WinBUGS and the SAS experimental procedure MCMC, frequentist approaches included R (lmer), MIXOR, SAS (GLIMMIX and NLMIXED), and also MLwiN.

Results
The packages gave similar parameter estimates for both the fixed effects and random effects, for both the binary and ordinal models. The software implementations differed however considerably in flexibility, computation time, and usability. There were also differences in the availability of additional tools for model evaluation, such as diagnostic plots. The experimental SAS procedure MCMC appeared to be inefficient.

Conclusion
All studied software implementations of logistic random effects regression models produce similar results. If there is no explicit preference for a frequentist or Bayesian approach, the choice for a particular implementation may solely depend on the desired flexibility, and the usability.

Introduction

Hierarchical, multilevel, or clustered data structures are often seen in medical, psychological and social research. Examples are: (1) individuals in households and households nested in geographical areas, (2) surfaces on teeth, teeth within the mouth, (3) children in classes, classes in schools, (4) multicenter clinical trials, in which individuals are treated in centers, (5) meta-analyses with individuals nested in studies. Multilevel data structures also arise in longitudinal studies where measurements are clustered within individuals.

The multilevel structure induces correlation among observations within a cluster, e.g. between patients from the same center. An approach to analyze clustered data is the use of multilevel or random effects regression analysis. There are several reasons to use a random effects model instead of a traditional fixed effects regression model.[1] First, we may want to estimate the effect of covariates at the group level, e.g. type of center (university versus peripheral center). Using a fixed effects model it is not possible to separate out group effects and effect of covariates on the group level. Secondly, random effects models treat the groups as a random sample from a population of groups. Using a fixed effects model, inferences cannot be made beyond the groups in the sample. Thirdly, statistical inference may be wrong. Using traditional regression techniques that not recognize the multilevel structures will cause the standard errors of regression coefficients to be wrongly estimated, leading to an overstatement or understatement of statistical significance for the coefficients of the higher-level covariates.

All this is common knowledge in the statistical literature,[2] but in the medical literature still multilevel data are often analyzed using fixed effects models.[3]

In this paper we use a multilevel dataset with an ordinal outcome, which we analysed as such, and dichotomized as a binary outcome. Relating patient and cluster characteristics to the outcome requires the use of a logistic random effects model. Such models are implemented in many different statistical packages, all with different features and using different computational approaches. Thus, it is possible that different packages give different parameter estimates. In this study we aim to compare different statistical software implementations, with regard to the results they give, the methods they use and their usability. The implementations include both frequentist and Bayesian approaches.

Methods

Data

The dataset we used for this study is the IMPACT (International Mission on Prognosis and Clinical Trial design in TBI) database. This dataset contains individual patient data from 9,205 patients with moderate and severe Traumatic Brain Injury (TBI) enrolled in eight Randomized Controlled Trials (RCTs) and three observational studies. The patients were treated in different centers, giving the data a multilevel structure. For more details on this study, we refer to Marmarou et al.[4], and Maas et al.[5].

Outcome

The outcome in our analyses is the Glasgow Outcome Scale (GOS), the commonly used outcome scale in TBI studies. GOS is an ordinal five point scale, with categories dead, vegetative state, severe disability, moderate disability and good recovery. We analyzed the GOS as a binary outcome, dichotomized into 'unfavourable' (dead, vegetative and severe disability) versus 'favourable' (good recovery and moderate disability), and as the original ordinal scale.

Covariates

At patient level, we included age, pupil reactivity and motor score at admission as predictors in the model, based on previous studies.[6] Age was treated as a continuous variable. Motor score and pupil reactivity were included as categorical variables (motor score: 1=none or extension, 2=abnormal flexion, 3=normal flexion, 4=localises or obeys, 5=untestable, and pupil reactivity: 1=both sides positive, 2=one side positive, 3=both sides negative).

We also included the variable 'study' since 11 studies were involved and the outcome may vary systematically among studies. In the random effects model, 'study' was turned into 11 dummy variables. The 231 centers were treated as a random effect (intercept) with variance σ^2.

Random effects models

In random effects models, the residual variance is split up in components that pertain to the different levels in the data.[7] In our study, a two-level model with grouping of patients within centers would include residuals at the patient and center level. Thus the residual variance is partitioned into a between-center component (the variance of the center-level residuals) and a within-center component (the variance of the patient-level residuals). The center residuals, often called 'center effects', represent unobserved center characteristics that affect patients' outcomes. Unique for binary response random effects models is that only the level-2 residuals are meaningful and no level-1 residuals are specified.

A dichotomous logistic random effects model has a binary outcome (Y=0 or 1) and models the log odds of the outcome probability as a function of various predictors to estimate the probability that Y=1 happens, given the random effects. The simplest dichotomous 2-level model is given by

$$\text{Log}it\,(\pi_{ij}) = \alpha_1 + \sum_{k=1}^{K}\beta_k x_{kij} + u_j \qquad (1)$$

$$u_j \sim N\,(0,\sigma^2) \qquad j = 1, 2, \ldots, J \qquad i = 1, 2, \ldots, n_j$$

With $\pi_{ij} = \text{Prob}(Y_{ij} = 1 \mid \text{covariates}, u_j)$ whereby Y_{ij} is here the dichotomized GOS (with $Y_{ij} = 1$ if GOS = 1,2,3) of the i th subject in j th center. Further, x_{kij} represent the (first and second level) predictors, α_1 is the intercept and β_k the kth regression coefficient. Furthermore, u_j is the random effect representing the effect of j th center. It is assumed that u_j follows a normal distribution with mean 0 and variance σ^2. In our research, x_{kij} represents the covariates age, motor score, pupil reactivity and study. The coefficients β_k can be interpreted as the log odds ratio of the predictor x_{kij} when increasing with one unit and controlling for the other predictors and for the random effects in the model. In other words, the odds ratio ($\exp(\beta_k) = OR_k$), is defined given u_j.

For an ordinal logistic multilevel model, we adopt the proportional odds assumption and hence we assume that:

$$\text{Ln}\left(\frac{P(Y_{ij}=1)}{P(Y_{ij}=2,3,4,5)}\right) = \alpha_1 + \sum_{k=1}^{K}\beta_k x_{kij} + u_j \qquad (2)$$

$$\text{Ln}\left(\frac{P(Y_{ij}=1,2)}{P(Y_{ij}=3,4,5)}\right) = \alpha_2 + \sum_{k=1}^{K}\beta_k x_{kij} + u_j$$

$$\text{Ln}\left(\frac{P(Y_{ij}=1,2,3)}{P(Y_{ij}=4,5)}\right) = \alpha_3 + \sum_{k=1}^{K}\beta_k x_{kij} + u_j$$

$$\text{Ln}\left(\frac{P(Y_{ij}=1,2,3,4)}{P(Y_{ij}=5)}\right) = \alpha_3 + \sum_{k=1}^{K}\beta_k x_{kij} + u_j$$

$$u_j \sim N\,(0,\sigma^2) \qquad j = 1, 2, \ldots, J \qquad i = 1, 2, \ldots, nj$$

In model (2), Y_{ij} is the GOS of i th subject in j th center. The difference of the four equations in model (2) is only in the intercept, the effect of the covariates is assumed to be the same for all outcome levels (proportional odds assumption). So the coefficient β_k can be interpreted as the log odds ratio of a higher GOS versus a lower GOS when

the predictor x_{kij} increases with one unit controlling for the other predictors and the random effects in the model.

Both for the binary and the ordinal analysis we assume a logit link function and a normal distribution for the random effects, but we will also consider different link functions and other random effects distributions below.

Software packages

We compared eight different programs and functions that have implemented logistic random effects regression models. The software packages can be divided according to the statistical approach upon which they are based: frequentist or Bayesian.

The frequentist approaches are included in the program MIXOR (the first program launched for the analysis of a logistic random effects model), the R-function lmer in the lme4 package, the SAS procedures GLIMMIX and NLMIXED (here we used SAS version 9.2), and the package MLwiN (version 2.13).

The frequentist approaches differ in the way the integral is approximated to establish the (restricted) maximum likelihood estimate (MLE) integrated over the random effects. Further, they differ in the optimization technique to arrive at the MLE.

With regard to the integral approximation, in MIXOR, only Gauss-Hermite quadrature is available and one can specify the number of quadrature points Q, depending on the desired accuracy.[8] The Gauss-Hermite method is also referred to as non-adaptive Gaussian quadrature. The R-function lmer is based on the Laplace technique, which is an Adaptive Gaussian Quadrature (AGQ) technique where the integral is numerically calculated over the whole support of the likelihood using Q quadrature points adapted to the data.[9] The SAS procedure NLMIXED allows for several integration approaches, we used AGQ as a default in this research.[10] The same holds for the SAS procedure GLIMMIX.[11] Finally, the package MLwiN adopts Marginal Quasi-Likelihood (MQL) or Penalised quasi-Likelihood (PQL) to achieve the approximation. Both of these two methods can be computed up to the 2nd order,[12] here we chose the 2nd order PQL procedure.

With regard to the optimization technique resulting in the MLE, a variety of techniques has been developed. In MIXOR, the Fisher-scoring algorithm is used. R function lmer uses NLMINB method which is a local minimiser for the smooth nonlinear function subject to bound-constrained parameters. SAS procedures NLMIXED and GLIMMIX both have a large number of optimization. We chose the Newton-Raphson algorithm for NLMIXED and the default Quasi-Newton approach for GLIMMIX. Finally, the package MLwiN adopts iterative generalised least squares (IGLS) or reweighted IGLS (RIGLS) optimization methods and we used the default IGLS.

The other three programs we studied are based on a Bayesian approach, for a general introduction to Bayesian analysis, see Lee.[13] The program most often used for Bayesian analysis is WinBUGS (latest version is 1.4.3). WinBUGS is based on the Gibbs Sampler, which is one of the Markov Chain Monte Carlo (MCMC) approaches.[14] The

package MLwiN allows for a multilevel analysis in the Bayesian way, it is based on a combination of Gibbs sampling and Metropolis-Hastings sampling. Finally, recently an experimental SAS procedure, called PROC MCMC, was launched in version 9.2 which uses the Metropolis-Hastings approach.[15]

In all Bayesian packages we used non-informative priors for all the regression coefficients, i.e. a normal distribution with mean 0 and a large variance (10^4). The random effect is assumed to follow a normal distribution and the variance of the random effects is given a uniform prior distribution between 0.01 and 100. The total number of iterations for the binary outcome was 10,000 with a burn-in part 3,000. For the ordinal model, the total number of iterations was 100,000 and the size of the burn-in part was 10,000.

For the Bayesian approaches, we checked convergence with the Brooks-Gelman-Rubin (BGR) method.[16] This method compares within-chain and between-chain variability for multiple chains starting at over-dispersed initial values. Convergence of the chain is indicated by a ratio close to 1.[14]

Analysis

We fitted binary and ordinal logistic random effects regression models to the IMPACT data, using the different statistical packages. All packages are able to deal with the binary logistic random effects model; MIXOR, MLwiN, NLMIXED, WinBUGS and SAS MCMC are able to analyze ordinal multilevel data. Syntax codes are provided in appendix 1.

We compared the packages with respect to the estimates of the parameters and the time needed to arrive at the final estimates. Further, we compared what extra facilities the software offers, what output is shown and how easy to use the program is. Finally, we looked at the flexibility of the software, whether it was possible to vary the model assumptions made in (1) and (2) such as replacing the logit link by other link functions, i.e. probit and log(-log) link functions or relaxing the assumption of normality for the random effects.

Results

Descriptive statistics

From the 9,205 patients in the original database, we excluded patients with missing 6 months GOS (n=484), only partial information on the GOS (n=35), missing age (n=2) or younger than 14 (n=175). This resulted in 8,509 patients in 231 centers in the analysis, of whom 2,396 (28%) died and 4,082 (48%) had an unfavourable outcome six months after injury (Table 4.1). The median age was 30 (interquartile range 21–45) years, 3522 patients (41%) had a motor score of 3 or lower (none, extension or abnormal flexion), and 1,989 patients (23%) had bilateral non-reactive pupils. The median number of patients per center was 19, ranging from 1 to 425.

Table 4.1 IMPACT study: Descriptive statistics of the study population

	TINT	**TIUS**	**SLIN**	**SAP**	**PEG**
Year of study	1992–1994	1991–1994	1994–1996	1995–1997	1993–1995
No. of patients	1131	1155	409	924	1574
No. of centers	50	36	50	57	29
Outcome(GOS)					
Dead	278(25%)	225(22%)	94(23%)	212(23%)	362(24%)
Vegetative	44(4%)	42(4%)	14(3%)	24(3%)	114(8%)
Severe disability	134(12%)	128(12%)	69(17%)	142(16%)	298(20%)
Moderate disab.	171(15%)	180(17%)	84(21%)	174(19%)	374(25%)
Good recovery	491(44%)	466(45%)	148(36%)	367(40%)	362(24%)
Predictor(age)					
Median(IQ range)	30(21–45)	30(23–41)	28(21–43)	32(23–47)	27(20–38)
Predictor(motor)					
None	5(0%)	9(1%)	0(0%)	141(15%)	475(32%)
Extension	136(12%)	143(14%)	55(13%)	123(13%)	180(12%)
Abnormal flexion	237(21%)	132(13%)	91(22%)	143(16%)	165(11%)
Normal flexion	327(29%)	300(29%)	127(31%)	223(24%)	334(22%)
Localises	384(34%)	406(39%)	134(33%)	286(31%)	309(21%)
Obeys command	29(3%)	51(5%)	2(1%)	0(0%)	47(3%)
Untestable	0(0%)	0(0%)	0(0%)	3(0%)	0(0%)
Predictor(pupil)					
Both side positive	806(72%)	703(68%)	315(77%)	619(67%)	784(52%)
One side positive	177(16%)	118(11%)	79(19%)	178(19%)	156(10%)
Both side negative	135(12%)	220(21%)	15(4%)	122(13%)	570(38%)

TINT = Tirilizad International (RCT), **TIUS** = Tirilizad US (RCT), **SLIN** = International Selfotel trial (RCT), **SAP** (RCT), Coma Data Bank (observational study), **SKB** = Bradycor SKB (RCT), **EBIC** = European Brain Injury Consortium Core

Binary model

Fitting the dichotomous model in the different packages gave similar results (Table 4.2). For the frequentist approaches the R-function lmer, the SAS procedures GLIMMIX and NLMIXED, and the programs MLwiN and MIXOR provided almost the same results for the variance of the random effects and fixed effects. One example is age, with estimated coefficients of 0.623, 0.618, 0.623, 0.611 and 0.625 respectively in the different programs and all estimated SDs being close to 0.028. Also the variance of the random effects was estimated similar: 0.101, 0.107, 0.102, 0.095 and 0.105, respectively.

The Bayesian programs WinBUGS MLwiN and SAS procedure MCMC gave similar posterior means and these were also close to the MLEs obtained from the frequentist software. For example, the posterior mean (SD) of the regression coefficient of age was 0.625 (0.029) 0.626 (0.028) and 0.630 (0.025) for MLwiN, WinBUGS and SAS procedure MCMC respectively. The posterior mean of the variance of the random effects was esti-

HIT I	**UK4**	**TCDB**	**SKB**	**EBIC**	**HIT II**	**Total**
1987–1989	1986–1988	1984–1987	1996	1995	1989–1991	
351	988	667	139	1005	852	8509
6	4	4	31	67	21	231
99(28%)	359(45%)	264(44%)	34(27%)	281(34%)	188(23%)	2396(28%)
10(3%)	13(2%)	34(6%)	6(5%)	18(2%)	32(4%)	351(4%)
62(18%)	146(19%)	95(16%)	30(24%)	123(15%)	108(13%)	1335(16%)
64(18%)	130(16%)	104(17%)	27(21%)	159(19%)	199(24%)	1666(20%)
115(33%)	143(18%)	107(18%)	29(23%)	241(29%)	292(36%)	2761(32%)
34(21–47)	36(22–55)	26(21–40)	27(20–39)	37.5(24–59)	33(22–49)	30(21–45)
122(35%)	113(14%)	136(23%)	34(27%)	150(18%)	210(26%)	1395(16%)
41(12%)	85(11%)	107(18%)	22(18%)	80(10%)	70(9%)	1042(12%)
45(13%)	37(5%)	74(12%)	14(11%)	55(7%)	92(11%)	1085(13%)
56(16%)	141(18%)	122(20%)	16(13%)	113(14%)	181(22%)	1940(23%)
77(22%)	191(24%)	113(19%)	21(17%)	182(22%)	199(24%)	2302(27%)
0(0%)	30(4%)	21(4%)	2(2%)	99(12%)	8(1%)	289(3%)
9(3%)	194(25%)	31(6%)	17(14%)	143(18%)	59(7%)	456(5%)
232(66%)	427(54%)	300(50%)	70(56%)	535(65%)	585(71%)	5376(63%)
53(15%)	115(15%)	55(9%)	35(28%)	79(10%)	99(12%)	1144(13%)
65(19%)	249(32%)	249(41%)	21(17%)	208(25%)	135(17%)	1989(23%)

PEG (RCT), **HIT I** = HIT I Nimodipine (RCT), **UK4** = UK Four Center Study (observational study), **TCDB** = Traumatic data study (observational study), **HIT II** = HIT II Nimodipine (RCT).

mated as 0.113 0.119 and 0.160 respectively with the standard deviation being close to 0.30. Hence, for the Bayesian methods the standard deviations for the posterior mean were somewhat higher than the frequentist standard errors. This is due to the fact that the Bayesian method acknowledges all uncertainty in the model by averaging over a prior distribution, while frequentist methods consider the parameters fixed values.

The random effects estimates of the 231 centers could easily be derived from R, SAS NLMIXED, GLIMMIX, MLwiN, and WinBUGS. The estimates from all packages were quite similar, for example the Pearson correlation for the estimated random effects from WinBUGS and R was 0.9999.

Table 4.2 Results from the binary models

		R(lme4)		**GLIMMIX**		**NLMIXED**		**MLwiN-Freq**	
computing time		8s		3s		24min		2s	
Random Effects*		Variance: 0.101		Variance: 0.107(0.027)		Variance: 0.102(0.027)		Variance: 0.095(0.024)	
	Covar.	*Coef*	*SE*	*Coef*	*SE*	*Coef*	*SE*	*Coef*	*SE*
FIXED EFFECTS	Inter	0.014	0.114	0.014	0.114	0.014	0.114	0.013	0.113
	Pupil2	0.656	0.074	0.650	0.074	0.656	0.075	0.643	0.074
	Pupil3	1.404	0.069	1.392	0.069	1.404	0.070	1.376	0.068
	Age	0.623	0.028	0.618	0.028	0.623	0.028	0.611	0.028
	Motor2	0.618	0.106	0.612	0.105	0.618	0.106	0.608	0.104
	Motor3	0.154	0.097	0.153	0.097	0.154	0.097	0.150	0.096
	Motor4	0.782	0.086	0.775	0.086	0.782	0.087	0.764	0.085
	Motor5	1.404	0.088	1.394	0.088	1.404	0.089	1.376	0.087
	Motor6	1.591	0.166	1.577	0.166	1.591	0.167	1.559	0.165
	Motor9	0.534	0.136	0.529	0.136	0.534	0.136	0.523	0.134
	Trial2	0.073	0.125	0.071	0.126	0.073	0.126	0.075	0.124
	Trial3	0.218	0.139	0.216	0.139	0.218	0.139	0.214	0.138
	Trial4	0.192	0.116	0.189	0.117	0.192	0.117	0.189	0.115
	Trial5	0.107	0.114	0.107	0.115	0.107	0.115	0.103	0.113
	Trial6	0.039	0.173	0.039	0.174	0.039	0.174	0.038	0.172
	Trial7	0.686	0.170	0.680	0.171	0.686	0.170	0.678	0.168
	Trial8	0.672	0.176	0.665	0.177	0.672	0.176	0.658	0.173
	Trial9	0.373	0.231	0.368	0.231	0.373	0.232	0.366	0.229
	Trial10	0.090	0.123	0.090	0.123	0.090	0.123	0.089	0.122
	Trial11	0.239	0.125	0.233	0.126	0.238	0.127	0.236	0.125

* The variance of the random effects with its standard error is given.

Ordinal model

Fitting the ordinal model in the different packages also gave similar results (Table 4.3). For the frequentist approach, the two SAS procedures GLIMMIX and NLMIXED, the packages MLwiN and MIXOR gave very similar estimates for the fixed effects parameters and the variance of the random effects. The estimate (SD) of e.g. the regression coefficient of age was 0.588 (0.023), 0.591(0.023), 0.581 (0.022) and 0.591 (0.027), respectively. The estimate of the variance (SD) of the random effects were 0.090 (0.021), 0.085 (0.020), 0.079 (0.018), and 0.098 (0.045), respectively.

MIXOR		**MLwiN-Bayesian**		**MCMC**		**WinBUGS**	
5s		15min		37h		53min	
Variance: 0.105(0.051)		Variance: 0.113(0.030)		Variance: 0.160(0.034)		Variance: 0.119(0.030)	
Coef	*SE*	*Coef*	*SE*	*Coef*	*SE*	*Coef*	*SE*
0.006	0.127	0.003	0.110	0.103	0.099	0.026	0.115
0.657	0.089	0.656	0.075	0.666	0.071	0.659	0.075
1.401	0.074	1.406	0.068	1.424	0.069	1.410	0.069
0.625	0.029	0.625	0.029	0.630	0.029	0.626	0.028
0.622	0.125	0.617	0.104	0.654	0.103	0.623	0.106
0.150	0.101	0.158	0.096	0.131	0.096	0.152	0.098
0.781	0.105	0.786	0.084	0.757	0.076	0.781	0.088
1.403	0.106	1.412	0.086	1.394	0.070	1.409	0.090
1.595	0.188	1.602	0.168	1.593	0.166	1.598	0.168
0.541	0.152	0.536	0.136	0.533	0.129	0.535	0.136
0.105	0.135	0.081	0.121	0.007	0.115	0.061	0.129
0.202	0.135	0.210	0.136	0.240	0.139	0.222	0.140
0.173	0.100	0.195	0.115	0.116	0.128	0.184	0.117
0.091	0.127	0.099	0.114	0.184	0.112	0.119	0.117
0.016	0.216	0.046	0.172	0.049	0.188	0.034	0.175
0.727	0.148	0.680	0.172	0.755	0.182	0.693	0.172
0.539	0.140	0.652	0.172	0.744	0.198	0.682	0.181
0.361	0.229	0.368	0.232	0.408	0.223	0.382	0.234
0.097	0.113	0.083	0.118	0.149	0.125	0.099	0.124
0.217	0.147	0.239	0.123	0.128	0.121	0.225	0.127

For the Bayesian approaches, WinBUGS and MLwiN produced similar results as the frequentist approaches. The posterior mean of the regression coefficient of age with WinBUGS was 0.551 and 0.592 in MLwiN, with SD = 0.023 in both cases. (same as the SAS frequentist result). The posterior mean of the variance of the random effects was 0.096 in WinBUGS and 0.093 in MLwiN and for both SD = 0.022, very close to the frequentist estimates. We stopped running the SAS MCMC procedure after 2,000 iterations because this already took 19 hours and the results of the last 1,000 iterations were far from having converged.

Table 4.3 Results from the ordinal models

		GLIMMIX		**NLMIXED**		**MLwiN-Freq**		**MIXOR**	
Computing time		6s		38min		1min		8s	
Random Effects*		Variance: 0.090 (0.021)		Variance: 0.085 (0.020)		Variance: 0.079 (0.018)		Variance: 0.098 (0.045)	
	covar.	*Coef*	*SE*	*Coef*	*SE*	*Coef*	*SE*	*Coef*	*SE*
FIXED EFFECTS	Pupil2	0.702	0.062	0.705	0.062	0.692	0.062	0.707	0.082
	Pupil3	1.396	0.057	1.401	0.057	1.378	0.056	1.403	0.062
	Age	0.588	0.023	0.591	0.023	0.581	0.022	0.591	0.027
	Motor2	0.275	0.083	0.277	0.083	0.274	0.083	0.281	0.092
	Motor3	0.295	0.081	0.296	0.081	0.288	0.081	0.293	0.078
	Motor4	0.843	0.072	0.846	0.072	0.833	0.072	0.842	0.074
	Motor5	1.365	0.073	1.369	0.073	1.347	0.072	1.369	0.080
	Motor6	1.565	0.133	1.572	0.137	1.557	0.133	1.578	0.161
	Motor9	0.628	0.112	0.630	0.111	0.622	0.111	0.630	0.115
	Trial2	0.066	0.106	0.067	0.107	0.067	0.104	0.054	0.114
	Trial3	0.251	0.116	0.252	0.117	0.248	0.116	0.240	0.115
	Trial4	0.120	0.098	0.122	0.099	0.128	0.097	0.107	0.083
	Trial5	0.190	0.097	0.189	0.097	0.185	0.095	0.200	0.107
	Trial6	0.051	0.147	0.051	0.146	0.045	0.145	0.074	0.143
	Trial7	0.768	0.144	0.772	0.142	0.770	0.142	0.780	0.161
	Trial8	0.900	0.149	0.901	0.148	0.885	0.146	0.993	0.179
	Trial9	0.339	0.193	0.341	0.190	0.338	0.192	0.337	0.185
	Trial10	0.264	0.102	0.265	0.102	0.263	0.101	0.269	0.090
	Trial11	0.044	0.105	0.047	0.106	0.042	0.104	0.029	0.093
	Inter1	1.188	0.098	1.190	0.098	1.169	0.097	1.186	0.096
	Inter2	0.930	0.097	0.931	0.098	0.914	0.096	0.927	0.096
	Inter3	0.040	0.097	0.040	0.098	0.036	0.096	0.035	0.098
	Inter4	1.025	0.097	1.026	0.098	1.012	0.096	1.032	0.099

* The variance of the random effects with its standard error

The random effects for the 231 centers from the SAS procedure NLMIXED, MLwiN (both frequentist and Bayesian) and WinBUGS were again quite similar with correlation again virtually 1.

Usability, flexibility and speed

The packages greatly differed in their usability. For instance, SAS is based on procedures for which certain options can be turned on and off. Understanding the different options in the statistical SAS procedures often requires a great deal of statistical

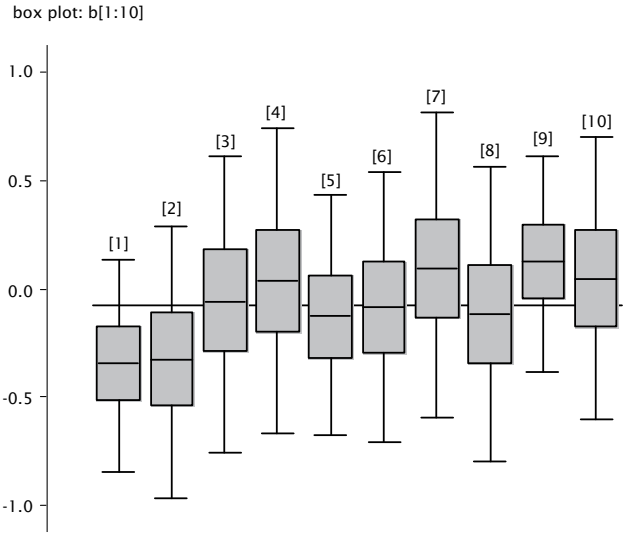
MLwiN-Bayesian		**WinBUGS**	
60min		32h	
Variance: 0.093 (0.022)		Variance: 0.096 (0.022)	
Coef	*SE*	*Coef*	*SE*
0.708	0.063	0.703	0.062
1.406	0.057	1.405	0.057
0.592	0.023	0.551	0.023
0.282	0.086	0.276	0.082
0.292	0.084	0.305	0.080
0.843	0.074	0.847	0.072
1.367	0.076	1.368	0.073
1.574	0.138	1.567	0.137
0.629	0.112	0.640	0.112
0.054	0.113	0.075	0.109
0.260	0.120	0.245	0.117
0.111	0.103	0.121	0.099
0.204	0.103	0.177	0.098
0.062	0.149	0.083	0.147
0.781	0.145	0.783	0.144
0.917	0.151	0.888	0.150
0.352	0.195	0.343	0.192
0.275	0.105	0.302	0.102
0.033	0.111	0.030	0.106
1.208	0.111	1.186	0.098
0.949	0.111	0.928	0.098
0.056	0.109	0.042	0.100
1.012	0.109	1.007	0.103

background since the procedures are often based on the most advanced and computationally powerful methods. Also SAS data management is quite powerful but is also associated with a steep learning curve. The SAS procedures NLMIXED and MCMC offer some programming facilities. The package R has gained a lot of attention in the last decade and is becoming increasingly popular among statisticians and non-statisticians. It requires programming but has many basic functions, and also the graphics are nicely incorporated. WinBUGS is the most popular general purpose package for Bayesian analyses. It is extremely popular with now more than 30,000 registered users. The package allows for a great variety of analyses using a programming language that resembles that of R. WinBUGS requires about the same programming skills as R. MIXOR needs no programming but provides very limited output. MLwiN has a clear and intuitive interface to specify a random effects model, but lacks a simple syntax file structure.

All packages require a good statistical background of the multilevel approach in order to analyze such data in a reliable manner.

The packages also differ in what they offer as standard output besides the parameter estimates. WinBUGS allows for the most extensive output, including diagnostic plots for model evaluation and plots of the individual random center effects. All packages except MIXOR can provide estimates of the random effects. In Figure 4.1 we show the box plots of the sampled random effects in WinBUGS for the first 10 centers of the binary logistic random effects model applied to the IMPACT data. Of course with packages

Figure 4.1 Box plot of a sample of the random effects (for center 1 to 10), that can be directly derived from WinBUGS. Each box represents a center with its random effects estimate and confidence interval.



like SAS and R the output of the statistical procedures can be saved and then processed by some other procedure or function to deliver the required graph or additional diagnostic analysis. For example, Figure 4.2 is produced with R and shows the histogram of the random effects of the binary IMPACT logistic random effects model.

Flexibility differs somewhat in the packages. All packages could handle a probit model and a log(-log) model except lme4 in R. But, only WinBUGS allows changing the distribution of the random effects into a t distribution or uniform distribution. Table 4.4 shows that WinBUGS has the largest flexibility in changing the model assumptions.

Figure 4.2 Histogram of the random effects in the binary model in R

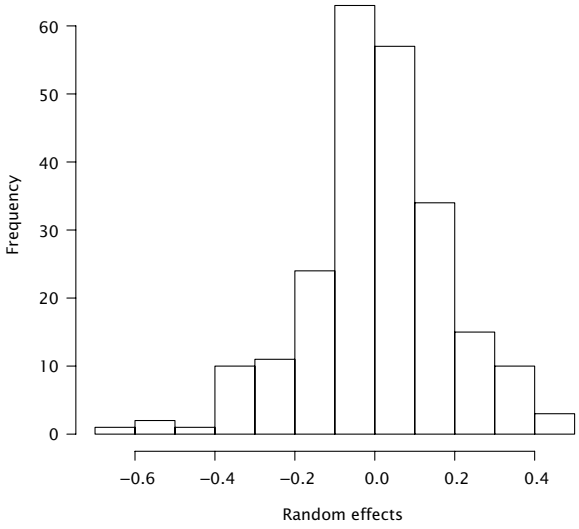


Table 4.4 Extra abilities of different packages

Package	Program/function	Link function		Obtaining the random effects	Other than normal random effects
		Probit model	Log(-log) model		
R	LME4			X	
MIXOR	MIXOR	X	X		
SAS	NLMIXED	X	X	X	
	GLIMMIX	X	X	X	
	MCMC	X	X	X	
MLwiN	MLwiN	X	X	X	
WinBUGS	WinBUGS	X	X	X	X

The speed of the computations varied widely. The computations were done on IBM t61 laptop with an Intel Core(TM) 2 Duo T7250 2.0 GHz CPU and 2 GB internal memory. For the binary logistic random effects models all frequentist approaches took only seconds, except for SAS NLMIXED which needed 24 minutes. The MLwiN procedure was the fasted, but GLIMMIX and MIXOR were almost as fast. The Bayesian approaches were considerably slower, which is known. MCMC sampling is time consuming, but also the checking for convergence in a Bayesian sense is far more difficult than in a frequentist sense. MLwiN was again the winner, but we considered all computation times as acceptable, except for the SAS MCMC procedure which took 37 hours for the binary model. Similar findings were obtained for the ordinal logistic random effects model, but the time to converge compared to the binary case increased considerably for some software. Now the winner in the frequentist software was GLIMMIX closely followed up by MIXOR. For the Bayesian software, MLwiN was again the winner, much faster than WinBUGS. The SAS procedure MCMC never got to convergence (we stopped it).

Discussion

In this study we compared different software implementations of logistic random effects regression models. We found that although results were very similar between the eight implementations, there are considerable variations in flexibility, computing time and usability.

Parameter estimates (both coefficients and standard errors of the fixed effects, and the random effects variance and estimates) were similar in all packages, for both the binary and ordinal logistic random effects models. In general the standard deviations of the posterior mean from a Bayesian approach were somewhat larger than the standard errors estimated from the frequentist approaches.

Frequentist and Bayesian approach

In the frequentist approach, probability is defined as a limiting relative frequency. That is, the probability of an event is the limit of the relative frequency of that event in a large number of studies. Further, in frequentist statistics one estimates the unknown but fixed model parameter θ. The estimate of θ is obtained by maximizing the likelihood. Prediction is done given the estimated θ and the uncertainty of the prediction is based on the sampling property of the estimated value of θ.[17]

In the Bayesian approach the parameter θ is given a probability distribution which expresses our prior knowledge about that parameter. There is still a true value for the parameter, but the parameter becomes stochastic because of our uncertainty. The Bayesian paradigm is based on Bayes' Theorem which combines a prior belief or probability with the actual observed data to arrive at an updated posterior probability.[18] The probability calculations heavily involve integration. Because the integrals are high

dimensional, the Bayesian approach was for about two centuries impossible to be used for real-life problems.

As can be deduced from above, the two approaches differ in their numerical approach: the frequentist involving maximization routines and the Bayesian involving numerical techniques that perform integration. Since 1989 a powerful class of numerical procedures, called Markov Chain Monte Carlo (MCMC) techniques,[19] were launched which revolutionalized the Bayesian approach. The MCMC approach is based on a sampling approach, i.e. the integral is approximated by Monte-Carlo sampling.[20] In fact there are two major classes of MCMC techniques: Gibbs sampling and Metropolis-Hastings sampling.

The Bayesian approach involves a prior distribution on the parameters and a likelihood. The posterior estimates depend on these two components. With the same data (likelihood), the posterior estimates may change heavily if different informative priors are used. On the other hand, if the prior is non-informative, such as flat or a normal distribution with large variance, the posterior estimates only depend on the likelihood. In this situation, the posterior modes are quite similar to the classic maximum likelihood estimates. This happened in our study as we used non-informative prior distributions of all the parameters. This is one of the reasons why the results from frequentist and Bayesian approaches are very similar.

It should be realized that logistic random effects models involve integration with both the frequentist and the Bayesian approach. In fact, models (1) and (2) yield conditional likelihoods, conditional on the values of the random effects. Since the random effects are not known, the marginal likelihood is determined which is the likelihood integrated over the distribution of the random effects. Random effects estimates from frequentist methods are often referred to as Empirical Bayes estimates.

Performance of each package

Although the parameter estimates were very similar between the eight implementations, we found considerable variations in computation time, usability and flexibility.

The frequentist approaches were all very fast, taking only seconds, with the SAS NLMIXED procedure as a major exception. Overall, the SAS procedure GLIMMIX, the program MIXOR and the package MLwiN were the winners. The fact that the SAS procedure NLMIXED took a much longer time likely has to do with the fact that is a general purpose program suitable for fitting a variety of complicated random effects models.

The Bayesian approaches were invariably slower than the frequentist approaches, which is due to the computational intensive MCMC approach and that convergence on chains is much harder to show than in a classical frequentist sense. However, we believe that the slowness of the Bayesian procedures here has also much to do with the size of the study and that the data management of the Bayesian procedures is not yet optimal. Nevertheless, the time to get convergence in WinBUGS for both the binary

(53 min) and the ordinal model (37 hours) will definitely prevent the user to do much on exploratory statistical research. MLwiN is for sure the winner for fitting binary and ordinal Bayesian logistic random effects models, taking only 15 minutes for the binary model and 60 minutes for the ordinal model. It is our software of choice from a computational point of view, if we focus on multilevel modeling only.

The SAS procedure MCMC is an experimental procedure available from version 9.2 onwards. It is a general purpose Markov Chain Monte Carlo simulation procedure that is designed to fit Bayesian models.[15] In our experience, however, this procedure was inefficient in dealing with mixed models. It was far too time consuming and difficult to converge to get stable estimates for both the regression coefficients and the variance of the random effects. At this moment, we cannot recommend this SAS procedure for fitting logistic random effects regression models.

The packages differ much in nature, with e.g. based on procedures with options to switch on and off and other software such as WinBUGS which is in fact a programming language. Which package to prefer from the aspect of usability is difficult to say since it very much depends on the nature of the user but also whether the logistic random effects model fitting is a stand-alone exercise. We know that in practice this is often not the case since we would like to process output of such an analysis to produce e.g. nice graphs. From this viewpoint MIXOR and WinBUGS score lower since they need the user to switch to other software, such as R, to produce additional output or better quality graphs. However, in the recent years other versions of WinBUGS have been created providing links to R, such as R2WinBUS. We did not consider these new software developments, however.

Regarding flexibility in statistical modelling, WinBUGS scores highest. Different distributions for the random effects (e.g. gamma, uniform, t-distribution) and different link functions such as probit and log(-log) model are possible. Different link functions are also possible in the SAS procedures GLIMMIX and NLMIXED, but none of these two packages allow other than normal distributions for the random effects. Note that the binary logistic random effect model was superior to the probit and log(-log) models according to Akaike Information Criterion (using GLIMMIX).

Other considerations

In this study we considered models with only a random intercept. Also random slopes (allowing the effect of the covariates to vary between the centers) or a cross-classified random effects structure such as patients in centers and in studies could be considered. The purpose of using the simplest random effects model was to show the different performances of packages quickly and effectively. The packages may act more differently dealing with more complex mixed models, but we consider that beyond the scope of this study.

Conclusion

We conclude that in our study the parameter estimates from logistic random effects regression models were not influenced by the choice of the statistical package. Therefore the choice for a certain statistical implementation should be determined by other factors, such as speed and desired flexibility. Based on our study, if there is no prior acquaintance with a certain package and preference is given to a frequentist approach, we can recommend MLwiN, the function lmer in R and the SAS procedure GLIMMIX. For a Bayesian implementation, we would recommend first MLwiN because of its efficiency; if the user is also interested in more comprehensive analyses than only multilevel modelling then he/she could choose WinBUGS.

References

1. Rasbash J: What are multilevel models and why should I use them? [*http://www.cmm.bristol.ac.uk/learning-training/multilevel-models/what-why.shtml]*
2. Molenberghs G: and Verbeke G. *Models for Discrete Longitudinal Data*. Berlin: Springer; 2005
3. Peter C, Jack V, David A: Comparing hierarchical modeling with traditional logistic regression analysis among patients hospitalized with acute myocardial infarction: should we be analyzing cardiovascular outcomes data differently? *Am Heart J* 2003, 145:27-35.
4. Marmarou A, Lu J, Butcher I, McHugh GS, Mushkudiani NA, Murray GD, Steyerberg EW: IMPACT Database of Traumatic Brain Injury: Design and Description. *J Neurotrauma* 2007, 24:239-250.
5. Maas AI, Marmarou A, Murray GD, Teasdale SG, Steyerberg EW: Prognosis and Clinical Trial Design in Traumatic Brain Injury: The IMPACT Study. *J Neurotrauma* 2007, 24:232-238
6. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS: Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. P*LoS Medicine* 2008, 5:1251-1261
7. Goldstein H: *Multilevel Statistical Models*, 2nd Edition. London: Edward Arnold; 1995
8. Donald H, Robert DG: MIXOR: a computer program for mixed-effects ordinal regression analysis. *Comput Meth Programs Biomed* 1996, 49:157-176
9. Bates D, Maechler M: *Package 'lme4'*. [http://lme4.r-forge.r-project.org/]. 2009
10. The NLMIXED procedure. *SAS/STAT User's Guide*, Version 9.2. 2009
11. The GLIMMIX procedure. *SAS/STAT User's Guide*, Version 9.2. 2009
12. Rasbash J, Steele F, Browne WJ, Goldstein H: *A User's Guide to MLwiN* (version 2.10). 2004
13. Lee PM: Bayesian *Statistics: An Introduction*. New York:Oxford University Press; 1989
14. Spiegelhalter D, Thomas A, Best N, Lunn D: *WinBUGS User Manual* (version 1.4.3). 2007
15. The MCMC procedure. *SAS/STAT User's Guide*, Version 9.2. 2009
16. Brooks SP and Gelman A: Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics* 1998, 7:434-455
17. Feller W: *An introduction to Probability Theory and its Applications*. New York: Wiley; 1957
18. Bernardo JM. Smith AFM: *Bayesian Theory*. London: Wiley; 1994
19. Gelfand A, Smith A: Sampling based approaches to calculating marginal densities. *J American Statist Assoc* 1990, 85:398–409.
20. Ripley B: *Stochastic Simulation*. New York: Wiley; 1987

Appendix 1

Note: programs for MIXOR and MLwiN are not listed because they do not provide straightforward syntax.

Variable coding

Variable label	variable name	coding
Age	Age	Continuous
Motor score	Motor	1='none' 2='extension' 3='abnormal flexion' 4='normal flexion' 5='localises' 6='obeys command' 9='untestable'
Pupil reactivity	Pupil	1='both side positive' 2='one side positive' 3='both side negative'
Unfavorable outcome	D_unfav	0='favorable' 1='unfavorable'
Glasgow Outcome Scale	GOS	1='dead' 2='vegetative status' 3='severe disability' 4='moderate disability' 5='good recovery'
Study	Trial	Using dummy variables

Binary logistic random effects model

SAS procedure nlmixed

```
proc nlmixed data=aa_std tech=newrap qpoints=5;
parms beta0=0, beta1=0, beta2=-2, beta3=-1, beta4=1, beta5=1, beta6=1, beta7=1,
beta8=1, beta9=1, beta10=1, beta11=1, beta12=1, beta13=1, beta14=1, beta15=1,
beta16=1, beta17=1, beta18=1, beta19=1, s2b=1;
eta=beta0+beta2*pupil2+beta3*pupil3+beta1*age+beta4*motor2+beta5*motor3+beta6
*motor4+beta7*motor5+beta8*motor6+beta9*motor9+beta10*trial2+beta11*trial3+bet
a12*trial4+beta13*trial5+beta14*trial6+beta15*trial7+beta16*trial8+beta17*trial9+beta1
8*trial10+beta19*trial11+b0;
mu=exp(eta)/(1+exp(eta));
model d_unfav ~ binary(mu);
random b0 ~ normal(0,s2b) subject=center_num out=bin;
run;
```

SAS procedure glimmix

```
proc glimmix data=aa_std;
class center_num;
model d_unfav(event=last)=pupil2-pupil3 age motor2-motor6 motor9 trial2-trial11/
dist=binary solution;
random intercept/ subject=center_num;
output out=i1 pred=p resid=r pred(NOBLUP)=p1;
run;
```

SAS procedure mcmc

```
proc mcmc data=aa_std outpost=postout seed=332786 nmc=10000 nbi=3000
monitor=(beta0-beta19 s2b);
array delta[231];
parms beta0 0 beta1 0  beta2 -2  beta3 -1 beta4 1 beta5 1 beta6 1 beta7 1 beta8 1
beta9 1 beta10 1;
parms beta11 1 beta12 1 beta13 1 beta14 1 beta15 1 beta16 1 beta17 1 beta18 1 beta19
1;
parms s2b 1;
%group_parms(delta,231, 20, 1);
prior beta:~normal(0,var=10000);
prior delta:~normal(beta0,var=s2b);
prior s2b~uniform(0.01,100);
w=beta1*age+beta2*pupil2+beta3*pupil3+beta4*motor2+beta5*motor3+beta6*motor4
+beta7*motor5+beta8*motor6+beta9*motor9+beta10*trial2+beta11*trial3+beta12*
trial4+beta13*trial5+beta14*trial6+beta15*trial7+beta16*trial8+beta17*trial9+beta18*
trial10+beta19*trial11;
pi=logistic(w+delta[center_num]);
model d_unfav ~ binary(pi);
run;
```

R fuction lmer

```
glmer(d_unfav~pupil2+pupil3+age+motor2+motor3+motor4+motor5+motor6+motor
9+trial2+trial3+trial4+trial5+trial6+trial7+trial8+trial9+trial10+trial11+(1|center_num),
nAGQ=1, family=binomial, data=total)
```

WinBUGS

```
model{
for (i in 1:N) {
agec[i]<-(age[i]-mean(age[]))/sd(age[]) #center age
logit(mu[i])<-beta[1]*pupil2[i]+beta[2]*pupil3[i]+beta[3]*agec[i]+beta[4]*motor2[i]+
beta[5]*motor3[i]+beta[6]*motor4[i]+beta[7]*motor5[i]+beta[8]*motor6[i]+beta[9]*
motor9[i]+beta[10]*trial2[i]+beta[11]*trial3[i]+beta[12]*trial4[i]+beta[13]*trial5[i]+
beta[14]*trial6[i]+beta[15]*trial7[i]+beta[16]*trial8[i]+beta[17]*trial9[i]+beta[18]*
trial10[i]+beta[19]*trial11[i]+b[center[i]]
d_unfav[i]~dbin(mu[i],1)
}

for (i in 1:Ncenter){
b[i]~dnorm(beta0,tau)
}

#the following prior distributions were chosen
beta0~dnorm(0,0.0001)
for (j in 1: 19){
beta[j]~dnorm(0,0.0001)
}
sigma~dunif(0.01,100)
tau <- pow(sigma,-1)
 }
```

Ordinal logistic random effects model

SAS procedure nlmixed

```
proc nlmixed data=aa_std tech=newrap qpoints=5;
parms beta1=0, beta2=-2, beta3=-1, beta4=1, beta5=1, beta6=1, beta7=1, beta8=1,
beta9=1, beta10=1, beta11=1, beta12=1, beta13=1, beta14=1, beta15=1, beta16=1,
beta17=1, beta18=1, beta19=1, s2b=1 i1=1 i2=1 i3=1, i4=1;
bounds i2>0,i3>0,i4>0;
eta=beta2*pupil2+beta3*pupil3+beta1*age+beta4*motor2+beta5*motor3+beta6*
motor4+beta7*motor5+beta8*motor6+beta9*motor9+beta10*trial2+beta11*trial3+
beta12*trial4+beta13*trial5+beta14*trial6+beta15*trial7+beta16*trial8+beta17*trial9+
beta18*trial10+beta19*trial11+b0;
if (gos=1) then p=1/(1+exp(-(i1+eta)));
else if (gos=2) then p=(1/(1+exp(-(i1+i2+eta))))-(1/(1+exp(-(i1+eta))));
else if (gos=3) then p=(1/(1+exp(-(i1+i2+i3+eta))))-(1/(1+exp(-(i1+i2+eta))));
```

```
else if (gos=4) then p=(1/(1+exp(-(i1+i2+i3+i4+eta))))-(1/(1+exp(-(i1+i2+i3+eta))));
else p=1-(1/(1+exp(-(i1+i2+i3+i4+eta))));
if (p > 1e-8) then ll = log(p);
else ll = -1e100;
model gos ~ general(ll);
random b0 ~ normal(0,s2b) subject=center_num out=ord;
estimate 'thresh1' i1;
estimate 'thresh2' i1+i2;
estimate 'thresh3' i1+i2+i3;
estimate 'thresh4' i1+i2+i3+i4;
run;
```

SAS procedure glimmix

```
proc glimmix data=aa_std;
class center_num;
model gos=pupil2-pupil3 age motor2-motor6 motor9 trial2-trial11/DIST=MULT
LINK=CLOGIT solution;
random intercept/ subject=center_num;
NLOPTIONS MAXIT=100;
output out=i1 pred=p resid=r pred(NOBLUP)=p1;
run;
```

WinBUGS

```
model{
for (i in 1:N) {
agec[i]<-(age[i]-mean(age[]))/sd(age[]) #center age
covc[i]<-beta[1]*pupil2[i]+beta[2]*pupil3[i]+beta[3]*agec[i]+beta[4]*motor2[i]+beta[5]
*motor3[i]+beta[6]*motor4[i]+beta[7]*motor5[i]+beta[8]*motor6[i]+beta[9]*motor9[i]
+beta[10]*trial2[i]+beta[11]*trial3[i]+beta[12]*trial4[i]+beta[13]*trial5[i]+beta[14]*
trial6[i]+beta[15]*trial7[i]+beta[16]*trial8[i]+beta[17]*trial9[i]+beta[18]*trial10[i]+
beta[19]*trial11[i]+b[center[i]]
for (j in 1:4){logit(f[i,j])<-a[j]+covc[i]}

#cumulative probability of response<=cutpoint
p[i,1]<-f[i,1];p[i,2]<-f[i,2]-f[i,1];p[i,3]<-f[i,3]-f[i,2];p[i,4]<-f[i,4]-f[i,3];p[i,5]<-1-f[i,4];
gos[i]~dcat(p[i,1:5])
}
for (i in 1:Ncenter){
b[i]~dnorm(a[1],tau)
}
```

```
a[1] ~ dnorm(0, 1.0E−06)I(,a[2])
a[2] ~ dnorm(0, 1.0E−06)I(a[1],a[3])
a[3] ~ dnorm(0, 1.0E−06)I(a[2],a[4])
a[4] ~ dnorm(0, 1.0E−06)I(a[3],)

for (j in 1: 19){
beta[j]~dnorm(0,0.0001)
}
sigma~dunif(0.01,100)
tau <- pow(sigma,−1)
```



III Prognostic models









5 Prognostic models in traumatic brain injury

Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AIR.
Early prognosis in traumatic brain injury: from prophecies to predictions.
Lancet Neurology 2010; 9:543-554.

'No head injury is too severe to despair of, nor too trivial to ignore' [Hippocrates]

Abstract

Traumatic Brain Injury (TBI) constitutes a heterogeneous disease, encompassing a broad spectrum of pathologies. Outcome can be highly variable, particularly in more severely injured patients. Despite the association of many variables with outcome, predictions are notoriously difficult. Multivariable analysis has identified age, clinical severity, Computerized Tomography abnormalities, systemic insults (hypoxia and hypotension) and laboratory parameters as relevant building blocks for combining variables into models to predict outcome in individual patients. A systematic literature search identified 16 studies reporting on prognostic models based upon admission characteristics; many of these showed shortcomings, which may partly explain the limited use of these models in clinical practice. Advances in statistical modelling and the availability of large datasets have facilitated the development of prognostic models with greater performance and generalizibility. Two prediction models are currently available, that have been developed on large datasets with state of the art methods, offering new opportunities. We see a great potential for use in clinical practice, in research, towards policy making and assessment of the quality of health care delivery. Continued development, refinement and validation is advocated together with assessment of the clinical impact of prediction models. Future directions should include the development of models to predict treatment response.

Introduction

Prognosis is the cornerstone of clinical medicine, since all diagnostic and therapeutic actions eventually aim to improve a subject's prognosis and outcome. Advances in statistical modelling and the availability of large databases have made it possible to consider diagnosis and prognosis nowadays in terms of probabilities rather than vague prophecies. Probability estimates can be applied towards clinical decision making, research and assessment of the quality of health care. Such quantitative estimates are of particular relevance to heterogeneous diseases such as Traumatic Brain Injury (TBI).

TBI poses a major public health problem with an estimated annual incidence of up to 500/100.000 and over 200 hospital admissions per 100.000 in Europe each year.[1,2] TBI is a heterogeneous disease in terms of cause, pathology, severity and prognosis. It poses diagnostic challenges and the heterogeneity makes it difficult to compare results between studies since case-mix and treatments may vary considerably.

Various outcomes can be considered in prediction research. A diagnostic perspective is taken in TBI studies assessing the probability of structural brain damage, the probability for developing an intracranial hematoma, or underpin recommendations for CT scanning.[3-5] A recent study identified a subset of children at such low risk for intracranial pathology that protection from unnecessary radiation exposure motivated not performing a CT scan.[6] These types of diagnostic outcomes are particularly relevant for patients with mild TBI. Predicting response to treatment would be highly relevant to patients in the intensive care setting, in whom intracranial pressure is monitored, but these have not (yet) been performed.

For patients with moderate and severe TBI, predicting clinical outcome is highly relevant. Typically, most studies performed have used mortality or functional outcome assessed with the Glasgow Outcome Scale (GOS)[7] as endpoint.

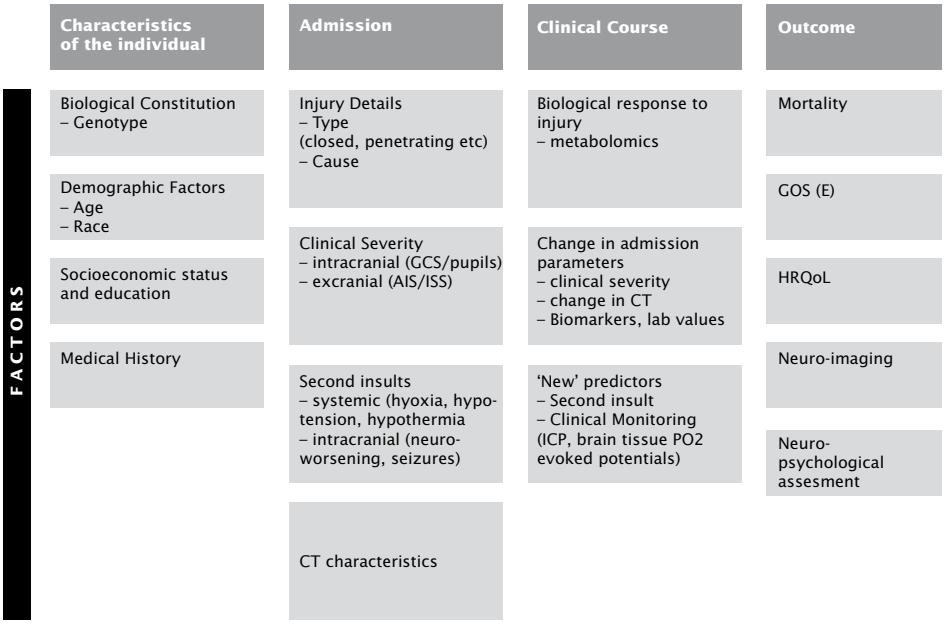
In this review, we focus on prediction of outcome in terms of mortality and functional outcome in patients with moderate and severe TBI.
We aim to:
- Describe the basics of prognostic analysis
- Review the current knowledge about traditional and newly recognized predictors for outcome in TBI
- Discuss prognostic modelling as a novel instrument in medicine
- Critically review prediction models in TBI
- Describe the applications for prognostic models in TBI
- Provide suggestions for the further development and improvement of prediction research in TBI.

Predictors of outcome

Much research has been conducted to identify early predictors of mortality and functional outcome as assessed by the Glasgow Outcome Scale (GOS) on admission after moderate or severe TBI. The GOS is usually dichotomized into good recovery and mild disability versus severe disability, vegetative state and mortality. This is a limitation since we cannot assume that predictors differentiate death from survival equally well as good recovery from worse outcomes. In this review we summarize findings from different studies using mortality and GOS by referring to 'outcome'.

For some predictors a large body of evidence exists, for others the relationship with outcome is less well established. Information obtained during the subsequent clinical course may further contribute to outcome prediction. An overview of the various components, or building blocks, of prognostic analysis is presented in Figure 5.1. This figure illustrates the complex relations between potential predictors, and highlights gaps in our knowledge (genomics, biomarkers).

Figure 5.1 Overview of building blocks of prognosis in Traumatic Brain Injury



Some basics of prediction research

Several steps in prediction research can be identified (panel 5.1).[8] First, the association between a single predictor and outcome can be studied in univariate analysis, relating a single predictor to the outcome of interest. Such an analysis is of limited value because it does not take the role of other confounding factors that may explain (part of) the observed association into account. Statistical analyses, such as logistic regression, are therefore required to adjust for confounders in the assessment of relative risks. Other statistical approaches to prognostic analysis include recursive partitioning (prediction trees) and neural network analysis. Odds Ratios (ORs) are commonly used to express the strength of prognostic effects. The relationship is statistically significant if the 95% confidence interval for the OR does not include the value 1. The OR does not

Panel 5.1 Steps in prognostic analysis in traumatic brain injury[8]

	Univariate analysis	Multivariable analysis	Prediction models
Aim	To estimate the relationship between a single predictor and outcome	To determine the prognostic value of a predictor, adjusting for confounding effects of other predictors.	To combine predictors into a model with the aim to estimate the risk of an outcome for individual patients
Limitations	Does not take the role of confounding factors into account that may explain (part of) the observed association	In individual patients predictors may influence outcome in opposite directions; does not take interactions or differential effects for specific subpopulations into account.	External validation essential to prove generalizibility outside of the development setting.
Performance measures	Sensitivity, specificity Positive predictive value, negative predictive value Odds ratio*	Odds ratio Relative Risk** R^2***	*Discrimination:* area under the receiver operating characteristic (AUC) *Calibration:* graphical representation Hosmer-Lemeshow goodness of fit test
Presentation	Tabular Graphical representation	Tabular Graphical representation	Web-based calculator Score chart

* *Sensitivity:* proportion of patients with the outcome that have the predictor (true positive) *Specificity:* proportion of patients without the outcome that do not have the predictor (true egative) *Positive predictive value (PPV)*: proportion of patients with the predictor that have the outcome *Negative predictive value (NPV):* proportion of patients without the predictor that do not have the outcome *Odds ratio (OR):* ratio of the odds for better versus poorer outcome in the presence of the parameter, compared to the odds in the absence of the parameter.

** *Relative risk (RR):* risk of outcome in group with the predictor versus group without the predictor

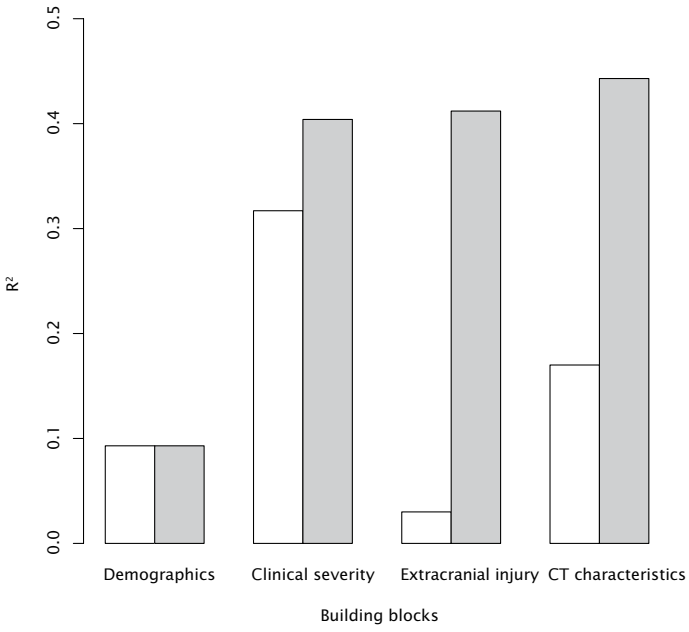
*** R^2*:* percentage of variability in outcome that is explained by the predictor. R^2 reflects predictive value better than OR, since also prevalence is taken into account.

account for the prevalence of a predictor. A predictor with a high OR but a very low prevalence is of limited predictive value. Predictive value is better reflected in measures such as explained variation (R^2).[9]

Admission characteristics

The prognostic strength of the main predictors in TBI is summarized in table 5.1. The prognostic value of the different building blocks for prognosis was quantified in the IMPACT data (N=8686)[10] (Figure 5.2). Clinical severity has the highest prognostic value (highest R^2), followed by CT characteristics, both separately and when these building blocks are added in the order of availability in clinical practice. The cumulative R^2 of the full model is 0.35.

Figure 5.2 Prognostic value of different building blocks of prognosis, expressed in univariate (white) and multivariate (grey) R^2, in the IMPACT dataset (n=8686)[10]



Demographic factors

Age is one of the strongest predictors of mortality and functional outcome in TBI, and many publications have shown that higher age is associated with poorer outcome. Most studies have analyzed the association between age and outcome with threshold values, using different thresholds varying from 30 to 60 years of age.[11-16]

Only a limited number of studies have analyzed the association between age and

Table 5.1 Strength of the association between predictors and outcome (full ordinal GOS) in TBI in the IMPACT database (n=8686)[10]

Predictor		Reference category	Univariate OR	Multivariate OR (adjusted for Age/ Motorscore/Pupils)
Demographics				
Age		25-75% i.q.r.	2.14 (2.00–2.28)	–
Gender		Male	1.01 (0.92–1.11)	0.94 (0.85–1.04)
Race	– Black	Caucasian	1.30 (1.09–1.56)	1.44 (1.08–1.93)
	– Asian		1.09 (0.78–1.51)	1.22 (0.84–1.78)
Clinical Severity				
Motor score		Localising/		
Absent		Obey commands	5.30 (3.49–8.04)	–
Abnormal extension			7.48 (5.6–9.98)	–
Abnormal Flexion			3.58 (2.71–4.73)	–
Flexion			1.74 (1.44–2.41)	–
Pupillary reactivity				
1 reacting		Both reacting	2.70 (2.07–3.53)	
Both non-reacting			4.77 (3.46–6.57)	
Secondary insults		Absent		
Hypotension		Absent	2.67 (2.09–3.41)	2.06 (1.64–2.59)
Hypoxia			2.08 (1.69–2.56)	1.65 (1.37–2.00)
Hypothermia		Absent	2.21 (1.56–3.15)	1.63 (1.11–2.40)
Structural abnormalities				
CT Classification		CT Class II		
CT Class I			0.45 (0.35–0.67)	0.47 (0.32–0.70)
CT Class III/IV			2.62 (2.13–3.21)	2.23 (1.83–2.72)
Mass lesion			2.18 (1.83–2.61)	1.48 (1.27–1.71)
tSAH present*		Absent	2.64 (2.42–2.89)	2.01 (1.83–2.21)
Type of intracranial lesion				
No epidural		No epidural	0.64 (0.56–0.72)	0.63 (0.55–072)
Laboratory parameters				
Glucose		25–75% i.q.r.	1.68 (1.54–1.83)	1.45 (1.36–1.55)
pH			0.80 (0.74–0.88)	0.84 (0.67–0.92)
Prothrombine time			1.41 (0.99–1.99)	1.63 (1.40–1.89)
Hb			0.69 (0.60–0.78)	0.76 (0.66–0.88)
Sodium < 137mmol/L		≥ 137 mmol/L	1.40 (1.22–1.60)	1.14 (0.91–1.43)

i.q.r. = interquartile range, tSAH = Traumatic Subarachnoid Hemorrhage, OR = Odds ratios from proportional odds analysis

outcome in a continuous way, reporting both a 'change point' around the age of 30-40, above which outcome becomes increasingly poorer, and a more or less continuous relation across all ages, which may be approximated by a linear function.[17-21]

Other demographic factors studied for their association with outcome include gender and race. Males are more prone to suffer from TBI due to higher risk for road traffic accidents and assaults. Although many studies did not find a relationship between gender and outcome after adjustment,[17, 20, 22, 23] a meta-analysis by Farace and Alves found poorer quality of life and functional outcomes in females surviving severe TBI compared to males.[24]

The association between race and outcome after TBI was controversial until a meta-analysis combining evidence from a substantial number of patients, showed that black patients have a poorer outcome. This association is confirmed in some recent studies[20, 25-27] The underlying reasons for this association can only be speculated upon, but may include differences in genetic constitution, causing a different response to injury and differences in access to acute and post-acute care. We consider this a priority for further research.

Clinical severity

Clinical severity relates both to extracranial and intracranial injuries. The overall severity of extracranial injuries is commonly assessed with the Abbreviated Injury Score (AIS)[28] or the Injury Severity Score (ISS).[29] Most studies about TBI and extracranial injury have studied patients with traumatic extracranial injury with or without TBI. The conclusion is that the coexistence of moderate traumatic brain injury with extracranial injury is associated with mortality and morbidity.[30-32]

In contrast, there is no consensus about the prognostic value of major extracranial injury in TBI patients. Some studies demonstrate that outcome mainly depends on the severity of the primary cerebral damage and is not worsened by the presence of extracranial injuries[33], but others show that the presence of major extracranial injuries is associated with a poorer outcome.[21, 37]

Recently we performed a meta analysis of individual patient data, and found that the conflicting results from previous studies may be explained by an interaction with the severity of brain injury. For patients with more severe brain injuries, the effect of extracranial injury on functional outcome was small, whereas in those with milder brain injuries, extracranial injuries had a more pronounced effect. This indicates that it is relevant to test for clinically plausible interaction effects. We also found that extracranial injuries mainly increase the chance of early mortality. Thus, the effect of extracranial injury found in registries that include patients who die early, will be larger than the effect found in trials that exclude these patients [van Leeuwen et al, submitted].

The clinical severity of intracranial injuries is reflected by the level of consciousness, assessed by the Glasgow Coma Scale (GCS).[7] Many studies have demonstrated an association between lower levels of the GCS and poorer outcome.

In patients with more severe injuries, the motor component of the GCS has the greatest predictive value, as in these patients eye and verbal response is commonly absent.[34] The prognostic value of the eye and verbal components of the GCS become more relevant in patients with less severe injuries who obey commands. It should be recognized that the GCS may fluctuate early after injury with some patients deteriorating, and others improving.[35] From a perspective of prognosis, assessment of the GCS should therefore be related to a given time period, commonly on admission after primary respiratory and hemodynamic stabilisation. Reliable assessment of the GCS however may be obscured in the acute setting by medical sedation, paralysis or intoxication.[36, 37]

Abnormalities in pupillary reactivity reflect brainstem damage or compression and are strongly associated with poorer outcome.[38] Pupillary reactivity is a more stable parameter in the early phase after injury than the GCS, being less prone to influences of sedation and paralysis.

Second insults

The injured brain is more vulnerable for systemic second insults, such as hypoxia and hypotension than a normal, healthy brain. Second insults are frequent after TBI, especially in the pre-hospital setting[39, 40] and can increase the degree of secondary damage. The association of second insults, in the pre-hospital setting or during acute care, with poorer outcome has been well established and various studies have shown that the combination of hypoxia and hypotension has a greater adverse effect on outcome than can be explained by either insult alone.[40-44]

Most studies have used a cutoff value for early hypotensive and hypoxic events (e.g. any episode with a systolic blood pressure < 90 mm/Hg). Detailed analysis of the association between the measured blood pressure on admission and outcome, however, showed that this relation is continuous: low blood pressure and high blood pressure are both associated with poorer outcome with a U-shaped relationship. Following adjustment for age, motor score and pupillary reactivity, the effects of higher blood pressure however largely disappear, suggesting that higher blood pressure values are merely reflective of more severe injuries and may possibly be caused by raised ICP (Cushing response).[43]

Structural abnormalities

The prognostic value of CT characteristics has been well documented, including the status of basal cisterns, midline shift, the presence and type of intracranial lesions and traumatic subarachnoid hemorrhage (tSAH). Obliteration of the basal cisterns and the presence of tSAH are the strongest CT predictors (BTF prognosis guidelines, available from http://www.braintrauma.org). In 1991 Marshall et al introduced a descriptive system of CT classification, which focuses on the presence or absence of a mass lesion and differentiates diffuse injuries by signs of increased intracranial pressure (compression of basal cisterns, midline shift).

Although the Marshall CT classification has prognostic value, combining individual CT characteristics in a model, such as in the Rotterdam CT score, provides better discrimination between patients with better versus poorer outcome than the descriptive Marshall classification (Panel 5.2).[44-46]

Panel 5.2 Marshall CT Classification[45]

Category	Definition
Diffuse injury I (no visible pathology)	No visible intracranial pathology seen on CT scan
Diffuse injury II	Cisterns are present with midline shift of 0–5 mm and/or lesions densities present; no high or mixed density lesion >25 cm^3 may include bone fragments and foreign bodies
Diffuse injury III (swelling)	Cisterns compressed or absent with midline shift of 0–5mm; no high or mixed density lesion >25 mm
Diffuse injury IV (shift)	Midline shift >5 mm; no high or mixed density lesion >25 cm^3
Evacuated mass lesion	Any lesion surgically evacuated
Non-evacuated mass lesion	High or mixed density lesion >25 cm^3; not surgically evacuated

Rotterdam Prognostic CT Score[46]

Predictor value	Score
Basal Cisterns	
– Normal	0
– Compressed	1
– Absent	2
Midline shift	
– No shift or shift ≤ 5 mm	0
– Shift > 5 mm	1
Epidural mass lesion	
– Present	0
– Absent	1
Intraventricular blood or tSAH	
– Absent	0
– Present	1
Sum score	+1

Prognostic studies have mainly focused on CT abnormalities and employed relatively broad categorizations. In tSAH for example, one of the strongest CT predictors, most studies have concentrated on presence or absence without differentiating as to the location (basal cisterns versus cortical) or extent. More detailed analysis and the use of advanced MRI imaging may yield additional prognostic information.

Laboratory values and biomarkers

Recently, interest in biomarkers, including laboratory parameters, is increasing. Biomarkers may be used to detect and track pathophysiological phenomena, as marker of injury severity, and to aid in prognosis assessment. In mild TBI, a biomarker that could establish the diagnosis, or predict the likelihood of secondary damage would have great clinical utility. In more severe injuries, the use of a biomarker to assess injury severity may avoid problems with unreliable GCS assessments in patients who are intoxicated or intubated. A number of putative serum, cerebrospinal fluid (CSF), and microdialysate biomarkers have been evaluated in clinical studies of TBI, with S100 and neuron-specific enolase (NSE) being among the most widely investigated.[47-52]

Although an association between several biomarkers and outcome has been established, the prognostic value is unclear due to relatively small numbers analyzed in univariate rather than multivariable analyses.[53] Levels of biomarkers may correlate with other clinical indicators such as GCS[54], and offer limited additional prognostic value over other predictors. The predictive value of biomarkers over and above other predictors has to be established in multivariable analysis.[55]

The prognostic value of laboratory parameters that are routinely measured has been investigated in larger numbers. High glucose values, low hemoglobin, low platelets as well as coagulation disturbances are the strongest predictors, independently related to poorer outcome.[56-59]

Laboratory values are potentially modifiable. The question of causality is relevant when attempts are made to correct abnormal values in the expectation to improve outcome. Based on the observed association between higher glucose levels and poorer outcome, two randomized trials were recently conducted to assess the effect of intensive insulin therapy to lower glucose levels. Both studies were however small (<150 patients) and results conflicting. [60-61] The risks of tight glucose control in TBI have been illustrated in microdialysis studies in the brain showing that normalization of blood glucose could lead to a depletion of glucose in the extracellular fluid of the brain, thus compromising cerebral metabolism. [62-64] Although an association between abnormal values and poorer outcome may exist, this does not by definition mean that correcting these abnormal values will indeed improve outcome. The observed abnormality may simply be an expression or surrogate marker of the severity of injury. Randomized controlled trials, are required to prove whether correcting abnormal values is of benefit.

Clinical course

Changes in admission parameters

A deterioration in neurologic function is a dire prognostic sign that generally indicates progressive brain damage. Early prognosis studies showed that the worst GCS over a given time period is especially predictive of poorer outcome. Deterioration in neurological function has been defined more objectively by Morris et al as neuroworsening (Panel 5.3)[65] and is highly predictive for poor outcome. In addition to the initial CT scan, follow-up scans also contain prognostic information. A survey among patients with moderate and severe TBI organized by the European Brain Injury Consortium (EBIC) showed that a substantial number of patients with diffuse injury (no mass lesions) at the first CT, demonstrated progressive intracranial pathologies on subsequent CT examinations.[66] The worst CT was more strongly correlated to outcome. Many other studies have confirmed the frequent occurrence of 'CT progression', but relatively few have addressed the prognostic significance. This is complex, as CT progression may often lead to therapeutic intervention.

Panel 5.3 Criteria for neuroworsening[65]

Criteria
▪ Spontaneous decrease GCS motor score >=2 points (compared with previous examination)
▪ New loss of pupillary reactivity
▪ Development of pupillary asymmetry of >=2 mm
▪ Other deterioration in neurological status sufficient to warrant immediate medical or surgical intervention

Second insults may occur in the clinical setting, despite all attempts to avoid them. Patients are particularly at risk for second insults during intra- and interhospital transport.[67] The depth, duration and number of second insults cumulate towards poorer outcome.[43, 44, 68]

As on admission, the strongest evidence for the prognostic value of laboratory parameters during the clinical course exists for glucose, platelets and coagulation disturbances Persistently high glucose levels are associated with poorer outcomes, also after adjustment for other important predictors.[56, 69, 70]

The lowest platelet count in the first 24 hours after admission is an independent predictor of outcome after six months.[56] A recent meta analysis showed that the prevalence of coagulopathy after TBI was 33% and that coagulopathy was related both to mortality and unfavourable outcome.[71]

Clinical monitoring

In more severely injured patients invasive and non-invasive monitoring in the ICU situation provides a wealth of information. Approaches to analysis have however remained relatively crude. It is difficult to draw clear conclusions on the predictive value of monitored parameters as the time of initiation and duration of monitoring vary greatly between and within studies. Summary measures, such as for ICP monitoring, include the highest, lowest and mean values overall or per day and the number of episodes or percentage of time that values are above a predefined threshold. This variability in analysis and reporting confounds comparisons between studies. Further, in repeated measurements predictive information may be better captured in patterns than in single values. Modern statistical approaches are available to analyze repeated measurements per patient, but have seldom been used in TBI studies and hence pose a challenge for future research.

Many studies have confirmed an association of high ICP, low CPP and decreased brain oxygen tension levels with poorer outcome.[56, 72-77] These associations, in combination with our understanding of pathophysiologic consequences form the rationale for guideline recommendations to avoid high intracranial pressure (ICP) and low cerebral perfusion pressure (CPP) (available from http://www.braintrauma.org).

It also has been suggested that outcome may be more dependent on ICP variability and on response to treatment of raised ICP than on absolute mean ICP values.[76, 78]

Electroencephalography and evoked potentials

In the past decades, there has been interest in electroencephalography (EEG) and evoked potentials or event-related potentials as predictors of outcome.[79, 80] A review, published in 2004, stated that the predictive ability of EEG is limited.[81] It was suggested that this may be because the TBI has greater impact on subcortical axonal fibers than on the cortical gray matter that generates most of the EEG signal. In the postacute phase the bispectral index has a higher correlation with behavioral scales than the EEG and may help in differentiating between a vegetative and minimally conscious state, also after TBI.[82]

Multiple studies have shown that somatosensory evoked potentials are useful predictors of outcome after TBI.[83-85] Lew et al (2003) reported that bilateral absence of cortically recorded median nerve SEPs within 8 days of severe TBI was strongly predictive of death or persistent vegetative state.[85] A meta-analysis showed that bilaterally negative SEPs had a 98.5% positive likelihood ratio for unfavourable outcome.[86] The false-positive rate for bilaterally negative SEPs may however be high in patients with focal lesions, subdural effusions and after recent decompressive craniectomies.

Although the results of research in this field are promising, the evidence regarding the prognostic effects of these clinical neurophysiological modalities is limited, and the added value over other clinical predictors is uncertain.

Prognostic models

Estimation of prognosis is by definition a multivariable challenge. Predictors should be considered jointly rather than on their own and can be combined in a multivariable prognostic model to quantify the risk for a particular outcome in individual patients.

Combining individual predictors into a model will increase the performance and generalizibility and is all the more important, as patients may have characteristics that affect the outcome in opposite directions. For example, for a 24-year old patient with fixed pupils, we would expect a favourable outcome based on age, but an unfavourable outcome based on pupil reactivity.

In an updated literature search to January 2010, we identified 27 prognostic models in 16 studies, meeting the following criteria: presenting a prognostic model for mortality >2 weeks post discharge or 6 month GOS in English language, with predictors measured within 24 hours after injury, and including over 200 patients with age>14 years, presenting with GCS<14 or motor score<6, and non-penetrating injury. Many of these models showed shortcomings, in particular a high risk of overfitting, e.g. that predictive performance is much poorer in new patients than expected from the development phase, and lack of external validation (Table 5.2).[11,17,87-101] The risk of overfitting was high in 10 of the 16 studies identified in this review. The number of considered predictors was mostly higher than the number included in the final model, and often too high in relation to the available sample size. As a rule of thumb, the maximum number of candidate predictors can be approximated by dividing the number of events (e.g. number of patients with poor outcome) by a factor 10, e.g. at most 10 predictors for 100 events.[102] Also overfitting is caused by using statistical techniques for predictors selection that are too much data driven, such as backward selection in a small dataset. Overfitting can be assessed by internal validation techniques, such as bootstrap resampling.[103] More important is external validation, i.e. testing model performance in another setting that differs in time or place.[104,105]

Table 5.2 Overview of prognostic models in moderate and severe TBI published between 1976 and 2009

Reference	Year	N (developm.)	N Predictors included	Severity	Outcome	Risk of overfitting	External validation
Jennett et al.[87]	1976	600	4	In coma for at least 6h	GOS 6 months	High	No
Braakman et al.[88]	1980	305	3	In coma for at least 6h	GOS 6 months	High	No
Choi et al.[89]	1983	264	4	Severe head injury and GMS ≤5	GOS 6 months	Intermediate	No
Lokkenberg and Grimes[90]	1984	254	2	GCS ≤8	GOS 6 months	Low	No
Braakman et al.[91]	1986	306	3	Severe head injury	GOS 6 months	Intermediate	No
Choi et al.[92]	1988	523	3	Severe head injury	GOS 6 months	High	No
Choi et al.[93]	1991	555	4	GCS ≤8	GOS 12 months	High	No
Fearnside et al.(2 models)[94]	1993	315/218	5	GCS ≤8	Mortality/ GOS	Intermediate	No
Marmelak et al.[95]	1996	672	3	GCS ≤8	GOS 6 months	Low	No
Quigley et al.[96]	1997	380	2	GCS 3-5	GOS 6 months	Low	No
Lang et al.[97]	1997	799	4	GCS ≤8	Mortality 6 months	High	No
Signorini et al.[11]	1999	372	5	GCS ≤12 and GCS >12 if ISS >15	Mortality 12 months	Intermediate	Yes
Ratanalert et al.[98]	2002	337	3	GCS ≤8	GOS 6 months	Intermediate	No
Hukkelhoven et al.(2 models)[99]	2005	2269	7	GCS ≤12	Mortality/ GOS 6 months	Low	Yes
Cremer et al.[100]	2006	304	5	GCS ≤8 and in coma for at least 24h	GOSE 12 months	Low	Yes
Perel et al. (4 models)[17]	2008	10008	4-9	GCS ≤14	Mortality/ GOS 14 days/ 6 months	Low	Yes
Steyerberg et al.(6 models)[101]	2009	8509	3-10	GCS ≤12	Mortality/ GOS 6 months	Low	Yes

External validation was only reported in 5 studies. These findings are consistent with reviews published by Perel and Mushkudiani.[106,107]

The models reported by the CRASH trial investigators and by the IMPACT study group are the most recent and developed on the largest patient numbers (10,008 and 8,509 respectively).[17, 101] Different sets of prediction models were developed with logistic regression analysis and cross-validated on each other. Models are available in score chart format and in a web based application (CRASH: http://www.crash2.lshtm.ac.uk/risk%20calculator/index.html; IMPACT: http://www.tbi-impact.org). Both studies showed that the largest amount of prognostic information is contained in a core set of three predictors (age, GCS or motor score, and pupillary reactivity) (Panel 5.4).

The CRASH models included also patients with milder injuries in the development, and are consequently also applicable to these. The IMPACT models focussed on moderate and severe TBI. Both models can be considered to represent the current state of the art in prognostic modelling in TBI as they were developed on large numbers and conformed to accepted quality criteria for model development, including external validation.

Panel 5.4 CRASH and IMPACT prediction models[17, 101]

	IMPACT	CRASH
Predicted outcome	6 month mortality 6 month unfavourable outcome	14 day mortality 6 month unfavourable outcome
Core model	Age Motor score Pupil reactivity	Age GCS Pupil reactivity Major extracranial injury
CT model	*Core model +* Hypoxia Hypotension CT Classification tSAH on CT Epidural mass on CT	*Core model +* petechial haemorrhages Obliteration of the third ventricle or basal cisterns Subarachnoid bleeding Midline shift Non-evacuated haematoma
Lab model	*CT model +* Glucose Hb	
Available at	http://www.tbi-impact.org	http://www.crash2.lshtm.ac.uk/risk%20calculator/index.html

Applications of prognostic models in TBI

Clinical Practice

Some estimation of prognosis is consciously or subconsciously employed by physicians, when informing relatives, making treatment decisions or allocating resources. Estimates derived from large datasets are preferable to relying on the subjective opinion of a physician whose experience, no matter how vast, can never match the information contained in the data of thousands of patients. The Canadian CT rule and the CHIP prediction rule for CT scanning in mild TBI are clear examples of how prediction models can provide evidence to better inform clinical decisions.[3,5] Caution is advocated when outcome prediction models are applied in individual patients. Prognostic estimates are necessarily only probabilities and cannot provide certainty on an actual outcome.

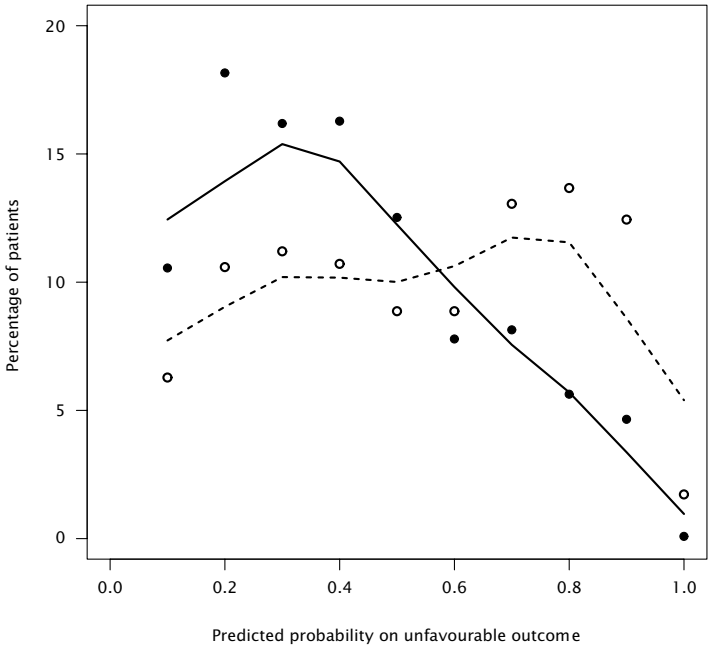
Research

The main research applications of prognostic models for outcome in TBI include classification and clinical trials. Prognostic risk estimation on admission provides a tool for classification of populations by their prognostic risk distribution (Figure 5.3).[108] We hence obtain insight into differences in the case-mix of different studies.

In the design and analysis of randomized controlled trials, prognostic models offer opportunities both in the enrolment and analysis phase. Traditionally, clinical trials employ relatively strict enrolment criteria. Some of these are motivated by safety and ethical considerations, but most criteria (e.g. age and disease severity) aim to exclude patients with a very good or very poor prognosis. Patients at these extremes are not likely to demonstrate benefit from the treatment under investigation. It is statistically more efficient to combine these criteria in a prognostic model.[109-112] The prognostic estimate can first be used to determine eligibility and next for the analysis.

In the analysis phase, prognostic models can adjust for baseline characteristics. This substantially increases statistical power, or equivalently, allows for a reduction in the required sample size (by more than 25%).[113] Prognostic analysis is further essential to the use of the sliding dichotomy, where the point of dichotomy of the GOS is differentiated according to the baseline prognostic risk.[114] For a patient with a very severe injury, survival may be relevant whilst for patients with less severe injuries any outcome worse than good recovery might be considered unfavourable. The sliding dichotomy approach has been adopted for the primary analysis of a number of phase III trials in TBI, stroke and intracerebral hemorrhage.[115-117]

Figure 5.3 Distribution of predicted probabilities on 6 month unfavourable outcome from the IMPACT Core model in a Randomized Controlled Trial ('Tirilizad International', solid points, solid line) and an observational study ('UK4', open points, dashed line). This figure shows the comparison of an observational study to a randomized controlled trial among TBI patients. The proportion of patients with a relative poor prognosis was smaller in the randomized controlled trial than in the observational study.[108]



Most prognostic studies in TBI have however analyzed the GOS arbitrarily dichotomized into unfavourable versus favourable outcome. As a result we cannot assume that predictors differentiating death very well from survival will perform similarly when asked to predict good recovery versus worse outcomes. To overcome this limitation, a proportional odds model can be used, as was done in the IMPACT studies. This approach uses the full GOS as outcome instead of a dichotomized GOS, assuming that the predictors differentiate equally well over each possible dichotomization (proportional odds assumption). A proportional odds model may be more relevant for all patients since it differentiates between death/survival in poor prognosis patients, but also between good recovery and anything worse in good prognosis patients. Moreover both the sliding dichotomy and proportional odds models substantially increase statistical power.[118] The proportional odds assumption may not be valid for all predictors. For example, the presence of severe extracranial injury discriminates between death and survival, but less well between good recovery and moderate disability. [van Leeuwen et al, in preparation].

Quality assessment of health care delivery
Comparison of observed and expected outcomes may give an indication of the quality of care delivered in a specific hospital or in a specific country. An example is the standardized mortality ratio (observed deaths/expected deaths, adjusted for baseline characteristics), which is used as a quality score in intensive care medicine. The expected mortality is commonly derived from scoring systems, such as Apache II, TRISS or SAPS II/III. These systems were however developed for a general ICU population, and their applicability to the indication TBI is doubtful. Prognostic models specific to TBI can better set baselines for clinical audits and benchmarking. These models are of great potential relevance for assessing the quality of health care delivery, as they have been developed not only for mortality, but also for functional outcome, as assessed by the GOS. It should be noted however that the cumulative R^2 of the IMPACT model is 0.35. This indicates that a great deal of unexplained variation still exists, so case-mix adjustment is incomplete by definition.

Discussion

This review illustrates that prognostic analysis and prognostic modelling have a great potential in TBI, both for diagnosis and prognosis. At the same time some of the gaps in our knowledge are identified, highlighting issues for further investigation.

Validated prognostic models have been based mainly on admission characteristics. Although, considerable insight has been gained into the prognostic value of variables obtained during the subsequent clinical course, these have not yet been widely included in prognostic models. Further research should focus on the quantification of the additional benefit that might be obtained for outcome prediction.

The epidemiology of TBI is changing and approaches to pre-hospital care, diagnostic capabilities and intensive care monitoring and treatment are continuously improving. Consequently, prognostic analysis should be seen as a continuing process requiring ongoing updating and validation in contemporary series.[119]

In the analyses of continuous variables such as age, blood pressure or laboratory parameters, many studies used threshold values, creating a dichotomy or categorization of continuous predictors, e.g. age <=50 vs. >50 years. Threshold values are increasingly used in clinical medicine towards goal-directed therapy. Threshold values are, however, not natural to biological systems. Collapsing continuous variables has many disadvantages.[120] It is recommended that future prediction studies analyze continuous predictors in a continuous way, possibly as a non-linear variable.[53]

A major gap in our knowledge concerns uncertainty how individuals may possibly respond differently to similar injuries. Such differences may in part be genetically determined and much research will be required in the fields of genomics and metabolomics to elucidate variability in response. An indication how relevant this may be, is given by

the observation that recovery is poorer in patients with stroke or TBI in the presence of APOEε4 genotype.[121] Other genes for which evidence exists for an association with poorer outcome are P53, COMT, DND2 and Cacna1a genes.[122] Also response to treatment varies between individuals. In oncology, one refers to characteristics that predict response to treatment as predictive factors, whilst prognostic factors more generally predict outcome (REMARK guidelines). Research on predictive factors in TBI is starting, including various biomarkers and imaging modalities. Predictive factors may lead to targeted therapies, considering individual mechanisms of disease.[123]

Further research is also required into more sensitive outcome measures, particularly in milder TBI.

Directly relevant to prognostic research in TBI is better standardization of data collection and coding to facilitate sharing of results and to permit meta-analysis of individual patient data across studies.[124] This will give the opportunity to improve, validate, and update prognostic models on larger numbers of patients.

The challenge for the immediate future is the implementation of prediction models in clinical practice. The tools are now provided by the availability of reliable and externally validated models, and it is up to clinicians and researchers to adopt these for general clinical and research applications, either to improve quality of care, or to beat the prognostic estimate.

Acknowledgments

Search strategy and selection criteria
References for this review were identified through searches of *PubMed*, by use of the search terms 'traumatic brain injury' or 'head injury' and other appropriate targets such as 'prognosis', 'prognostic models' up to January 2010. Papers were also identified from the authors own files and from references cited in relevant articles. An electronic search of resources, such as book chapters, was also done. We considered only publications written in English and Dutch. The final reference list was generated on the basis of relevance to the topics covered in this review.

Funding
This work was supported by NIH (NS-042691).

References

1. Maas AI, Marmarou A, Murray GD, Teasdale SG, Steyerberg EW. Prognosis and clinical trial design in traumatic brain injury: the IMPACT study. *J Neurotrauma* 2007 Feb;24(2):232-8.
2. Styrke J, Stalnacke BM, Sojka P, Bjornstig U. Traumatic brain injuries in a well-defined population: epidemiological aspects and severity. *J Neurotrauma* 2007 Sep;24(9):1425-36.
3. Smits M, Dippel DW, Steyerberg EW, de Haan GG, Dekker HM, Vos PE, et al. Predicting intracranial traumatic findings on computed tomography in patients with minor head injury: the CHIP prediction rule. *Ann Intern Med* 2007 Mar 20;146(6):397-405.
4. Haydel MJ, Preston CA, Mills TJ, Luber S, Blaudeau E, DeBlieux PM. Indications for computed tomography in patients with minor head injury. *N Engl J Med* 2000 Jul 13;343(2):100-5.
5. Stiell IG, Wells GA, Vandemheen K, Clement C, Lesiuk H, Laupacis A, et al. The Canadian CT Head Rule for patients with minor head injury. *Lancet* 2001 May 5;357(9266):1391-6.
6. Kuppermann N, Holmes JF, Dayan PS, Hoyle JD,Jr, Atabaki SM, Holubkov R, et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *Lancet* 2009 Oct 3;374(9696):1160-70.
7. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974 Jul 13;2(7872):81-4.
8. Steyerberg EW. Clinical Prediction Models. *A Practical Approach to Developement, Validating and Updating*. New York: Springer; 2009.
9. McHugh GS, Butcher I, Steyerberg EW, Lu J, Mushkudiani N, Marmarou A, et al. Statistical approaches to the univariate prognostic analysis of the IMPACT database on traumatic brain injury. *J Neurotrauma* 2007 Feb;24(2):251-8.
10. Murray GD, Butcher I, McHugh GS, Lu J, Mushkudiani NA, Maas AIR, Marmarou A, Steyerberg EW. Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007 Feb;24(2):251-8.
11. Signorini DF, Andrews PJ, Jones PA, Wardlaw JM, Miller JD. Predicting survival using simple clinical variables: a case study in traumatic brain injury. *J Neurol Neurosurg Psychiatry* 1999 Jan;66(1):20-5.
12. Ono J, Yamaura A, Kubota M, Okimura Y, Isobe K. Outcome prediction in severe head injury: analyses of clinical prognostic factors. *J Clin Neurosci* 2001 Mar;8(2):120-3.
13. Andrews PJ, Sleeman DH, Statham PF, McQuatt A, Corruble V, Jones PA, et al. Predicting recovery in patients suffering from traumatic brain injury by using admission variables and physiological data: a comparison between decision tree analysis and logistic regression. *J Neurosurg* 2002 Aug;97(2):326-36.
14. Demetriades D, Murray J, Martin M, Velmahos G, Salim A, Alo K, et al. Pedestrians injured by automobiles: relationship of age to injury type and severity. *J Am Coll Surg* 2004 Sep;199(3):382-7.
15. Ratanalert S, Chompikul J, Hirunpat S. Talked and deteriorated head injury patients: how many poor outcomes can be avoided? *J Clin Neurosci* 2002 Nov;9(6):640-3.
16. Bahloul M, Chelly H, Ben Hmida M, Ben Hamida C, Ksibi H, Kallel H, et al. Prognosis of traumatic head injury in South Tunisia: a multivariate analysis of 437 cases. *J Trauma* 2004 Aug;57(2):255-61.

17. MRC CRASH Trial Collaborators, Perel P, Arango M, Clayton T, Edwards P, Komolafe E, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ* 2008 Feb 23;336(7641):425-9.
18. Gomez PA, Lobato RD, Boto GR, De la Lama A, Gonzalez PJ, de la Cruz J. Age and outcome after severe head injury. *Acta Neurochir* (Wien) 2000;142(4):373,80; discussion 380-1.
19. Hukkelhoven CW, Steyerberg EW, Rampen AJ, Farace E, Habbema JD, Marshall LF, et al. Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. *J Neurosurg* 2003 Oct;99(4):666-73.
20. Mushkudiani NA, Engel DC, Steyerberg EW, Butcher I, Lu J, Marmarou A, et al. Prognostic value of demographic characteristics in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007 Feb;24(2):259-69.
21. Tokutomi T, Miyagi T, Ogawa T, Ono J, Kawamata T, Sakamoto T, et al. Age-associated increases in poor outcomes after traumatic brain injury: a report from the Japan Neurotrauma Data Bank. *J Neurotrauma* 2008 Dec;25(12):1407-14.
22. Colantonio A, Escobar MD, Chipman M, McLellan B, Austin PC, Mirabella G, et al. Predictors of post-acute mortality following traumatic brain injury in a seriously injured population. *J Trauma* 2008 Apr;64(4):876-82.
23. Utomo WK, Gabbe BJ, Simpson PM, Cameron PA. Predictors of in-hospital mortality and 6-month functional outcomes in older adults after moderate to severe traumatic brain injury. *Injury* 2009 Sep;40(9):973-7.
24. Farace E, Alves WM. Do women fare worse? A metaanalysis of gender differences in outcome after traumatic brain injury. *Neurosurg Focus* 2000;8(1):e6.
25. Sorani MD, Lee M, Kim H, Meeker M, Manley GT. Race\ethnicity and outcome after traumatic brain injury at a single, diverse center. *J Trauma* 2009 Jul;67(1):75-80.
26. Shafi S, Marquez de la Plata C, Diaz-Arrastia R, Shipman K, Carlile M, Frankel H, et al. Racial disparities in long-term functional outcome after traumatic brain injury. *J Trauma* 2007 Dec;63(6):1263,8; discussion 1268-70.
27. Arango-Lasprilla JC, Rosenthal M, Deluca J, Komaroff E, Sherer M, Cifu D, et al. Traumatic brain injury and functional outcomes: does minority status matter? *Brain Inj* 2007 Jun;21(7):701-8.
28. Medicine AftAoA. *The Abbreviated Injury Scale*, 1990 Revision. Association for the Advancement of Automotive Medicine. 15-24. 1990. Des Plaines, IL.
29. Baker SP, O'Neill B, Haddon W,Jr, Long WB. The injury severity score: a method for describing patients with multiple injuries and evaluating emergency care. J Trauma 1974 Mar;14(3):187-96.
30. Gennarelli TA, Champion HR, Copes WS, Sacco WJ. Comparison of mortality, morbidity, and severity of 59,713 head injured patients with 114,447 patients with extracranial injuries. *J Trauma* 1994 Dec;37(6):962-8.
31. Gennarelli TA, Champion HR, Sacco WJ, Copes WS, Alves WM. Mortality of patients with head injury and extracranial injury treated in trauma centers. *J Trauma* 1989 Sep;29(9):1193,201; discussion 1201-2.

32. McMahon CG, Yates DW, Campbell FM, Hollis S, Woodford M. Unexpected contribution of moderate traumatic brain injury to death after major trauma. *J Trauma* 1999 Nov;47(5):891-5.
33. Sarrafzadeh AS, Peltonen EE, Kaisers U, Kuchler I, Lanksch WR, Unterberg AW. Secondary insults in severe head injury – do multiply injured patients do worse? *Crit Care Med* 2001 Jun;29(6):1116-23.
34. Marmarou A, Lu J, Butcher I, McHugh GS, Murray GD, Steyerberg EW, et al. Prognostic value of the Glasgow Coma Scale and pupil reactivity in traumatic brain injury assessed pre-hospital and on enrollment: an IMPACT analysis. *J Neurotrauma* 2007 Feb;24(2):270-80.
35. Murray GD, Teasdale GM, Braakman R, Cohadon F, Dearden M, Iannotti F, et al. The European Brain Injury Consortium survey of head injuries. *Acta Neurochir* (Wien) 1999;141(3):223-36.
36. Balestreri M, Czosnyka M, Chatfield DA, Steiner LA, Schmidt EA, Smielewski P, et al. Predictive value of Glasgow Coma Scale after brain trauma: change in trend over the past ten years. *J Neurol Neurosurg Psychiatry* 2004 Jan;75(1):161-2.
37. Stocchetti N, Pagan F, Calappi E, Canavesi K, Beretta L, Citerio G, et al. Inaccurate early assessment of neurological severity in head injury. *J Neurotrauma* 2004 Sep;21(9):1131-40.
38. Stocchetti N, Colombo A, Ortolano F, Videtta W, Marchesi R, Longhi L, et al. Time course of intracranial hypertension after traumatic brain injury. *J Neurotrauma* 2007 Aug;24(8):1339-46.
39. Stocchetti N, Furlan A, Volta F. Hypoxemia and arterial hypotension at the accident scene in head injury. *J Trauma* 1996 May;40(5):764-7.
40. Chesnut RM, Marshall LF, Klauber MR, Blunt BA, Baldwin N, Eisenberg HM, et al. The role of secondary brain injury in determining outcome from severe head injury. J Trauma 1993 Feb;34(2):216-22.
41. Signorini DF, Andrews PJ, Jones PA, Wardlaw JM, Miller JD. Adding insult to injury: the prognostic value of early secondary insults for survival after traumatic brain injury. *J Neurol Neurosurg Psychiatry* 1999 Jan;66(1):26-31.
42. Walia S, Sutcliffe AJ. The relationship between blood glucose, mean arterial pressure and outcome after severe head injury: an observational study. Injury 2002 May;33(4):339-44.
43. McHugh GS, Engel DC, Butcher I, Steyerberg EW, Lu J, Mushkudiani N, et al. Prognostic value of secondary insults in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007 Feb;24(2):287-93.
44. Chesnut RM. Secondary brain insults after head injury: clinical perspectives. *New Horiz* 1995 Aug;3(3):366-75.
45. Marshall LF, Bowers S, Klauber MR, Van Berkum M, Eisenberg HM, Jane JA, Luerssen TG, Marmarou A, Foulkes MA. A new classification of head injury based on computerised tomography. *Journal of Neurosurgery.* 75:S14-S20 (1991).
46. Maas AI, Hukkelhoven CW, Marshall LF, Steyerberg EW. Prediction of outcome in traumatic brain injury with computed tomographic characteristics: a comparison between the computed tomographic classification and combinations of computed tomographic predictors. *Neurosurgery* 2005 Dec;57(6):1173,82; discussion 1173-82.
47. Sawauchi S, Taya K, Murakami S, Ishi T, Ohtsuka T, Kato N, et al. Serum S-100B protein and neuron-specific enolase after traumatic brain injury. *No Shinkei Gek*a 2005 Nov;33(11):1073-80.

48. Naeimi ZS, Weinhofer A, Sarahrudi K, Heinz T, Vecsei V. Predictive value of S-100B protein and neuron specific-enolase as markers of traumatic brain damage in clinical use. *Brain Inj* 2006 May;20(5):463-8.
49. Nylen K, Ost M, Csajbok LZ, Nilsson I, Hall C, Blennow K, et al. Serum levels of S100B, S100A1B and S100BB are all related to outcome after severe traumatic brain injury. *Acta Neurochir* (Wien) 2008 Mar;150(3):221,7; discussion 227.
50. Schultke E, Sadanand V, Kelly ME, Griebel RW, Juurlink BH. Can admission S-100beta predict the extent of brain damage in head trauma patients? *Can J Neurol Sci* 2009 Sep;36(5):612-6.
51. Beaudeux JL. S100B protein: a novel biomarker for the diagnosis of head injury. *Ann Pharm Fr* 2009 May;67(3):187-94.
52. Rainey T, Lesko M, Sacho R, Lecky F, Childs C. Predicting outcome after severe traumatic brain injury using the serum S100B biomarker: results using a single (24h) time-point. *Resuscitation* 2009 Mar;80(3):341-5.
53. Kövesdi E, Lückl J, Bukovics P, Farkas O, Pál J, Czeiter E, Szellár D, Dóczi T, Komoly S, Büki A. Update on protein biomarkers in traumatic brain injury with emphasis on clinical use in adults and paediatrics, *Acta Neurochirurgica*, 2010.
54. Pineda JA, Lewis SB, Valadka AB, Papa L, Hannay HJ, Heaton SC, et al. Clinical significance of alphaII-spectrin breakdown products in cerebrospinal fluid after severe traumatic brain injury. *J Neurotrauma* 2007 Feb;24(2):354-66.
55. Kattan MW. Judging new markers by their ability to improve predictive accuracy. *J Natl Cancer Inst* 2003 May 7;95(9):634-5.
56. Lannoo E, Van Rietvelde F, Colardyn F, Lemmerling M, Vandekerckhove T, Jannes C, et al. Early predictors of mortality and morbidity after severe closed head injury. *J Neurotrauma* 2000 May;17(5):403-14.
57. Van Beek JG, Mushkudiani NA, Steyerberg EW, Butcher I, McHugh GS, Lu J, et al. Prognostic value of admission laboratory parameters in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007 Feb;24(2):315-28.
58. Saggar V, Mittal RS, Vyas MC. Hemostatic abnormalities in patients with closed head injuries and their role in predicting early mortality. J Neurotrauma 2009 Oct;26(10):1665-8.
59. Rovlias A, Kotsou S. The blood leukocyte count and its prognostic significance in severe head injury. *Surg Neurol* 2001 Apr;55(4):190-6.
60. Yang M, Guo Q, Zhang X, Sun S, Wang Y, Zhao L, et al. Intensive insulin therapy on infection rate, days in NICU, in-hospital mortality and neurological outcome in severe traumatic brain injury patients: a randomized controlled trial. *Int J Nurs Stud* 2009 Jun;46(6):753-8.
61. Bilotta F, Caramia R, Cernak I, Paoloni FP, Doronzio A, Cuzzone V, et al. Intensive insulin therapy after severe traumatic brain injury: a randomized clinical trial. *Neurocrit Care* 2008;9(2):159-66.
62. Bilotta F, Caramia R, Paoloni FP, Delfini R, Rosa G. Safety and efficacy of intensive insulin therapy in critical neurosurgical patients. *Anesthesiology* 2009 Mar;110(3):611-9.
63. Vespa PM. Intensive glycemic control in traumatic brain injury: what is the ideal glucose range? *Crit Care* 2008;12(5):175.

64. Vespa PM. The implications of cerebral ischemia and metabolic dysfunction for treatment strategies in neurointensive care. *Curr Opin Crit Care* 2006 Apr;12(2):119-23.
65. Morris GF, Juul N, Marshall SB, Benedict B, Marshall LF. Neurological deterioration as a potential alternative endpoint in human clinical trials of experimental pharmacological agents for treatment of severe traumatic brain injuries. Executive Committee of the International Selfotel Trial. *Neurosurgery* 1998 Dec;43(6):1369,72; discussion 1372-4.
66. Servadei F, Murray GD, Penny K, Teasdale GM, Dearden M, Iannotti F, et al. The value of the "worst" computed tomographic scan in clinical studies of moderate and severe head injury. European Brain Injury Consortium. *Neurosurgery* 2000 Jan;46(1):70,5; discussion 75-7.
67. Gentleman D. Causes and effects of systemic complications among severely head injured patients transferred to a neurosurgical unit. *Int Surg* 1992 Oct-Dec;77(4):297-302.
68. Manley G, Knudson MM, Morabito D, Damron S, Erickson V, Pitts L. Hypotension, hypoxia, and head injury: frequency, duration, and consequences. *Arch Surg* 2001 Oct;136(10):1118-23.
69. Rovlias A, Kotsou S. The influence of hyperglycemia on neurological outcome in patients with severe head injury. *Neurosurgery* 2000 Feb;46(2):335,42; discussion 342-3.
70. Salim A, Hadjizacharia P, Dubose J, Brown C, Inaba K, Chan LS, et al. Persistent hyperglycemia in severe traumatic brain injury: an independent predictor of outcome. *Am Surg* 2009 Jan;75(1):25-9.
71. Harhangi BS, Kompanje EJ, Leebeek FW, Maas AI. Coagulation disorders after traumatic brain injury. *Acta Neurochir* (Wien) 2008 Feb;150(2):165,75; discussion 175.
72. Huang SJ, Hong WC, Han YY, Chen YS, Wen CS, Tsan YS, et al. Clinical outcome of severe head injury in different protocol-driven therapies. *J Clin Neurosci* 2007 May;14(5):449-54.
73. Struchen MA, Hannay HJ, Contant CF, Robertson CS. The relation between acute physiological variables and outcome on the Glasgow Outcome Scale and Disability Rating Scale following severe traumatic brain injury. *J Neurotrauma* 2001 Feb;18(2):115-25.
74. Juul N, Morris GF, Marshall SB, Marshall LF. Intracranial hypertension and cerebral perfusion pressure: influence on neurological deterioration and outcome in severe head injury. The Executive Committee of the International Selfotel Trial. *J Neurosurg* 2000 Jan;92(1):1-6.
75. Kirkness CJ, Burr RL, Cain KC, Newell DW, Mitchell PH. Relationship of cerebral perfusion pressure levels to outcome in traumatic brain injury. *Acta Neurochir Suppl* 2005;95:13-6.
76. Kirkness CJ, Burr RL, Mitchell PH. Intracranial pressure variability and long-term outcome following traumatic brain injury. *Acta Neurochir Suppl* 2008;102:105-8.
77. Vik A, Nag T, Fredriksli OA, Skandsen T, Moen KG, Schirmer-Mikalsen K, et al. Relationship of 'dose' of intracranial hypertension to outcome in severe traumatic brain injury. *J Neurosurg* 2008 Oct;109(4):678-84.
78. Treggiari MM, Schutz N, Yanez ND, Romand JA. Role of intracranial pressure values and patterns in predicting outcome in traumatic brain injury: a systematic review. *Neurocrit Care* 2007;6(2):104-12.
79. Narayan RK, Greenberg RP, Miller JD, Enas GG, Choi SC, Kishore PR, et al. Improved confidence of outcome prediction in severe head injury. A comparative analysis of the clinical examination, multimodality evoked potentials, CT scanning, and intracranial pressure. *J Neurosurg* 1981 Jun;54(6):751-62.

80. Barelli A, Valente MR, Clemente A, Bozza P, Proietti R, Della Corte F. Serial multimodality-evoked potentials in severely head-injured patients: diagnostic and prognostic implications. *Crit Care Med* 1991 Nov;19(11):1374-81.
81. Wang JT, Young GB, Connolly JF. Prognostic value of evoked responses and event-related brain potentials in coma. *Can J Neurol Sci* 2004 Nov;31(4):438-50.
82. Schnakers C, Ledoux D, Majerus S, Damas P, Damas F, Lambermont B, et al. Diagnostic and prognostic use of bispectral index in coma, vegetative state and related disorders. *Brain Inj* 2008 Nov;22(12):926-31.
83. Sleigh JW, Havill JH, Frith R, Kersel D, Marsh N, Ulyatt D. Somatosensory evoked potentials in severe traumatic brain injury: a blinded study. *J Neurosurg* 1999 Oct;91(4):577-80.
84. Mazzini L, Pisano F, Zaccala M, Miscio G, Gareri F, Galante M. Somatosensory and motor evoked potentials at different stages of recovery from severe traumatic brain injury. *Arch Phys Med Rehabil* 1999 Jan;80(1):33-9.
85. Lew HL, Dikmen S, Slimp J, Temkin N, Lee EH, Newell D, et al. Use of somatosensory-evoked potentials and cognitive event-related potentials in predicting outcomes of patients with severe traumatic brain injury. *Am J Phys Med Rehabil* 2003 Jan;82(1):53,61; quiz 62-4, 80.
86. Carter BG, Butt W. Review of the use of somatosensory evoked potentials in the prediction of outcome after severe brain injury. *Crit Care Med* 2001 Jan;29(1):178-86.
87. Jennett B, Teasdale G, Braakman R, Minderhoud J, Knill-Jones R. Predicting outcome in individual patients after severe head injury. *Lancet.* 1976;1:1031-1034
88. Braakman R, Gelpke GJ, Habbema JD, Maas AI, Minderhoud JM. Systematic selection of prognostic features in patients with severe head injury. *Neurosurgery.* 1980;6:362-370
89. Choi SC, Ward JD, Becker DP. Chart for outcome prediction in severe head injury. *J Neurosurg.* 1983;59:294-297
90. Lokkeberg AR, Grimes RM. Assessing the influence of non-treatment variables in a study of outcome from severe head injuries. *J Neurosurg.* 1984;61:254-262
91. Braakman R, Habbema JD, Gelpke GJ. Prognosis and prediction of outcome in comatose head injured patients. *Acta Neurochir Suppl* (Wien). 1986;36:112-117
92. Choi SC, Narayan RK, Anderson RL, Ward JD. Enhanced specificity of prognosis in severe head injury. *J Neurosurg.* 1988;69:381-385
93. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head-injured patients. *J Neurosurg.* 1991;75:251-255
94. Fearnside MR, Cook RJ, McDougall P, McNeil RJ. The westmead head injury project outcome in severe head injury. A comparative analysis of pre-hospital, clinical and ct variables. *Br J Neurosurg.* 1993;7:267-279
95. Marmelak AN, Pitts LH, Damron S. Predicting survival from head trauma 24 hours after injury: A practical method with therapeutic implications. *J Trauma.* 1996;41:91-99
96. Quigley M, Vidovich D, Cantella D, Wilberger J, Maroon J, Diamond D. Defining the limits of survivorship after very severe head injury. *J Trauma.* 1997;42:7-10

97. Lang EW, Pitts LH, Damron SL, Rutledge R. Outcome after severe head injury: An analysis of prediction based upon comparison of neural network versus logistic regression analysis. *Neurol Res.* 1997;19:274-280

98. Ratanalert S, Chompikul J, Hirunpat S, Pheunpathom N. Prognosis of severe head injury: An experience in thailand. *Br J Neurosurg.* 2002;16:487-493

99. Hukkelhoven CWPM, Steyerberg EW, Habbema JDF, Farace E, Marmarou A, Murray GD, Marshall LF, Maas AIR. *Outcome after severe or moderate traumatic brain injury: Development and validation of a prognostic score based on admission characteristics.*

100. Cremer OL, Moons KG, van Dijk GW, van Balen P, Kalkman CJ. Prognosis following severe head injury: Development and validation of a model for prediction of death, disability, and functional recovery. *J Trauma.* 2006;61:1484-1491

101. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, Murray GD, Marmarou A, Roberts I, Habbema JD, Maas AI. Predicting outcome after traumatic brain injury: Development and international validation of prognostic scores based on admission characteristics. *PLoS Med.* 2008;5:e165; discussion e165

102. Steyerberg EW, Eijkemans MJ, Harrell FE,Jr, Habbema JD. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. *Med Decis Making* 2001 Jan-Feb;21(1):45-56.

103. Steyerberg EW, Harrell FE,Jr, Borsboom GJ, Eijkemans MJ, Vergouwe Y, Habbema JD. Internal validation of predictive models: efficiency of some procedures for logistic regression analysis. *J Clin Epidemiol* 2001 Aug;54(8):774-81.

104. Bleeker SE, Moll HA, Steyerberg EW, Donders AR, Derksen-Lubsen G, Grobbee DE, et al. External validation is necessary in prediction research: a clinical example. *J Clin Epidemiol* 2003 Sep;56(9):826-32.

105. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med* 1999 Mar 16;130(6):515-24.

106. Perel P, Edwards P, Wentz R, Roberts I. Systematic review of prognostic models in traumatic brain injury. *BMC Med Inform Decis Mak* 2006 Nov 14;6:38.

107. Mushkudiani NA, Hukkelhoven CW, Hernandez AV, Murray GD, Choi SC, Maas AI, et al. A systematic review finds methodological improvements necessary for prognostic models in determining traumatic brain injury outcomes. J Clin Epidemiol 2008 Apr;61(4):331-43.

108. Lingsma HF, Maas AIR, Steyerberg EW, Prognostication of moderate and severe traumatic brain injury. *Ned Tijdschr Geneeskd.* 2009;154(3):A739.

109. Machado SG, Murray GD, Teasdale GM. Evaluation of designs for clinical trials of neuroprotective agents in head injury. European Brain Injury Consortium. *J Neurotrauma* 1999 Dec;16(12):1131-8.

110. Roozenbeek B, Maas AI, Lingsma HF, Butcher I, Lu J, Marmarou A, et al. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med* 2009 Oct;37(10):2683-90.

111. Kent D, Hayward R. Subgroup analyses in clinical trials. *N Engl J Med* 2008 Mar 13;358(11):1199; author reply 1199-200.

112. Hayward RA, Kent DM, Vijan S, Hofer TP. Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. *BMC Med Res Methodol* 2006 Apr 13;6:18.

113. Hernandez AV, Steyerberg EW, Taylor GS, Marmarou A, Habbema JD, Maas AI. Subgroup analysis and covariate adjustment in randomized clinical trials of traumatic brain injury: a systematic review. *Neurosurgery* 2005 Dec;57(6):1244,53; discussion 1244-53.

114. Murray GD, Barer D, Choi S, Fernandes H, Gregson B, Lees KR, et al. Design and analysis of phase III trials with ordered outcome scales: the concept of the sliding dichotomy. *J Neurotrauma* 2005 May;22(5):511-7.

115. Mendelow AD, Gregson BA, Fernandes HM, Murray GD, Teasdale GM, Hope DT, et al. Early surgery versus initial conservative treatment in patients with spontaneous supratentorial intracerebral haematomas in the International Surgical Trial in Intracerebral Haemorrhage (STICH): a randomised trial. *Lancet* 2005 Jan 29-Feb 4;365(9457):387-97.

116. Maas AI, Murray G, Henney H,3rd, Kassem N, Legrand V, Mangelus M, et al. Efficacy and safety of dexanabinol in severe traumatic brain injury: results of a phase III randomised, placebo-controlled, clinical trial. *Lancet Neurol* 2006 Jan;5(1):38-45.

117. den Hertog HM, van der Worp HB, van Gemert HM, Algra A, Kappelle LJ, van Gijn J, et al. The Paracetamol (Acetaminophen) In Stroke (PAIS) trial: a multicenter, randomised, placebo-controlled, phase III trial. *Lancet Neurol* 2009 May;8(5):434-40.

118. McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, et al. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clin Trials* 2010 Jan;7(1):44-57.

119. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJ, Habbema JD. Validation and updating of predictive logistic regression models: a study on sample size and shrinkage. *Stat Med* 2004 Aug 30;23(16):2567-86.

120. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med* 2006 Jan 15;25(1):127-41.

121. Alexander S, Kerr ME, Kim Y, Kamboh MI, Beers SR, Conley YP. Apolipoprotein E4 allele presence and functional outcome after severe traumatic brain injury. *J Neurotrauma* 2007 May;24(5):790-7.

122. Jordan BD. Genetic influences on outcome following traumatic brain injury. *Neurochem Res* 2007 Apr-May;32(4-5):905-15.

123. Saatman KE, Duhaime AC, Bullock R, Maas AI, Valadka A, Manley GT, et al. Classification of traumatic brain injury for targeted therapies. *J Neurotrauma* 2008 Jul;25(7):719-38.

124. Maas AI. Standardisation of data collection in traumatic brain injury: key to the future? *Crit Care* 2009 Dec 16;13(6):1016.

6 Prognostic value of extracranial injury in traumatic brain injury

van Leeuwen N, Lingsma HF, Perel P, Lecky F, Roozenbeek B. Lu J, Shakur H, Weir J, Steyerberg EW, Maas AIR. Prognostic Value of Major Extracranial Injury in Traumatic Brain Injury: An Individual Patient Data Meta-analysis in 39,274 Patients. Submitted.

Abstract

Objective
Major extracranial injury (MEI) is common in Traumatic Brain Injury (TBI) patients. The aim of this study is to assess the prognostic value of MEI on mortality after TBI.

Design and setting
Individual patient data meta-analysis
Patients: Individual patients from three observational TBI studies (IMPACT), a randomized controlled trial (CRASH), and a trauma registry (TARN).

Methods
MEI was defined as extracranial injury with an Abbreviated Injury Score ≥ 3 or 'requiring hospital admission on its own'. We related MEI to mortality with logistic regression analysis, adjusted for age, GCS motor score and pupil reactivity and stratified by brain injury severity. We pooled odds ratios (ORs) with random effects meta-analysis methods.

Results
We included 39,274 patients in total, 17,136 with severe, 7,229 with moderate, and 14,909 with mild TBI. Mortality was 25% and 32% had MEI. MEI was a strong prognostic factor for mortality in TARN, with adjusted ORs and 95% confidence intervals (95%CI) of 2.81 (2.44-3.23) in mild, 2.18 (1.80-2.65) in moderate and 2.14 (1.95-2.35) in severe TBI patients. The prognostic effect was smaller in IMPACT and CRASH with pooled adjusted ORs and 95%CIs of 2.14 (0.93-4.91) in mild, 1.46 (1.14-1.85) in moderate and 1.18 (1.03-1.55) in severe TBI patients. When patients who died within 6 hours after injury were excluded from TARN, the effects of MEI were comparable to those observed in IMPACT and CRASH.

Conclusions
MEI is an important prognostic factor for mortality in patients with TBI. However, the strength of the effect is smaller in patients with more severe brain injury. Also the strength of the effect decreases when only considering patients who survive the early phase after injury, instead of considering all patients, starting from the time of injury.

Introduction

Major extracranial injury (MEI) is frequently present in patients with traumatic brain injury (TBI). Relatively few studies have however focused on the effect of MEI on mortality after TBI. Most studies concerning TBI and MEI have investigated patients with extracranial trauma, with or without TBI. These studies show that the coexistence of traumatic brain injury with extracranial injury is associated with both increased mortality and morbidity.[1-4]

In contrast, there is no consensus on the degree to which the presence of MEI worsens outcome in TBI patients. Some studies demonstrate that outcome mainly depends on the severity of the primary cerebral damage and is not worsened by the presence of extracranial injuries.[5,6] Other studies suggest that the presence of MEI does predict a poorer outcome in TBI patients.[7-10] Differences between studies might be due to patient population, setting and study design. Determining the importance of MEI in outcome after TBI has relevance for understanding and potentially improving the patient pathway, and for improving prognostic models that might be used to benchmark care,[4] or to inform relatives and medical decisions.

We report a collaborative analysis on a large number TBI patients with and without documented MEI, including data from the International Mission on Prognosis and Clinical Trial design in TBI (IMPACT) study, the Medical Research Council Corticosteroid Randomization after Significant Head Injury (MRC CRASH) trial, and the Trauma Audit & Research Network (TARN) registry. Our aim was to determine the role of MEI as a prognostic factor for mortality after TBI. We hypothesize that the presence of MEI is associated with higher mortality in patients with TBI.

Methods

Patient population and data collection

We included individual patient data from the International Mission on Prognosis and Clinical Trail design in TBI (IMPACT) study, the Medical Research Council Corticosteroid Randomization after Significant Head Injury (MRC CRASH) trial, and the Trauma Audit & Research Network (TARN).

IMPACT combines individual patient data from randomized controlled trials (RCTs) and three observational studies in moderate and severe TBI, mainly from the US and Europe. Here we focused on the three observational studies (the European Brain Injury Consortium study (EBIC), the UK four center study (UK4), and the Traumatic Coma Databank (TCDB)), as the presence of MEI was not an exclusion criterium for these studies. Patients were enrolled in these studies between 1984 and 1995.

The CRASH trial is a trial with broad inclusion criteria studying the effect of corticosteroids on death and disability after head injury. CRASH was conducted in both high and

low/middle income countries. In CRASH we analyzed low/middle income countries and high income countries separately, as trauma organizations may be different.[7] CRASH enrolled patients between 1999 and 2005.

TARN is a hospital based trauma registry in England and Wales including all patients with trauma resulting in immediate admission to hospital for three days or longer or death. From these, we selected TBI patients defined as having an Abbreviated Injury Scale for the Head Region of 3 or higher, which was not resulting from scalp laceration, scalp avulsion or penetrating injury. The patients from TARN included in this study were enrolled between 1990 and 2009.

Detailed descriptions of all the studies and data collection and management can be found in previous publications.[11-13]

Outcome and major extracranial injury

The primary outcome examined in this analysis was mortality at six months in IMPACT and CRASH and discharge mortality in TARN. Where 6 month outcome was missing or systematically not recorded in IMPACT, the mortality at three months was substituted instead. CRASH had also 14 day mortality available.

Major Extracranial Injury (MEI) was defined as 'Abbreviated Injury Scale (AIS) $\geq$ 3' or 'an injury requiring hospital admission on its own'.

Statistical analyses

The strength of the association between MEI and mortality was analyzed univariably and multivariably using binary logistic regression models. We adjusted for core prognostic parameters: age, GCS motor score (1=makes no movements, 2=extension to painful stimuli, 3=abnormal flexion to painful stimuli, 4 =flexion/withdrawal to painful stimuli, 5=localizes painful stimuli, 6=obeys commands) and pupil reactivity (1= both responsive, 2=one responsive, 3=both unresponsive) at admission. We also adjusted for hypotension to better understand the pathway of the prognostic effect of MEI. In IMPACT we additionally adjusted for study since it consists of three studies. In CRASH we also adjusted for treatment since there was a significant treatment effect.

Results were expressed as odds ratio for mortality with MEI compared to absent MEI, with 95% confidence intervals. An overall summary measure was derived using random effects meta-analysis (Der Simonian-Laird pooling). TARN was not included in the pooled analysis because of the different nature of the study and the different time point of the outcome. Tests of heterogeneity were performed to assess consistency of effects across studies. Forest plots were used to display consistency of findings across the datasets.

We calculated partial R^2 statistics to indicate the amount of variance explained by MEI, both univariable and multivariable. In CRASH and IMPACT we corrected the univariable and multivariable R^2s for the variance explained by study and treatment.

Absolute risks of patients with and without MEI were calculated from the models

by taking the mean of the probabilities predicted by the multivariable models, again stratified for brain injury severity.

Missing data were imputed for the motor score of the Glasgow Coma Scale (GCS), pupil reactivity and MEI with multiple imputation using all relevant prognostic factors and outcome. Imputations were done separately for TARN, CRASH and IMPACT.

Analyses were performed with R statistical software 2.7.1 (R Foundation for Statistical Computation, Vienna) using packages Rmeta and Design, and SPSS 15.0 (SPSS Inc, Chicago).

Sensitivity analyses

In preliminary analysis we found a large difference between IMPACT and CRASH versus TARN in terms of the effect of MEI on outcome. We hypothesized that this might be due to the different setting (TBI studies versus a trauma registry), the different distribution of TBI severity across the studies (only moderate and severe TBI in IMPACT, many mild TBI patients in TARN), or the different time point of outcome assessment (discharge versus 6 month). We tested these hypotheses by three approaches.

1. We tested for interaction between MEI and brain injury severity (GCS), by adding an interaction term between MEI and GCS to the binary logistic regression model containing age, GCS motor score, pupil reactivity, MEI and GCS as main effects. We assed the p-value of the interaction term and consequently stratified the analyses for brain injury severity, defining mild TBI as Glasgow Coma Scale (GCS) 13-15, moderate TBI as GCS 9-12 and severe traumatic brain injury as GCS 3-8.
2. We excluded the patients from TARN who died within 6 hours after injury since the majority of these patients is not likely to be included in IMPACT or CRASH.
3. We analyzed in CRASH both 14 day and 6 month mortality.

Results

Patient population

We included 2,218 patients from IMPACT (791 from UK4, 603 from TCDB, and 824 from EBIC), 9,554 from CRASH (7,205 from low/middle income countries, and 2,349 from high income countries), and 27,504 from TARN. This resulted in 39,274 patients for the analysis. For all variables missing was less than 10%, except for TARN where 90% of the pupil reactivity data was missing since this variable was only recorded from 2005 onwards.

Patient characteristics

The majority of the patients (17,132, 44%) had severe TBI. A total of 7,229 (18%) had moderate and 14,909 (38%) had mild TBI. The IMPACT study included mainly severe TBI patients (81%) and TARN mainly mild (43%) and severe (42%) TBI patients. In CRASH the distribution of brain injury severity was more equal (30% mild, 30% moderate, 40% severe). In IMPACT mortality was 41%, compared to 24% in CRASH and 28% in TARN. In IMPACT 41% of the patients had MEI, in CRASH this was 23% and in TARN and 34%. MEI was observed more frequently in patients with severe TBI (30-46%), than in those with mild TBI (14-41%). (Table 6.1).

Major Extracranial Injury and mortality

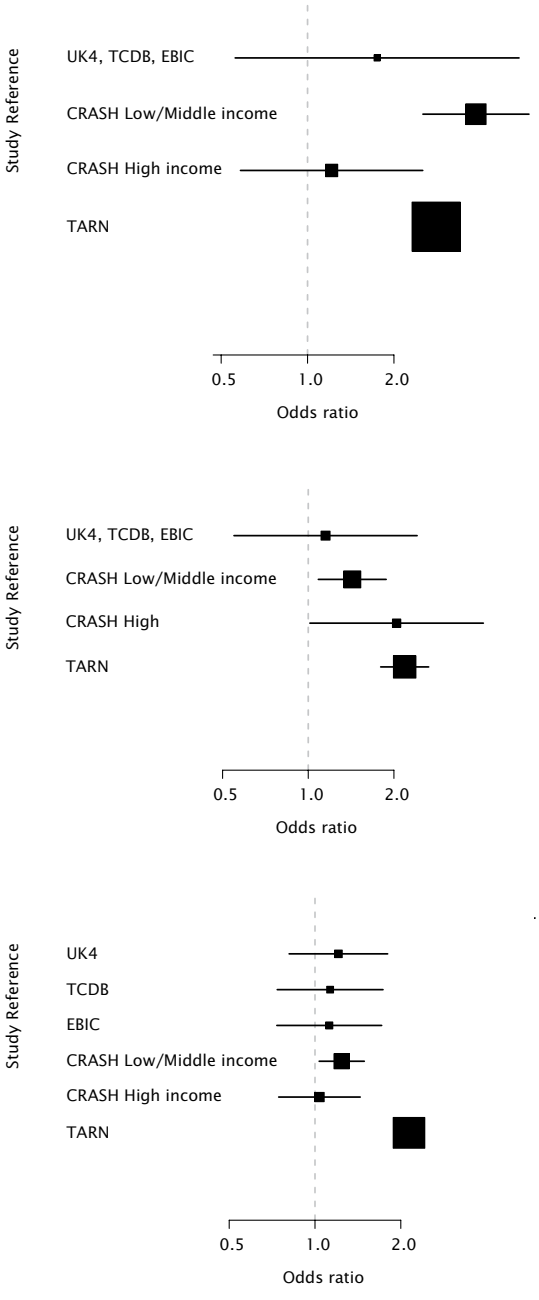
We found a moderate prognostic effect of MEI in IMPACT and CRASH with pooled adjusted ORs and 95% confidence intervals (95%CIs) of 2.14 (0.93-4.91) in mild, 1.46 (1.14-1.85) in moderate and 1.18 (1.03-1.55) in severe TBI patients. In TARN MEI was a strong prognostic factor for mortality, with adjusted odds ratios (OR) and 95%CIs of 2.81 (2.44-3.23) in mild, 2.18 (1.80-2.65) in moderate and 2.14 (1.95-2.35) in severe TBI patients (Figure 6.1 and Table 6.2). The unadjusted ORs were all smaller than adjusted ORs, indicating that the effect of MEI on mortality was independent of other predictors of mortality.

Table 6.1 Patient characteristics of 3 studies from the IMPACT study, the CRASH trial and the TARN registry.

	Age	**GCS score**	**Motor score**	**Pupillary reactivity***	**Major extracranial injury**	**Mortality**
	Median (25th–75th percentile)	*Mild (GCS 13–15)* *Moderate (GCS 9–12)* *Severe (GCS 3–8)*	*None* *Extension* *Abnormal Flexion* *Normal flexion* *Localize/obeys* *Untestable/ missing*	*Both responsive* *One responsive* *Both unresponsive*	*Yes*	*Dead*
UK4 (n=791)	36 (22–55)	24 (3%) 83 (11%) 684 (87%)	113 (14%) 85 (11%) 37 (5%) 141 (18%) 221 (28%) 194 (26%)	434 (55%) 113 (14%) 244 (31%)	303 (38%)	359 (45%)
TCDB (n=603)	26 (21–40)	22 (4%) 45 (8%) 536 (89%)	136 (23%) 107 (18%) 74 (12%) 121 (20%) 134 (22%) 31 (5%)	299 (50%) 55 (9%) 249 (41%)	280 (46%)	264 (44%)
EBIC (n=822)	37 (24–59)	73 (9%) 168 (20%) 581 (71%)	150 (18%) 80 (10%) 55 (7%) 113 (14%) 281 (34%) 143 (17%)	532 (65%) 80 (10%) 210 (26%)	316 (38%)	281 (34%)
CRASH LOW/ MIDDLE INCOME (n=7,205)	32 (24–45)	2108 (29%) 2331 (32%) 2766 (38%)	356 (5%) 403 (6%) 531 (7%) 891 (12%) 5024 (70%) 0 (0%)	6135 (85%) 450 (6%) 620 (9%)	1694 (23%)	1854 (26%)
CRASH HIGH INCOME (n=2,349)	37 (24–54)	760 (32%) 551 (24%) 1038 (44%)	429 (18%) 112 (5%) 128 (5%) 290 (12%) 1390 (59%) 0 (0%)	1965 (84%) 147 (6%) 237 (10%)	522 (23%)	469 (20%)
TARN (n=27,504)	39 (24–60)	11922 (43%) 4051 (15%) 11531 (42%)	4117 (15%) 838 (3%) 973 (4%) 1449 (5%) 11892 (43%) 8235 (30%)	21548 (78%) 1630 (6%) 4326 (16%)	9452 (34%)	7673 (28%)

*Pupil reactivity in TARN was imputed for 90% of the patients

Figure 6.1 Forest plots showing the strength of the adjusted association between major extracranial injury and mortality in mild (upper), moderate (middle) and severe (lower) TBI patients



Adjusting the effect of extracranial injury for hypotension led to a small decrease of the prognostic effect (ORs decreasing by 0.1-0.4) of MEI, indicating that hypotension indeed explains part of the relationship between extracranial injury and outcome. Hypotension itself was a strong prognostic factor for mortality, independent of MEI (adjusted ORs 2.9 to 3.6).

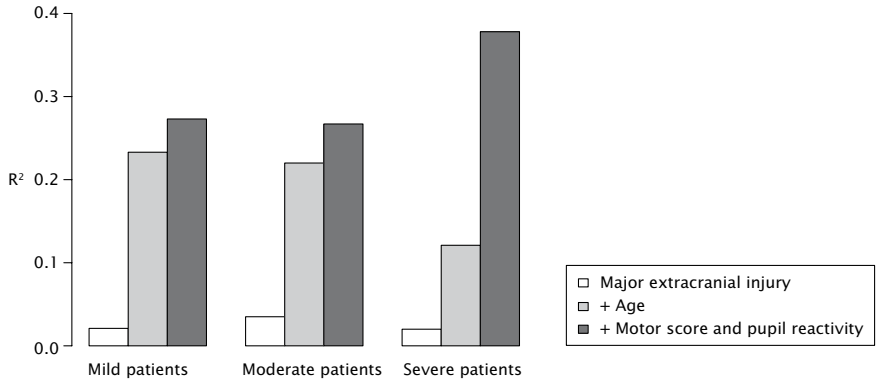
The prognostic value of MEI in terms of univariable R^2 (Figure 6.2) varied from 0.0% (in severe patients in IMPACT and CRASH) to 3.4% (in severe patients in TARN), and was considerably smaller than the prognostic value of core predictors as age, GCS motor score and pupil reactivity.

Table 6.2 Associations between major extracranial injury (versus no and minor extracranial injury) and mortality

	Mild TBI (GCS 13–15)[1]				**Moderate TBI (GCS 9–12)**	
	N Mortality (%)	*N Major extra-cranial injury (%)*	*Unadjusted OR (95% CI)[2]*	*Adjusted OR[3] (95% CI)*	*N Mortality (%)*	*N Major extra-cranial injury (%)*
UK4	10 (42)	9 (38)			34 (41)	33 (40)
TCDB	6 (27)	9 (41)	1.14 (0.46–2.86)	1.75 (0.56–5.46)	9 (20)	24 (53)
EBIC	11 (15)	19 (26)			26 (16)	51 (30)
CRASH L-M income	112 (5)	309 (14)	3.96 (2.64–5.94)	3.86 (2.52–5.91)	348 (15)	530 (22)
CRASH High income	64 (8)	115 (15)	1.39 (0.73–2.64)	1.21 (0.58–2.52)	81 (15)	111 (20)
Pooled CRASH & IMPACT	203 (7)	453 (15)	1.96 (0.84–4.59)	2.14 (0.93–4.91)	498 (16)	737 (23)
TARN	1132 (10)	3147 (26)	2.24 (1.98–2.54)	2.81 (2.44–3.23)	764 (19)	1178 (29)

1. GCS = Glasgow Coma Scale
2. OR (95% CI) = Odds ratio (95% confidence interval)
3. Adjusted analyses – adjusted for age, pupil reactivity and GCS motor score. In IMPACT and CRASH also adjusted for respectively study and treatment.
In mild and moderate TBI the logistic regression analyses were done together for UK4, TCDB and EBIC because of low numbers of patients

Figure 6.2 The prognostic value of major extracranial injury (MEI), univariable and in combination with age and brain injury severity (GCS motor score and pupil reactivity), expressed in percentage explained variance (R^2)



		Severe TBI (GCS 3–8)			
Unadjusted OR (95% CI)	*Adjusted OR (95% CI)*	*N Mortality (%)*	*N Major extra-cranial injury (%)*	*Unadjusted OR (95% CI)*	*Adjusted OR (95% CI)*
		315 (46)	261 (38)	0.90 (0.66–1.23)	1.21 (0.81–1.80)
0.70 (0.39–1.26)	1.15 (0.55–2.41)	249 (47)	247 (46)	0.98 (0.70–1.38)	1.13 (0.74–1.73)
		244 (42)	246 (42)	0.79 (0.56–1.10)	1.12 (0.73–1.71)
1.48 (1.15–1.92)	1.43 (1.09–1.88)	1393 (50)	852 (30)	1.16 (0.99–1.37)	1.24 (1.03–1.49)
1.16 (0.66–2.04)	2.04 (1.01–4.12)	324 (31)	296 (30)	0.98 (0.73–1.31)	1.04 (0.75–1.44)
1.13 (0.73–1.75)	1.46 (1.14–1.85)	2525 (45)	1904 (34)	1.00 (0.86–1.15)	1.18 (1.03–1.55)
1.68 (1.42–1.80)	2.18 (1.80–2.65)	5777 (50)	5127 (45)	1.92 (1.78–2.07)	2.14 (1.95–2.35)

Absolute risks

In CRASH and IMPACT, the increase in absolute risk on mortality associated with MEI was 8% (6% vs. 14%) in mild, 4% (15% vs. 19%) in moderate and 1% (45% vs. 46%) in severe TBI patients. The prevalence of MEI in TBI patients was larger in TARN for all brain injury severities than in IMPACT and CRASH, as was the increase in absolute risks on mortality. The increase in absolute risk on mortality associated with MEI was 8% (7% vs. 15%) in mild, 9% (16% vs. 25%) in moderate and 16% (43% vs. 59%) in severe TBI patients in TARN (Table 6.3).

Table 6.3 Absolute risks on mortality stratified for MEI vs. no MEI and for TBI severity groups in IMPACT & CRASH vs. TARN.

		Mild TBI patients	Moderate TBI patients	Severe TBI patients
IMPACT & CRASH	No major extra-cranial injury	5.5% (5.2–5.8)	14.8% (0.142–0.153)	44.8% (44.1–45.6)
	Major extracranial injury	13.9% (12.6–15.2)	18.7% (0.177–0.198)	45.5% (44.5–46.6)
TARN	No major extra-cranial injury	7.4% (7.2–7.4)	16.4% (15.8–17.1)	42.9% (42.2–43.6)
	Major extra-cranial injury	15.3% (14.7–1.58)	24.8% (23.5–26.0)	59.1% (58.3–59.8)

Differences between CRASH, IMPACT and TARN

There was a significant interaction between MEI and brain injury severity in CRASH ($p<0.001$) and TARN ($p=0.029$) but not in IMPACT.

Since we found, also after stratification, a considerable difference in the prognostic effect of MEI between IMPACT-CRASH and TARN across all TBI severities, we excluded 912 patients from TARN who died within 6 hours after injury since the majority of these patients would not have been included in IMPACT or CRASH. This resulted in decreased ORs of MEI for mortality: 2.4 in mild, 1.8 in moderate and 1.6 in severe TBI (IMPACT and CRASH: 2.1 in mild, 1.6 in moderate and 1.2 in severe TBI).

To assess the difference between IMPACT-CRASH and TARN further, we analyzed 14 day mortality in CRASH. In low/middle income countries MEI was less strongly related to 14 day mortality than to 6 month mortality (ORs 0.1-1 point lower for 14 day mortality). In high income countries however, effects were opposite (ORs 0.1 to 0.4 points higher for 14 day mortality).

Discussion

Our study shows that MEI is a prognostic factor in patients with TBI. However the strength of the effect interacts with brain injury severity and varies by the population studied, and by study design. In TBI patients included in a registry (TARN) MEI is strongly associated with mortality after adjustment for age, GCS motor score and pupil reactivity. In patients included in TBI studies (broadly selected RCTs or observational studies) the incremental prognostic value of MEI compared to known predictors of mortality is limited, particularly in more severe TBI.

We found a large difference in prognostic effect between TARN and IMPACT and CRASH. The larger effect in TARN was largely explained by inclusion of patients who died before or shortly after admission. The ORs in IMPACT and CRASH thus could be interpreted as the effect of MEI when a TBI patient survives the early stage (first hours) after trauma. The effect in TARN could be interpreted as the effect of MEI in the unselected TBI population. For example: a victim of a road traffic accident with severe TBI and MEI has an odds for mortality 2.14 fold that of a similar patient without MEI. When this patient survives the early stage, the prognostic effect of MEI is reduced to a 1.18 fold increased risk.

Our study shows that the magnitude of the effect of MEI on mortality depends on the study design. This is also explanation for the disagreement in the literature about the prognostic effect of MEI. Studies demonstrating that outcome is not worsened by MEI only included (often severe) patients admitted to an intensive-care unit.[5,6] These studies are mostly comparable to IMPACT and CRASH with regard to study population and results. The studies showing an effect of MEI in TBI patients, obtained the data from a Trauma Registry like TARN.[8-10]

The prognostic effect of MEI thus depends on the population studied. This means that it is also dependent on the application of a prognosis in a clinical setting. For counseling of relatives of severe TBI patients in the hospital for example, MEI is more likely to be a highly relevant prognostic factor in the Emergency Department than a few hours later if the patient has survived the immediate risk of death from haemorrhage caused by major extracranial injury and has been admitted to intensive care. Thus, this study demonstrates that it is important not only to formulate a clear research question but also to define the specific patient population, which is often not done in prognostic research. To interpret results of a prognostic study and to determine applicability to a particular setting it is important to be aware of the study population and design.

We reported absolute risks in the different studies and the different strata of patients, which further provide some relevant clinical insights. For example patients with mild TBI & MEI have a similar risk on mortality to one with moderate TBI and no MEI. Absolute risks on mortality were higher in TARN than in IMPACT and CRASH across all TBI severities. This is probably partly due to the previously mentioned difference in pa-

tient population. Further, differences in mortality between the studies might be caused by differences in health care system and resources (low/middle income countries in CRASH) and by the time of data collection (varying between 1984 for TCDB and 2009 for the most recent patients in TARN).

It might be expected that MEI is more associated with early mortality than with late mortality. This is supported by our finding that ORs decrease when excluding early deaths in TARN. In CRASH we analyzed both 14 day and 6 month mortality, with inconsistent results. In high income countries the ORs for 14 day mortality were indeed higher than those for 6 month mortality, in low/middle income countries it was the other way round. An explanation might be that within high income countries trauma deaths after 14 days are rare, while lack of resources and also a greater level of underlying comorbidity make late trauma deaths more prevalent in low/middle income countries. MEI will have an impact there because it will often cause immobility, resulting from e.g. limb and pelvic fractures, which may cause mortality in less resourced settings. In general, the prognostic effect of MEI was larger in low/middle income countries, which might be partly explained by structure and processes of care (e.g. longer times to admissions, less resources). These findings illustrate the necessity to take resources and post acute facilities into account when including patients in TBI studies from regions where resources may be more limited. This is particularly important as a tendency has been noted for pharmaceutical companies and researchers to involve centers from other regions of the world in TBI studies, because of higher patient potential and lower cost.[14]

The unadjusted ORs were all smaller than adjusted ORs. This means that the effect of MEI on mortality was not explained by other predictors of mortality. Adjusting only for brain injury severity lead to a small decrease in the effect of MEI since patients with MEI have more severe brain injury, which is also related to mortality. Adjusting for age lead to an increase of the effect of MEI since patients with MEI are younger on average, which is related to less mortality.

Hypotension explained a small part of the association between MEI and mortality. This was expected since systemic injuries can cause major bleedings and thus hypotension. The finding that the ORs of MEI change only very little after adjustment for hypotension and that hypotension is also a strong predictor of mortality independent of MEI suggests that the threshold values for defining hypotension may be too restrictive, or that other mechanisms, such as inflammatory response to multiple injuries, play a role in the relationship between extracranial injury and mortality.

Previous studies have shown that TBI increases the risk of both mortality and morbidity in the general trauma population.[1-3] We find that the presence of MEI is also associated with increased mortality in patients with TBI. Whether this effect may be greater or smaller than in the general trauma population can not be answered from our study, since we only included patients with TBI. Within the TARN registry work is currently ongoing to analyse the effect of TBI in the general trauma population. It is however an

artificial distinction between patients with TBI and patients with MEI. In clinical practice there are patients with trauma and they have often multiple injuries, both extracranial and intracranial. Based on our results and findings from previous studies we would provisionally conclude that both MEI and TBI carry a high risk of mortality, and that a combination of both further increases this risk. The relation is however multidimensional and interaction effects exist with the severity of brain injury.

We used a very simple definition of MEI, since extracranial injury severity was reported differently in each dataset. Analysis of the prognostic value of the full AIS or Injury Severity Scale (ISS) in TBI patients may provide additional insights in the mechanism of effect. On the other hand the definition of AIS ≥ 3 we use is quite common, easy to use in practice and showed to discriminate well.

A limitation of our study is the imputation of missing variables, although imputation is better than deleting missing variables.[15] For TARN, where pupil reactivity was imputed in the majority of patients, we compared the results in complete cases with the results in the imputed data, which gave similar results.

It could be argued that another limitation is the heterogeneity between the three studies used in the meta-analysis, in timing of outcome, setting and patient population. However this heterogeneity allowed us to disentangle the effects of MEI on mortality and to explain to some extent the conflicting results in the current literature.

The strength of this study is obviously the many patients included in the study. Also, the meta-analysis is based on individual patient data.

In conclusion, this meta-analysis demonstrates that MEI is a prognostic factor for increasing mortality in patients with TBI. However, the strength of the effect is smaller in patients with more severe brain injury. Also the strength of the effect decreases when only considering patients who survive the early phase after injury, instead of considering all patients, starting from the time of injury.

Acknowledgments

This research was funded by National Institute of Health (NS-42691).

References

1. McMahon CG, Yates DW, Campbell FM, Hollis S, Woodford M: Unexpected contribution of moderate traumatic brain injury to death after major trauma. *J Trauma* 1999, 47(5):891-895.
2. Gennarelli TA, Champion HR, Sacco WJ, Copes WS, Alves WM: Mortality of patients with head injury and extracranial injury treated in trauma centers. *J Trauma* 1989, 29(9):1193-1202.
3. Gennarelli TA, Champion HR, Copes WS, Sacco WJ: Comparison of mortality, morbidity, and severity of 59,713 head injured patients with 114,447 patients with extracranial injuries. *J Trauma* 1994, 37(6):962-968.
4. Patel HC, Bouamra O, Woodford M, King AT, Yates DW, Lecky FE: Trends in head injury outcome from 1889 to 2003 and the effect of neurosurgical care: an observational study. *Lancet* 2005, 366(9496):1538-44.
5. Sarrafzadeh AS, Peltonen EE, Kaisers U, Kuchler I, Lanksch WR, Unterberg AW: Secondary insults in severe head injury - do multiply injured patients do worse? *Crit Care Med* 2001, 29(6):1116-1123.
6. Heinzelmann M, Platz A, Imhof HG: Outcome after acute extradural haemstoma, influence of additional injuries and neurological complications in the ICU. *Elsevier Ltd* 1996, 27(5):345-349.
7. Perel P, Arango M, Clayton T, for the MRC CRASH Trial Collaborators: Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ* 2008, 336(7641):425-429.
8. Lefering R, Paffrath T, Linker R, Bouillon B, Neugebauer EAM: Head injury and outcome – What influence do concomitant injuries have? *J Trauma* 2008, 65(5):1036-1044.
9. Jacobs B, Beems T, Stulemeijer M: Outcome prediction in mild traumatic brain injury: age and clinical variables are stronger predictors than CT abnormalities. *J Neurotrauma* 2010, 27(4):655-68.
10. Ho KM, Burrell M, Rao S: Extracranial injuries are important in determining mortality of neurotrauma. *Crit Care Med* 2010, 38(7):1562-8.
11. Marmarou A, Lu J, Butcher I: IMPACT database of traumatic brain injury: design and description. *J Neurotrauma* 2007, 24(2):239-250.
12. Edwards P, Farrell B, Lomas G, for the CRASH Trial Pilot Study Collaborative Group: The MRC CRASH Trial: study design, baseline data, and outcome in 1000 randomised patients in the pilot phase. *Emerg Med J* 2002,19(6):510-514.
13. Lecky F, Woodford M, Yates DW: Trends in trauma care in England and Wales 1989-97. UK Trauma Audit and Research Network. *Lancet* 2000, 355(9217):1771-1775.
14. Maas AI, Roozenbeek B, Manley GT: Clinical trials in traumatic brain injury: past experience and current developments. *Neurotherapeutics* 2010, 7(1):115-26.
15. Schafer J L, Graham JW: Missing data: our view of the state of the art. *Psychol Methods* 2002, 7(2):147-177.

7 Prediction of respiratory insufficiency in Guillain-Barré syndrome

Walgaard C, Lingsma HF, Ruts L, Drenthen J, van Koningsveld R, Garssen MJP, van Doorn PA, Steyerberg EW, Jacobs BC. Prediction of respiratory insufficiency in Guillain-Barré syndrome. *Annals of Neurology* 2010; 67:781-787

Abstract

Objective
Respiratory insufficiency is a frequent and serious complication of the Guillain-Barré syndrome (GBS). We aimed to develop a simple but accurate model to predict the chance of respiratory insufficiency in the acute stage of disease based on clinical characteristics available at hospital admission.

Methods
Mechanical ventilation (MV) in the first week of admission was used as an indicator of acute stage respiratory insufficiency. Prospectively collected data from a derivation cohort of 397 GBS patients were used to identify predictors of MV. A multivariable logistic regression model was validated in a separate cohort of 191 GBS patients. Model performance criteria comprised discrimination (area under receiver operating curve, AUC) and calibration (graphically). A scoring system for clinical practise was constructed from the regression coefficients of the model in the combined cohorts.

Results
In the derivation cohort 22% needed MV in the first week of admission. Days between onset of weakness and admission, MRC sumscore and presence of facial and/or bulbar weakness were the main predictors of MV. The prognostic model had a good discriminative ability (AUC 0.84). In the validation cohort 14% needed MV in the first week of admission and both calibration and discriminative ability of the model were good (AUC 0.82). The scoring system ranged from zero to seven with corresponding chances of respiratory insufficiency from 1 to 91%.

Interpretation
This model accurately predicts development of respiratory insufficiency within one week in patients with GBS, using clinical characteristics available at admission. After further validation, the model may assist in clinical decision-making, e.g. on patient transfer to an ICU.

Introduction

Respiratory insufficiency is a life threatening manifestation of the Guillain-Barré syndrome (GBS) that occurs in 20-30% of patients and is associated with poor functional outcome.[1-4] Respiratory insufficiency often develops insidiously in GBS. This may explain the relatively high frequency of nocturnal and emergency intubations.[5,6] Moreover, 60% of intubated patients develop major complications including pneumonia, sepsis, and pulmonary embolism.[7] Delaying intubation may increase the risk of pneumonia due to aspiration and worsens outcome.[8, 9] Specific treatments for GBS may not have reduced mortality and length of hospital stay among ventilated GBS patients.[10] Prediction of respiratory insufficiency is important to triage patients to the appropriate unit (general ward or ICU) and avoid respiratory distress.

Previous studies identified various risk factors for respiratory insufficiency in GBS, including cranial nerve deficits,[5, 11-13] disability grade on admission,[8, 11, 14] rapid progressive motor weakness,[5, 14] areflexia,[8] descending weakness,[15] dysautonomia,[5] EMG features of nerve conduction block,[11, 16] positive CMV serology,[17] anti-GQ1b antibodies,[12] and increased liver enzymes.[11, 14] Only one validated model for the prediction of respiratory insufficiency in clinical practice is available, based on information about the vital capacity and the ratio of the proximal to distal peroneal nerve compound muscular amplitude potential.[11] In this study electrophysiological testing generally was done within six days after admission, while most intubations in GBS occur in the first week of admission. Prediction models for respiratory insufficiency should be available as early as possible, preferably at hospital admission, and based on readily available information. Previous studies showed that clinical parameters in the progressive phase are highly predictive of the clinical course of GBS.[18, 19]

The aim of the current study was to develop a simple and accurate model using clinical features available at hospital admission to predict the occurrence of respiratory insufficiency in the acute stage of GBS. Model performance was validated in an independent cohort of patients with GBS.

Methods

Patients

Prospectively collected data from a cohort of 397 patients with GBS were used to identify risk factors for respiratory insufficiency in the acute stage. This derivation cohort consisted of patients included in two treatment trials and one pilot study. The first study was a multicenter, double-blind randomized controlled trial that compared plasma exchange (PE) with intravenous immunoglobulin (IVIg) for which 147 patients were included between 1985 and 1991.[20] The second study was a pilot study in 25 Dutch patients to determine the additional therapeutic effect of methylprednisolone with IVIg.[21] In the third study this combination was tested in a multicenter, double-blind randomized controlled trial including 225 patients between 1994 and 2000.[22] Most patients were randomized in Dutch hospitals, the others in two German and two Belgium hospitals. Same inclusion and exclusion criteria were used in these three studies. Inclusion criteria were fulfilment of the NINDS diagnostic criteria for GBS[23], being unable to walk unaided ten metres across an open space (GBS disability score three or more) and onset of weakness within two weeks before randomization. Exclusion criteria were age below six years, previous GBS, known severe allergic reaction to properly matched blood products, pregnancy, known selective IgA deficiency, previous steroid therapy, severe concurrent disease, inability to attend follow-up, or contraindications for corticosteroid treatment (not in first trial).

To validate the model we used prospectively collected data from a cohort of 191 patients enrolled in one pilot study[24] and one observational study. The pilot study determined the additional therapeutic effect of mycophenolate mofetil to IVIg and MP and for this study 27 patients were included between 2002 and 2005. The same in- and exclusion criteria were used as in the derivation cohort. Regarding the observational study, 168 GBS patients were included between 2005 and 2008 to assess pain and autonomic dysfunction. This study also included patients with a milder course (able to walk throughout the course of the disease) (n=33) or the Miller Fisher syndrome (n=18). Patients with additional central nervous system involvement (n=4) were excluded. All patients in the validation cohort were included in Dutch hospitals. Patients who were intubated before the day of admission in the participating hospital were excluded from the derivation and validation set.

Data collection

Baseline characteristics (age, gender, pre-existing chronic pulmonary disease), preceding diarrhoea or symptoms of an upper respiratory tract infection, day of onset of weakness, cranial nerve dysfunction, MRC sumscore, GBS disability score, and sensory deficit at study entry were collected prospectively. Most patients entered the study within one day of hospital admission (interquartile range 0-1 days). The Medical

Research Council (MRC) sumscore is defined as the sum of MRC scores of six different muscles measured bilaterally, resulting in a score ranging from 0 (tetraplegic) to 60 (normal).[25] The GBS disability score is a widely accepted scale to assess functional status of GBS patients, ranging from zero (normal) to six (death; Supplementary Text).[26] Additional serological screening was performed to determine recent infections with *Campylobacter jejuni*, cytomegalovirus (CMV), Epstein-Barr virus (EBV), and *Mycoplasma pneumonia* and antibodies to the gangliosides GM1, GD1a, and GQ1b. The serum samples used were obtained within four weeks from onset of weakness and before start of treatment. Liver enzymes (ASAT, ALAT) were considered abnormal when the ratio between measured values and the upper limit of normal was > 1.5.

Endpoint

The main endpoint in our study was mechanical ventilation (MV) in the first week of hospital admission, as an indicator of acute stage respiratory insufficiency. The decision to intubate was based on the discretion of the treating physician.

Statistical analysis

Potential predictors of MV within one week were first considered in logistic regression models in the derivation cohort. Predictors that were statistically significant in univariable analysis and available at admission were further analysed in a multivariable logistic regression model. A backward stepwise selection procedure was done with a p value of 0.1 as selection criterion. Variables with more than 15% missing data were omitted from analysis. Missing values in other variables were imputed using a multiple imputation method.[27] Odds ratios of univariable analysis were compared between the imputed dataset and the unimputed dataset. Model performance was quantified with respect to discrimination (area under the receiver operating curve, AUC). The AUC ranges from 0.5 to 1.0 for sensible models. Internal validity of the model was assessed using bootstrapping techniques, and included the selection of predictors. The model was applied to the validation dataset for external validation. Model performance in the validation set was quantified with respect to discrimination (AUC) and calibration. Calibration was assessed graphically by plotting observed frequencies against predicted probabilities. A final scoring system was constructed based on the regression coefficients of the multivariable model in a dataset where the derivation and validation sets were combined for larger reliability. Statistical analyses were done with SPSS for Windows, and R statistical software.

Results

In the derivation cohort, 20 (5%) of the 397 patients were intubated before referral to one of the participating hospitals and excluded from the current study. Eighty three (22%) of the remaining 377 patients required MV in the first week of hospital admission and 16 (4%) after the first week. In the validation cohort three (2%) of the 191 GBS patients were excluded because of intubation before referral to a trial hospital. Twenty seven (14%) of the remaining 188 patients required MV in the first week of hospital admission and two (1%) after the first week.

Strong associations with MV in the first week from admission were found for the following clinical parameters available at hospital admission: MRC sumscore, GBS disability score, rate of initial disease progression (indicated by the number of days between onset of weakness and hospital entry), facial weakness, bulbar weakness, and areflexia of arms and legs (Table 7.1).

Facial weakness and bulbar weakness elaborately overlapped in these GBS patients and were combined as a single predictor for multivariable analysis. Areflexia was left out of the multivariable logistic regression analysis because data were missing in 30% of patients. For the remaining parameters data were missing in less than 3% and were imputed using multiple imputation. In multivariable logistic regression analysis strong predictors of MV in the first week of hospital admission were MRC sumscore at admission ($p<0.001$), days between onset of weakness and admission ($p<0.001$), and facial and/or bulbar weakness at admission ($p<0.001$). GBS disability score was not associated with respiratory insufficiency in multivariable analysis. A model to predict respiratory insufficiency was constructed using these three statistically significant clinical parameters and showed a very good discriminative ability (AUC = 0.84) and good calibration (Figure 7.1). After excluding the 18 patients intubated within 24 hours from hospital admission, the discriminative ability remains very good (AUC = 0.83).

The model developed in the derivation cohort was further tested in the independent validation cohort and showed an equally good discriminative ability (AUC = 0.82) and calibration (Figure 7.1).

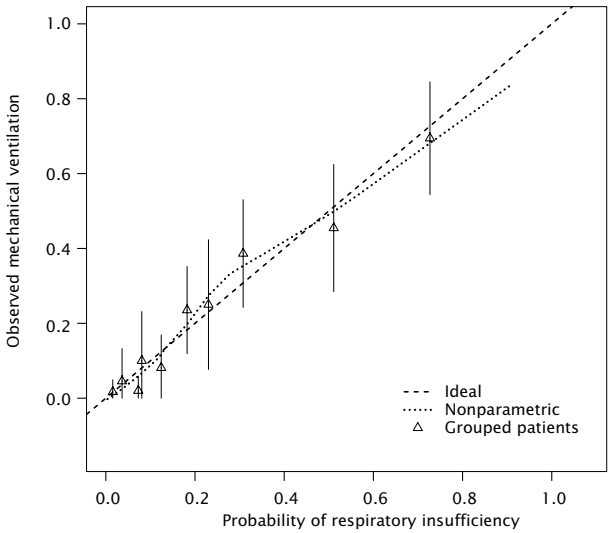
Table 7.1 Characteristics of the derivation set of 377 patients with GBS in relation to mechanical ventilation in the first week of hospital admission.

	N	MV (%) within 1 week	Univariable OR (95% CI)	P value	Multivariable OR (95% CI)	P value
Demographic features						
Total	377	83(22%)				
Age (years)				0.3		
≤ 40	131/377	24(18%)	1			
40–60	109/377	29(27%)	1.6(0.9–3.0)			
> 60	137/377	30(22%)	1.3(0.7–2.3)			
Gender (male)	209/377	49(23%)	1.2(0.7–2.0)	0.5		
Chronic pulmonary disease	11/243	1(9%)	0.5(0.1–4.1)	0.5		
Neurological deficits at entry						
Onset weakness – entry (days)				<0.001		<0.001
> 7	96/376	7(7%)	1		1	
4–7	147/376	28(19%)	3.0(1.3–7.2)		3.5(1.3–9.3)	
≤ 3	133/376	47(35%)	6.9(3.0–16)		9.2(3.4–25)	
Cranial nerve involvement						
Facial and/or bulbar weakness	119/377	39(33%)	2.4(1.4–3.9)	0.001	3.9(2.1–7.3)	<0.001
Bulbar weakness	37/377	18(49%)	4.0(2.0–8.1)	<0.001		
Facial weakness	112/377	8(32%)	1.7(0.7–4.2)	0.002		
Ophthalmoplegia	25/377	36(32%)	2.2(1.3–3.6)	0.2		
MRC sumscore				<0.001		<0.001
60–51	48/375	1(2%)	1		1	
50–41	180/375	26(14%)	8.1(1.1–61)		6.3(0.8–50)	
40–31	77/375	16(21%)	12(1.6–97)		9.8(1.2–81)	
30–21	46/375	22(48%)	44(5.6–346)		29(3.4–246)	
≤ 20	24/375	18(75%)	144(16–1281)		87(9.1–830)	
GBS disability score				<0.001		0.2
3	92/377	6(7%)	1		1	
4 or 5	285/377	77(27%)	5.3(2.2–13)		1.9(0.7-5)	
Sensory deficits	244/371	53(22%)	1.1(0.6–1.8)	0.8		
Pain	181/375	37(20%)	0.8(0.5–1.4)	0.5		
Areflexia (both arms and legs)	149/265	47(32%)	2.9(1.5–5.4)	0.001		
Infection and serology						
*Symptoms of preceding infection**						
Diarrhea	85/375	18(21%)	0.9(0.5–1.7)	0.8		
Upper respiratory tract infection	137/369	28(20%)	0.9(0.5–1.5)	0.5		
Infection serology†						
Campylobacter jejuni	97/333	24(25%)	1.3(0.7–2.2)	0.4		
Cytomegalovirus	42/332	14(33%)	2.0(1.0–4.0)	0.06		
Epstein-Barr virus	42/332	10(24%)	1.1(0.5–2.4)	0.8		
Mycoplasma pneumoniae	17/332	3(18%)	0.8(0.2–2.7)	0.7		
Anti-ganglioside IgM/IgG antibodies†						
GM1	72/333	11(15%)	0.6(0.3–1.2)	0.1		
GD1a	16/333	6(38%)	2.2(0.8–6.4)	0.1		
GQ1b	21/333	6(29%)	1.5(0.6–3.9)	0.5		
Liver dysfunction†						
ALAT	55/357	17(31%)	1.7(0.9-3.1)	0.1		
ASAT	37/357	12(32%)	1.7(0.8-3.7)	0.1		

MV(%) = mechanical ventilated in the first week after hospital admission. MRC = Medical Research Counsel. Ig = immunoglobulin. ALAT = alanine aminotransferase. ASAT = aspartate aminotransferase.
* Symptoms of infection in the four weeks preceding the onset of weakness.
† Using pre-treatment serum samples obtained at entry.

Figure 7.1 Calibration plots for the developed model in the derivation (a) and validation (b) cohort.

a. Derivation cohort (N=377)



b. Validation cohort (N=188)

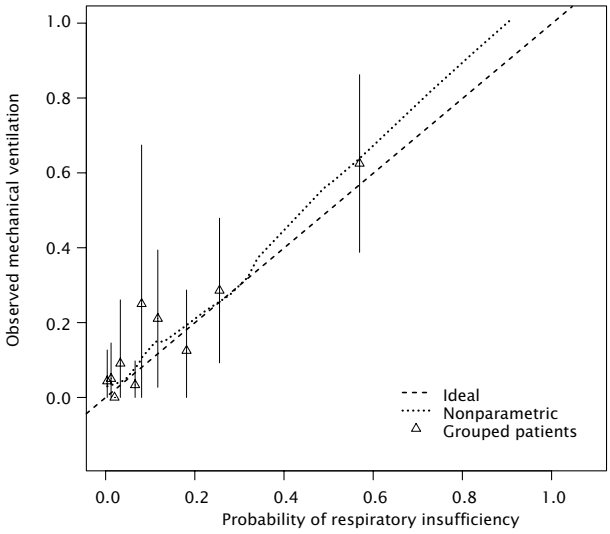
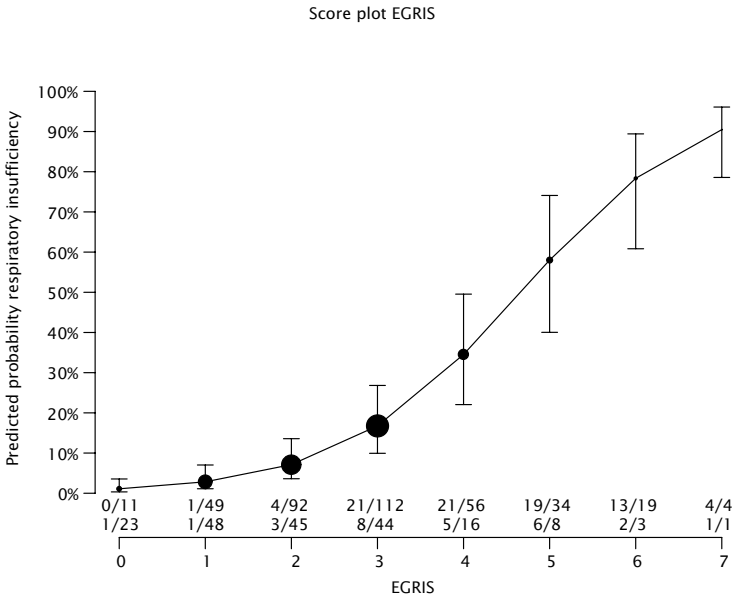


Figure 7.2 Predicted probability of respiratory insufficiency and observed percentage of mechanical ventilation (MV) in derivation and validation cohorts according to the EGRIS.

The black line reflects the predicted probability of respiratory insufficiency derived from the combined cohorts. Size of bullets in the graph reflects size of patient group with corresponding EGRIS score in the combined cohorts (N=565). Above x-axis are the numbers of patients requiring MV of patients with a defined EGRIS in each cohort.



The Erasmus GBS Respiratory Insufficiency Score (EGRIS) was based on the regression coefficients of the three predictors in the multivariable model in the combined cohorts (n=565). Scores ranged from zero to seven, with five categories for the MRC sumscore at admission, three categories for days between onset of weakness and hospital entry, and two categories for facial and/or bulbar weakness at admission, with corresponding chances for respiratory insufficiency within one week ranging from 1 to 91% (Table 7.2 and Figure 7.2). Median duration of MV was 27 days (interquartile range 12-53 days). The duration of MV was not associated with the EGRIS (data not shown).

Table 7.2 Erasmus GBS Respiratory Insufficiency Score (EGRIS).

	Categories	Score
Days between onset of weakness and hospital admission	> 7 days	0
	4–7 days	1
	≤ 3 days	2
Facial and/or bulbar weakness at hospital admission	Absence	0
	Presence	1
MRC sumscore at hospital admission	60–51	0
	50–41	1
	40–31	2
	30–21	3
	≤ 20	4
EGRIS		0–7

As an example we consider two hypothetical patients at the emergency department with a MRC sumscore of 25 (3 points). The first patient had weakness since 1 day (2 points) and facial weakness (1 point), while the second patient had weakness since 10 days (0 points) and no facial or bulbar weakness (0 points). The EGRIS for the first patient is 6 points, corresponding to a risk of respiratory insufficiency in the first week of admission of 77% (95% CI 61-89%; Fig). The EGRIS for the second patient is 3, corresponding to a much lower risk of respiratory insufficiency of 17% (95% CI 10-27%; Fig). For further illustration, patients were divided into three clinically relevant risk groups (Table 7.3). Only 10 (4%) of 268 patients with a low EGRIS (0-2) had respiratory insufficiency in the first week, compared to 45 (65%) of 69 patients with a high EGRIS (5-7).

Table 7.3 Risk categories for respiratory insufficiency according to EGRIS.
Probability of respiratory insufficiency in the first week of hospital admission in the derivation, validation, and combined sets stratified for EGRIS and expressed as number of mechanically ventilated patients/total number of patients (%). 95% CI: 95% confidence interval for combined sets.

	Derivation set	Validation set	Combined sets
Low risk (EGRIS 0–2)	5 / 152 (3%)	5 / 116 (4%)	10 / 268 (4%, 95% CI 1–6%)
Intermediate risk (EGRIS 3–4)	42 / 168 (25%)	13 / 60 (22%)	55 / 228 (24%, 95% CI 19–30%)
High risk (EGRIS 5–7)	36 / 57 (63%)	9 / 12 (75%)	45 / 69 (65%, 95% CI 54–76%)
Total	83 / 377 (22%)	27 / 188 (14%)	110 / 565 (19%, 95% CI 16–23%)

Discussion

In the current study a prognostic model was developed that accurately predicts respiratory insufficiency in the early stage of GBS using three clinical characteristics readily available at hospital admission. The most important predictors of MV in the first week of admission were the rate of disease progression, indicated by the number of days between onset of weakness and hospital admission, the MRC sumscore, and the presence of facial or bulbar weakness. A multivariable prediction model proved valid in an independent cohort of GBS patients. The proposed eight point EGRIS accurately predicts the probability of respiratory insufficiency in the first week of hospital admission in individual GBS patients, ranging from 1 to almost 90%.

Our study confirms findings by others that respiratory insufficiency in GBS is associated with a high GBS disability score at hospital admission[8, 11, 14], a rapid disease progression[5, 14], presence of cranial nerve deficit[5, 11-13], and areflexia[8]. In our cohort no data were available on dysautonomia[5] and descending weakness[15], both previously reported to be predictors of respiratory insufficiency. Also very limited information was available regarding vital capacity or electrophysiology at admission, so we were unable to validate the model of Durand et al.[11] Vital capacity and electrophysiological measurements at hospital admission may further improve the EGRIS. Measurement of vital capacity may be confounded by bilateral facial weakness, occurring in more than half of GBS patients, and a low vital capacity may reflect impending or established respiratory insufficiency rather than an increased risk of future respiratory insufficiency. Moreover, electrophysiology may not be available at admission, and the results may be highly variable in the first week of GBS.[28]

The clinical risk factors for acute stage respiratory insufficiency partly differ from those for a poor long-term outcome. In a previous study, using the same derivation cohort of patients, the ability to walk unaided after six months depended on age, presence of preceding diarrhoea, and GBS disability score at two weeks after admission.[19] A low GBS disability score was associated with respiratory insufficiency, however lost significance in the multivariable regression analysis together with MRC sumscore. In addition, MV is incorporated in the GBS disability score rendering this score less suitable to predict respiratory insufficiency. Age and preceding diarrhoea were not associated with respiratory insufficiency in the current and previous studies. Probably, age influences the capacity to recover more than the disease severity in the acute phase. Preceding diarrhoea in GBS is frequently caused by infections with *C. jejuni*, and associated with a severe, pure motor, and axonal variant.[29, 30] In this form of GBS the proximal muscles and cranial nerves are relatively spared, which may explain why this phenotype is not predisposing to respiratory insufficiency. The frequency of respiratory insufficiency in GBS patients may be lower in Japan, where the *C. jejuni* or axonal form of GBS is predominant,[31] in contrast to Western countries, where the demyelinating forms are predominant. In a

Japanese cohort of patients with severe GBS only 10% needed MV,[15] compared to 19% in the combined cohorts from the current study. This may support the hypothesis that in GBS severe demyelination is associated with respiratory insufficiency.[11, 16]

The EGRIS has some limitations. First, the derivation and validation cohorts differed with respect to the proportion of patients requiring MV (22% versus 14%). The lower frequency of MV in the validation cohort is explained by different inclusion criteria which allowed inclusion of patients with mild forms of GBS. However, the EGRIS model developed in the derivation cohort performed equally well in the validation cohort, demonstrating its wide clinical applicability. Second, the endpoint in our study was MV which is only an indirect indicator of respiratory insufficiency. In fact, the decision to intubate is relatively arbitrary and based on the discretion of the treating physician, supported by previously published general criteria for intubation in GBS.[32] Our results could be biased by the long time span of data acquisition, during which the practice of intubation may have changed. However, no trend was found in our dataset regarding the frequency of MV and the performance of EGRIS. More detailed information is required in future studies regarding respiratory parameters, especially at the time of intubation. Third, model development focussed on the prediction of MV in the first week after admission. The first week of the disease reflects probably the most unpredictable period of GBS with the highest frequency of acute respiratory insufficiency. In our cohorts 3% of the patients were intubated after the first week of admission. The EGRIS predicted the need of MV, irrespective of the time point during clinical course, accurately with an AUC of 0.80. Fourth, time from onset of weakness to hospital admission is probably influenced by social factors. The time from onset of weakness to loss of ambulation is possibly less arbitrary but was not documented in our cohorts. Since most patients were included in the trials shortly after losing ambulation, the moment of study entry usually equals that of losing ambulation. Lastly, most patients included in our studies were Dutch Caucasians and the EGRIS may not be applicable in patients from other geographical areas or ethnic origin. Prospective studies in more diverse populations of patients are required to determine the general validity of the EGRIS.

How to apply the EGRIS in clinical practice? Based on the model, respiratory insufficiency in the first week of admission cannot be excluded in an individual patient with GBS. Even in the low-risk subgroup, with an EGRIS score of 2 or less, 4% (95% CI of 1 to 6%) of the patients developed respiratory insufficiency, which required MV. This underlines that the clinical course in individual GBS patients can by highly variable and stresses the importance of regular pulmonary function monitoring (VC, respiratory frequency), initially every 2-6 hours in the progressive phase and every 6-12 hours in the plateau phase.[30] Nonetheless, the EGRIS model holds great promise to be used as a practical tool to inform patients and their families and assist physicians in decision-making. For examples, patients with an increased risk of respiratory insufficiency may be transferred to an ICU, or be considered for early elective intubation.

Acknowledgements

This work was supported by a scientific research grant from the Dutch Prinses Beatrix Fonds (grant PBF WAR07-28) and a Clinical Fellowship grant from the Netherlands organization for health research and development (grant ZonMW 907-00-111). We thank the Dutch Guillain-Barré Study Group and the Plasma Exchange/Sandoglobulin Guillain-Barré Syndrome Trial Group for providing the data for this analysis.

References

1. Dhar R, Stitt L, Hahn AF. The morbidity and outcome of patients with Guillain-Barré syndrome admitted to the intensive care unit. *Journal of the neurological sciences*. 2008;264:121-128
2. Fletcher DD, Lawn ND, Wolter TD, Wijdicks EF. Long-term outcome in patients with Guillain-Barré syndrome requiring mechanical ventilation. *Neurology.* 2000;54:2311-2315
3. Rees JH, Thompson RD, Smeeton NC, Hughes RA. Epidemiological study of Guillain-Barré syndrome in south east England. *Journal of neurology, neurosurgery, and psychiatry.* 1998;64:74-77
4. Winer JB, Hughes RA, Osmond C. A prospective study of acute idiopathic neuropathy. I. Clinical features and their prognostic value. *Journal of neurology, neurosurgery, and psychiatry.* 1988;51:605-612
5. Lawn ND, Fletcher DD, Henderson RD et al. Anticipating mechanical ventilation in Guillain-Barré syndrome. *Archives of neurology.* 2001;58:893-898
6. Wijdicks EF, Henderson RD, McClelland RL. Emergency intubation for respiratory failure in Guillain-Barré syndrome. *Archives of neurology.* 2003;60:947-948
7. Henderson RD, Lawn ND, Fletcher DD et al. The morbidity of Guillain-Barré syndrome admitted to the intensive care unit. *Neurology.* 2003;60:17-21
8. Cheng BC, Chang WN, Chang CS et al. Predictive factors and long-term outcome of respiratory failure after Guillain-Barré syndrome. *The American journal of the medical sciences.* 2004;327:336-340
9. Orlikowski D, Sharshar T, Porcher R et al. Prognosis and risk factors of early onset pneumonia in ventilated patients with Guillain-Barré syndrome. *Intensive care medicine.* 2006;32:1962-1969
10. Souayah N, Nasar A, Suri MF, Qureshi AI. National trends in hospital outcomes among patients with Guillain-Barré syndrome requiring mechanical ventilation. *J Clin Neuromuscul Dis*. 2008;10:24-28
11. Durand MC, Porcher R, Orlikowski D et al. Clinical and electrophysiological predictors of respiratory failure in Guillain-Barré syndrome: a prospective study. *Lancet neurology.* 2006;5:1021-1028
12. Kaida K, Kusunoki S, Kanzaki M et al. Anti-GQ1b antibody as a factor predictive of mechanical ventilation in Guillain-Barré syndrome. *Neurology.* 2004;62:821-824
13. Orlikowski D, Terzi N, Blumen M et al. Tongue weakness is associated with respiratory failure in patients with severe Guillain-Barré syndrome. *Acta neurologica Scandinavica.* 2008
14. Sharshar T, Chevret S, Bourdain F, Raphael JC. Early predictors of mechanical ventilation in Guillain-Barré syndrome. *Critical care medicine*. 2003;31:278-283
15. Funakoshi K, Kuwabara S, Odaka M et al. Clinical predictors of mechanical ventilation in Fisher/Guillain-Barré overlap syndrome. *Journal of neurology, neurosurgery, and psychiatry.* 2009;80:60-64
16. Durand MC, Lofaso F, Lefaucheur JP et al. Electrophysiology to predict mechanical ventilation in Guillain-Barré syndrome. *Eur J Neurol.* 2003;10:39-44
17. Visser LH, van der Meche FG, Meulstee J et al. Cytomegalovirus infection and Guillain-Barré syndrome: the clinical, electrophysiologic, and prognostic features. Dutch Guillain-Barré Study Group. *Neurology*. 1996;47:668-673
18. The prognosis and main prognostic indicators of Guillain-Barré syndrome. A multicenter prospective study of 297 patients. The Italian Guillain-Barré Study Group. *Brain.* 1996;119 (Pt 6):2053-2061

19. van Koningsveld R, Steyerberg EW, Hughes RA et al. A clinical prognostic scoring system for Guillain-Barré syndrome. *Lancet neurology.* 2007;6:589-594
20. van der Meche FG, Schmitz PI. A randomized trial comparing intravenous immune globulin and plasma exchange in Guillain-Barré syndrome. Dutch Guillain-Barré Study Group. *The New England journal of medicine.* 1992;326:1123-1129
21. Treatment of Guillain-Barré syndrome with high-dose immune globulins combined with methylprednisolone: a pilot study. The Dutch Guillain-Barré Study Group. *Annals of neurology.* 1994;35:749-752
22. van Koningsveld R, Schmitz PI, Meche FG et al. Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barré syndrome: randomised trial. *Lancet.* 2004;363:192-196
23. Asbury AK, Cornblath DR. Assessment of current diagnostic criteria for Guillain-Barré syndrome. *Annals of neurology.* 1990;27 Suppl:S21-24
24. Garssen MP, van Koningsveld R, van Doorn PA et al. Treatment of Guillain-Barré syndrome with mycophenolate mofetil: a pilot study. *Journal of neurology, neurosurgery, and psychiatry.* 2007;78:1012-1013
25. Kleyweg RP, van der Meche FG, Schmitz PI. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barré syndrome. *Muscle & nerve.* 1991;14:1103-1109
26. Hughes RA, Newsom-Davis JM, Perkin GD, Pierce JM. Controlled trial prednisolone in acute polyneuropathy. *Lancet.* 1978;2:750-753
27. Steyerberg EW. *Clinical Prediction Models.* 1st ed: Springer-Verlag New York Inc., 2008
28. Hadden RD, Cornblath DR, Hughes RA et al. Electrophysiological classification of Guillain-Barré syndrome: clinical associations and outcome. Plasma Exchange/Sandoglobulin Guillain-Barré Syndrome Trial Group. *Annals of neurology.* 1998;44:780-788
29. Dourado ME, Duarte RC, Ferreira LC et al. Anti-ganglioside antibodies and clinical outcome of patients with Guillain-Barré Syndrome in northeast Brazil. *Acta neurologica Scandinavica.* 2003;108:102-108
30. van Doorn PA, Ruts L, Jacobs BC. Clinical features, pathogenesis, and treatment of Guillain-Barré syndrome. *Lancet neurology.* 2008;7:939-950
31. Hughes RA, Cornblath DR. Guillain-Barré syndrome. *Lancet.* 2005;366:1653-1666
32. Ropper AH, Kehne SM. Guillain-Barré syndrome: management of respiratory failure. *Neurology.* 1985;35:1662-1665

8 Prediction of outcome in Guillain-Barré syndrome

Walgaard C, Lingsma HF, Ruts L, van Doorn PA, Steyerberg EW, Jacobs BC. Early recognition of poor prognosis in Guillain-Barré syndrome. Submittted

Abstract

Background
Guillain-Barré syndrome (GBS) has a highly diverse clinical course and outcome, yet patients are treated with standard therapy. Patients with poor prognosis may benefit from additional treatment, provided they can be identified early, when nerve degeneration is potentially reversible and treatment is most effective. We developed a clinical prognostic model for early prediction of outcome in GBS applicable for clinical practice and future therapeutic trials.

Methods
Data collected prospectively from a derivation cohort of 397 GBS patients were used to identify risk factors of being unable to walk at 4 weeks, 3 months and 6 months. Potential predictors of poor outcome (unable to walk unaided) were considered in univariable and multivariable logistic regression models. The clinical model was based on the multivariable logistic regression coefficients of selected predictors and externally validated in an independent cohort of 158 GBS patients.

Results
High age, preceding diarrhea and low MRC sum score at hospital admission and at one week were independently associated with being unable to walk at 4 weeks, 3 months and 6 months (all P 0.05-0.001). The model can be used at admission and at day 7 of admission, the latter having a better predictive ability for the 3 endpoints; the area under the receiver operating curve (AUC) is 0.84-0.87 and at admission the AUC is 0.73-0.77. The model proved to be valid in the validation cohort.

Conclusions
A clinical prediction model applicable early in the course of the disease accurately predicts outcome in GBS.

Introduction

Guillain-Barré syndrome (GBS) is a monophasic polyradiculoneuropathy with a highly variable clinical severity and outcome. Intravenous immunoglobulin (IVIg) and plasma exchange are beneficial in patients who are severely affected, however one-third recovers incompletely.[1] These patients need more effective treatment, but the clinical diversity and the rarity of the disease hamper good and well-powered RCTs in this patient group. To early identify patients with a poor outcome, who are eligible for additional treatment, prognostic models are needed. Prognostic models can also increase the power of therapeutic studies by adjusting for prognostic factors.[2] Ultimately, such prediction models can be used to individualize therapy in accordance with the expected outcome.

Previous studies have identified patient characteristics associated with poor outcome in GBS.[3-9] The Erasmus GBS Outcome Score (EGOS) is a prognostic model based on age, diarrhea and GBS disability score at two weeks after hospital admission that accurately predicts the chance of being able to walk independently at 6 months.[7] However, prognostic models to optimize treatment in GBS should be applicable in the earliest phase of the disease, when treatment is considered to be most effective. Such models should also be designed to predict the primary endpoints used in most treatment trials in GBS; i.e. the clinical recovery on the GBS disability score at 4 weeks.[10-14] The aim of the current study was to develop readily applicable prognostic models for accurate selection of patients with a poor prognosis, based on clinical information available in the first week of hospital admission.

Methods

Patients

Data collected prospectively from a cohort of 397 GBS patients were used to identify predictors for outcome. This derivation cohort consisted of patients, who had been included in two treatment trials and one pilot study. The first study was a multicentre double-blind randomized controlled trial; this included 147 patients between 1985 and 1991, that compared plasma exchange (PE) with intravenous immunoglobulin (IVIg).[11] The second study was a pilot study in 25 Dutch patients to determine the additional therapeutic effect of methylprednisolone (MP) to IVIg.[15] This combination was tested in the third study: a multicentre double-blind randomized controlled trial in 225 patients included between 1994 and 2000.[14] Most patients were included in Dutch hospitals, the others in two German and two Belgian hospitals. All three studies used the same inclusion and exclusion criteria. Inclusion criteria were fulfillment of the NINDS diagnostic criteria for GBS,[16] inability to walk unaided ten meters across an open space (GBS disability score three or more) and onset of weakness within two weeks before randomization. Exclusion criteria were age below 6 years, pregnancy, previous GBS, known severe allergic reaction to properly matched blood products, known selective IgA deficiency, previous steroid therapy, severe concurrent disease, inability to attend follow-up, or contraindications for corticosteroid treatment (not in first trial).

To validate the model we used data collected prospectively from a cohort of 191 patients enrolled in a pilot study[17] and an observational study[18] in GBS patients, both performed in the Netherlands. The pilot study evaluated the additional therapeutic effect of mycophenolate mofetil to IVIg and MP in 27 patients included between 2002 and 2005. The same inclusion and exclusion criteria were used as in the derivation cohort. Between 2005 and 2008 164 GBS patients were included in the observational study, which assessed pain and autonomic dysfunction (GRAPH study).[18] Patients with a mild form of GBS (able to walk throughout the course of the disease) (N=33) were also included in this study, but not used for validation. Approval was received by an ethical standards committee on human experimentation for all the studies mentioned above. Written informed consent to participate in one of the studies was obtained from all patients.

Data collection

Data were collected prospectively at hospital admission on the following: age, gender, diarrhea or symptoms of an upper respiratory tract infection in the 4 weeks preceding onset of weakness, day of onset of weakness, cranial nerve dysfunction, Medical Research Counsel (MRC) sum score[19], GBS disability score[20], and sensory deficit. In addition, data on the MRC sum score and GBS disability score were collected prospectively at day 7 of hospital admission. The MRC sum score is defined as the sum of MRC scores of 6 different muscles measured bilaterally, which results in a score ranging from 0

(tetraplegic) to 60 (normal; appendix e-1)[19]. The GBS disability score is a widely accepted scale for assessing the functional status of GBS patients; it ranges from 0 (normal) to 6 (death; appendix e-1)[20]. Serological screening was performed to identify recent infections with Campylobacter jejuni, cytomegalovirus (CMV). The serum samples were obtained within 4 weeks of onset of weakness and before start of treatment.

Outcome measures

This study used walking ability as outcome measure. Poor outcome was defined as the inability to walk unaided 10 meters across an open space (GBS disability score of 3 or higher). Outcome was assessed at 4 weeks, 3 months and 6 months after inclusion in one of the studies. An additional outcome measure in this study was the improvement of one or more points on the GBS disability score in the first 4 weeks after inclusion. No improvement was considered as poor outcome. Both outcome measures have been used as primary endpoint in previous treatment trials in GBS.

Model development

Potential prognostic factors of outcome at 4 weeks, 3 months and 6 months after inclusion were first analyzed in the derivation cohort by univariable logistic regression analysis. Statistically significant predictors for poor outcome at all time points were further analyzed for their independent predictive value using multivariable logistic modeling.

Missing values were imputed using a multiple imputation method.[21] Odds ratios (OR) were used to express the strength of prognostic effects and were compared between the imputed and the complete case analyses. Predictive value was also measured using the likelihood ratio chi square test (LR chi^2), to account for the prevalence of the predictor. Variables which added significant predictive information were selected for use in a multivariable model.

The model was fitted using the ability to walk unaided at 4 weeks after hospital admission as outcome measure. The model was constructed based on the multivariable logistic regression coefficients in the derivation dataset.

Predictive performance of the model was quantified with respect to discrimination (area under the receiver operating curve, AUC). The AUC ranges from 0.5–1.0 for sensible models. The internal validity of the model was assessed by bootstrapping techniques, including both the selection of predictors and estimation of the coefficients[21]. The model was applied to the validation dataset for external validation. Model performance in the validation set was quantified with respect to discrimination (AUC) and calibration. Calibration was assessed graphically by plotting observed frequencies against predicted probabilities.

Statistical analyses used SPSS version 15.0 for Windows, Stata version 11, and R statistical software (version 2.7, using the Design library).

Results

Three (<1%) of the 397 patients in the derivation cohort died in the first week after hospital admission and were excluded from the current study. In this cohort the primary endpoint was missing at 3 months for 3 (<1%) patients and at 6 months for 12 (3%) patients. Fifty-five percent had a poor outcome at 4 weeks, 30% at 3 months and 19% at 6 months after hospital admission. In the validation cohort, none of the patients died in the first 4 weeks of follow-up. Due to the slightly different follow-up structure of the observational study, outcome was unavailable for 38 (24%) patients at 4 weeks, 14 (9%) patients at 3 months and 7 (4%) patients at 6 months after hospital admission. These patients were excluded from the study. Of the remaining patients in the validation cohort 54% had poor outcome at 4 weeks, 29% at 3 months and 15% at 6 months after hospital admission.

In univariate analysis 6 predictors of outcome – at 4 weeks, 3 months and 6 months- were identified: age, disease progression (expressed as number of days between onset of weakness and hospital entry), MRC sum score and GBS disability score, diarrhea in the 4 weeks preceding GBS, and *C. jejuni* serology (all P<0.05-0.001) (table 8.1). *C. jejuni* serology was excluded for multivariable analysis because in clinical practice serology results will be difficult to obtain shortly after hospital admission. For further modeling, the MRC sum score was selected over the GBS disability score, because the model using the MRC sum score had a substantially better performance (LR statistic 69.75 versus 46.49 at admission and 195.27 versus 154.35 at one week). Disease progression lost its predictive ability when analyzed in a multivariable model with age, diarrhea, and MRC sum score. The results of the multivariable analyses of the remaining prognostic factors are shown in table 8.2.

Table 8.1 Risk of poor outcome, defined as inability to walk unaided at 4 weeks, 3 months and 6 months after entry to the hospital, according to potential predictors in the derivation set of 394 GBS patients based on univariable regression analysis.

Inability to walk unaided at		**4 weeks**		**3 months**		**6 months**	
	N	OR (95% CI)	P Value	OR (95% CI)	P Value	OR (95% CI)	P Value
Total	394						
Demographic features							
Age (years)			0.003		0.01		<0.001
≤ 40	138	1 (ref)		1 (ref)		1 (ref)	
40–60	114	1.9 (1.2–3.2)		1.6 (0.9–2.8)		2.2 (1.0–4.6)	
> 60	142	2.2 (1.4–3.5)		2.3 (1.3–3.9)		4.0 (2.1–7.9)	
Gender (male)	215	0.9 (0.6–1.3)	NS	1.2 (0.8–1.9)	NS	1.2 (0.7–2.0)	NS
Clinical severity at admission							
Onset weakness-admission (per day increase)		0.9 (0.9–1.0)	0.02	0.9 (0.8–1.0)	0.003	0.9 (0.8–1.0)	0.006
Bulbar weakness	43	1.4 (0.7–2.7)	NS	1.1 (0.5–2.2)	NS	0.6 (0.3–1.0)	0.05
Facial weakness	125	1.2 (0.8–1.9)	NS	0.6 (0.4–1.0)	0.06	1.1 (0.5–2.5)	NS
MRC sum score			<0.001		<0.001		<0.001
60–51	47	1 (ref)		1 (ref)		1 (ref)	
50–41	180	2.8 (1.3–5.8)		5.9 (1.4–25)		6.1 (0.8–46)	
40–31	83	6.8 (3.0–15)		14 (3.3–64)		19 (2.4–144)	
≤ 30	83	14 (5.8–32)		23 (5.2–101)		26 (3.4–198)	
GBS disability score			<0.001		<0.001		0.002
0,1 or 2	0	0		0		0	
3	91	1 (ref)		1 (ref)		1 (ref)	
4	265	3.6 (2.1–6)		3.9 (1.9–7.9)		2.7 (1.2–5.8)	
5	38	10.5 (4.1–27)		7.3 (2.9–18)		6.1 (2.3–16)	
Sensory deficits	258	1.1 (0.7–1.7)	NS	1.0 (0.6–1.6)	NS	1.1 (0.6–1.9)	NS
Pain	187	1.0 (0.7–1.6)	NS	1.2 (0.7–1.8)	NS	0.9 (0.6–1.5)	NS
C*linical severity 7 days after admission*							
MRC sum score			<0.001		<0.001		<0.001
60–51	95	1 (ref)		1 (ref)		1 (ref)	
50–41	119	5.0 (2.5–10)		3.6 (1.2–11)		2.5 (0.7–9.5)	
40–31	76	19 (8.8–43)		11 (3.5–32)		6.3 (1.7–23)	
≤ 30	104	137 (46–405)		47 (16–139)		30 (8.8–99)	
GBS disability score			<0.001		<0.001		<0.001
0,1 or 2	33	0		0		0	
3	79	1 (ref)		1 (ref)		1 (ref)	
4	186	10.6 (5.4–21)		8.5 (3.0–24)		8.6 (2.0–37)	
5	96	36 (15–83)		21.3 (7.2–63)		25 (5.8–109)	
I*nfection and serology*							
Symptoms of preceding infection*							
Diarrhea	89	1.6 (1.0–2.6)	0.05	1.8 (1.1–3.0)	0.02	2.3 (1.3–3.9)	0.003
Upper respiratory tract infection	147	0.5 (0.4–0.8)	0.003	0.7 (0.5–1.2)	NS	0.5 (0.3–0.8)	0.006
Infection serology†							
Campylobacter jejuni	114	1.7 (1.1–2.6)	0.02	2.2 (1.4–3.4)	0.001	2.6 (1.5–4.3)	<0.001
Cytomegalovirus	45	2.2 (1.1–4.3)	0.02	2.4 (1.3–4.6)	0.006	0.9 (0.4–2.0)	NS

OR = odds ratio; CI = confidence interval; NS = non significant; MRC = Medical Research Council; GBS = Guillain-Barré syndrome.

* Symptoms of an infection in the 4 weeks preceding the onset of weakness.

† Using pre-treatment serum samples obtained at entry.

Table 8.2 Multivariable analysis of main predictors of poor outcome, defined as being unable to walk at 4 weeks after hospital admission and as no improvement on the GBS disability score in the first 4 weeks after admission.

	Unable to walk unaided at 4 weeks after hospital admission			**No improvement on GBS disability score 4 weeks after hospital admission**		
	OR (95% CI)	**P value**	**AUC**	**OR (95% CI)**	**P value**	**AUC**
At admission			0.73			0.71
Age (years)		0.006			0.001	
≤ 40	1 (ref)			1 (ref)		
40-60	1.9 (1.1–3.3)			1.9 (1.1–3.3)		
> 60	2.3 (1.3–3.8)			2.7 (1.6–4.5)		
MRC sum score		<0.001			<0.001	
60-51	1 (ref)			1 (ref)		
50-41	2.8 (1.3–6.2)			5.0 (2.0–13)		
40-31	6.1 (2.5–14)			11 (4.0–29)		
≤ 30	9.6 (3.8–24)			13 (4.7–34)		
Preceding diarrhea*	1.7 (1.0–2.9)	0.07		1.8 (1.1–3.1)	0.02	
Seven days after admission			0.87			0.87
Age (years)		0.008			0.001	
≤ 40	1 (ref)			1 (ref)		
40-60	2.1 (1.0–4.2)			2.0 (1.0–3.8)		
> 60	2.8 (1.4–5.4)			3.2 (1.7–5.9)		
MRC sum score		<0.001			<0.001	
60–51	1 (ref)			1 (ref)		
50–41	3.8 (1.7–8.4)			8.0 (2.9–22)		
40-31	10 (4.2–26)			35 (12–99)		
≤ 30	58 (18–188)			110 (38–320)		
Preceding diarrhea*	2.1 (1.0–4.4)	0.04		1.9 (1.0–3.5)	0.05	

Abbreviations: OR = odds ratio; AUC = Area Under the Receiver Operating Characteristic (ROC) Curve; MRC = Medical Research Council.
* Diarrhea in the 4 weeks preceding the onset of weakness.

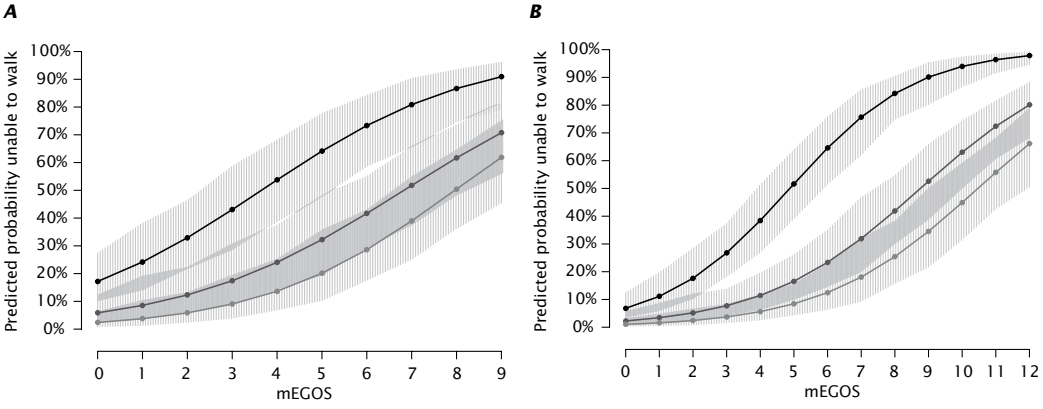
Age, diarrhea and MRC sum score were used to develop the model for clinical practice. The model can be applied already at hospital admission and at day 7 of hospital admission. When used at admission, the model scores ranged from 0-9 with 4 categories for the MRC sum score, 3 categories for age and 2 categories for preceding diarrhea (modified EGOS; Table 8.3 and Figure 8.1A) The predictive ability of the model was better when it is used at day 7 of admission, because the MRC sum score at this time point predicts outcome more accurately. Therefore, the MRC sum score was weighted stronger in the model when used at one week and the scores range from 0-12 (modified EGOS; Table 8.3 and Figure 8.1B).

Table 8.3 The modified Erasmus GBS Outcome Scores.

Prognostic factors	Categories	Score	Prognostic factors	Categories	Score
Age at onset (years)	≤ 40	0	Age at onset (years)	≤ 40	0
	41–60	1		41–60	1
	> 60	2		> 60	2
Preceding diarrhea*	Absence	0	Preceding diarrhea*	Absence	0
	Presence	1		Presence	1
MRC sum score (at hospital admission)	51–60	0	MRC sum score (at day 7 of admission)	51–60	0
	41–50	2		41–50	3
	31–40	4		31–40	6
	0–30	6		0–30	9
mEGOS		0–9	mEGOS		0–12

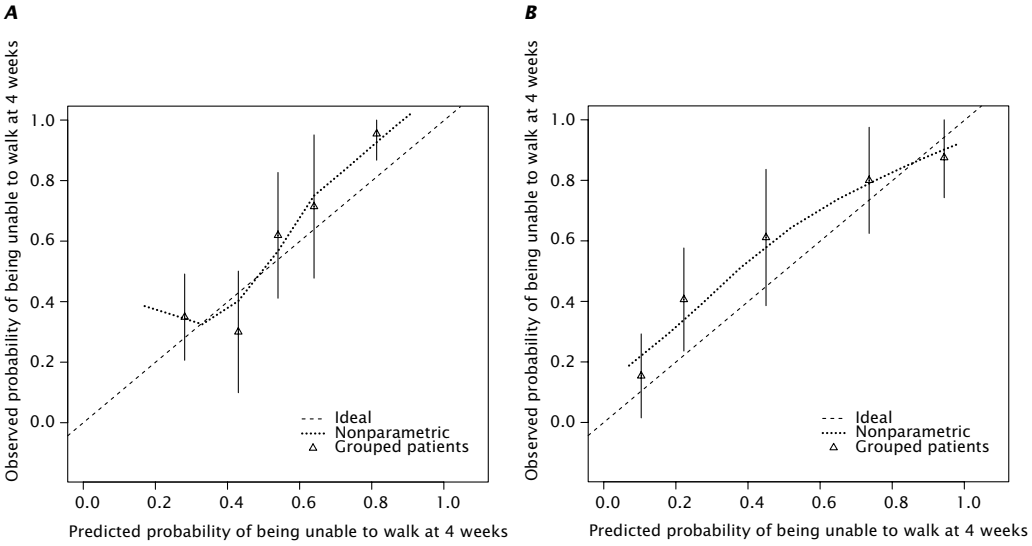
MRC = Medical Research Council; mEGOS = modified Erasmus GBS Outcome Score.
* Diarrhea in the 4 weeks preceding the onset of weakness.

Figure 8.1 Predicted fraction of patients unable to walk independently according to mEGOS at 4 weeks (black), 3 months (light grey) and 6 months (dark grey) on the basis of the mEGOS at hospital admission (A) and at day 7 of admission (B). The grey areas around the coloured lines represent 90% confidence intervals.



The performance of mEGOS when used at admission was good for prediction of outcome at 4 weeks (AUC 0.73), at 3 months (AUC 0.73) and at 6 months (AUC 0.77) and was excellent when used at day 7 of admission, with AUCs for predicting outcome at these three time points of 0.87, 0.84 and 0.84 respectively. The model was validated in an independent cohort and showed a good calibration in the independent validation cohort (figure 8.2A and 8.2B), and a good discriminative ability for predicting outcome at all three time points (admission: AUC = 0.75, 0.73 and 0.75, one week: AUC = 0.81, 0.70 and 0.77).

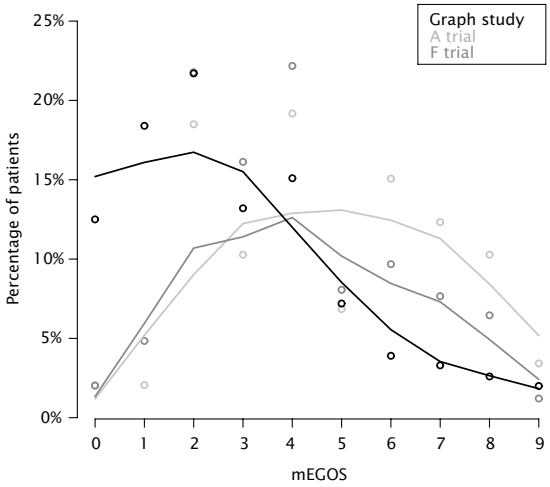
Figure 8.2 Calibration plots for external validation of mEGOS at admission (A) and at day 7 of admission (B)



Age, preceding diarrhea and MRC sum score in multivariable analysis were also independently associated with another endpoint which is frequently used in therapeutic trials in GBS: the improvement of one or more points on the GBS disability score at 4 weeks after hospital admission (Table 8.2). In addition, the mEGOS model predicted the failure to improve on the GBS disability score at 4 weeks with high accuracy (AUC of 0.71 and 0.87).

The current model can also be used to compare populations of patients included in various therapeutic trials and for covariate adjustments. To illustrate this, we compared 3 study populations[11, 14, 18] with respect to the distribution of the patients over the mEGOS categories (Figure 8.3). The figure shows that the GBS populations in the two clinical trials were comparable with respect to prognosis at hospital admission and before start of treatment. However, patients included in the observational study overall had a better prognosis than the patients in the two trial populations, most likely explained by the different inclusion criteria. Also relatively mildly-affected patients (able to walk through out the disease) were included in the observational study; those patients were excluded for validation of the models.

Figure 8.3 Comparing three therapeutic study populations with respect to prognostic factors at baseline using mEGOS, at admission. Points represent the percentages of patients with a specific mEGOS in a therapeutic trial comparing PE versus IVIg (light grey), a therapeutic trial comparing IVIg/placebo versus IVIg/methylprednisolone (dark grey) and an observational study (black). Smoothed lines represent the distribution of the study population over the total mEGOS.



Discussion

The diversity in clinical severity and outcome in GBS patients hampers optimizing of treatment, because RCT populations will always have a large variability in baseline risk for outcome. To avoid large problems with statistical power, we should deal with this diversity properly. Slow inclusion rates are inherent in this rare disease, which is an additional challenge in conducting RCTs in GBS. New therapies and treatment modalities for GBS may not further improve outcome in patients who already recover sufficiently after standard treatment. Therefore selective treatment trials should focus on a more homogeneous subgroup of patients with poor recovery despite current standard treatment. In this study a prognostic model is presented which early identifies patients with poor outcome and can be used for future therapeutic trials. The main predictors of being unable to walk independently at 4 weeks, 3 months and 6 months were MRC sum score, age and preceding diarrhea in our study. Based on these predictors a model was constructed which proved to be valid in an independent cohort of GBS patients. The model is applicable at hospital admission as well as at day 7 of hospital admission and is therefore suitable to study treatments which should be started immediately as well as after standard treatment in patients with poor prognosis. The model may provide a first step towards individualized treatment in GBS.

This prognostic model originates from the EGOS, which can be used in clinical prac-

tice at two weeks after hospital admission to predict outcome at 6 months and is based on the predictors age, preceding diarrhea and GBS disability score[7]. The EGOS is a simple, accurate and validated prognostic model, but less suited for treatment development because of the delay of two weeks and the predicted outcome measure. The new prognostic model (mEGOS) was primarily designed for trials in GBS and for this purpose has important advantages. First, the mEGOS model can be applied already in the first week of admission, when treatment is considered to be most effective. Second, the mEGOS predicts reaching independent walking or improving on the GBS disability score at 4 weeks, which are the two primary endpoints most frequently used in therapeutic trials in GBS. Thirdly, the mEGOS also accurately predicts long-term GBS disability scores, which were important secondary endpoints in previous trials. Because of these features the mEGOS model can be used to early identify patients with poor prognosis for future selective therapeutic studies. In addition, this model can be used for covariate adjustment, which is a powerful tool in heterogeneous patient populations to estimate the effect of treatment in individuals and to increases the statistical power of therapeutic trials.[2, 22-23] For example, adjustment for the effect of age on outcome results in an estimated treatment effect for a patient of a given age instead of an average age. If these selective trial results in patients with poor prognosis are positive, the mEGOS may also be used to individualize treatment of GBS patients in routine clinical practice.

Our study confirms that poor outcome is associated with older age,[4-5, 7-8] rapid disease progression,[8] severe disease indicated by GBS disability score or MRC sum score, [3-4,7] preceding diarrhea, positive *C. jejuni* serology,[3, 5, 7] positive CMV serology,[9] and no symptoms of a preceding respiratory tract infection.[3-4] Two of these studies used partly the same data as in this study.[7, 9] For the purpose of this study, we selected age, preceding diarrhea, and MRC sum score which are readily available at hospital admission of the patient. Prognostic biomarkers may further improve those models in the future. Promising candidates are infection serology, anti-ganglioside antibodies and serum IgG-level increase after IVIg treatment, which were all related to outcome.[3,5-7,9] The need of accurate prediction models for outcome has also been acknowledged for traumatic brain injury patients[24] and for stroke patients.[25-26] Those neurological conditions resemble GBS in the sense that they are acute and monophasic and have a highly variable clinical course.

Our study had several limitations. First, the prognostic model was derived from cohorts of Dutch Caucasians, which may restrict the application to those patients. Second, information on outcome at 4 weeks was not available in 24% of patients from the validation cohort. For this cohort data were used from an observational study, in which 4 weeks was not a standardized evaluation time point. However, percentages of patients with a poor outcome at 4 weeks in the derivation and validation cohort were comparable (55% and 54%), so it is unlikely that this caused bias. A third limitation is

that the model only predicts the ability to walk independently, and not the full ordinal GBS disability scores, as this would have provided maximum statistical power.[27] However, this specific outcome measure we used is highly relevant for patients and was previously used by most therapeutic trials in GBS.

In conclusion, mEGOS is an accurate and validated model for prediction of outcome at several time points in the first 6 months after onset of GBS. An important advantage above existing models is that the mEGOS can be used in the early phase of disease when the process of nerve damage is ongoing and possibly reversible. This model predicts commonly-used trial endpoints in GBS and can be used to conduct new trials selectively in patients with poor outcome. In addition the model can be used to compare patient populations with respect to prognostic factors and expected outcome. This model may hold great promise to assist clinicians in optimizing individual treatment for GBS patients.

Acknowlegdements

Mrs. Walgaard is funded by a scientific research grant from the Dutch Prinses Beatrix Fonds (PBF WAR07-28). Mrs. Lingsma is funded by NIH grant #NS-42691. Dr. Steyerberg is funded by NIH grant #NS-42691 and the Netherlands Organization for Scientific Research (912.08.004) Dr. Jacobs receives research support from the Netherlands Organization for Health Research and Development, Erasmus MC, the Dutch Prinses Beatrix Fonds, GBS-CIDP Foundation International and Baxter Biopharmaceutics.

References

1. van Doorn PA, Ruts L, Jacobs BC. Clinical features, pathogenesis, and treatment of Guillain-Barre syndrome. *Lancet Neurol* 2008;7:939-950.
2. Roozenbeek B, Maas AI, Lingsma HF, et al. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med* 2009;37:2683-2690.
3. The prognosis and main prognostic indicators of Guillain-Barre syndrome. A multicentre prospective study of 297 patients. The Italian Guillain-Barre Study Group. *Brain* 1996;119 (Pt 6):2053-2061.
4. Chio A, Cocito D, Leone M, Giordana MT, Mora G, Mutani R. Guillain-Barre syndrome: a prospective, population-based incidence and outcome survey. *Neurology* 2003;60:1146-1150.
5. Hadden RD, Karch H, Hartung HP, et al. Preceding infections, immune factors, and outcome in Guillain-Barre syndrome. *Neurology* 2001;56:758-765.
6. Kuitwaard K, de Gelder J, Tio-Gillen AP, et al. Pharmacokinetics of intravenous immunoglobulin and outcome in Guillain-Barre syndrome. A*nn Neurol* 2009;66:597-603.
7. van Koningsveld R, Steyerberg EW, Hughes RA, Swan AV, van Doorn PA, Jacobs BC. A clinical prognostic scoring system for Guillain-Barre syndrome. *Lancet Neurol* 2007;6:589-594.
8. Winer JB, Hughes RA, Osmond C. A prospective study of acute idiopathic neuropathy. I. Clinical features and their prognostic value. *J Neurol Neurosurg Psychiatry* 1988;51:605-612.
9. Visser LH, van der Meche FG, Meulstee J, et al. Cytomegalovirus infection and Guillain-Barre syndrome: the clinical, electrophysiologic, and prognostic features. Dutch Guillain-Barre Study Group. *Neurology* 1996;47:668-673.
10. Efficiency of plasma exchange in Guillain-Barre syndrome: role of replacement fluids. French Cooperative Group on Plasma Exchange in Guillain-Barre syndrome. *Ann Neurol* 1987;22:753-761.
11. van der Meche FG, Schmitz PI. A randomized trial comparing intravenous immune globulin and plasma exchange in Guillain-Barre syndrome. Dutch Guillain-Barre Study Group. *N Engl J Med* 1992;326:1123-1129.
12. Double-blind trial of intravenous methylprednisolone in Guillain-Barre syndrome. Guillain-Barre Syndrome Steroid Trial Group. *Lancet* 1993;341:586-590.
13. Randomised trial of plasma exchange, intravenous immunoglobulin, and combined treatments in Guillain-Barre syndrome. Plasma Exchange/Sandoglobulin Guillain-Barre Syndrome Trial Group. *Lancet* 1997;349:225-230.
14. van Koningsveld R, Schmitz PI, Meche FG, Visser LH, Meulstee J, van Doorn PA. Effect of methylprednisolone when added to standard treatment with intravenous immunoglobulin for Guillain-Barre syndrome: randomised trial. *Lancet* 2004;363:192-196.
15. Treatment of Guillain-Barre syndrome with high-dose immune globulins combined with methylprednisolone: a pilot study. The Dutch Guillain-Barre Study Group. *Ann Neurol* 1994;35:749-752.
16. Asbury AK, Cornblath DR. Assessment of current diagnostic criteria for Guillain-Barre syndrome. *Ann Neurol* 1990;27 Suppl:S21-24.

17. Garssen MP, van Koningsveld R, van Doorn PA, et al. Treatment of Guillain-Barre syndrome with mycophenolate mofetil: a pilot study. *J Neurol Neurosurg Psychiatry* 2007;78:1012-1013.
18. Ruts L, Drenthen J, Jacobs BC, van Doorn PA, Dutch GBSSG. Distinguishing acute-onset CIDP from fluctuating Guillain-Barre syndrome: a prospective study. *Neurology* 2010;74:1680-1686.
19. Kleyweg RP, van der Meche FG, Schmitz PI. Interobserver agreement in the assessment of muscle strength and functional abilities in Guillain-Barre syndrome. *Muscle Nerve* 1991;14:1103-1109.
20. Hughes RA, Newsom-Davis JM, Perkin GD, Pierce JM. Controlled trial prednisolone in acute polyneuropathy. *Lancet* 1978;2:750-753.
21. Steyerberg EW. Clinical Prediction Models, 1st ed: Springer-Verlag New York Inc., 2008.
22. Hernandez AV, Steyerberg EW, Habbema JD. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. *J Clin Epidemiol* 2004;57:454-460.
23. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? *Am Heart J* 2000;139:745-751.
24. Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AI. Early prognosis in traumatic brain injury: from prophecies to predictions. *Lancet Neurol* 2010;9:543-554.
25. Reid JM, Gubitz GJ, Dai D, et al. Predicting functional outcome after stroke by modelling baseline clinical and CT variables. *Age Ageing* 2010;39:360-366.
26. Uchino K, Billheimer D, Cramer SC. Entry criteria and baseline characteristics predict outcome in acute stroke trials. *Stroke* 2001;32:909-916.
27. Maas AI, Steyerberg EW, Marmarou A, et al. IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics* 2010;7:127-134.

9 Prediction of two month mortality after aneurysmal subarachnoid haemorrhage

Risselada R, Lingsma HF, Bauer-Mehren A, Friedrich CM, Molyneux AJ, Kerr RSC, Yarnold J, Sneade M, Steyerberg EW, Sturkenboom MCJM. Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT). *European Journal of Epidemiology* 2010; 25:261–266.

Abstract

Introduction

Background and Purpose

Aneurysmal subarachnoid haemorrhage (aSAH) is a devastating event with substantial case-fatality. Our purpose was to examine which clinical and neuro-imaging characteristics, available on admission, predict 60-day case-fatality in aSAH and to evaluate performance of our prediction model.

Methods

We performed a secondary analysis of patients enrolled in the International Subarachnoid Aneurysm Trial (ISAT), a randomised multicenter trial to compare coiling with clipping in aSAH patients. Multivariable logistic regression analysis was used to develop a prognostic model to estimate the risk of dying within 60 days from aSAH based on clinical and neuro-imaging characteristics. The model was internally validated with bootstrapping techniques.

Results

The study population comprised of 2,128 patients who had been randomised to either endovascular coiling or neurosurgical clipping. In this population 153 patients (7.2%) died within 60 days. World Federation of Neurosurgical Societies (WFNS) grade was the most important predictor of case-fatality, followed by age, lumen size of the aneurysm and Fisher grade. The model discriminated reasonably between those who died within 60 days and those who survived (c statistic = 0.73), with minor optimism according to bootstrap re-sampling (optimism corrected c statistic = 0.70).

Conclusion

Several strong predictors are available to predict 60 day case-fatality in aSAH patients who survived the early stage up till a treatment decision; after external validation these predictors could eventually be used in clinical decision making.

Introduction

Subarachnoid haemorrhage (SAH) is a devastating event, which is marked by sudden onset of severe headache, causing substantial case-fatality. In 85% of the patients, the SAH is caused by rupture of an aneurysm (aSAH).[1,2] From those who survive the first month, approximately one third remains dependent with respect to daily activities during their remaining lifetime.[1] Amongst patients who regain independency, quality of life remains reduced.[3]

Early prediction of short term outcome in terms of case-fatality may support clinical decision making and may provide realistic and evidence based expectations to patients and relatives. Predictions may also be used to classify patients according to prognostic risk, which may be useful to compare outcome between different patient series, to study treatment results over time, or to stratify patients for randomised clinical trials (RCT).

Several other prognostic studies on outcome after aSAH have been performed, but most included relatively small numbers. Two included substantial numbers allowing analysis of the effects of multiple independent factors.[4,5] However, these studies did not examine case-fatality, but arbitrarily dichotomized ordinal outcome scales (modified Rankin Scale or Glasgow Outcome Scale). Moreover, due to their design, these studies were unable to examine the effect of treatment on prediction of outcome.

Our aim was to develop a prognostic model for 60-day case-fatality, based on clinical features and neuro-imaging, regularly readily available on admission to a neurological or neurosurgical unit. These data were obtained from a large RCT conducted in mainly European countries.

Methods

Patients

Data were collected prospectively by the Medical Research Council funded International Subarachnoid Aneurysm Trial (ISAT) (International Standard Randomised Controlled Trial, number ISRCTN49866681). Full details of the ISAT study are available elsewhere.[6] The aim of the trial was to determine whether treatment using endovascular coiling reduced the risk of patients being dependent or dead at one year by 25 percent when compared with neurosurgical treatment (clipping).

Predictors and outcome

We considered all patient characteristics that can be collected easily and reliably within the first hours after hospital admission and that were also present in the ISAT database. These included age, gender, previous occurrence of SAH, CT scan Fisher grading, lumbar puncture, World Federation of Neurosurgical Societies (WFNS) grading, number of intracranial aneurysms, location of the aneurysm, maximum lumen size of the aneurysm, vasospasm on angiography, and intended treatment by randomization. Fisher grading of blood visible on a plain CT scan runs from grade 1 ('no blood visible') up to grade 4 ('intraventricular or intraparenchymal blood'). Lumbar puncture was not performed in all participants. If it was performed it was graded 1 ('no blood in cerebrospinal fluid') or 2 ('xanthochromia or blood'); 0 otherwise ('no lumbar puncture'). WFNS scale runs from grade 1 ('Glasgow Coma Scale (GCS) 15 and no motor deficit') to grade 5 ('GCS 3-6 with or without motor deficit'). One category additional to the standard WFNS scale was created in ISAT for those in whom WFNS could not be assessed; 'grade 6'. The number of aneurysms was categorized in 1, 2, and 3 or more intracranial aneurysms. We discerned four aneurysm locations: Anterior Cerebral Artery (ACA), Internal Carotid Artery (ICA), Middle Cerebral Artery (MCA), and Posterior Circulation (PC). The maximum lumen size of the aneurysm was expressed in millimetres. Vasospasm was examined on angiography and categorized as 'none', 'mild', 'moderate', or 'severe'. Treatment was either neurosurgical clipping or endovascular coiling; we used treatment as allocated by the randomization procedure. We developed the model based on cases with a complete set of data. The outcome was 60-day case-fatality.

Model

We used univariate logistic regression analysis to estimate the association between single predictors and outcome, expressed as an odds ratio (OR). Predictors have a statistically significant effect if the 95 percent confidence interval (95% CI) does not include the value one. The prediction model was developed with multivariable logistic regression with backward stepwise selection. All potential predictors were entered into the model and those that met Akaike's Information Criterion (AIC) were selected into the model. AIC compares models based on how well they fit the data, but penalizes for the complexity of the model. AIC requires that the increase in model χ^2 when entering a new predictor has to be larger than two times the degrees of freedom: $\chi^2 > 2\ df$. When considering a predictor with 1 *df*, such as gender, this implies that χ^2 has to exceed 2, equivalent to $p<0.157$. When considering a predictor with 2 *df*, $\chi^2 > 4$, or $p<0.135$; and in case of 4 *df*, $\chi^2 > 8$, or $p<0.092$.[7]

Performance

The performance of the model was assessed with respect to calibration and discrimination. Calibration is the ability of the model to produce unbiased estimates of the

probability of the outcome. Calibration was examined with a goodness of fit test, which assesses agreement between predicted and observed risks over the full range of predicted probabilities.[8]

Discrimination is the model's ability to separate patients with different outcomes. To quantify the discrimination, we used the concordance (*c*) statistic. For binary outcomes, *c* is identical to the area under the receiver operating characteristic curve.[7] The *c* statistic evaluates whether those with higher predicted risk are more likely to die within 60 days among all possible pairs of patients with different outcomes. A model with a *c* statistic of 0.5 has no discriminative power at all, for example a coin flip. A *c* statistic of 1.0 reflects perfect discrimination.

Model validation

The performance of a prediction model is generally worse in new patients than initially expected. This 'optimism' can be studied with internal validation techniques.[7] Internal validity of our model was assessed with standard bootstrapping procedures.[7] Bootstrapping involves drawing samples of patients with replacement from the study population. Each sample can be considered as if one is repeating the data collection with the same number of patients and under identical circumstances as the original. The multivariable logistic regression coefficients were re-estimated in 300 bootstrap samples. Each of these 300 models was evaluated on the original sample. The average difference in the *c* statistic was determined to indicate the optimism in the initially estimated discriminative ability.[7] A shrinkage factor was estimated from the bootstrap validation procedure and we shrunk the regression coefficients to provide better predictions for future patients.[7]

All statistical analyses were performed using R software, version 2.8.1 (R Foundation for Statistical Computing, Vienna, Austria).

Results

A total of 2,143 patients were recruited into the ISAT trial by 43 neurosurgical centers, mainly in Europe. CT scans of 14 patients were not performed or available, and in one patient no information on vasospasm was available. We excluded these cases from our analysis. Data on the outcome were available for all patients. Thus, we performed complete case analysis on 2,128 patients (99.3%) of whom 153 (7.2%) died within 60 days.

The distribution of patient characteristics of the study population is presented in table 9.1. For reasons of small numbers in the 'severe' category of vasospasm, we aggregated data from the 'moderate' and 'severe' categories into one category. Univariate analysis showed a statistically significant relation with 60-day case-fatality for age, lumen size, Fisher grade, lumbar puncture, WFNS grade, and vasospasm. Sex, location and number of aneurysms and intended treatment were not significantly associ-

ated with 60-day case-fatality (Table 9.1). In the multivariable model with stepwise backward selection age, lumen size, Fisher grade, and WFNS grade met AIC and were included in the final model. In table 9.2 the chi square statistics with corresponding p-values are presented as well as the point estimate of the OR. Age and WFNS grade were the most important predictors.

The goodness of fit test yielded a p-value of 0.86, suggesting that the model fitted the data in which it was developed well. The c statistic of the original model was 0.73, meaning that the model discriminates reasonably between patients who die within 60 days from onset of the SAH and those who survive this period.
Validation by means of 300 bootstrap samples resulted in a shrinkage factor of 0.85, which was applied to the betas of the model. The c statistic of the internally validated model was 0.70. Details of the final prognostic model for 60-day case-fatality are described in the appendix.

Table 9.1 Population characteristics and univariate association with 60-day case fatality

Predictor	Alive n=1975		Death n=153						
	median	IQR	median	IQR	beta	SE	OR	CI_{min}	CI_{max}
Age [10years]	5.2	4.3–5.9	5.6	5.0–6.5	0.36	0.08	1.43	1.23	1.66
Lumensize [mm]	5	4–7	6	4–8	0.10	0.02	1.10	1.05	1.15
	n	**%**	**n**	**%**					
Sex									
female	1236	63	103	67			1		
male	739	37	50	33	−0.21	0.18	0.81	0.57	1.15
Previous SAH									
yes	129	7	6	4	−0.54	0.43	0.58	0.25	1.35
no	1846	93	147	96			1		
Fisher's grade									
1	112	6	2	1			1		
2	350	18	10	7	0.47	0.78	1.60	0.35	7.41
3	840	43	62	41	1.42	0.73	4.13	1.00	17.13
4	673	34	79	52	1.88	0.72	6.57	1.59	27.13
Lumbar puncture									
xantochromia or blood	217	11	7	5	−0.95	0.39	0.39	0.18	0.84
no blood	5	0	0	0	−5.19	20.8	0.01	0.00	∞
no puncture	1753	89	146	95			1		
WFNS grade									
1	1270	64	54	35			1		
2	495	25	51	33	0.89	0.20	2.42	1.63	3.60
3	120	6	13	8	0.94	0.32	2.55	1.35	4.80
4	55	3	19	12	2.09	0.30	8.12	4.51	14.63
5	13	1	7	5	2.54	0.49	12.66	4.86	33.02
6 (not assessable)	22	1	9	6	2.26	0.42	9.62	4.23	21.89
n of aneurysms detected									
1	1555	79	116	76			1		
2	314	16	29	19	0.21	0.22	1.24	0.81	1.89
>=3	106	5	8	5	0.10	0.38	1.10	0.52	2.32
Location									
ACA	1008	51	71	46	−0.16	0.19	0.85	0.59	1.23
ICA	638	32	53	35			1		
MCA	277	14	23	15	0.00	0.26	1.00	0.60	1.66
PC	52	3	6	4	0.33	0.45	1.39	0.57	3.38
Vasospasm									
none	1575	80	109	71			1		
mild	218	11	24	16	0.46	0.24	1.59	1.00	2.53
moderate/severe	182	9	20	13	0.77	0.26	2.15	1.30	3.55
Intended treatment									
clip	983	50	83	54			1		
coil	992	50	70	46	−0.18	0.17	0.84	0.60	1.16

IQR = inter quartile range; beta = regression coefficient in the logistic regression model; SE = standard error; OR = odds ratio; CI_{min} = lower limit of the 95% confidence interval; CI_{max} = upper limit of the 95% confidence interval.

Table 9.2 Statistical parameters of the final model

Factor	X^2	*df*	p-value	OR	95% CI
WFNS grade	51	5	<0.001	grade 1 = ref.	
2				1.87	1.23–2.83
3				1.70	0.87–3.32
4				4.87	2.60–9.14
5				7.00	2.54–19.28
6				5.75	2.41–13.73
Age [10 yrs]	17	1	<0.001	1.32	1.13–1.55
Lumen size [mm]	12	1	<0.001	1.08	1.03–1.13
Fisher grade	8	3	0.04	grade 1 = ref.	
2				1.43	0.27–7.65
3				2.67	0.53–13.51
4				2.76	0.54–14.14

X^2 is the chi square test statistic for the predictor in the final model; *df* = degrees of freedom; 95% CI was calculated based on the S.E. of the estimates of the coefficients in the full model to avoid underestimation of uncertainty.

Discussion

We developed a prognostic model to predict the risk of 60 day case-fatality in individual patients after aSAH. Predictions were based on characteristics that are regularly readily available on admission to a neurological or neurosurgical unit and which were collected in a large clinical trial. The full model yielded a c statistic of 0.73.

Previously, several models to estimate the probability of unfavourable outcome after aSAH have been developed. Our model was similar to those; we included roughly the same predictors: age, clinical status, and lumen size.[4,5,9] However, our study is of added value because of the substantial size and the inclusion of both clipped and coiled patients. The studies by Hoh et al.[4] (n=515) and Mocco et al.[9] (n=148) contained relatively few patients. The small numbers of coiled patients (79 and 35, respectively) and the design of the study did not allow for taking the effect of treatment in consideration. The study of Rosengart et al.[5] (n=3667) was not able to do that either, since patients treated with Guglielmi or other detachable coils were excluded. All three studies used a dichotomized ordinal scale as an outcome, for which the cut off can be (arbitrarily) chosen in different studies. In a sense, examining case-fatality is also a dichotomization of an ordinal scale, though less arbitrary. Therefore, we are convinced that logistic regression is well suited for an outcome that is by its nature dichotomous, whereas for an ordinal outcome we would prefer specific modelling techniques.

Several limitations of this study should be acknowledged. This study used data from one large trial on a selected population of patients who survived the early stage up till a treatment decision and who were in equipoise regarding that decision on treatment with either endovascular coiling or neurosurgical clipping, which may limit external validity. The model may perform well in this development sample, but worse when applied to other groups of patients, for example, a less strictly selected population. Nonetheless, according to a recently published paper, the ISAT population proved to be quite similar to the population admitted with an aSAH to neurosurgical units in the United Kingdom.[10] Although in ISAT, a lower proportion of poor grade patients were enrolled. Validation of a prognostic model in independent patient series is therefore considered an essential next step.[11] However, since large samples of systematically collected data on aSAH are sparse, assessment of external validity is difficult. For now the external validity of our model remains to be established. This will be a topic of future research.

Although our model represents knowledge obtained from 2,128 SAH patients in equipoise regarding treatment, statistical models can never replace the clinician with regard to decision making; they can only assist with this task. A prediction for an individual aSAH patient a particular situation is always subject to uncertainty.

The model makes certain structural assumptions and statistical interaction terms were not included. It is hence possible that specific patterns of risk factors are inadequately reflected in the model predictions. Therefore, predictions should be regarded with care and not directly be applied for treatment limiting decisions.

Although the performance of the presented model was satisfactory, it might potentially be improved by including neuro-imaging biomarkers other than lumen size, location, Fisher grade on plain CT scan, and vasospasm on angiography. Biomarkers regarding anatomy and morphology might be considered, as well as aneurysm characteristics obtained from three and four dimensional angiography.[12,13] Performance may also be improved by inclusion of subsequent information obtained after admission, including temporal course, neuro-imaging at later time points, eventual rebleeding of the aneurysm, delayed ischemic deficit, and other parameters such as hydrocephalus. The objective of the present study, however, was to investigate prognostic models that predict 60-day case-fatality with predictors available on admission.

Statistical testing for calibration has a number of drawbacks. First, the null hypothesis is of good calibration. Hence, if we test calibration in a small study, we have low power and will not reject the null hypothesis unless miscalibration is very severe. On the other hand, even a model with very good, though not perfect, calibration will fail the test in case of a sufficiently large sample. Moreover, reported goodness-of-fit tests are usually non-significant if they reflect apparent validation on the data that were also used to construct the model. Such non-significant results may contribute to the face validity of a model, but have no real scientific meaning.[7]

In conclusion, we presented a prognostic model for predicting 60-day case-fatality after aneurysmal SAH. Our model contained age, lumen size, Fisher grade, and WFNS grade as predictors. After calibration and internal validation, our model showed reasonable performance, although external validity of our model remains to be established.

Acknowledgments

This study was performed within the scope of the @neurIST-project ('@neurIST, Integrated biomedical informatics for the management of cerebral aneurysms', www.aneurist.org), funded by the European Commission 6th Framework Programme.

The International Subarachnoid Aneurysm Trial was supported by grants from: the Medical Research Council, UK; and Programme Hospitalier de Recherche Clinique 1998 of the French Ministry of Health (AOM 98150). It was sponsored by Assistance Publique, Hôpitaux de Paris (AP-HP); the Canadian Institutes of Health Research; and the Stroke Association of the UK for the Neuropsychological assessments.

References

1. Hop JW, Rinkel GJ, Algra A, van Gijn J. Case-fatality rates and functional outcome after subarachnoid hemorrhage: A systematic review. *Stroke.* 1997;28:660-664
2. Van Gijn J, Kerr RS, Rinkel GJ. Subarachnoid haemorrhage. *Lancet.* 2007;369:306-318
3. Hop JW, Rinkel GJ, Algra A, van Gijn J. Changes in functional outcome and quality of life in patients and caregivers after aneurysmal subarachnoid hemorrhage. *J Neurosurg.* 2001;95:957-963
4. Hoh BL, Topcuoglu MA, Singhal AB, Pryor JC, Rabinov JD, Rordorf GA, Carter BS, Ogilvy CS. Effect of clipping, craniotomy, or intravascular coiling on cerebral vasospasm and patient outcome after aneurysmal subarachnoid hemorrhage. *Neurosurgery.* 2004;55:779-786; discussion 786-779
5. Rosengart AJ, Schultheiss KE, Tolentino J, Macdonald RL. Prognostic factors for outcome in patients with aneurysmal subarachnoid hemorrhage. *Stroke.* 2007;38:2315-2321
6. Molyneux A, Kerr R, Stratton I, Sandercock P, Clarke M, Shrimpton J, Holman R. International Subarachnoid Aneurysm Trial (ISAT) of neurosurgical clipping versus endovascular coiling in 2143 patients with ruptured intracranial aneurysms: A randomised trial. *Lancet.* 2002;360:1267-1274
7. Steyerberg EW. *Clinical prediction models. A practical approach to development, validation, and updating.* New York: Springer. 2009
8. Harrell FE Jr. Resampling model calibration; the calibrate function in library(Design) in R: *A language and environment for statistical computing.* Vienna (Austria): R Foundation for Statistical Computing; 2006. Available: http://www.R-project.org/. Accessed 12 January 2010
9. Mocco J, Ransom ER, Komotar RJ, Schmidt JM, Sciacca RR, Mayer SA, Connolly ES, Jr. Preoperative prediction of long-term outcome in poor-grade aneurysmal subarachnoid hemorrhage. *Neurosurgery.* 2006;59:529-538; discussion 529-538
10. Langham J, Reeves BC, Lindsay KW, van der Meulen JH, Kirkpatrick PJ, Gholkar AR, Molyneux AJ, Shaw DM, Copley L, Browne JP. Variation in outcome after subarachnoid hemorrhage: A study of neurosurgical units in uk and ireland. *Stroke.* 2009;40:111-118
11. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. *Ann Intern Med.* 1999;130:515-524
12. Hayakawa M, Katada K, Anno H, Imizu S, Hayashi J, Irie K, Negoro M, Kato Y, Kanno T, Sano H. Ct angiography with electrocardiographically gated reconstruction for visualizing pulsation of intracranial aneurysms: Identification of aneurysmal protuberance presumably associated with wall thinning. *AJNR Am J Neuroradiol.* 2005;26:1366-1369
13. Zhang C, Villa-Uriol MC, De Craene M, Pozo J, Frangi A. Morphodynamic analysis of cerebral aneurysm pulsation from time-resolved rotational angiography. *IEEE Trans Med Imaging.* 2009;28:1105-1116

Appendix

Details of the Prognostic Model

The probability of dying within 60 days is calculated according to the logistic formula: $1/(1 + \exp^{-LP})$. The linear predictor (LP) takes the form of LP = intercept + regression coefficients × predictor values.

LP for 60-day case-fatality = −5.812 + 0.2762 × age + 0.3572 × [Fisher grade II] + 0.9756 × [Fisher grade III] + 1.008 × [Fisher grade IV] + 0.6216 × [WFNS grade 2] + 0.5261 × [WFNS grade 3] + 1.574 × [WFNS grade 4] + 1.934 × [WFNS grade 5] + 1.738 × [WFNS grade not assessable] + 0.07662 × lumen size of aneurysm.

Coding of the predictors was as follows: age in decades, lumen size in millimetres; all other predictors, 1 if true and 0 if false.

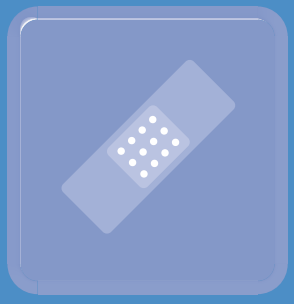


IV Applications









10 Between-center differences in outcome after traumatic brain injury

Lingsma HF, Roozenbeek B, Bayoue L, Lu J, Weir J, Butcher I, Marmarou A †, Murray GD, Maas AIR, Steyerberg EW. Large between-center differences in outcome after moderate and severe traumatic brain injury in the IMPACT study. *Neurosurgery.* In press.

Abstract

Introduction
Differences between centers in patient outcome after traumatic brain injury (TBI) are of importance for multicenter studies, and have seldom been studied. We aimed to quantify these differences in centers enrolling patients in randomized clinical trials (RCTs) and surveys.

Methods
We analyzed individual patient data from 9578 patients with moderate and severe TBI enrolled in ten RCTs and three observational studies. We used random effects logistic regression models to estimate the between-center differences in unfavourable outcome (dead, vegetative state or severe disability measured with the Glasgow Outcome Scale) at six months, adjusted for differences in patient characteristics. We calculated the difference in odds of unfavourable outcome between the centers at the higher end versus those at the lower end of the outcome distribution. We analyzed the total database, Europe and the US separately, and four of the larger RCTs.

Results
The 9578 patients were enrolled at 265 centers, and 4629 (48%) had an unfavourable outcome. After adjustment for patient characteristics, there was a 3.3 fold difference in the odds of unfavourable outcome between the centers at the lower end of the outcome distribution (2.5^{th} percentile) versus those at the higher end of the outcome distribution (97.5^{th} percentile) ($p<0.001$). In the four larger RCTs, the differences between centers were similar. However, differences were smaller between centers in the US (2.4 fold) than between centers in Europe (3.8 fold).

Conclusion
Outcome after TBI differs substantially between centers, particularly in Europe. Further research is needed to study explanations for these differences to suggest where quality of care might be improved.

Introduction

Interest in differences in quality of care between health care providers is increasing throughout the medical world.[1-3] Such differences may be studied by comparing patient outcomes between centers, for example the number of patients dying within 30 days after a myocardial infarction or surgical procedure. Observed between-center differences in outcome can be caused by various reasons, such as a different patient population (e.g. a different age distribution of the patients) or simply by random variation. Remaining differences may be explained by bias; eg registration bias or residual confounding but also by structural differences (e.g. referral system in a country) and by differences in process (e.g. adherence to guidelines). Both aspects relate to quality of care. In randomized clinical trials (RCTs) large between-center differences may adversely affect the chances of detecting a treatment effect. More insight in the magnitude of the between-center differences might hence help to improve quality of care and to improve the design of RCTs.

Considerable between-center differences in outcome have been reported in various diseases and disciplines.[4-6] Few studies have reported on such differences in traumatic brain injury (TBI), and these were usually based on single studies. For example, considerable between-center differences in six month unfavourable outcome were found in the NABIS hypothermia trial.[7] Further thorough study however is necessary in moderate and severe TBI, which is a major problem worldwide and leads to high mortality and permanent disability in predominantly young patients, causing high costs to society.[8]

We aim to quantify differences in outcome between centers enrolling patients in randomized clinical trials (RCTs) and surveys.

Methods

Patients and data collection

We used the IMPACT database (International Mission on Prognosis and Clinical Trial design in TBI), which currently contains data on over 11,989 individual patients with moderate and severe TBI, both from randomized controlled trials (RCTs) and observational studies.

We excluded patients with missing outcome (n=601), missing age (n=6), younger than 14 (n=359) and with missing center (n=172). We excluded patients from one single-center study (n= 756) since it could not contribute information to estimate between-center differences. Ultimately 9,578 patients were analyzed, and from one study (n=517) for which we did not know the treating center.

Details of the development of the IMPACT database and the constituent studies have been previously reported.[9] Also a more extensive description of the data used in this study can be found in table 10.1.

Outcome and measures

The primary endpoint was an unfavourable outcome after six months, measured with the Glasgow Outcome Scale (GOS), which was dichotomized as favourable (good recovery or moderate disability) versus unfavourable outcome (severe disability, vegetative state or death).[10] We adjusted analyses for the main predictors of outcome in TBI: age, pupillary reactivity and Glasgow Coma Scale motor score.[11, 12, 13, 14] GCS motor untestable was included in the model as a separate category to deal with patients being sedated at admission. Missing values for pupillary reactivity were imputed with a multiple imputation procedure in 13.9% of the patients. The imputation was based on 15 relevant covariates and the outcome, as described before.[14, 15]

We developed a common center code over all studies in the database, so when a center participated in multiple studies, it had one unique code.

Statistical analyses

In the quantification of between-center differences we need to account for random variation caused by low numbers and for differences in patient characteristics. This was achieved using a random effects logistic regression model, which estimates 'fixed' coefficients β for covariates at the patient level i in center j (X_{ij}) and 'random' coefficients for the centers j(θ_j). The parameter θ_j is assumed to be normally distributed with mean μ and variance τ^2:

$$\text{Logit}\,(P(Y_{ij}=1|X_{ij})) = \beta X_{ij} + \theta_j \text{ with } \theta_j \sim N(\mu, \tau^2).$$

The variance (τ^2) estimated in the random effects model is a measure of the between-center differences, and indicates the spread of the estimated proportions of unfavourable outcome of the individual centers. In a random effects model, outcomes for small centers are drawn towards the mean to avoid too extreme estimates. Therefore τ^2 can be interpreted as the unexplained between-center differences, beyond what would be expected based on random variation.[16, 17]

We aimed to illustrate the concept of between-center differences and random variation graphically, by calculating the distribution of unfavourable outcome that would be expected based on just random variation. We therefore simulated outcome per center without any between-center differences, i.e. a constant risk of unfavourable outcome of 48%, which was the average percentage unfavourable outcome over all centers. Histograms were created 1000 times and averaged to obtain a stable estimate of what might be expected given random variation only.

To facilitate interpretation of the estimated between-center differences, we compared the centers at the higher end of the outcome distribution (the 97.5th percentile) with the centers at the lower end (the 2.5th percentile) of the outcome distribution. The relative difference in odds on unfavourable outcome in these two groups of centers can

be calculated from the parameter τ^2: 95% OR range = exp (3.92*τ). The value 3.92 is the Z value corresponding to the width of the 95% confidence interval in a normal distribution (2*1.96). For example, a τ^2 of 0.09 means that centers at the higher end of the outcome distribution (the 'worst' centers) have a 3.2 times higher odds of unfavourable outcome than centers at the lower end of the outcome distribution (the 'best' centers): exp (3.92*$\sqrt{}$0.09) = 3.2. This calculation is derived from Spiegelhalter et al who proposed a similar interpretation.[18] If there would be no unexplained between-center differences beyond random variation, τ^2 would be 0 and the 95% OR range would be 1.

All analyses in the total database were stratified by study to adjust for any systematic study effects, such as calendar time and inclusion criteria. We first considered random effects logistic regression models for crude between-center differences, and subsequently adjusted for any differences in patient characteristics between the centers by extending the regression model with three patient characteristics: age (as a continuous linear variable), pupillary reactivity (3 categories: none, one, both reacting), and GCS motor score (6 categories: none, extension, abnormal flexion, withdrawal, localizing/obeying commands, or untestable.[11]

We analyzed centers from Europe and the US separately, motivated by a previous analysis that found that patient outcomes were better in the US.[19] We further analyzed four of the larger RCTs separately, arbitrarily defined as having at least 20 centers and at least 20 patients per center on average since between-center effects can be more reliably estimated with larger numbers of centers and larger numbers of patients per center. As a sensitivity analysis we excluded centers with 3 patients or less, centers with 10 patients or less, and centers with 50 patients or less. We fitted the models in the total database with adaptive Gaussian quadrature with ten q points as an alternative to our default 'Laplace' method, leading to a better goodness of fit, but very similar estimates of the between-center variance. Analyses were performed with R statistical software 2.7.2 (R Foundation for Statistical Computation, Vienna) and SPSS 15.0 (SPSS Inc, Chicago).

Results

Patients and centers

Of the total 9,578 patients included in the study, 2,603 (27%) died and 4,629 (48%) had an unfavourable outcome six months after injury. The median age was 30 (interquartile range 21-45) years, 3900 patients (41%) had a GCS motor of 3 or lower (none, extension or abnormal flexion), and 1,914 patients (20%) had bilateral non-reactive pupils on admission (Table 10.1). The majority of patients were from Europe (n=5,705) and the US (n=3,325). The other patients were from Israel (n=225), Canada (n=152), Australia (n=147), Turkey (n=12), Argentina (n=8), Hong Kong (n=3) and South Korea (n=1).

There were 265 unique centers, with greatly varying patient numbers. The smallest centers treated only one patient and the largest center 453 patients (Figure 10.1).

Figure 10.1 Observed number of patients per center in 265 centers. Numbers vary from 1 to 453 with median 17 and interquartile range 6-45.

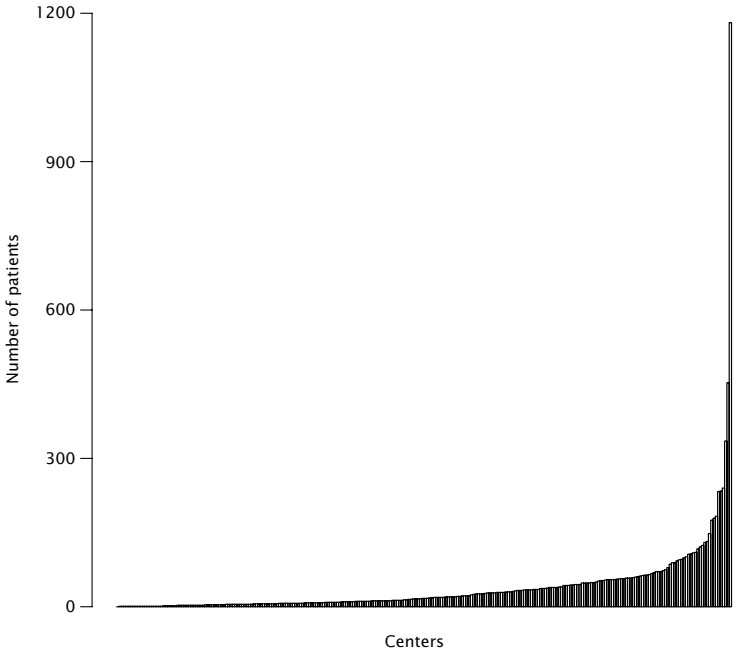


Table 10.1 Descriptive statistics of the studies in the IMPACT database used for analyses of between-center differences.

	TCDB*	UK4	EBIC	TIUS	TINT	SLIN
Study period	1984–87	1986–88	1995	1991–94	1992–94	1999
Original publication	Foulkes et al 1991	Murray et al 1999	Murray at al 1999		Marshall et al 1998	Morris et al 1999
n patients	604	791	822	1041	1118	409
N centers	4	4	64	34	39	50
Age						
Median (IQ range)	26(21–40)	36(22–55)	38(24–59)	30(23–41)	30(21–45)	28(21–43)
Motor						
1–2	243 (40%)	198 (25%)	230 (28%)	152 (15%)	141(12%)	55 (13%)
3	74 (12%)	37 (5%)	55 (7%)	132 (13%)	237(21%)	91 (22%)
4	122 (20%)	141 (18%)	113 (14%)	300 (29%)	317(29%)	127 (31%)
5–6	134 (22%)	221 (28%)	281 (34%)	457 (44%)	413 (37%)	136 (33%)
Untestable	31 (5%)	194 (25%)	143 (17%)	0 (0%)	0 (0%)	0 (0%)
Pupils						
Both reactive	300(50%)	434(55%)	532(65%)	708(68%)	807(72%)	314(77%)
One reactive	55(9%)	113(14%)	80(10%)	119(11%)	174(16%)	80(20%)
None reactive	249(41%)	244(31%)	210(26%)	214(21%)	137(12%)	15(4%)
Outome						
Unfavorable	393(65%)	518(65%)	422(51%)	395(38%)	456(41%)	177(43%)
Mortality	264(44%)	359(45%)	281(34%)	225(22%)	278(25%)	94(23%)

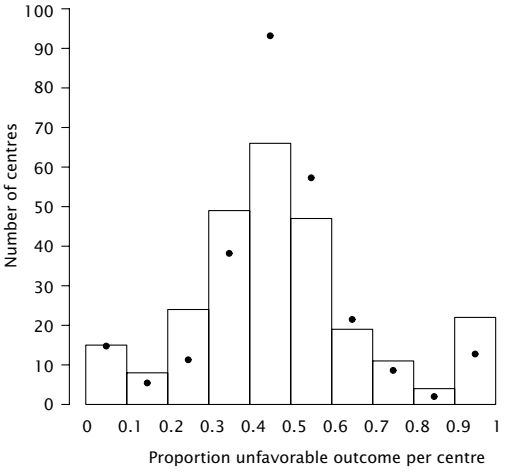
*__TCDB__ = Traumatic Coma Data Bank (observational study), __UK4__ = UK Four Center Study (observational study), __EBIC__ = European Brain Injury Consortium Core data study (observational study), __TINT__ = Tirilizad International (RCT), __TIUS__ = Tirilizad US (RCT), __SLIN__ = International Selfotel trial (RCT), __Saphir__ (RCT), __PEGSOD__ (RCT), __HIT I__ = HIT I Nimodipine (RCT), __HIT II__ = HIT II Nimodipine (RCT), __SKB__ = Bradycor SKB (RCT), __NABIS__ = NABIS Hypothermia (RCT), __Pharmos__ = Pharmos Dexanabinol (RCT)

Between-center differences

We found a large variation in outcome (from 0% to 100%), represented by the bars in figure 10.2. Part of this variation is explained by random variation between the centers caused by low numbers of patients in some centers. The points in figure 10.2 reflect the distribution of the 265 centers around the average proportion of unfavourable outcome of 48% that would be expected based on only random variation, given the sample sizes. We note that some centers are expected to have 0% or 100% unfavourable outcome due to random variation. But in the observed distribution (the bars), more centers are at the extremes of the distribution and fewer are in the center, reflecting between-center differences not explained by random variation (τ^2).

SAPHIR	PEGSOD	HIT I	HIT II	SKB	NABIS	Pharmos	Total
1995–97	1993–95	1987–89	1989–91	1996	1994–98	2000–04	
	Young et al 1996	Bailey et al 1991	Eur Study Group 1994	Marmarou et al 1999	Clifton et al 2001	Maas et al 2006	
919	1510	350	819	126	213	856	9578
51	72	6	21	26	6	86	265
32(20–38)	27(20–38)	34(21–47)	33(22–49)	27(20–39)	30(22–40)	33(23–45)	30(21–45)
264 (29%)	655 (43%)	163(47%)	280 (34%)	56(43%)	85(39%)	134(16%)	2656(28%)
143 (16%)	165 (11%)	45(13%)	92 (11%)	14(11%)	23(11%)	136(16%)	1244(13%)
223 (24%)	334 (22%)	56(16%)	181 (22%)	16(13%)	43(20%)	225(26%)	2208(23%)
286 (31%)	356 (24%)	77(22%)	207 (25%)	23(19%)	58(27%)	235(27%)	2884(31%)
3 (0%)	0 (0%)	9(3%)	59 (7%)	17(13%)	4(2%)	126(15%)	586(6%)
655(71%)	792(52%)	232(66%)	579(71%)	76(60%)	141(66%)	666(78%)	6236(65%)
264(29%)	156(10%)	50(14%)	102(12%)	50(40%)	31(15%)	154(18%)	1428(15%)
0(0%)	562(37%)	68(19%)	138(17%)	0(0%)	41(19%)	36(4%)	1914(20%)
378(41%)	774(51%)	171(48%)	328(40%)	70(56%)	110(52%)	437(51%)	4629(48%)
212(23%)	362(24%)	99(28%)	188(23%)	34(27%)	62(29%)	145(17%)	2603(27%)

Figure 10.2 Observed proportion of patients with unfavourable outcome per center (bars). Points represent what would be expected if there were only random differences between the 265 centers, i.e. if they all had a constant risk of unfavourable outcome of 48%. This distribution was created by drawing 1000 samples from a binomial distribution with probability of 48%. The points at the extremes of the distribution are higher since there are centers with only 1 or 2 patients that are expected to have a 0% or 100% unfavourable outcome



Stratified for study, the between-center differences in outcome were 2.4 fold, meaning that the odds on unfavourable outcome in centers at the higher end of the outcome distribution was 2.4 times higher than in centers at the lower end. Patient characteristics were highly predictive for outcome (Table 10.2).

Table 10.2 Effects on unfavourable outcome of predictors in the adjustment model

	OR	95% CI	P value
Age (per decade)	1.43	1.39–1.49	<0.001
GCS motorscore			<0.001
None	1.0 (ref)		
Extension	1.90	1.57–2.31	
Abnormal flexion	0.95	0.79–1.14	
Withdrawal	0.51	0.44–0.60	
Localizing/obeying commands	0.28	0.16–0.31	
Untestable	0.61	0.48–0.77	
Pupil reactivity			<0.001
Both reactive	1.0 (ref)		
One reactive	2.13	1.87–2.43	
None reactive	4.09	3.57–4.70	

Patients who were localizing or obeying commands had an odds of unfavourable outcome 0.28 times the odds of patients with an absent motor response. Patients with extension had almost twice the odds of having unfavourable outcome compared to patients with an absent motor response. Patients with non reactive pupils had 4.09 times the odds of having unfavourable outcome compared to patients with two reactive pupils.

Patient characteristics varied largely between the centers. The mean predicted probability of unfavourable outcome based on age, motor and pupils ranged from 13% to 93%. After adjustment for these three patient characteristics, the between-center differences increased to a 3.3 fold difference in the odds of unfavourable outcome between the 'best' and the 'worst' centers. The adjusted between-center differences for patient characteristics were smaller in the US (2.4 fold difference in odds) than in Europe (3.8 fold difference in odds). There was a small difference in outcome between the USA and Europe. In Europe the percentage unfavourable outcome was 48%, in the USA it was 50%. After adjustment for patient characteristics and study there was no significant difference in outcome between the USA and Europe.

The adjusted between-center differences for patient characteristics were smaller in the RCTs (3.3 fold difference in odds) than in the observational studies (13.1 fold difference in odds).

The TIUS, TINT, PEGSOD and HIT II studies had at least 20 centers and at least 20 patients per center. In TIUS and TINT, there was a 3.1 fold adjusted difference in the odds

of unfavourable outcome between the 'best' and the 'worst' centers. The between-center differences were larger in HIT II (4.7 fold) and smaller in PEGSOD (1.8 fold). (Table 10.3)

The sensitivity analysis excluding centers with low numbers of patients resulted in only slightly smaller between-center differences, indicating robustness of the results.

Table 10.3 Between center differences in total database and within studies.

	Unadjusted		**Adjusted for age + motor + pupils**	
	Tau^2	Difference in the odds on unfavourable outcome between centers at the 97.5th and the 2.5th percentile of the outcome distribution	Tau^2	Difference in the odds on unfavourable outcome between centers at the 97.5th and the 2.5th percentile of the outcome distribution
Total* (n=9578)	0.052	2.4 fold	0.095	3.3 fold
USA (n=3325)	0.033	2.0 fold	0.046	2.4 fold
Europe (n=5706)	0.052	2.4 fold	0.115	3.8 fold
Obs. studies (n=2217)	0.309	8.8 fold	0.431	13.1 fold
RCTs (n=7361)	0.045	2.3 fold	0.093	3.3 fold
HIT II (n=819)	0.074	2.9 fold	0.157	4.7 fold
TIUS (n=1041)	0.071	2.9 fold	0.080	3.1 fold
TINT (n=1118)	0.045	2.3 fold	0.083	3.1 fold
PEGSOD (n=1510)	0.000	1.0 fold	0.02	1.8 fold

*Models in the total database were adjusted for study

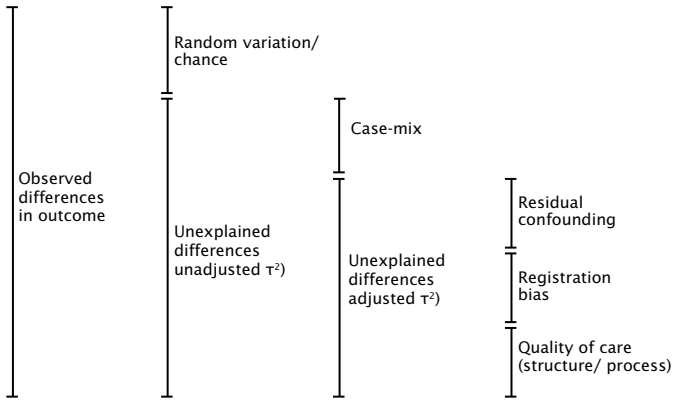
Discussion

In this study we quantified the differences in patient outcome after severe and moderate traumatic brain injury between centers enrolling patients in randomized controlled trials (RCTs) and surveys. After adjustment for patient characteristics and taking account of random variation, there was a 3.3 fold difference in the odds of unfavourable outcome after six months between the centers at the lower end and those at the higher end of the outcome distribution.

Limitations of the study include that our data consist of multiple individual studies, including RCTs, with varying inclusion criteria and varying in calendar time. However all analyses were adjusted for study and a substantial number of centers participated in multiple studies. None of the included trials showed a significant treatment effect, so it is not likely that differences in study treatment have influenced the results. Although it has been suggested that outcomes have improved over time, we did not detect a time effect in this dataset after adjustment for patient characteristics (odds ratio 0.99 per year after adjustment for patient characteristics, p=0.34). Another limitation is that even with the large number of patients and centers in the current analysis, estimation of τ^2 is associated with some uncertainty, which we did not calculate.

Conceptually, the observed between-center differences consist partly of random variation, which we accounted for with the random effect models (Figure 10.3).

Figure 10.3 Schematic partitioning of between-center differences. Observed differences consist partly of chance (2nd bar). Remaining differences consist partly of patient characteristics (3rd bar). The then remaining differences, as quantified in this study might be due to residual confounding, registration bias or differences in quality of care. Magnitude of the bars does not represent real numbers.



Contrary to our expectations the between-center differences did not decrease, but even increased after adjustment for patient characteristics. This implies that some centers with patients with good prognoses have unexpectedly unfavourable outcomes and some centers with patients with unfavourable prognosis have unexpectedly good outcomes. A likely explanation for for these differences – at least in part – is quality of care (Figure 10.3).

Another explanation might be that the model we used included too few patient characteristics to adequately adjust for confounding. We therefore also used a more extended model (10 predictors including secondary insults, CT characteristics and lab values) in the Tirilizad trials (TIUS and TINT), which had more data available. This led only to a small change in between-center differences, compared to adjustment for only age, motor score and pupils.

Between-centers differences in other medical fields are reported in numerous ways: observed percentages of unfavourable o utcome or p-values for the overall between center-differences.[4, 5] Therefore it is difficult to compare our results to these other studies. Steyerberg et al reported a 95% OR range for the regional differences in 30 day mortality after myocardial infarction in the GUSTO I trial. They found a 95% OR range from 0.93 to 1.07, which corresponds to a 1.2 fold difference between the 'best' and the 'worst' regions, so much smaller than we found in our study.[20]

The IMPACT database consists only of experienced centers participating in large RCTs and observational studies. Moreover trials have strict protocols with regard to patient inclusion and treatment, which reduces heterogeneity. Consequently our findings may underestimate 'real world' between-center differences across all centers treating patients with TBI.

In this study we adjusted the between-center differences for patient characteristics and we accounted for random variation. The question remains of what causes the remaining unexplained differences. There might be residual confounding, but there also might be differences in registration between centers causing apparent differences in outcome (figure 2). The only between-center differences that are potentially avoidable are those caused by differences in quality of care. These differences can occur in two domains: structure and process. Structure concerns for example volume (the number of patients treated) and the referral pattern in a region. Process concerns actions in individual patients, such as surgical and medical management.

We found that between-center differences in the US were smaller than in Europe. This might be due to the fact that in Europe differences between countries exist in the organization of neurosurgical care or that more variation exists in the approaches to treatment. Geographic region or country may therefore indirectly explain part of the between-center differences, and should be further investigated. The percentage unfavourable outcome was similar in the US and Europe after adjustment for patient characteristics in contrast to previous analysis of only the Tirilizad trials.[19]

We found that between-center differences were much larger in the observational studies than in RCTs. We do not know however whether this is the result of less (unobserved) patient variation, more uniform registration, or less treatment variation in RCTS. The substantial differences in the between-center differences between the trials we analysed separately (e.g. 4.7 fold differences in HIT II, 1.8 fold differences in PEG-SOD) confirm that there are RCTs with less variation but again it is unknown whether this is the results of less variation in treatment.

Previous studies in traumatic brain injury also give directions for possible causes of between-center differences in outcome. One study found an 18% larger discharge mortality in centers with a non-aggressive approach to treatment compared to centers with an aggressive approach, suggesting an effect of different treatment policies.[21] A study from the United Kingdom found that patients with severe head injury who were treated in a non-neurosurgical center had case-mix adjusted odds of death of 2.15 times that of patients treated in a neurosurgical center, pointing to the possible relevance of structural indicators such as of trauma organization and infrastructure.[22] In the NABIS Hypothermia trial it was found that both treatment and outcome varied significantly between centers, particularly between small and large centers.[7] In our study, the different sensitivity analyses excluding centers with few patients might be interpreted as an indication of volume-effects. However, the number of patients enrolled in a study is largely determined by the inclusion criteria and the duration of the study, and does not reflect the actual volume of a center. Detailed analysis is complex and we consider this a topic for future research.

In the NABIS Hypothermia trial the relevance of between-center differences for trial design was addressed. The large between-center differences observed in our study emphasize this important issue and raise the question whether the current practice of increasing the number of participating centers to reduce recruitment time is desirable. It is important to achieve balance in the randomization within each center, and to standardize treatment across centers as much as possible, to avoid that the treatment effect is confounded by the large between-center differences.

The large between-center differences furthermore implicate that there is considerable room for improvement of quality of care and reduction of unfavourable outcome. It is essential to understand possible causes of the observed differences and thus research into this should be prioritized.

Acknowledgements

This study was funded by NIH (NS-42691)

References

1. Kelly A, Thompson JP, Tuttle D, Benesch C, Holloway RG. Public reporting of quality data for stroke: is it measuring quality? *Stroke* 2008;39(12):3367-3371.
2. Krumholz HM, Keenan PS, Brush JE, Jr., et al. Standards for measures used for public reporting of efficiency in health care: a scientific statement from the American Heart Association Interdisciplinary Council on Quality of Care and Outcomes Research and the American College of Cardiology Foundation. *Circulation* 2008;118(18):1885-1893.
3. Wright J, Shojania KG. Measuring the quality of hospital care. *Bmj* 2009;338:b569.
4. Biau DJ, Halm JA, Ahmadieh H, et al. Provider and center effect in multicenter randomized controlled trials of surgical specialties: an analysis on patient-level data. *Ann Surg* 2008;247(5):892-898.
5. Bradley EH, Herrin J, Elbel B, et al. Hospital quality for acute myocardial infarction: correlation among process measures and relationship with short-term mortality. *Jama* 2006;296(1):72-78.
6. Krumholz HM, Chen J, Wang Y, Radford MJ, Chen YT, Marciniak TA. Comparing AMI mortality among hospitals in patients 65 years of age and older: evaluating methods of risk adjustment. *Circulation* 1999;99(23):2986-2992.
7. Clifton GL, Choi SC, Miller ER, et al. Intercenter variance in clinical trials of head trauma--experience of the National Acute Brain Injury Study: Hypothermia. *J Neurosurg* 2001;95(5):751-755.
8. Maas AI, Marmarou A, Murray GD, Teasdale SG, Steyerberg EW. Prognosis and clinical trial design in traumatic brain injury: the IMPACT study. *J Neurotrauma* 2007;24(2):232-238.
9. Marmarou A, Lu J, Butcher I, et al. IMPACT database of traumatic brain injury: design and description. *J Neurotrauma* 2007;24(2):239-250.
10. Jennett B, Bond M. Assessment of outcome after severe brain damage. Lancet 1975;1(7905):480-484.
11. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet* 1974;2(7872):81-84.
12. Perel P, Arango M, Clayton T, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *Bmj* 2008;336(7641):425-429.
13. Murray GD, Butcher I, McHugh GS, Lu J, Mushkudiani NA, Maas AI, et al. Multivariable prognostic analysis in traumatic brain injury: results from the IMPACT study. *J Neurotrauma* 2007;24(2):329-337.
14. Steyerberg EW, Mushkudiani N, Perel P, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med* 2008;5(8):e165; discussion e165.
15. McHugh GS, Butcher I, Steyerberg EW, et al. Statistical approaches to the univariate prognostic analysis of the IMPACT database on traumatic brain injury. *J Neurotrauma* 2007;24(2):251-258.
16. DeLong ER, Peterson ED, DeLong DM, Muhlbaier LH, Hackett S, Mark DB. Comparing risk-adjustment methods for provider profiling. *Stat Med* 1997;16(23):2645-2664.
17. Timbie JW, Newhouse JP, Rosenthal MB, Normand SL. A cost-effectiveness framework for profiling the value of hospital care. *Med Decis Making* 2008;28(3):419-434.
18. Spiegelhalter DJ AK, Myles JP. *Bayesian Approaches to Clinical Trials and Health-Care Evaluation.* Chichester, UK: John Wiley, 2004.

19. Hukkelhoven CW, Steyerberg EW, Farace E, Habbema JD, Marshall LF, Maas AI. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. *J Neurosurg* 2002;97(3):549-557.
20. Steyerberg EW, Eijkemans MJ, Boersma E, Habbema JD. Applicability of clinical prediction models in acute myocardial infarction: a comparison of traditional and empirical Bayes adjustment methods. *Am Heart J* 2005;150(5):920.
21. Bulger EM, Nathens AB, Rivara FP, Moore M, MacKenzie EJ, Jurkovich GJ. Management of severe head injury: institutional variations in care and effect on outcome. *Crit Care Med* 2002;30(8):1870-1876.
22. Patel HC, Bouamra O, Woodford M, King AT, Yates DW, Lecky FE. Trends in head injury outcome from 1989 to 2003 and the effect of neurosurgical care: an observational study. *Lancet* 2005;366(9496):1538-1544.

11 Between-center differences and treatment effects in randomized controlled trials

Lingsma HF, Roozenbeek B, Perel B, Roberts I, Maas AIR, Steyerberg EW. Between-center differences and treatment effects in randomized controlled trials: a case study in traumatic brain injury. Submitted.

Abstract

Introduction
Large between-center differences in outcome exist in Traumatic Brain Injury (TBI). The aim of this study was to assess the influence of such differences on the estimated treatment effect in a large randomized controlled trial (RCT).

Methods
We used data from the MRC CRASH trial on the efficacy of corticosteroid infusion in patients with TBI. We analyzed the effect of the treatment on 14 day mortality with fixed effect logistic regression. Next we used random effects logistic regression with a random intercept to estimate the treatment effect taking into account between-center differences in outcome. Between-center differences in outcome were expressed with a 95% range of odds ratios for centers compared to the average, based on the variance of the random effects (tau^2). A random effects logistic regression model with random slopes was used to allow the treatment effect to vary by center. The variation in treatment effect between the centers was expressed in a 95% range of the estimated treatment ORs.

Results
In 9978 patients from 237 centers, 14-day mortality was 19.5%. Mortality was higher in the treatment group (OR=1.22, p=0.00010). Using a random effects model showed large between-center differences in outcome (95% range of center effects: 0.27- 3.71), but did not substantially change the estimated treatment effect (OR=1.24, p=0.00003). There was limited between-center variation in the treatment effect (OR=1.22, 95% treatment OR range: 1.17-1.26).

Conclusion
Large between-center differences in outcome do not necessarily affect the estimated treatment effect in RCTs.

Introduction

Traumatic brain injury (TBI) is a major health and socio-economic problem throughout the world. It is the field with one of the greatest unmet needs in medicine and public health.[1] Not only is TBI a major cause of death and disability, incurring great personal suffering to victims and relatives, but it also leads to huge direct and indirect costs to society.[2]

Many randomized controlled trials (RCTs) have been performed to investigate the effectiveness of new therapies in TBI, but very few have convincingly demonstrated benefit.[3] Multiple factors may have contributed to this disappointing picture, including RCTs in TBI being too small to detect or refute reliably moderate but clinically important benefits or hazards of treatment.[4] To design trials of sufficient size to detect moderate treatment effects, participation of multiple centers is required.

Considerable between-center differences in patient outcome have been reported in TBI.[5-7] Recently it was shown that a 3.3-fold difference between centers in the odds of having an unfavourable outcome exist ($p<0.001$), which was not explained by random variation or patient characteristics.[8]

It has been hypothesised that such between-center differences in outcome influence the chances of demonstrating a treatment effect in RCTs.[7,9] The aim of this study is to assess the effect of between-center differences on estimates of the treatment effect in a large RCT in TBI.

Methods

Data

We used the individual patient data of the MRC CRASH trial. The CRASH trial (corticosteroid randomisation after significant head injury) is a large, international, randomised placebo-controlled trial of the effect of early administration of 48 h infusion of corticosteroids (methylprednisolone) on risk of death and disability after head injury. Patients from 239 centers in 48 countries were enrolled between April 1999 and May 2004, when the steering committee stopped recruitment because of a higher 14 day mortality rate in the treatment group.[10]

Analysis

We first assessed whether there were differences in outcome between the centers in the CRASH trial, using a random effect logistic regression model (Appendix 1). In this model the outcome of a patient is only determined by the center that treats the patient. Since some centers only treat a small number of patients, part of the between-center differences are caused by random variation. The random effect model estimates the between-center differences beyond random variation. The between-center differences

are expressed as τ^2, which is the variance of the random effects.

Part of the differences between centers may be caused by the fact that centers are from a particular country. To separate between-center differences from between-country differences we extended the random effect model with a country level.

Because part of the between-center effect may be explained by differences in patient characteristics, we adjusted the between-center differences in outcome for age, Glasgow Coma Scale (GCS) and pupil reactivity at admission. These are the three main generally accepted prognostic factors in TBI.[11, 12] Age and GCS (a scale from 1-15) where treated as continuous variables and pupil reactivity as a binary variable (both pupils reactive versus one or both unreactive). So now the outcome of a patient is determined by patient characteristics and center.

The differences between centers in outcome were expressed in a 95% range of odds ratios for centers compared to the average. To avoid confusion with the odds ratio of the treatment effect we refer to this range as the 95% center effect range.

Next we estimated the treatment effect with and without taking the between-center differences into account. We first analyzed the univariate effect of the treatment on 14 day mortality with usual fixed effect logistic regression. Center effects were ignored, which is a common approach also in multicenter trials. We considered this as the reference strategy.

We furthermore use a random effect model to estimate the treatment effect. The outcome is also determined by the center, so the treatment effect is adjusted for between center-differences. This approach assumes a uniform treatment effect across centers. This means we expect the treatment to have equal effects in each center. As a second approach we used a random effect logistic regression model with interaction between center and treatment to asses whether the treatment effect varied between the centers. The variation in estimated treatment effect was expressed in a 95% range of the estimated treatment effect across centers. We compared the estimates of the treatment effect and the p-values in the two approaches with the reference strategy.

The random effect estimates of the individual centers for both outcome and treatment effect were plotted with 95% posterior intervals.

Statistical analysis where performed in R statistical software 2.7.2 using the Design and lme4 libraries (R Foundation for Statistical Computation, Vienna). Random effect models were fitted with Adaptive Gaussian Quadrature with 10 qpoints.

Results

Descriptives

In total 10,008 patients were included in the RCT. We excluded 30 patients with missing 14 day outcome, leaving 9978 patients from 237 centers for the analyses. After 14 days 1,948 (19.5%) of the patients had died, with higher mortality in the treatment group. (Table 11.1)

Table 11.1 Baseline characteristics and 14 day mortality of patients enrolled in the CRASH trial with mortality data available (n=9978).

	Corticosteroid (n=4991)	**Placebo (n=4987)**
Age (median, interquartile range)	33, 23–47	32, 23–47
Gender		
Male	4060 (81.3%)	4016 (80.5%)
Glasgow Coma Scale		
Severe (3–8)	1966 (39.4%)	1966 (39.4%)
Moderate (9–12)	1554 (31.1%)	1479 (29.7%)
Mild (13–14)	1471 (29.5%)	1542 (30.9%)
Pupillary reactivity		
Both reactive to light	4272 (85.6%)	4016 (80.5%)
14 day mortality		
Dead	1053 (21.1%)	895 (17.9%)

Between-center differences

There was a large difference between centers in outcome ($\tau^2_{\text{outcome, centre}} = 0.447$, $p<0.00001$).The corresponding 95% range of center effects was 0.27- 3.71 (Table 11.2). This means that in centers with the lowest mortality (2.5th percentile) the odds of dying was 0.27 times the average, while in the centers the highest mortality (97.5th percentile) the odds of dying was 3.71 times the average. After adjustment for age, GCS and pupil reactivity the between-center in outcome increased to $\tau^2_{\text{outcome, centre}} = 0.620$ ($p<0.00001$) with a corresponding 95% range of center effects of 0.21- 4.68. Figure 11.1 shows the estimated adjusted odds ratios for mortality for each center, compared to the average, with 95% posterior intervals.

Figure 11.1 Differences between centers in mortality, adjusted for age, GCS, pupil reactivity and treatment in a random effects model. A center with average mortality has log odds 0, a positive log odds indicates higher mortality. Lines indicate 95% posterior interval.

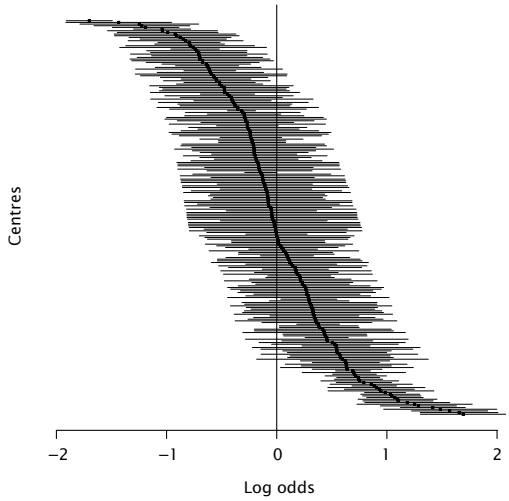


Table 11.2 Between-center and between-country variation in 14 day mortality, unadjusted and adjusted for treatment, age, GCS, and pupillary reactivity.

	Unadjusted		**Adjusted (Conditional)**	
	Tau^2	95% range	Tau^2	95% range
Between-centers	0.447 (p<0.00001)	0.27–3.71	0.620 (p<0.00001)	0.21–4.68
Between-counties	0.385 (p<0.00001)	0.30–3.37	0.642 (p<0.00001)	0.21–4.81
Combined:				
Between-centers	0.331 (p<0.00001)	0.32–3.09	0.235 (p<0.00001)	0.39–2.58
Between-counties	0.142 (p<0.00001)	0.48–2.09	0.470 (p<0.00001)	0.26–3.88

Part of the differences in outcome between centers were actually differences between countries. When taking into account that centers are from a particular country, the range of between-center differences decreased to 0.39-2.58 ($\tau^2_{\text{outcome, centre | country}}$ =0.235, p<0.00001). The range of between-country differences was 0.26 to 3.88 ($\tau^2_{\text{outcome, centre | country}}$ =0.470, p<0.00001).

Treatment effect

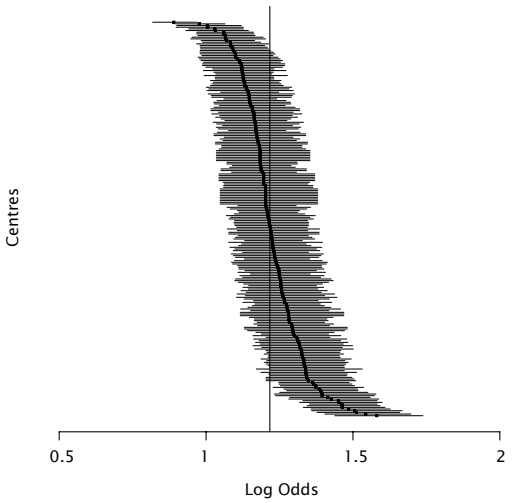
In the reference strategy, the univariate fixed effect logistic regression odds ratio (OR) for treatment was 1.22 (p=0.0001, Table 11.3).

Table 11.3 Estimated unadjusted treatment effects (odds ratio (OR) and p value) on 14 day mortality with different approaches taking into account between-center differences.

Approach	Model	OR unadjusted	P value tx effect
– Uniform treatment effect over centers – No adjustment for between-center differences	Fixed effect logistic regression	1.22	0.00010
– Uniform treatment effect over centers – Adjustment for between-center differences	Random effect logistic regression with random intercept	1.24	0.00003
– Varying treatment effect over centers – Adjustment for between-center differences	Random effect logistic regression with random slope	1.22 (95% range: 1.17–1.26)	0.00029

Our first approach of adjusting for the between-center heterogeneity resulted in an OR for the treatment effect of 1.24 (p=0.00003). With our second approach we estimated a varying treatment effect between the centers. The mean OR was 1.22 (p= 0.00029). The treatment effect heterogeneity was small, but statistically significant ($\tau^2_{\text{treatment effect}}$ =0.02, p<0.00001). The corresponding 95% range of the estimated treatment effects across centers was 1.17–1.26 (Figure 11.2)

Figure 11.2 Between-center differences in treatment effect, in a random effect model. The overall treatment effect is log odds=0.20 (OR=1.22). Lines indicate 95% posterior interval.



Discussion

Although we found large between-center differences in outcome in the CRASH trial, taking these into account did not substantially change the estimated treatment effect. Neither did we see major differences in treatment effect by center. This study provides no support for the hypothesis that between-center differences in outcome affect the chances of demonstrating a treatment effect in RCTs, in contrast to current beliefs in this clinical area.[7, 9]

Considering differences between centers in outcome and in estimated treatment effect could be of importance from two perspectives. First, between-center heterogeneity in the treatment effect between may indicate limited generalizability, which is of importance for example when registering a drug in a particular country. In our study there was no meaningful heterogeneity in overall treatment effect. Although the between-center differences in the treatment effect were statistically significant, the 95% range was small (1.17-1.26). Clearly, determining generalizability is not solely a statistical issue but requires a clinical judgement to the extent to which the trial results might apply to another population.

Some trials have estimated the heterogeneity of the treatment effect between centers or countries or regions, but did not use random effect modelling. The PLATO study (The Study of Platelet Inhibition and Patient Outcomes) compared two platelet inhibitors (Ticagrelor versus Clopidogrel) for prevention of cardiovascular events in patients with acute coronary syndrome. The overall treatment effect was a hazard ratio (HR) of 0.84 in favour of Ticagrelor. The treatment effect was also tested in four different geographic regions separately; Asia-Australia (N=1714), Central-South America (N=1237), Europe-Middle East-Africa (N=13859), and North America (N=1814). In Europe the estimated HR was 0.80 (95% CI: 0.72-0.90). The HRs in Asia-Australia, Central-South America were 0.80 and 0.86, both non statistically significant. The estimated HR in North America was however 1.25 (95% CI: 0.93-1.67). The authors state that 'the difference in results between patients enrolled in North America and those enrolled elsewhere raises the questions of whether geographic differences between populations of patients or practice patterns influenced the effects of the randomized treatments, although no apparent explanations have been found.'

This interpretation shows the importance to distinct statistical from clinical reasoning. Although the statistical analysis showed significant differences between geographic regions in the PLATO trial, which could be an indication of limited generalizability, the authors have no biological or mechanistic explanation for the heterogeneity of the treatment effect and no heterogeneity was expected on beforehand. It thus might have been a better choice to use a random effect model to estimate the between-region differences in treatment effect.

Second, it is thought that heterogeneity between centers might reduce statistical

power to detect the treatment effect.[9] Providing that a trial is large enough, randomization will ensure that the intervention and control group are similar with regard to known and unknown confounders.[10] As expected, our study showed that taking into account between-center differences did not affect statistical significance.

Several explanations can be given for our findings. First, differences in outcome between centers in RCTs may be caused by patient characteristics, which we adjusted for in this analysis. We may not expect that patient characteristics result in differences in treatment effect between centers if the treatment is assumed to work for all patients included in the trial. Secondly there may be differences in care. If these only affect the baseline event rate (e.g. fewer ICU capacity) the treatment effect is not likely to be influenced. In contrast there could be differences in care interacting with the treatment, e.g. if time to hospital arrival is structurally longer in some places, an acute treatment may be less effective. If such an interaction is expected, it would usually be captured in inclusion criteria, such as inclusion within a certain time after injury. In our study we found large differences in outcome between the centers but limited variability in the treatment effect. In other words, there was no substantial interaction between center and treatment, although such an interaction might have been expected since the CRASH trial comprised an acute treatment and was conducted in low- to high- income countries. This is also an important finding from the perspective of standardisation of care in trials, which some consider very important.[9] Our study suggests that if non-standardized care only influences the absolute risk and does not interact with the treatment, there is no reason to put much effort in standardizing care.

We consider our results to be applicable to drug interventions, which work on physiological mechanisms. Trials investigating a more complex intervention such as surgery or a complex treatment strategy may be more sensitive to differences in quality of care. We recognize that further studies are required to confirm or refute these findings for other types of interventions and for other diseases. Moreover it is crucial to think in advance on the mechanism of the treatment, and whether heterogeneity or homogeneity of the treatment effect by center is expected.

In this study we have assessed heterogeneity of the treatment effects on a relative scale, but we can also use an absolute scale (risk difference). We found that there is no heterogeneity on the relative scale, despite heterogeneity in the absolute risks per center. This combination implies that there is heterogeneity in treatment effects on an absolute scale, which is important to realize when considering treatment for individuals.[13]

The demonstration of hetero- or homogeneity in treatment effects by country or center in the single study is conceptually the same as demonstration hetero- or homogeneity is a meta-analysis. The CRASH trial could be seen as a prospective meta-analysis of 40 trials in 40 different countries. A simple way showing the heterogeneity in treatment effects would be to present the results of a forest plot meta-analysis and

test for heterogeneity. This was done for the CRASH trial (data not shown), also not indicating heterogeneity.

Part of the between-center differences were actually between-county differences. This could be an indication of center-differences being caused by structural differences between countries such as availability of resources and organisation of trauma care.

Our study has some limitations. First, we did not consider differences in data quality between the centers, which might affect the treatment effect.[7] Second, the CRASH might be considered an exception in the sense that the treatment was harmful. However, it is unlikely that our results would depend on the direction of the treatment effect.

Summarizing, our study shows that there were large between center differences in the CRASH trial, which did not affect the estimated treatment effect. Between-center differences do not affect the chances of demonstrating a treatment effect, which supports the conduct of large, multi-center trials.

Acknowledgments

This work was supported by the National Institutes of Health [NS-01923521]

References

1. Maas AI, Stocchetti N, Bullock R. Moderate and severe traumatic brain injury in adults. *Lancet Neurol* 2008; 7(8): 728-41.
2. Finkelstein EA, Corso PS, Miller TR. *The incidence and economic burden of injuries in the United States*. New York: Oxford University Press; 2006.
3. Maas AI, Marmarou A, Murray GD, Teasdale SG, Steyerberg EW. Prognosis and clinical trial design in traumatic brain injury: the IMPACT study. *J Neurotrauma* 2007; 24(2): 232-8.
4. Dickinson K, Bunn F, Wentz R, Edwards P, Roberts I. Size and quality of randomised controlled trials in head injury: review of published studies. *BMJ* 2000; 320(7245): 1308-11.
5. Maas AI, Murray G, Henney H,3rd, Kassem N, Legrand V, Mangelus M, et al. Efficacy and safety of dexanabinol in severe traumatic brain injury: results of a phase III randomised, placebo-controlled, clinical trial. *Lancet Neurol* 2006; 5(1): 38-45.
6. Clifton GL, Drever P, Valadka A, Zygun D, Okonkwo D. Multicenter trial of early hypothermia in severe brain injury. *J Neurotraum*a 2009; 26(3): 393-7.
7. Clifton GL, Choi SC, Miller ER, Levin HS, Smith KR,Jr, Muizelaar JP, et al. Intercenter variance in clinical trials of head trauma--experience of the National Acute Brain Injury Study: Hypothermia. *J Neurosurg* 2001; 95(5): 751-5.
8. Lingsma HF, Roozenbeek B, Bayoue L, Lu J, Weir J, Butcher I, et al. Large between-center differences in outcome after moderate and severe traumatic brain injury in the IMPACT* study. *Neurosurgery* In Press; .
9. Kirkpatrick PJ. On guidelines for the management of the severe head injury. *J Neurol Neurosurg Psychiatry* 1997; 62(2): 109-11.
10. Roberts I, Yates D, Sandercock P, Farrell B, Wasserberg J, Lomas G, et al. Effect of intravenous corticosteroids on death within 14 days in 10008 adults with clinically significant head injury (MRC CRASH trial): randomised placebo-controlled trial. *Lancet* 2004; 364(9442): 1321-8.
11. MRC CRASH Trial Collaborators, Perel P, Arango M, Clayton T, Edwards P, Komolafe E, et al. Predicting outcome after traumatic brain injury: practical prognostic models based on large cohort of international patients. *BMJ* 2008; 336(7641): 425-9.
12. Steyerberg EW, Mushkudiani N, Perel P, Butcher I, Lu J, McHugh GS, et al. Predicting outcome after traumatic brain injury: development and international validation of prognostic scores based on admission characteristics. *PLoS Med* 2008; 5(8): e165; discussion e165.
13. Rothwell PM, Mehta Z, Howard SC, Gutnikov SA, Warlow CP. Treating individuals 3: from subgroups to individuals: general principles and the example of carotid endarterectomy. *Lancet* 2005; 365(9455): 256-65.

Appendix 1

Random effect logistic regression with random intercept for center

$$\text{Logit}\,(p(Y_{ij}=1)) = \beta_0 + (u_{0j} + e_{0ij}) \quad (1)$$

with Y_{ij} the outcome for patient j in center i, β_0 the intercept, u_{0j} the random intercept for the center, and e_{0ij} the residuals. The random intercepts are assumed to be normally distributed with $\tau^2_{0j} = \text{var}(u_{0j})$.

Random effect logistic regression with random intercepts for center and country

$$\text{Logit}\,(p(Y_{ij}=1)) = \beta_0 + (u_{0j} + u_{0k} + e_{0ijk}) \quad (2)$$

with u_{0k} the random intercept for the country, and e_{0ijk} the residuals. The random intercepts are assumed to be normally distributed with $\tau^2_{0j} = \text{var}(u_{0j})$ and $\tau^2_{0kj} = \text{var}(u_{0k})$.

Random effect logistic regression with random intercept for center, including patient characteristics

$$\text{Logit}\,(p(Y_{ij}=1)) = \beta_0 + \beta_1 x_{ij} + (u_{0j} + e_{0ij}) \quad (3)$$

with patient characteristics x_{ij}

Range of the center effects

$$95\%\ \text{center effect range} = \exp(1.96*\tau^2_{0j})\ ;\ \exp(1.96* - \tau^2_{0j}) \quad (4)$$

Fixed effect logistic regression

$$\text{Logit}\,(p(Y_{ij}=1)) = \beta_0 + \beta_1 x_{ij} + e_{ij} \quad (5)$$

with x_{ij} the treatment and β_1 the treatment effect.

Random effect logistic regression with random intercept for center, including treatment

$$\text{Logit}\,(p(Y_{ij}=1)) = \beta_0 + \beta_1 x_{ij} + (u_{0j} + e_{0ij}) \quad (6)$$

with x_{ij} the treatment and β_1 the treatment effect, and random intercept u_{0j}

Random effect logistic regression with random slope of the treatment effect per center

$$\text{Logit}\,(p(Y_{ij}=1)) = \beta_0 + \beta_1 x_{ij} + (u_{1j} + e_{1ij}) \quad (7)$$

with u_{1j} as the random slope. The random slopes are assumed to be normally distributed with $\tau^2_{1j} = \text{var}(u_{1j})$

Range of the estimated treatment effect across centers

$$95\%\ \text{treatment effect range} = \exp(\beta_1 + 1.96*\tau^2_{1j})\ ;\ \exp(\beta_1 + 1.96* - \tau^2_{1j}) \quad (8)$$

12 Variation between hospitals in outcome after stroke is only partly explained by differences in quality of care.

Lingsma HF, Dippel DWJ, Hoeks S, Steyerberg EW, Franke CL, van Oostenbrugge RJ, de Jong G, Simoons ML, Scholte op Reimer WJM, and The Netherlands Stroke Survey investigators. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. *J Neurol Neurosurg Psychiatry* 2008;79:888–894

Abstract

Background and purpose
Patient outcome is often used as an indicator of quality of hospital care. The aim of this study is to investigate whether there is a straightforward relationship between quality of care and outcome and whether outcome measures could be used to assess quality of care after stroke.

Methods
In 10 centers in the Netherlands, 579 patients with acute stroke were prospectively and consecutively enrolled. Poor outcome was defined as a score on the modified Rankin scale ≥3 at 1 year. Quality of the care was assessed by relating diagnostic, therapeutic and preventive procedures to indication. Multiple logistic regression models were used to compare observed proportions of patients with poor outcome with expected proportions, after adjustment for patient characteristics and quality of care parameters.

Results
271 (47%) patients were dead or disabled at 1 year. Poor outcome varied across the centers from 29% to 78%. Large differences between centers were also observed in clinical characteristics, prognostic factors and quality of care. For example, between hospital quartiles based on outcome, age ≥70 varied from 50% to 65%, presence of vascular risk factors from 88% to 96%, intravenous fluids when indicated from 35% to 81%, and antihypertensive therapy when indicated from 60% to 85%. The largest part of variation in patient outcome between centers was explained by differences in patient characteristics
(Akaike's Information Criterion (AIC) = 134.0). Quality of care parameters explained a small part of the variation in patient outcome (AIC = 5.5).

Conclusions
Patient outcome after stroke varies largely between centers and is for a substantial part explained by differences in patient characteristics at time of hospital admission. Only a small part of the hospital variation in patient outcome is related to differences in quality of care. Unadjusted proportions of poor outcome after stroke are not valid as indicators of quality of care.

Introduction

Assessment of quality of care is becoming more and more important in medical practice. Donabedian has argued that quality in health care can be viewed as a function of three components: structure, process and outcome.[1] The first scientific forum of the American Heart Association and the American College of Cardiology on assessment of health care quality in cardiovascular disease and stroke has elaborated this framework for stroke care.[2] They proposed series of possible performance measures in the three different domains, based on the existing guidelines for stroke.

Despite this suggestion, quality of care in stroke is still often evaluated by use of outcome measures, usually (standardized) mortality rates on hospital level.[3]

Outcome assessment is generally easier than assessment of process measures, and it is often assumed that outcome measures reflect the relative importance of the different aspects of the care process, which makes them most relevant for patients. Those in favour of process indicators often express doubt about whether outcome really reflects quality of care since outcome largely depends on patient characteristics. Furthermore, flaws in care and well-performed care may cancel each other, and may not be reflected in overall outcome. This problem of outweighing good and bad performance becomes even larger in analyses on hospital level, as these also averages quality scores of individual patients.[3, 4]

Several studies have investigated the validity and feasibility of outcome data as indicators of quality of stroke care, with diverging results and conclusions, mostly due to limited information on the quality of care process.[5-8] For this study, data were derived from the Netherlands Stroke Survey, in which detailed data on both patient characteristics and process of care are available. Therefore this survey offers a unique opportunity to investigate whether there is a straightforward relationship between quality of care and outcome and whether outcome measures could be used to assess quality of care after stroke.

Methods

Study Population

The Netherlands Stroke survey was conducted in 10 centers in The Netherlands: 2 in the North, 4 in the Middle, and 4 in the Southern regions. The participating sites comprised 1 small (<400 beds), 4 intermediate (400 to 800 beds) and 5 large centers (>800 beds). Two centers were University hospitals. All centers had a neurology department, a neurologist with expertise in stroke, and a multidisciplinary stroke team. All but one hospital had a stroke unit, 8 were participating in a regional stroke service, and 9 were equipped for thrombolytic therapy. These institutions deliver care to approximately 10% of all acute stroke patients in The Netherlands, and their size and stroke expertise

can be considered representative of hospital-based stroke care in the Netherlands.[9]

All patients who were admitted to the neurology department with suspected acute stroke between October 2002 and May 2003 were screened. Patients were enrolled consecutively and prospectively if the initial diagnosis of first or recurrent acute brain ischemia was confirmed by the neurologist's assessment and if symptom onset was less then 6 months ago. All patients were admitted to the neurology department and were followed throughout their hospital stay. All patients or their proxies provided informed consent and the Medical Ethics Committees and Review Boards of the participating hospitals approved the study.

Data Collection

Trained research assistants collected data from the patients' hospital charts, within 5 days after discharge. At 1 year, survival status was obtained through the Civil Registries. In all survivors a telephonic interview was conducted based on a structured questionnaire, which was sent in advance. Follow up was complete in 96% of the patients. More details on the study population and methods of data collection can be found in an earlier publication on this survey.[9*]

Clinical characteristics and prognostic factors

Stroke subtype (brain infarction, hemorrhagic brain infarction, transient ischemic attack or amaurosis fugax) was defined by the treating neurologist based on clinical features and brain imaging (computed tomography (CT) or magnetic resonance imaging (MRI)) data. Previous stroke was defined present if ischemia of brain or eye or cerebral haemorrhage was noted in the medical history. Level of consciousness was assessed with the Glasgow Coma Scale[10], and disability in activities of daily living with the Barthel Index.[11] Atrial fibrillation and ischemic heart disease were marked if diagnosed by physical examination or if detected on ECG or if noted in the patient's medical history. Also peripheral vascular disease, diabetes mellitus, and hypertension were based on patient's medical history, or scored if diagnosed during hospitalization. Hyperlipidemia was defined present if total serum cholesterol exceeded 5 mmol/L, or if low-density lipoprotein exceeded 3.2 mmol/L or if hyperlipidemia was noted in the patient's medical history. The presence of carotid stenosis ≥70% was assessed by carotid imaging.

Quality of care

To measure quality, we distinguished between acute stroke treatment, sub-acute stroke care, and prevention. Quality of care parameters in acute stroke treatment involved the use of CT or MRI, electrocardiogram (ECG), appropriate laboratory tests, the administration of acetylsalicylic acid within 48 hours and thrombolytic therapy within 3 hours.

* This previous paper presented by mistake 11 centers (16 hospitals) instead of 10 centers (16 hospitals).

Sub-acute care included the administration of intravenous fluids, swallowing test, percutaneous endoscopic gastrostomy tube (PEG tube) insertion when indicated, early mobilisation, and early physiotherapy. Prevention included assessment of risk factors, measurement of serum cholesterol, carotid endarterectomy within 6 months, antiplatelet therapy, oral anticoagulants, antihypertensive therapy and cholesterol lowering therapy. These quality of care parameters and their indications were selected from national guidelines and most of them are also mentioned in the report from the American Heart Association/American College of Cardiology on assessment of healthcare quality in cardiovascular disease and stroke.[2] Each parameter was considered present in a certain patient when the diagnostic or therapeutic procedure was carried out and was indicated, or was not carried out and was not indicated. Otherwise, the indicator was considered absent. The quality of care parameters have been described more extensively in an earlier publication of this survey.[9]

Outcome measures

Poor outcome was defined as dead or disabled at 1 year, i.e. a score on the modified Rankin scale ≥ 3.[12] Additional outcome measures were dead or disabled at discharge, 30-day mortality and 1-year mortality.

Statistical analyses

To assess differences between centers in clinical characteristics, prognostic factors, quality of care parameters and outcome measures, centers were grouped in quartiles based on the percentage of patients dead or disabled after 1 year. These quartiles were fixed for all further analyses. P-values were derived from chi square tests for differences between the 10 centers.

We performed stepwise logistic regression analysis with backward elimination of predictors to construct prediction models for poor outcome. The selection criterion for inclusion was $P < 0.157$.[13] In step 1 only clinical characteristics (age, sex and duration of symptoms) were entered into the model. In step 2, patient-related prognostic factors were added: stroke severity, consciousness level at hospital arrival, Barthel Index at hospital arrival, previous stroke, atrial fibrillation, history of ischemic heart disease, peripheral vascular disease, diabetes mellitus, hypertension, hyperlipidemia, admission glucose $\geq$ 11 mmol/L, and independent pre-stroke living arrangement. In step 3, the mentioned quality of care parameters were added to the model. The main interest was not on the relationship of individual predictors with outcome but on the predictive strength of the different steps (clinical characteristics, other patient related factors and quality of care). The contribution of each step was expressed by Akaike's Information Criterion (AIC), which corresponds to the χ^2 of the step (or the difference in -2 log likelihood between the model with and without that step) minus 2 times the degrees of freedom.[13]

The discriminative ability of the model was expressed by the area under the receiver operating characteristic (ROC) curve. This area represents the probability that, within pairs of one patient with and one without the outcome, the patient with the higher prediction actually had the outcome.[14]

We calculated w scores to estimate the absolute differences in the number of patients with poor outcome between centers, before and after adjustment for clinical patient characteristics, prognostic factors, and quality of care parameters.[5] The w score of a hospital expresses the difference between the observed and predicted number with poor outcome per 100 patients and is calculated by the formula $[(o\text{-}p)/n]*100$, where o is the observed number of patients with poor outcome, p the predicted number of patients with poor outcome and n is the number of patients. For the unadjusted w scores, we derived p at each hospital by multiplying the number of patients (n) by the proportion patients with poor outcome in the total population. For the adjusted w scores, we derived p at each hospital by summing the individual predicted probabilities generated by the logistic regression models.[6] 95% CI's for the w scores were calculated using the method described by Parry et al.[15] The total variation between centers was also expressed as the percentage of patients with a different outcome then expected. We performed all analyses using SPSS 13.0 for Windows and Microsoft Excel.

Sensitivity analyses

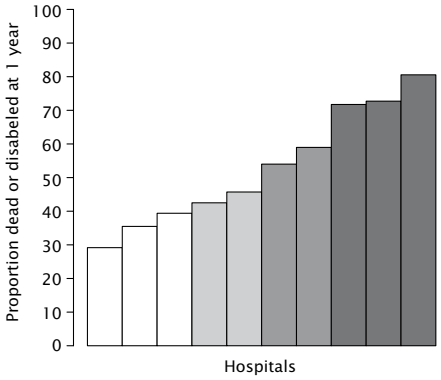
We repeated the logistic regression analysis for three alternative outcome measures: dead or disabled at discharge, 30-day mortality and 1 year mortality. Furthermore we repeated the logistic regression analysis by modeling the three steps (clinical characteristics, prognostic factors and quality of care parameters) in different orders.

Results

Outcome

The study population consisted of 579 patients who were admitted to the hospital because of stroke. Of all patients, 59 (10%) died during hospital stay. Of the remaining 520 patients, 206 (39%) were disabled at discharge. At 1 year, 143 patients (25%) were dead and 128 of the remaining 436 patients (29%) were disabled (modified Rankin scale 3, 4 or 5). So, the total number of patients with poor outcome at 1 year after stroke was 271 (47%). This percentage increased from 37% in hospital quartile 1 to 75% in hospital quartile 4. (Figure 12.1) For all outcome measures (mortality, disability and composite), both short term (discharge and 30 days) and 1-year, we observed the same trend across the hospital quartiles.

Figure 12.1 Outcome after ischemic stroke by hospital and quartile division (N=579)



Clinical characteristics and prognostic factors

Of all patients (n=579), 90% was admitted within 48 hours after symptom onset, and 95% within 1 week. Mean age was 70.4 (±13.2), 311 patients (54%) were male, the majority of patients (510, 88%) was diagnosed with brain infarction and 536 patients (93%) had one or more vascular risk factors. Regarding the symptoms of stroke, 13% of the patients had a lowered consciousness level and 89% were ADL (Activities of Daily Living) dependent at hospital admission (Table 12.1).

A number of differences in relative frequency of patient characteristics between the hospital quartiles was observed. Some were moderate, for example the presence of vascular risk factors, and some were large, for example age ≥70, lowered consciousness level and ADL dependency at hospital admission (Table 12.1).

Table 12.1 Variation in clinical characteristics and prognostic factors by hospital.

	Hospital quartiles based on patient outcome (% Rankin Scale ≥ 3 at 1 year)*					
	Total (N%)	1 (Lowest) (N%)	2 (N%)	3 (N%)	4 (Highest) (N%)	P Value (χ^2)†
Number of patients	579	179	127	101	172	
Number of centers	10	3	2	2	3	
Age ≥ 70	334 (58)	90 (50)	68 (54)	65 (64)	111 (65)	<0.001
Male gender	311 (54)	110 (62)	65 (51)	53 (53)	83 (48)	0.336
Vascular risk factors	536 (93)	131 (93)	121 (95)	97 (96)	152 (88)	0.049
Atrial fibrillation	99 (17)	21 (15)	18 (14)	21 (21)	33 (19)	0.619
Ischemic heart disease	116 (20)	31 (17)	24 (19)	20 (20)	41 (24)	0.430
Peripheral vascular disease	57 (10)	15 (8)	13 (10)	11 (11)	18 (11)	0.460
Diabetes Mellitus	119 (21)	36 (20)	26 (21)	17 (17)	40 (23)	0.198
Hypertension	346 (60)	132 (74)	65 (51)	62 (61)	87 (51)	<0.001
Hyperlipidemia	335 (58)	99 (53)	88 (69)	65 (64)	83 (48)	<0.001
Previous stroke / TIA	144 (25)	43 (24)	31 (24)	28 (28)	42 (24)	0.873
Independent prestroke living arrangement	513 (89)	163 (92)	112 (88)	89 (88)	149 (87)	0.493
Hospital arrival <48 hours after symptom onset	518 (90)	158 (88)	112 (88)	89 (88)	159 (92)	0.936
Stroke subtype						
Brain infarction	510 (88)	160 (90)	100 (79)	92 (91)	158 (92)	<0.001
TIA	60 (10)	17 (10)	25 (20)	7 (7)	11 (6)	
Amaurosis fugax	3 (1)	1 (1)	1 (1)	1 (1)	0 (0)	
Hemorrhagic infarction	6 (1)	1 (1)	1 (1)	1 (1)	3 (2)	
Severe stroke‡	92 (16)	27 (15)	15 (12)	21 (21)	29 (17)	0.309
Lowered consciousness level°	75 (13)	9 (5)	17 (13)	16 (16)	33 (19)	0.010
ADL independent° (Barthel Index=20)	119 (21)	54 (30)	37 (29)	15 (15)	13 (8)	<0.001
Incontinent°	169 (30)	48 (28)	28 (23)	29 (29)	64 (38)	0.123
Glucose ≥ 11 mmol/L	57 (10)	19 (11)	12 (10)	4 (4)	22 (14)	0.207

* Centers were divided into quartiles based on the percentage of patients that were dead or disabled (Rankin Scale ≥ 3) at 1 year;
† χ^2 for differences between 10 centers;
‡ Paresis of arm, leg and face, homonymous hemianopia and aphasia or other cortical function disorder;
° At hospital arrival.

Quality of care

The majority of the patients received the recommended diagnostic investigations and medical treatment in the acute phase, with the exception of thrombolytic therapy. Performance of a 12-lead ECG, provision of acetylsalicylic acid within 48 hours and thrombolytic therapy differed between the centers (P= 0.045, P<0.001 and P<0.001 respectively), performance of CT/MRI and laboratory tests did not (P=0.494 and P=0.624 respectively) (Table 12.2).

Table 12.2 Variation in acute management of ischemic stroke by hospital

	Hospital quartiles based on patient outcome (% Rankin Scale ≥ 3 at 1 year)*					
	Total (N%)	1 (Lowest) (N%)	2 (N%)	3 (N%)	4 (Highest) (N%)	P Value (χ^2)†
Number of patients	579	179	127	101	172	
Number of centers	10	3	2	2	3	
Diagnostic Investigations						
CT/MRI	567 (98)	178 (99)	124 (98)	99 (98)	166 (97)	0.494
12-lead ECG	555 (97)	166 (97)	120 (95)	100 (99)	169 (98)	0.045
Laboratory tests	564 (97)	175 (98)	124 (98)	100 (99)	165 (96)	0.624
Medical treatment						
Acetylsalicylic acid <48 hours	479 (83)	156 (87)	99 (78)	83 (82)	141 (82)	<0.001
In patients without OAC that arrived <48 hours‡	393/431 (91)	123/130 (95)	77/86 (90)	72/79 (91)	121/136 (89)	<0.001
Thrombolytic therapy	40 (7)	9 (5)	9 (7)	4 (4)	18 (11)	<0.001
Sub-acute care						
Intravenous fluids°§	198 (48)	48 (35)	56 (81)	30 (38)	64 (50)	<0.001
Swallowing test°	203 (40)	68 (43)	48 (48)	31 (34)	55 (35)	<0.001
PEG tube insertion°#	7 (21)	0 (0)	3 (25)	3 (38)	1 (25)	0.516
Mobilisation on day 1°	121 (24)	42 (26)	38 (38)	30 (33)	11 (7)	<0.001
Physiotherapy on day 1°	106 (21)	51 (32)	28 (28)	13 (14)	14 (9)	<0.001

* Centers were divided into quartiles based on the percentage of patients that were dead or disabled (Rankin Scale ≥ 3) at 1 year;
† χ^2 for differences between 10 centers;
‡ Oral anticoagulation;
° In patients with brain infarction;
§ In patients without parenteral feeding;
# In patients with swallow problems for more then 2 weeks.

Sub-acute care was less often performed in adherence to national guidelines. Of all 510 patients with a brain infarction 203 (40%) underwent a swallowing test, 121 (24%) were mobilised on the first day and 106 (21%) had physiotherapy during the first day. Of the 413 patients with brain infarction and no parenteral feeding, 198 (48%) received intravenous fluids. For all sub-acute process measures differences between centers were observed (P values <0.001), with the exception of PEG tube insertion.

Performance of secondary prevention varied also considerably between centers. The proportion of patients that underwent carotid imaging when indicated varied between 33% and 92% across the centers quartiles (P<0.001). Only 9 of 52 patients (17%) with carotid stenosis ≥70% underwent carotid endarterectomy within 6 months. The number of patients without atrial fibrillation that received antiplatelet therapy was high (93%), but there was still a significant difference between the centers (P<0.001). The proportion of patients that received oral anticoagulants and antihypertensive therapy when indicated also differed across centers (P=0.048 and P=0.029), while laboratory tests and cholesterol lowering therapy in patients with indication did not (P=0.304 and P=0.085 respectively) (Table 12.3).

Relation between clinical characteristics, prognostic factors, quality of care and outcome
Predictive factors in the model were age, sex, duration of symptoms, severe stroke, lowered consciousness level at hospital arrival, Barthel Index at hospital arrival, previous stroke, atrial fibrillation, ischemic heart disease, diabetes mellitus, hypertension, hyperlipidemia, ECG performed, mobilisation on day 1, antiplatelet therapy and oral anticoagulation.

Age, sex and duration of symptoms explained a large part of the variation (AIC=54.7, P<0.001) and another substantial part was explained by prognostic factors (AIC=79.3, P<0.001). Quality of care explained a relatively small part of the variation (AIC=5.5, P=0.009). The area under the curve of the model with only patient characteristics was 0.80 and that of the complete model 0.82, indicating a reasonable predictive performance (Table 12.4).

Table 12.3 Variation in secondary prevention after ischemic stroke by hospital

	Hospital quartiles based on patient outcome (% Rankin Scale ≥ 3 at 1 year)*					
	Total N(%)	1 (Lowest) (N%)	2 (N%)	3 (N%)	4 (Highest) (N%)	P Value (χ^2)†
Number of patients	579	179	127	101	172	
Number of centers	10	3	2	2	3	
Diagnostic Investigations						
Carotid Imaging	363 (63)	143 (80)	94 (74)	54 (54)	72 (42)	<0.001
In patients with indication‡	89/115 (77)	45/49 (92)	25/33 (76)	14/18 (78)	5/15 (33)	<0.001
Laboratory tests	560 (97)	174 (97)	123 (97)	82 (81)	163 (95)	0.304
Total cholesterol	430 (78)	135 (86)	114 (93)	5 (82)	99 (58)	<0.001
LDL cholesterol	323 (61)	101 (70)	104 (91)	22 (22)	96 (57)	<0.001
Glucose	545 (97)	167 (98)	119 (98)	99 (99)	160 (94)	<0.001
Treatment						
Carotid endarterectomy within 6 months	12 (2)	2 (1)	6 (5)	1 (1)	3 (2)	0.124
In patients with carotid stenosis ≥ 70%	9/52 (17)	2/20 (10)	3/13 (23)	1/10 (10)	3/9 (33)	0.012
Antiplatelet therapy	512 (88)	161 (90)	112 (88)	91 (90)	148 (86)	0.004
In patients without AF˚	448/480 (93)	146/152 (96)	101/109 (93)	76/80 (95)	125/139 (90)	<0.001
Oral anticoagulants	94 (16)	31 (18)	25 (20)	14 (14)	24 (14)	0.349
In patients with AF˚	59/99 (60)	19/27 (70)	13/18 (72)	10/21 (48)	17 (52)	0.048
Antihypertensive therapy	330 (57)	114 (65)	51 (42)	68 (67)	97 (57)	<0.001
In hypertensive patients	258/340 (76)	101/130 (78	37/62 (60)	47/62 (76)	73/86 (85)	0.029
Cholesterol lowering therapy	220 (39)	79 (44)	61 (50)	27 (27)	53 (31)	<0.001
In patients with indication§	134/187 (72)	46/59 (78)	43/53 (81)	16/33 (55)	27/42 (64)	0.085

* Centers were divided into quartiles based on the percentage of patients that were dead or disabled (Rankin Scale ≥ 3) at 1 year;
† χ^2 for differences between 10 centers;
‡ Barthel Index >18 and no brainstem or cerebellar symptoms or isolated hemianopia;
˚ Atrial fibrillation;
§ Hyperlipidemic patients <75 years (females) or <70 years (males) with a history of ischemic heart disease, carotid stenosis, peripheral vascular disease, or high cardiovascular risk profile.

Table 12.4 Multivariate analysis: Predictors of outcome (dead or disabled at 1 year) after ischemic stroke

		AIC (χ^2-2*df)*			
		Step	Model	P value	AUC†
Step 1:	Age, sex and duration of symptoms	54.68	54.68	<0.001	69.0
Step 2:	Stroke severity and risk factors	79.33	134.01	<0.001	80.4
Step 3:	Quality of care	5.46	139.47	0.009	81.5

* Akaike's Information Criterion;
† Area under the ROC curve.

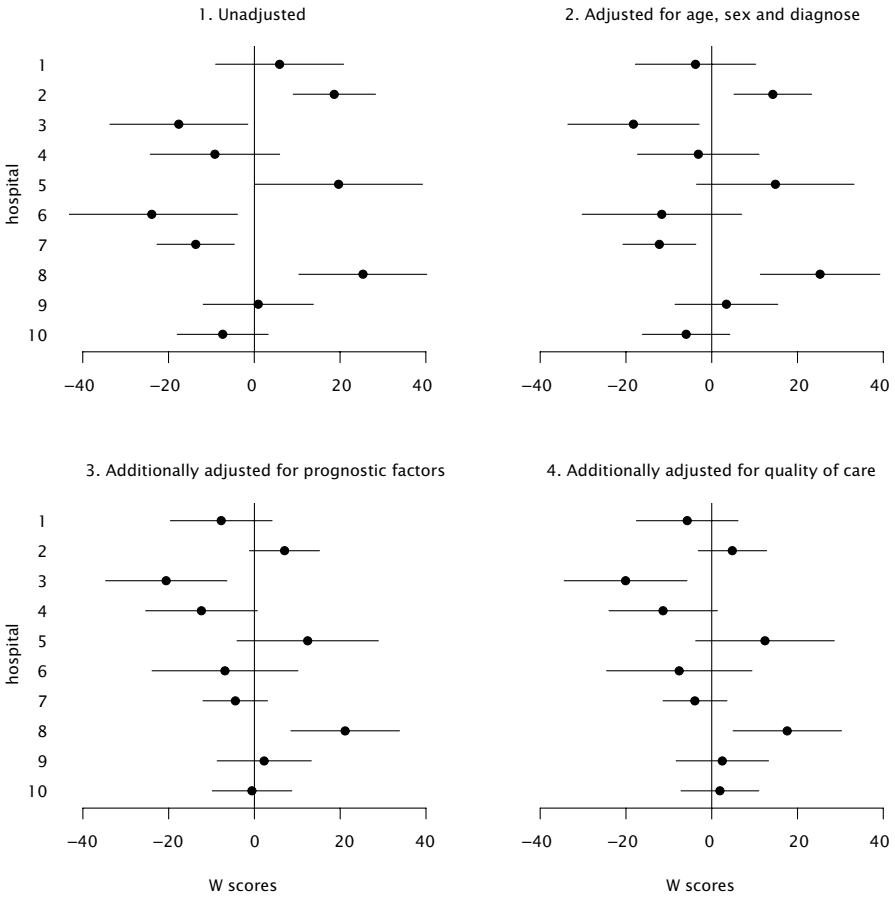
W scores

The differences in outcome between centers were also expressed in w scores. Before any adjustment the sum of the absolute w scores across the ten centers was 142, indicating that over all centers 14.2% of the patients had a different outcome (better or worse) then expected. After adjustment for age, sex and duration of symptoms this percentage was reduced to 11.2%. After further adjustment for prognostic factors the total variation declined further to 9.5%. After adjustment for quality of care, 8.8% of the patients still had a different outcome then expected, which could not be explained by the variables taken into account in this study (Figure 12.2).

Sensitivity analyses

Results were not affected by changing the dependent variable of the logistic regression into dead or disabled at discharge, 30-day mortality or 1-year mortality. The backward elimination of predictors resulted in slightly different predictors remaining in the model, but the different steps had approximately the same predictive power as in the initial model. Changing the order of the steps did not affect the results; the predictive strength of patient characteristics remained much larger than that of quality of care parameters.

Figure 12.2 Differences in observed number of patients with poor outcome and predicted number (W score) per hospital



Discussion

We explored the validity of patient outcome as an indicator of differences in quality of care between centers. We compared observed with expected outcome in 10 representative centers and we investigated whether differences in patient characteristics and quality of care could explain differences in outcome between the centers. We found that clinical characteristics and prognostic factors explain a relatively large part of the variation in outcome while quality of care parameters explain a much smaller part.

Also previous studies observed considerable variation between centers in outcome after stroke and were unable to explain this variation with differences in quality of care.[5-8, 16] The strength of our study is the detailed data on quality of care parameters that have been considered important in evidence based guidelines. Despite this we were also not able to demonstrate a clear and consistent relationship between quality of care and outcome on top of patient characteristics. Our sample size was small however and we may have had insufficient power to detect small effects.

Data were collected in 2002 and 2003. Improvements in the process of care might have lead to a stronger relationship with outcome. However, since the time of the study, no major changes in he care process have taken place.

An explanation for our results that evidence and consensus based measurements of the process of care appear to have such little impact on outcome, could be that treatment effects are generally modest. This is reflected in the fact that large RCTs are needed to identify a benefit of treatment. In quality of care studies, however, we are looking for differential use of established treatments and the resulting differences in outcome will therefore be even smaller. Also, not all items of care or treatments apply to all patients and so cannot be expected to have a large impact on aggregated outcomes made up of all patients.

It should be noted, that we defined poor patient outcome as Rankin scale ≥ 3 and one could question whether it is justified to put patients with Rankin scale 3 and patients who died into the same category. On the other hand, sensitivity analysis with 1-year mortality as outcome did not change the results. Furthermore, pre-stroke modified Rankin scale could have explained part of the variation in outcome but we could not adjust for it since it was not available in our dataset. We did however adjust for previous stroke and independent pre-stroke living arrangement, as a proxy for pre-stroke functional status.

Variation in stroke patients' outcome between centers was determined more by clinical characteristics and prognostic factors, than by hospital variation in quality of care. It would therefore make more sense to monitor the process of care directly in order to assess quality of care. There are several examples were this is being done, for example the RIKS stroke register from Sweden,[17] the National Sentinel Audit from England, Wales and Northern Ireland,[18] the Scottish Stroke Care Audit.[19] In the Netherlands however,

unadjusted 7-day day mortality rates were published on the internet until 2006.[20] A clear advantage of measuring process parameters instead of outcome is that it directly identifies opportunities for improvement in all hospitals, not only in those with poor outcome. This approach has successfully been applied in England and Wales in the form of regular national audits of stroke care using the Intercollegiate Stroke Audit Package.[21]

Those in favour of outcome assessment, advocate that quality assessment on process level requests a too detailed data collection, and conclusions on quality depend largely on the selection of process measures.[3] Our study shows, on the other hand, that using outcome assessment for quality measurement, is only valid after adjustment for patient characteristics. This approach is increasingly adopted, e.g. in the United Kingdom, where total hospital mortality rates are adjusted for some key patient characteristics[22] and in the United States where adjusted mortality rates after acute myocardial infarction and heart failure are used to compare centers.[23, 24] However, even if complete adjustment for patient characteristics is possible, this may not be sufficient for a for a meaningful comparison of outcomes between centers. Centers with patients with a good prognosis, small deficits and less co morbidity may still be more likely to deliver good quality of care compared to centers with more complex patients. The former ones have fewer patients with an indication for certain interventions, and hence they are less likely to withhold these interventions. For example, a patient without swallowing problems does not need a PEG tube, so it cannot be withheld unjustly. There are simply less opportunities to deliver substandard care. This implies that adjustment for patient characteristics may also be necessary when process measures are used.

Recently, attention is given to the development of prognostic models for outcome after stroke, which may be useful for quality assessment through proper outcome adjustment.[25-30] These models should be validated, however, in databases from the concerning country and they should be updated regularly. An important question is also whether a model is feasible in the sense that it can be fed by routinely collected data. Besides the discussion on which patient characteristics should be included in models for adjustment, also the feasibility and validity of different methodological and statistical approaches should be investigated and discussed.

Part of the variation in outcome remained unexplained. It might be so that we failed to measure important aspects of care e.g. how well complications are identified and treated. However, it seems implausible that such aspects of care are likely to have a huge impact on outcome. Another explanation is that there might be differences in patient characteristics that we cannot quantify. It remains unexplained how quite large residual variation in outcomes remains after adjusting for all known factors. More research is needed to clarify this phenomenon.

We conclude that patient outcome largely varies between centers and is for a substantial part explained by differences in patient characteristics at time of hospital

admission. Only a small part of the hospital variety in patient outcome is related to differences in quality of the care process. Unadjusted proportions of poor outcome after stroke are not valid as indicators of quality of care.

References

1. Donabedian A. The quality of care. How can it be assessed? *Jama.* 1988;260:1743-1748
2. Measuring and improving quality of care : A report from the american heart association/american college of cardiology first scientific forum on assessment of healthcare quality in cardiovascular disease and stroke. *Stroke*. 2000;31:1002-1012
3. Mant J. Process versus outcome indicators in the assessment of quality of health care. *Int J Qual Health Care*. 2001;13:475-480
4. Rubin HR, Pronovost P, Diette GB. The advantages and disadvantages of process-based measures of health care quality. *Int J Qual Health Care.* 2001;13:469-474
5. Weir N, Dennis MS. Towards a national system for monitoring the quality of hospital-based stroke services. *Stroke.* 2001;32:1415-1421
6. McNaughton H, McPherson K, Taylor W, Weatherall M. Relationship between process and outcome in stroke care. *Stroke*. 2003;34:713-717
7. Mohammed MA, Mant J, Bentham L, Raftery J. Comparing processes of stroke care in high- and low-mortality hospitals in the west midlands, uk. *Int J Qual Health Care*. 2005;17:31-36
8. Wolfe CD, Tilling K, Beech R, Rudd AG. Variations in case fatality and dependency from stroke in western and central europe. The european biomed study of stroke care group. *Stroke.* 1999;30:350-356
9. Scholte op Reimer WJ, Dippel DW, Franke CL, van Oostenbrugge RJ, de Jong G, Hoeks S, Simoons ML. Quality of hospital and outpatient care after stroke or transient ischemic attack: Insights from a stroke survey in the netherlands. *Stroke.* 2006;37:1844-1849
10. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. *Lancet.* 1974;2:81-84
11. Wade DT, Collin C. The barthel adl index: A standard measure of physical disability? *Int Disabil Stud.* 1988;10:64-67
12. Bonita R, Beaglehole R. Recovery of motor function after stroke. *Stroke.* 1988;19:1497-1500
13. Akaike H. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory.* 1973
14. Harrell FE, Jr. *Regression modeling strategies: With applications to linear models, logistic regression, and survival analysis*. New York: Springer; 2001.
15. Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO. Annual league tables of mortality in neonatal intensive care units: Longitudinal study. International neonatal network and the scottish neonatal consultants and nurses collaborative study group. *Bmj.* 1998;316:1931-1935
16. Rudd AG, Irwin P, Rutledge Z, Lowe D, Wade DT, Pearson M. Regional variations in stroke care in england, wales and northern ireland: Results from the national sentinel audit of stroke. Royal college of physicians intercollegiate stroke working party. *Clin Rehabil.* 2001;15:562-572
17. Asplund K, Hulter Asberg K, Norrving B, Stegmayr B, Terent A, Wester PO, Riks-Stroke C. Riks-stroke - a swedish national quality register for stroke care. *Cerebrovasc Dis.* 2003;15 Suppl 1:5-7

18. Clinical Effectiveness and Evaluation Unit Royal College of Physicians of London. National sentinel stroke audit, phase 1 (organisational audit) 2006, phase 2 (clinical audit) 2006. *http://www.healthcarecommission.org.uk/_db/_documents/Stroke_audit_Public_full_report__2006v2.pdf.* 2007
19. Dennis MF, R. McDowall, M. . National report on stroke services in scottish hospitals data relating to 2005/2006. *http://www.strokeaudit.scot.nhs.uk/Downloads/files/2007%20National%20Report.pdf.* 2007
20. *http://www.kiesbeter.nl/ziekenhuizen/Kwaliteit/Resultaat/.*
21. Rudd AG, Lowe D, Irwin P, Rutledge Z, Pearson M, Intercollegiate Stroke Working P. National stroke audit: A tool for change? *Qual Health Care.* 2001;10:141-151
22. Jarman B, Gault S, Alves B, Hider A, Dolan S, Cook A, Hurwitz B, Iezzoni LI. Explaining differences in english hospital death rates using routinely collected data. *Bmj.* 1999;318:1515-1520
23. Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, Roman S, Normand SL. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with an acute myocardial infarction. *Circulation.* 2006;113:1683-1692
24. Krumholz HM, Wang Y, Mattera JA, Wang Y, Han LF, Ingber MJ, Roman S, Normand SL. An administrative claims model suitable for profiling hospital performance based on 30-day mortality rates among patients with heart failure. *Circulation.* 2006;113:1693-1701
25. Weimar C, Ziegler A, Konig IR, Diener HC. Predicting functional outcome and survival after acute ischemic stroke. *J Neurol.* 2002;249:888-895
26. Tilling K, Sterne JA, Rudd AG, Glass TA, Wityk RJ, Wolfe CD. A new method for predicting recovery after stroke. *Stroke.* 2001;32:2867-2873
27. Johnston KC, Connors AF, Jr., Wagner DP, Haley EC, Jr. Predicting outcome in ischemic stroke: External validation of predictive risk models. *Stroke.* 2003;34:200-202
28. Counsell C, Dennis M, McDowall M, Warlow C. Predicting outcome after acute and subacute stroke: Development and validation of new prognostic models. *Stroke.* 2002;33:1041-1047
29. Counsell C, Dennis M, McDowall M. Predicting functional outcome in acute stroke: Comparison of a simple six variable model with other predictive systems and informal clinical prediction. *J Neurol Neurosurg Psychiatry.* 2004;75:401-405
30. Collaboration GSS. Predicting outcome after acute ischemic stroke: An external validation of prognostic models. *Neurology.* 2004;62:581-585

13 Effectiveness of statin treatment after a recent TIA or stroke in everyday clinical practice

Lingsma HF, Steyerberg EW, Scholte op Reimer WJM, van Domburg R, Dippel DWJ, and The Netherlands Stroke Survey investigators. Statin treatment after a recent TIA or stroke: is effectiveness shown in randomized clinical trials also observed in everyday clinical practice? *Acta Neurol Scand* 2010; 122:15–20.

Abstract

Aim and background
The benefit of statin treatment in patients with a previous ischemic stroke or TIA has been demonstrated in randomized clinical trials (RCT). However, the effectiveness in everyday clinical practice may be decreased because of a different patient population and less controlled setting. We aim to describe statin use in an unselected cohort of patients, identify factors related to statin use and test whether the effect of statins on recurrent vascular events and mortality observed in RCTs , is also observed in everyday clinical practise.

Methods
In 10 centers in the Netherlands, patients admitted to the hospital or visiting the outpatient clinic with a recent Transient Ischemic Attack (TIA) or ischemic stroke were prospectively and consecutively enrolled between October 2002 and May 2003. Statin use was determined at discharge and during follow up. We used logistic regression models to estimate the effect of statins on the occurrence of vascular events (stroke or myocardial infarction) and mortality within 3 years. We adjusted for confounders with a propensity score, that relates patient characteristics to the probability of using statins.

Results
Of the 751 patients in the study, 252 (34%) experienced a vascular event within 3 years. Age, elevated cholesterol levels and other cardiovascular risk factors were associated with statin use at discharge. After 3 years, 109 of 280 (39%) of the users at discharge had stopped using statins. Propensity score adjusted analyses a beneficial effect of statins on the occurrence of the primary outcome (OR 0.8, 95%CI: 0.6–1.2).

Conclusion
In our study we found poor treatment adherence to statins. Nevertheless, after adjustment for the differences between statin users and non-statin users, the observed beneficial effect of statins on the occurrence of vascular events within 3 years, although not statistically significant, is compatible with the effect observed in clinical trials.

Introduction

The benefit of the use of 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitors (statins) in patients with coronary heart disease and in those with an increased risk for cardiovascular disease is already firmly established. Statin use reduces mortality and the risk of strokes and cardiovascular events.[1-5] Until recently, limited data were available on the impact of statin therapy in patients with symptomatic cerebrovascular disease but no known coronary heart disease.[6,7] However, the Heart Protection Study (HPS) and the Stroke Prevention by Aggressive Reduction in Cholesterol Levels (SPARCL) trial addressed the clinical question whether patients with stroke or Transient Ischemic Attack (TIA) without coronary heart disease would benefit from treatment with a statin.[8,9] These trials revealed that statin treatment is beneficial in patients with a recent stroke or TIA or with pre-exsiting cerebrovascular disease without known coronary disease.

An important question however, is whether the results of the trials can also be observed in everyday clinical practice.[10] This is not self-evident, for a number of reasons. First the patient population in clinical trials is different from the population in daily practice. Second a trial takes place in a controlled setting with strict control of therapy adherence while in every day clinical practice, reported adherence rates for statins vary between 25% and 40% after 2 years and 26% after 5 years.[11,12]

It is difficult to test the effectiveness of a therapy in daily practice because patients are not randomized. Patient with and without the treatment may be different in many respects and this can cause confounding. Therefore, in an observational study, the effect of a therapy can only be tested after adjustment for confounding, e.g. by a propensity score.[13]

The aim of our study is to describe statin use, to identify factors related to statin use and to assess the effect of statins on outcome in an unselected cohort of patients, in order to find out whether the effect observed in randomized clinical trials holds true in circumstances that reflect daily practice.

Methods

Study Population

The Netherlands Stroke survey was conducted in 10 hospitals in The Netherlands: 2 in the North, 4 in the Middle, and 4 in the Southern regions. The participating sites comprised 1 small (<400 beds), 4 intermediate (400 to 800 beds) and 5 large centers (>800 beds). Two centers were University hospitals and all centers had a neurology department. All but one hospital had a stroke unit. These institutions deliver care to approximately 10% of all acute stroke patients in The Netherlands, and their size and stroke expertise can be considered representative of hospital-based stroke care in the Netherlands.[14]

All patients who were admitted to the neurology department or visited the neurological outpatient clinic with suspected ischemia of the brain between October 2002 and May 2003 were screened. Screening per hospital could be discontinued when at least 30 admitted patients and 30 outpatients were enrolled. Patients with ischemic stroke or TIA and symptom onset within the last 6 months were enrolled consecutively and prospectively if the initial diagnosis of first or recurrent ischemia of the brain or eye was confirmed by the neurologist's assessment. All patients or their proxies provided informed consent. The Medical Ethics Committees of the participating hospitals approved the study. We excluded patients with missing outcome data and patients who died during hospital admission.

Data Collection

Trained research assistants collected data from the patients' hospital charts, within 5 days after discharge. At 1 year, survival status was obtained through the Civil Registries. In all survivors or their relatives a telephonic interview was conducted based on a structured questionnaire, which was sent in advance. At 3 years the procedure was repeated. More details on the study population and methods of data collection can be found in earlier publications on this survey.[15,16]

Measures

The diagnosis of ischemic stroke or TIA by the treating neurologist was based on clinical features and brain imaging data (computed tomography (CT) or magnetic resonance investigation (MRI). Previous stroke was defined present if cerebral ischemia of or cerebral haemorrhage was noted in the medical history. Disability in activities of daily living (ADL) was measured with the Barthel Index.[17] Atrial fibrillation and ischemic heart disease were marked if diagnosed by physical examination, if detected on 12-lead electrocardiogram (ECG) or if noted in the patient's medical history. Also peripheral vascular disease, diabetes mellitus, and chronic hypertension were based on patient's medical history or scored if diagnosed during hospitalization. Hyperlipidemia was

defined present if total serum cholesterol exceeded 5 mmol/L, or if low-density lipoprotein exceeded 3.2 mmol/L or if hyperlipidemia was noted in the patient's medical history.

Statin use was defined as statin use at discharge, as noted in the hospital chart. It was also assessed pre-stroke from the medical history, and at 1 and 3 years through the follow-up interviews.

The primary outcome in this study was the occurrence of death, non-fatal stroke or myocardial infarction (MI). Secondary outcome was mortality alone. Both outcomes were assessed after 1 year and after 3 years follow up. Mortality was obtained through the Civil Registries, non-fatal stroke and MI were assessed through the follow-up interviews.

Statistical analysis

Continuous variables were compared by Student's t-tests and categorical variables by chi-square tests. Logistic regression analysis was performed to study the effect of statin use on outcome.

We used a propensity score to adjust for potential confounders.[9] The propensity score of a patient is the probability of that patient being treated with statins at discharge given baseline characteristics (confounders). To calculate the propensity score we used a logistic regression model with statin use yes/no as outcome and potential confounders as explaining variables. Variables included in the model were: age, sex, hospital, admission, previous stroke, stroke subtype, stroke severity, lowered consciousness level at admission, current smoking, body mass index>25, peripheral vascular disease, ischemic heart disease, diabetes mellitus, hypertension, hyperlipidemia, cholesterol measured during hospital admission, LDL cholesterol>3.2 during hospital admission, total cholesterol>5 during hospital admission, pregnant or breastfeeding, severe renal dysfunction, atrial fibrillation, prosthetic heart valves and ADL dependency at discharge. We subsequently added the propensity score to the initial logistic regression model. This adjusts for confounding by conditioning on the individual probability of being treated with statins.

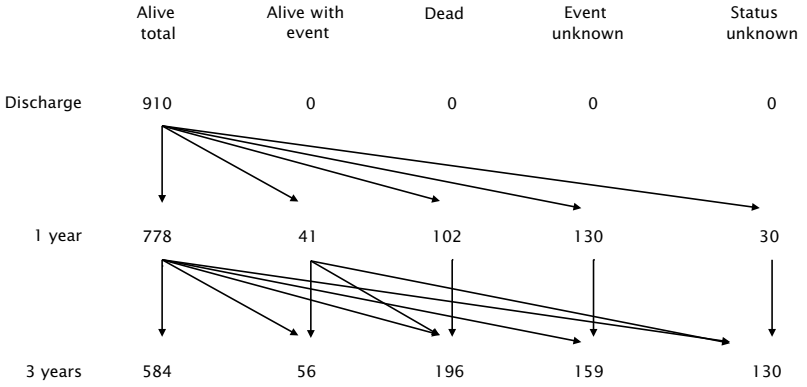
We performed all analyses using SPSS 15.0 for Windows. A two-sided p-value of $p<0.05$ was considered statistically significant.

Results

Study population

In total 972 patients who were admitted to hospital or visited outpatient clinic because of stroke were included in the study. Of these patients 62 (6%) were excluded because they died in the hospital, so the study population consisted of 910 patients (Figure 13.1). Regarding mortality, 1 year follow up was complete in 880 patients, 3 year follow up was complete in 780 patients. With regard to vascular events, 1 year follow-up was complete in 880 patients, 3 year follow-up was complete in 751 patients. We observed no significant differences between patients with and without missing outcome data with regard to baseline characteristics and statin use at discharge.

Figure 13.1 Status of the study population.



Outcome

At 1 year, 102 patients (12%) had died and 41 (5%) had a new stroke or MI. The total number of patients with the composite primary outcome event within 1 year amounted to 143 (18%). After 3 years 196 patients (25%) had died and 56 (7%) had suffered from new stroke or MI. So the total number of patients with the primary outcome event within 3 years was 252 (34%).

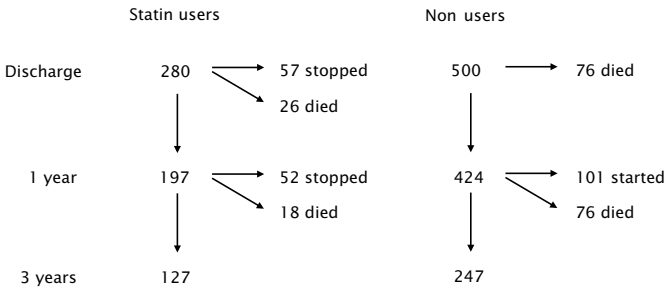
Patient characteristics

Patients with statins at discharge were on average younger (mean age 65 vs 71, $p<0.01$), more often male (62% vs. 53%, $p<0.01$), had more ischemic heart disease (24% vs. 16%, $p<0.01$), hypertension (66% vs. 54%, $p<0.01$) and hyperlipedimia (91% vs. 48%, $p<0.01$), had less severe strokes (4 vs. 7% with lowered consciousness level, $p=0.02$), were more often smoking (32% vs. 23%, $p<0.01$) and overweight (18% vs. 8%, $p<0.01$) and had less often atrial fibrillation (7% vs. 15%, $p<0.01$) (Table 13.1).

Tabel 13.1 Patient characteristics in statin users and non users.

N (%)	Total 910 (100%)	Statins 354 (39%)	No statins 556 (61%)	P value
Age (mean)	68	65	71	<0.01
Male gender	514 (57%)	219 (62%)	295 (53%)	<0.01
Hospital admission	515 (57%)	208 (59%)	307 (55%)	0.29
Brain infarction	569 (63%)	225 (64%)	344 (62%)	0.61
Previous stroke	197 (22%)	78 (22%)	119 (21%)	0.82
Ischemic heart disease	172 (19%)	86 (24%)	86 (16%)	<0.01
Peripheral vascular disease	93 (10%)	43 (12%)	50 (9%)	0.13
Diabetes	158 (18%)	71 (20%)	87 (16%)	0.09
Hypertension	534 (59%)	233 (66%)	301 (54%)	<0.01
Hyperlipidemia	587 (65%)	322 (91%)	265 (48%)	<0.01
Current smoking	238 (26%)	112 (32%)	126 (23%)	<0.01
Lowered consciousness level	52 (6%)	12 (4%)	40 (7%)	0.02
Severe stroke	80 (9%)	27 (8%)	53 (10%)	0.32
Body Mass Index>25	103 (11%)	62 (18%)	41 (8%)	<0.01
Atrial fibrillation	105 (12%)	23 (7%)	82 (15%)	<0.01
1 year mortality (n=880)	102 (11%)	26 (8%)	76 (14%)	<0.01
1 year mortality, non-fatal stroke or myocardial infarction (n=880)	143 (16%)	43 (13%)	100 (19%)	0.02
3 year mortality (n=780)	196 (25%)	56 (20%)	140 (28%)	0.01
3 year mortality, non-fatal stroke or myocardial infarction (n=751)	252 (34%)	83 (27%)	169 (38%)	<0.01

Figure 13.2 Statin use in the study population.



Statin use

Many patients stopped using statins during follow up. Figure 13.2 shows 280 patients who were using statins at discharge 500 patients who were not. Note that we show only the 780 patients with complete follow up in this figure. During the first year, 57 patients stopped using statins and in the second and third year, again 52 patients

stopped. So after 3 years 109 of 280 (39%) of the users at discharge had stopped using statins. There were also 101 patients who started using statins during follow up.

Propensity score

Table 13.2 shows that older patients had a lower probability of using statins at discharge (OR: 0.7 per 10 years, 95%CI: 0.6 to 0.8). Other important factors were hyperlipidemia (OR= 6.6, 95%CI = 4.7–9.4), admission to the hospital (OR=1.7, 95%CI = 1.1–2.5), ischemic heart disease (OR=1.7, 95%CI = 1.1–2.6), hypertension (OR=1.5, 95%CI = 1.1–2.1), body mass index>25 (OR=1.9, 95%CI = 1.1–3.1) and LDL cholesterol>3.2 in the hospital (OR=1.9, 95%CI = 1.3–2.9). These factors gave an increased probability of using statins at discharge.

Tabel 13.2 Propensity score, odds ration (OR) for having statins at discharge (N=910).

	OR	95% CI
Age per 10 year increase	0.7	0.6 to 0.8
Sex	1.3	0.9 to 1.8
Hospital admission	1.7	1.1 to 2.5
Acute ischemic stroke	1.1	0.7 to 1.7
Previous stroke	1.1	0.7 to 1.6
Peripheral vascular disease	1.6	1.0 to 2.8
Ischemic heart disease	1.7	1.1 to 2.6
Diabetes	1.5	1.0 to 2.3
Hypertension	1.5	1.1 to 2.1
Current smoking	1.1	0.8 to 1.6
ADL dependent at discharge	0.6	0.3 to 1.4
Lowered consciousness level	0.5	0.2 to 1.0
Severe stroke	0.8	0.4 to 1.4
Body Mass Index>25	1.9	1.1 to 3.1
Cholesterol measured during hospital stay	1.0	0.6 to 1.6
Total cholesterol above 5 mmol/l in hospital	1.1	0.7 to 1.7
LDL cholesterol above 3.2 mmol/l in hospital	1.9	1.3 to 2.9
Hyperlipidemia	6.6	4.7 to 9.4
Pregnant or breastfeeding	0.2	0.0 to 1.1
Atrial fibrillation	0.8	0.4 to 1.6
Prosthetic heart valves	0.7	0.2 to 2.1
Severe renal dysfunction	0.6	0.2 to 1.7

Relationship between statins and outcome

In the unadjusted analysis, statin use had a beneficial effect on all outcomes studied, with significant odds ratios around 0.6. Propensity score adjusted analyses also showed a beneficial effect of statins on the primary outcome, however less pronounced, and not significant (OR=0.8, 95%CI = 0.6–1.2, Table 13.3).

Tabel 13.3 Unadjusted and adjusted effect of statins on outcome.

	OR	95% CI
1 year mortality (n=880)		
Unadjusted	0.5	0.3 to 0.8
Adjusted	0.9	0.5 to 1.5
1 year mortality, non-fatal stroke or myocardial infarction (n=880)		
Unadjusted	0.6	0.4 to 0.9
Adjusted	0.9	0.6 to 1.4
3 year mortality (n=780)		
Unadjusted	0.6	0.5 to 0.9
Adjusted	1.1	0.7 to 1.7
3 year mortality, non-fatal stroke or myocardial infarction (n=751)		
Unadjusted	0.6	0.4 to 0.8
Adjusted	0.8	0.6 to 1.2

Discussion

In this study we observed a beneficial effect of statins on the occurrence of vascular events within 3 years compatible with the effect found in clinical trials, despite poor treatment adherence.

The number of patients that started statins after stroke was low (39%), but the national guidelines in the Netherlands recommended statins at the time of data collection only for patients who had a high risk of cardiovascular events and a serum cholesterol exceeding 5.0 mmol/L, or a previous MI. Statin use was indeed associated with age, hospital admission, elevated cholesterol levels and other cardiovascular risk factors. The association with elevated cholesterol and cardiovascular risk factors can be expected, since the guidelines specifically target high risk patients. With regard to age, Simpson et al also found that older patients were less likely to receive statins after a stroke.[18] They also reported that females were less likely to receive statins but after adjustment for other patient characteristics this was not the observed in our study. Our results again suggest that older patients need to be targeted for secondary prevention therapy with statins, since there is no reason to assume that they will benefit less from therapy.[19]

Patients admitted to the hospital got more often statins than patients visiting the outpatient clinic and reasons for withholding were unclear in the majority of outpatients.[16] It is however important to get more insight in the reasons for differing management in age groups and between settings. Also since Ovbiagele et al. showed that unless statin therapy is started at or before hospital discharge, secondary prevention is poor.[20]

The low adherence rate we found is in line with previous studies.[11,12] Also the SPARCL trial and HPS study reported varying compliance to treatment, although to a lesser extent than in our study.[8,9] It is of paramount importance to study ways to improve adherence so that a higher effectiveness of statins in everyday clinical practice can be obtained. Unfortunately we did not have any information on why patients stopped using statins. It is however important to get insight in the reasons for discontinuing treatment (costs, side effects, other treatment, normalization of cholesterol levels) to distinguish 'good' reasons from 'bad' reasons. Besides, as different medications may have different adherence rates, the real dividend in clinical practice can vary and this may even influence funding.

The propensity score adjusted odds ratio (OR) of 0.8 for the effect of statins on the occurrence of vascular events and mortality within 3 years is compatible with the estimated hazard ratio (HR) of 0.75 to 0.8 in RCTs.[8,9] Although with the same underlying treatment effect, estimated ORs will be different from HRs and the RCTs had a mean follow-up of around 5 years. Our results are also compatible with a systematic review by Law et al.[21] They pooled a large number of trials and observational studies

and calculated both the effect of LDL cholesterol reduction on stroke and the effect of statins on LDL cholesterol. Combining these two, they found an odds ratio of 0.85 for the effect of statin use on the occurrence of stroke.

Logistic regression was performed where Cox regression might have been preferable, but in some cases we did not have the exact time of recurrent events. We did perform Cox regression with mortality as outcome and found hazard ratios very similar to the odds ratios from the logistic regression with 3 year mortality as outcome.

In our study we found only a very small effect of statin use at 1 year. This finding is in line with findings from RCTs, which also found no effect at 1 year.[8,9,22] Our results again provide evidence that statins are becoming more effective after long term use; this stresses even more the need to increase adherence rates.

We analyzed our data according the intention-to-treat approach that is also followed in the majority of randomised clinical trials and is also advised in the CONSORT guidelines.[23] To take into account treatment change during follow up ('starters and stoppers') it is also suggested to use 'per protocol' analysis, which only includes patients who complete study therapy in the final analysis.[24] Another possibility is to calculate a 'compliance factor' based on the percentage starters and stoppers and apply this factor to the odds ratio. This approach was demonstrated in the HPS, were an adherence rate of 85% was observed.

Besides lower treatment adherence, effectiveness can be influenced by different patient populations in RCTs and clinical practice. Although SPARCL and HPS included relatively broad patient selections, we observed some differences between the trials with our patients, who were for example older, more often smoking, and had more often hypertension. But we did not study this phenomenon in detail.

We adjusted for confounding with the propensity score. We included an extensive number of patient characteristics in the propensity score, but there might be more unknown or unmeasured confounders that we were unable to adjust for.

Other possible limitations of the study are that the vascular events were patient-reported, we did not consider the dose or the type of statins and we had missing outcomes.

In conclusion, this study showed the effectiveness of statins in patients with a recent stroke or TIA in everyday clinical practice. To further increase effectiveness, specific subgroups (older patients, outpatients) should be targeted and ways to improve adherence should be studied.

References

1. Randomised trial of cholesterol lowering in 4444 patients with coronary heart disease: the Scandinavian Simvastatin Survival Study (4S). *Lancet* 1994;344:1383-1389
2. Prevention of cardiovascular events and death with pravastatin in patients with coronary heart disease and a broad range of initial cholesterol levels. The Long-Term Intervention with Pravastatin in Ischaemic Disease (LIPID) Study Group. *N Engl J Med* 1998; 339:1349-1357
3. Colhoun HM, Betteridge DJ, Durrington PN, Hitman GA, Neil HA, Livingstone SJ, Thomason MJ, Mackness MI, Charlton-Menys V, Fuller JH. Primary prevention of cardiovascular disease with atorvastatin in type 2 diabetes in the Collaborative Atorvastatin Diabetes Study (CARDS): multicenter randomised placebo-controlled trial. *Lancet* 2004; 364:685-696
4. Plehn JF, Davis BR, Sacks FM, Rouleau JL, Pfeffer MA, Bernstein V, Cuddy TE, Moye LA, Piller LB, Rutherford J, Simpson LM, Braunwald E. Reduction of stroke incidence after myocardial infarction with pravastatin: the Cholesterol and Recurrent Events (CARE) study. The Care Investigators. *Circulation* 1999; 99:216-223
5. Sever PS, Dahlof B, Poulter NR, Wedel H, Beevers G, Caulfield M, Collins R, Kjeldsen SE, Kristinsson A, McInnes GT, Mehlsen J, Nieminen M, O'Brien E, Ostergren J. Prevention of coronary and stroke events with atorvastatin in hypertensive patients who have average or lower-than-average cholesterol concentrations, in the Anglo-Scandinavian Cardiac Outcomes Trial--Lipid Lowering Arm (ASCOT-LLA): a multicenter randomised controlled trial. *Lancet* 2003; 361:1149-1158
6. Goldstein LB, Amarenco P, Bogousslavsky J, Callahan AS, Hennerici MG, Welch KM, Zivin J, Sillesen H. Statins for secondary stroke prevention in patients without known coronary heart disease: the jury is still out. *Cerebrovasc Dis* 2004; 18:1-2
7. Amarenco P, Labreuche J, Lavallee P, Touboul PJ. Statins in stroke prevention and carotid atherosclerosis: systematic review and up-to-date meta-analysis. *Stroke* 2004; 35:2902-2909
8. Heart Protection Study Collaborative. MRC/BHF Heart Protection Study of cholesterol lowering with simvastatin in 20,536 high-risk individuals: a randomised placebo-controlled trial. *Lancet* 2002; 360:7-22
9. Amarenco P, Bogousslavsky J, Callahan A, 3rd, Goldstein LB, Hennerici M, Rudolph AE, Sillesen H, Simunovic L, Szarek M, Welch KM, Zivin JA (Stroke Prevention by Aggressive Reduction in Cholesterol Levels). High-dose atorvastatin after stroke or transient ischemic attack. *N Engl J Med* 2006; 355:549-559
10. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet* 2005; 365:82-93
11. Benner JS, Glynn RJ, Mogun H, Neumann PJ, Weinstein MC, Avorn J. Long-term persistence in use of statin therapy in elderly patients. *JAMA* 2002; 288:455-461
12. Jackevicius CA, Mamdani M, Tu JV. Adherence with statin therapy in elderly patients with and without acute coronary syndromes. *JAMA* 2002; 288:462-467
13. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; 17:2265-2281

14. Verschoor H, Stolker DHCM, Franke CL. *Stroke Services anno 2003*. Netherlands Heart Foundation, The Hague 2004.
15. Lingsma HF, Dippel DW, Hoeks SE, Steyerberg EW, Franke CL, van Oostenbrugge RJ, de Jong G, Simoons ML, Scholte Op Reimer WJ. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. *J Neurol Neurosurg Psychiatry* 2008; 79:888-894
16. Scholte op Reimer WJ, Dippel DW, Franke CL, van Oostenbrugge RJ, de Jong G, Hoeks S, Simoons ML. Quality of hospital and outpatient care after stroke or transient ischemic attack: insights from a stroke survey in the Netherlands. *Stroke* 2006; 37:1844-1849
17. Wade DT, Collin C. The Barthel ADL Index: a standard measure of physical disability? *Int Disabil Stud* 1988; 10:64-67
18. Simpson CR, Wilson C, Hannaford PC, Williams D. Evidence for age and sex differences in the secondary prevention of stroke in Scottish primary care. *Stroke* 2005; 36:1771-1775
19. CBO *Conceptrichtlijn Diagnostiek, behandeling en zorg voor patiënten met een beroerte*. 2008.
20. Ovbiagele B, Saver JL, Fredieu A, Suzuki S, McNair N, Dandekar A, Razinia T, Kidwell CS. PROTECT: a coordinated stroke treatment program to prevent recurrent thromboembolic events. *Neurology* 2004; 63:1217-1222
21. Law MR, Wald NJ, Rudnicka AR. Quantifying effect of statins on low density lipoprotein cholesterol, ischaemic heart disease, and stroke: systematic review and meta-analysis. *BMJ* 2003; 326:1423
22. Collins R, Armitage J, Parish S, Sleight P, Peto R (Heart Protection Study Collaborative) Effects of cholesterol-lowering with simvastatin on stroke and other major vascular events in 20536 people with cerebrovascular disease or other high-risk conditions. *Lancet* 2004; 363:757-767
23. Moher D, Schulz KF, Altman DG. The CONSORT statement: revised recommendations for improving the quality of reports of parallel-group randomised trials. *Lancet* 2001; 357:1191-1194
24. Schoenfeld PS. Evidence-based medicine in practice: applying intention-to-treat analysis and perprotocol analysis. *Am J Gastroenterol* 2005; 100:3-4

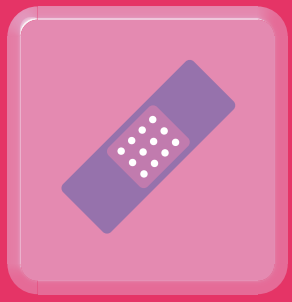


V Discussion









14 General discussion

General discussion

The aim of this thesis was to study methods to measure quality of care with outcome measures, and to apply these methods to different acute neurological diseases. Three specific questions were addressed:

1. What is the role of statistical uncertainty in measuring quality of care with outcome measures?

1a. How large is the effect of statistical uncertainty on between-hospital comparisons?

1b. How should statistical uncertainty be incorporated in outcome measures?

2. What is the role of case-mix variation in measuring quality of care with outcome measures?

2a. How large is the effect of case-mix on between-hospital comparisons?

2b. How can case-mix variation be captured for between-hospital comparisons?

3. How do outcome measures relate to processes of care?

First we found that statistical uncertainty is often large when measuring quality of care with outcome measures. Different methods were proposed to account for statistical uncertainty and to quantify statistical uncertainty. Traditional integer rankings disregard statistical uncertainty and should be avoided. Rankings should incorporate statistical uncertainty, e.g. using the 'expected rank'.

Second, case-mix often varies between hospitals and has a large effect on between-hospital comparisons. Case-mix adjustment is crucial in measuring quality of care with outcome measures. We developed several prognostic models that could be used for case-mix adjustment.

Thirdly, we found that process measures and outcome in acute neurological diseases were only moderately related.

We conclude that outcome measures for quality of care should be case-mix adjusted random effect estimates, which are related to processes of care.

In this final chapter the results and interpretation are discussed and summarized per research question. And we discuss the implications of the results for policy and research.

Statistical uncertainty

When comparing hospitals based on outcome, some variation always exists just by chance, because of statistical uncertainty. The first research question addressed the role of statistical uncertainty.

Random effects models

Variation between hospitals in binary outcomes is traditionally modelled in a fixed effects logistic regression model. In such a model the estimates are based solely on the observed outcome and statistical uncertainty is ignored. In chapter 2 we presented an alternative approach which does account for statistical uncertainty; random effects regression models.[1-3] In a random effects model the estimates for each hospital are based on its observed outcome, the uncertainty, and the average outcome.

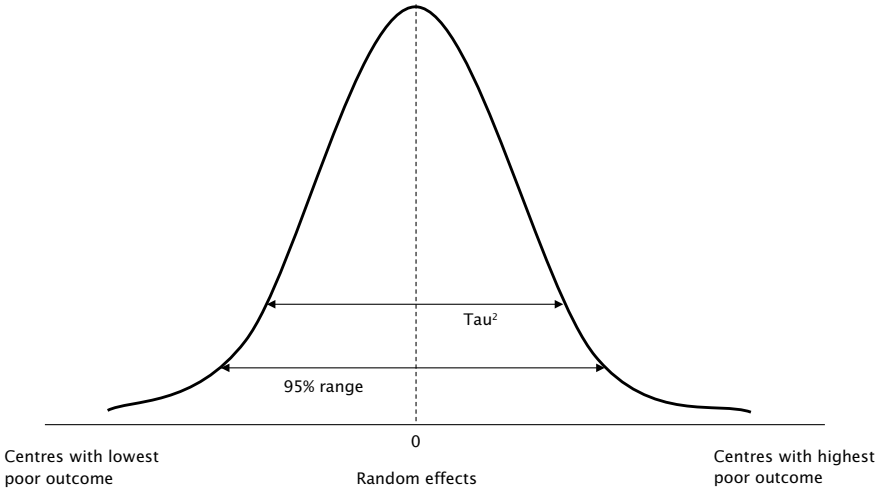
The amount of statistical uncertainty is mainly determined by the number of outcomes per hospital, which is determined by the number of patients and the frequency of the outcome. When the number of outcomes per hospital is small, the estimates are very uncertain. Fixed effect estimates disregard this uncertainty, which leads to overinterpretation of the differences in outcome. Particularly smaller hospitals can have an extreme outcome caused rather by chance than by quality of care. On the other hand, it is unlikely that a small hospital will be identified as a good or bad performer when using random effects models. When the numbers of patient per hospital are large, the estimates from the random and the fixed effects models will be similar.

The overall between-hospital differences are also more conservatively estimated with random effects models. The estimator of between-hospital differences is the variance of the random effects, and is labelled tau^2. Tau^2 can be interpreted as the between-hospital differences beyond chance. Because it is difficult to interpret, we proposed in chapter 10 to express tau^2 in a 95% range of odds ratios for the odds of poor outcome in each hospital, compared to the average odds of poor outcome. The 95% OR range is an attractive way to express the magnitude of between-hospital differences, beyond random variation (Figure 14.1). It can also be translated in an X fold difference in outcome between the hospitals on the lower end of the distribution and the hospitals on the higher end, were we propose to use the 2.5^{th} and 97.5^{th} percentiles of the distribution.

Others propose to express tau^2 in an odds ratio for the odds of poor outcome if treated at a hospital 1 standard deviation above the average relative to that if treated at a hospital 1 standard deviation below the average.[4] This corresponds to the 34th versus 66th percentiles. Another possibility is to present the overall range, the interquartile range, or top and bottom deciles of the random effects, or the proportions of poor outcome estimated by the random effects model.[5]

A limitation of our work is that we have not explored the uncertainty of tau^2. A standard deviation of tau^2 is however estimated in the model. A confidence interval could also be created with bootstrapping.

Figure 14.1 Schematic representation of variance of the random effects (tau^2) and the 95% range of odds ratios



In chapter 4 it was shown that logistic random effects model can nowadays be fit with all commonly available statistical available software packages. Even with many small hospitals in the dataset, the results were very stable over the different packages. Only the additional features and the usability differed between the packages. We therefore do not consider technical or computational difficulties a limitation for using random effect models.

Rankability

The concept of rankability quantifies the amount of uncertainty in relation to the magnitude of the differences in outcome (chapter 2). We defined it as the percentage of the observed differences between hospitals not due to statistical uncertainty. When the differences are large but the uncertainty too, rankability will be low. Only the combination of large differences and limited uncertainty gives a high rankability, which indicates that the hospitals can be distinguished from each other in terms of outcome. When rankability is low, it is meaningless to compare hospitals based on outcome; the differences will only represent random variation. Attempts to compare hospitals based on outcome should present the rankability. In chapter 3 and 12 it was shown that rankability is actually low in many practical examples, typically around 50%. The rankability can also be seen as the ratio between signal and (statistical) noise. E.g. a rankability of 50% then indicates 50% signal and 50% noise.

Expected Rank

A particular form of comparing hospital based on outcome is ranking them. Rankings disregard the magnitude of the relative differences between the hospitals. One hospital has to be first and one has to be last, even if the differences are small. Similarly, rankings disregard variation that exists by chance, also if the between-hospital differences are estimated with a random effects model.

In chapter 2 we followed van Houwelingen and others with a method to incorporate both the magnitude and the uncertainty in rankings: the expected rank (ER). ERs have theoretically the same range as the integer rakings (1 to the number of hospitals), but in practice the ERs will be shrunken towards the median rank. The amount of shrinkage is dependent on the number of patients and outcome events (which determines the uncertainty) and magnitude of the differences, and directly relates to the rankability. A high rankability is associated with little shrinkage; a low rankability is associated with large shrinkage.

We can scale the expected ranks ER between 0 and 100% with percentiles based on expected rank (PCER) for easy interpretation and to make the ranks independent of the number of clinics.

The PCER can be interpreted as the probability (as a percentage) that a hospital is worse than a randomly selected other hospital.

It remains a value judgment whether ranking is appropriate. We would suggest that any ranking is meaningless when rankability is low (<50%), that the ER should be used when rankability is moderate (>50% and <75%) and that simple integer ranks are only appropriate when rankability is high (>75%). ERs and integer ranks will then be very similar.

Between-hospital differences in acute neurological diseases

We estimated the between-hospital variation in outcome in TBI in chapter 10 and 11. We found that the variation between hospitals in observed proportion unfavourable outcome was very extreme, i.e. between 0 and 100%. Because of the small number of patients per hospital, a large part of these differences was attributable to random variation. However, substantial differences in outcome between the hospitals remained beyond random variation.

In chapter 13 we showed that considerable between-hospital differences existed in the observed proportion poor outcome among stroke patients (between 29 and 78%). Again a large part of these differences was attributable to random variation. The random effects model showed that the between-hospital differences beyond random variation were rather small.

Rankability of the hospitals was low to moderate in stroke (55%), which was due to a combination of limited difference in outcome and low numbers. This was also reflected in the expected ranks: 6 of the 10 centers were shrunken very close to the median rank of 5.5. While integer ranks would have suggested that one of these hospitals was ranked second, and one was ranked seventh.

Summary on statistical uncertainty

Statistical uncertainty is often large when comparing hospitals with outcome measures. Ignoring statistical uncertainty leads to overestimation of the overall differences in outcome and to too extreme estimates of individual hospital performance, especially for smaller hospitals. In both stroke and TBI apparent large differences in outcome are partly attributable to statistical uncertainty. Random effects models take into account statistical uncertainty and are therefore the preferred approach for analysis of between-hospital differences in outcome. Technically random effect models are currently perfectly feasible. Rankability indicates the amount of uncertainty in hospital comparisons and should be reported.

The ER or PCER can be used to incorporate both the magnitude and the uncertainty in rankings, which is ignored by integer rankings. When rankability is lower then 50%, no ranking should be attempt at all. When rankability is between 50% and 75% ERs might be reported. When rankability is over 75%, ERs will be similar to interger ranks.

Case-mix adjustment

Between-hospital differences in outcome may reflect differences in case-mix, i.e. the type or mix of patients treated by a hospital.[6] Since patient characteristics, such as demographics and disease severity may be strong predictors of outcome, even small differences in case-mix may be important to consider when comparing outcomes between hospitals. The third research question addressed the role of case-mix variation.

Prognostic models

Taking into account case-mix differences in hospital comparisons ('case-mix adjustment') can be done with a prognostic model. Prognostic models combine a number of patient characteristics to predict an outcome of interest, most often using regression models.[7]

Although it has been argued that case-mix adjustment models have to be context-specific,[6] a model from the literature might be used. For several diseases commonly accepted case-mix adjustment models are available. A well known example is the TRISS model to evaluate trauma care.[8] When a published model is used it is important to assess whether the model can be expected to be applicable to the setting in which it will be used. Such an assessment is ideally a combination of a qualitative comparison with

the development setting and a quantitative model validation. If results are unsatisfactory, the model can be adapted to the current setting. Possible methods include, ordered from less to more adaptation to the new setting, re-calibration, model revision or model extension.[7] The most extreme variant of adapting is developing a new model on the available data.

Development of a valid prognostic model generally includes seven logically distinct steps: data inspection, coding of the predictors, model specification, model estimation, assessment of the performance, internal and/or external validation and presentation.[7] When developing a prognostic model for case mix adjustment the last step is less relevant.

Once a prognostic model is chosen or developed, the actual case-mix adjustment can be done in several ways. The first approach is to include it in the (random effects) regression model that estimates the between-hospital differences in outcome, either all predictors separate, or the linear predictor of the complete model. The (random effects) estimates are the estimates of the outcome in the different hospitals.

Another approach is to use the prognostic model to estimate a probability on poor outcome for each patient and to sum these up per hospital. This will give the expected number of unfavourable outcomes in a hospital. The expected number can be compared with the observed number, and can be presented as a ratio or as an absolute difference. The Hospital Standardized Mortality Rate (HSMR) is an example of a ratio, the W score is an example of an absolute difference.[9]

The two approaches will give the same results when the same model is used, although they differ in practical aspects. The main difference is that the first approach requires all individual hospital-effects to be estimated at once, and thus all the individual data to be available together. With the second approach this does not have to be the case. Given the availability of a commonly accepted prognostic model, an individual hospital can calculate its expected number of poor outcome and compare that with the observed number.

Pitfalls in case-mix adjustment

The approaches above require patient level data on predictors. If only aggregated data are available, one might consider using these. But aggregated data may introduce ecological bias.[10] Ecological bias can arise from the assumption that relationships observed for groups hold for individuals. This is not necessarily the case since, since not all patients within a group have exactly the mean value.

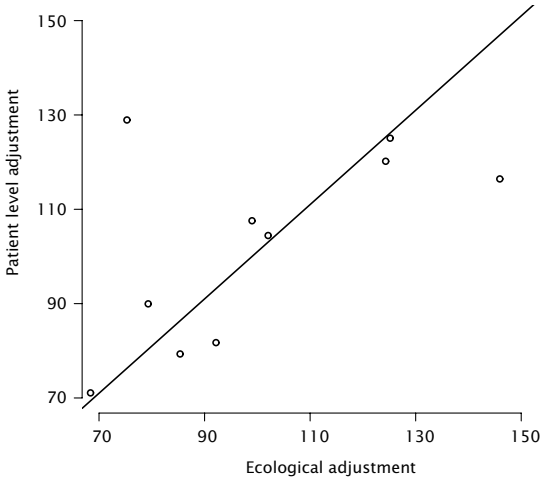
To assess the influence of ecological bias we calculated the standardized mortality ratio's in the stroke data in two ways. First, the individual patient characteristics age, sex and stroke severity were included in the adjustment model. Second, each patient was given the mean hospital value of the linear predictor of the same model.

Figure 14.2 shows that the results of the two approaches differ considerably, with an R^2 of only 0.34. Thus, individual level patient data are needed for case-mix adjustment rather than aggregated data.

Another possible problem in case-mix adjustment the 'constant risk fallacy', which arises when the risk associated with the variable on which adjustment is made varies across the units being compared.[11, 12]

It is important to realize that although with a valid prognostic model it is possible to adjust for patient characteristics to some extent, even the best models can never explain 100% variation in outcome. That means that a great deal of variance in outcome remains unexplained. The error of attributing differences in case-mix adjusted outcomes to quality of care is called 'the case-mix adjustment fallacy'.[13]

Figure 14.2 Standardized rates of poor outcome after stroke for 10 hospitals calculated with individual patient data (x-axis) and aggregated data (y-axis)



Prognostic models in acute neurological diseases

As described in chapter 5, in TBI two good quality prognostic models are available. Both are developed in large numbers of patients. External validation showed good performance. These models were used for case-mix adjustment of the between-center differences in outcome in chapter 10 and 11. In chapter 6 we found that TBI models could be slightly improved by including information on extracranial injury, dependent on the patient population.

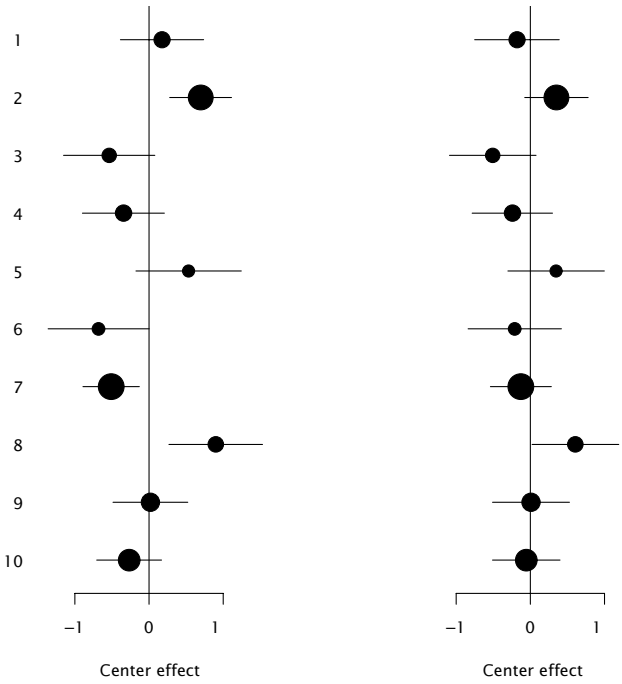
In chapter 8 we developed a model that predicts poor outcome on 4 weeks, 6 months and 12 months in Guillain-Barré syndrome. The model showed very good performance and could be used for case-mix adjustment when comparing hospitals treating GBS patients.

For aneurysmal subarachnoid haemorrhage we developed a model to predict mortality that could be used for case-mix adjustment (chapter 9). The model showed moderate performance after internal validation, which indicates that case mix adjustment possibilities are limited when comparing hospitals on outcome after aneurysmal subarachnoid haemorrhage.

Importance of case-mix adjustment

In chapter 12 and 13 we found that in stroke there were large differences between the hospitals in case-mix, and that not adjusting for these would have led to a very different result of the hospital comparisons. One hospital (hospital 1) even appeared to perform less then average, due to its unfavourable case-mix (Figure 14.3).

Figure 14.3 Unadjusted and adjusted random effect estimates for unfavourable outcome after stroke of 10 hospitals



In TBI there were also large differences between the hospitals in patient characteristics. Adjusting for these led however in both TBI studies (chapter 10 and 11) to an unexpected increase in the estimated between-hospital differences in outcome. This raises the idea that there are hospitals with poor prognosis patients who have actually better outcomes than some hospitals with good prognosis patients.

Other applications of prognostic models

Besides case-mix adjustment, there are numerous other applications of prognostic models. Another application that has relevance for quality of care is the use of prognostic models to identify patients that are likely to benefit from a particular treatment, and to predict whether a patient may require a certain treatment. This application fits into the framework of personalized medicine and was shown in this thesis for GBS in chapter 7 and 8.

An application of models with more indirect relevance for quality of care is the use of prognostic models for design and analysis of studies that aim to assess the effectiveness of a certain treatment. In randomized clinical trials prognostic models can be used for covariate adjustment or prognostic targeting to increase statistical power. The use of the TBI model for this purpose was already shown.[14]

The GBS model developed in chapter 8 will be used for covariate adjustment in an currently ongoing trial for the effectiveness of a Second IVIg Dose in Guillain-Barré syndrome patients with poor prognosis (SID-GBS trial).

In observational studies prognostic models can be used to adjust the treatment effect for differences in prognosis between the treatment groups, either through standard confounder adjustment or through a propensity score, as was shown in chapter 14 for stroke. Furthermore prognostic models can be used for specification of study populations. Such comparisons were shown in chapter 5 for TBI and chapter 8 for GBS.

Summary on case-mix adjustment

The prognostic models described in this thesis show that patient characteristics are highly predictive of outcome. Case-mix may vary between hospitals, and influence between-hospital comparisons on outcome. Case-mix adjustment is therefore absolutely essential when measuring quality of care with outcomes. Prognostic models can be used for case-mix adjustment. For TBI and GBS we reviewed and developed well performing prognostic models. For aSAH the performance of the models was poorer. Besides case-mix adjustment, prognostic models have numerous other applications relevant for quality of care.

Relationship between process and outcome

Outcomes are one aspect of quality of care, processes of care such as certain treatment and interventions are another aspect. When both represent quality of care outcome and process measures are expected to be related. The third research question was whether outcome measures are related to processes of care.

Process and outcome in acute neurological diseases

We found that in stroke, process measures such as thrombolytic therapy were only moderately related to outcome (chapter 12). Since we used generally excepted and evidence-based treatments and interventions, it is unlikely that we have selected the wrong process measures.

Part of the explanation for evidence and consensus based measurements of the process of care having such little impact on outcome, could be that treatment effects are generally modest. Large randomized controlled trials (RCTs) are usually needed to identify a benefit of treatment. In clinical practice, treatment effects may be even smaller, as was shown for statin use in chapter 14. This might be due to lower adherence rates, and different patient populations in RCTs than in clinical practice. Moreover, in quality of care studies we are looking for differential use of established treatments and the resulting differences in outcome will therefore be even smaller.

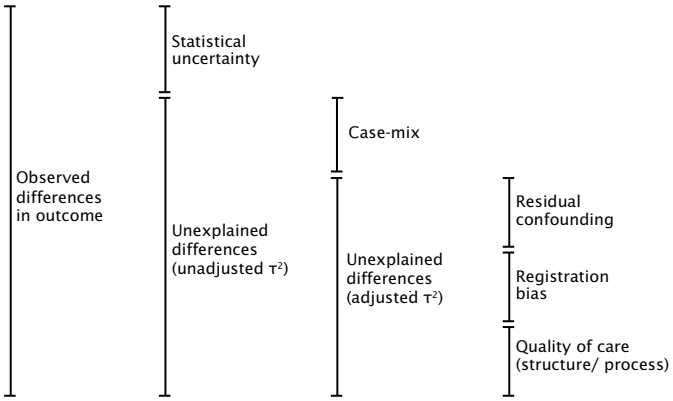
A second explanation might be that process measures are measured on hospital level, e.g. percentage of patients receiving thrombolytic therapy. Not all items of care or treatment apply to all patients. The average of the hospital can not be expected to have a large impact on an individual patient's outcome. This can be seen as ecological bias; the failure of group level associations to properly reflect individual-level associations.[10]

In TBI we did not find strong relationships between process measures and outcome neither, which again may be explained by the effect of treatment being small of even absent, or by methodological problems. One particular example that shows the methodological difficulties in relating process to outcome on a hospital level is the analysis of the effect of time to hospital admission in TBI. We hereto performed analyses in one study in the IMPACT database, the European Brain Injury Consortium study (EBIC)[15] We related time to hospital admission to unfavourable outcome, adjusted for patient characteristics. We found that a long time to admission in individual patients is related to poor outcome (OR 1.02 per hour, p=0.07), as expected. But a long mean time to hospital admission of the particular hospital was related to good outcome (OR=0.94 per hour, p=0.04). This is probably explained by hospitals with longer times to admissions being more specialized in brain injury.

Other explanations for differences in outcome

It might not be very surprising that outcome and process measures are not strongly related. After taking into account statistical uncertainty and case-mix, the remaining unexplained differences in outcome do not solely represent processes of care (Figure 14.4).

Figure 14.4 Schematic partitioning of observed between-hospital differences in outcome



First their may be registration bias. When hospitals measure parameters of interest different, either the outcome or predictors in the case-mix adjustment model, artificial differences are created. Such differences can strongly influence the estimated outcome for an individual centers. High quality, uniformly collected data are therefore a prerequisite for every attempt to compare hospitals.

Further differences in outcome could be caused by residual confounding. As mentioned before, no case-mix adjustment model is perfect, and there will always be unmeasured confounders.

Altogether, observed differences in outcome may consist of a relatively small signal of quality of care differences and a large amount of noise from statistical uncertainty, measured and unmeasured confounders, and registration bias.

These considerations also come back in guidelines with respect to statistical modelling for comparing health care providers based on outcome. These present 7 attributes to take into account: (1) clear and explicit definition of patient sample, (2) clinical coherence of model variables, (3) sufficiently high-quality and timely data, (4) designation of a reference time before which covariates are derived and after which outcomes are measured, (5) use of an appropriate outcome and a standardized period of outcome assessment, (6) application of an analytical approach that takes into account the

multilevel organization of data, and (7) disclosure of the methods used to compare outcomes, including disclosure of performance of risk-adjustment methodology in derivation and validation samples.[16, 17]

Process or outcome?

A potential solution for the poor correlation between outcome and process is measuring process of care directly. This is the topic of a more general debate on whether outcome or process measures should be used to measure quality of care. Outcome assessment is generally easier than assessment of process measures, and requires less detailed data collection. Furthermore outcome measures reflect the relative importance of the different aspects of the care process, which makes them most relevant for patients.

A disadvantage of outcome measures is that flaws in care and well-performed care may cancel each other out, and may not be reflected in overall outcome. It would therefore make more sense to monitor the process of care directly in order to assess quality of care. Another clear advantage of measuring process parameters instead of outcome is that it may directly identify opportunities for improvement in all hospitals, not only in those with poor outcome. Also process measures lack stigma. The message is 'improve X', not 'you are a poor performer'. For this reason they are less likely to prompt perverse solutions. A disadvantage of process measures is that conclusions on quality depend largely on the selection of processes monitored (Table 14.1).[18, 19, 20, 21, 13]

Table 14.1 (Dis)advantages of process and outcome measures for quality of care

Process measures	Outcome measures
+ directly identifies possible improvements	+ easy to collect
+ no stigma	+ relevant for patients
– sensitive for choice of process measures	– flaws could be cancelled out
	– sensitive for case-mix differences

Summary on relationship between process and outcome

It is difficult to clearly relate outcome to processes of care in acute neurological diseases, even when these processes are evidence-based, as in stroke. This implies that even after reasonable case-mix adjustment, low proportions of poor outcome do not necessarily represent high quality processes of care. Neither do high proportions of poor outcome represent poor quality. Observed differences in outcome may be largely attributable to noise from statistical uncertainty, measured and unmeasured confounders, and registration bias, and only for a small part to differences in process. An alternative approach might be to measure process directly, which also has disadvantages.

Implications for policy

The final aim of measuring quality of care is improvement of outcomes, including mortality, morbidity, and general health status. This aim can be achieved through three potential pathways.

1. In the change pathway, quality of care information helps providers to identify areas in which they underperform and improve their quality of care.[22, 23]
2. In the selection pathway patients or their intermediaries use quality of care information to identify better performing hospitals and reward these by 'selecting' the provider.
3. In the reputation pathway, providers attempt to improve quality because they are concerned about their reputation.[24]

For the change pathway, process measures might be more feasible then outcome measures. First it does not require public availability of quality of care in information. This makes the advantage of outcome measures, patient relevance, less important. Also process measures directly identify opportunities for improvement in all hospitals, not only in those with poor outcome. Research has shown however that the change pathway is relatively weak.[24] Moreover this thesis has shown that in acute neurological diseases, process measures are difficult to relate to outcomes which are relevant to patients.

The selection and reputation pathways require public availability of quality information. Studies on public reporting generally find that patients do consider it important to have access to comparative information between hospitals[25-27] but they do not actually compare the information of different hospitals before choosing one. So quality of care will also not be very effectively improved through the selection pathway.[24] For the reputation pathway, patients do not have to use quality information in their choice for a hospital. They only need to form an opinion about the good and poor performing hospitals and discuss it with others. Nevertheless the reputation pathway requires quality information, so patients can form their opinion. The reputation pathway is considered the most promising.[24]

One of the reasons patients do not actually use quality of care information to base choices on is that they have problems processing a large volume of information into a choice.[24] Outcome measures are attractive for their simplicity. But we have shown in this thesis that outcome measures cannot easily be assumed to represent quality of care.

In the reputation pathway quality of care information is only used roughly, and in combination with other input. It could be argued that outcome measures do not have to be perfect as long as they distinguish the best hospitals from the worst.

Quality of care measurement in the Netherlands

In the Netherlands there are several ongoing projects aiming to measure quality of care and making the information publically available. The largest national projects are the performance indicators from the Health Care Inspectorate, the program 'Zichtbare Zorg' ('Visible Care') and the Hospital Standardized Mortality ratio (HSMR). The Health Care Inspectorate and 'Zichtbare Zorg' collect large numbers of quality measures, on structure, process and outcome for specific diseases, and (plan to) make their data publically available on the internet.

A combination of different indicators per diseases might give a reasonable impression of the quality of care for a particular hospital for a particular disease. However the large body of information makes interpretation difficult and process and outcome measures do not necessarily correlate, as was shown in this thesis. For some diseases the number of patients might be too small to draw any conclusions, as shown in chapter 3 of this thesis for the current performance indicators of the Health Care Inspectorate.

The HSMR is an outcome measure, reflecting the ratio between the observed mortality in a hospital and the expected mortality based on patient characteristics. It is attractive for its simplicity, but it is questioned whether it represents quality of care. First, case-mix adjustment might not be sufficient while this thesis has shown that sufficient case-mix adjustment is crucial. A second problem is the high aggregation level. A high level of aggregation is attractive because of the large numbers and hence limited statistical uncertainty, but it complicates interpretation. One department may have good outcomes, while another has poor outcomes. Such differences level out when only overall mortality is measured at the hospital level.

Preceding the public release of the HSMR in 2011, unadjusted mortality rates of all Dutch hospitals were published by the Dutch hospital associations NVZ and NFU in 2010. As this thesis has shown such unadjusted outcome measures for sure do not represent quality of care.

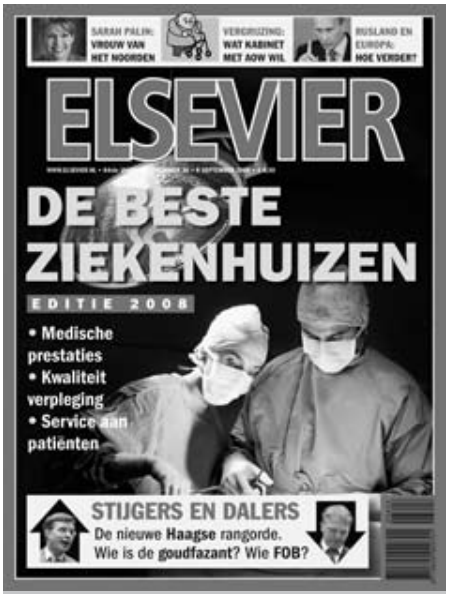
Lay press rankings

Another form of comparing hospitals is publishing rankings, which is popular in the lay press.[28] In the Netherlands one magazine (*Elsevier*, Figure 14.5) and one daily newspaper (*Algemeen Dagblad, AD*) present a yearly 'Hospital top 100'. In other countries there are also plenty of examples of lay press rankings.

Elsevier bases its ranking on expert opinion from doctors, nurses, quality managers, and board members, who are asked what they consider poor and good performing hospitals in their specialty. *AD* bases its ranking on quality indicators from the Healthcare Inspectorate combined with patient satisfaction measurements. Both claim to present quality of care information. We assessed the content validity (do they measure what they aim to?) and the construct validity (do the results represent what they aim to measure?) of these two lay press rankings in the years 2005-2008.

Content validity was assessed by testing consistency between the successive years of each ranking by predicting the rank of a hospital in 2008 based on its rank in the previous years in multivariable regression analyses. We found that only the ranks in 2007 were a significant predictor of the ranks in 2008, for both *Elsevier* and *AD*.

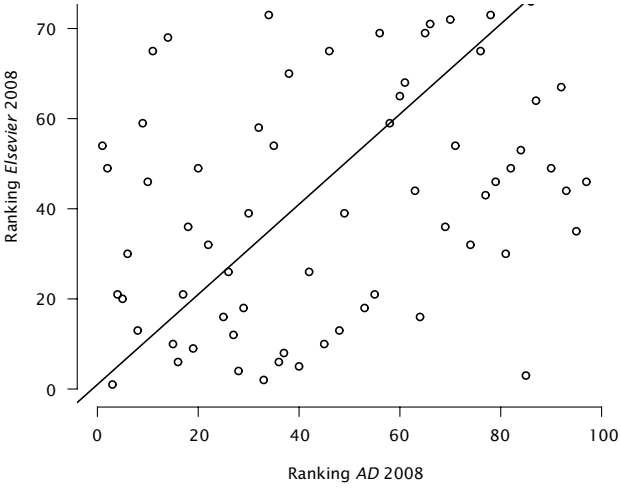
Figure 14.5 Cover of *Elsevier* 2007



Construct validity was assessed by testing the consistency between the *Elsevier* and *AD* rankings by computing correlation coefficients. We found that there was only minor correlation between *AD* and *Elsevier* rankings (Spearman correlation coefficients <0.4, Figure 14.6).

So both content and construct validity were at best moderate, which makes it unlikely that these lay press ranking represent the quality of the hospital.

Figure 14.6 Correlation between rakings *AD* en *Elsevier*



Thus, currently the Dutch general public has the access to different process and outcome measures from different sources, of which none represent quality of care. They are either not interpretable or methodologically invalid. Publishing such measures is useless since they have no potential to improve quality of care. While a large amount of money is spent to produce the figures, they only lead to unfair comparisons. Publishing outcome measures without sufficient case-mix adjustment might even lead to hospitals trying to avoid high risk patients.

Summary on implications for policy

Ideally, quality measures have the potential to improve quality of care though all three quality improvement pathways. This requires specific, comprehensive, detailed process measures for clinicians for the change pathway. And simple outcome measures to make publically available for the selection and reputation pathway. This thesis has shown that when such outcome measures are estimated with a random effects model, sufficiently adjusted for case-mix, and have a moderate to high rankability, they can be considered to at least partly represent quality of care. Since none of the currently available quality of care measures in the Netherlands fulfils these requirements, they should not be made publically available.

Implications for research

Research is an ongoing process. This thesis has provided some answers, but many questions are still open, and new questions are generated.

Statistical uncertainty

We have demonstrated the importance of taking into account statistical uncertainty when comparing hospitals based on outcome. The only way to decrease statistical uncertainty is to increase statistical power. This can be done by either increasing the number of patients, or by using more efficient statistical techniques. Patient numbers can be increased by presenting figures at a higher level of aggregation. This is not the preferred approach since it reduces interpretability. Another approach is to increase time period to measure more patients. E.g. combing multiple years when outcomes are continuously measured for a certain set of hospitals. The disadvantage of this approach is that changes over time become invisible. Also statistical power might be increased by combining multiple outcome measures in one model, by creating composite endpoints, or by using a multivariate model. With a composite endpoint, the outcomes are combined in advance, e.g. by taking any cardiovascular event as outcome instead of only strokes or myocardial infarctions. That composite endpoint is then analysed with a standard regression model. In a (random effects) multivariate model multiple outcomes can be specified, and then both the differences in the outcomes between the hospitals and the correlation between the outcomes can be estimated.

Which of these alternative approaches is preferable is a topic for future research. Such research should include an overview of the practical advantages and disadvantages of the approaches and an assessment of possible statistical methods.

The calculation of the actual sample sizes required is challenging in quality of care research, since this an assumption needs to be made on the magnitude of the differences. In RCTs the assumption is often halving the outcome incidence. Deciding on the magnitude of differences between hospitals that should be detectable is less straightforward. They could be based on the 95% range of odds ratios, e.g a two fold difference corresponds to a range of 0.7 to 1.4. Some studies exist on the actual calculation of the sample sizes required for random effects models.[29] More theoretical work and implementation of the findings to studies comparing hospitals is required in this area. One specific research question is what the effect is of the number of patients per hospital and the number of hospitals on statistical power in a random effects model.

Case-mix adjustment

We have shown the importance of case-mix adjustment when comparing hospitals based on outcome, using prognostic models. There is a strong believe that biomarkers will be able to further improve prognostic models. But since biomarkers will in most

cases be related to disease severity, which is already included in current models with different severity scales, expectations should not be too optimistic. Probably the largest amount of currently unexplained variation in outcome can be attributed to events in the post-acute period, such as secondary or recurrent events. However since quality of care might influence the occurrence of such events, we would not include them in a case-mix adjustment model.

Some different technical aspects of case-mix adjustment still have to be explored in future research, such a fitting one model in the total dataset or stratified per hospital, or different categorisations of variables, or using direct versus indirect standardization. But this will probably lead only to minor improvements.

In the United States case-mix adjustment models are often based on administrative claim data, e.g. from Medicare. This gives the possibility to analyze very large numbers of patients.[5, 17] Models developed on administrative data need to be validated against models based on medical record data. Assessing the possibilities of developing prognostic models on Dutch insurance data and compare these to models based on clinical data is a research topic of high interest.

Process measures

We have shown that even evidence based treatments are at most moderately related to outcome measures when measured at hospital level in clinical practice. This finding implicates that the effects of quality of care on outcomes are small. Good outcomes do not assure good processes. It is unknown to what extent the currently publically available quality measures in the Netherlands, such as the morality rates and the performance indicators from the Health Care Inspectorate correlate. Future research should address this question.

Collecting process measures is however found to be time consuming. This might change with more possibilities for automated data collection, e.g. from electronic patient files, becoming available in the future. Automated data collection will greatly enhance the possibilities for quality of care research, since larger amounts of high-quality data could be made available with fewer resources. Modern possibilities of database linkage, e.g. linking in-hospital data to late outcome data from the civil registry, offer large opportunities.

Specific research steps towards hospital comparisons with automated data include; validation of process measures and patient characteristics from electronic patient files against clinical (study) data, exploring the practical possibilities of linking clinical data to external data sources, and validation of outcome data from external sources against clinical (study) data.

Data quality

A remaining challenge in quality of care research is reducing the noise from registration bias. When data form different hospitals are incomparable, comparisons are useless. A Dutch study showed that large differences existed between hospitals in the coding of data that are used to calculate the Dutch Hospital Standardized Mortality Ratio (HSMR). E.g. the percentage of patients coded as acute admission, which was a predictor in the adjustment model, varied between 29% and 51%, caused by different interpretations by the coding teams.[30] No study has structurally explored the effects of misclassification of patient characteristics, misclassification of outcome, and including different patient groups.

Another source of bias is residual confounding, which could be due to too little adjustment for patient characteristics. Future research should study what the added value is of including more patient characteristics in an adjustment model. Not in terms of model performance, but in differences in results of the hospital comparisons.

Residual confounding could also be due to hospital or region specific policies such as referral patterns or termination of treatment in terminal patients. Another research question is what the magnitude of the effect of such differences on hospital comparisons is.

Level of aggregation

A final topic for future quality of care research is the level of aggregation on which figures should be calculated and presented. High levels of aggregation are attractive because of larger numbers and simplicity, but complicate interpretability.

A single number for a hospital levels out possible differences in quality of care between different departments. A specific research question is to what extent the outcomes e.g. mortality rates of different departments in a hospital differ and contribute the overall mortality rate.

For clinicians, a high level of aggregation does not provide directions to improve their quality of care. Therefore it is important to develop disease specific quality measures that do directly relate to processes of care that might be improved. This could be either specific outcome measures such as disease specific mortality, re-admissions, and disease recurrence, or process measures. Relevant process measures that are easy to obtain and correlated to properly adjusted outcome measures may be good indicators of quality of care.

A challenge for all clinical fields is to develop such disease specific quality of care measures, in collaboration between clinicians and quality of care experts.

Study design

Between-hospital differences in outcome are not only of interest from the perspective of quality improvement. Another relevant area is study design. Currently most randomized trials are multi-center trials, and are conducted in multiple countries. The presence of differences in outcome between the centers may influence the chances of demonstrating a treatment effect in randomized controlled trials (RCTs). In chapter 12 we found that the large between-hospital differences in TBI did not influence the estimate of the treatment effect in a randomized controlled trial. Also the treatment effect did not vary very much between the centers. These findings suggest that between hospitals differences in outcome are not very important for clinical trial design. While in TBI there is a strong belief that between-hospital differences are one of the causes of the many the negative RCTs in the field. We studied only one RCT, so future research has to confirm our results.

Traumatic brain injury

The failure of RCTs to advance the field of TBI has lead to a focus on new approaches, including comparative effectiveness research.

Although many studies are conducted to indicate that a treatment is efficacious relative to a placebo, there are few that directly compare the different available alternatives or that have examined their impacts in populations of the same age, gender/sex, and ethnicity or with the same comorbidities as the patient. Comparative effectiveness research is designed to fill this knowledge gap.[31]

Randomized controlled trials – generally considered to be the gold standard – address efficacy rather than effectiveness. Efficacy reflects the degree to which an intervention produces the expected result under carefully controlled conditions chosen to maximize the likelihood of observing an effect if it exists. The study population and setting of efficacy studies may differ in important ways from those settings in which the interventions are likely to be used. In contrast, comparative effectiveness research intends to measure the benefits and harms of an intervention in ordinary settings and broader populations, and therefore can often be more relevant to policy evaluation and the health care decisions of providers and patients.

Comparative effectiveness research can employ many different methods including database studies, pragmatic randomized trials, health technology assessments, and observational studies. The large between-hospital differences in TBI in outcome presented in chapter 10 and 11 could be considered worrying. From the perspective of CER however, they provide a major opportunity to compare alternative interventions and treatment strategies that all are possible best practices, in every day clinical practice. But as all hospital comparisons, CER studies require uniformly collected, prospective, high quality data, which are currently lacking in TBI. With such data CER may have the ability to answer questions in TBI that will not be answered with RCTs, but are impor-

tant for decision making in clinical practice. A specific challenge for the field of TBI is to perform a well conducted methodologically sound CER study in the coming years.

Summary on implications for research

This thesis has provided methods to account for statistical uncertainty and differences in case-mix. Future research on methods to measure quality of care should mainly aim at reducing statistical uncertainty and at exploring possibilities for automated data collection of patient characteristics, processes and outcomes. Another important research aim is the development of diseases specific quality measures. In prognostic models for case-mix adjustment some methodological improvement might be possible.

References

1. Shahian DM, Normand SL, Torchiana DF, Lewis SM, Pastore JO, Kuntz RE, et al. Cardiac surgery report cards: comprehensive review and statistical critique. *Ann Thorac Surg* 2001; 72(6): 2155-68.
2. Marshall EC, Spiegelhalter DJ. Reliability of league tables of in vitro fertilisation clinics: retrospective analysis of live birth rates. *BMJ* 1998; 316(7146): 1701,4; discussion 1705.
3. Smits JM, De Meester J, Deng MC, Scheld HH, Hummel M, Schoendube F, et al. Mortality rates after heart transplantation: how to compare center-specific outcome data? *Transplantation* 2003; 75(1): 90-6.
4. Krumholz HM, Wang Y, Chen J, Drye EE, Spertus JA, Ross JS, et al. Reduction in acute myocardial infarction mortality in the United States: risk-standardized mortality rates from 1995-2006. *JAMA* 2009; 302(7): 767-73.
5. Krumholz HM, Merrill AR, Schone EM, Schreiner GC, Chen J, Bradley EH, et al. Patterns of hospital performance in acute myocardial infarction and heart failure 30-day mortality and readmission. *Circ Cardiovasc Qual Outcomes* 2009; 2(5): 407-13.
6. Iezzoni LI, editor. *Risk Adjustment for Measuring Healthcare Outcomes*. 3th ed. Chigago: AcademyHealth/HAP Book; 2003.
7. Steyerberg EW. *Clinical Prediction Models. A Practical Approach to Developement, Validating and Updating*. New York: Springer; 2009.
8. Boyd CR, Tolson MA, Copes WS. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score. *J Trauma* 1987; 27(4): 370-8.
9. Parry GJ, Gould CR, McCabe CJ, Tarnow-Mordi WO. Annual league tables of mortality in neonatal intensive care units: longitudinal study. International Neonatal Network and the Scottish Neonatal Consultants and Nurses Collaborative Study Group. *BMJ* 1998; 316(7149): 1931-5.
10. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *Int J Epidemiol* 1989; 18(1): 269-74.
11. Nicholl J. Case-mix adjustment in non-randomised observational evaluations: the constant risk fallacy. *J Epidemiol Community Health* 2007; 61(11): 1010-3.
12. Mohammed MA, Deeks JJ, Girling A, Rudge G, Carmalt M, Stevens AJ, et al. Evidence of methodological bias in hospital standardised mortality ratios: retrospective database study of English hospitals. *BMJ* 2009; 338: b780.
13. Lilford R, Mohammed MA, Spiegelhalter D, Thomson R. Use and misuse of process and outcome data in managing performance of acute medical care: avoiding institutional stigma. *Lancet* 2004; 363(9415): 1147-54.
14. Roozenbeek B, Maas AI, Lingsma HF, Butcher I, Lu J, Marmarou A, et al. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med* 2009; 37(10): 2683-90.
15. Murray GD, Teasdale GM, Braakman R, Cohadon F, Dearden M, Iannotti F, et al. The European Brain Injury Consortium survey of head injuries. *Acta Neurochir* (Wien) 1999; 141(3): 223-36.

16. Krumholz HM, Normand SL, Spertus JA, Shahian DM, Bradley EH. Measuring performance for treating heart attacks and heart failure: the case for outcomes measurement. *Health Aff* (Millwood) 2007; 26(1): 75-85.
17. Krumholz HM, Brindis RG, Brush JE, Cohen DJ, Epstein AJ, Furie K, et al. Standards for statistical models used for public reporting of health outcomes: an American Heart Association Scientific Statement from the Quality of Care and Outcomes Research Interdisciplinary Writing Group: cosponsored by the Council on Epidemiology and Prevention and the Stroke Council. Endorsed by the American College of Cardiology Foundation. *Circulation* 2006; 113(3): 456-62.
18. Rothen HU, Takala J. Can outcome prediction data change patient outcomes and organizational outcomes? *Curr Opin Crit Care* 2008; 14(5): 513-9.
19. Mant J. Process versus outcome indicators in the assessment of quality of health care. *Int J Qual Health Care* 2001; 13(6): 475-80.
20. Rubin HR, Pronovost P, Diette GB. The advantages and disadvantages of process-based measures of health care quality. *Int J Qual Health Care* 2001; 13(6): 469-74.
21. Lilford R, Pronovost P. Using hospital mortality rates to judge hospital performance: a bad idea that just won't go away. *BMJ* 2010; 340: c2016.
22. Berwick DM, James B, Coye MJ. Connections between quality measurement and improvement. *Med Care* 2003; 41(1 Suppl): I30-8.
23. Fung CH, Lim YW, Mattke S, Damberg C, Shekelle PG. Systematic review: the evidence that publishing patient care performance data improves quality of care. Ann Intern Med 2008; 148(2): 111-23.
24. Hibbard JH. What can we say about the impact of public reporting? Inconsistent execution yields variable results. *Ann Intern Med* 2008; 148(2): 160-1.
25. Lubalin JS, Harris-Kojetin LD. What do consumers want and need to know in making health care choices? *Med Care Res Rev* 1999; 56 Suppl 1: 67,102; discussion 103-12.
26. Vonberg RP, Sander C, Gastmeier P. Consumer attitudes about health care acquired infections: a German survey on factors considered important in the choice of a hospital. *Am J Med Qual* 2008; 23(1): 56-9.
27. Losina E, Plerhoples T, Fossel AH, Mahomed NN, Barrett J, Creel AH, et al. Offering patients the opportunity to choose their hospital for total knee replacement: impact on satisfaction with the surgery. *Arthritis Rheum* 2005; 53(5): 646-52.
28. Wang W, Dillon B, Bouamra O. An analysis of hospital trauma care performance evaluation. *J Trauma* 2007; 62(5): 1215-22.
29. Normand SL, Zou KH. Sample size considerations in observational health care quality studies. *Stat Med* 2002; 21(3): 331-45.
30. Van den Bosch WF, Silberbusch J, Roozendaal K, wagner C. Variatie in codering patientgegevens beinvloedt gestandaardiseerd ziekenhuis sterftecijfer (HSMR). *Nederlands Tijdschrift voor Geneeskunde* 2010; 154(A1189).
31. IOM (Institute of Medicine). *Initial National Priorities for Comparative Effectiveness Research*. Washington; The National Academies Press; 2009.

Summary

I Introduction

Over the past 20 years, quality of care has become a major topic in health care. Only two decades ago, physicians could be confident that they alone had a social mandate to judge and manage the quality of care. In contrast, in the current era of evidence-based medicine, medical practice is continuously and critically evaluated by different stakeholders.

But measuring quality of care is complex, since no uniform definition exists and many factors determine the outcome of a patient. The care provided by one specific health care provider is only one of these factors. So far there has been no generally accepted approach or method to measure quality of health care, and to compare health care providers.

Quality of care comprises structure, process and outcome. Structure relates to organisation of care, such as number of beds in a hospital. Process relates to actual actions of care, such as whether the patient receives medication within a certain time frame. Outcome includes patient outcome measures such as mortality. Outcome measures are attractive since they are most relevant to the final aim of measuring quality of care; improvement of patient outcomes. Measuring quality of care with outcomes poses two major methodological problems: statistical uncertainty and differences in the type of patients ('case-mix') between hospitals.

With regard to statistical uncertainty, there will always be some variation in outcome between hospitals caused by chance (statistical uncertainty). Ignoring this may lead to overinterpretation of differences in outcome between hospitals, especially if the numbers are small.

With regard to case-mix, outcomes will differ regardless of quality of care when hospitals have a different patient population e.g. in terms of age and disease severity. To account for patient characteristics that influence outcome, a prognostic model can be used. Prognostic models combine a number of patient characteristics to predict the outcome of interest, most often using regression models. Comparisons between providers could then be done with adjustment for each patient's risk as estimated by the prognostic model.

The aim of this thesis was to develop methods to measure quality of care with outcome measures. Specific questions included:

1. What is the role of statistical uncertainty in measuring quality of care with outcome measures?

1a. How large is the effect of statistical uncertainty on between-hospital comparisons?

1b. How should statistical uncertainty be incorporated in outcome measures?

2. What is the role of case-mix variation in measuring quality of care with outcome measures?

2a. How large is the effect of case-mix on between-hospital comparisons?

2b. How can case-mix variation be captured for between-hospital comparisons?

3. How do outcome measures relate to processes of care?

The methods studied in this thesis are applied to acute neurological diseases, including traumatic brain injury (TBI), stroke, Guilllain-Barré syndrome (GBS) and aneurysmal subarachnoid haemorrhage (aSAH).

II Statistical uncertainty

In part II of the thesis methods to take into account statistical uncertainty when measuring quality of care with outcome measures are presented.

In chapter 2 it was shown how statistical uncertainty can be incorporated in rankings. Rankings based on outcome are often used to present hospital performance. These rankings do however not reflect that part of the variation in outcome between providers is caused by statistical uncertainty or by differences in case-mix, and not by any differences in quality of care.

We compared percentages of poor outcome after stroke (mortality and disability at 1 year) between 10 hospitals in the Netherlands. There where substantial differences in outcome between hospitals in unadjusted analysis, but adjustment for case mix differences led to halving of these differences. Further changes were seen when statistical uncertainty was taken into account, especially for smaller hospitals. We calculated the Expected Rank (ER), an estimate of the rank of a provider beyond statistical uncertainty, and a measure for rankability ρ, which is the part of variation between providers that is due to true differences (as opposed to statistical uncertainty).

Using the Expected Rank led to shrinkage of the original ranks of 1 to 10 towards the median rank of 5.5 and to a different order of the hospitals. The rankability was 55%, which we interpret as approximately half of the differences between hospitals being due to noise rather than signal.

In comparing and ranking hospitals, case-mix adjusted random effect estimates and the Expected Ranks are more robust alternatives to traditional estimates and simple rankings. The Expected Rank provides a way to combine the attractiveness of a rank-

ing, namely a single number and easy interpretation, with reliable analyses that does justice to the providers.

Rankability was also used in chapter 3, to show how much statistical uncertainty is present in outcome measures that are currently used by the Dutch Healthcare Inspectorate to assess the quality of hospital care.

With the official data on seven outcome indicators of 97 Dutch hospitals of the year 2007, it was shown that sample sizes were typically small (median 2-21 cases per hospital). This caused substantial uncertainty and only poor (< 50%) to moderate (50-75%) rankability.

Thus the currently used Dutch outcome indicators are not suitable for ranking hospitals. When judging hospital quality the influence of statistical uncertainty must be accounted for to avoid overinterpretation of the numbers in the quest for more transparency in health care. Adequate sample size is a prerequisite in attempting reliable ranking.

Statistical uncertainty can be taken into account by estimating the hospital outcomes with random effects logistic regression models. In chapter 4 different statistical software packages fitting such models are compared.

The data of 8509 patients with traumatic brain injury from 231 centers were used to fit dichotomized and ordinal logistic random effects regression models in different statistical packages.

The studied software packages produced similar results, but differed somewhat in usability and flexibility. The choice for a particular implementation may therefore solely depend on the desired flexibility, and the usability.

III Prognostic models

In part III, I address the problem of case-mix differences between hospitals when measuring outcomes. Several prognostic models that could be used for case-mix adjustment in hospital comparisons are developed and described (Table 1).

Table 1 Overview of the prognostic models described and developed in this thesis

Chapter	Model	Disease	Outcome	Predictors
5	IMPACT (basic model)	TBI	6 month unfavorable outcome	– Age – GCS motor score – Pupil reactivity
5	CRASH (basic model)	TBI	6 month unfavorable outcome and 14 day mortality	– Age – GCS – Pupil reactivity – Extracranial injury
7	EGRIS	GBS	Mechanical ventilation	– Days between onset of weakness and admission – MRC sumscore – Presence of facial and/or bulbar weakness
8	mEGOS	GBS	Ability to walk after 4 weeks, 3 months and 6 months	– MRC sum score – Age – Preceding diarrhea
9	–	aSAH	60-day mortality	– WFNS grade – Age – Lumen size of the aneurysm – Fisher grade

TBI=Traumatic brain injury, GCS= Glasgow Coma Scale, GBS=Guillain-Barré syndrome, aSAH=Aneurysmal subarachnoid, haemorrhage, MRC=Medical research Council, WFNS=World Federation of Neurosurgical Societies

Chapter 5 gives an overview of the prognostic models available to predict outcome in TBI patients and some general considerations about prognostic model development. Despite the association of many variables with outcome, making predictions for individuals is notoriously difficult. A systematic literature search identified 16 studies reporting on prognostic models based upon admission characteristics; many of these showed shortcomings, which may partly explain the limited use of these models in clinical practice. Two high quality prediction models are currently available, that have been developed on large datasets with state of the art methods. There is great potential for use in clinical practice, in research, towards policy making and assessment of the quality of health care delivery.

In chapter 6 the potential improvement of case-mix adjustment models in TBI by including extracranial injury as a predictor is assessed. Major extracranial injury is common in TBI patients, but its prognostic value is not well known.

In pooled data of five studies major extracranial injury was related to mortality with logistic regression analysis, adjusted for known important predictors.

It was found that major extracranial injury is an important prognostic factor for mortality in patients with TBI. However, the strength of the effect was smaller in

patients with more severe brain injury. Also the strength of the effect decreases when only considering patients who survive the early phase after injury, instead of considering all patients, starting from the time of injury.

Chapter 7 presents a prognostic model that can be used to identify GBS patients that will require mechanical ventilation. Respiratory insufficiency is a frequent and serious complication of GBS, and prediction of respiratory insufficiency is important to triage patients to the appropriate unit and avoid respiratory distress.

In this study a prognostic model to predict the chance of respiratory insufficiency in the acute stage of disease was developed, which showed good performance at external validation in an independent cohort. After further validation, the model may assist in clinical decision-making, e.g. on patient transfer to an ICU.

A second prognostic model for GBS is presented in chapter 8, predicting ability to walk. GBS is a heterogeneous disease, yet patients are treated with standard therapy. This is insufficient for some patients. These patients may benefit from additional therapy but this requires early identification. A prognostic model to predict walking ability at 4 weeks, 3 months and 6 months after hospital admission was developed. Discriminative ability was good, as well as calibration in an independent external validation cohort.

With this model poor outcome can be accurately predicted in an early phase of the disease and it can therefore be used to select patients for additional therapy, and to improve quality of care.

In chapter 9 a prognostic model to predict outcome in aSAH patients is developed. Aneurysmal subarachnoid haemorrhage is a devastating event with substantial case-fatality. In this study a model to predict 60-day case-fatality using clinical and neuroimaging characteristics, available on admission was developed and evaluated. The model was internally validated with bootstrapping techniques, showing reasonably discriminative ability.

After external validation these predictors could eventually be used in clinical decision making.

IV Applications

In the fourth part of the thesis the methods and models presented in part II and III are applied to traumatic brain injury (TBI) and stroke.

Chapter 10 studies between-hospital differences in outcome in TBI. In this study we analyzed patients with moderate and severe TBI from 265 centres. Using random effects logistic regression models, the between-centre differences in outcome were

estimated, adjusted for differences in patient characteristics. Taking into account statistical uncertainty, there was a 3.3 fold difference in the odds of unfavourable outcome between the centers at the lower end of the outcome distribution (2.5th percentile) versus those at the higher end of the outcome distribution (97.5th percentile).

Further research is needed to study explanations for these differences and to suggest where quality of care might be improved.

Chapter 11 investigates whether the between-hospital differences in outcome after TBI affect the estimation of the treatment effect in clinical trials. We hereto analyzed a large randomized controlled trial (the CRASH trial).

We studied the effect of the treatment on 14 day mortality in patients from 237 centers, using different statistical approaches taking into account between-center differences in outcome and between-center differences in treatment effect. The 14-day mortality was higher in the treatment group. If center differences were ignored, the odds ratio was 1.22. ($p=0.00010$). Although there were large between-center differences in outcome, these did not substantially change the estimated treatment effect (OR=1.24, $p=0.00003$). The between-center variation in the treatment effect was limited.

Thus large between-center differences in outcome do not necessarily affect the estimated treatment effect in RCTs.

In chapter 12 it is assessed whether the between-hospital differences in outcome after stroke are more related to patient characteristics or to process measures.

In patients with acute stroke from 10 centers in the Netherlands, poor outcome was related to patient characteristics and quality of the care (diagnostic, therapeutic and preventive procedures in patients with indication) with logistic regression models. The proportion of patients with poor outcome varied across the centers from 29% to 78%. The largest part of variation in patient outcome between centers was explained by differences in patient characteristics. Quality of care parameters explained a small part of the variation in patient outcome.

Hence, unadjusted proportions of poor outcome after stroke are not valid as indicators of quality of care.

Chapter 13 studied the relation between one particular process measure in stroke, treatment with statins, and outcome. The benefit of statin treatment in patients with a previous ischemic stroke or TIA has been demonstrated in RCTs. However, the effectiveness in everyday clinical practice may be decreased because of a different patient population and less controlled setting.

In this study we described statin use in an unselected cohort, identified factors related to statin use, and tested whether the effect of statins on outcome observed in RCTs, is also observed in everyday clinical practice.

Age, elevated cholesterol levels and other cardiovascular risk factors were associated with statin use at discharge. After 3 years, 39% of the users at discharge had stopped using statins. After adjustment for the differences between statin users and non-statin users, statins had a small but beneficial effect of on the occurrence of recurrent vascular events and mortality (OR 0.8, 95% CI: 0.6–1.2).

Despite the poor treatment adherence to statins, the observed beneficial effect of statins in this study is compatible with the effect observed in clinical trials.

V Discussion

The aim of this thesis was to study methods to measure quality of care with outcome measures, and to apply these methods to different acute neurological diseases. In the first half of the thesis I described and developed methods to face the two major methodological challenges in measuring quality of care with outcome measures; statistical uncertainty and case-mix adjustment. In the second half these methods were applied to different acute neurological diseases.

We found that statistical uncertainty is often large when comparing hospitals with outcome measures. Ignoring statistical uncertainty leads to overestimation of the overall differences in outcome and to too extreme estimates of individual hospital performance, especially for smaller hospitals. Different methods were proposed to account for statistical uncertainty and to quantify statistical uncertainty, including the use of random effect models, and calculation of rankability and the Expected Rank (ER). Random effects models take into account statistical uncertainty and are therefore the preferred approach for analysis of between-hospital differences in outcome. Technically random effect models are currently perfectly feasible. Rankability indicates the amount of uncertainty in hospital comparisons and should become a standard element in reporting. The Expected Rank can be used to incorporate both the magnitude and the uncertainty in rankings, which is ignored by integer rankings. When rankability is lower then 50%, no ranking should be attempted at all. When rankability is between 50% and 75% ERs might be reported. When rankability is over 75%, ERs will be similar to interger ranks.

In both stroke and TBI we found that apparent large differences in outcome are partly attributable to statistical uncertainty.

With regard to case-mix, the prognostic models described in this thesis show that patient characteristics are highly predictive of outcome. Case-mix may vary between hospitals, and influence between-hospital comparisons based on outcome. Case-mix adjustment is therefore absolutely essential when measuring quality of care with outcomes. Prognostic models can be used for case-mix adjustment, although they will

never be perfect. For TBI and GBS we reviewed and developed well performing prognostic models. For aSAH the performance of the models was poorer. Note that prognostic models have numerous applications relevant for quality of care, besides case-mix adjustment in between provider comparisons.

In stroke a large part of the between-hospital differences was attributable to case-mix differences.

We further found that it is difficult to clearly relate outcome to processes of care in acute neurological diseases, even when these processes are evidence-based, as in stroke. This implies that even after reasonable case-mix adjustment, low proportions of poor outcome do not necessarily represent high quality processes of care. Neither do high proportions of poor outcome represent poor quality.

Summarizing, given reliable data, observed differences in outcome may be largely attributable to noise from statistical uncertainty, and measured and unmeasured differences in case-mix, and only for a small part to differences in process. We conclude that outcome measures for quality of care should be case-mix adjusted random effect estimates, which are related to processes of care.

The findings in this thesis have implications for both policy and research. From a policy perspective, good quality measures have the potential to improve quality of care though three different quality improvement pathways (the change, selection and reputation pathway). This requires specific, comprehensive, detailed process measures for clinicians for the change pathway, and simple outcome measures to make publically available for the selection and reputation pathway. This thesis has shown that when such outcome measures are estimated with a random effects model, sufficiently adjusted for case-mix, and have a moderate to high rankability, they can be considered to at least partly represent quality of care. Since none of the currently available quality of care measures in the Netherlands fulfils these requirements, they should not be made publically available.

From a research perspective, this thesis has provided methods to account for statistical uncertainty and differences in case-mix. An important research aim for the future is the development of diseases specific quality measures. In prognostic modelling for case-mix adjustment some methodological improvement might be possible. But future research on methods to measure quality of care should mainly aim at reducing statistical uncertainty and at exploring possibilities for automated data collection on patient characteristics, processes and outcomes.

Samenvatting

I Introductie

In de laatste 20 jaar is kwaliteit van zorg een steeds belangrijker onderwerp geworden binnen de gezondheidszorg. Vóór die tijd werd het bewaken en verbeteren van kwaliteit van zorg volledig aan artsen overgelaten. Tegenwoordig is er echter veel behoefte aan transparantie en proberen verschillende belanghebbenden inzicht te krijgen in kwaliteit van zorg. Dat blijkt ondermeer uit de ziekenhuisranglijsten die worden gepubliceerd door Elsevier en het Algemeen Dagblad. Ook de Inspectie voor de Gezondheidszorg publiceert kwaliteit van zorg informatie.

Het meten van kwaliteit van zorg is lastig. Ten eerste bestaat er geen eenduidige definitie van kwaliteit en ten twee is gezondheidszorg complex. Vele factoren bepalen hoe het uiteindelijk gaat met een patiënt. Dat is dus niet alleen toe te schrijven aan de kwaliteit van zorg van één specifieke zorgaanbieder, zoals het ziekenhuis. Tot nu toe is er geen algemeen geaccepteerde methode om kwaliteit van zorg te meten en zorgaanbieders te vergelijken.

Kwaliteit van zorg kan verdeeld worden in 3 domeinen: structuur, proces en uitkomst. Een structuurmaat is bijvoorbeeld het aantal bedden dat een ziekenhuis heeft, of de aanwezigheid van een spoedeisende hulp. Procesmaten zijn de daadwerkelijke zorgactiviteiten, bijvoorbeeld het zo snel mogelijk geven van bloedverdunners aan patiënten met een herseninfarct, of het maken van een CT scan. Uitkomstmaten zijn uitkomsten van patiënten, zoals sterftecijfer of het percentage patiënten dat weer zelfstandig kan wonen na een herseninfarct. Om kwaliteit te meten zijn uitkomstmaten aantrekkelijk; ze zijn gemakkelijk te meten, en relevant voor patiënten. Bovendien is het uiteindelijke doel van meten en verbeteren van kwaliteit van zorg het verbeteren van uitkomsten. Het voorschrijven van bloedverdunners is immers geen doel op zich, maar een middel om sterfte te reduceren.

Maar het meten van kwaliteit van zorg met uitkomstmaten brengt twee problemen met zich mee. Ten eerste kunnen uitkomsten beïnvloed worden door toeval. Als de kans op sterfte in een bepaald ziekenhuis 50% is, kan het toch voorkomen dat bijvoorbeeld 14 van de 20 patiënten (70%) overlijden. Dat is het gevolg van statistische onzekerheid en vergelijkbaar met het opgooien van een munt. Hoewel we weten dan de kans op kop 50% is, kan het voorkomen dat in 20 worpen de uikomst 14 keer kop is. Hoe kleiner het aantal worpen, en dus het aantal patiënten, hoe groter de statistische onzekerheid. Als er geen rekening wordt gehouden met statistische onzekerheid lijken verschillen in uitkomst tussen zorgaanbieders groter dan ze zijn.

Ten tweede kunnen verschillen in uitkomst veroorzaakt worden door verschillen in de patiëntenpopulaties. Als een ziekenhuis bijvoorbeeld veel oudere patiënten behandelt, of patiënten die ernstig ziek zijn, zal het sterftecijfer hoger zijn dan in een ziekenhuis met een jongere patiëntenpopulatie, ongeacht de kwaliteit van zorg. Om rekening te houden met verschillen in patiëntenpopulatie kan een prognostisch model worden gebruikt. In een prognostisch model wordt de kans op overlijden (of een andere uitkomstmaat) van een patiënt geschat op basis van de patiëntkarakteristieken zoals leeftijd, geslacht, ernst van de ziekte etc. De gemiddelde kans op overlijden van alle patiënten in een bepaald ziekenhuis kan dan vergeleken worden met de daadwerkelijke sterfte.

In dit proefschrift worden methoden voor het meten van kwaliteit van zorg met uitkomstmaten onderzocht. Daarbij gaat het specifiek om de rol van statistische onzekerheid, van verschillen in patiëntenpopulatie en om de relatie tussen uitkomstmaten en procesmaten. De methoden die worden besproken zijn toegepast op verschillende acute neurologische aandoeningen: traumatisch hersenletsel, herseninfarct, aneurysmale hersenbloeding en Guillain-Barré syndroom. Traumatisch hersenletsel is hersenletsel door een externe oorzaak, bijvoorbeeld een val of een verkeersongeluk. Een herseninfarct is een verstopping van een bloedvat waardoor een deel van de hersenen geen zuurstof meer krijgt. Een aneurysmale hersenbloeding is een bloeding in de hersenen die wordt veroorzaakt door een uitstulping in een bloedvat, een aneurysma, dat knapt. Guillain-Barré syndroom is een acute aandoening van de zenuwen die de spieren aansturen, die kan leiden tot spierzwakte of verlamming.

II Statistische onzekerheid

Dit deel van het proefschrift beschrijft de rol van toeval als gevolg van statistische onzekerheid bij het meten van kwaliteit van zorg met uitkomstmaten.

Eerst is onderzocht hoe rekening kan worden gehouden met statistische onzekerheid in uitkomstmaten. Hiertoe zijn tien Nederlandse ziekenhuizen vergeleken op basis van het aantal patiënten dat 1 jaar na een herseninfarct overleden of gehandicapt was. Er bleken grote verschillen te zijn tussen de ziekenhuizen. Het grootste deel van die verschillen werd echter veroorzaakt door statistische onzekerheid. Vooral de sterftepercentages van kleine ziekenhuizen werden sterk beïnvloed door statistische onzekerheid. Slechts 55% van de oorspronkelijke verschillen in uitkomst was toe te schrijven aan daadwerkelijke verschillen, de rest aan statistische onzekerheid.

De Inspectie voor de Gezondheidszorg gebruikt ook uitkomstmaten om kwaliteit van ziekenhuiszorg te meten. In het volgende hoofdstuk is onderzocht hoe groot de rol van statistische onzekerheid hierin is. Het bleek dat de uitkomsten berekend worden op vaak kleine patiëntenaantallen per ziekenhuis (tientallen), wat de invloed van statistische onzekerheid vergrootte. Van de verschillen in uitkomsten tussen ziekenhuizen was in alle gevallen minder dan 75% veroorzaakt door echte verschillen, de rest door statistische onzekerheid.

Als geen rekening wordt gehouden met statistische onzekerheid leidt dat tot een te hoge schatting van verschillen in uitkomst tussen ziekenhuizen. Er zijn statistische methoden beschikbaar om rekening te houden met statistische onzekerheid.

III Prognostische modellen

In het volgende deel van het proefschrift is onderzocht wat de rol is van verschillen in patiëntenpopulatie bij het meten van kwaliteit van zorg met uitkomstmaten.

Eerst is een overzicht gegeven van de prognostische modellen die beschikbaar zijn om uitkomst bij patiënten met traumatisch hersenletsel te voorspellen. Uit dit overzicht bleek dat er twee goede prognostische modellen ontwikkeld zijn die op basis van leeftijd en ernst van het hersenletsel kunnen voorspellen of een patiënt overlijdt of gehandicapt blijft na traumatisch hersenletsel. Verder bleek dat de prognostische modellen nog licht verbeterd kunnen worden door ook het al dan niet hebben van letsel aan de rest van het lichaam als voorspeller mee te nemen.

Voor patiënten met Guillain-Barré syndroom zijn twee prognostische modellen ontwikkeld. Patiënten met Guillain-Barré kunnen last krijgen van ademhalingsproblemen. Zij moeten in dat geval beademend worden op de intensive care. Er is een prognostisch model ontwikkeld dat op basis van spierkracht en de mate van progressie van de ziekte voorspelt welke patiënten een grote kans hebben op ademhalingsproblemen. Deze patiënten kunnen eventueel uit voorzorg naar de intensive care worden gebracht. Ook is een prognostisch model gemaakt dat op basis van leeftijd, spierkracht en aanwezigheid van diaree (wat een indicatie is voor aanwezigheid van een bepaalde bacterie) voorspelt of patiënten 4 weken, 3 maanden en 6 maanden na het ontstaan van de ziekte weer kunnen lopen.

In het volgende hoofdstuk is een prognostisch model gemaakt dat voorspelt of patiënten met een hersenbloeding na twee maanden zijn overleden. De ernst van de bloeding, leeftijd, de grootte van het aneurysma en of de bloeding zichtbaar is op een CT scan waren de sterkste voorspellers voor het overlijden van een patiënt.

Verschillen in patiëntenpopulatie beïnvloeden uitkomstmaten, ongeacht kwaliteit van zorg. Het is daarom cruciaal dat bij het vergelijken van ziekenhuizen op basis van uitkomst rekening wordt gehouden met patiëntenpopulatie. De prognostische modellen voor traumatisch hersenletsel en Guillain-Barré kunnen daarvoor gebruikt worden. Omdat het model voor hersenbloedingen slechts redelijk voorspelde, kan het slechts met enig voorbehoud worden gebruikt bij het meten van kwaliteit van zorg.

IV Toepassingen

In het vierde deel van het proefschrift zijn met behulp van de methoden uit de voorgaande delen ziekenhuizen vergeleken op basis van uitkomstmaten.

Er zijn 265 ziekenhuizen uit de hele wereld vergeleken die patiënten met middelzwaar tot ernstig traumatisch hersenletsel behandelden. Er bleken grote verschillen (meer dan factor 3) tussen de ziekenhuizen te bestaan in het percentage patiënten dat zes maanden na het hersenletsel overleden of gehandicapt was, ook als rekening werd gehouden met statistische onzekerheid. Deze verschillen werden niet veroorzaakt door verschillen in patiëntenpopulatie. Een deel ervan wordt dus mogelijk veroorzaakt door verschillen in kwaliteit van zorg, hoewel altijd rekening moet worden gehouden met de (on)betrouwbaarheid van de dataverzameling.

Verder zijn tien Nederlandse ziekenhuizen vergeleken op basis van het percentage patiënten met een herseninfarct dat na 6 maanden overleden of gehandicapt was. Naast deze uitkomstmaat werden ook procesmaten gemeten, bijvoorbeeld het percentage patiënten waarbij op tijd een CT scan werd gemaakt en het percentage dat de juiste medicatie kreeg voorgeschreven. Er bleken grote verschillen in uitkomst te bestaan; het percentage patiënten dat na 6 maanden overleden of gehandicapt was varieerde van 29% tot 78%. De verschillen werden grotendeels veroorzaakt door verschillen in patiëntenpopulatie. Slechts een klein deel werd veroorzaakt door verschillen in zorgprocessen. Dus zelfs als rekening is gehouden met verschillen in patiëntenpopulatie betekenen goede uitkomsten niet per definitie kwalitatief goede zorgprocessen. Evenmin als slechte uitkomsten slechte zorgprocessen betekenen.

V Discussie

Het doel van dit proefschrift was om methoden voor het meten van kwaliteit van zorg met uitkomstmaten te onderzoeken. Eerst zijn methoden onderzocht die gebruikt kunnen worden om rekening te houden met statistische onzekerheid en verschillen in patiëntenpopulatie. Vervolgens zijn deze toegepast op acute neurologische ziekten.

De conclusie is dat alleen van uitkomstmaten die rekening houden met statistische onzekerheid en verschillen in patiëntenpopulatie en die samen hangen met relevante procesmaten kan worden aangenomen dat ze – tenminste gedeeltelijk – kwaliteit van zorg meten.

Deze bevindingen hebben ten eerste implicaties voor onderzoek. In dit proefschrift zijn methoden beschreven om rekening te houden met statistische onzekerheid en verschillen in patiëntenpopulatie. Toekomstig onderzoek zou zich voornamelijk moeten richten op het verkleinen van statistische onzekerheid, bijvoorbeeld door verschillende uitkomstmaten te combineren of door het meten van uitkomsten over langere periodes.

Ten tweede zijn er implicaties voor beleid. Geen van de huidige uitkomstmaten voor

kwaliteit van zorg die momenteel beschikbaar is voor het Nederlandse publiek, houdt namelijk rekening met statistische onzekerheid en verschillen in patiëntenpopulatie. Het heeft dan ook geen zin om deze uitkomstmaten openbaar te maken, ze leveren geen informatie over kwaliteit van zorg en zijn enkel misleidend.

Dankwoord

Enerzijds is het afronden en verdedigen van dit proefschrift een mijlpaal. Anderzijds zie ik de jaren van mijn promotieonderzoek, voor mij tevens de eerste jaren van het werkende bestaan, als een levensfase. Een fase waarin ik veel heb geleerd, maar ook veel leuke dingen heb meegemaakt. Het is dan ook een genoegen om terug te kijken op de afgelopen jaren en de mensen te noemen aan wie ik veel te danken heb, en die het leven leuk maken, of allebei.

Als eerste mijn promotor prof. dr. Ewout Steyeberg. Ewout, ik had me geen betere promotor dan jij kunnen wensen. Ik heb verschrikkelijk veel van je geleerd, niet alleen inhoudelijk. Je bent op vele vlakken een voorbeeld. Ik stel het ook op prijs dat je altijd hebt meegedacht over mijn ontwikkeling en me zoveel ruimte en kansen hebt gegeven. Maar bovenal waardeer ik ons goede contact, de altijd gezellige tripjes en vele hardlooprondjes. Bedankt voor alles, ik hoop dat we nog lang kunnen samenwerken.

Ik heb de eer gehad te mogen samenwerken met verschillende wijze heren, van wie ik veel heb kunnen leren en aan wie ik veel te danken heb.

Prof. dr. Andrew Maas. Andrew, zonder jou had mijn promotieperiode er een stuk saaier uitgezien en had ik nog geen tiende van de mensen ontmoet die ik nu ken. Daarnaast heb ik veel geleerd van je strategische en sociale vaardigheden. Heel veel dank voor het vertrouwen en de vele geboden kansen.

Dr. Diederik Dippel. Diederik, mede dankzij jouw goede en enthousiaste begeleiding bij mijn afstudeer onderzoek ben ik de lol van onderzoek in gaan zien. Daarnaast heb je me veel geleerd over allerlei aspecten van onderzoek. Bedankt voor dit alles, en ik ben blij dat ik met jou en je onderzoeksgroep kan blijven samenwerken.

Prof. dr. Gordon Murray. Gordon, first of all thank you for travelling to Rotterdam to take place in the committee. I would also like to thank you for the nice collaboration within the IMPACT group and your contribution to the papers in my thesis. I have learned a lot from your sharp comments and your social skills. And I enjoyed all the 6 am runs.

Prof. dr. ir. Eric Boersma, dr. Bart Jacobs en prof. dr. Roland Bal, hartelijk dank voor de goede samenwerking en voor het plaatsnemen in de commissie.

Dr. Wilma Scholte op Reimer. Wilma, ondanks mijn keuze voor een onderzoeksmaster ben ik onderzoek pas leuk gaan vinden tijdens mijn afstudeeronderzoek bij jou en Diederik. Daarnaast heb je me geholpen na te denken over wat ik verder wilde doen en me in contact gebracht met de juiste mensen. Veel dank daarvoor.

Naast de reeds genoemde mensen hebben verschillende co-auteurs bijgedragen aan artikelen in dit proefschrift. Christa, Roelof, Anne-Margreet, Li en Nikki, dank dat ik

jullie papers heb mogen gebruiken voor mijn proefschrift, en vooral bedankt voor de gezelligheid en goede samenwerking.

Bob, ik heb genoten van ons werk samen, zelfs van TLN versie 35. Om nog maar niet te spreken over ons gezamenlijke tripje naar Santa Barbara, fun fun fun! En natuurlijk bedankt dat je mijn paranimf wilt zijn.

The 'IMPACT familiy' Jim, Izzy, Morag, Anne-Claire, and Juan. Thank you for your contributions to my thesis and for the good company during the IMPACT meetings. Prof. dr. Mamarou I will remember you as a dedicated and always friendly person. The CRASH and TARN investigators, thank you for your contributions to this thesis and the nice collaboration. I hope we keep working together.

Binnen het Erasmus MC heb ik in de afgelopen jaren met veel verschillende mensen samen gewerkt, meer of minder intensief, maar in alle gevallen leuk en leerzaam. Heleen, Maaike, Anne-Marie, Madeleine, Hanane, Marjolein, Juna, Joost, Iain, Ruben, Tahlita, Willemijn, Mario, Yorick, Hilde, bedankt voor de goede samenwerking en de gezelligheid.

Thijs en 'tante' Jet, bedankt voor jullie bijdrage aan de Nederlandse samenvatting.

Speciale vermelding verdient Sanne Hoeks. Sanne, ik kan je voor veel dingen bedanken. Voor het regelen van mijn afstudeeronderzoek bij jullie groep, voor de vele 'werk' besprekingen in de koffiecorner, voor het feit dat ik mee mocht op wintersport... In ieder geval had mijn leven er zonder jou heel anders uit gezien!

Ook de leuke (ex) collega's op MGZ dragen bij aan het feit dat ik dagelijks met plezier het Erasmus MC inwandel. In willekeurige volgorde, Lenneke, Frans-Willem, Carolien, Carlijn, Merel, Lidy, Suzan, Nicolien, Noortje, Sanne, Sonja, Anja, Farsia, Esther de B, Suzanne, Frank, Gladys, Caspar, Gerard, en alle anderen. Bedankt voor alle hulp en ondersteuning, praatjes op de gang, koppen koffie, Italiaanse broodjes, MGZ wandelingen, fietstochten, sectie-uitjes en volleybalwedstrijden. Rolf, onze tijd in AE-138 was top! Bedankt voor de alle lol, goede gesprekken en ongevraagde wijze adviezen. DMS, ik ben blij dat ik mijn proefschrift heb kunnen afmaken voordat onze etentjes teveel hersencellen hebben doen sneuvelen. Dat er nog vele mooie avonden mogen volgen.

Maar gelukkig wandel ik ook dagelijks met plezier het Erasmus MC weer uit. Dankzij allerlei mensen die ik wil bedanken omdat ze – zoals gezegd – het leven leuk maken.

Mijn atletiekmaatjes waaronder trainer Richard, Manon het sprintkanon, en natuurlijk de MILAs. Samen rennen, fietsen, rennen, zwemmen, rennen, koffie drinken, rennen, uit eten, rennen... Misschien denken mensen dat we gek zijn, maar wij weten wel beter! Robi en Luci ook bedankt voor jullie werk aan mijn proefschrift.

Inge B, opruimen zal nooit meer zo leuk zijn als in de Schout Heynricstraat! Nu zien we elkaar gelukkig regelmatig in en zonnige zuiden. En de rest van Missy, om een Nederlandse rapper te citeren: 'Had ik maar een tijdmachine. Maar die heb ik niet'. Dus hoop ik gewoon dat we contact houden.

Gem, we hebben leuke en mooie jaren gehad samen, en je hebt me geholpen goede keuzes te maken in en na mijn studie.

Inge K, in een jaar of 15 zijn we via basketball, Joppe, kamperen met de rubberboot, skiën, glansrollen in 'de Doos', vele etentjes, feestjes, en nog veel meer, terecht gekomen in de Senaatzaal van de Erasmus Universiteit. Vanwaar we straks snel doorgaan naar het volgende etentje en feestje. Bedankt dat je mijn paranimf wilt zijn!

Mijn aller-allerliefste pappie en mammie en Hessel. En mijn favoriete schoonzus Fiona. Ik heb het getroffen met zo'n leuke familie, niet alleen vroeger maar zeker ook nu. Bedankt voor alle gezellige etentjes en weekendjes, hulp, en onvoorwaardelijke steun, ook aan de minder leuke Hester die ik thuis soms was. Of ben?

En natuurlijk Stefan. Want ik ben iedere dag weer blij en gelukkig dat ik met jou samen ben!

Curriculum vitae

Hester Floor Lingsma was born in Amstelveen, the Netherlands, on the 11th of October 1982. After finishing secondary school at the Alkwin Kollege in Uithoorn in 2002, she started studying Policy and Management in Health Care at the Erasmus University Rotterdam. In addition, she started studying History at the University Leiden in 2003 and finished the first year ('propedeuse') in 2005. She obtained her BSc in Policy and Management in Health Care also in 2005 and started a Master of Science in Clinical Epidemiology, specialization Health Services Research, at the Netherlands Institute for Health Sciences (NIHES). She obtained her MSc degree in 2007 with a thesis on the relation between processes of care and patient outcome after stroke, under supervision of dr. Wilma Scholte op Reimer (Department of Cardiology, Erasmus MC, Rotterdam) and dr. Diederik Dippel (Department of Neurology, Erasmus MC, Rotterdam). In July 2007 she started a PhD project at the Department of Public Health of the Erasmus MC under supervision of prof. dr. Ewout Steyerberg, resulting in this thesis.

Hester Floor Lingsma werd geboren op 11 oktober 1982 in Amstelveen. Na het behalen van haar VWO diploma op het Alkwin Kollege in Uithoorn begon ze in 2002 aan de studie Beleid en Management van de Gezondheidszorg aan de Erasmus Universiteit in Rotterdam. Daarnaast ging ze in 2003 Geschiedenis studeren aan de Universiteit Leiden, en behaalde haar propedeuse in 2005. In 2005 behaalde ze ook de Bachelor Beleid en Management van de Gezondheidszorg en begon aan een Master Clinical Epidemiology, specialisatie Health Services Research, aan het Netherlands Institute for Health Sciences (NIHES). In 2007 studeerde ze af met een onderzoek naar de relatie tussen zorgprocessen en uitkomsten na een herseninfarct, onder begeleiding van dr. Wilma Scholte op Reimer (Thoraxcentrum, Erasmus MC, Rotterdam) en dr. Diederik Dippel (Afdeling Neurologie, Erasmus MC, Rotterdam). In juli 2007 begon ze op de afdeling Maatschappelijke Gezondheidszorg van het Erasmus MC onder begeleiding van prof. dr. Ewout Steyerberg aan het promotieonderzoek dat heeft geleid tot dit proefschrift.

Publications

2010

Wijnhoud AD, Maasland L, Lingsma HF, Steyerberg EW, Koudstaal PJ, Dippel DW. Prediction of Major Vascular Events in Patients With Transient Ischemic Attack or Ischemic Stroke; A Comparison of 7 Models. *Stroke*. 2010 Sep 2.

Lingsma H, Roozenbeek B, Steyerberg E, On Behalf Of The Impact Investigators. Covariate adjustment increases statistical power in randomized controlled trials. *J Clin Epidemiol*. 2010 Aug 27.

Korte MR, Fieren MW, Sampimon DE, Lingsma HF, Weimar W, Betjes MG; on behalf of the investigators of the Dutch Multicentre EPS Study. Tamoxifen is associated with lower mortality of encapsulating peritoneal sclerosis: results of the Dutch Multicentre EPS Study. *Nephrol Dial Transplant*. 2010 Jun 27.

Roozenbeek B, Lingsma HF, Steyerberg EW, Maas AI; the IMPACT Study Group. Underpowered trials in critical care medicine: how to deal with them? *Crit Care*. 2010 Jun 8;14(3):423.

Hoeks SE, Scholte Op Reimer WJ, Lingsma HF, van Gestel Y, van Urk H, Bax JJ, Simoons ML, Poldermans D. Process of Care Partly Explains the Variation in Mortality Between Hospitals After Peripheral Vascular Surgery. *Eur J Vasc Endovasc Surg*. 2010 May 21.

Walgaard C, Lingsma HF, Ruts L, Drenthen J, van Koningsveld R, Garssen MJ, van Doorn PA, Steyerberg EW, Jacobs BC. Prediction of respiratory insufficiency in Guillain-Barré syndrome. *Ann Neurol*. 2010 Jun;67(6):781-7.

Zuiverloon TC, van der Aa MN, van der Kwast TH, Steyerberg EW, Lingsma HF, Bangma CH, Zwarthoff EC. Fibroblast growth factor receptor 3 mutation analysis on voided urine for surveillance of patients with low-grade non-muscle-invasive bladder cancer. *Clin Cancer Res*. 2010 Jun 1;16(11):3011-8.

Lingsma HF, Roozenbeek B, Steyerberg EW, Murray GD, Maas AI. Early prognosis in traumatic brain injury: from prophecies to predictions. *Lancet Neurol*. 2010 May;9(5):543-54.

McHugh GS, Butcher I, Steyerberg EW, Marmarou A, Lu J, Lingsma HF, Weir J, Maas AI, Murray GD. A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. *Clin Trials*. 2010;7(1):44-57.

Risselada R, Lingsma HF, Bauer-Mehren A, Friedrich CM, Molyneux AJ, Kerr RS, Yarnold J, Sneade M, Steyerberg EW, Sturkenboom MC. Prediction of 60 day case-fatality after aneurysmal subarachnoid haemorrhage: results from the International Subarachnoid Aneurysm Trial (ISAT). *Eur J Epidemiol.* 2010 Apr;25(4):261-6.

Maas AI, Steyerberg EW, Marmarou A, McHugh GS, Lingsma HF, Butcher I, Lu J, Weir J, Roozenbeek B, Murray GD. IMPACT recommendations for improving the design and analysis of clinical trials in moderate to severe traumatic brain injury. *Neurotherapeutics.* 2010 Jan;7(1):127-34.

Lingsma HF, Steyerberg EW, Scholte op Reimer WJ, van Domburg R, Dippel DW; Netherlands Stroke Survey Investigators. Statin treatment after a recent TIA or stroke: is effectiveness shown in randomized clinical trials also observed in everyday clinical practice? *Acta Neurol Scand.* 2010 Jul;122(1):15-20.

Lingsma HF, Steyerberg EW, Eijkemans MJ, Dippel DW, Scholte Op Reimer WJ, Van Houwelingen HC; Netherlands Stroke Survey Investigators. Comparing and ranking hospitals based on outcome: results from The Netherlands Stroke Survey. *QJM.* 2010 Feb;103(2):99-108.

Capelle LG, Van Grieken NC, Lingsma HF, Steyerberg EW, Klokman WJ, Bruno MJ, Vasen HF, Kuipers EJ. Risk and epidemiological time trends of gastric cancer in Lynch syndrome carriers in the Netherlands. *Gastroenterology.* 2010 Feb;138(2):487-92.

van Dishoeck AM, Lingsma HF, Steyerberg EW. [Performance indicators: the role of 'task uncertainty']. *Ned Tijdschr Geneeskd.* 2010;154:A1775. (in Dutch)

Lingsma HF. Ziekenhuiszorg verbetert niet door openbare sterftecijfers. *Medisch Contact* 2010; 2:57. (in Dutch)

2009

Roozenbeek B, Maas AI, Lingsma HF, Butcher I, Lu J, Marmarou A, McHugh GS, Weir J, Murray GD, Steyerberg EW; IMPACT Study Group. Baseline characteristics and statistical power in randomized controlled trials: selection, prognostic targeting, or covariate adjustment? *Crit Care Med.* 2009 Oct;37(10):2683-90.

Korte MR, Boeschoten EW, Betjes MG; EPS Registry. The Dutch EPS Registry: increasing the knowledge of encapsulating peritoneal sclerosis. *Neth J Med.* 2009 Sep;67(8):359-62. (in Dutch)

Lingsma HF, Eijkemans MJ, Steyerberg EW. Incorporating natural variation into IVF clinic league tables: The Expected Rank. *BMC Med Res Methodol.* 2009 Jul 16;9:53.

Pons H, Lingsma HF, Bal R. De ranglijst is een slechte raadgever. *Medisch Contact* 2009; 47:1969-71. (in Dutch)

Roozenbeek B, Maas AI, Marmarou A, Butcher I, Lingsma HF, Lu J, McHugh GS, Murray GD, Steyerberg EW; IMPACT Study Group. The influence of enrolment criteria on recruitment and outcome distribution in traumatic brain injury studies: results from the impact study. *J Neurotrauma.* 2009 Jul;26(7):1069-75.

van Dishoeck AM, Lingsma HF, van der Kolk M, Steyerberg EW, Robben P, Mackenbach JP. *Haalbaarheidstudie naar de kwantificering van gezondheidseffecten van toezicht gehouden door de Inspectie voor de Gezondheidszorg.* Rotterdam 2009. (in Dutch)

2008

Lingsma HF, Dippel DW, Hoeks SE, Steyerberg EW, Franke CL, van Oostenbrugge RJ, de Jong G, Simoons ML, Scholte op Reimer WJ. [Differences between hospitals in outcome after a stroke are only partially explained by differences in the quality of care]. *Ned Tijdschr Geneeskd.* 2008 Sep 27;152(39):2126-32. (in Dutch)

Maas AI, Lingsma HF; IMPACT Study Group. New approaches to increase statistical power in TBI trials: insights from the IMPACT study. *Acta Neurochir Suppl.* 2008;101:119-24.

Lu J, Murray GD, Steyerberg EW, Butcher I, McHugh GS, Lingsma HF, Mushkudiani N, Choi S, Maas AI, Marmarou A. Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials. *J Neurotrauma.* 2008 Jun;25(6):641-51.

Steyerberg EW, Lingsma HF. Predicting citations: Validating prediction models. *BMJ.* 2008 Apr 12;336(7648):789.

Lingsma HF, Dippel DW, Hoeks SE, Steyerberg EW, Franke CL, van Oostenbrugge RJ, de Jong G, Simoons ML, Scholte Op Reimer WJ; Netherlands Stroke Survey investigators. Variation between hospitals in patient outcome after stroke is only partly explained by differences in quality of care: results from the Netherlands Stroke Survey. *J Neurol Neurosurg Psychiatry.* 2008 Aug;79(8):888-94.

PhD Portfolio

Summary of PhD training and teaching activities

Name PhD student: Hester F. Lingsma
Erasmus MC Department: Public Health
PhD period: 2007-2010
Promotor: Prof dr E.W. Steyerberg

	Year	Workload (ECTS)
1. PhD training		
General academic skills		
Biomedical English Writing and Communication	2008	4
Research Integrity	2009	1
NWO talent classes	2010	1
Research skills		
'Mixed Models: Models for Longitudinal and Incomplete data'	2008	1
'Bayesian statistics'	2010	1
Seminars and workshops		
Consultation Center for Patient Oriented research (CPO)	2007–2010	1
COEUR Research seminars	2007–2010	1
Seminars department of Public Health	2007–2010	3
'Meta-analysis in Prognosis' Freiburg, Germany	2008	1
Invited participation 'Personalized Medicine' Brussels, Belgium	2010	1
Presentations		
Presentations within Erasmus MC	2007–2010	5
National conferences	2007–2010	2
International conferences	2007–2010	8
International conferences		
European Stroke conference	2007/2008	2
Society for Medical Decision Making	2008/2010	2
International Neurotrauma Society	2009/2010	2
International Brain Injury Association	2010	1
Society for Clinical Trials	2010	1
2. Teaching activities		
Lecturing		
'Prognostic Research' (Msc course University of Utrecht)	2008	1
NIHES course 'Clinical Epidemiology'	2009	1
Introductie survival analyse (Geneeskunde keuzeonderwijs Oncologie)	2010	1
Supervising practicals		
NIHES course 'Classical Methods for Data Analysis'	2007–2010	3
NIHES course 'Clinical Epidemiology'	2007–2008	2
NIHES course 'Advanced Analysis of Prognosis Studies'	2008–2010	3
Supervising Master's theses		
BSc and MSc students EUR and VU Amsterdam	2009–2010	4