Prognosis after traumatic brain injury

Chantal Hukkelhoven

Prognosis after traumatic brain injury

Prognose na traumatisch hersenletsel

Chantal Hukkelhoven

Printing of this thesis is partly realised with financial support of the Department of Public Health of the Erasmus MC, University Medical Center Rotterdam.

Prognosis after traumatic brain injury

Prognose na traumatisch hersenletsel

Proefschrift

ter verkrijging van de graad van doctor
aan de Erasmus Universiteit Rotterdam
op gezag van de rector magnificus
Prof.dr. S.W.J. Lamberts
en volgens besluit van het College voor Promoties.

De openbare verdediging zal plaatsvinden op
donderdag 22 december 2005 om 13.30 uur

door

Chantal Wilhelmina Paula Maria Hukkelhoven

geboren te Sittard

Prognosis after traumatic brain injury / Hukkelhoven, Chantal
Thesis Erasmus MC, University Medical Center Rotterdam –
With summary in English and Dutch

Cover illustration: Jan Wille
The drop in the water is symbolic for traumatic brain injury; a sudden event with far-reaching consequences for the patient and his family and friends
Design & lay-out: Albert Epping
Printed by Print Partners Ipskamp, Enschede

ISBN-10: 9090201181
ISBN-13: 9789090201184

Promotiecommissie

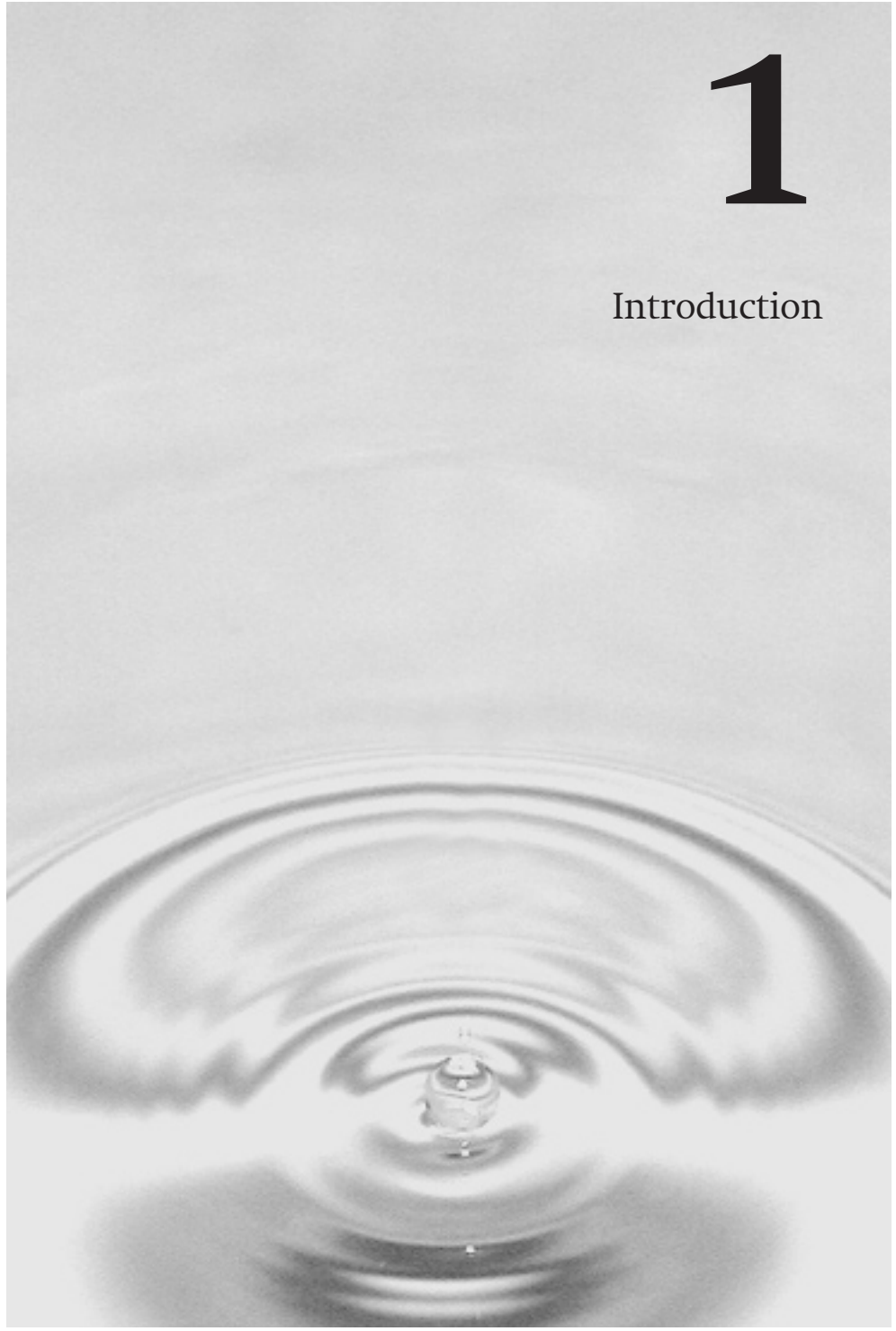
Promotor:	Prof.dr. J.D.F. Habbema
Overige leden:	Prof.dr. M.M.B. Breteler Prof.dr. Th. Stijnen Prof.dr. C.J.J. Avezaat
Copromotoren:	Dr. E.W. Steyerberg Dr. A.I.R. Maas

Voor mijn vader
Ter nagedachtenis aan mijn moeder

Contents

1

Introduction



1.1 Traumatic brain injury

Traumatic brain injury (TBI) is an important public health care problem in the western world. It is one of the most common causes of death in young adults and it can affect people's lives enormously.

TBI is generally defined as an injury to the brain caused by an external physical force. Often, the term 'head injury' is used synonymously with TBI. Since 'head injury' may refer to injury of the skull only, in this thesis the term 'traumatic brain injury' is used.

Symptoms of TBI can be various, depending on the extent of damage to the brain. Specific characteristics of TBI, however, are the presence of amnesia and/or a loss of consciousness after the injury. In mild TBI, post-traumatic amnesia is an indicator of the severity of the injury; in patients with more severe injuries depth and duration of this post-traumatic loss of consciousness is a better indicator of the severity. In patients with more severe injuries the level of consciousness can be depressed for weeks or months.

TBI evolves over time[1,2]. The sudden and often profound mechanical damage that occurs at the time of injury is considered the primary damage. This primary damage initiates a complex sequence of events, causing secondary damage. Such events may be the development of a haematoma in the intracranial compartment or result from pathophysiological mechanisms. Secondary damage can be further exacerbated by systemic insults, such as hypoxia and hypotension.

1.2 Prevalence, etiology and impact

In the USA at least 5 million people (2% of the population) currently live with disabilities resulting from TBI, and each year at least 1,4 million sustain a TBI[3]. Of these, about 50.000 die, 85 per 100.000 persons are hospitalized and 390 per 100.000 inhabitants are treated and released from an emergency department[4]. In the Netherlands, each year about 60 per 100.000 inhabitants require hospitalization and around 90 per 100.000 persons with an age of 20 years or older visit the emergency room of a hospital because of a TBI[5,6]. In 2002 around 950 (6 per 100.000) Dutch persons died because of a TBI[7].

The risk of experiencing a TBI is not equally divided among all age groups. Adolescents, young adults and persons older than 70 years have the highest risk of TBI (Figure 1).

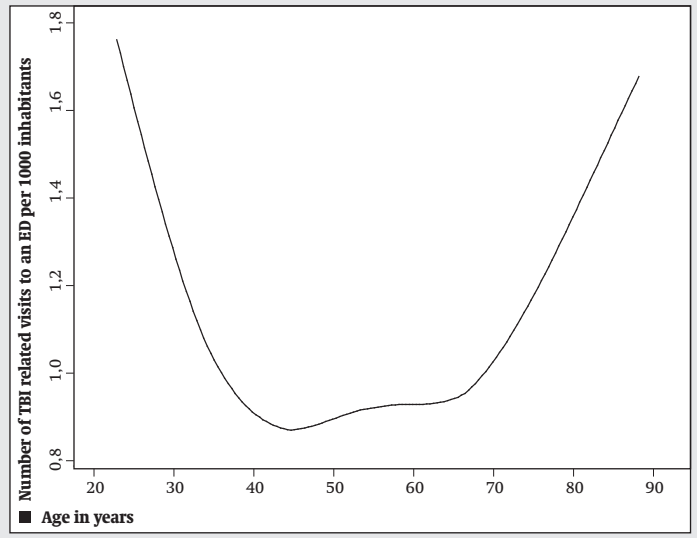
In the Netherlands, most persons who visited the emergency room of a hospital because of a TBI were injured by a traffic accident (42%). Other main causes of TBI were a fall (35%), an accident at home (5%), a sports accident (9%), an accident at work (4%) and a suicide attempt (5%)[5].

The societal costs of TBI can be substantial. In the United States the joint direct and indirect costs of TBI are estimated at $56,3 billion in 1995[8]. In the Netherlands, around 9.600 patients were admitted to a hospital because of a TBI in 1999[5,6]. The financial burden of direct medical costs due to TBI in the Netherlands was calculated to be € 72 million in 1999. These costs exclude the number of lost working days (more than 17.000 days).

1.3 Consequences

The consequences of a TBI on the life of a TBI victim and family can be substantial. Dependent on the severity of the TBI, a TBI victim may – sometimes temporarily – suffer from problems with cognition (concentration, memory, judgment, mood), movement abilities (strength, coordination and balance), sensation (tactile sensation and special senses such as vision) and emotional stability[3]. Social consequences of TBI can also be considerable, including increased risk of suicide, divorce, chronic unemployment and economic strain. These consequences are tragic, not only to the TBI victims, but also to their relatives. Family members report depression, social isolation and anger, and the family life is often disrupted. The social consequences of TBI are illustrated by the story of Marco, a TBI victim, who permitted us to describe the impact that TBI had on his life (www.nietaangeborenhersenletsel.tk).

Figure 1. Number of TBI related visits to an emergency department (ED) per 1000 persons per age category in the Netherlands in 1999[5]



Personal story of Marco

The accident happened in July 1999. While driving on my scooter, I was hit by a car and landed with my head against a house. By ambulance I was brought to a general hospital, where I stayed for one day. Subsequently I was transported to the Dijkzigt Hospital in Rotterdam. There, I have been in coma for two weeks. In total I have spent four weeks in the Dijkzigt Hospital. Afterwards, I have been taken to several rehabilitation centers, where I stayed for two months in total. Finally, when I was allowed to go home again, I followed daily therapy in a rehabilitation center near my house.

After the accident, I had to learn everything again: talking, sitting, walking, and going to the toilet by myself, really everything. This was very hard for me and I often wished I had never woken up again. I know this sounds cruel, but this is how I felt at that moment. Nevertheless, I am very glad that my family and previous girlfriend helped me to live further. Happily I am physically more or less healthy again; I only experience some problems with my memory.

After about two years I went back to my former employer (I used to be a cook at a rather high level in the catering industry). I liked working again very much, since I had something to do. Together with the GAK and a reintegration company I set up a plan to return completely to my old profession. Everything went well, but as soon as I was absent for some time, e.g. for a holiday, I had to start all over again. I did not know what to do. Since the catering industry was too hectic for me, I started looking for another job, together with a work reintegration company. This wasn't easy, since I had to think of my wishes and capabilities. I had (and have) problems with my memory, so not every job was and is suitable. Maybe I had to receive a new training. At that time I suffered from: tiredness, forgetting names or things easily, speaking not audibly, feeling angry or offended easily, being unable to practice my profession. Before the accident I had many friends, but I lost most of them, although nowadays I don't consider them as my real friends.

I found it really hard to contact 'unfamiliar' people, people I did not know before. People cannot immediately see that I suffer(ed) from TBI, so they often treat me rudely or start pestering me. Therefore, I advice other TBI victims to overcome their initial hesitation and tell everybody what happened. By telling people they are given the opportunity to respond adequately. People act differently when they know, although some people also start patronizing.

After performing a test to choose a profession, it appeared that I loved to work with computers. Therefore, I started a new training in SPW-3. After having followed this training I started working for a foundation. Basically, everything went well, but I still experienced too much problems, e.g. with my memory. Now, I am looking for a job that I like and that I can handle. I think this will be computerwork.

1.4 Classification of TBI

A TBI can be classified according to several aspects, such as mechanism, clinical severity and morphology.

Mechanically, TBI can be classified into closed and penetrating TBI. A closed TBI occurs when a blunt object strikes the head or when brain damage results from acute acceleration/ deceleration or rotational forces. A compound skull fracture may be present in closed TBI. A penetrating TBI occurs when an object, such as a bullet, pierces the skull and enters the brain tissue.
Both types have a different pathology. Penetrating TBI primarily results in direct damage to the cerebral tissue and hemorrhaging from the penetrating object. In closed TBI the resulting brain injury may be focal (restricted to one area of the brain) or diffuse (involving different areas of the brain).

To evaluate clinical severity, various grading systems have been developed. The most often used system is the Glasgow Coma Scale (GCS)[9], which is composed of three parameters, i.e. eye opening, motor response and verbal response (Table 1).

For purposes of classification the scores of the three parameters can be added up, yielding a total score between 3 and 15. This total score summarizes the patients' level of consciousness and is used for purposes of classification. At the bedside, however, the individual GCS parameters are considered more informative than the total GCS. According to the total GCS, about 80% of the TBI patients receiving medical attention can be categorized as mild (GCS 13-15), 10% as moderate (GCS 9-12) and 10% as severe (GCS 3-8) at the time of injury[10].

Table 1. Glasgow Coma Scale (GCS)

Eye opening	Spontaneously	4
	To verbal command	3
	To pain	2
	No response	1
Best motor response	Obeying commands	6
	Localizing	5
	Flexion withdrawal	4
	Abnormal flexion	3
	Extension	2
	No response	1
Verbal response	Oriented	5
	Confused conversation	4
	Inappropriate words	3
	Incomprehensible sounds	2
	No response	1
Total		3-15

Table 2. Diagnostic categories of types of abnormalities visualized on computed tomography (CT) scanning

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

The extent of damage can be assessed morphologically with a computed tomography scan (CT scan). The CT scan creates a series of cross-sectional X-ray images of the head and brain and can show bone fractures, as well as the presence of lesions, contusions and brain tissue swelling. Abnormalities on the CT scan are often morphologically categorized according to the presence or absence of a 'mass lesion' or according to the Marshall CT-classification[11] (Table 2). After the initial assessment and treatment of the TBI, magnetic resonance imaging (MRI) may be used to detect more subtle changes in the brain tissue.

1.5 Grading outcome

Different scales to grade outcome after TBI are available. One of the most widely used outcome scales is the Glasgow Outcome Scale (GOS)[12]. This is a five level classification scale that assesses overall outcome after TBI (Table 3). The scale is often dichotomized into two groups: favorable and unfavorable outcome. Favorable outcome includes the categories 'moderate disability' and 'good recovery', while unfavorable outcome includes 'death', 'vegetative state' and 'severe disability'. Soon after the injury, individual patients move frequently from one GOS category to another, but from six months onwards the GOS seems to stabilize[13].

Table 3. Glasgow Outcome Scale (GOS)

Score	Description
Death	
Vegetative state	Awake, but not aware
Severe disability	Conscious, but disabled
Moderate disability	Disabled, but independent
Good recovery	Resumption of normal life, even though there may be minor neurological or psychological deficits.

1.6 Prognosis and rationale of this thesis

The outlook for patients with a mild TBI is generally a good recovery, while patients with a severe TBI have a substantial risk to die. Predicting outcome for very good or very severe patients is therefore rather easy. However, for most severe and moderate TBI patients the outcome is not so easy to predict, while such predictions would be helpful. For example, they may support clinical decision-making and provide realistic and evidence-based expectations to relatives (counseling) and caregivers. Outcome predictions may also be applied to classify patients according to prognostic risk, which may be useful to compare outcome between patient series from different centers or to study treatment results over time. Furthermore, the design and analysis of randomized clinical trials (RCTs) may be improved; prognostic risks may be used to define enrollment criteria and for risk-stratification to estimate covariate-adjusted treatment effect[14-19].

Individual outcome predictions can be obtained by prognostic models, created by multivariable analysis of patient data, including multiple clinical risk factors (predictors), such as age and GCS. By filling in these factors for a specific patient, the model can provide an individual probability of an outcome, e.g. the likelihood that the patient will die during the next six months[20,21].

In the past, several multivariable models have been developed with the purpose to predict outcome for TBI patients. The methods used for model development, however, are often inconsistent or even contradictory, e.g. with regard to the selected predictors and the way predictors are included in the model. This may limit the value and potential application of the previously developed models. Proper evaluation of prognostic models in TBI requires an understanding of issues affecting the design, conduct, analysis, reporting and validation of such models. In addition, many models that predict long-term outcome after TBI have been developed on relatively small cohorts from one single place or region, so that their applicability – i.e. their performance on more recent patients or patients from different centers – is questionable. Also the validity of prognostic models in TBI has seldom been evaluated in independent patients. These restraints limit the confidence we may have in previously developed prognostic models in TBI.

Next to predicting outcome more accurately, it may be useful to predict the need of specialized intensive care. Nowadays a high proportion of the severe and moderate TBI patients are first transported to a general hospital and later to a level I trauma center (secondary referral). This secondary referral may delay the institution of appropriate therapy and increase the risk of adverse events and systemic insults during inter-hospital transport. A more efficient triage may be aided by early identification of patients in need of specialized intensive care. This may be facilitated by the use of prognostic modeling.

1.7 Objectives of this thesis

This thesis studies the prognosis of patients with a severe or moderate TBI, with a focus on predicting long-term mortality and unfavorable outcome, using baseline demographic, clinical and CT characteristics. We have the following objectives:

1. **To study methodological developments in prognostic modeling in TBI.**
   We aim to derive guidelines for model development and validation in TBI, based on a review of methodological aspects in previously published prognostic studies in TBI.

2. **To develop and validate prognostic models that predict long-term outcome for patients with severe or moderate TBI.**
   We use baseline demographic, clinical and CT characteristics to predict six-month mortality and unfavorable outcome. The models are validated on the study population (internal validation) and on several other TBI cohorts of substantial size (external validation). Furthermore, we compared the external performance of our models to that of four previously developed prognostic models.

3. **To predict the need of specialized intensive care.**
   We investigate the feasibility of predicting the risk of potential operable lesions and the risk of raised intracranial pressure. Both outcome measures are indicators of the need of specialized intensive care.

Table 4. Overview of the patient series used in this thesis

Series	N	Period data collection	Region*	No of centers	TBI severity#	Thesis chapter
Multi-center trials						
Tirilazad trial	2269	1991-1994	Eur, Aus, N-Am	76	sev, mod	3 - 7
Selfotel trial	427	1994-1995	Eur, Aus, N-Am, S-Am	52	sev	7
Unselected multi-center series						
EBIC survey‡	796	1995	Eur	67	sev, mod	6, 7
TCDB‡	746	1984-1988	N-Am	4	sev	6, 7
Single-center series						
ErasmusMC	275	2000-2003	NL	1	mainly sev, mod	8

* Eur = Europe, Aus = Australia, N-Am = North-America, S-Am = South-America, NL = the Netherlands
\# sev = severe, mod = moderate
‡ EBIC = European Brain Injury Consortium, TCDB = Traumatic Coma Data Bank

1.8 Patient series

To address the objectives we use five patient series; two multi-center series of patients included in randomized controlled trials (RCTs) in TBI, two relatively unselected multi-center series of patients and one patient series collected in a single center in the Netherlands (Table 4).

Multi-center trials:

- The Tirilazad trials consist of patients included in the International and the North American multi-center (phase III) RCTs on the drug Tirilazad Mesylate[22] in TBI[23]. The International Tirilazad trial (n=1120) was conducted in 40 centers in Europe, Israel and Australia from 1992 to 1994 and the North American Tirilazad trial (n=1149) in 36 centers in the USA and Canada from 1991 to 1994. Both trials enrolled patients aged 15-65 years, with a severe or moderate closed TBI. Patients with an absent motor score or with a moderate TBI and a normal CT scan (Diffuse Injury I) were excluded. All patients were admitted to a neurosurgical unit within four hours after injury. Recommendations for patient management were similar across the centers.
- The Selfotel trial consists of patients included in the International Selfotel trial (n=427), a phase III RCT, investigating the competitive NMDA-glutamate antagonist Selfotel[24]. The trial was conducted in Europe, Canada, Australia and Argentina between 1994 and 1995. Enrollment criteria were: age 16 – 65 years, closed TBI with a GCS of 4 to 8, presence of abnormalities on the CT scan and at least one reactive pupil. Recommendations for patient management were similar across the centers.

Relatively unselected multi-center series of patients:

- The survey by the European Brain Injury Consortium (EBIC) included 796 patients with severe or moderate TBI[25], consecutively collected between February and April 1995 from 67 European centers in which the six months outcome (GOS) assessment was routinely performed. Patients were admitted within 24 hours after the injury.
- The National Traumatic Coma Data Bank (TCDB) contains data on 746 patients with closed severe TBI admitted to four centers in the United States[26]. Data acquisition occurred from 1984 to 1988. Patients were included if they deteriorated to a condition meeting enrollment criteria within 48 hours after the injury.

Single-center series of patients:

- The ErasmusMC cohort consists of 275 patients with mainly severe or moderate TBI, admitted to the trauma center of the Erasmus MC, Rotterdam, the Netherlands, between 2000 and 2003.

1.9 Outline of this thesis

Chapter 2 addresses objective 1. In this chapter we review earlier developed prognostic models in TBI in order to gain insight into methodological developments in prognostic modeling in TBI. This insight is used as a background to our modeling efforts. Furthermore, we propose guidelines to develop and validate prognostic models in TBI.

The **chapters 3 till 7** describe five studies to predict long-term mortality and unfavorable outcome (objective 2), using baseline demographic, clinical and CT characteristics. The patients included in the Tirilazad trials were central to address this research question. Chapter 3 studies the characteristics of the Tirilazad patients. In this chapter we also study regional differences in patient characteristics, case management and outcome, which may cause variation in outcome prediction, independent of demographical and clinical patient characteristics. In chapter 4 and 5 the optimal way to include well-known and important predictors of outcome are studied. Chapter 4 examines the predictor age and chapter 5 studies the prognostic performance of the Marshall CT-classification in comparison with alternative easily applicable classification including CT characteristics. The purpose of chapter 6 was to develop prognostic models that estimates six-month outcome after severe or moderate TBI. These models are validated internally and externally on several other TBI cohorts. Chapter 7 describes the external validity of prognostic models considered in chapter 2 and 5. We relate the external validity of the models developed by us to the external validity of four previously developed models that used baseline demographic, clinical and CT characteristics to predict outcome at six months or later after severe or moderate TBI.

Chapter 8 describes an explorative study for characteristics that predict the need for specialized intensive care (objective 3). This need was expressed in the risk of potentially operable lesions and the risk on raised intracranial pressure (ICP). We developed prognostic models to estimate these risks.

In **chapter 9**, the main findings of the preceding chapters are summarized and discussed. Finally, conclusions and recommendations for further research are given.

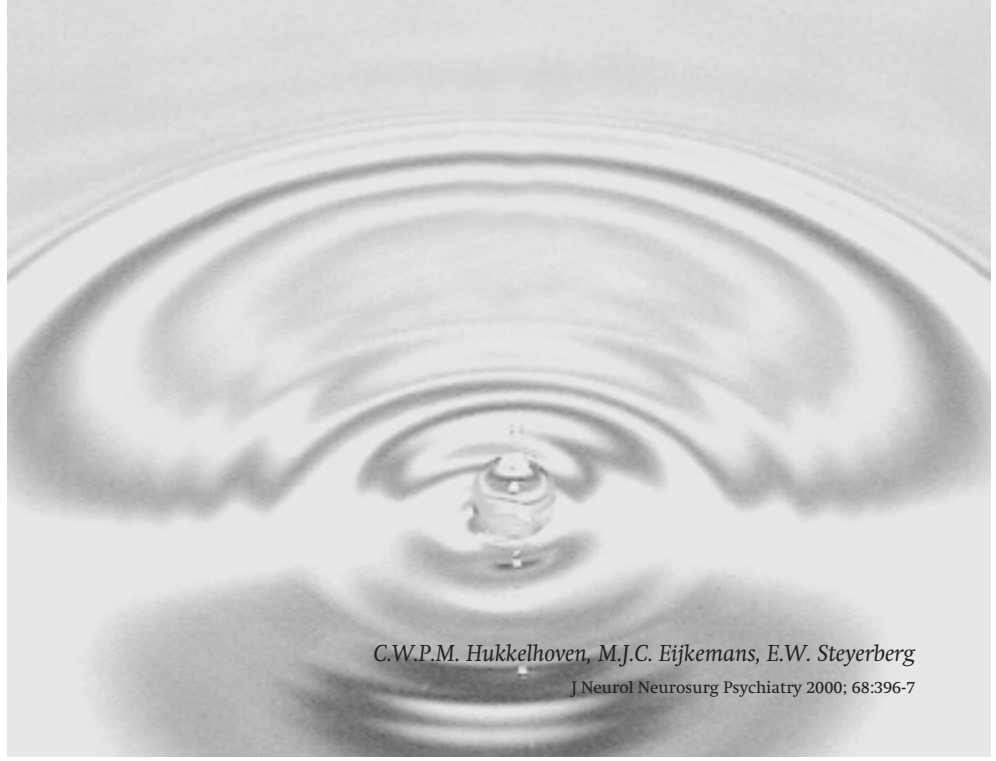
References

1. Ghajar J. Traumatic brain injury. Lancet 2000;356(9233):923-9.
2. Minderhoud JM. Traumatische hersenletsels. Houten: Stafleu Van Loghum, 2003.
3. Thurman DJ, Alverson C, Dunn KA, Guerrero J, Sniezek JE. Traumatic brain injury in the United States: A public health perspective. J Head Trauma Rehabil 1999;14(6):602-15.
4. Langlois J, Rutland-Brown W, Thomas K. Traumatic Brain Injury in the United States: Emergency Department Visits, Hospitalizations, and Deaths. Atlanta (GA): Centers for Disease Control and Prevention, National Center for Injury Prevention and Control, 2004.
5. Dutch Injury Surveillance System (LIS). Database of 1999. Amsterdam: Consumer Safety Institute, 1999.
6. National Database of Hospitalized Patients (LMR). Database of admissions and operations. Utrecht: Prismant, 1999.
7. Statistics Netherlands. Cause of Death Statistics. Voorburg/Heerlen, 2002.
8. Thurman D. The epidemiology and economics of head trauma. In: Miller L, Hayes R, eds. Head Trauma: Basic, Preclinical, and Clinical Directions. New York: Wiley and Sons, 2001.
9. Teasdale G, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974;2(7872):81-4.
10. Kraus JF, Black MA, Hessol N, et al. The incidence of acute brain injury and serious impairment in a defined population. Am J Epidemiol 1984;119(2):186-201.
11. Marshall LF, Bowers Marshall S, Klauber MR, et al. A new classification of head injury based on computerized tomography. J Neurosurg 1991;75:S14-S20.
12. Jennett B, Bond M. Assessment of outcome after severe brain damage. Lancet 1975;1(7905):480-4.
13. Choi SC, Barnes TY, Bullock R, Germanson TA, Marmarou A, Young HF. Temporal profile of outcomes in severe head injury. J Neurosurg 1994;81(2):169-73.
14. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 1998;52(1-3):289-303.
15. Choi SC. Sample size in clinical trials with dichotomous endpoints: use of covariables. J Biopharm Stat 1998;8(3):367-75.
16. 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(5):454-60.
17. Maas AIR, Steyerberg EW, Murray GD, et al. Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 1999;44(6):1286-98.
18. 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;16(12):1131-8.
19. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am Heart J 2000;139(5):745-51.
20. 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;21(1):45-56.
21. Harrell FJ. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer, 2001.

22. Hall ED, Yonkers PA, Andrus PK, Cox JW, Anderson DK. Biochemistry and pharmacology of lipid antioxidants in acute brain and spinal cord injury. J Neurotrauma 1992;9 Suppl 2:S425-42.
23. Marshall LF, Maas AIR, Marshall SB, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 1998;89(4):519-25.
24. Morris GF, Bullock R, Marshall SB, Marmarou A, Maas A, Marshall LF. Failure of the competitive N-methyl-D-aspartate antagonist Selfotel (CGS 19755) in the treatment of severe head injury: results of two phase III clinical trials. The Selfotel Investigators. J Neurosurg 1999;91(5):737-43.
25. Murray GD, Teasdale GM, Braakman R, et al. The European Brain Injury Consortium survey of head injuries. Acta Neurochir (Wien) 1999;141(3):223-36.
26. Marshall LF, Becker DP, Bowers SA, et al. The National Traumatic Coma Data Bank. Part 1: Design, purpose, goals, and results. J Neurosurg 1983;59(2):276-84.

2

Prognostic models in traumatic brain injury: a systematic review of methodological developments and a proposal for guidelines



C.W.P.M. Hukkelhoven, M.J.C. Eijkemans, E.W. Steyerberg

J Neurol Neurosurg Psychiatry 2000; 68:396-7

Abstract

Context
Early prediction of outcome after traumatic brain injury (TBI) is important for several purposes. We aimed to review previous TBI studies to gain insight into methodological developments in prognostic modeling in TBI.

Methods
We searched the National Library of Medicine's PubMed database for relevant publications from 1966 till July 2004. Studies were selected if they developed a prognostic model with the purpose to be used in clinical practice to predict the Glasgow Outcome Score (GOS) – or a categorization of it – for patients with severe or moderate TBI.

Results
We selected 26 papers. Twenty-two studies developed models on relatively small patient series ($n < 500$) or on patient series originating from one single place or region. The type and number of candidate predictors varied considerably, with mostly 8 to 20 candidate predictors per model. Most studies used a stepwise selection procedure with a p-value < 0.05 to select predictors, which resulted in a median of 4 selected predictors. The most frequently used statistical technique was (logistic) regression analysis. The handling of missing values was often not reported (18 studies) or, if reported, patients with missing values were omitted from the study population. For most models only the apparent performance was examined. Seven studies assessed internal validity, usually by splitting the development dataset in a derivation- and test-set. External validation occurred by computer simulation (1 study) or on more recent patients from the same setting (3 studies). No models were validated on patients from another place.

Conclusion
This systematic review highlighted limitations in the development and validation of models that predict outcome after TBI. Adherence to modern insights in prognostic research is required to improve the validity of outcome predictions. Guidelines for developing and validating prognostic models are proposed, which if followed should contribute to improve the quality of prognostic modeling and validation in future studies.

Introduction

Outcome prediction after TBI is important for communication with relatives and caregivers, for resource allocation and clinical decision-making, and may further be used for classifying patients according to prognostic risk and for stratification in clinical trials. Outcome prediction is usually not accurately possible with one single risk factor. Multiple risk factors need to be considered jointly, preferably in a prognostic model. Such models are usually created by multivariable analysis, including several risk factors (predictors), such as age or Glasgow Coma Scale (GCS)[1]. By filling in these factors for a specific patient, the model can provide an individual probability of a certain outcome, e.g. the likelihood that the patient will die during the next six months.

Predicting outcome after TBI has been the object of intensive research. In 1976 Jennett et al.[2] explored the possibility of predicting outcome in individual TBI patients by using a prognostic model, and many others followed them. Although their work moved this field towards a more realistic statistical base and away from clinical guesswork, prognostic models in TBI have not had a widespread impact on the clinical management of TBI patients. Partly, this may be caused by the fact that clinicians are not familiar with prognostic modeling, but – probably more important – the validity of prognostic TBI models has not been demonstrated clearly. Clinicians may also feel reluctant to use outcome predictions for clinical decision-making, e.g. treatment limiting decisions.

In this study, we systematically review previous TBI studies to gain insight into methodological developments in prognostic modeling in TBI. We also evaluate the quality of the used development and validation methods by comparison with recently developed methodological standards. We conclude with a proposal for guidelines on sensible development and validation strategies for prognostic modeling in TBI.

Methods

We searched for relevant medical publications, using the National Library of Medicine's PubMed to access the MEDLINE database (period: 1966 – 2004). Tracking citations in the reference lists of the selected papers retrieved additional papers. For PubMed the following key words were used: 'prognosis' or 'prediction', 'model' or 'prediction rule', 'mortality' or 'outcome' or 'Glasgow Outcome Score'(GOS) and 'head injury' or 'traumatic brain injury'. Studies were selected if they complied with the following criteria:

1) a prognostic model was developed, using multiple clinical characteristics to predict outcome of patients with severe or moderate TBI;
2) the model predicted GOS or a (dichotomized) categorization of it, such as mortality;
3) the model was developed with the purpose to be used in clinical practice;
4) the study was published in the English language.

A data extraction form was designed to collect the required information. We focused on five subjects, i.e. study population (number of patients, inclusion criteria, period data collection, place data collection), predictors (number and type of candidate predictors, number and type of selected predictors, coding of the predictors, time-point assessment predictors), outcome (coding of the GOS, time-point assessment outcome), model development (type of model, way of selecting predictors, handling missing predictor values, used performance measures) and model validation (type of validation(s), number and origin of patients used for external validation). One reviewer (C.H.) extracted all the data. For articles reporting more than one prognostic model, preference was given to the most extensive model that fulfilled the criteria mentioned above.

Statistical terms, relevant for prognostic modeling and used in this study, are explained in the **Glossary**.

Results

Of more than 100 studies initially considered, 26[2-27] were selected for detailed assessment of development and validation aspects. Below, we present the methodology of these 26 studies according to five subjects, i.e. patient series used for developing the model, (candidate) predictors, predicted outcome measure, statistical techniques used to develop the model and validation aspects. In several studies more than one model was presented.

Study population

Many studies (11 of the 26) developed models on patient data collected more than 20 years ago (end of data collection < 1985) (Table 1). Twenty-two of the 26 publications developed models on relatively small patient series (< 500 patients). The largest study population consisted of 799 patients[20]. Furthermore, study populations often originated from one single place or region; only 4 patient samples originated from multiple centers and only 2 from multiple countries. The study populations always consisted of patient series, collected in hospitals (cohorts), but the inclusion criteria varied considerably. For instance, Quigley et al.[21] selected only patients with a GCS ≤ 5, while Andrews et al.[26] selected all patients with a GCS ≤ 12, or a GCS > 12, provided that the Injury Severity Score was above 15.

Predictors

The number of candidate predictors for inclusion into a prognostic model varied considerably among the models and could be very large; up to 82 candidate predictors were considered[20]. Mostly, the number of candidate predictors varied between 8 and 21.
Also the number of chosen predictors varied considerably among the previously developed models (Table 2). For several models, such as the models developed by Narayan et al.[7] and Schreiber et al.[25], selection was based on univariable p-values of the predictors. For other models all candidate predictors were included into the model without further selection, which sometimes resulted in many included predictors, e.g. 12 predictors in the model by Stablein et al.[5]. Most studies (n=16) used a forward or backward stepwise selection procedure to select predictors. During this procedure a p-value of 0.05 or, incidentally, a p-value of 0.10 (model by Combes et al.[19]) was used. The median number of selected predictors was four. Some model developers included more complex characteristics. For instance, Narayan et al.[7] included motor evoked potentials.

Several predictors were included in nearly all models (Table 3). Age is one of them; 22 studies included this predictor. Other predictors that were often included are the GCS, or one or more of its components (all studies), pupillary reactivity (17 studies) and computerized tomography (CT) scan abnormalities (17 studies). The coding of predictors varied among studies. For instance, Stewart et al.[3] included age as a dichotomous variable (below or above 55 years) while Signorini et al.[23] included age as a piecewise transformation, that is, no effect of age until 50 years and a continuous linear effect above this age.

– continues on page 30 –

Table 1. Study population, outcome and validation population, used in the previously developed prognostic models

First author		Study population					Outcome		Validation population		
	Year of publication	Period data collection	Participating countries	No of participating centres	Inclusion criteria	N development	Outcome measure	Time-point assement outcome	N internal validation	N external validation	Type and origin external validation
Jennett	1976	1968 - 1976	UK, NL	3	Severe TBI	400	Several categorizations of GOS	6 months	200	-	-
Steward	1979	1965 - 1972	USA	1	TBI	352	Mortality	<7 days - >5 years	-	-	-
Braakman	1980	1973 - 1978	NL	2	GCS ≤ 8	Six models: n=177 till 305	Survival and several categorizations of GOS	Admission - 28 days	-	-	-
Stablein	1980	Not reported	USA	1	Severe TBI	115	GR+MD+SD/V+D	3 months	-	-	-
Young	1981	Not reported	USA	Nr	TBI	170	Unfav	1 year	-	-	-
Narayan	1981	1976 - 1979	USA	1	Unclear	100	Unfav	Unclear	-	-	-
Choi	1983	1976 - 1981	USA	1	Severe TBI	264	Unfav	6 months	-	Not reported	Computer simulation
Lokkeberg	1984	1981	USA	3	GCS **≤ 8**	254	GOS	Discharge	-	-	-
Williams	1984	Not reported	USA	1	TBI	**96**	GOS	Discharge	-	-	-
Born	1985	1978 - 1982	Belgium	1	GCS **≤ 7**	107	Mortality	6 months	-	-	-
Choi	1988	1976 - 1986	USA	1	GCS **≤ 7**	349	GR/ MD/SD/V+D	6 months	174	-	-
Braakman	1988	Before 1980	UK, USA, NL	4	TBI and vegetative after 1 month	140	V+D/SD+MD+GR	1 year	-	-	-
Gibson	1989	1983 - 1987	UK	1	TBI	187	Mortality	Probably discharge	-	52	Same center, collected in 1988
Choi	1991	1976 - 1989	USA	1	GCS **≤ 8**	555	GOS	1 year	-	-	-
Waxman	1991	1985 - 1987	USA	1	GCS **≤ 10**	**306**	GOS	Discharge	-	-	-
Fearnside	1993	Not reported	Australia	1	GCS **≤ 8**	< 315	Mortality; GR+MD/SD+V	6 months	-	-	-
Mamelak	1996	1978 - 1993	USA	1	GCS **≤ 8**	672	D+V/SD +MD+GR	6 months	108	-	-
Combes	1996	1989 - 1992	France	1	GCS **≤ 8**	132	Unfav	48 hours	66	-	-
Lang	1997	1977 - 1992	USA	1	Mainly severe TBI	799	Mortality	6 months	267	-	-
Quigley	1997	1986 - 1991	USA	1	GCS **≤ 5**	375	D+V/SD+MD+GR	At least 6 months. Mean: 2,8 years	-	-	-
Sakellaropoulos	1999	1994 - 1996	Greece	1	TBI	525	GOS	24 hours	75	-	-
Signorini	1999	1989 - 1991	UK	1	GCS **≤ 12 or GCS >** 12 and ISS > 15	372	Survival	1 year	-	520	Same center, collected between 1991 - 1996
Lannoo	2000	1993 - 1996	Belgium	1	GCS **≤ 6**	Two models; n=45 and n=68	Mortality; GR+MD/SD+V	6 months	-	-	-
Schreiber	2002	1994 - 2000	USA	1	Motor score **≤ 5**	213 - 325	Mortality	Discharge	-	-	-
Andrews	2002	1989 - 1991	UK	1	GCS < 13 or (GCS > 12 and ISS > 15)	124	GOS	1 year	10-fold cross validation	-	-
Pillay	2003	1993 - 1998	India	1	GCS **≤ 8 and diffuse** injury but no mass lesion on CT	289	D+V/SD+MD+GR	Development: 1 month; validation: 6 months	-	26	Same center, collected between 1999 -2000

Nr = not reported, GCS = Glasgow Coma Scale, ISS = Injury Severity Scale, GOS = Glasgow Outcome Score, GR = good recovery, MD = moderate disability, SD = severe disability, V = vegetative state, D = death, UK = United Kingdom, NL = the Netherlands, USA = United States of America

Table 2. Overview of the methodology, used to develop the prognostic models

Methodology	Study																								
	Jennett et al., 1976	Steward et al., 1979	Braakman et al., 1980	Stablein et al., 1980	Young et al., 1981	Narayan et al., 1981	Choi et al., 1983	Lokkeberg et al., 1984	Williams et al., 1984	Born et al., 1985	Choi et al., 1988	Braakman et al., 1988	Gibson et al., 1989	Choi et al., 1991	Waxman et al., 1991	Fearnside et al., 1993	Mamelak et al., 1996	Combes et al., 1996	Lang et al., 1997	Quigley et al., 1997	Sakellaropoulos et al., 1999	Signorini et al., 1999	Lannoo et al., 2000	Schreiber et al., 2002	Andrews et al., 2002	Pillay et al., 2003
Predictors																									
No. of candidate predictors	> 8		70			9	21	7	7		21	nr		23	17	nr	5	12	82		8	8	21	23	20	9
No. of selected predictors	nr	5	3-5	12	2	6-7	3	2-7	2	4-5	3	3	7	4	3	2	3	3	4-20	2	7	7	8-14	6	3-8	3
Time-point assessment																									
Injury or admission			x	x	x	x		x	x		x			x	x		x	x	x	x		x				
< 12 hours													x		x		x									
24 hours - 7 days	x	x	x		x	x	x					x							x				x			
> 7 days - 28 days			x				x					x														
Not reported / unclear										x						x					x			x	x	x
Model development																									
Type of model																									
Logistic regression				x	x	x									x		x	x	x			x	x	x		x
Regression		x						x	x																	
Tree model														x											x	
Bayesian approach	x																				x					
Discriminant analysis			x							x	x															
Neural network																			x							
Other																				x						
Not reported							x		x			x	x													
Method for predictor selection																										
Forward or backward stepwise selection			x				x	x	x		x	x			x	x	x	x	x		x	x	x			x
Univariate or multivariate regression						x				x														x	x	
Other														x												
No selection		x		x	x								x							x						
Not reported	x																									
Handling missing values																										
Patient omission								x								x							x	x	x	x
Not reported	x	x	x	x	x	x	x		x	x	x	x	x		x		x	x	x	x	x					
Not applicable														x											x	
Performance measures																										
Accuracy rate	x	x	x	x		x	x			x		x	x	x		x			x	x	x	x	x			x
Sensitivity / Specificity																x	x	x	x				x			x
False positive / False negative						x																				x
Area under ROC-curve																		x				x		x		
Goodness-of-fit																						x		x		
Other			x		x			x	x						x							x				
Not reported					x				x																	

nr = not reported

Table 3. Number and type of the predictors, selected for the prognostic models

Selected predictors	Study																											
	Total	Jennett et al., 1976	Steward et al., 1979	Braakman et al., 1980	Stablein et al., 1980	Young et al., 1981	Narayan et al., 1981	Choi et al., 1983	Lokkeberg et al., 1984	Williams et al., 1984	Born et al., 1985	Choi et al., 1988	Braakman et al., 1988	Gibson et al., 1989	Choi et al., 1991	Waxman et al., 1991	Fearnside et al., 1993	Mamelak et al., 1996	Combes et al., 1996	Lang et al., 997	Quigley et al., 1997	Sakellaropoulos et al., 1999	Signorini et al., 1999	Signorini et al., 1999	Lannoo et al., 2000	Schreiber et al., 2002	Andrews et al., 2002	Pillay et al., 2003
Demographics																												
Age	22	x	x	x	x		x	x	x			x	x	x	x		x	x	x	x	x	x	x	x	x	x	x	
Gender	4				x																		x	x			x	
Cause of injury	4																					x	x	x			x	
Clinical characteristics																												
Hypoxia	1																		x									
Systolic blood pressure	7				x									x		x	x			x						x	x	
Mean arterial pressure	2																					x			x			
Lucid interval or neurological deterioration	2	x	x																									
GCS	17	x	x			x	x	x	x	x				x		x				x	x	x	x	x	x	x	x	
Motor score	13	x		x	x		x				x	x	x		x		x	x	x	x								x
Eye score	6	x		x	x						x							x		x								
Verbal score	4			x	x						x									x								
Pupillary reactivity	17	x		x	x		x			x	x	x	x	x	x			x		x			x	x	x	x	x	
Pupillary size	1				x																							
Brain stem signs	6	x			x		x	x			x																	x
CT scan																												
CT classification or derivate	7	x	x		x		x		x													x					x	
Midline shift	4					x											x									x		x
Brain swelling	1																x											
Traumatic subarachnoid haemorrhage	2																x								x			
Intraventricular blood	2																x								x			
Haematoma*	7	x												x	x		x						x	x	x			
Nr of abnormal CT findings	1															x												
Raised intracranial pressure	2																								x	x		
Concomitant injuries / ISS / AIS#	5													x		x						x	x	x				
No. of other predictors	9		1	1	4		1		3											2		1			7		3	

* More than one type of haematoma (subdural, epidural, intracerebral) can be included
\# ISS = Injury severity scale, AIS = abbreviated injury scale

Additionally, predictors were often measured at different time-points, e.g. before admission to a neurosurgical unit or after three days (Table 2). Fourteen models used predictors that were assessed at the place of injury or at admission to the hospital. For three models predictors were assessed relatively late, i.e. between 7 to 28 days after injury.

Outcome

Inherent to the criteria used for selecting papers for this review, all TBI models predicted mortality, the GOS[28] or a categorization of the GOS. Nineteen studies dichotomized the GOS, although different dichotomizations were used. As shown in table 1 the time of assessing the outcome measure varied from 'discharge' or 48 hours to 24 months post-injury.

Model development

Several statistical techniques were used to develop prognostic TBI models (Table 2). Before 1990 several models have been developed with discriminant analysis. Later on, this technique was replaced by (logistic) regression analysis. Next to (logistic) regression (15 studies), models were also constructed with recursive partitioning (tree) (n = 2). Incidentally, other methods, such as neural networks[20], were used to develop a model.

Most statistical techniques cannot accommodate missing values. Often, the articles did not mention how patients with missing data were handled. If mentioned (6 studies), patients with missing data were omitted, so that the models were developed on patients with known values for all considered predictors (Table 2).

To assess performance, many measures were used. In the past, performance of TBI models has often been expressed in the accuracy rate or − its complement − the error rate (Table 2). Other often-used measures were sensitivity, specificity, false positive rate and false negative rate. The area under the receiver operating characteristic curve (AUC) and, incidentally, a goodness-of-fit test were used to determine performance of more recent models.

Model validation

For most models only the apparent performance was examined. Seven studies assessed internal validity, usually by splitting the development dataset. For instance, Choi et al.[12] used 2/3 of the patients to develop the model, and the left 1/3 of the patients to assess its internal validity. Only the models by Choi et al.[8], Gibson et al.[14], Signorini et al.[23] and Pillai et al.[27] were externally validated. External validation occurred by computer simulation[8] or on more recent patients from the same setting[14,23,27]. No models were validated on patients from another place.

Discussion

Study population

Several models were developed on patient series collected more than 20 years ago. These models may be not useful, since diagnostic and therapeutic management have changed considerably since that time. For instance, the CT scan has taken a central role in current clinical practice. Furthermore, most models were developed on relatively small patient series. Consequently, the precision to quantify a prognostic model was rather small and generalizability of the developed model may be limited. The fact that the study population often originated from one single place or region may also limit generalizability of the model.

Predictors

Overfitting

The number and type of chosen predictors varied considerably among the previously developed models. Five studies included all candidate predictors (Table 3). Since -in the field of TBI- many candidate predictor variables have been suggested, including all of them may be attractive. From a prognostic point of view, however, including many predictors may not be very sensible. Some candidate predictors may be strongly correlated with each other, so that part of them adds little prognostic value. Furthermore, including too many potential predictors may introduce overfitting, such as occurred in the model developed by Signorini et al.[23,29]. At validation of the model by Signorini et al.[23] on a more recent group of patients than the study population, the Hosmer-Lemeshow statistic showed a significant lack of calibration ($p<0.0001$), with an overly pessimistic prediction in the patients with a poor prognosis but also a too optimistic prediction for patients with a better prognosis. This phenomenon is typical for 'overfitting'. Overfitting can be limited by several procedures. One of them is that, as a rough estimate, no more than $m/10$ predictor degrees of freedom (df) should be analyzed to construct a multiple regression model, where m is the number of events (for example, deaths)[30,31]. For instance, in the study population used by Signorini et al.[23] 87 patients died within 1 year. Consequently, less than 87/10 = 8.7 df could be examined in the analysis without risk of overfitting (see **Appendix**).

Definition

Next, it is important that candidate predictors can be defined precisely and that their value is not substantially influenced by any treatment or therapy. Variables as age and cause of injury are unambiguous. A problem with the CT classification, however, is that it differentiates between patients with evacuated and patients with non-evacuated mass lesions. Many have argued that this reflects a clinical decision and does not in itself constitute a CT parameter, and in clinical practice this has led to confusion. Furthermore, TBI patients are more and more frequently paralyzed and intubated, by which especially the verbal- and eye score of the GCS become not interpretable. Consequently, these predictors should preferably not be included in prognostic TBI models.

Predictors were often measured at different time-points (Table 2). This hampers comparison of the performance of the different models, since measurements performed at a later time may be expected to correlate better with the outcome[4]. Which time-point is appropriate depends on the aim of the model.
Furthermore, complex predictors, such as motor evoked potentials[7], may have substantial predictive power, but are not always practical and easily obtainable in emergency situations.

Outcome

All TBI models considered here predict mortality or (a categorization of) the GOS. Both outcome measures are relatively easy to define and measure, and represent clinically relevant endpoints. Mortality is unambiguous, but for the GOS some degree of inter-observer variability has been reported[32]. This inter-observer variability may partially be diminished by using a structured interview during assessment[33].
The time of assessing the outcome measure varied considerably, i.e. from 'discharge' to 24 months post-injury. Early outcome measures are less stable, making several prognostic models clinically less relevant. From 6 months onwards GOS measurements are considered constant[34].

Model development

Statistical techniques
Several statistical techniques were used to develop prognostic TBI models, such as discriminant analysis, logistic regression analysis, neural networks and recursive partitioning.

Compared to discriminant analysis logistic regression analysis requires fewer assumptions in theory (independent variables do not need to be normally distributed, linearly related, or have equal within-group variances), is more statistically robust in practice, handles categorical as well as continuous variables and has coefficients which many find easier to interpret. Next to logistic regression, recently developed models were also constructed with recursive partitioning (2 models) and, occasionally, neural networks (1 model) (Table 2). Although logistic regression, recursive partitioning and neural networks each have their own advantages and drawbacks, their performance is generally similar[35-38].

Selection and inclusion of predictors
Many TBI models that were developed with stepwise selection techniques used a p-value of ≤ 0.05 to select predictors. Such a p-value is stringent and may lead to the exclusion of many -also informative- characteristics, especially in small development samples. Higher p-values, for instance p-values of 0.20 or 0.50, have been advocated to increase the external validity of the model[39].

Also the way predictor variables, e.g. age, are included in the models varied greatly. From a prognostic point of view, coding of predictor variables should be as detailed as possible, since categorization of predictors often leads to loss of information. On the other hand, categorization of predictor variables may diminish complexity of the model and improve implementation in clinical practice.

Missing values
Missing predictor values form a common problem in prognostic studies. They can introduce bias, depending on the missing data mechanism and the adopted missing data approach. In TBI models patients with missing predictor values were mostly omitted, so that the models were developed on complete cases only. Such an approach assumes that the missingness in the predictors is not associated with the outcome[40]. This assumption, however, is often not realistic. Furthermore, patients with missing values usually miss only one or a few of the predictor values, so that dropping incomplete patients leads to a waste of precious information available in the other predictor variables and the outcome[31,41].

Patients with initially missing predictors can be preserved for analysis by considering the missing data as a separate category[42], by single imputation[43], in which a single value is substituted for each missing value, and by multiple imputation[44,45], where several independently completed data sets are obtained.

Performance measures
Before a model can be used in clinical practice, its performance needs to be appropriate. Three aspects of validity can be distinguished; clinical usefulness, calibration and discrimination. Most of the reported performance measures, i.e. accuracy rate, error rate, sensitivity, specificity, false positive rate, false negative rate, determine the clinical usefulness of a model. The accuracy rate and the error rate have been used to assess performance of most, and especially older, TBI models. Currently, the accuracy rate is generally considered inappropriate for validation purposes, because this rate is greatly influenced by the outcome distribution of the patient series. For instance, in a representative TBI patient population with an average mortality of 20 percent, the accuracy rate will already be 80 percent if all patients would be labeled as survivors. In patient series with such a skewed outcome distribution, high accuracy rates are typical. An exception is the model by Lannoo et al.[24]; this model reported a high accuracy rate (93%) while the mortality rate was 51%.

Nowadays, the performance of prognostic models is often expressed with respect to calibration and discrimination. In contrast to measures of clinical usefulness, calibration and discrimination measures evaluate the performance of a prediction model over the whole range of predicted probabilities.
Although the performance of most models was sufficient to high, this performance was often assessed only at apparent validation, sometimes even on part of the study population.

Model validation

The purpose of a prognostic model is to provide valid outcome predictions to new patients. Therefore, it is essential that the validity of a model is assessed. Before a model can be implemented safely in clinical practice, all different types of validation – apparent, internal and especially external validation – need to be satisfactory.

For most models only the apparent validation was examined (Table 1). The major drawback of assessing only this type of validation, however, is that the assessment will always be optimistic, since both development (estimation of regression coefficients, selection of predictors) and testing are performed on the same patients.

For few models the internal validity was assessed, usually by splitting the dataset. The major disadvantage of data splitting, however, is that only part of the patients is used for model development and only part for model validation. Statistically more efficient techniques, that uses all available patients for model development while also the internal validity can be assessed, are re-sampling methods, such as bootstrapping[31,39,46].

External validation of models is essential to support general applicability[47]. When simply tested on the development patients, the (apparent) performance may be excellent. However, when testing on new patients the performance of the model may be considerably poorer[48]. Surprisingly, only four model developers[8,14,23,27] performed an external validation of their model. External validation occurred, however, only on more recent patients from the same setting, while – besides time aspects – also place aspects may affect the external validity of a prognostic model. For instance, other medical centers may treat patients with different characteristics, such as more severe patients. A model that is valid for patients from one single place (specific 'case-mix') is not automatically valid for patients from another place. Therefore, a model should preferably be repeatedly validated on various patient series, differing in time and place.

Quality of used development and validation methods in TBI modeling

This review has highlighted important limitations in the methods used to develop and validate previous prognostic models in TBI. Many studies were limited by old, small and relatively homogeneous study populations and the way predictors were obtained and selected. Furthermore, rather crude statistical methods were used. In the majority of the studies the extent of missing predictor values was not reported and, if reported, patients with missing values were omitted. Additionally, validation was often only performed on the study population itself and the performance was only seldom tested on new patients from another time and not on patients from another place.

Partly the methodological deficiencies can be explained by the fact that most models have been developed many years ago. Especially in the last decade knowledge about methodological aspects of prognostic modeling, such as bootstrapping, has considerably increased[30,31].

Before a prognostic model is widely accepted and applied clinically, we believe attention should be paid to a number of checkpoints. In table 4 we propose guidelines for good prognostic modeling and validation. These guidelines are based on the findings described in this article, but also on other modeling studies[30,49-54] and on practical considerations.

In conclusion, it is evident that quality in the development and validation methods used for prognostic modeling in TBI can be improved, especially with respect to procedures for predictor selection, dealing with missing predictor values, and validation. The proposed guidelines for developing and validating a prognostic model, if followed, should contribute to improve the quality of prognostic models and valid outcome predictions in future patients.

Table 4. Guidelines for developing and validating prognostic models in TBI

Study population	• large and well-defined cohort of consecutive patients • heterogeneous, i.e. including a clinically broad spectrum of patients from multiple centers • representative for current clinical practice
Predictors	• plausible, based on previous studies and expert opinion • precisely defined (in order to minimize inter-observer variability) • readily available or easily obtainable
Outcome	• relevant for clinical practice • precisely defined • measurable with minor observer variability
Model development	• use of appropriate statistical techniques to model prediction-outcome relationships and to deal with missing predictor values • use of sensible performance measures, evaluating calibration and discrimination aspects • presentation in a readily applicable format
Model validation	• internal validation • external validation on patients managed by different protocols or at different times and places

Appendix

Letter regarding the model developed by Signorini et al.[23]; Predicting survival using simple clinical variables: a case study in traumatic brain injury

Signorini et al.[23] developed a prognostic model to predict survival at 1 year for patients with traumatic brain injury. A strong point is that this model uses variables, which are easy and cheap to measure. A thorough statistical analysis was performed, including tests for goodness-of-fit and checks for influential observations. The model was also validated externally in a more recent group of patients. However, during the external validation the Hosmer-Lemeshow statistic showed a significant lack of calibration ($p < 0.0001$). This implies that the model does not give accurate predictions of the survival of 'new' patients. The lack of calibration is especially due to an overly pessimistic prediction in the patients with a poor prognosis but also to a too optimistic prediction for patients with a better prognosis (Figure 2)[23]. This is typical for 'overfitting'—that is, that a model tends to predict too extreme probabilities in new patients. Overfitting can be limited by several procedures. One of them is that, as a rough estimate, no more than $m/10$ predictor degrees of freedom (df) should be analyzed to construct a multiple regression model, where m is the number of events (for example, deaths)[30]. As 87 patients died within 1 year, 87/10 = 8.7 df could be examined during the course of analysis without risk of overfitting. In the paper 6 df were used by the final multivariate prognostic model. However, age was fitted as a piecewise linear variable after using a generalized additive model, requiring an unknown number of df, but always more than one. Furthermore, we assume that easy to achieve variables such as sex (1 df) and cause of injury (3 df) were considered but dropped during model construction. Also some of the candidate variables originate from combined variables when, after initial assessment, it seemed that some categories could be collapsed. Altogether this means that probably much more than 8.7 df were examined. The overfitting could have been corrected by multiplying each regression coefficient in the model with a shrinkage factor. This factor can be estimated by a heuristic formula[55], by cross validation, or by a bootstrap re-sampling procedure. This can be done with the Design library[56], which was already used by the authors. The shrinkage factor is close to unity when there is no overfitting. When the selection of predictors is unstable or predictors have small effects, a lower shrinkage factor might be found— for example, 0.8. We regret that the model is presented as giving 'reasonable accurate predictions of long term survival', especially because the external validation showed a significant lack of calibration. Correction with a shrinkage factor would have resulted in a recalibration of the probability of survival in the nomogram presented in the paper (Figure 3)[23] and in the formula used in a subsequent paper[57]. We hope that modern modeling techniques will increasingly be applied in clinical prediction problems such as traumatic brain injury, such that prognostic models are developed that reliably support the physician in clinical decision-making.

C.W.P.M. Hukkelhoven
M.J.C. Eijkemans
E.W. Steyerberg

Glossary

Accuracy rate and error rate
The accuracy rate is defined as the proportion of patients with a certain outcome that was predicted correctly, whereas the error rate is equal to 1 minus the accuracy rate.

Area under the receiver operating curve
The area under the receiver-operating curve (AUC) is often used to quantify the discriminative ability of a prognostic model, developed with logistic regression analysis. The receiver-operating curve is a plot of the sensitivity versus the false positive rate (or 1-specificity), evaluated at consecutive threshold values of the predicted probability. The AUC evaluates whether those patients with higher predicted risk are more likely to have a poor outcome (mortality/unfavorable outcome) among all possible pairs of patients with different outcomes. A model with an AUC of 0.50 has no discriminative power at all (such as a coin flip), and an AUC of 1.0 reflects perfect discrimination[30,31].

Bootstrap validation
Bootstrap validation involves drawing samples of patients with replacement from the development sample. 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. Regression models were estimated in each of the bootstrap samples and evaluated on the original sample. The average difference in performance indicated the optimism (overfitting)[30,31]. Subsequently, the coefficients can be corrected (shrunk) for predictive purposes. In this way, nearly unbiased predictions of the outcome can be obtained for future but similar patients[30,31,39].

Calibration
Calibration (or reliability) refers to the agreement between the observed outcome frequencies in the data and the predicted probabilities of the model. For example, if a group of patients (with certain characteristics) are predicted to have a 10% chance of mortality, the actually observed mortality of this group should on average be 10%.

Clinical usefulness
Clinical usefulness refers to the ability of the model to improve the decision making process. For instance, determining which patients have such a high risk of developing raised intracranial pressure or a potentially operable lesion that they need to be admitted to a specialized neurosurgical trauma center.

Discriminant analysis
Discriminant analysis is a technique for classifying a set of observations, e.g. a patient population, into predefined groups (classes). The purpose is to determine the class of an observation (patient) based on a set of characteristics (predictors). The basic idea underlying discriminant analysis is to determine whether groups differ with regard to the mean(s) of the predictor(s), and then to use these predictors to prognosticate group membership[58]. Accordingly, an attempt is made to maximize between group variance while minimizing within group variance.

Discrimination
Discrimination refers to the model's ability to separate patients with different outcomes. A good discriminating model for mortality will predict high probabilities for patients who die and low probabilities for patients who survive. Discrimination is often quantified by the area under the receiver operating characteristic curve (AUC).

False positive rate and false negative rate, sensitivity and specificity
The false positive rate is the proportion of patients who are actually negative (e.g. survive), but who are classified by the model as positive (e.g. outcome death), whereas the false negative rate is the proportion of patients who are who are actually positive, but predicted as negative (Figure 1).
Sensitivity refers to the proportion of patients with the outcome (e.g. favorable outcome) who are correctly classified by the model as having the outcome. Specificity refers to the proportion of patients without the outcome who are correctly classified by the model as not having the outcome (Figure 1).

Figure 1. Calculation of false positive rate, false negative rate, sensitivity and specificity

		Outcome	
		Yes	No
Outcome Prediction	Yes	a	b
	No	c	d

False positive = b/ (b + d)
False negative = c/ (a + c)
Sensitivity = a/ (a + c)
Specificity = d/ (b + d)

Goodness-of-fit test
Goodness-of-fit tests are used to test the calibration of a prognostic model. An example of such a test is the Hosmer-Lemeshow goodness-of-fit test for logistic regression analysis[59], which assesses agreement between predicted and observed risks over the full range of predicted probabilities. Patients are often grouped per decile of predicted risk to perform the test, which means that each group contains 10% of the patients.

Imputation, single and multiple
Single and multiple imputation approaches assume that the missingness of data is related to the observed data (the other predictors), but does not depend on unobserved data or values of the predictors itself. A simple method of single imputation is to use the median or mean of the predictor for the missing values, while a more complex method is to estimate the missing values by using regression models including the values of the other predictors. However, if a large proportion of the values is missing, single imputation may overstate the available sample size, leading to an underestimation of the variance and hence too narrow confidence intervals and more significant p-values[45]. Multiple imputation may then be a statistically better approach[44]. On the other hand, if only a small part of the predictors is missing, multiple imputation will give similar results to single imputation.

Logistic regression
Logistic regression is a frequently used statistical method in prognostic research. It relates one or more characteristics (predictors) of patients ($X = \{X_1 \dots X_k\}$) to an outcome (Y) by multiplying the characteristic(s) with regression coefficients ($\beta = \{\beta_i \dots \beta_k\}$). These regression coefficients represent the strength of the association between a patient characteristic and the outcome. The outcome variable is also called the dependent variable and the predictors the independent variables.

In logistic regression analysis the outcome variable is generally dichotomous, that is, the outcome variable can take the value 1 with a probability of outcome p, or the value 0 with a probability of other outcome 1-p.
The relationship between the predictor and outcome variable is defined by the logistic regression function, which is the logit transformation of the probability of the outcome given predictor X (p(X)):

$$p(X) = \frac{e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}{1 + e^{(\alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k)}}$$

where α = the intercept of the model (constant), β_i the array of regression coefficients and X_i the array of patient characteristics (predictors).

An alternative form of the logistic regression equation is:

$$\text{Logit } [p(X)] = \log [p(X)/(1-p(X))] = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k$$

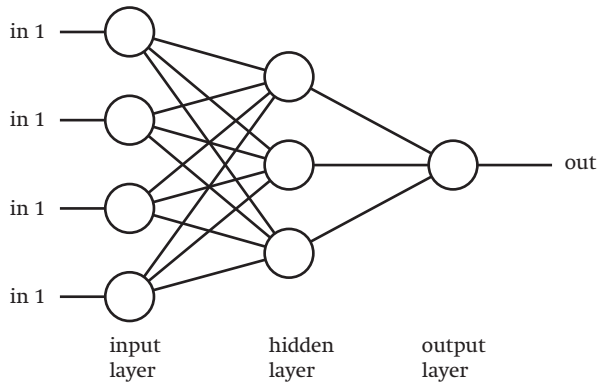
where α, β_i and X_i defined as above.

A logistic regression model is usually estimated with maximum likelihood methods, assuming that the outcome Y follows a binomial distribution.

Neural networks
Artificial neural networks are mathematical structures designed to mimic the information processing functions of a network of neurons in the brain[60,61]. The network is composed of a large number of interconnected units (neurons) working in parallel.

The most common neural network model is the multilayer perceptron (Figure 2). In this type of network, the neurons are arranged in a layered configuration containing an input layer, usually one 'hidden' layer, and an output layer. The values of input variables (e.g. patient characteristics) are imported into the network via the input layer and multiplied with the weights of the connections. These multiplied values constitute the input of the next (hidden) layer, from where the process is continued to produce the output variables (e.g. risk of mortality) in the output layer.

Figure 2. Multilayer perception



A neural network does not use any preliminary information about the links between the input and output variables; the relationships between input and output variables are determined by the data. Neural networks learn by example; the errors from the initial prediction of the first record (e.g. patient) are fed back into the network and the weights are adjusted to minimize the error the second time around, and so on for many iterations. The process from input to output layer is repeated many times. The hidden layer makes the network more flexible by enabling it to recognize more patterns, compared to a logistic regression model. A neural network without a hidden layer and with one output variable is equivalent to a logistic regression model.

Recursive partitioning

Recursive partitioning[62] is a method to construct binary trees. The method is based on statistically optimal splitting (partitioning) of the patients into pairs of smaller subgroups. Splits are based on cut-off levels of the predictors, which produce maximum separation among two subgroups and a minimum variability with these subgroups with respect to the outcome. The predictor causing the largest separation is situated at the top of the tree, followed by the predictor causing the next largest separation, and so on. For instance, in the tree developed by Choi et al.[15], the patients are first split on the basis of their pupillary response; patients with a bilaterally normal response are separated from patients having unilaterally or bilaterally absent responses. Thus, the predictor pupillary response causes largest separation when dividing the population into two subgroups. Splitting continues until the subgroups reach a minimum size or until no improvement can be obtained. As the full tree developed may be too complex it is usually pruned using cross-validation to prevent overfitting. Prediction of outcome for a patient is accomplished by simply running that patient down the prediction tree, according to the values of the predictors.

Overfitting

Overfitting can be defined as the tendency of models to perform very well on the development population, but to predict too extreme probabilities in new patients. Overfitting can be (partly) corrected by multiplying each regression coefficient in the model with a shrinkage factor. This factor can be estimated by a heuristic formula[55], by cross validation, or by a bootstrap resampling procedure. The latter can, for instance, be performed easily with the Design library[56]. The shrinkage factor is close to unity when there is no overfitting.

Stepwise selection

In stepwise selection candidate predictors are added (forward stepwise selection) or deleted (backward stepwise selection) to assess their additional prognostic value. A candidate predictor is usually selected if its additional prognostic value is statistically significant.

Validity

Various types of validity are distinguished;

a) apparent validity, in which model validity is assessed on exactly the same patients as used to develop the model,
b) internal validity, in which model validity is assessed on patients similar to those from the development population, for example by using bootstrapping,
c) external validity, in which model validity is assessed in patients from another setting, time or place.

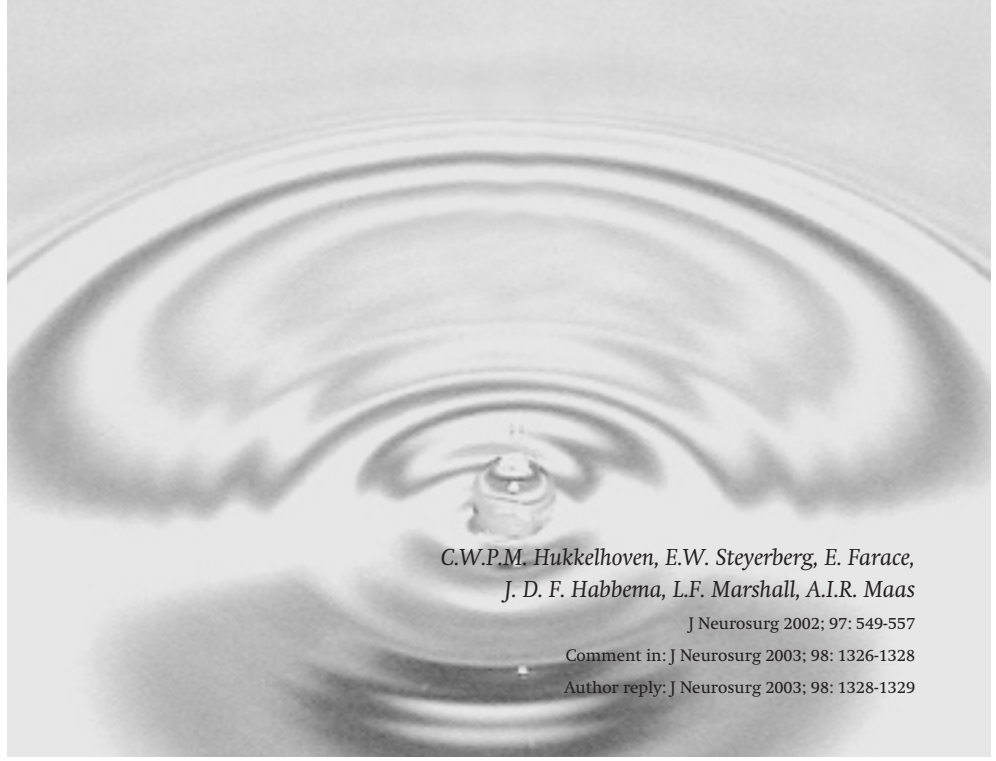
References

1. Teasdale GM, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974;2(7872):81-4.
2. Jennett B, Teasdale GM, Braakman R, Minderhoud J, Knill-Jones R. Predicting outcome in individual patients after severe head injury. Lancet 1976;1(7968):1031-4.
3. Stewart WA, Litten SP, Sheehe PR. A prognostic model for head injury. **Acta Neurochir (Wien)** 1979;45(3-4):199-208.
4. Braakman R, Gelpke GJ, Habbema JDF, Maas AIR, Minderhoud JM. **Systematic selection of** prognostic features in patients with severe head injury. Neurosurgery 1980;6(4):362-70.
5. Stablein DM, Miller JD, Choi SC, Becker DP. Statistical methods for determining prognosis in severe head injury. Neurosurgery 1980;6(3):243-8.
6. Young B, Rapp RP, Norton JA, Haack D, Tibbs PA, Bean JR. Early prediction of outcome in head-injured patients. **J Neurosurg 1981;54(3):300-3.**
7. Narayan RK, Greenberg RP, Miller JD, 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;54(6):751-62.
8. Choi SC, Ward JD, Becker DP. Chart for outcome prediction in severe head injury. **J Neurosurg** 1983;59(2):294-7.
9. Lokkeberg AR, Grimes RM. **Assessing the influence of non-treatment variables in a study of** outcome from severe head injuries. J Neurosurg 1984;61(2):254-62.
10. Williams JM, Gomes F, Drudge OW, Kessler M. Predicting outcome from closed head injury by early assessment of trauma severity. J Neurosurg 1984;61(3):581-5.
11. Born JD, Albert A, Hans P, Bonnal J. Relative prognostic value of best motor response and brain stem reflexes in patients with severe head injury. Neurosurgery 1985;16(5):595-601.
12. Choi SC, Narayan RK, Anderson RL, Ward JD. Enhanced specificity of prognosis in severe head injury. **J Neurosurg 1988;69(3):381-5.**
13. Braakman R, Jennett WB, Minderhoud JM. **Prognosis of the posttraumatic vegetative state. Acta** Neurochir (Wien) 1988;95(1-2):49-52.
14. Gibson RM, Stephenson GC. Aggressive management of severe closed head trauma: time for reappraisal. Lancet 1989;2(8659):369-71.
15. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head-injured patients. J Neurosurg 1991;75(2):251-5.
16. Waxman K, Sundine MJ, Young RF. Is early prediction of outcome in severe head injury possible? Arch Surg 1991;126(10):1237-41; discussion 1242.
17. 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(3):267-79.
18. Mamelak AN, Pitts LH, Damron S. Predicting survival from head trauma 24 hours after injury: a practical method with therapeutic implications. J Trauma 1996;41(1):91-9.
19. Combes P, Fauvage B, Colonna M, Passagia JG, Chirossel JP, Jacquot C. Severe head injuries: an outcome prediction and survival analysis. Intensive Care Med 1996;22(12):1391-5.
20. 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(3):274-80.
21. Quigley MR, Vidovich D, Cantella D, Wilberger JE, Maroon JC, Diamond D. Defining the limits of survivorship after very severe head injury. J Trauma 1997;42(1):7-10.
22. Sakellaropoulos GC, Nikiforidis GC. Development of a Bayesian Network for the prognosis of head injuries using graphical model selection techniques. Methods Inf Med 1999;38(1):37-42.
23. 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;66(1):20-5.**
24. Lannoo E, Van Rietvelde F, Colardyn F, et al. **Early predictors of mortality and morbidity after** severe closed head injury. J Neurotrauma 2000;17(5):403-14.
25. Schreiber MA, Aoki N, Scott BG, Beck JR. Determinants of mortality in patients with severe blunt head injury. **Arch Surg 2002;137(3):285-90.**
26. Andrews PJ, Sleeman DH, Statham PF, 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;97(2):326-36.
27. Pillai SV, Kolluri VR, Praharaj SS. Outcome prediction model for severe diffuse brain injuries: development and evaluation. Neurol India 2003;51(3):345-9.
28. Jennett B, Bond M. Assessment of outcome after severe brain damage. **Lancet 1975;1(7905):480-4.**
29. Hukkelhoven CWPM, Eijkemans MJC, Steyerberg EW. **Predicting survival using simple clinical** variables: a case study in traumatic brain injury. J Neurol Neurosurg Psychiatry 2000;68(3):396-7.
30. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
31. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer, 2001.
32. Maas AIR, Braakman R, Schouten HJ, Minderhoud JM, van Zomeren AH. **Agreement between** physicians on assessment of outcome following severe head injury. J Neurosurg 1983;58(3):321-5.
33. Wilson JT, Pettigrew LE, Teasdale GM. Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J Neurotrauma 1998;15(8):573-85.
34. Choi SC, Barnes TY, Bullock R, Germanson TA, Marmarou A, Young HF. Temporal profile of outcomes in severe head injury. **J Neurosurg 1994;81(2):169-73.**
35. Borque A, Sanz G, Allepuz C, Plaza L, Gil P, Rioja LA. **The use of neural networks and logistic** regression analysis for predicting pathological stage in men undergoing radical prostatectomy: a population based study. J Urol 2001;166(5):1672-8.
36. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 2001;29(2):291-6.
37. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17(21):2501-8.
38. Titterington DM, Murray GD, Murray LS, et al. Comparison of discrimination techniques applied to a compex data set of head injured patients. J Roy Stat Soc, Series A 1981;144:145-175.

39. Steyerberg EW, Eijkemans MJC, Harrell FE, Jr., Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
40. Vach M, Blettner M. Missing data in epidemiological studies. In: Armitage P, Colton T, eds. Encyclopedia of Biostatistics. New York: John Wiley and Sons, 1998: 2641-2654.
41. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227-1237.
42. Greenland S, Finkle W. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142:1255-1264.
43. Little R, Rubin D. Statistical analysis with missing data. New York: John Wiley and Sons, 1987.
44. Schafer J. Analysis of Incomplete Multivariate Data. New York: Chapman&Hall, 1997.
45. Rubin D. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons, 1987.
46. Efron B, Tibshirani R. An Introduction to the Bootstrap. New York: Chapman and Hall, 1993.
47. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(6):515-24.
48. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(4):453-73.
49. Spiegelhalter DJ. Probabilistic prediction in patient management and clinical trials. Stat Med 1986;5(5):421-33.
50. Steyerberg EW, Eijkemans MJC, Harrell FE, Jr., Habbema JDF. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 2001;21(1):45-56.
51. Simon R, Altman DG. Statistical aspects of prognostic factor studies in oncology. Br J Cancer 1994;69(6):979-85.
52. Altman DG, Lyman GH. Methodological challenges in the evaluation of prognostic factors in breast cancer. Breast Cancer Res Treat 1998;52(1-3):289-303.
53. Riley RD, Abrams KR, Sutton AJ, et al. Reporting of prognostic markers: current problems and development of guidelines for evidence-based practice in the future. Br J Cancer 2003;88(8):1191-8.
54. Laupacis A, Sekar N, Stiell IG. Clinical prediction rules. A review and suggested modifications of methodological standards. JAMA 1997;277(6):488-94.
55. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9(11):1303-25.
56. Design: S-plus functions for biostatistical epidemiologic modelling, testing, estimation, validation, graphics, prediction, and typesetting by storing enhanced model design attributes in the fit [program].
57. 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;66(1):26-31.
58. Davis J. Statistics and Data Analysis in Geology. Toronto: John Wiley and Sons, 1986.
59. Hosmer D, Lemeshow S. Applied Logistic Regression. New York: John Wiley & Sons, 1989.
60. Jensen BA. Expert systems - neural networks, Instrument Engineers' Handbook, 3rd ed. Radnor, Pennsylvania: Chilton, 1994.
61. Hinton GE. How neural networks learn from experience. Sci American 1992;267:144-151.
62. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, California: Wadsworth, 1984.

3

Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials



C.W.P.M. Hukkelhoven, E.W. Steyerberg, E. Farace, J. D. F. Habbema, L.F. Marshall, A.I.R. Maas

J Neurosurg 2002; 97: 549-557

Comment in: J Neurosurg 2003; 98: 1326-1328

Author reply: J Neurosurg 2003; 98: 1328-1329

Abstract

Object

Regional differences have been shown in patient characteristics and case management within multiple unselected series of patients suffering from traumatic brain injury (TBI). One might expect that such regional heterogeneity would be small in a more selected population of a randomized clinical trial. The goal of this study was to examine what regional differences in patient characteristics, case management, and outcomes exist between continents and among countries within a patient population included in a randomized clinical trial.

Methods

Data were extracted from two concurrently conducted randomized clinical trials of the drug tirilazad; the designs of these studies were similar. The studies included 1701 patients with severe and 476 patients with moderate TBI. Differences were primarily investigated between studies performed in Europe and North America, but also among European regions and between Canada and the United States. Associations among regions and outcomes (6-month mortality rates and Glasgow Outcome Scale scores) were studied using multivariable logistic regression analysis.

Results

Comparisons between continents and among regions within Europe showed differences in the distribution of patient ages, causes of injury, and several clinical characteristics (motor score, pupillary reactivity, hypoxia, hypotension, intracranial pressure [ICP]), and findings on computerized tomography scans. Secondary referrals occurred 2.5 times more frequently in Europe. Within Europe secondary referral was mainly associated with an increased proportion of patients with mass lesions (46% in the European Study compared with 40% in the North American Study). Therapy for lowering ICP was more frequently applied in North America. After adjustments for case-mix and management, mortality and unfavorable outcomes were significantly higher in Europe (odds ratios = 1.58 and 1.46, respectively). Significant differences in outcome between regions within Europe or within North America were not observed.

Conclusions

Despite the use of a strict study protocol, considerable differences in patient characteristics and case management exist between continents and among countries, reflecting variations in social, cultural, and organizational aspects. Outcomes of TBI may be worse in Europe compared with North America, but this finding requires further study.

Introduction

Traumatic brain injury is a heterogeneous disease that affects patients with wide ranges of clinical severity and varying clinical symptoms. General surveys[1-3] and studies on unselected series[4-6] have reported that differences in patient characteristics, demographics and management are partly caused by regional factors.

Within more selected series, such as a randomized clinical trial, the population is more clearly defined by inclusion and exclusion criteria. Recommendations for basic case management and therapeutic approaches serve to minimize treatment variability. These factors may be expected to result in a more homogeneous patient population, in which regional differences in patient characteristics, management, and outcomes are diminished. Nevertheless, heterogeneity of the population has been proposed as one factor contributing to difficulties in demonstrating the efficacy of new therapeutic agents or approaches in the field of TBI[7,8].

The purpose of the present study was to examine the contribution of regional factors to the possible heterogeneity of a patient population included in a randomized clinical trial. We studied regional differences in patient characteristics (demographics and clinical characteristics), causes of injury, case management (referral policy and treatment), and outcomes in 2177 patients included in the tirilazad trials conducted in patients with TBI[9]. Because these two trials were conducted within the same period of time (1991 – 1994) and according to largely similar protocols, they offer the opportunity to study differences between and within North America and Europe without the interference of external circumstances, such as variation in recruitment criteria, definitions of variables, or changes in case management over time.

Clinical material and methods

Patient population

All patients included in this study had been enrolled in one of two multicenter prospective phase III randomized clinical trials on the use of the drug tirilazad for TBI. One trial was conducted in North America from 1991 to 1994 (NAS) and the other in Europe, Israel, and Australia from 1992 to 1994 (ES). The inclusion and exclusion criteria were virtually identical. In both trials the patients who were enrolled were 15 to 65 years of age and had endured a severe (Glasgow Coma Scale [GCS] 4-8) or moderate (GCS 9-12) closed TBI. The proportion of patients with moderate TBIs was 15% in the study comprising Europe, Israel, and Australia, whereas this proportion was 30% in the North American trial. These proportions were attributed to requirements in the protocol and therefore were not a result of regional variation in the patient population. In the former study patients suffering from moderate TBI, in whom the computerized tomography (CT) scan revealed normal findings, were excluded, whereas the NAS patients, in whom the CT scan yielded a normal finding and the GCS motor score was 5 or 6, were only excluded if their blood alcohol level exceeded 0.2 g/dl. All patients were admitted to the neurosurgical trauma center within four hours after injury. Recommendations for case management, especially those for intracranial pressure (ICP)-lowering therapy, were similar across trials. Participating centers

included neurosurgical trauma centers involved in teaching and research related to trauma care. In both trials the efficacy of tirilazad mesylate, an aminosteroid medication that displays an anti-oxidant effect[10], was studied against that of placebo. Details on the international tirilazad trial have been reported[9].

In neither trial was a significant difference between the tirilazad- and placebo-treated groups shown for the primary outcome measure (Glasgow Outcome Scale [GOS][11] score 6 months after injury). Therefore, data from both treatment groups were combined in the present analysis.

Data extraction

Data pertaining to the first 24 hours after injury were extracted on the following items: patient demographics (age and sex); causes of injury; clinical characteristics (GCS score, hypoxia, hypotension, and pupillary reactivity – all determined before or at admission to the neurosurgical trauma center – and maximum ICP within the first 24 hours after injury. [Pupillary reactivity was differentiated as 'both pupils reactive', 'one pupil reactive' or 'no pupillary reactivity']); characteristic appearance of the CT scan (CT scan classification[12], status of basal cisterns, and presence of a traumatic subarachnoid hemorrhage (SAH), mass lesion or a midline shift. [Because observer variation may occur in distinguishing between the CT scan classes of 'mass lesion evacuated' and 'mass lesion not evacuated', we combined both categories into the classification 'mass lesion']; referral policy and early case management (use of ICP-lowering therapy and time intervals between injury and referral to the study center, acquisition of a CT scan, or performing a surgery); the number of patients included in each center; and outcomes, that is, mortality and unfavorable outcome (severe disability, vegetative state, or death according to the GOS) 6 months after injury. A GOS measurement between 5 and 7 months after injury was considered to be a measurement at 6 months. For 16% of North American patients and 6.7% of European patients, the GOS was measured outside the time interval of 5 to 7 months after injury. For these patients outcome was imputed according to a specific algorithm, in which GOS values at other points in time were considered. This algorithm reduced the frequency of missing GOS scores to 9.6% in the North American population and 3.4% in the European population. The detailed algorithm is presented in the **Appendix.**

Regional differences

Regional differences in patient demographics, clinical characteristics, referral policy, and case management approach were studied between Europe and North America for both severe and moderate TBI. Patients enrolled in Australia and Israel were not included in the present analysis of intercontinental differences because of the relatively low numbers (64 patients in Australia and 28 in Israel). For analysis of differences among regions within Europe, the countries were grouped together by geographic location if the numbers of patients in individual countries were judged insufficient for a country-based analysis. Patients enrolled in Switzerland (23 patients) were excluded from this analysis because the policies in that country governing patient referral and treatment in cases of trauma as well as geographic circumstances precluded meaningful grouping with another country. Within North America, differences were studied between Canada and the United States (US). Regional comparisons of separate case management approaches for

patients suffering from severe and moderate TBI were judged not to be meaningful because of small sample sizes.

Statistical analysis

For a comparison of continuous variables between continents, we used Student t-tests and Wilcoxon tests; to compare differences among regions within Europe we applied analysis of variance or Kruskal-Wallis tests. For categorical variables, frequencies were analyzed in contingency tables with chi-square statistics. Probability values lower than 0.05 were considered statistically significant. Regional differences in the quality of bare follow-up data may lead to variations in the exclusion of patients who died after a certain period of time. To check this, the time interval between injury and death was compared between both continents. No difference was observed.

Differences in outcomes at 6 months between the two continents and between the two countries within North America were studied using uni- and multivariable logistic regression analysis, with adjustment for variation in case mix (dissimilarity between patients in different regions, such as possible differences in age, motor score, pupillary reactivity, CT scan classification, hypoxia, hypotension, and proportion of patients with moderate TBI) and case management (time interval between injury and admission to study center, and application of sedation, paralysis, and ICP-lowering therapy). Because center volume may be associated with patient outcome, we also adjusted for the number of patients included by each center. Values of missing confounders (2.6% of the required values) were imputed per patient, based on the values of the nonmissing confounders[13,14]. We used computer software for this analysis, i.e. Design and Hmisc library from S-plus (Version 2000; MathSoft Inc., Seattle, WA). The associations among regions within Europe and outcome 6 months after injury were tested with a likelihood ratio test. The robustness of the logistic regression analyses was confirmed by performing two sensitivity analyses; 1) repetition of regression analyses in which patients whose primary cause of death was not likely to be related to TBI were excluded; and 2) repetition of regression analyses in which patients whose CT scans revealed normal findings were excluded. Calculations were performed with the aid of a commercially available statistical computer software package (SAS Version 6.12; SAS Institute INC., Cary, NC).

Results

The study population included 2177 patients, 1701 patients with a severe TBI and 476 with a moderate TBI. Table 1 shows that the percentage of patients with severe TBI was lower in North America (72%) than in Europe (85%).

Comparison of Europe and North America

Patient characteristics and causes of injury

The characteristics of patients enrolled in the NAS and the ES are shown in Table 2. Patient ages (mean 32.8 years in North America and 33.7 years in Europe) and sex distributions (79% male patients in North America compared with 76% male patients in Europe) were similar in both studies. On both continents patients were most frequently injured in road traffic accidents (57% in North America compared with 61% in Europe [Table 3]); however, in North America accidents more frequently involved motor vehicles, whereas in Europe a larger proportion of

Table 1. Distribution of patients with severe TBI by country

Region	Total	No. of patients with severe TBI (%)*	
All	2177	1701	(78)
Continent			
North America	1149	826	(72)
Europe	1028	875	(85)
Country			
US	1029	741	(72)
Canada	119	85	(71)
Norway	20	16	(80)
Sweden	56	48	(86)
Denmark	55	43	(78)
Finland	60	49	(82)
The Netherlands	80	72	(90)
Belgium	58	50	(86)
Germany	196	165	(84)
Switzerland	23	19	(83)
France	103	86	(83)
Italy	167	143	(86)
Portugal	20	13	(65)
Spain	92	80	(87)
UK	98	91	(93)

* Severe TBI is defined as a GCS score of 3 to 8 at admission

Table 2. Characteristics of 2177 patients with severe or moderate TBI

Characteristic	No. of Patients (%)			
	North America		Europe	
Total no. of patients	1149		1028	
Demographics				
Age in yrs (mean ± standard deviation)	32.8 ±12.4		33.7 ±14.6	
Sex				
Male	904	(79)	781	(76)
Female	245	(21)	247	(24)
Clinical features				
Motor score				
1 – 3	352	(39)	346	(37)
4 – 5	488	(53)	515	(56)
6	73	(8)	67	(7.2)
Pupillary reactivity*				
Both reactive	641	(68)	698	(72)
One Reactive	111	(12)	156	(16)
No reaction	186	(20)	113	(12)
Hypotension*	240	(21)	141	(14)
Hypoxia*	280	(28)	137	(15)
ICP ≥ 20 mm Hg within 1st 24 hrs postinjury*	436	(58)	466	(64)
Findings on admission CT scan				
Initial CT classification*				
Diffuse injury I	69	(6.2)	16	(1.6)
Diffuse injury II	378	(34)	326	(32)
Diffuse injury III	176	(16)	167	(16)
Diffuse injury IV	35	(3.2)	40	(3.9)
Mass lesion	454	(40)	471	(46)
Midline shift > 5 mm*	185	(17)	229	(23)
Absent or compressed basal cisterns	495	(44)	485	(48)
Traumatic SAH	610	(54)	544	(54)

* p-value < 0.05 between continents

traffic accidents involved bicycles and mopeds. Traumatic brain injuries caused by falls were more common in Europe (20% in the ES compared with 15% in the NAS), whereas assaults were more common in North America (11% in the NAS and 2.6% in the ES [Table 3]). The proportion of patients with secondary insults (hypoxia or hypotension) or the absence of pupillary reactivity was larger in North America than in Europe. The incidence of traumatic SAH and the status of the basal cisterns were similar between continents; however, in Europe mass lesions were noted more frequently (46% in the ES compared with 40% in the NAS [Table 2]). Likewise, midline shift greater than 5 mm was more common in Europe (23% in the ES and 17% in the NAS [Table 2]).

Table 3. Causes of injury in 2177 patients with severe or moderate TBI*

Cause of injury	No. of Patients (%)			
	North America		Europe	
All causes	1149		1028	
Traffic accident	651	(57)	625	(61)
Motor vehicle	511	(78)	345	(55)
Motorcycle	118	(18)	117	(19)
Bicyle/moped	22	(3.4)	163	(26)
Fall	171	(15)	201	(20)
Sports/recreation	20	(1.7)	24	(2.3)
Assault	126	(11)	27	(2.6)
Other#	181	(16)	151	(15)

* p-value < 0.05
\# Other includes accidents at work and injuries within a building

Table 4. Comparison of health care process parameters

Referral policy	Median (Q1-Q3)*							
	Primary Referral				Secondary Referral			
	North America		Europe		North America		Europe	
No. of patients#	982	(90%)‡	628	(61%)‡	167	(10%)‡	400	(39%)‡
Time from injury to arrival at study center (mins)	39#	(29-55)	50	(35-73)	133	(110-165)	140	(115-180)
Time from injury to acquisition of CT scan (mins)#	95	(73-124)	108	(84-136)	180	(135-223)	120	(76-174)
Time from injury to surgery (mins)	155	(123-210)	160	(135-200)	241	(186-310)	215	(180-265)

* Q1 = first quartile; Q3 = third quartile
\# p < 0.05 between continents
‡ Percentage per continent

These analyses were repeated separately for patients suffering from severe and moderate TBI. The aforementioned differences in patients' characteristics between the two continents remained for both groups, except for pupillary reactivity, for which no differences were observed in the group of patients with moderate TBI.

Case management

In Europe more patients were admitted to the study center following secondary referral (Table 4). The median times between injury and arrival at the study center and between injury and examination of the CT scan were longer in Europe for patients with a primary referral (Table 4). Intracranial pressure was monitored in 66% of patients in North America and in 73% of patients in Europe. The guidelines for the protocol of the trials advised ICP monitoring in patients with a GCS score of 6 or less, as well as in patients in whom there was evidence of raised ICP on the CT scan (absent or compressed basal cisterns or mass lesion). Measurements of ICP were available in 80% of the patients in the NAS and 88% of the patients in the ES with these clinical characteristics. The method of ICP monitoring varied between continents: ventricular fluid pressure monitoring was used more commonly in North America (37%) than in Europe (24%). The approach to ICP-lowering management was significantly different between the continents, being on average more intensive in North America (Table 5). In both continents sedation was induced in more than 90% of the patients. In North America a larger proportion of patients received neuromuscular blocking agents (Table 5).

Patient outcomes

Mortality at 6 months post-TBI was higher in Europe than in North America (25% in the ES and 20% in the NAS, p=0.002). Because this difference may be caused by the larger proportion of patients with moderate TBI in the NAS population, we adjusted for moderate and severe TBI. The separate analyses still showed a significantly greater probability that death would occur among patients in Europe (odds ratio [OR] = 1.24, $p < 0.05$; Table 6). The difference in mortality rates may also be explained by the demonstrated differences in patient characteristics, time intervals, and use of ICP-lowering medication. Characteristics indicating a more severe case mix in Europe were greater proportions of patients with mass lesions, midline shifts, and raised ICP. On the other hand, patients in the NAS more often displayed papillary abnormalities and an increased incidence of hypoxia and hypotension. After adjustment for variations in case mix and management, the observed difference in outcomes between both studies became even more pronounced (OR = 1.56, p = 0.007, Table 6).

The difference in unfavorable outcome between the continents (43% in Europe compared with 38% in North America) was 5%. After the difference was adjusted for severe and moderate TBI, it was not significant (OR = 1.04, p = 0.68; Table 6). After adjustment for variations in case mix and case management, including the proportion of patients with moderate TBI, the regional difference in unfavorable outcome became more clear and reached statistical significance (OR = 1.46, p = 0.007). Subgroup analyses for patients with severe and moderate TBIs confirmed that outcomes were poorer in Europe compared with North America (all ORs > 1, Table 6).

Table 5. Treatment for patients in different ICP categories by continent

Treatment	No. of patients (%)											
	Total			ICP < 20 mm Hg			20 mm Hg ≤ ICP < 30mm Hg			ICP ≥ 30 mm Hg		
	North America		Europe	North America		Europe	North America		Europe	North America		Europe
No. of patients	757		732	321		266	227		225	209		241
Sedation	681	(90%)*	688 (94%)	289	(90%)*	253 (95%)	216	(95%)	215 (96%)	176	(84%)*	220 (91%)
Paralysis	492	(65%)*	393 (54%)	177	(55%)*	119 (45%)	169	(74%)*	133 (59%)	146	(70%)*	141 (59%)
Drainage	232	(31%)*	106 (15%)	73	(23%)*	26 (9.8%)	79	(35%)*	30 (13%)	80	(38%)*	50 (21%)
Mannitol	350	(46%)*	269 (37%)	66	(21%)*	36 (14%)	127	(56%)*	83 (37%)	157	(75%)*	150 (62%)
Hyperventilation												
Intensive (≤ **30 mm Hg**)	616	(85%)*	474 (66%)	239	(79%)*	168 (64%)	194	(87%)*	144 (65%)	183	(91%)*	162 (69%)
Very intensive (≤ **25 mm Hg**)	393	(54%)*	202 (28%)	140	(46%)*	65 (25%)	121	(55%)*	65 (29%)	132	(66%)*	72 (31%)
Barbiturates	72	(9.5%)	85 (12%)	11	(3.4%)	8 (3.0%)	17	(7.5%)	11 (4.9%)	44	(21%)	66 (27%)

* p < 0.05 between continents

Table 6. Outcome 6 months postinjury in patients with severe or moderate TBI

Outcome	Europe (1028 patients)	North America (1149 patients)	Crude OR (95% CI) #	Adjusted OR* (95% CI) #
Severe and moderate				
Mortality	259 (25%)	225 (20%)	1.24 (1.00 – 1.52)†	1.56 (1.13 – 2.16)
Unfavorable outcome ‡	428 (43%)	407 (38%)	1.04 (0.87 – 1.25)†	1.46 (1.11 – 1.93)
Severe				
Mortality	242 (28%)	199 (24%)	1.21 (0.97 – 1.51)	1.54 (1.09 – 2.18)
Unfavorable outcome ‡	401 (47%)	359 (46%)	1.02 (0.84 – 1.24)	1.48 (1.09 – 2.00)
Moderate				
Mortality	17 (11%)	26 (8.1%)	1.49 (0.78 – 2.84)	2.36 (0.88 – 6.32)
Unfavorable outcome ‡	27 (19%)	48 (17%)	1.16 (0.69 – 1.95)	1.38 (0.64 – 2.97)

CI = confidence interval
* Adjusted for severe or moderate TBI, causes of injury, age, clinical characteristics (motor score at admission, hypotension, hypoxia, CT classification, compressed or absent basal cisterns, midline shift > 5 mm, raised ICP, and type lesion), and case management (difference in time between injury and arrival at study center, ICP-lowering therapy, and number of patients per center). The effect of region on outcome is statistically significant (p-value < 0.05) if the 95% CI does not include the value one
† Adjusted for severe or moderate TBI
‡ Unfavorable outcome includes death, vegetative state, and severe disability according to the GOS

Comparison of regions within Europe

Patient characteristics and causes of injury

Comparisons among regions within Europe are summarized in Table 7. Road traffic accidents more frequently were the cause of injury in Italy, Portugal and Spain, and the Netherlands and Belgium. Falls as a cause of injury were more common in Germany and Finland. The mean age of patients varied from 30 years in France to 38 years in Germany (Table 8). Pupillary abnormalities were more frequently seen in Germany and the United Kingdom (UK). Hypoxia was more common in France and in the Netherlands and Belgium. The proportion of patients with intracranial hypertension was lowest in the UK (49%) and highest in the Netherlands and Belgium (73%). Patients with mass lesions were frequently observed in Germany (60%) and Italy (52%).

Case management

The proportion of patients who were secondarily referred for treatment ranged from 25% in France to 58% in Italy. The median time between injury and arrival at the study center for patients primarily referred ranged from 36 minutes in Portugal and Spain to 100 minutes in France. For patients secondarily referred, the time to arrival at the study center ranged from 128 minutes in Italy to 180 minutes in France. The median time between injury and acquisition of a CT scan for primarily referred patients was lowest in Italy (82 minutes) and highest in France (139 minutes). In patients in the secondary referral group, the median time to acquisition of a CT scan was lowest in the Netherlands and Belgium (89 minutes) and highest in Portugal and Spain (176 minutes).

Table 7. Cause of injury in patients with moderate or severe TBI stratified by region

Cause of injury*	European study								North American study	
	Norway, Sweden, Denmark	Finland	The Nether-lands and Belgium	Germany	France	Italy	Portugal and Spain	UK	USA	Canada
No. of patients	131	60	138	196	103	167	112	98	1029	119
Traffic accident (%)	63	35	70	48	67	75	69	47	56	60
Of which caused by motorvehicle	51	57	55	62	65	44	57	54	78	80
Of which caused by motorcycle	7.2	0	14	12	29	23	35	24	19	14
Of which caused by bicycle/moped	42	43	31	27	5.8	33	7.8	22	3.1	5.6
Fall (%)	21	32	19	31	15	11	13	16	15	18
Sports/recreation (%)	3.1	1.7	2.9	2.0	1.0	1.8	0.9	5.1	1.8	0.8
Assault (%)	2.3	13	0.7	3.6	1.0	1.8	1.8	2.0	12	3.4
Other (%)	11	18	8.0	16	17	10	16	30	15	18

* $p < 0.05$ within Europe

There was no clear difference among regions in the time between injury and the start of surgical procedures in patients who underwent surgery.

Patient outcomes

Table 8 shows that the mortality rate varied between 20 and 31 % within Europe and unfavorable outcomes varied between 36 en 56%. These differences did not prove to be statistically significant in either unadjusted or adjusted analyses (likelihood ratio test: $p = 0.32$ for death and $p = 0.23$ for unfavorable outcome).

Comparisons within North America

Patient characteristics and causes of injury

Ages of patients, causes of injury, and most clinical characteristics were similar for both Canada and the US. Only the occurrence of traumatic SAH was more frequent in patients from the US (56%) than from Canada (42%, Table 8).

Case management

In Canada significantly more patients were admitted following secondary referral (35% in Canada compared with 12% in the US). The median time between injury and arrival at the study center was similar in Canada and the US, both for patients who were primarily referred (approximately 40 minutes), and for patients who were secondarily referred (approximately 135 minutes). Among patients who were primarily referred, in Canada delays before CT scans were obtained were significantly longer than those in the US (median 128 minutes for Canada compared with 93 minutes in the US; $p < 0.05$; data not shown); among patients who were secondarily referred, surgery was performed later (median 333 minutes post-TBI in Canada compared with 202 minutes post-TBI in the US; $p < 0.05$, data not shown).

Patient outcomes

The mortality rate was 21% in Canada and 19% in the US and the rates of unfavorable outcome were 37 and 38%, respectively. Differences in both outcome measures were not significant in either unadjusted or adjusted analyses ($p = 0.68$ for the mortality rate and $p = 0.83$ for the rate of unfavorable outcome in the unadjusted analysis).

Table 8. Characteristics of patients with moderate or severe TBI stratified by region

Characteristic	European Study								North American Study	
	Norway, Sweden, Denmark	Finland	Nether-lands and Belgium	Germany	France	Italy	Portugal and Spain	United Kingdom	USA	Canada
No. of patients	131	60	138	196	103	167	112	98	1029	119
Demographics										
Mean age (years)*	35	37	30	38	30	34	31	31	33	34
Sex* (%)										
Male	68	75	80	80	87	72	75	73	79	74
Female	32	25	20	20	13	28	25	21	21	26
Clinical features (%)										
Motor score*										
1 – 3	37	31	47	30	36	33	27	60	38	41
4 – 5	58	55	47	63	58	56	64	39	54	51
6	4.4	14	6.9	7.8	6.1	11	9.3	1.1	8.0	7.8
Pupillary reactivity*										
Both pupils react	82	70	70	59	81	84	76	62	69	65
One pupil reacts	15	22	15	17	15	15	11	18	20	17
No pupil reacts	3.7	8.3	15	24	4.9	1.2	13	20	11	18
Hypotension	14	15	17	17	11	12	19	10	22	19
Hypoxia*	12	6.7	22	10	21	13	15	20	27	32
ICP ≥ 20 mm Hg within first 24 hrs after injury*	51	66	7	70	56	71	65	49	57	67
Findings on admission CT scan (%)										
Initial CT classification*										
Diffuse injury I	4.6	1.7	1.5	0.5	1.0	0	0.9	3.1	6.4	4.4
Diffuse injury II	34	35	29	22	35	32	44	35	34	38
Diffuse injury III	18	10	22	14	24	13	9.8	19	16	14
Diffuse injury IV	3.1	5.0	5.9	3.1	4.0	3.0	3.6	5.1	3.3	1.8
Mass lesion	42	48	41	60	37	52	42	38	41	42
Midline shift > 5 mm	21	32	18	29	19	25	17	21	16	19
Compressed or absent basal cisterns	43	52	50	52	45	52	37	55	44	44
Traumatic SAH* #	40	45	45	66	69	49	59	57	56	42
Outcome (%)										
Mortality	24	30	31	30	22	20	25	20	19	21
Unfavorable outcome	40	56	46	48	36	36	44	41	38	37

* $p < 0.05$ within Europe
\# $p < 0.05$ within North America

Discussion

The present study shows that, despite strict inclusion and exclusion criteria and recommendations for case management, major regional variations may exist in a patient population included in a geographically diverse clinical trial. These differences relate to causes of injury, patient characteristics (demographics and clinical characteristics), case management (referral policy, time intervals, and therapy) and outcomes. The differences may pertain to social and cultural differences, as well as to aspects of local policies governing responses to trauma, both of which are outside the influence of investigators. Regional differences in measuring instruments or in clinical scoring may also add to the observed heterogeneity.

Social and cultural differences, such as mode of transport, influence the causes of injuries that are sustained. In North America relatively more patients who were enrolled in the study were injured while driving a motor vehicle, which resulted in many high-velocity injuries. In Europe relatively more patients sustained a low-velocity injury, caused by a bicycle, moped, or fall. The cause of injury was associated with clinical characteristics: patients who sustained a high-velocity injury had a higher incidence of abnormal pupils, hypoxia, and hypotension. Findings on the CT scan indicative of raised ICP (absent or compressed basal cisterns, midline shift, or mass lesions) were more common in patients injured as the result of a fall or a bicycle accident (data not shown). These patterns are in accordance with those of previous reports[6,15,16]. As the cause of injury, falls were more frequent in Europe, particularly in Finland and Germany (Tables 3 and 7). Patients who sustained injuries from falls were commonly older and had a higher incidence of obliteration of basal cisterns, midline shifts, and mass lesions were detected on CT scans in these patients (data not shown). These clinical characteristics were related to outcome. Consequently, social and cultural differences influence the composition of the population under study (case mix) and relate to outcome.

Likewise, varying approaches to basic TBI management, as demonstrated between North America and Europe for ICP-lowering therapy, may be considered evidence of a cultural difference in attitude: ICP-lowering medication was less frequently administered in Europe than in North America to patients in whom ICP was 20 mm Hg or higher, thus demonstrating that a more conservative approach was taken in Europe (Table 2). A limiting factor of the present study is that ICP-lowering therapy was only recorded after initiation of ICP monitoring. It is conceivable that therapy may have been instituted before initiation of monitoring, thus affecting the observed ICP. If in North America relatively many patients received therapy before the start of monitoring, this would lower their observed ICP. In this case the 'real' ICP of North American patients would be higher, thus decreasing the observed difference in the frequency of ICP-lowering therapy. Both in North America and in Europe ICP-lowering medication was also given to patients in whom the ICP was lower than 20 mm Hg, although the study protocol and also more recent North American evidence-based guidelines for severe TBI management[17] advise that ICP-lowering therapy be initiated only at a threshold of 20 to 25 mm Hg.

Secondary referral occurred more frequently in Europe than in North America (Table 4). We performed an additional analysis of which patient characteristics and external factors were associated with secondary referral. Both in North America and Europe, the distance from the

injury site to a neurosurgical trauma center was related to secondary referral. In North America patients who were secondarily referred more frequently had sustained injuries as motor vehicle occupants, reflecting a protocol for dealing with trauma in which policy dictates referral of patients with high-velocity injuries to a neurosurgical trauma center. In contrast, in Europe mass lesions were significantly more frequently seen in patients who were secondarily referred and, consequently, surgery was also performed in a higher percentage of patients. Hence, the European referral policy appears to be based more on abnormalities observed on the CT scan, resulting in the selection of patients with a higher incidence of mass lesions.

Secondary referral occurred also more often in Canada than in the US. The longer delays before CT scans were obtained or surgery was performed in Canada, compared with the US, are most likely caused by local in-hospital protocols for dealing with trauma, because the delays between injury and arrival at the study center were similar both for patients primarily and secondarily referred.

Remarkably, outcomes in the European population were worse than in those in the North American population, with regard to death and, to a lesser extent, unfavorable outcomes (Table 6). Several causes contributing to these differences in outcome need to be considered, such as regional variation in enrollment criteria, regional differences in case mix, and regional variation in management. Various adjustments, including those of major predictive variables[18], were made with statistical models, but the difference in outcome could not be explained. Sensitivity analyses (excluding patients in whom there were normal findings on CT scans or patients whose deaths were not likely to have been related to TBI) provided results similar to the overall analysis. Care should be taken, however, before concluding that treatment results in North America are definitely better. Many other factors can also affect the difference. These include unknown confounders, traumas other than TBI, center effects, and perhaps to a lesser extent, regional differences in scoring or measuring instruments. In an extensive analysis of the National Acute Brain Injury Study of hypothermia, Clifton et al.[19] demonstrated significant intercenter differences. Our results could not confirm this, however; in the multivariable analyses the number of patients included in each center had no significant relationship with outcome.

In the present study we found no clear differences in outcomes among regions within Europe, or between Canada and the US. The lack of statistical significance may be caused by the small sample sizes in the different regions, which result in large confidence intervals and limited statistical power.

In the recent past two large series contained patient data on TBI that were collected in different centers distributed over a continent, one series in Europe (the European Brain Injury Consortium survey, EBIC)[6] and the other in the US (the Traumatic Coma Data Bank, TCDB)[20,21]. We used these more unselective series for a comparison with the present study. Differences among regions within Europe in the present study are similar to those reported in the EBIC survey[6]. In the EBIC survey, patients injured in Spain, the Netherlands, and Belgium were also younger, on average, and more frequently sustained injuries in road traffic accidents. Conversely, in Finland and Scandinavia, patients were older, and more frequently sustained injuries due to a fall. In the UK, France, Scandinavia, Portugal, and Spain secondary referrals were more frequent, but were not

always related to a higher incidence of surgical operations. Detailed comparisons of the EBIC survey and the present study are complicated, because different inclusion and exclusion criteria were used. For example, in the present study all patients older than 65 years were excluded, resulting in a lower mean age (34 years compared with 42 years in the EBIC survey).

The TCDB[20,21] only included patients with severe TBIs. When compared with patients who sustained severe TBI from the NAS, the population in the TCDB was somewhat younger (median age 30 years in the present study compared with 25 years in TCDB). Variations in causes of injury were more pronounced (75% traffic accidents and 5% assaults in the TCDB compared with 61% traffic accidents and 11% assaults in the present study). The difference in the proportion of direct transfers was also remarkable, 61% in the TCDB compared with 90% in the present study. The observed differences may reflect changes occurring over time in mobility, society, or protocols for dealing with trauma.

The distribution of patients in the various categories of CT classification differed among the EBIC, the TCDB, and the present study. It remains uncertain whether these differences are real or may be explained by variations in selection criteria and observer variation in scoring the CT scan classification.

The observed regional differences can be seen to argue for blocked randomization in each center and for stratification of patients by prognostic risk. Such procedures will ensure a balanced treatment assignment, with respect to prognostic risk in each stratum. Prerandomization stratification by prognostic risk may be difficult to accomplish in emergency situations, however. Alternatively, covariate adjustment could be included in the final statistical analysis for predictive characteristics and regional effects. Such an adjustment will increase the statistical power of the study to detect a treatment effect[22,23]. Moreover, heterogeneity in the severity of injury in the patients can be reduced by limiting enrollment to patients with an intermediate prognosis, for example, between a 20 and 80% risk for unfavorable outcome. This leads to a focus on patients for whom treatment effects can be better determined[8]. Designing a randomized clinical trial in this way may limit the disturbing effects of heterogeneity and increase chances of showing the benefit of a new therapy[7].

Conclusions

We have shown that regional differences in patient characteristics, case management, and outcomes exist also in the context of a randomized clinical trial. This aspect merits additional attention to the design of such a study. The observed poorer outcome in Europe, compared with North America, is a remarkable finding and requires further study.

Appendix

Imputation of outcome at 6 months postinjury in both tirilazad trials.

Outcome assessment of GOS scores at 6 months was considered to be applicable if the assessments were performed between 5 and 7 months after the date of injury. For 164 patients, the GOS score was measured outside this time interval. In these cases imputation of outcome (favorable or unfavorable) was performed, in which we used assessments obtained at 3 or 12 months postinjury and gave each imputation a weight factor. We used the following algorithm:

- If the GOS score is available for the period 7 to 8 months postinjury, impute that score for the 6-month postinjury time point (weight=1).
- If the GOS score is still missing, but is available for the period 4 to 5 months postinjury, impute that score for the 6-month postinjury time point (weight=1).
- If the GOS score is still missing, but is available for the period 2.5 to 3.5 months postinjury, then do one of the following:
  1) If the GOS score at 2.5 to 3.5 months postinjury was 'good recovery', impute a score of unfavorable outcome for the 6-month postinjury time point (weight=1).
  2) If the GOS score at 2.5 to 3.5 months postinjury was 'vegetative state', impute a score of unfavorable outcome for the 6-month postinjury time point (weight=1).
  3) If the GOS score at 2.5 to 3.5 months postinjury was 'severe disability', then do one of the following:
     i) if the GOS score at 11 to 13 months is missing, give the GOS score at 6 months postinjury a favorable outcome value (weight=0.33) and an unfavorable outcome value (weight=0.67).
     ii) if the GOS score at 11 to 13 months is 'good recovery', give the GOS score at 6 months postinjury a favorable outcome value (weight=0.75) and an unfavorable outcome value (weight=0.25).
     iii) if the GOS score at 11 to 13 months is 'moderate disability', give the GOS score at 6 months postinjury a favorable outcome value (weight=0.5) and an unfavorable outcome value (weight=0.5).
- If the GOS score is still missing, but the outcome for the period at 11 to 13 months is 'severe disability', give the GOS score at 6 months an unfavorable outcome (weight=1).

Comment on paper

To the editor:

As European Principal Investigators and members of the Executive Committee of the Tirilazad Europe (Australia Study), we read with the interest the paper by Hukkelhoven, et al.[24].

The issue of regional differences among patients enrolled in large multicenter studies is important and these authors are to be congratulated for the detailed description of the characteristics of a large database composed by combining information about patients recruited into two separate projects: European and American tirilazad trials[9]. Their findings confirm some of those previously reported concerning the demographic, clinical, and radiological profile of head-injured patients enrolled in clinical trials.

An intriguing, and even unprecedented finding is the apparent difference in outcome in patients recruited from the different continents (Europe and the US) in which the two trials were performed, even after accounting for 'variations in case mix and management.' Previous prospective comparisons between populations of head-injured patients in Europe and North America did not show these differences. For example, Murray[25] demonstrated that the outcomes in patients treated at a trauma center in California were almost identical to those predicted in patients treated in centers in the United Kingdom and The Netherlands.

A number of issues must be addressed before considering that differences in approaches to case management, which may account for differences in observed outcome. These include questions about the comparability of patients recruited for the two trials, the convenience of case management approaches within the different continents, and inconsistencies of approaches to assessment of outcome.

Comparability in the pattern and severity of brain damage is a fundamental prerequisite in the investigations of potential influences on outcome. Thus, at a comparable clinical level of severity, outcome is worse in a patient who has a mass lesion; however, patients recruited in Europe had a higher frequency of mass lesions than those in North America. Likewise, at a given clinical level of severity, outcome varies according to the time when the patient is assessed postinjury and the extent of resuscitation previously carried out. As a reflection of this, neurosurgical unit personnel who accept patients only as secondary referrals (after previous assessment and resuscitation in other hospitals) are likely to recruit patients already selected as having more serious injuries. Only 10% of patients in North American centers were secondary referrals, whereas 39% of patients in European centers were referred secondarily (as much as 60% in some regions of Europe).

Although complete information regarding time from injury to admission to the study center is unavailable for the two cohorts, it is clear that patients took longer to reach the European centers.

Therefore, it would be interesting to know the results of comparisons of outcome between these two categories of patients. Furthermore, comparability in the findings on each patient's

initial computerized tomography (CT) scan may still allow differences to emerge in association with the development of delayed worsening, including the possibility of the development of surgically significant mass lesions in a patient previously considered to have a diffuse injury[26]. A significant difference in the incidence of traumatic subarachnoid hemorrhage (SAH) a powerful indicator of poor prognosis) was demonstrated in patients directly admitted, compared with patients transferred from another hospital[27]. Center personnel accepting a large number of secondary referrals are likely to select more critically ill patients (with evolving mass lesions and traumatic SAH).

Validity of the concept of a continent-specific approach to case management can be questioned. Thus, wide variations in practices among different centers in North America have been reported[19,28], with only 16% of centers complying fully with published guidelines[29]. Similar variations have been reported among European countries and among centers within a country[2]. Intercenter differences apply even among centers taking part in sophisticated clinical trials[19].

In the clinical trials involving tirilazad, researchers included patients with severe head injury as well as those considered to have moderate head injury. The identification of patients with moderate head injury is less firmly established, there is more variability in the translation of Glasgow Coma Scale (GCS) observations into a qualifying GCS score in this group, and there are varying approaches to preadmission assessment, intubation, resuscitation, and criteria for interhospital transfer likely to affect this group. These factors may have resulted in differences in the total populations included in the European and North American studies.

The use of very intensive hyperventilation in 46% of patients with an intracranial pressure (ICP) below 20 mm Hg in North American centers is particularly curious.

The possibility that more frequent use of 'more aggressive' therapy, particularly aimed at lowering ICP, is proposed as a possible reason for the difference in outcome. This is controversial. For instance, extreme hyperventilation ($PCO2 \leq 25$ mm Hg) was used twice as frequently in patients treated in North American centers (including in patients with a normal ICP) even though there is evidence that this practice is associated with worsening of patient outcome. The avoidance of chronic, prolonged hyperventilation is regarded as a standard in the treatment of severely head injured patients in North America[30]. Researchers of other observational studies[28,31] have found the greater use of ICP monitoring or therapy not to be associated with substantially improved favorable outcomes.

Reliable comparisons between results in cohorts of patients require that outcome is assessed using comparable methods, follow up is as complete as possible, and occurs at identical intervals after injury. Variations between different observers, related to experience and/or professional background, are well recognized. The impressionistic use of the Glasgow Outcome Scale, known to be associated with high rates of interobserver variation, has been superseded by a structured approach to score assignment[32], but this approach was not used in the tirilazad studies. Although the proformas used to record outcome information were identical in the two trials, no advance steps were taken to ensure that they were applied and interpreted identically on the two continents.

In conclusion, we have significant reservations about the assertion that mortality and unfavorable outcomes were significantly higher in Europe. At the very least, this statement requires qualification with an acknowledgment that the observation may reflect differences in the type of patients recruited to the trials, in approaches to reporting outcome, and differences in case management. Finally, the debate about the merits of different approaches to the treatment of severely brain injured patients, which has festered for more than two decades, seems unlikely to be resolved by comparing outcomes of cohorts in different circumstances. We strongly endorse the view[31] that randomized, prospective trials are required, first, to establish the principle that ICP lowering/ decompressive therapy is beneficial, and second, to evaluate the merits of different approaches.

Franco Servadei, M.D.
Albino Bricolo, Ph.D.
Jacques Lagarrigue, M.D.
Ramiro Lobato, M.D.
Lennart Persson, Ph.D.
Members of the Executive Committee of the Tirilazad Europe–Australia Study

Author reply

Dear Editor,

We would like to thank the principle investigators and members of the Executive Committee of the Tirilazad Europe-Australian study for their constructive comments on our paper on regional differences in patient characteristics, case management and outcomes in traumatic brain injury[24]. In this paper we compared North America with Europe and found on average a better outcome in North American patients. Surprisingly, this continental difference in outcome remained after adjustment for a large set of potential confounders. In fact, the difference became even larger after adjustment, since the North American patients were more severe with respect to several confounders, such as the presence of hypoxia or hypotension. We did not wish to place too much emphasis on this finding, as the underlying reasons for this difference were unclear. The finding however is intriguing.

Outcome comparison between continents in a multivariable analysis has seldom been performed, which makes comparison of our findings with previously observed results difficult. In the past, Murray et al.[25] compared six-month mortality between Glasgow and San Francisco in a multivariable analysis, but, in contrast to our results, no continental difference in the outcome was observed. However, this study[25] was rather a comparison of two centers than of two continents and data were collected about 20 to 30 years ago, i.e. between 1970 and 1985.

Servadei et al. wondered whether the continental difference in outcome is real. Regarding the low p-values corresponding with the adjusted odds ratios in Table 6[24], it is not very plausible that the difference is caused by chance. The data might however be biased for comparison of continents. Methods and/or time points of severity assessment and additional investigations may have varied structurally between continents, possibly affecting the estimated values of the severity parameters. Servadei et al. also suggested that the continental difference in outcome might be explained by a structural continental difference in outcome assessment. This is unlikely, since the difference in outcome is also observed for the unambiguous outcome mortality.

Besides the possibility of bias, several other factors could have contributed to the continental difference in outcome. Among them were unknown clinical factors – not included as a confounder – , differences in management approach and center effects[24].

In our paper we adjusted for many potential confounders in order to estimate the 'pure' effect of continent on outcome. Data on many previously identified potential confounders were available. However, it is always possible that other important confounders were missed. Examples of such confounders are extracranial injuries, which were not consistently reported in the Tirilazad trials, or as yet unidentified confounders.

Servadei et al. suggested the dissimilar distributed referrals as possible confounder. When adjusting for referral, all characteristics that correlate with referral policy – and that are not already adjusted for – are clustered under this referral-factor. Although we did not correct for referral in itself, we adjusted for difference in time between injury and arrival at the study center

in our paper. By this, the latter obviated both referral policy and duration of the transport by ambulance, which could also possibly worsen the clinical state of the patient. Further subgroup analyses, separate for patients with primary and secondary referral, learned that referral policy could not explain the outcome difference between Europe and North America (all ORs > 1.0, Table 9, analogous to Table 6[24]). In fact, the outcome difference was even slightly larger (larger ORs) for primarily referred patients than for those secondarily referred. Separate analyses for severe and moderate TBI patients showed the same pattern (Table 9).

Further, the present analyses included only characteristics measured at admission. Patients with a similar severity at admission may develop differently in time, thus contributing to the observed continental difference in outcome.

Another possible factor contributing to the continental difference in outcome may be variation in management. The question arose whether the identified outcome difference may be a structural difference between continents or primarily a difference between centers with more centers performing favorably in North America. Wide variations in practices between different centers within North America and Europe have been reported[2,6,19,28]. The outcome difference

Table 9. Comparison of mortality 6 months postinjury, patients with severe or moderate TBI

Referral	North America		Europe		Crude OR	Adjusted OR#
	n	mortality	n	mortality	(95% CI)	(95% CI)
Severes and moderates						
All referrals	1028	25%	1149	20%	1.24 (1.00–1.52)*	1.56 (1.13–2.16)
Primary referral	628	26%	982	20%	1.27 (0.99–1.62)*	1.59 (1.12–2.28)
Secondary referral	400	24%	167	20%	1.20 (0.77–1.89)*	1.47 (0.82–2.62)
Severes						
All referrals	875	28%	826	24%	1.21 (0.97–1.51)	1.54 (1.09–2.18)
Primary referral	540	28%	698	24%	1.23 (0.95–1.59)	1.57 (1.07–2.30)
Secondary referral	335	27%	128	23%	1.20 (0.75–1.93)	1.47 (0.79–2.74)
Moderates						
All referrals	153	11%	323	8%	1.49 (0.78–2.84)	2.36 (0.88–6.32)
Primary referral	88	13%	284	8%	1.70 (0.79–3.65)	2.41 (0.82–7.04)
Secondary referral	65	9%	39	8%	1.22 (0.29–5.19)	1.84 (0.27–12.5)

* Adjusted for severe or moderate TBI

\# Adjusted for severe or moderate TBI, cause of injury, age, clinical characteristics (motor score at admission, hypotension, hypoxia, CT classification, compressed or absent basal cisterns, midline shift > 5 mm, raised ICP, type lesion), management (difference in time between injury and arrival at study center, ICP lowering therapy, number of patients per center), referral policy

OR = Odds Ratio (Europe versus North America), 95% CI = 95% Confidence Interval. The effect of region on mortality is statistically significant (p-value < 0.05) if the 95% CI does not include the value one

being 'center-specific' instead of 'continent-specific' is possible, especially if inter-center-variability in outcome would be larger than inter-continent-variability. Unfortunately, power was not large enough to study these inter-center-differences in a valid way – even not with our relatively large sample size.

Regarding management, we adjusted for intensity of ICP therapy. Therefore the ICP management can not be the reason for the difference in outcome. We regret that some physicians may have interpreted our paper as if we were advocating a more aggressive therapy, as this was certainly not reported as such in our paper. In 1983 our group[33] reported in this journal a clear difference in 1-year survival between two Dutch centers, with the higher survival rate at the center with the more conservative management regimen. However, other studies reported opposite findings[28,34]. We agree with Servadei et al. that randomized, prospective trials are required to offer convincing evidence whether intensive therapy, e.g. for lowering ICP, is beneficial. This, however, is a complicated issue as in general more severely injured patients will receive more intensive therapy[30].

Overall, the observed continental difference in outcome is an unexpected and intriguing result. This difference may be caused by coincidence and may have been overestimated in the tirilazad data. Other explanations include regional differences in scoring or measuring instruments, (yet) unknown clinical factors, regional differences in management approach and center effects. We hope our paper and the discussion presented here will challenge neurosurgeons to think further about this remarkable finding and to explore possibilities that may lead to improvement in outcome of TBI patients.

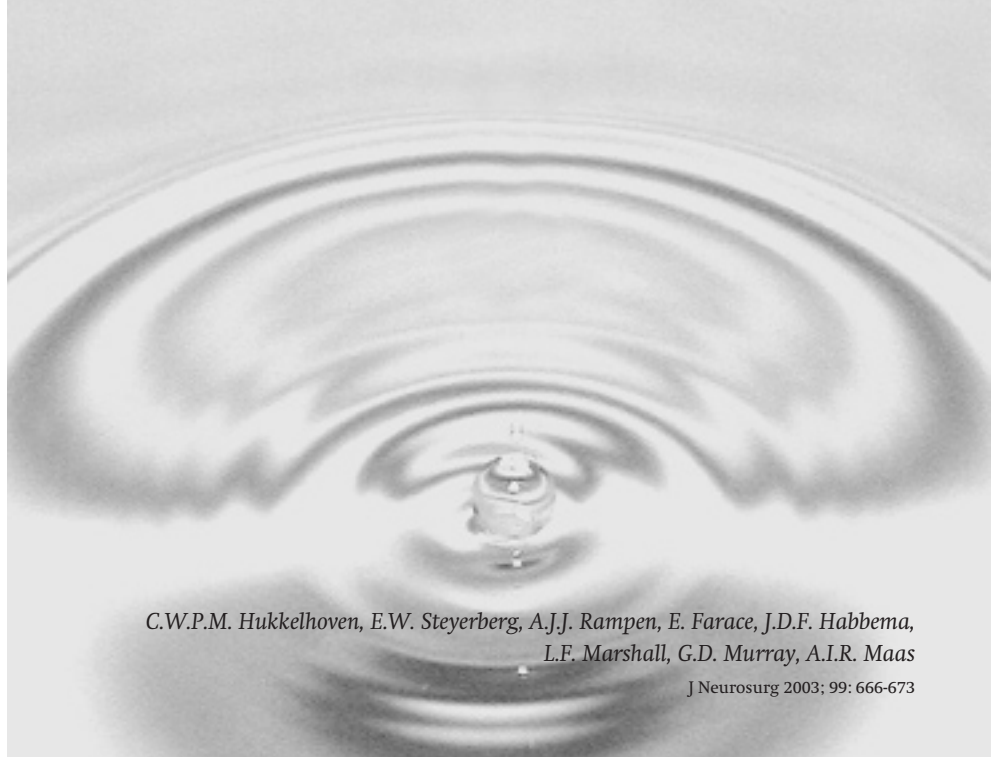
References

1. Ghajar J, Hariri RJ, Narayan RK, Iacono LA, Firlik K, Patterson RH. Survey of critical care management of comatose, head-injured patients in the United States. Crit Care Med 1995;23(3):560-7.
2. Jeevaratnam D, Menon D. Survey of intensive care of severely head injured patients in the United Kingdom. BMJ 1996;312:944-947.
3. Matta B, Menon D. Severe head injury in the United Kingdom and Ireland: a survey of practice and implications for management. Crit Care Med 1996;24:1743-1748.
4. Jennett B, Teasdale GM, Galbraith S, et al. Severe head injury in three countries. J Neurol Neurosurg Psychiatry 1977;40:291-198.
5. Murray L, Teasdale GM, Murray GD, Miller D, Pickard J, Shaw M. Head injuries in four British neurosurgical centers. Br J Neurosurg 1991;13:566-571.
6. Murray GD, Teasdale GM, Braakman R, et al. The European Brain Injury Consortium survey of head injuries. Acta Neurochir (Wien) 1999;141:223-236.
7. Maas AIR, Steyerberg EW, Murray GD, et al. Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 1999;44(6):1286-98.
8. Machado S, Murray GD, Teasdale GM. Evaluation of designs for clinical trials of neuroprotective agents in head injury. J Neurotrauma 1999;16:1131-1138.
9. Marshall LF, Maas AIR, Marshall S, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 1998;89:519-525.
10. Hall E, Yonkers P, Andrus P, Cox J, Anderson D. Biochemistry and pharmacology of lipid antioxidants in acute brain and spinal cord injury. J Neurotrauma 1992;9(Suppl 2):S425-S442.
11. Jennett B, Bond M. Assessment of outcome after severe brain damage. Lancet 1975;1(7905):480-4.
12. Marshall LF, Bowers Marshall S, Klauber MR, et al. A new classification of head injury based on computerized tomography. J Neurosurg 1991;75 (Suppl):S14-S20.
13. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
14. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227 - 1237.
15. Gennarelli TA, Thibault LE. Biomechanics of acute subdural hematoma. J Trauma 1982;22(8):680-6.
16. Gutman MB, Moulton RJ, Sullivan I, Hotz G, Tucker WS, Muller PJ. Risk factors predicting operable intracranial hematomas in head injury. J Neurosurg 1992;77(1):9-14.
17. Bullock R, Chesnut R, Clifton G, et al. Management and prognosis of severe traumatic brain injury. Part 1: Guidelines for the management of severe traumatic brain injury. J Neurotrauma 2000;17(6/7):451-553.
18. Chesnut R, Ghajar J, Maas AIR, et al. Management and prognosis of severe traumatic brain injury. Part 2: early indicators of prognosis in severe traumatic brain injury. J Neurotrauma 2000;17(6/7):557-627.
19. 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):751-5.
20. Foulkes M, Eisenberg H, Jane J, Marmarou A, Marshall LF, and the Traumatic Coma Data Bank Research Group: design, methods and baseline characteristics. J Neurosurg 1991;75:S8-S13.

21. Marshall LF, Becker D, Bowers S, et al. The National Traumatic Coma Data Bank Part 1. J Neurosurg 1983;59:276-284.
22. Robinson L, Jewell N. Some surprising results about covariate adjustment in logistic-regression models. Int Stat Rev 1991;59:227-240.
23. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am Heart J 2000;139(5):745-51.
24. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. J Neurosurg 2002;97(3):549-57.
25. Murray GD. Use of an international data bank to compare outcome following severe head injury in different centres. Stat Med 1986;5(2):103-12.
26. Servadei F, Murray GD, Penny K, 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;46(1):70-5; discussion 75-7.
27. Mattioli C, Beretta L, Gerevini S, et al. Traumatic subarachnoid hemorrhage on the computerized tomography scan obtained at admission: a multicenter assessment of the accuracy of diagnosis and the potential impact on patient outcome. J Neurosurg 2003;98(1):37-42.
28. 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):1870-6.
29. Hesdorffer DC, Ghajar J, Ianoco L. Predictors of compliance with the evidence-based guidelines for traumatic brain injury care: a survey of the United States trauma centers. J Trauma 2002;52:1202-1209.
30. Stocchetti N, Penny KI, Dearden M, et al. Intensive care management of head-injured patients in Europe: a survey from the European brain injury consortium. Intensive Care Med 2001;27(2):400-6.
31. The use of hyperventilation in the acute management of severe traumatic brain injury. J Neurotrauma 1996;1996:699-703.
32. Wilson JT, Edwards P, Fiddes H, Stewart E, Teasdale GM. Reliability of postal questionnaires for the Glasgow Outcome Scale. J Neurotrauma 2002;19(9):999-1005.
33. Gelpke GJ, Braakman R, Habbema JD, Hilden J. Comparison of outcome in two series of patients with severe head injuries. J Neurosurg 1983;59(5):745-50.
34. Marshall LF, Smith RW, Shapiro HM. The outcome with aggressive treatment in severe head injuries. Part I: the significance of intracranial pressure monitoring. J Neurosurg 1979;50(1):20-5.

4

Age and outcome following severe traumatic brain injury: an analysis of 5600 patients



C.W.P.M. Hukkelhoven, E.W. Steyerberg, A.J.J. Rampen, E. Farace, J.D.F. Habbema, L.F. Marshall, G.D. Murray, A.I.R. Maas

J Neurosurg 2003; 99: 666-673

Abstract

Objective

Increasing age is associated with poorer outcome in patients with closed traumatic brain injury (TBI). It is uncertain whether critical age thresholds exist, however, and the strength of the association has yet to be investigated across large series. The authors studied the shape and strength of the relationship between age and outcome, that is, the 6-month mortality rate and unfavorable outcome based on the Glasgow Outcome Scale.

Methods

The shape of the association was examined in four prospective series with individual patient data (2664 cases). All patients had a closed TBI and were of adult age (96% < 65 years). The strength of the association was investigated in a meta-analysis of the aforementioned individual patient data (2664 cases) and aggregate data (2948 cases) from TBI studies published between 1980 and 2001 (total 5612 cases). Analyses were performed with univariable and multivariable logistic regression.

Results

Proportions of mortality and unfavorable outcome increased with age: 21 and 39%, respectively, for patients younger than 35 years and 52 and 74%, respectively, for patients older than 55 years. The association between age and both mortality and unfavorable outcome was continuous and could be adequately described by a linear term and expressed even better statistically by a linear and a quadratic term. The use of age thresholds (best fitting threshold 39 years) in the analysis resulted in a considerable loss of information. The strength of the association, expressed as an odds ratio per 10 years of age, was 1.47 (95% confidence interval [CI] 1.34 – 1.63) for death and 1.49 (95% CI 1.43 – 1.56) for unfavorable outcome in univariable analyses and 1.39 (95% CI 1.30 – 1.50) and 1.46 (95% CI 1.36 – 1.56), respectively, in multivariable analyses. Thus, the odds for a poor outcome increased by 40 to 50% per 10 years of age.

Conclusions

An older age is continuously associated with a worsening outcome after TBI; hence, it is disadvantageous to define the effect of age on outcome in a discrete manner when we aim to estimate prognosis or adjust for confounding variables.

Introduction

Increasing age is associated with worse outcome in patients with systemic diseases such as cancer, coronary heart disease, and neurological diseases such as subarachnoid hemorrhage (SAH), traumatic brain injury (TBI), and dementia[1-10]. Traumatic brain injury is a major health and socioeconomic problem throughout the world and is the leading cause of death and disability in younger patients in more economically developed countries. It remains unclear how the association between patient age and outcome after closed TBI can be described best, however. In some studies researchers have treated outcome as a continuous function of age[2,11,12], whereas others have identified age threshold values between 30 and 60 years of age[3-5,13-17].

In a study on early indicators in the management and prognosis of severe TBI, Chesnut et al.[17], provided a detailed overview of published data on the association between patient age and outcome following TBI. These authors concluded that the probability of a poor outcome increased with patient age in a stepwise manner, suggesting an age threshold of 60 years. Note, however, that they recognized that this threshold might be an artifact of the age grouping used by various researchers in converting continuous data into categorical data.

Establishing the association between patient age and outcome more precisely is important in being able to predict outcome and understand how to adjust for age in epidemiological studies. Obtaining more knowledge about the shape of this association may also help to explain the relationship itself. Furthermore, identifying threshold values may be relevant to clinical research, for example, for purposes of stratification in randomized clinical trials or prognostic modeling.

The primary aim of this study was to describe the relationship between age and outcome in patients with TBI, which fit the data well and was simple (that is, low-dimensional) and easily communicated and applied in clinical practice. We also examined whether a meaningful age threshold value could be determined. Furthermore, we quantified the strength of the effect of patient age on outcome in a meta-analysis.

Table 1. Studies included in the analysis of the association between age and 6-month outcome in patients with severe TBI

Authors and year	Period of data collection	Place of data collection	Patient age limits (yrs)	Age coding (yrs)*	No. of patients	Mortality rate	Rate of unfavorable outcome# (%)	Specific criteria
Individual patient data (n=2664)								
Marshall, et al., 1998 and Hukkelhoven, et al., 2002[19,20]	1992 – 94	Europe, Australia and Israel	14 - 65	Individual	959	27%	46%	GCS motor score > 1, abnormal CT scan
Morris, et al., 1999[21]	1994 - 95	Europe, Canada Australia, Argentina	16 - 65	Individual	409	23%	43%	Abnormal CT scan
Murray, et al.,[22]	1995	Europe	≥ 15	Individual	471	40%	60%	NA
Hukkelhoven, et al., 2002[19]	1991 -94	North America	14 - 65	Individual	825	24%	46%	GCS motor score > 1, abnormal CT scan
Aggregated data (n=2948)								
Braakman, et al.,1980[1]	1973-78	Rotterdam and Groningen, the Netherlands	≥ 21	21-30,31-40, ..., >70	180	59%	NA	NA
Klauber, et al., 1981[26]	1978-79	San Diego, USA	≥ 20	20-39, 40-59, ≥60	74	47%	NA	GCS ≤ 7
Gale, et al., 1983[28]	1980-81	Seattle and King County, USA	≥ 15	15-24, 25-39, 40-59, ≥ 60	58	79%	NA	GCS ≤ 7
Lobato, et al., 1988[25]	1977-86	Madrid, Spain	≥ 20	20-29, 30-39, ,..., ≥60	41	37%	37%	Epidural hematoma
Jaggi, et al.,1990[23]	UA	Philadelphia, USA	21 - 85	21-40,>40	67	55%	NA	NA
Wilberger, et al.,1990[29]	1982-87	Pittsburgh, USA	≥ 14	14-34, 35-65, > 65	101	65%	81%	Acute subdural hematoma
Choi, et al., 1991[24]	1976-89	Richmond, USA	UA	Partly divided into ≤26, 27-61, >61 and partly into ≤33, >33	434	27%	41%	NA
Vollmer, et al.,1991[5]	1984-87	USA (TCDB)	≥ 16	16-25, 26-35, ,..., ≥ 56	587	43%	NA	NA
Gentleman, et al., 1992[27]	1979-90	Glasgow, UK	no limits	<40, 40-64, ≥65	572	NA	49%	NA
Resnick, et al., 1997[15]	1991-94	Pittsburgh, USA	UA	< 30,30-39,≥40	37	35%	62%	Prolonged raised ICP
Gómez, et al., 2000[3]	1987-96	Madrid, Spain	≥ 15	15-25, 26-35, ..., >65	797	50%	64%	NA

NA = not applicable; UA = unavailable

* Indicates how age was classified in the original paper; either as individual numbers or as age categories. Ellipses indicate that intermediate age ranges are included in the study mentioned

\# Severe disability, vegetative state, or death according to the GOS[18]. Outcome was determined at least six months postinjury.

Clinical material and methods

Patient population

Two data sources were used in the present study: 1) individual patient data (2664 cases) from four different patient series; and 2) aggregate data (2948 cases) extracted from TBI outcome studies published between 1980 and 2001 (Table 1). Adults (patients > 14 years) with severe closed TBI (Glasgow Coma Scale [GCS] scores 3-8) were selected for analysis.

The shape of the association between patient age and outcome was studied in the individual patient data by using both continuous age transformations and age as a threshold value (for example, an age ≥ 40 years compared with an age < 40 years). The strength of the association was considered in a meta-analysis that included both aggregate data and individual patient data. Outcome measures at 6 months postinjury were death and the Glasgow Outcome Scale [GOS] score[18] dichotomized into unfavorable outcome (death, vegetative state, and severe disability) and favorable outcome (moderate disability and good recovery).

Data collection

Individual patient data included populations from three multicenter phase III randomized clinical trials[19-21] and one prospective series[22] of patients with closed TBI. Six-month outcome data were available from 2500 patients. In 164 patients outcome had not been assessed at 6 months postinjury, but could be assigned according to a specific algorithm that used GOS results obtained at other points in time[19].

We searched for relevant published studies by using the PubMed service to access the MEDLINE database of citations (period 1980-2001) with the key words: 'age', 'outcome', 'head injury' or 'traumatic brain injury' and 'relation' or 'association'. Additional papers were retrieved by tracking citations in the reference lists of the aforementioned reports. Studies published in the English language were selected if they included the following: 1) 35 or more adult patients with severe closed TBI; 2) frequency data on at least two age categories together with 6-month mortality data or 6-month unfavorable outcome data; and 3) patient data from Europe, Israel, North America, or Australia to ensure comparability with the data from the individual patient series. Of more than 100 studies initially considered, 11[1,3,5,15,23-29] met these criteria and constitute the aggregate data. Studies selected for the meta-analysis are listed in Table 1.

Shape of the association

The shape of the association between patient age and outcome was studied using the following age transformations:

1. A smoothing spline[30-32]. This is a very flexible way to describe an association, but cannot be expressed easily in a parametric formula. A penalty factor, which prevents wild oscillation of the spline, is based on the default nominal degrees of freedom. Because the smoothing spline represents maximal goodness-of-fit but cannot be expressed in a formula easily, it was used as a reference for the performance of other age transformations.

2. We also considered a variety of continuous transformations: linear (age), linear plus quadratic (age plus age^2), linear plus cubic (age plus age^3), square root (age$^{0.5}$), logarithmic (log[age]), exponential (exp[age]) and reciprocal (age^{-1}). We also analyzed a previously suggested piecewise transformation, that is, no effect of age until 50 years and a linear effect above this age[4].

3. In addition, we determined age threshold values: all 60 threshold values between 15 and 76 years of age were evaluated, thus including thresholds discussed in previous studies[3-5,13-17]. The 95% confidence intervals [CIs] were constructed with profile log-likelihood methods[33] around the best fitting (optimal) threshold.

We included all transformations of age in logistic regression models, both univariably ones and those adjusted for potential confounders. The confounding variables considered were cause of injury, sex, geographical region, GCS, hypoxia, hypotension, pupillary reactivity, raised intracranial pressure [ICP] (defined as ICP ≥ 20 mm Hg), traumatic SAH, and CT scanning-based classification[34]. Interactions between the best fitting age transformation and the confounding variables were not statistically significant.

Values of missing confounders (4.6% of all values) were assigned to each patient, based on correlations with existing confounders[35,36]. The fit of the models was expressed on the log-likelihood scale of the model chi-square function (deviance)[37]. The higher the model chi-square, the better the model fits the data. The fit of the different age transformations was compared with the fit of the smoothing spline and with the fit of age as a continuous linear term.

Strength of the association

The strength of the association was analyzed with univariable (aggregate data and individual patient data) and multivariable (individual patient data) logistic regression, including age as a continuous linear term. Results were expressed as odds ratios [ORs] for every 10 years of age. In the aggregate data, age was typically reported in categories with a range of values, for example, 10 to 20 years or 15 to 25 years. Consequently, the association with age as a continuous linear term could not be estimated directly. By using the overall age distribution (mean age and standard deviation) in the study under consideration, however, we could validly assign a mean age to each age category[38]. If the overall age distribution was not reported, the age distribution from the combined data from the tirilazad trials[19,20] was ascribed, matched to specific patient characteristics (for example, patients with epidural hematoma).

All analyses were performed on available data sets separately as well as on pooled data sets. Because the association between patient age and death was statistically significantly different across the studies with aggregate data (test for heterogeneity, p = 0.0001), pooled effects were estimated using a random effects model[39], including 'study' as a factor in the analyses. For all other associations, homogeneity in effects (p > 0.10)[40] could be assumed and fixed effects methods were used.

Calculations were performed using commercially available software (SAS, version 6.12; SAS Institute Inc., Cary, NC or S-plus, version 2000; Insightful Inc., Seattle, WA).

Results

In the four individual patient series, the mean rate of death varied from 23 to 40% and the mean rate of unfavorable outcome varied from 43 to 60%, with the highest proportions in the unselected series of the European Brain Injury Consortium [EBIC] (Table 1). Complete outcome data and the distribution of patients across age groups are listed in Table 2. Patient populations from the 11 studies gathered from the literature were diverse, some being relatively unselected and others highly selected (for example, patients with acute hematomas). Consequently, outcome varied considerably: 22 to 79% for mortality and 41 to 85% for unfavorable outcome.

The proportion of several confounding variables, such as low GCS score, traumatic SAH, and mass lesions, increased with age (Table 3). In contrast, no age-related effects were observed for unreactive pupils, hypotension, and hypoxia.

Shape of the association among individual patient data

Poor outcome increased with age (Table 2 and Figure 1); for example, the mortality rate increased from 21% at an age younger than 35 years to 72% at an age older than 65 years. For unfavorable outcome, these percentages were 39 and 85%, respectively.
The smoothing splines look partly linear and partly quadratic (Figure 1). Adding a slightly more or less liberal penalty factor did not clearly alter the shape of the curve. For ages older than 65 years, the splines were based on only 101 patients (4%), thus implying that the curve is uncertain

Table 2. Frequency of six-month outcome across age groups in four prospective series of patients with severe TBI

Variable (%)	Age Groups (yrs)						
	Total	15-24	25-34	35-44	45-54	55-64	≥ 65
*Outcome**							
No. of patients	2664	908	682	424	324	225	101
Mortality	28	21	21	28	36	43	72
Unfavorable outcome	48	38	41	50	64	69	85
GOS score							
No. of patients	2500	867	622	391	306	214	100
Good Recovery	32	44	35	25	20	17	4
Moderate Disability	19	18	22	23	14	14	11
Severe Disability	15	12	16	17	20	21	9
Vegetative State	4	4	3	5	7	4	3
Death	30	22	23	30	39	45	73

* If the 6-month GOS score[18] was not available (164 cases), unfavorable outcome was assigned, based on GOS values at other points in time

Table 3. Frequency of demographics, clinical characteristics, CT parameters, and 6-month outcome across age groups in four prospective series of patients with severe TBI

Variable (%)	Age Group (yrs)						
	Total	15-24	25-34	35-44	45-54	55-64	≥ 65
No. of patients	2664	908	682	424	324	225	101
Demographics							
High-velocity injury	62	78	65	53	45	35	42
Male sex	77	78	80	77	79	75	62
Clinical characteristics							
GCS score 3-5	38	37	35	40	36	42	52
Pupillary reactivity							
both reactive	64	69	64	63	63	60	44
one reactive	18	14	19	21	17	20	16
both unreactive	18	17	17	16	20	20	40
Hypotension	27	25	26	29	33	25	29
Hypoxia	37	37	37	36	38	40	31
ICP ≥ 20 mm Hg	54	55	53	53	56	53	35
CT scanning parameters							
Traumatic SAH	57	49	57	59	68	69	63
CT scanning classification*							
I or II	36	42	38	31	30	26	26
III or IV	21	25	22	18	17	13	6.9
V or VI	42	32	39	48	52	59	67

* According to Marshall, et al., 1991[34]

above this age. Among survivors, the relative proportion of poor outcomes increased with age: severe disability and vegetative state occurred in just over 20% of the survivors at age 20 years and in approximately 50% of the survivors at age 60 years (Figure 1 and Table 2).

Several continuous transformations of age resulted in a good fit: age as a linear variable reached 90 and 96% of the optimum (the smoothing spline is the reference) for rates of mortality and unfavorable outcome, respectively. The linear plus quadratic (age plus age^2) and linear plus cubic (age plus age^3) transformations resulted in even better fits (Table 4), which were statistically significant for mortality in both the univariable and multivariable analyses and for unfavorable outcome in the multivariable analysis. The absolute differences in estimated probabilities of poor outcome in a comparison of linear and linear plus quadratic, and linear and linear plus cubic age transformations were small, however (Figure 1). The difference in the estimated mortality rate in a comparison of age linear and age linear plus age quadratic transformations was a maximal 3.3% at age 47 years. Unfavorable outcome results were similar (Figure 1).

Table 4. Comparison of the fit of different age transformations in the univariable association between (transformed) age and outcome at 6 months post-TBI

Models for age transformations	Mortality rate				Rate of unfavorable outcome			
	Performance vs smoothing spline		Performance vs age linear		Performance vs smoothing spline		Performance vs age linear	
	Model $\chi 2^{\#}$	% of Optimum†	Gain‡	P-value§	Model $\chi 2^{\#}$	% of Optimum†	Gain‡	P-value§
Age as a smoothing spline	134.4	100	13.7	0.003	168.8	100	6.1	0.08
Continuous transformations								
Age	120.7	90	reference	NA$^{||}$	162.7	96	reference	NA$^{||}$
Age linear plus quadratic	128.9	96	8.2	0.004	166.3	99	3.7	0.06
Age linear plus cubic	128.8	96	8.1	0.004	166.2	98	3.5	0.06
Age thresholds based on the literature								
At 30 years[13,15]	69.5	52	-51.2	NA$^{||}$	100.2	59	-62.4	NA$^{||}$
At 35 and at 55 years[5]	92.2	69	-28.5	< 0.001‡	135.0	80	-27.6	< 0.001‡
At 40 years[14]	96.4	72	-24.3	NA$^{||}$	131.1	78	-31.6	NA$^{||}$
At 50 years[16]	69.5	52	-51.3	NA$^{||}$	95.1	56	-67.6	NA$^{||}$
At 55 years[3]	66.1	49	-54.6	NA$^{||}$	78.3	46	-84.3	NA$^{||}$
At 60 years[17]	64.7	48	-56.0	NA$^{||}$	68.3	40	-94.3	NA$^{||}$
Zero until age 50, linear above this age[4]	96.3	72	-24.4	NA$^{||}$	101.7	60	-61.0	NA$^{||}$
Best fitting age threshold (39 years)	98.0	73	-22.7	NA$^{||}$	141.2	84	-21.5	NA$^{||}$

$^{\#}$ Additional model chi square of the age variable to the fit of the model
§ For a likelihood-ratio test, which tests the gain of the age-transformation
† The fit of the model using the smoothing spline was considered to be optimal (100%) and was used as a reference for this column
‡ The gain of a model is defined as the additional model chi square for the model in question minus that for the model including age as a linear term. If the gain is positive the transformation fits better than age linear; a negative number implies that it fits worse
$^{||}$ Zero degrees of freedom
NA = not applicable

Figure 1. Graph demonstrating the univariable association between age and 6-month outcome in 2664 patients with severe TBI. Age was described as a continuous linear term (age linear), an age linear plus quadratic term, and a smoothing spline. The vertical strokes at the base of the graph indicate the age distribution. For ease of interpretation, the probability scale is presented in this figure, rather than the logistical log-odds scale generally used in logistic regression models. A linear association on the log-odds scale corresponds to a sigmoid curve on the probability scale. Model parameters for age linear (age per 10 years) were as follows: logit (mortality) = -2.18 + 0.34 × age and logit (unfavorable outcome) = -1.34 + 0.37 × age. Model parameters for age linear plus age quadratic (age per 10 years) were as follows: logit (mortality) = -1.26 − 0.18 × age + 0.06 × age^2 and logit (unfavorable outcome) = -0.77 + 0.03 × age + 0.04 × age^2

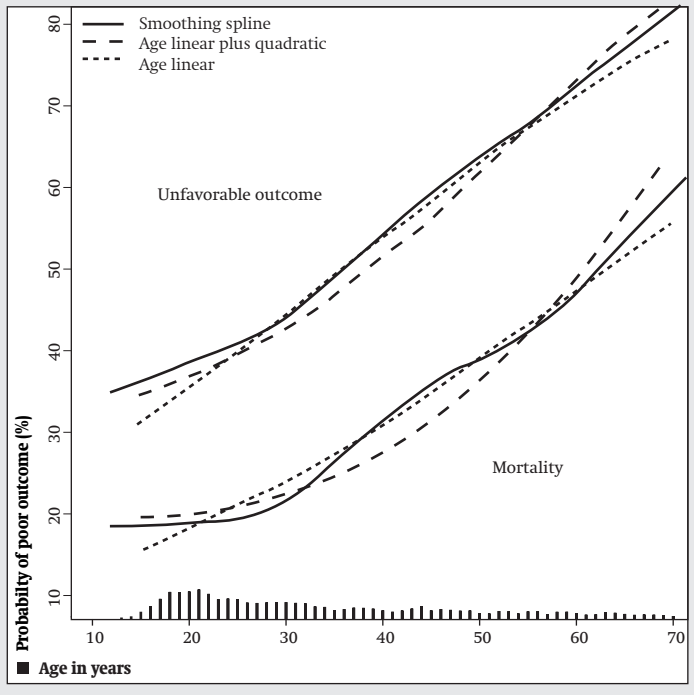


Figure 2. Graph displaying the univariable association between age and 6-month outcome in 2664 patients with severe TBI. Age was described as a discrete variable with a threshold at 39 years. The vertical strokes at the base of the graph indicate the age distribution

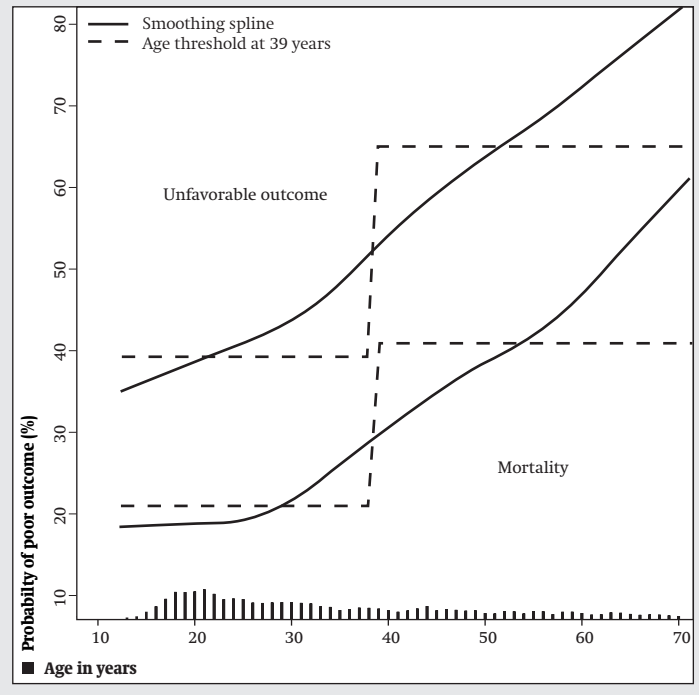
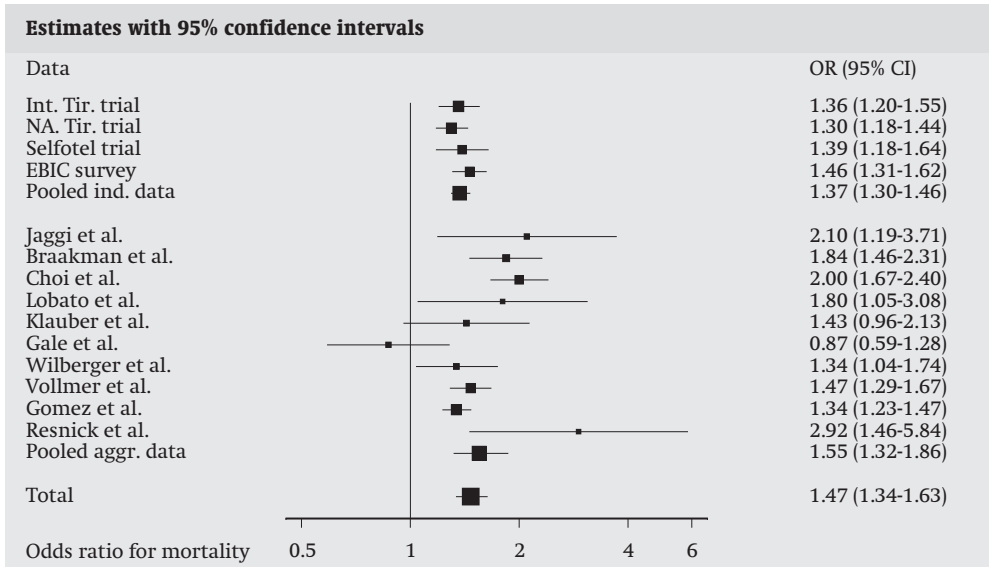
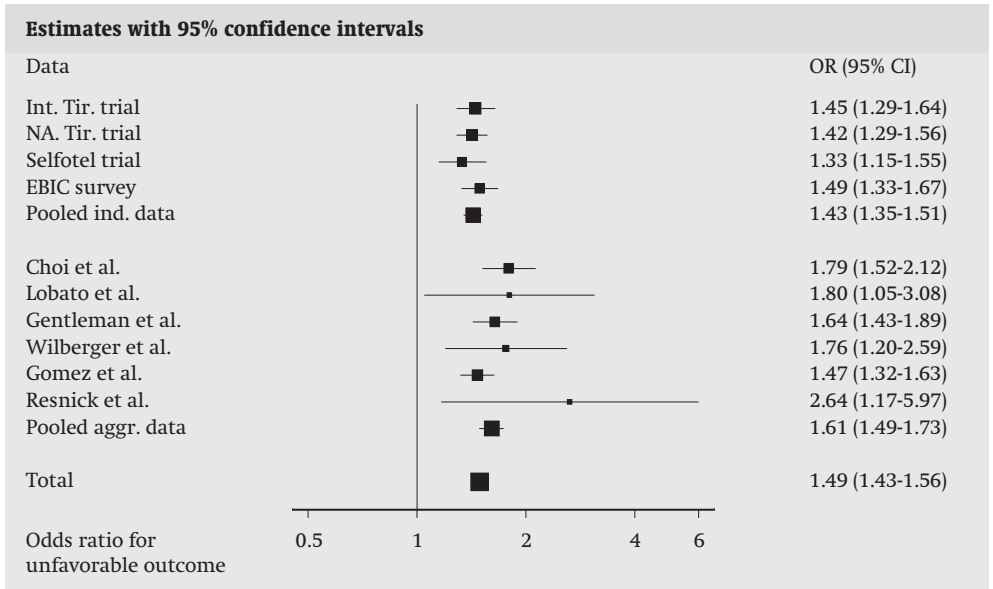


Figure 3. Graphs demonstrating a comparison of the strength of the effect of patient age on mortality (upper) and unfavorable outcome (lower) obtained from individual patient data and aggregate data. Solid squares denote the values for the estimated ORs. Horizontal lines extending to the right and left of the solid squares indicate the 95% CIs. The variation in the CIs is, for the most part, a function of the different sample sizes. Aggr. = aggregate; ind. = individual; Int. Tir. trial = Marshall, et al., 1998 and Hukkelhoven, et al., 2002; NA. Tir. trial = Hukkelhoven, et al., 2002; EBIC survey = Murray, et al., 1999; Selfotel trial = Morris, et al., 1999; OR = odds ratio; CI = confidence interval





The fit of the linear, linear plus quadratic, and linear plus cubic age transformations was consistent in each individual study and remained constant after adjustment for confounding variables, indicating robustness of the findings. After adjustment, age linear resulted in 78% of the optimal goodness-of-fit for the mortality rate and 93% of the optimal fit for unfavorable outcome. Linear plus quadratic and linear plus cubic age transformations yielded 94 and 98% of the optimal goodness-of-fit for rates of mortality and unfavorable outcome, respectively. Other continuous transformations, such as $age^{0.5}$, log(age), and the piecewise transformation[4] (a linear effect above the age 50 years), performed worse.

Optimal age thresholds could be identified accurately, that is, 39 years for both death (95% confidence interval [CI] 39-40 years) and unfavorable outcome (95% CI 39-39 years). These threshold transformations resulted in a maximal 73% (mortality) and 84% (unfavorable outcome) of the optimal fit (Table 3), however. The age threshold value at 39 years is graphically reflected in Figure 2.

Strength of the association

Among the individual patient data, the odds ratios (ORs) were similar, that is, 1.30 to 1.46 for mortality and 1.33 to 1.49 for unfavorable outcome per 10 years of age (Figure 3). Meta-analyses of these data yielded a pooled OR of 1.37 (95% CI 1.30 - 1.46) for mortality and 1.43 (95% CI 1.35 - 1.51) for unfavorable outcome. Thus, the effect of a 10-year increase in age was a multiplication of the odds for mortality with 1.37 and for unfavorable outcome with 1.43. In the individual patient data, we adjusted the age-outcome association for potential confounding variables, as shown in Table 2. The pooled adjusted ORs – 1.39 (95% CI 1.30 - 1.50) for mortality and 1.46 (95% CI: 1.36 - 1.56) for unfavorable outcome – were very similar to the pooled unadjusted ORs. Given that the 95% CIs did not include the value 1, age was independent of other risk factors in patients with severe TBI.

In the aggregate data, the ORs varied considerably: mortality, 0.87 to 2.92; and unfavorable outcome, 1.47 to 2.64 per 10-year increase in age (Figure 3). The pooled ORs were 1.55 (95% CI 1.32 - 1.86) for mortality (10 studies, 2376 cases) and 1.61 (95% CI 1.49 - 1.73) for unfavorable outcome (6 studies, 1982 cases) per 10 years of age.

When combining aggregate and individual patient data, total ORs were 1.47 (95% CI 1.34 – 1.63) for mortality and 1.49 (95% CI 1.43 – 1.56) for unfavorable outcome (Figure 3, *upper* and *lower*).

Discussion

We compared various age transformations to identify simple and accurate descriptions of the associations between age and mortality and age and unfavorable outcome in patients with severe TBI. We found that these associations were continuous. Statistically, age linear plus quadratic transformations fit significantly better than age linear ones. This was primarily caused by the slightly better fit in younger patients, who constituted a large part of the study population. Nonetheless, the absolute difference in the estimated probability of poor outcome comparing age linear and age linear plus quadratic transformations was at most a few percent, which we consider clinically unimportant; therefore, both age linear and age linear plus quadratic transformations are adequate descriptions of the association between patient age and six-month outcome following TBI. A linear relationship between age and outcome has also been reported in patients with aneurysmal SAH[41,42].

The smoothing splines (Figure 1) may also be interpreted as consisting of two linear parts. We found an optimal change point at 60 years for mortality and at 29 years for unfavorable outcome. These change points varied considerably across populations and contained broad CIs, however. Therefore, these piecewise linear transformations are less appropriate to describe the association between patient age and outcome. We also found a clear effect of age in those younger than 50 years, in contrast to a previous conclusion based on data in 372 patients[4].

The use of age thresholds for describing the relationship between patient age and outcome resulted in a considerable loss of information and is therefore not recommended. Hence, our findings challenge the conclusions of authors who have published guidelines claiming that the association between patient age and outcome can be described in a stepwise manner[17]. The reason why authors of previous studies have identified age thresholds with many different values[3-5,13-17] is probably a consequence of the statistical methods used. Arbitrary categorization of age and relatively small numbers of patients in specific age categories means that few patients can change proportion of poor outcome considerably.

Further, the value of the identified age threshold is determined by the distribution of age in the examined patient population. In the present study, with a close to linear association and a very large proportion of the patients between 15 and 65 years of age, thresholds were situated approximately midway, at 39 years, which was partly induced by the age distribution of the examined patient population. When separately analyzing the unselected population of the EBIC study, which contained a relatively greater number of older patients, threshold values included higher ages (that is, 59 years for mortality and 45 years for unfavorable outcome).

It may be hypothesized that increased mortality at an older age is in part caused by an increased number of (possibly extracerebral) medical complications, which would be expected to increase late mortality. The median time to death did not differ between patients younger than and older than 50 years of age, however ($p = 0.75$, tirilazad data). Moreover, in both age groups the primary cause of death was cerebral, and no clear difference was noted in the frequency of extracerebral (systemic) causes of death.

In accordance with several data from several other studies[3,5,42] , we observed that the proportion of survivors with poor outcome (for example, severe disability or vegetative state) increased with age and that the proportion of patients with favorable outcomes declined. These results support the hypothesis that the adult brain has a decreased capacity for repair as it ages[43], because of a decreasing number of functioning neurons and a greater exposure to minor repetitive (often subclinical) insults to the brain as age increases. In adults, however, diminished cognitive or behavioral function may be influenced beneficially by regeneration or plasticity of the brain[44,45]. Further investigation of the physiological and pathophysiological features in the aging brain is required to identify new medical interventions that perhaps could prevent the poorer outcome associated with advanced age.

Several limitations of our analyses should be acknowledged. First, the individual patient data consisted mainly of selected populations from randomized clinical trials; therefore specific subgroups, such as patients with GCS motor score of 1 or an age older than 65 years, were underrepresented. Whether our findings may be extrapolated to these or other subgroups of patients with closed TBI is uncertain. Nonetheless, applying results to specific patient categories may be valid, given that statistical interactions between patient age and important confounding variables were not significant. Second, the less detailed study parameters – for example, the age categories – in the aggregate data may have led to less reliable ORs. The ORs of the pooled individual data did not differ considerably from the ORs of the pooled aggregate data, however. The ORs were generally 1.4 to 1.5 per 10 years of age and univariable and adjusted ORs were similar, although we cannot exclude possible effects of other confounding factors that were not (consistently) present in the data set, such as extracranial injuries[4].

Our study has a number of implications. First, a better estimate for the odds on poor outcome is obtained, that is, a 40 to 50% increase per 10 years of age, which is independent of the presence of risk factors. Furthermore, the existence of threshold values was not supported. Finally, the association between patient age and outcome after severe TBI is a continuous function, which can be adequately described by an age linear term or even statistically better by an age linear plus quadratic term. We therefore advise applying one of these transformations in future studies on adults with severe TBI for purposes of prognostic modeling or adjusting for confounding variables.

Acknowledgements

We gratefully acknowledge the permission granted by the principal investigators for accessing the data sets for purposes of the current study. We greatly appreciate all the work performed by the investigators and study personnel, without whose input the current analysis could not have been performed.

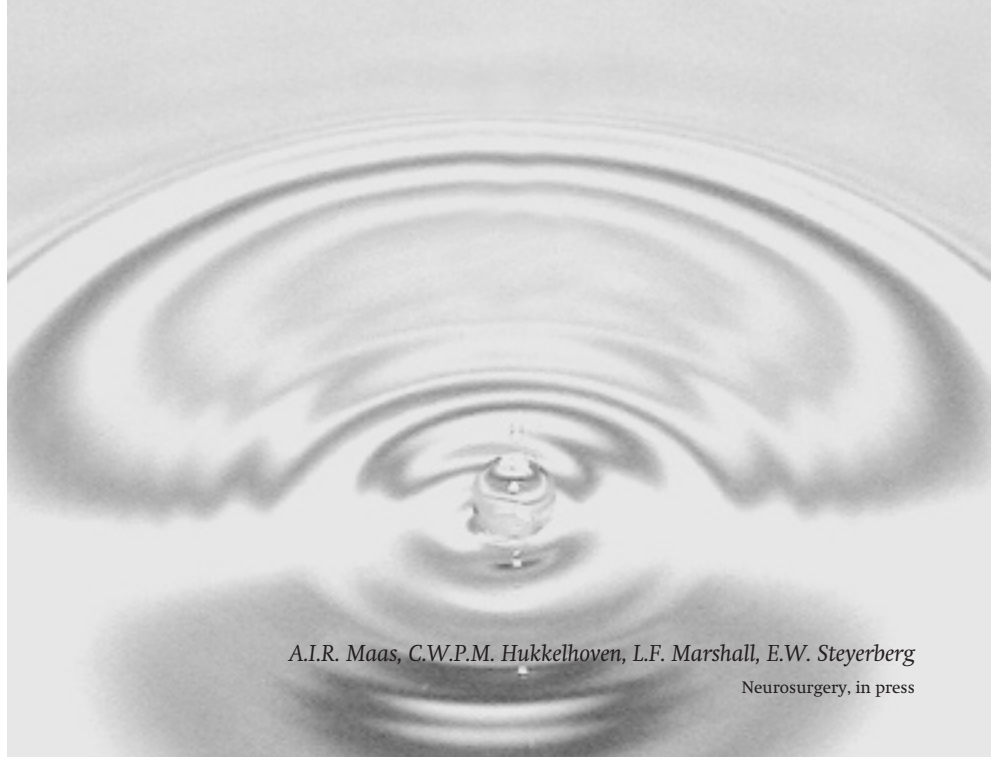
References

1. Braakman R, Gelpke G, Habbema JDF, Maas AIR, Minderhoud J. Systematic selection of prognostic features in patients with severe head injury. Neurosurgery 1980;6:362-370.
2. Choi S, Ward J, Becker D. Chart for outcome prediction in severe head injury. J Neurosurg 1983;59(2):294-297.
3. Gómez P, Lobato R, Boto G, De la Lama A, González P, de la Cruz J. Age and outcome after severe head injury. Acta Neurochir (Wien) 2000;142:373-381.
4. Signorini D, Andrews P, Jones P, Wardlaw J, Miller J. Predicting survival using simple clinical variables: a case study in traumatic brain injury. J Neurol Neurosurg Psychiatry 1999;66:20-25.
5. Vollmer D, Torner J, Jane J, et al. Age and outcome following traumatic coma: why do older patients fare worse? J Neurosurg 1991;75:S37-S49.
6. Laketta E. Age-associated cardiovascular changes in health: impact on cardiovascular disease in older persons. Heart Fail Rev 2002;7:29-49.
7. van Reekum R, Simard M, Cohen T. The prediction and prevention of Alzheimer's disease - towards a research agenda. J Psychiatry Neurosci 1999;24:413-430.
8. Seeman T, Guralnik J, Kaplan G, Knudson L, Cohen R. The Alameda County Study: the health consequences of multiple morbidity in the elderly. J Aging Health 1989;1:50-66.
9. Guralnik J. Assessing the impact of comorbidity in the older population. Ann Epidemiol 1996;6:376-380.
10. Yancik R, Wesley M, Ries L, Havlik R, Edwards R, Yates J. Effect of age and comorbidity in postmenopausal breast cancer patients aged 55 years and older. JAMA 2001;285:885-892.
11. Narayan R, Greenberg R, Miller J, et al. Improved confidence of outcome prediction in severe head injury. J Neurosurg 1981;54:751-762.
12. Teasdale GM, Skene A, Parker L, Jennett B. Age and outcome of severe head injury. Acta Neurochir Suppl (Wien) 1979;28:140-143.
13. Overgaard J, Hvid-Hansen O, Land A, et al. Prognosis after head injury based on early clinical examination. Lancet 1973;2(7830):631-635.
14. Piek J, Chesnut R, Marshall LF, et al. Extracranial complications of severe head injury. J Neurosurg 1992;77:901-907.
15. Resnick D, Marion D, Carlier P. Outcome analysis of patients with severe head injuries and prolonged intracranial hypertension. J Trauma 1997;42(6):1108-1111.
16. Waxman K, Sundine M, Young R. Is early prediction of outcome in severe head injury possible? Arch Surg 1991;126:1237-1242.
17. Chesnut R, Ghajar J, Maas AIR, et al. Management and prognosis of severe traumatic brain injury. Part 2: early indicators of prognosis in severe traumatic brain injury. J Neurotrauma 2000;17(6/7):557-627.
18. Jennett B, Bond M. Assessment of outcome after severe brain damage. A practical scale. Lancet 1975;1:480-484.
19. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, management and outcome: experience from the Tirilazad trials. J Neurosurgery 2002;97:549-557.
20. Marshall LF, Maas AIR, Marshall S, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 1998;89:519-525.

21. Morris G, Bullock R, Marshall S, Marmarou A, Maas AIR, Marshall LF. Failure of the competitive N-methyl-D-aspartate antagonist Selfotel (CGS 19755) in the treatment of severe head injury: result of two phase III clinical trials. The Selfotel investigators. J Neurosurg 1999;91:737-743.
22. Murray GD, Teasdale GM, Braakman R, et al. The European Brain Injury Consortium survey of head injuries. Acta Neurochir (Wien) 1999;141:223-236.
23. Jaggi J, Obrist W, Gennarelli T, Langfitt T. Relationship of early cerebral blood flow and metabolism to outcome in acute head injury. J Neurosurg 1990;72:176-182.
24. Choi S, Muizelaar J, Barnes T, Marmarou A, Brooks D, Young H. Prediction tree for severely head-injured patients. J Neurosurg 1991;75:251-255.
25. Lobato R, Rivas J, Cordobes F, et al. Acute epidural hematoma; an analysis of factors influencing the outcome of patients undergoing surgery in coma. J Neurosurg 1988;68:48-57.
26. Klauber M, Marshall LF, Barrett-Connor E, Bowers S. Prospective study of patients hospitalized with head injury in San Diego County, 1978. Neurosurgery 1981;9:236-241.
27. Gentleman D. Causes and effects of systemic complications among severely head injured patients transferred to a neurosurgical unit. Int Surg 1992;77:297-302.
28. Gale J, Dikmen S, Wyler A, Temkin N, McLean A. Head injury in the Pacific Northwest. Neurosurgery 1983;12:487-491.
29. Wilberger J, Harris M, Diamond D. Acute subdural hematoma: morbidity and mortality related to timing of operative intervention. J Trauma 1990;30(6):733-736.
30. Royston P. A strategy for modelling the effect of a continuous covariate in medicine and epidemiology. Statist Med 2000;19:1831-1847.
31. Reinsch C. Smoothing by spline functions. Numer Math 1967;10:177-183.
32. Hastie T, Tibshirani R. Generalised additive models. London: Chapman and Hall, 1990.
33. Hardy R, Thompson S. A likelihood approach to meta-analysis with random effects. Stat Med 1996;15:619-629.
34. Marshall LF, Gautille T, Klauber M, et al. The outcome of severe closed head injury. J Neurosurg 1991;75:S28-S36.
35. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
36. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227 - 1237.
37. Clayton D, Hills M. Statistical Models in Epidemiology. Oxford: Oxford University Press, 1993.
38. Steyerberg EW, Kievit J, de Mol Van Otterloo J, van Bockel J, Eijkemans MJC, Habbema JDF. Perioperative mortality of elective abdominal aortic aneurysm surgery. A clinical prediction rule based on literature and individual patient data. Arch Intern Med 1995;155:1998-2004.
39. DerSimonian R, Laird N. Meta-analysis in clinical trials. Control Clin Trials 1986;7:177-188.
40. Petitti D. Approaches to heterogeneity in meta-analysis. Statist Med 2001;20:3625-3633.
41. Kassell N, Torner J, Jane J, Adams H, Kongable G. The International Cooperative Study on the Timing of Aneurysm Surgery. Part 1: Overall management results. J Neurosurgery 1990;73:18-36.
42. Lanzino G, Kassell N, Germanson T, et al. Age and outcome after aneurysmal subarachnoid hemorrhage: why do older patients fare worse? J Neurosurgery 1996;85:410-418.
43. Carlsson C, Von Essen C, Lofgren J. Factors affecting the clinical course of patients with severe head injuries. I. Influence of biological factors. II Significance of post-traumatic coma. J Neurosurg 1968;29:242-248; 248-251.
44. Marshall LF. Head injury: past, present and future. J Neurosurgery 2000;47:546-561.
45. Peterson D. Stem cells in brain plasticity and repair. Curr Opin Pharmacol 2002;2:34-42.

5

Prediction of outcome in traumatic brain injury with CT characteristics: a comparison between the CT-classification and combinations of CT predictors



A.I.R. Maas, C.W.P.M. Hukkelhoven, L.F. Marshall, E.W. Steyerberg

Neurosurgery, in press

Abstract

Background and objective
The Marshall CT-classification identifies six groups of patients with Traumatic Brain Injury (TBI), based on morphologic abnormalities on the CT scan. This classification is increasingly used as a predictor of outcome. We aimed to examine the predictive value of the Marshall CT-classification in comparison with alternative CT models.

Methods
The predictive value was investigated in the Tirilazad trials (n=2269). Alternative models were developed with logistic regression analysis and recursive partitioning. Six-month mortality was used as outcome measure. Internal validity was assessed with bootstrapping techniques and expressed as the area under the receiver-operating curve (AUC).

Results
The Marshall CT-classification showed reasonable discrimination (AUC = 0.67), which could be improved by rearranging the underlying individual CT characteristics (AUC = 0.71). Performance could be further increased by adding intraventricular and traumatic subarachnoid hemorrhage and by a more detailed differentiation of mass lesions and basal cisterns (AUC = 0.77). Models developed with logistic regression analysis and recursive partitioning showed similar performance. For clinical application we propose a simple CT score, which permits a more clear differentiation of prognostic risk, particularly in patients with mass lesions.

Conclusion
It is preferable to use combinations of individual CT predictors rather than the Marshall CT-classification for prognostic purposes in TBI. Such models should include at least the following parameters: status of basal cisterns, shift, traumatic subarachnoid or intraventricular hemorrhage, and presence of different types of mass lesions.

Introduction

Classification of traumatic brain injury (TBI) is necessary to accurately describe patient series and requires grouping of patients according to specific characteristics. In clinical practice, the clinical severity of TBI is generally classified as severe, moderate or mild according to the level of consciousness as measured with the Glasgow Coma Scale (GCS). The increased use of early sedation, intubation and ventilation in more severe patients has decreased the value of the full GCS for purposes of classification[1-3]. Alternatively, in more severe patients, TBI can be classified according to morphologic criteria based on CT or MRI investigations. Although MRI may be more sensitive for detecting small white matter lesions in a later phase after TBI, CT examination remains the investigation of choice in the acute phase[4,5].

Conventional classification of TBI with CT findings differentiates between focal and diffuse injuries[6,7]. In 1991 Marshall et al.[8], following analysis of the Traumatic Coma Data Bank, proposed a CT-classification for grouping patients with TBI according to multiple CT characteristics. This CT-classification identifies six different groups of patients with TBI, based on the type and severity of several abnormalities on the CT scan. It differentiates between patients with and without mass lesions and permits a further discrimination of patients with diffuse injuries into four categories, taking into account signs of raised ICP (i.e. compressed or absent basal cisterns, midline shift). Since its introduction this CT-classification has become widely accepted for descriptive purposes, and is also increasingly being used as major predictor of outcome in TBI. Various studies have confirmed the predictive value of the CT-classification[9-11], and the international guidelines on prognosis include the CT-classification as a major CT predictor based on class I evidence[12]. Whether the Marshall CT-classification is best suited for prediction or whether other combinations of CT parameters may be more appropriate for this specific purpose has not been investigated in detail.

The aim of the present study was to examine the prognostic performance of the Marshall CT-classification in comparison with other combinations of CT predictors in TBI, by re-weighting and refining the CT characteristics used to determine this classification and by including additional CT parameters.

Patients and methods

The Marshall CT-classification

The Marshall CT-classification is presented in Table 1. Discriminating features in this classification are 1) presence or absence of mass lesions, 2) presence or absence of intracranial abnormalities 3) CT signs of raised intracranial pressure (status of basal cisterns, shift), and 4) planned evacuation of mass lesions.

To facilitate comparison with alternative classifications we translated the Marshall CT-classification into a binary tree (Figure 1).

Figure 1 Marshall CT-classification presented in a tree structure

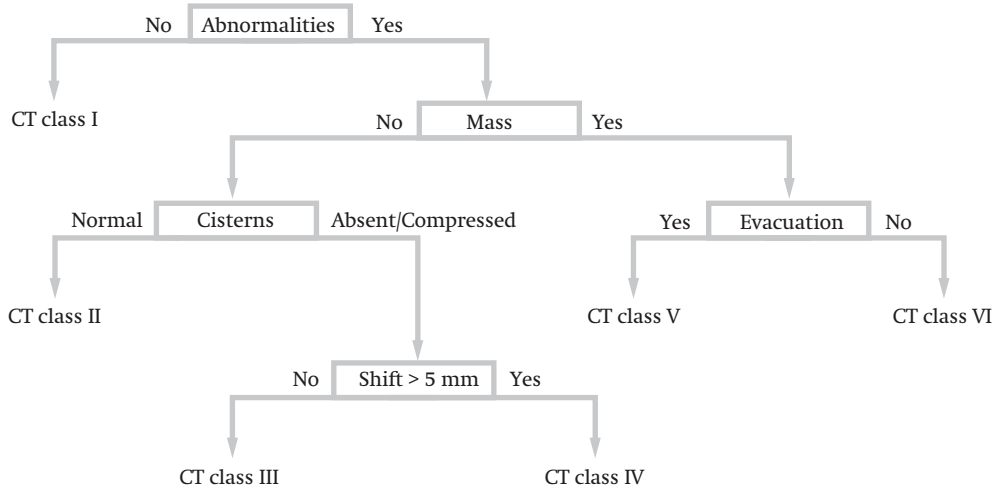


Table 1 Diagnostic categories of types of abnormalities visualized on computed tomography (CT) scanning

Category	Definition
Diffuse injury I	no visible intracranial pathology seen on CT scan
Diffuse injury II	cisterns are present with midline shift 0-5 mm and/or: - lesion densities present, - no high- or mixed-density lesion > 25 cc, - may include bone fragments and foreign bodies
Diffuse injury III	cisterns compressed or absent with midline shift 0-5 mm, no high- or mixed-density lesion > 25 cc
Diffuse injury IV	midline shift > 5 mm, no high- or mixed-density lesion > 25 cc
Evacuated mass lesion	any lesion surgically evacuated
Non-evacuated mass lesion	high- or mixed-density lesion > 25 cc, not surgically evacuated

Patients

Our studies were conducted on the combined data sets of the International and North American Tirilazad trials (n=2269). Details on the Tirilazad trials have been reported elsewhere[13,14]. Inclusion and exclusion criteria of the two trials were similar. Both trials included patients between 15 and 65 years of age with severe (GCS 3-8) or moderate (GCS 9-12) closed TBI. With respect to the CT characteristics the inclusion criteria varied slightly: the international study excluded moderate TBI patients with a normal CT scan, whereas the North American study excluded such patients only when the blood alcohol level exceeded 0,2 g/dl. Protocols and recommendations for management were comparable for both trials.

In both trials the efficacy of Tirilazad mesylate, an amino-steroid that displays an antioxidant effect, was studied against that of placebo. We combined data from placebo and treatment groups, since in neither trial a significant difference between the Tirilazad and the placebo treated group was shown for the primary outcome measure, i.e. mortality and unfavorable outcome on the Glasgow Outcome Scale.

Definitions of CT characteristics and outcome

We based our studies on data recorded for admission CT scans performed within the first 4 hours after injury. Full data on CT characteristics on admission were available in 2249 patients. CT data were extracted on the following items:

- CT-classification
- presence of abnormalities
- presence and size of midline shift
- status of basal cisterns
- presence of intraventricular blood (IVH)
- presence of traumatic subarachnoid hemorrhage (tSAH)
- presence and type of mass lesions and expected evacuation of mass lesion
- single versus multiple non-mass lesions

The characteristics ‘any abnormalities’, ‘intraventricular blood’, ‘tSAH’ and ‘expected evacuation of mass lesions’ were available in the data sets as binary data and scored as present or absent, without further differentiation. The other CT characteristics were classified into several categories with increasing differentiation: midline shift was classified in two ways: (i) shift ≤ 5 mm versus shift > 5 mm and (ii) no shift, shift of 1-5 mm, shift of 6-10 mm, or shift > 10 mm. The status of basal cisterns was categorized in two ways as (i) normal versus abnormal (compressed and absent) and (ii) normal versus compressed versus absent. Presence and type of mass lesions were categorized in 3 ways as (i), mass lesion present or absent (ii), absent mass lesion, epidural mass lesion, intradural (intracerebral plus subdural) mass lesion and (iii) absent mass lesion, epidural mass lesion (EDH), subdural mass lesion (SDH), intracerebral mass lesion.

For several patients values of some of the predictors were missing (4,8% of the required values). These values were statistically estimated with regression models including the other predictors and subsequently imputed[15,16]. This approach is considered preferable to complete case analysis, in which patients with missing values are excluded from analysis[15]. The outcome measure was mortality at six months post-injury.

Table 2 Distribution of CT parameters and mortality in the International and in the North American sample

CT parameters and outcome		Total (n=2249)		International sample (n=1112)		North American sample (n=1137)	
		n	%	n	%	n	%
CT parameters							
Abnormalities	No	173	8%	52	5%	121	11%
	Yes	2076	92%	1060	95%	1016	89%
	Missing	0		0		0	
Traumatic SAH	No	1030	47%	519	48%	511	46%
	Yes	1171	53%	573	52%	604	54%
	Missing	42		20		22	
Intraventricular blood	No	1746	79%	863	79%	883	79%
	Yes	473	21%	235	21%	238	21%
	Missing	30		14		16	
Basal cisterns	Normal	1194	54%	576	52%	618	56%
	Compressed	709	32%	375	34%	334	30%
	Absent	296	13%	143	13%	153	14%
	Missing	50		18		32	
Midline shift	None	1426	64%	654	59%	772	70%
	1 - 5 mm	362	16%	208	19%	154	14%
	6 - 10 mm	233	11%	121	11%	112	10%
	> 10 mm	190	9%	118	11%	72	6.5%
	Missing	38		11		27	
Lesion	No	360	16%	117	11%	243	22%
	Yes	1872	84%	995	89%	877	78%
	One, not mass	691	37%	429	29%	262	30%
	Multiple, not	585	31%	329	33%	256	29%
	Mass lesion	729	39%	370	37%	359	41%
	Epidural*	204	28%	129	35%	75	21%
	Subdural*	418	57%	205	55%	213	59%
	Intracerebral*	584	80%	277	75%	307	86%
	Missing	21		4		17	
CT-classification	I	173	8%	52	5%	121	11%
	II	833	37%	425	38%	408	36%
	III	426	19%	219	20%	207	18%
	IV	88	4%	46	4%	42	4%
	V	539	24%	289	26%	250	22%
	VI	190	8%	81	7%	109	10%
	Missing	0		0		0	
Outcome							
Mortality	Yes	491	22%	270	24%	221	19%
	No	1758	78%	842	76%	916	81%
	Missing	0		0		0	

* More than one (type of) mass lesion on the CT scan was possible

Statistical analysis and performance of models

To test whether the arrangements of the CT characteristics within the Marshall CT-classification was reasonable for predictive purposes, we developed alternative models with the same variables. Subsequently, we investigated whether performance could be improved by adding CT characteristics or by separating already included characteristics into smaller categories. Each model was developed with two methods: recursive partitioning (CART)[17] and logistic regression analysis. We chose these two approaches for different reason: logistic regression analysis is a standard statistical procedure, in which interactions and relative importance of predictors are considered. Results are generally robust, but the underlying methodology may remain unclear to non-statisticians. On the other hand recursive partitioning, in which prediction trees are created, has a greater clinical appeal, but carries some risk of overfitting, especially in more complex trees. To correct for this, the trees were pruned using cross validation.

Modeling was performed with SAS software (version 6.12, SAS Institute INC., Cary, NC) and S-plus (version 2000, Insightful Corporation, Seattle, WA), using the RPART library. The RPART library (rpart2.zip) and manual (rpart2doc.zip) can be found at http://www.stats.ox.ac.uk/pub. Swin.

Internal validity of the original CT-classification and alternative models was assessed with bootstrapping procedures. Internal validity assesses whether the models perform well for a population of patients similar to those for whom the model was developed. Bootstrapping involved taking samples 100 times with replacement from the development sample. Each sample can be considered as repeating the data collection with the same number of patients and under identical circumstances as the original. In each of the 100 bootstrap samples a regression model was estimated, and evaluated on the original sample, applying a shrinkage factor to correct for optimism[15,18-20]. In this way nearly unbiased predictions of outcome can be obtained for future but similar patients[15,19,21].

Performance of the models was assessed with respect to discrimination, which can be quantified by the area under the receiver-operating curve (AUC). For a randomly chosen pair of patients, the AUC represents the probability that a patient who dies has a higher predictive probability for mortality. The higher the AUC, the better the model discriminates. A model with an AUC of 0.50 has no discriminative power, while a model with an AUC of 1.0 reflects perfect discrimination.

Application in clinical practice

Presentation of a classification according to a prediction tree is readily understandable for a clinical audience. Interpretation of a logistic regression model is more complicated. To facilitate application of these models we created a score chart to estimate the outcome probability based on the values of the regression coefficient, which were re-scaled and rounded to whole numbers.

Results

Individual CT characteristics and outcome

The distribution of CT characteristics and outcome is presented in Table 2. Mortality was lower in the North American patients (19%) than in the International patients (24%), partly reflecting a different distribution of severe versus moderate patients included in both trials. More than 90% of the patients had abnormalities on the admission CT scan: 84% showed evidence of parenchymal or extracerebral lesions, 45% had abnormal basal cisterns, 53% tSAH and 21% had intraventricular blood. Mass lesions were present in 39% of the population, and of these 80% had an associated intracerebral lesion. 74% of all mass lesions were evacuated and of these 84% were evacuated within 4 hours of injury.

Midline shift, basal cisterns, intraventricular blood and traumatic subarachnoid hemorrhage were identified as significant predictors of mortality (Table 3). In the multivariable analysis the full differentiation of lesions was not clearly associated with differences in mortality, but the differentiation between epidural and intradural lesions was highly relevant (Table 3, Table 4).

Prognostic value of CT-classification versus alternative groupings of individual components

Figure 2 presents the classification of our patient population according to the Marshall CT-classification in a prediction tree format with mortality figures per class. The percentage mortality in patients with no abnormalities (CT class 1) and in patients with diffuse injuries without radiological signs of raised ICP (CT class 2) was low (6.4 and 11% respectively). The highest mortality rate was observed in patients with absent or compressed basal cisterns and a midline shift larger than 5 mm (CT class 4): 44%. Mortality rates for patients with mass lesions were 30% for those with evacuated mass lesions and 34% for those with non-evacuated mass lesions. Analysis of the discriminatory properties of the Marshall CT-classification showed an AUC of 0.669.

Figure 3 presents a prediction tree constructed with recursive partitioning, using the same characteristics and the same number of terminal nodes as used in the Marshall CT-classification. We found that a primary division according to status of the basal cisterns yields the strongest discrimination. Subsequently for patients with present basal cisterns a split on abnormalities and for patients with absent or compressed basal cisterns a split on shift > 5 mm caused the maximum reduction in heterogeneity. Discriminatory analysis showed an AUC 0.705, considerably higher than found for the original CT-classification.

Alternative models with additional variables

We investigated whether models could be developed with better discriminative properties by adding additional CT predictors not originally included in the Marshall CT-classification or by further separation of already included CT characteristics.

– continues on page 103 –

Table 3 Multivariable analysis of CT characteristics, pooled Tirilazad patients

Characteristics	Categorization	Mortality	OR (95% CI)*
Abnormalities	No	6.4%	Reference
	Yes	23%	1.0 (0.5 – 2.0)
Shift	No shift	17%	Reference
	0 - 5 mm	26%	1.4 (1.0 – 1.9)
	6 – 10 mm	36%	1.6 (1.1 – 2.4)
	> 10 mm	49%	2.0 (1.3 – 3.1)
Basal cisterns	Normal	15%	Reference
	Compressed	27%	2.0 (1.5 – 2.7)
	Absent	55%	5.7 (4.0 – 8.0)
Intraventricular blood	No	19%	Reference
	Yes	31%	2.0 (1.5 – 2.6)
TSAH	No	12%	Reference
	Yes	30%	2.0 (1.5 – 2.5)
Lesions	No	12%	Reference
	Single non-mass	15%	0.9 (0.6 – 1.4)
	Multiple non-mass	23%	1.3 (0.9 – 1.9)
	Epidural mass	17%	0.5 (0.4 – 0.9)
	Subdural mass	40%	1.4 (0.9 – 2.1)
	Intracerebral mass	35%	1.1 (0.7 – 1.7)

* OR = Odds Ratio, 95% CI = 95% Confidence Interval

Figure 2 Mortality for each category of the Marshall CT-classification, Tirilazad cohort (n=2249)

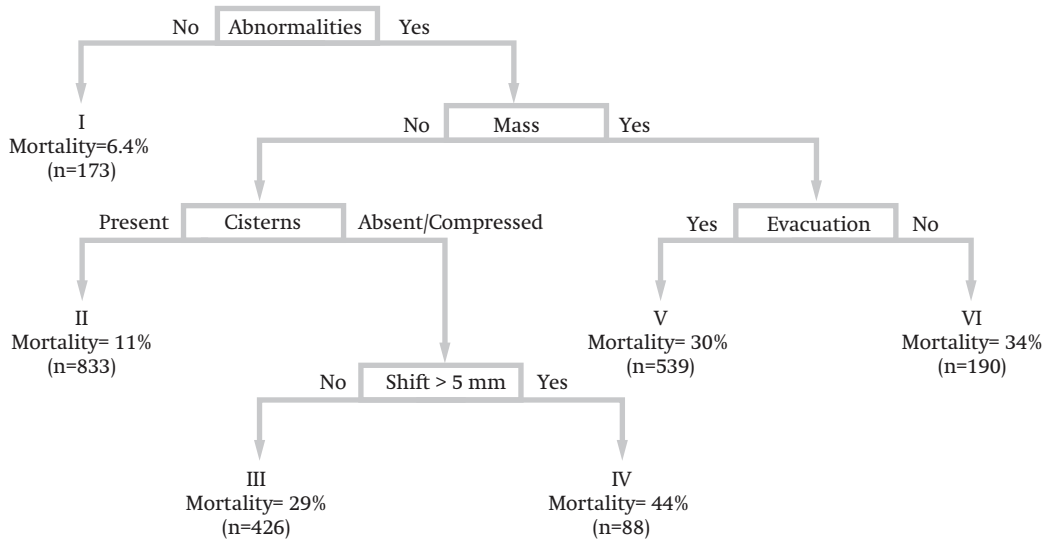


Figure 3 Mortality for each category of the CT prediction tree constructed with recursive partitioning, using the same characteristics and the same number of terminal nodes as in the Marshall CT-classification. Construction occurred on the Tirilazad cohort (n=2249)

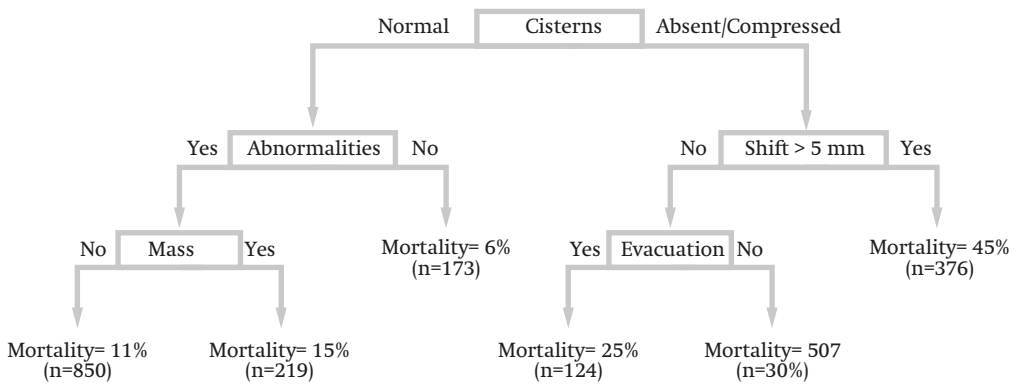


Table 4 Added discriminative value of a) extra CT characteristics and b) further differentiation of CT characteristics

Models	Discrimination (AUC*)	
	Logistic regression	Recursive partitioning
Basic model #	0.703	0.705
Added CT characteristic		
Evacuation mass lesion	0.714	0.712
Dropped CT characteristics		
Any abnormalities	0.703	0.701
Added CT characteristics		
Non-mass lesions (single or multiple)	0.714	0.712
Blood (tSAH and/or intraventricular blood)	0.730	0.737
Separation of already included CT characteristics		
Mass as epidural versus intradural	0.720	0.719
Mass as epidural vs subdural vs intracerebral	0.722	0.721
Cisterns (normal, compressed, absent)	0.726	0.727
Shift (no, 1-5 mm, 6-10 mm, > 10 mm)	0.710	0.716
All ‡	0.769	0.794

* AUC = Area under the receiver operating curve
\# The basic model contains all characteristics included in the Marshall CT-classification, except evacuation mass lesion
‡ All: Characteristics of the basic model plus single non mass lesions, multiple non mass lesions, mass as epidural vs. subdural vs. intracerebral, cisterns as normal vs. compressed vs. absent, shift as 1-5 mm vs. 6-10 mm vs. > 10 mm, blood as tSAH vs. intraventricular blood

Table 5 Prognostic score chart for the probability of mortality in patients with severe or moderate TBI according to their CT characteristics

Predictor	Value	Score
Basal cisterns	Normal	0
	Compresssed	1
	Absent	2
Midline shift	No shift or shift ≤ 5 mm	0
	Shift > 5mm	1
Epidural mass lesion	Present	0
	Absent	1
Intraventricular blood or tSAH	Absent	0
	Present	1
Sum score*: add relevant scores		..... +1

* The sum score can be used to obtain the predicted probability of mortality from the formula below. We chose to add plus 1 to make the grading numerically consistent with the grading of the motor score of the GCS and with the Marshall CT-classification. The corresponding probabilities are calculated with the formula: Probability (mortality) = $1/[1+ e^{-(-2.60+ 0.80 *Sumscore)}]$

Table 6 CT-classification by prediction score

Score	Nr of patients	Actual mortality	
		n	%
1	36	0	0
2	600	41	6.8
3	773	122	16
4	465	121	26
5	261	138	53
6	114	69	61

Table 7 Marshall CT-classification compared to the CT prediction score

CT prediction score	Marshall CT-classification						
	1	2	3	4	5	6	Total
1	0	0	0	0	35	1	36
2	173	336	0	0	65	26	600
3	0	492	107	5	95	74	773
4	0	5	249	19	136	56	465
5	0	0	70	37	134	20	261
6	0	0	0	27	74	13	114
Total	173	833	426	88	539	190	2249

The added benefit of additional parameters was assessed versus the basic model presented in Figure 3. Results are summarized in Table 4. The discriminative ability could be improved considerably by adding tSAH and intraventricular blood, and by further differentiating the basal cisterns, midline shift and mass lesions into several categories. Dropping the characteristic 'any abnormalities' had negligible influence on the AUC. No statistically significant interactions were observed between the selected characteristics (p = 0.42).

The results of this multivariable analysis showed the potential for developing an alternative model with added characteristics. Including all discriminating variables in a logistic regression model yielded an AUC of 0.769, and of 0.794 in the prediction tree. Such a model is however complex and resulted in 15 terminal nodes in the prediction tree. Searching for an appropriate compromise between good discrimination and easy clinical applicability we chose a simpler model based on the following characteristics: midline shift (subdivided into 0-5 mm, > 5 mm), basal cisterns (subdivided into absent, compressed and present), mass lesion (subdivided into epidural and intradural), traumatic subarachnoid hemorrhage and/or intraventricular blood. The apparent validity of this model was 0.750 and on internal validation we obtained an AUC of 0.748.

Clinical application

For clinical application we translated the logistic regression model into a score chart, with which the probability for mortality according to the CT characteristics can be estimated by adding the scores for individual patients (Table 5). We chose to add plus 1 to the sum score in order to make the grading numerically consistent with the grading of the motor score of the GCS and with the Marshall CT-classification. Table 6 shows the application of this score chart for classifying the study population according to prognostic risk. The difference in observed mortality rates between patients from the lowest and patients from the highest risk group is 61%, which is considerably larger than the maximal difference in mortality in the Marshall CT-classification (38%, Figure 2). Table 7 illustrates the better discrimination for prognostic risk assessment of the CT prediction score in comparison to the Marshall CT-classification, particularly in patients with mass lesions.

Discussion

We confirmed the predictive value of the Marshall CT-classification in a large series of patients (n=2249), but showed that a better discrimination can be obtained by making fuller use of the individual CT characteristics underlying the Marshall CT-classification. Discrimination could be further improved by adding intraventricular and traumatic subarachnoid hemorrhage and by a more detailed differentiation of mass lesions and basal cisterns (AUC = 0.77). We do not wish to detract from the general validity and appeal of the Marshall CT-classification when used for descriptive purposes. This classification however was not developed from the perspective of prognosis, and the question whether the categorization of variables in the Marshall CT-classification is appropriate for predictive purpose is relevant. For instance in the Marshall CT-classification, radiological signs of raised ICP (status of basal cisterns and presence of shift) are only used for further differentiation of patients with diffuse injuries whilst these parameters may also be expected to be of prognostic value particularly in patients with mass lesions. Indeed, Table 7 shows a better prognostic discrimination of the proposed prognostic CT score over the Marshall CT-classification, particularly in patients with mass lesions.

The presence of tSAH has been shown to be a strong predictor both for outcome and mortality in TBI, but is not included in the Marshall CT-classification[22-28]. The predictive value of tSAH in TBI is confirmed in our study and we have additionally shown that including this parameter in a predictive model significantly increases discrimination; we also found IVH to be an independent predictor, in contrast to other studies in which the relation of IVH to poorer outcome was mainly caused by the association with other predictors[22,29,30]. Further, the Marshall CT-classification does not permit any distinction on type of mass lesion. Many studies have shown that prognosis in patients with an EDH is much better than in those with a subdural or intracerebral hematoma[6,12]. Bricolo et al.[31] postulated that mortality should approach zero in patients with an uncomplicated EDH. As shown in Table 6 we found zero mortality in such patients (mass lesion with a prognostic score of 1). A further problem with the original CT-classification is that it differentiates between patients with evacuated versus non-evacuated mass lesions. Many have argued that this reflects a clinical decision and does not in itself constitute a CT-parameter, and in clinical practice this has led to confusion and it has been proposed not to include this differentiation[11]. Nevertheless we do note a 4% difference in mortality between patients with evacuated versus non-evacuated mass lesions. Further in depth adjusted analysis however will be required to determine whether the baseline characteristics of these two groups were similar or not.

Previous studies have shown that the Marshall CT-classification is a strong predictor in TBI[3,9-12,25], with high inter- and intra observer reliability[32]. Wardlaw et al.[33], however, found in a retrospective analysis of 425 patients of varying severity that the Marshall CT-classification did not remain a significant independent outcome predictor on multivariate analysis when clinical features were included, in contrast to tSAH and a newly suggested, ill-defined variable describing 'overall appearance'. For the present study we did not include clinical characteristics in our models, but in a previous study describing a prediction model for TBI, we found that both the Marshall CT-classification and tSAH remained as statistically significant predictors in multivariate analysis, following adjustment for clinical variables[9].

Consistent with other prediction studies in TBI[34], we found that performance of the models was more dependent on the variables included than on the statistical approach. We found no clear statistical benefit in the use of a prediction tree compared to logistic regression models. We considered the use of a prediction tree in the current analysis appropriate as the Marshall CT-classification can be readily presented as prediction tree, which may be appealing for the clinician. Further, a tree can capture and correct easily for interaction, i.e. different relations between predictors in different subgroups. Interaction, if present, is easily detected by a better discriminative ability of the tree as compared to a logistic model. We did not observe such differences in our studies. The clinical appeal of a prediction tree method is however also dependent on the number of terminal nodes. The limited number of nodes (n=6) in the Marshall CT-classification and in the basic model makes this type of presentation appropriate. When additional variables were added to our model we found an optimal number of 15 terminal nodes, which significantly decreases the clinical appeal. For this reason we would prefer the logistic regression model, also as discriminative properties are similar. We realize that a logistic regression model may have less clinical appeal and therefore suggest to translate it into a score chart as proposed in table 6. Although this score chart performed well in our study population, assessment of its general applicability will require validation in other data sets.

A number of limitations of our study should be recognized. First, our studies were performed on a large patient series including only patients with severe and moderate injury. Results can therefore not be extrapolated towards patients with mild injuries. Second, we focused our studies on analysis of data from the initial CT examination performed within 4 hours after injury. Other studies[10,11] have shown that the 'worst' CT scan obtained during the clinical course has greater predictive value. Also within the current data set we found that the final CT-classification, based on the worst CT following admission, yielded better discrimination (AUC = 0.692 for Marshall CT-classification and 0.716 for basic model). The intent of our studies, however, was to investigate the use of the CT-classification and CT predictors toward a prognostic classification of TBI on admission. Such classification is considered useful to establish the baseline characteristics and prognostic risk of TBI patients on admission. Third, the predictive analysis presented was conducted versus six-month mortality. For these studies we chose mortality rather than the GOS dichotomized into unfavorable versus favorable as this constitutes a hard and objective endpoint without any missing outcome data. As a sensitivity analysis we additionally calculated the discriminative properties of the Marshall CT-classification, the basic model in which the individual parameters of the CT-classification were rearranged and the extended model versus unfavorable outcome and found similar results.

In summary, we conclude that the Marshall CT-classification has strong predictive power, but greater discrimination can be obtained if the individual CT parameters, underlying the CT-classification are included in a prognostic model. Consequently for prognostic purposes we recommend the use of individual characteristics rather than the CT-classification. Performance of CT models for predicting outcome in TBI can be significantly improved by including more details of variables and by adding other variables to the model. We suggest that such models should include the following characteristics: status of basal cisterns, shift, tSAH and/or IVH and presence of mass lesions with differentiation between EDH versus intradural lesions. For more easy clinical application models can be translated into a score chart.

Acknowledgements

The authors express their gratitude to all of the study participants and the principal investigators of the Tirilazad trials whose work made this report possible. The authors further wish to thank Marja van Gemerden for administrative assistance. Grant support was provided by NIH-NS 42691.

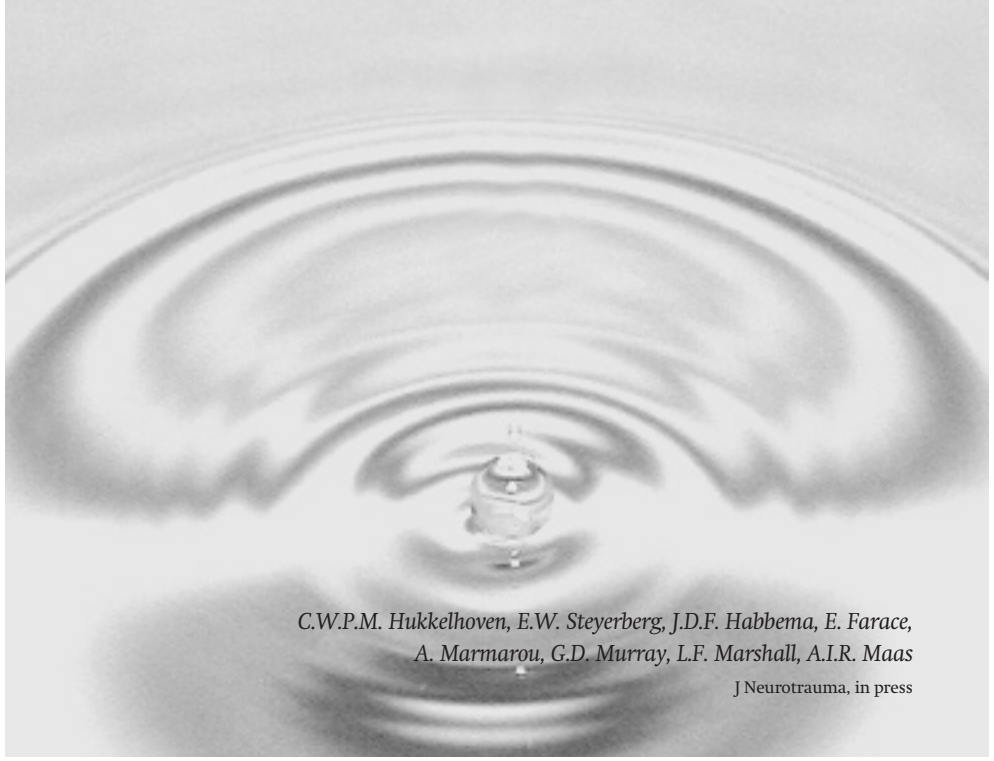
References

1. Buechler CM, Blostein PA, Koestner A, Hurt K, Schaars M, McKernan J. Variation among trauma centers' calculation of Glasgow Coma Scale score: results of a national survey. J Trauma 1998;45(3):429-32.
2. Balestreri M, Czosnyka M, Chatfield DA, et al. Predictive value of Glasgow Coma Scale after brain trauma: change in trend over the past ten years. J Neurol Neurosurg Psychiatry 2004;75(1):161-2.
3. Moskopp D, Stahle C, Wassmann H. Problems of the Glasgow Coma Scale with early intubated patients. Neurosurg Rev 1995;18(4):253-7.
4. Firsching R, Woischneck D, Klein S, Reissberg S, Dohring W, Peters B. Classification of severe head injury based on magnetic resonance imaging. Acta Neurochir (Wien) 2001;143(3):263-71.
5. Uchino Y, Okimura Y, Tanaka M, Saeki N, Yamaura A. Computed tomography and magnetic resonance imaging of mild head injury - is it appropriate to classify patients with Glasgow Coma Scale score of 13 to 15 as "mild injury"? Acta Neurochir (Wien) 2001;143(10):1031-7.
6. Gennarelli TA, Spielman GM, Langfitt TW, et al. Influence of the type of intracranial lesion on outcome from severe head injury. J Neurosurg 1982;56(1):26-32.
7. Lobato RD, Cordobes F, Rivas JJ, et al. Outcome from severe head injury related to the type of intracranial lesion. A computerized tomography study. J Neurosurg 1983;59(5):762-74.
8. Marshall LF, Marshall SB, Klauber MR, et al. A new classification of head injury based on computerized tomography. J Neurosurg 1991;75:S14-S20.
9. Hukkelhoven CWPM, Steyerberg EW, Habbema JDF, et al. Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics. J Neurotrauma 2005;in press.
10. Lobato RD, Gomez PA, Alday R, et al. Sequential computerized tomography changes and related final outcome in severe head injury patients. Acta Neurochir (Wien) 1997;139(5):385-91.
11. Servadei F, Murray GD, Penny K, 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;46(1):70-5; discussion 75-7.
12. Chesnut R, Ghajar J, Maas AIR, et al. Management and prognosis of severe traumatic brain injury. Part 2: early indicators of prognosis in severe traumatic brain injury. J Neurotrauma 2000;17:557-627.
13. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. J Neurosurg 2002;97(3):549-57.
14. Marshall LF, Maas AIR, Marshall SB, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 1998;89(4):519-25.
15. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
16. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227-1237.
17. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Belmont, California: Wadsworth International Group, 1984.
18. Efron B, Tibshirani F. An introduction to the bootstrap. New York: Chapman and Hall, 1993.
19. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer-Verlag Inc., 2001.

20. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9(11):1303-25.
21. Steyerberg EW, Eijkemans MJC, Harrell FE, Jr., Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
22. Eisenberg HM, Gary HE, Jr., Aldrich EF, et al. Initial CT findings in 753 patients with severe head injury. A report from the NIH Traumatic Coma Data Bank. J Neurosurg 1990;73(5):688-98.
23. Greene KA, Jacobowitz R, Marciano FF, Johnson BA, Spetzler RF, Harrington TR. Impact of traumatic subarachnoid hemorrhage on outcome in nonpenetrating head injury. Part II: Relationship to clinical course and outcome variables during acute hospitalization. J Trauma 1996;41(6):964-71.
24. Greene KA, Marciano FF, Johnson BA, Jacobowitz R, Spetzler RF, Harrington TR. Impact of traumatic subarachnoid hemorrhage on outcome in nonpenetrating head injury. Part I: A proposed computerized tomography grading scale. J Neurosurg 1995;83(3):445-52.
25. Kakarieka A, Braakman R, Schakel EH. Clinical significance of the finding of subarachnoid blood on CT scan after head injury. Acta Neurochir (Wien) 1994;129(1-2):1-5.
26. 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;8(2):120-3.
27. Selladurai BM, Jayakumar R, Tan YY, Low HC. Outcome prediction in early management of severe head injury: an experience in Malaysia. Br J Neurosurg 1992;6(6):549-57.
28. Servadei F, Murray GD, Teasdale GM, et al. Traumatic subarachnoid hemorrhage: demographic and clinical study of 750 patients from the European brain injury consortium survey of head injuries. Neurosurgery 2002;50(2):261-7; discussion 267-9.
29. Cordobes F, de la Fuente M, Lobato RD, et al. Intraventricular hemorrhage in severe head injury. J Neurosurg 1983;58(2):217-22.
30. Lee JP, Lui TN, Chang CN. Acute post-traumatic intraventricular hemorrhage analysis of 25 patients with emphasis on final outcome. Acta Neurol Scand 1991;84(2):85-90.
31. Bricolo AP, Pasut LM. Extradural hematoma: toward zero mortality. A prospective study. Neurosurgery 1984;14(1):8-12.
32. Vos PE, van Voskuilen AC, Beems T, Krabbe PF, Vogels OJ. Evaluation of the traumatic coma data bank computed tomography classification for severe head injury. J Neurotrauma 2001;18(7):649-55.
33. Wardlaw JM, Easton VJ, Statham P. Which CT features help predict outcome after head injury? J Neurol Neurosurg Psychiatry 2002;72(2):188-92; discussion 151.
34. Titterington DM, Murray GD, Murray LS, et al. Comparison of discrimination techniques applied to a complex data set of head injured patients. J Royal Statist Soc 1981;144:145-175.

6

Predicting outcome after traumatic brain injury: development and validation of a prognostic score based on admission characteristics



C.W.P.M. Hukkelhoven, E.W. Steyerberg, J.D.F. Habbema, E. Farace, A. Marmarou, G.D. Murray, L.F. Marshall, A.I.R. Maas

J Neurotrauma, in press

Abstract

Context
Early prediction of outcome after traumatic brain injury (TBI) is important for several purposes, but no prognostic models have yet been developed with proven generalizability across different settings.

Objective
To develop and validate prognostic models that use information available at admission to estimate six-month outcome after severe or moderate TBI.

Main outcome measures
Mortality and unfavorable outcome, i.e. death, vegetative state or severe disability on the Glasgow Outcome Scale, at six months post-injury.

Design, setting and patients
We used prospectively collected data on 2269 patients from two multi-center clinical trials to develop prognostic models for each outcome with logistic regression analysis. We included seven predictive characteristics, i.e. age, motor score, pupillary reactivity, hypoxia, hypotension, CT-classification and traumatic subarachnoid hemorrhage. The models were validated internally with bootstrapping techniques. External validity was determined in prospectively collected data from two relatively unselected surveys in Europe (n=796) and in North America (n=746). We evaluated the discriminative ability, i.e. the ability to distinguish patients with different outcomes, with the area under the receiver operating characteristic curve (AUC). Further, we determined calibration, i.e. agreement between predicted and observed outcome, with the Hosmer-Lemeshow goodness-of-fit test.

Results
The models discriminated well in the development population (AUC 0.78 to 0.80). External validity was even better (AUC 0.83 to 0.89). Calibration was less satisfactory, with poor external validity in the North American survey ($p<0.001$). Especially, observed risks were higher than predicted for poor prognosis patients. A score chart was derived from the regression models to facilitate clinical application.

Conclusions
Relatively simple prognostic models using baseline characteristics can accurately predict six-month outcome in patients with severe or moderate TBI. The high discriminative ability indicates the potential of this model for classifying patients according to prognostic risk.

Introduction

Traumatic brain injury (TBI) is a major cause of death and disability among a predominantly young population. In the USA each year about 52.000 people die and 80.000 remain severely disabled after TBI[1]. Early prediction of outcome may support clinical decision-making and resource allocation and may provide realistic and evidence-based expectations to 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 randomized clinical trials (RCTs)[2].

Several prognostic models have been developed to predict long term outcome for patients with severe or moderate TBI[3-16]. Many of the models pre-date the general availability of Computed Tomography (CT) scans or include data obtained after admission. Also, the prognostic value of the commonly included eye- and verbal score of the Glasgow Coma Scale is nowadays restricted due to current management approaches such as early sedation and intubation[17-19]. Furthermore, many studies used relatively small sample sizes originating from a single center or region, which limits generalizability. The model may perform well in the development sample, but poorly when applied to other groups of patients, e.g. patients from another region. Validation of a prognostic model on another patient series should therefore be considered essential[20]. Unfortunately, such external validity has seldom been assessed in the field of TBI[21]. We can therefore only have limited confidence in previously developed prognostic models for outcome prediction in current patients with TBI.

We aimed to develop and validate prognostic models for six-month outcome, i.e. mortality and unfavorable outcome according to the Glasgow Outcome Scale (GOS)[22]. We used large patient series from a broad range of western countries. We consider a model based on easily accessible clinical features and CT scanning, available on admission.

Patients and methods

Patients

Two selected patient populations were used for model development, i.e. the patients included in the International and in the North American multi-center (phase III) RCTs on the drug Tirilazad Mesylate in TBI[23,24]. These are the largest available data sets on severe and moderate TBI. The models were validated in two relatively unselected populations of patients consecutively admitted with severe or moderate TBI, i.e. the core data survey conducted by the European Brain Injury Consortium (EBIC)[25] and the Traumatic Coma Data Bank (TCDB)[26].

The protocols and the inclusion and exclusion criteria of the two Tirilazad trials were virtually identical. Both trials enrolled patients aged 15-65 years, with a severe (Glasgow Coma Scale[27] (GCS) 3-8) or moderate (GCS 9-12) closed TBI. Patients with an absent motor score or with a moderate TBI and a normal CT scan were excluded. All patients were admitted to a neurosurgical unit within four hours after injury. Recommendations for patient management were similar across the centers. The International Tirilazad trial (n=1120) was conducted in 40 centers in Europe, Israel and Australia from 1992 to 1994 and the North American Tirilazad trial (n=1149) in 36 centers in the USA and Canada from 1991 to 1994. By protocol, the proportion of patients with moderate TBI was lower in the International trial (14% versus 28% in the North American trial). Details on the Tirilazad trials have been reported[23,24]. Both Tirilazad trials studied the efficacy of Tirilazad Mesylate, an amino-steroid with anti-oxidant effect, against a placebo[28]. We combined data from placebo and treatment groups, since in both trials no statistically significant difference between the Tirilazad and the placebo treated group was shown for the primary outcome measure.

The survey by the European Brain Injury Consortium (EBIC) included 796 patients with severe or moderate TBI[25], consecutively collected between February and April 1995 from 55 European centers in which the six-month outcome assessment was routinely performed. The National Traumatic Coma Data Bank (TCDB) contained data on 746 patients with non penetrating severe TBI (GCS ≤ 8) admitted to four centers in the USA[29]. Data acquisition occurred from 1984 to 1988. In the surveys patients were included if they deteriorated to a condition meeting enrollment criteria within 24 or 48 hours respectively.

Definitions of potential predictors and outcome

We considered patient characteristics that were previously identified as important predictors[30,31], and that could be determined easily and reliably within the first hours after injury. These included age, gender, cause of injury, motor score, pre-admission hypotension and hypoxia, pupillary reactivity, CT-classification[32] and the presence of traumatic subarachnoid hemorrhage (tSAH). Cause of injury was categorized into traffic accidents, falls and other causes. Hypotension and hypoxia were considered present when documented by a systolic blood pressure below 90 mm Hg or a pO_2 below 60 mm Hg respectively, or if supported by strong clinical suspicion. Pupillary reactivity and motor score were measured on admission. For the motor score, we combined category 'no response' with 'extensor response' and category 'localizing' with 'obeying commands' to increase numbers per class.

The CT-classification and the presence of tSAH were derived from the admission CT scan, obtained within the first four hours after the injury. The CT-classification[32] systematically groups a spectrum of abnormalities on the CT scan, such as diffuse injury, midline shift, obliterated basal cisterns and mass lesions into six categories. Since only few patients were classified into category I ('no visible abnormalities'), this category was combined with category II. The categories V ('mass lesion evacuated') and VI ('mass lesion non-evacuated') were joined to eliminate possible differences in decision-making concerning the evacuation of mass lesions.

For several patients, values of some of the predictors were missing (4.8% of the required values). These values were statistically estimated with regression models including the other predictors, and subsequently imputed[33,34]. Such imputation is recommended as more efficient than dropping cases[33]. The true variability among predictor values is only slightly underestimated because of the small numbers of missing values. We also developed models with complete cases only, and found only minor differences in the model characteristics.

The two outcome measures were mortality and unfavorable outcome at six months post-injury. Data on mortality were available for all 2269 Tirilazad patients. Data on unfavorable outcome, derived from the GOS measurement, were available for 2137 patients, including 120 estimated outcomes based on GOS measurements at other points in time[23]. The EBIC survey contained data on six-month mortality and unfavorable outcome. In the TCDB the GOS assessment was frequently outside the window of six months +/- one month. We therefore chose to limit this analysis to mortality, since mortality generally occurs early and almost always within the first few months.

Model development

We used logistic regression analysis to estimate the association between a predictor and outcome, expressed as an odds ratio (OR). Predictors have statistically significant effects ($p < 0.05$) if the 95% confidence interval (95% CI) of the OR does not include the value one.

For both outcome measures (mortality and unfavorable outcome) a multivariable model was developed containing predictors that had a p-value < 0.20 in a backward stepwise procedure[35]. In a relatively large patient population, this liberal p-value will exclude only those characteristics with low predictive value[36].

Age was included as a linear and a squared term[37]. Regression coefficients of the clinical and CT predictors were similar across the two Tirilazad populations, as confirmed by statistically non-significant interaction terms between predictors and population. Therefore, merging of the two populations was considered legitimate.

Performance

The performance of the models 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. For example, if patients with certain characteristics are predicted to have a 10% chance of mortality, the actually observed mortality should also be 10%. Calibration was assessed graphically

by plotting observed outcome against the predicted probability. A smooth, non-parametric calibration line was created with the lowess algorithm[36]. Calibration was tested with the Hosmer-Lemeshow goodness-of-fit test, which assesses agreement between predicted and observed risks over the full range of predicted probabilities. Patients were grouped per decile of predicted risk to perform the test, which means that each group contained 10% of the patients.

Discrimination is the model's ability to separate patients with different outcomes. To quantify the discrimination we used the area under the receiver operating characteristic curve (AUC). The AUC evaluates whether those with higher predicted risk are more likely to have a poor outcome (mortality/unfavorable outcome) among all possible pairs of patients with different outcomes. A model with an AUC of 0.50 has no discriminative power at all (such as a coin flip), and an AUC of 1.0 reflects perfect discrimination (such as a test without false-positive or false-negative results).

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[35]. Internal validity of the models was assessed with standard bootstrapping procedures[36,38]. Bootstrapping involves drawing samples of patients with replacement from the development sample. 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. Regression models were estimated in 200 bootstrap samples. These models were each evaluated on the original sample. The average difference in AUC was determined to indicate the optimism in the initially estimated discriminative ability[33]. In addition, we reduced the regression coefficients to provide better predictions for future patients[33,35]. The reduction was based on a shrinkage factor as estimated from the bootstrap validation procedure.

We further assessed the external validity of the models. Predictions of outcome were generated with the internally validated Tirilazad models for patients in the EBIC and TCDB data sets. Discriminative ability and calibration were assessed to indicate the performance of the developed models in patients from another setting. As a secondary analysis, we adjusted the model intercepts so that the average predictions were correct for the EBIC and TCDB patients. This adjustment improves calibration, but does not influence discrimination. Calculations were performed using SAS software (version 6.12, SAS Institute INC., Cary, NC) and S-plus (version 2000, Insightful Corporation, Seattle, WA).

Application in clinical practice

To facilitate application of the models in clinical practice, we created a score chart to estimate the outcome probability. The score chart was based on the values of the shrunk regression coefficients, which were re-scaled and rounded to whole numbers[39]. The re-scaling and rounding was such that the performance remained similar to that of the original model. We estimated 95% confidence intervals for predicted probabilities with each score, based on the average standard errors of predictions for patients with similar scores.

Table 1 Patient characteristics and outcome in the International and the North American Tirilazad trials, the survey of the European Brain Injury Consortium (EBIC) and the North American survey of the Traumatic Coma Data Bank (TCDB)

Characteristics and outcome		Development patients				Validation patients			
		International Tirilazad trial (n = 1120)		North American Tirilazad trial (n = 1149)		European survey (EBIC) (n = 796)		North American survey (TCDB) (n = 746)	
		n	%	n	%	n	%	n	%
Potential predictors									
Age	Mean (SD)	33.6 (14.6)		32.8 (12.4)		40.9 (20.6)		33.1 (16.4)	
Gender	Male	853	76%	904	79%	594	75%	575	77%
	Missing	0		0		1		0	
Cause of injury	Traffic accidents	668	60%	651	57%	422	53%	560	75%
	Falls	215	19%	171	15%	96[#]	12%[#]	132	18%
	Other	237	21%	327	28%	278	35%	45	6%
Motor score	None/extensor	141	13%	162	14%	203	34%	295	42%
	Abnormal flexion	236	21%	138	12%	48	8%	89	13%
	Withdraws	329	29%	330	29%	97	16%	142	20%
	Localizes/obeys	414	37%	519	45%	251	42%	177	25%
	Missing	0		0		197		43	
Pupillary reactivity	Both pupils reacted	760	72%	641	68%	482	66%	262	63%
	One pupil reacted	168	16%	111	12%	65	8.9%	27	7%
	No pupil reacted	122	12%	186	20%	186	25%	129	31%
	Missing	70		211		63		328	
Hypoxia	Yes or suspected	149	15%	280	28%	213	27%	360	48%
	Missing	130		133		5		0	
Hypotension	Yes or suspected	155	14%	240	21%	180	23%	310	42%
	Missing	30		31		6		0	
CT-classification*	I - II	477	43%	529	47%	324	41%	197	29%
	III	219	20%	207	18%	72	9%	145	21%
	IV	46	4%	42	4%	19	2%	36	5%
	V - VI	370	33%	359	32%	367	47%	309	45%
	Missing	8		12		14		59	
Traumatic subarachnoid haemorrhage	Yes	575	52%	515	46%	314	41%	148	23%
	Missing	22		24		37		113	
Six-month outcome									
Mortality	Yes	275	25%	225	20%	244	31%	326	44%
Unfavorable outcome	Yes	457	42%	405	38%	388	49%	na[‡]	na[‡]
	Missing	33		87		0			

* CT-classification: I = no visible intracranial pathology on CT scan, II = midline shift 0-5 mm, III = cisterns compressed or absent with midline shift 0-5 mm, IV = midline shift > 5 mm, V = any lesion surgically evacuated, VI = high- or mixed-density lesion > 25 mm, not surgically evacuated

[#] Only falls under influence of alcohol were registered as falls

[‡] na = not applicable.

Table 2 Associations between predictors and six-month outcome in the pooled Tirilazad patients

Predictors	Coding	Mortality (n = 2269)				Unfavorable outcome (n = 2149)			
		n	mortality	OR‡uni (95% CI§)	OR‡multi (95% CI§)	n	unfavorable outcome	OR‡uni (95% CI§)	OR‡multi (95% CI§)
Age*	22 years	na#	na#	Reference	Reference	na#	na#	Reference	Reference
	43 years	na#	na#	1.7 (1.5 – 2.0)	1.6 (1.4 – 1.9)	na#	na#	1.9 (1.7 – 2.2)	2.0 (1.7 – 2.3)
Gender	Male	1757	379 (22%)	Reference	-	1650	651 (39%)	Reference	-
	Female	512	121 (24%)	1.1 (0.9 – 1.4)	-	499	211 (42%)	1.1 (0.9 – 1.4)	-
Cause of injury	Traffic	1319	263 (20%)	Reference	-	1271	469 (37%)	Reference	-
	Falls	386	119 (31%)	1.8 (1.4 – 2.3)	-	370	172 (46%)	1.5 (1.2 – 1.9)	-
	Other causes	564	118 (21%)	1.1 (0.8 – 1.4)	-	508	221 (44%)	1.3 (1.1 – 1.6)	-
Motor score	None/extensor	303	133 (44%)	5.9 (4.3 – 7.9)	3.3 (2.3 – 4.7)	293	212 (72%)	9.0 (6.6 – 12)	5.7 (4.0 – 8.1)
	Abnormal flexion	374	118 (32%)	3.5 (2.6 – 4.6)	2.6 (1.9 – 3.7)	365	205 (56%)	4.4 (3.4 – 5.7)	3.8 (2.8 – 5.1)
	Withdraws	659	139 (21%)	2.0 (1.5 – 2.6)	1.7 (1.3 – 2.3)	622	249 (40%)	2.3 (1.8 – 2.9)	2.0 (1.6 – 2.6)
	Localizes/obeys	933	110 (12%)	Reference	Reference	869	196 (23%)	Reference	Reference
Pupillary reactivity	Both pupils reacted	1623	251 (15%)	Reference	Reference	1518	459 (30%)	Reference	Reference
	One pupil reacted	320	104 (33%)	2.7 (2.0 – 3.6)	1.4 (1.1 – 2.0)	317	178 (56%)	2.9 (2.3 – 3.8)	1.5 (1.1 – 2.0)
	No pupil reacted	326	145 (44%)	4.5 (3.5 – 5.9)	1.9 (1.4 – 2.6)	314	225 (72%)	5.9 (4.5 – 7.7)	2.5 (1.6 – 2.8)
Hypoxia	No	1818	346 (19%)	Reference	Reference	1709	612 (36%)	Reference	Reference
	Yes or suspected	451	154 (34%)	2.3 (1.8 – 2.9)	1.5 (1.2 – 2.0)	440	250 (57%)	2.5 (2.0 – 3.1)	1.6 (1.3 – 2.1)
Hypotension	No	1872	355 (19%)	Reference	Reference	1764	637 (36%)	Reference	Reference
	Yes or suspected	397	145 (37%)	2.5 (2.0 – 3.2)	2.2 (1.7 – 2.9)	385	225 (58%)	2.5 (2.0 –3.2)	2.1 (1.6 – 2.8)
CT-classification||	I + II	1006	104 (10%)	Reference	Reference	932	238 (26%)	Reference	Reference
	III	426	123 (29%)	3.5 (2.6 – 4.6)	2.3 (1.7 – 3.2)	411	203 (49%)	2.8 (2.2 – 3.6)	1.9 (1.5 – 2.5)
	IV	88	39 (44%)	6.6 (4.2 – 10.5)	4.1 (2.5 – 6.8)	86	48 (56%)	3.5 (2.2 – 5.5)	1.9 (1.1 – 3.2)
	V + VI	729	225 (31%)	3.8 (3.0 – 4.9)	2.4 (1.8 – 3.2)	701	361 (51%)	3.0 (2.5 – 3.7)	1.9 (1.5 – 2.4)
Traumatic subarachnoid haemorrhage	No	1056	132 (13%)	Reference	Reference	994	272 (27%)	Reference	Reference
	Yes	1213	368 (30%)	3.1 (2.5 – 3.8)	2.0 (1.5 – 2.5)	1155	590 (51%)	2.8 (2.3 – 3.3)	1.8 (1.5 – 2.3)

* 22 years is the 25%-percentile and 43 years the 75%-percentile, including age as a continuous linear plus quadratic term
\# na (not applicable), since percentiles were used for describing the association between age and outcome
‡ OR = Odds Ratio, uni = via univariable logistic regression analysis, multi = via multivariable logistic regression analysis
§ 95% CI = 95% Confidence Interval. If this interval does not include the value 1, the factor has a statistically significant effect on the outcome
|| CT-classification I = no visible intracranial pathology on CT scan, II = midline shift 0-5 mm, III = cisterns compressed or absent with midline shift 0-5 mm, IV = midline shift > 5 mm, V = any lesion surgically evacuated, VI = high- or mixed-density lesion > 25 mm, not surgically evacuated

Results

Patient characteristics and outcome

The distribution of patient characteristics and outcome in the four patient populations is presented in Table 1. Poor outcome was lowest in the North American Tirilazad population (20% mortality and 38% unfavorable outcome). In the International Tirilazad population more patients died (25%) or had an unfavorable outcome (42%). We observed the poorest outcomes in the relatively unselected populations: 31% mortality and 49% unfavorable outcome in the EBIC and 44% mortality in the TCDB data sets.

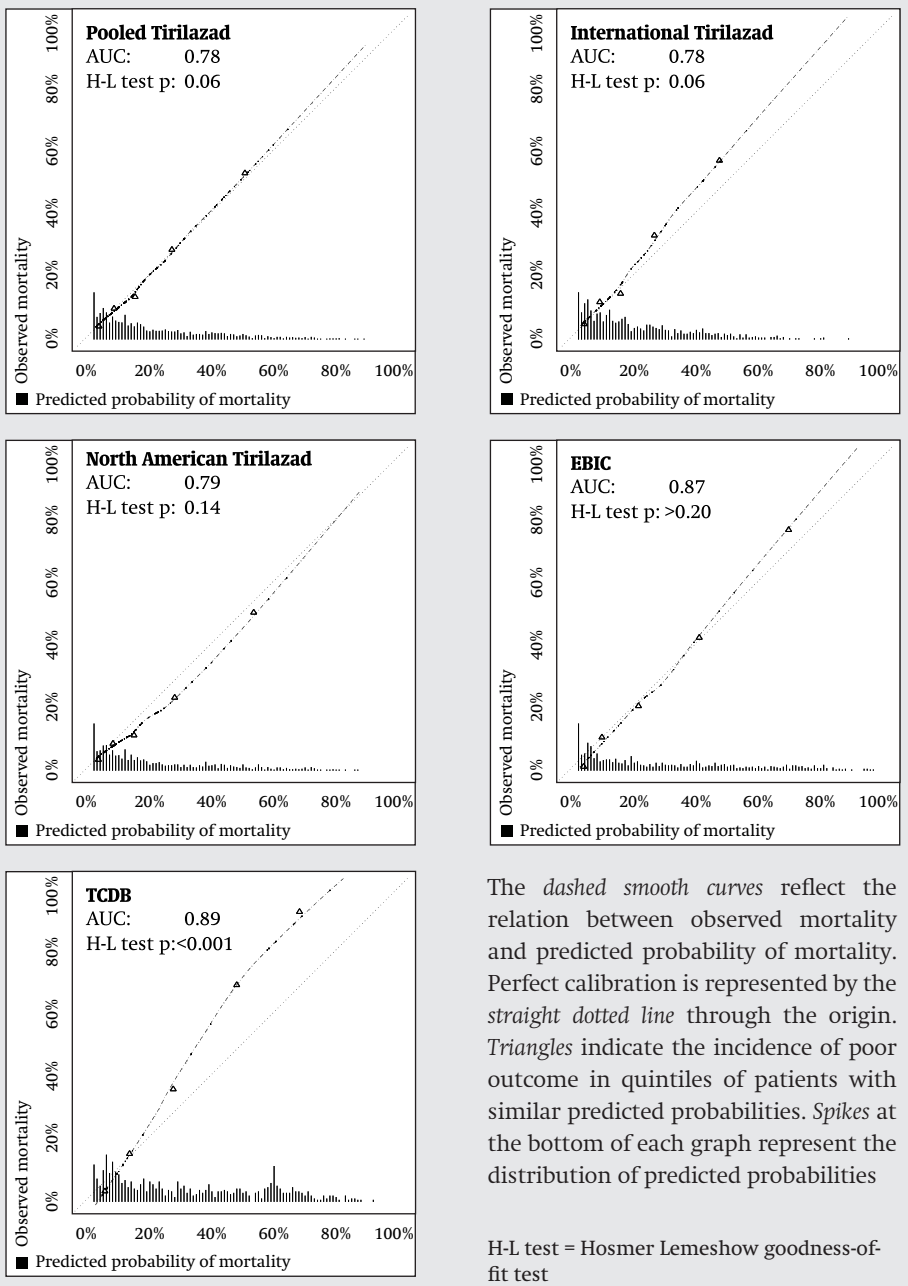
The distribution of predictors differed between the four patient populations. The EBIC patients were generally older than the Tirilazad and TCDB patients (41, 34 and 33 years respectively). Regarding the motor score, the EBIC and TCDB populations were the most heterogeneous, containing both many severely (motor score 'none' and 'extensor response') and many less severely injured patients (motor score 'localizing' and 'obeying commands'). The proportion of patients with hypoxia, hypotension and bilaterally absent pupillary reactivity was relatively low in the International Tirilazad population (15%, 14% and 12% respectively) and high in the TCDB (48%, 42% and 31% respectively). Overall, the Tirilazad populations had the most homogeneous case-mix, whereas the EBIC and TCDB populations were more heterogeneous.

Table 3 Discrimination and calibration of the prognostic models. The models were developed on the pooled Tirilazad patients. Internal validation occurred on the pooled Tirilazad population and on the International and North American Tirilazad patients separately. External validation was performed in the unselected European Brain Injury Consortium (EBIC) survey and the North American survey of the Traumatic Coma Data Bank (TCDB)

	Mortality		Unfavorable outcome	
	Discrimination AUC † (95% CI) ‡	Calibration P-value §	Discrimination AUC† (95% CI) ‡	Calibration P-value §
Internal validation				
Pooled Tirilazad	0.78 (0.76 – 0.81)	0.06	0.80 (0.78 – 0.82)	0.02
International Tirilazad	0.78 (0.75 – 0.82)	0.06	0.80 (0.78 – 0.83)	0.01
North American Tirilazad	0.79 (0.75 – 0.82)	0.14	0.80 (0.78 – 0.83)	0.11
External validation				
European survey (EBIC)	0.87 (0.84 – 0.89)	0.42	0.83 (0.80 – 0.86)	0.05
North American survey (TCDB)	0.89 (0.87 – 0.91)	<0.001	-	-

† AUC = Area under the receiver operating curve
‡ 95% CI = 95% Confidence Interval of the AUC
§ Hosmer-Lemeshow goodness-of-fit test, low p-values indicate poor goodness-of-fit

Figure 1 Validation of the prognostic models in the pooled Tirilazad population (n=2269), the International Tirilazad population (n=1120), the North American Tirilazad population (n=1149), the European Brain Injury Consortium Survey (n=796) and the Traumatic Coma Data Bank (n=746)



The *dashed smooth curves* reflect the relation between observed mortality and predicted probability of mortality. Perfect calibration is represented by the *straight dotted line* through the origin. *Triangles* indicate the incidence of poor outcome in quintiles of patients with similar predicted probabilities. *Spikes* at the bottom of each graph represent the distribution of predicted probabilities

H-L test = Hosmer Lemeshow goodness-of-fit test

Univariable analysis showed a statistically significant relation to both outcome measures for age, motor score, pupillary reactivity, hypoxia, hypotension, and CT parameters (Table 2). Multivariable effects of these predictors were slightly smaller, but still substantial (ORs ≥ 1.4 and 95% confidence intervals > 1, Table 2). Gender and cause of injury had no clear prognostic effects in the multivariable analysis and were therefore not included in further model development. Age, motor score and the CT parameters were the most important predictors. The CT-classification was more important to predict mortality, while age, motor score and pupillary reactivity had a slightly stronger association with unfavorable outcome. Details of the multivariable prognostic models for mortality and unfavorable outcome are described in the **Appendix**.

Performance of the models

The discriminative ability of the two models was good in the Tirilazad patients (AUC 0.78 – 0.80, Table 3). Performance was even better in the EBIC and TCDB patients (external validation, AUC 0.83 – 0.89). Internally, calibration of the models for mortality was satisfactory, with higher mortality for those with higher predictions (Figure 1). In the International and North American patients, the observed mortalities were 25% and 20%, while the average predicted risks were 21% and 23% respectively. These deviations from the ideal calibration were close to statistical significance (Hosmer-Lemeshow test: p=0.06 and p=0.14).

Externally, calibration was good for the EBIC patients. For the TCDB patients, the calibration curve was above the line representing identical predictions and observations. On average, the observed mortality was 44%, while 34% mortality was expected. The deviation was especially large for high predictions, e.g. patients with a predicted mortality risk of 50% actually had an observed mortality of around 70%.

The calibration curves for unfavorable outcome were similar to those shown in Figure 1. Despite this positive graphical impression of calibration, deviations reached statistical significance for the total and International Tirilazad patients (Hosmer-Lemeshow test p<0.05, Table 3). With adjustment of the model intercepts, the calibration for both mortality and unfavorable outcome improved. A significant deviation remained however for mortality predictions in the TCDB patients. This implies that the poor calibration was not only due to a problem in prediction of the average outcome for these patients.

Clinical application

We developed a score chart to estimate the probability of mortality and unfavorable outcome (Table 4). To predict outcome for an individual patient, the scores in the chart need to be added. The corresponding probability of a poor outcome can be read from figure 2. For example, consider a 30-year patient with motor score 'withdrawal', presence of pupillary reactivity, no hypoxia or hypotension, a CT classification of II, and no tSAH. This patient has a sum score of 1 point for mortality and 1 point for unfavorable outcome, which corresponds with a probability around 5% for mortality and 12% for unfavorable outcome (Figure 2). If additionally hypotension (add 2 points for mortality, 1 for unfavorable outcome) would have been present, and the CT showed both tSAH (add 2 points for mortality, 1 for unfavorable outcome) and signs of raised ICP (CT classification III) (add 2 points for mortality, 1 for unfavorable outcome), the probability of poor outcome increases to 32% for mortality (sum score: 7) and to 45% for unfavorable outcome (sum score: 4) (Table 4, Figure 2).

We may apply the score chart for classifying patients, e.g. into five risk categories. Table 5 demonstrates the considerable heterogeneity of patients with TBI, even within the setting of an RCT with strict enrollment criteria. Of the Tirilazad patients, 34% had a predicted risk of unfavorable outcome of less than 20% (14% observed), while 6% of patients had a predicted risk larger than 80% (89% observed).

Table 4 Prognostic score chart for the probability of mortality and unfavorable outcome in patients with severe or moderate TBI according to the prognostic models

Predictor	Value	Mortality	Unfavorable outcome
Age	15 – 39	0	0
	40 – 54	1	1
	55 – 64	2	2
	≥ 65	3	3
Motor score	None/extensor	3	3
	Abnormal flexion	2	2
	Withdraws	1	1
	Localizes/obeys	0	0
Pupillary reactivity	Both react	0	0
	One reacts	1	1
	None reacts	2	2
Hypoxia	No	0	0
	Yes	1	1
Hypotension	No	0	0
	Yes	2	1
CT-classification#	I or II	0	0
	III	2	1
	IV	4	1
	V or VI	2	1
Traumatic subarachnoid haemorrhage	No	0	0
	Yes	2	1
Sum score*: add relevant scores		.....	.....

* The sum score can be used to obtain the predicted probability of mortality or unfavorable outcome from Figure 2

\# CT-classification: I = no visible intracranial pathology on CT scan, II = midline shift 0-5 mm, III = cisterns compressed or absent with midline shift 0-5 mm, IV = midline shift > 5 mm, V = any lesion surgically evacuated, VI = high- or mixed-density lesion > 25 mm, not surgically evacuated

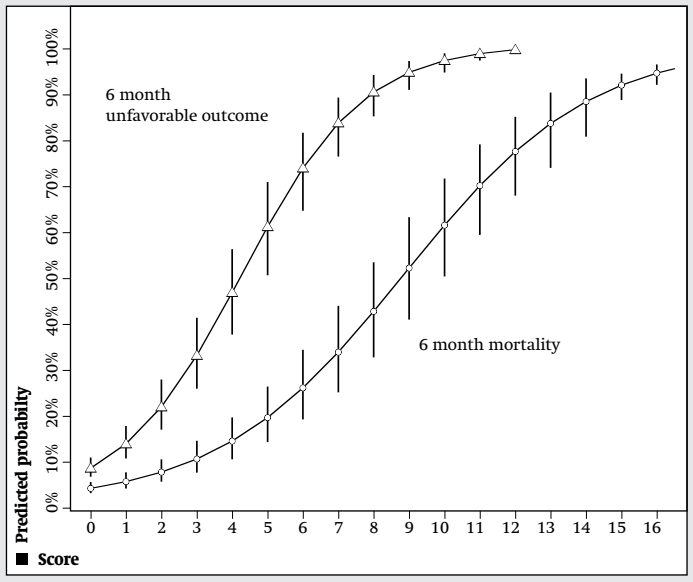
Table 5 Classification of Tirilazad patients (n=2149) according to their severity and prognosis into five risk categories. Categories were defined according to risk of unfavorable outcome, as estimated by the prognostic models

Classification	Risk of unfavorable outcome	Score*	n (%)	Mean mortality (%)	Mean unfavorable outcome (%)
Good	0 - 20%	0-2	741 (34%)	7%	14%
Relatively good	21 - 40%	3	581 (27%)	15%	34%
Intermediate	41 - 60%	4-5	374 (17%)	30%	55%
Poor	61 - 80%	6-7	321 (15%)	47%	76%
Very poor	81 - 100%	≥ 8	132 (6%)	68%	89%

* Score according to the score chart in Table 4

Figure 2 Predicted probability of mortality and unfavorable outcome corresponding to the sum scores from Table 5

After calculating a sum score for a patient with moderate or severe TBI, this graph can be used to determine the corresponding predicted probability with the 95% confidence interval. The exact probability of mortality or unfavorable outcome can be calculated with the formulas shown in the appendix



Discussion

We developed and validated prognostic models to predict the risk of six-month mortality and unfavorable outcome in individual patients after moderate or severe TBI. Predictions were based on characteristics that were readily available on admission to the neurosurgical unit. The models discriminated well between patients with poor and good outcome in the development population. External discrimination on relatively unselected North American (TCDB) and European (EBIC) patients was even better. This can be explained by the greater heterogeneity in case-mix in these surveys, enabling a clearer separation between patients with good from those with poor outcome.

The mortality in the North American TCDB patients was considerably higher than predicted (44% vs. 34%). This poor calibration may be explained by the historic aspect of the data set, which originates from the 1980s, and the improvement of treatment standards, including trauma organization, diagnostic facilities such as CT scanning, and critical care management. Also, inclusion criteria were towards more severe conditions, e.g. patients with neurological worsening. In contrast, the observed frequency of poor outcome was slightly lower than predicted for the North American Tirilazad patients (20% vs. 23% for mortality, 38% vs. 41% for unfavorable outcome respectively). This systematic positive effect has been discussed before[23]. It may have resulted from chance, or be explained by an as yet unidentified regional factor, e.g. a clinical characteristic or a difference in trauma care. This observation also points to the desirability for updating a prognostic model according to specific population characteristics, such as calendar year, treatment setting or inadvertent selection influenced by local trauma organization[40].

In the past, several models have been derived to estimate the probability of hospital mortality of adult intensive care unit patients with physiological characteristics collected during the first day(s), including APACHE, SAPS and MPM[41]. Our models differ in several aspects, since we predicted six-month mortality or unfavorable outcome, only for TBI patients, and only with baseline characteristics. For comparison, we considered three TBI-specific models that were potentially useful for current clinical practice (Table 6). These were described in such detail that we could evaluate their discriminative ability[5,6,16]. Choi et al.[5] predicted twelve-month mortality and unfavorable outcome using age, motor score, pupillary reactivity and the presence of intracerebral lesions at admission. Signorini et al.[6] predicted twelve-month survival with age, GCS, pupillary reactivity, injury severity score (ISS) and the presence of haematoma on the first CT scan. Andrews et al.[16] predicted twelve-month unfavorable outcome with age, cause of injury, GCS and pupillary reactivity in a tree model. We found that the discriminative ability of these previous models was lower in the EBIC and TCDB populations than that of our models (Table 7). The AUC of the model of Choi et al.[5] was 0.82 for mortality and 0.78 for unfavorable outcome in the EBIC data and 0.78 for mortality in the TCDB. Signorini's model[6] performed better, i.e. AUC equal to 0.85 and 0.81 in the EBIC and TCDB respectively. The AUC of the model by Andrews et al.[16] was 0.75 in the EBIC population. On validation of Signorini's model[6] we excluded the effect of the ISS, as ISS was not consistently registered in the surveys. It is conceivable that the AUC would have been higher if this variable had been available, although the additional predictive value of extracranial injuries may be small[42,43].

The poorer discriminative ability of previously described models may have several reasons. The previously developed models were developed on patient samples originating from a single

Table 6 Characteristics of three previously developed TBI-specific models considered for comparison to the Tirilazad model

Model	Outcome	Predictors	Model	Comment
Choi et al., 1991	12-month mortality and unfavorable outcome	Age, motor score, pupillary reactivity, intracerebral lesions	Tree model	Reasonable number of patients (n=555), only major predictors (4 out of 23 considered), only internal validation
Signorini et al., 1999	12-month survival	Age, GCS, pupillary reactivity, ISS, hematoma	Logistic regression	Reasonable number of patients (n=372), reasonable number of predictors (n=5), limited external validation
Andrews et al., 2002	12-month mortality and unfavorable outcome	Age, cause of injury, GCS, pupillary reactivity	Tree model	Small number of patients (n=121), only major predictors (n=4), only internal validation
This study	6-month mortality and unfavorable outcome	Age, motor score, pupillary reactivity, hypoxia, hypotension, CT classification, tSAH	Logistic regression	Large number of patients (n=2269), larger number of predictors (n=7), internal and external validation

GCS = Glasgow Coma Scale, ISS = Injury Severity Score, CT = Computerized Tomograph, tSAH = traumatic subarachnoid haemorrhage

Table 7 Comparison of the discriminative ability of three previously described models and the Tirilazad model

Model	Discriminative ability (AUC*)#		
	Pooled Tirilazad population (n=2269)	European survey (n=796)	North American survey (n=746)
Choi et al.[5]	0.68 / 0.72	0.82 / 0.78	0.78 / -
Signorini et al.[6]	0.71/ -	0.85 / -	0.81 / -
Andrews et al.[16]	- / 0.66	- / 0.75	- / -
This study	0.78 / 0.80	0.87 / 0.83	0.89 / -

* AUC = Area under the receiver operating curve
\# First AUC refers to the model predicting mortality, second AUC to the model predicting unfavorable outcome.

institute or the same region[5,6,16], which may limit their generalizability. Further, patient samples were all much smaller than the 2200 patients used for our models. Thus, the precision to quantify a prognostic model was smaller and 'overfitting' may have occurred. This is the phenomenon that a model predicts outcomes well in the development population but tends to predict too extreme probabilities in new patients. On the other hand, 'underfitting' – i.e. the erroneous omission of predictors with weaker effects – may also have occurred in small samples by lack of power. Finally, the previously developed models included at most five predictors, compared to the seven predictors in our models. It is hence quite plausible that our proposed models truly have a higher predictive accuracy[6,21,33,35,36].

The presented score chart may support clinicians in their initial assessment of the severity and prognosis of a TBI patient. This is for example important for informing relatives. Although it is unlikely that any clinician has the equivalent and systematic experience of treating the 2200 patients underlying our models, statistical models can never replace the clinician. A prediction for an individual TBI patient always has uncertainty, as shown in Figure 2. 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. However in specific circumstances, such as multiple casualty situations or in settings where treatment facilities are very limited it may provide guidance to resource allocation.

We propose that the prime application of the prognostic models is towards a more accurate classification of TBI than is currently possible with the GCS or CT-classification alone. Considerable heterogeneity may still exist between patients classified by the GCS as severe (GCS 3-8) or moderate (GCS 9-12) TBI. This heterogeneity makes it difficult to compare different series and treatment results. Classification of patients according to baseline prognostic risk will facilitate a more accurate comparison of series over time and place. The risk estimate may also be used as a quality control instrument to evaluate the efficacy of health care in TBI. It may be considered a challenge for clinicians to obtain better results than predicted.

Classification of TBI patients according to prognostic risk estimate can play an important role for more efficient design and conduct of RCTs[2]. Until now RCTs have included TBI patients who may do poorly no matter how good the treatment or patients who may do well no matter what treatment is given. Inclusion of patients with such an extreme prognosis in a RCT will decrease statistical power[44]. Restricting the inclusion of patients within the Tirilazad trials to those with an intermediate risk, e.g. between 20 and 80% of unfavorable outcome, would for instance have excluded 40% of the patients (Table 5). Accordingly, costs and efforts would have been decreased, while maintaining nearly the same power to detect a treatment effect. Similar reductions (30%) in sample size have been suggested before[2].

Several limitations of our analyses should be acknowledged. First, the Tirilazad studies, on which we developed our models, concern selected trial populations. Thus, specific subgroups were relatively infrequent, such as patients with absent motor score, those obeying commands, and those with a normal CT scan or age above 65 years. Discrimination was higher in populations that included more patients from these subgroups, such as the relatively unselected EBIC and TCDB surveys.

Second, limitations in assessing the full GCS, which forms an integral part of the Revised Trauma Score, the Trauma and Injury Severity Score and the APACHE scores in the intensive care settings, are well recognized[17,19]. It has been proposed to impute pre-sedation GCS values for calculation of the APACHE scores[45]. Nevertheless, in current clinical practice the motor score is not always available due to effects of early sedation/ paralysis and artificial ventilation[46]. Substituting the admission motor score by earlier observations, such as the motor score at the site of the accident, may alter the prognostic value of the models, and is therefore not advised. Further studies are needed to elucidate this important issue.

Although the performance of the presented models was satisfactory, it might potentially be improved by including CT parameters other than the CT-classification or tSAH. Performance may also be improved by inclusion of subsequent information obtained after admission, including temporal course, MRI scans at later time points, and other parameters such as raised ICP[47]. In a secondary analysis we found that substituting the admission CT-classification by the worst CT in the models significantly improved the performance. The objective of the present study, however, was to investigate prognostic models that predict long-term outcome with baseline predictors only.

In conclusion, prediction models were developed which provide high discrimination between patients with good and poor six-month outcome. Using large patient populations originating from many countries increased the generalizability of the models. These models may be useful for providing realistic information to relatives on expectations of outcome, for quantifying and classifying the severity of brain injury, for stratification of patients in clinical trials, or as a reference for evaluating quality of care.

Acknowledgements

The authors express their gratitude to all of the study participants and the Principal Investigators of the Tirilazad trials and the EBIC and TCDB surveys whose work made this report possible. The authors further acknowledge collaboration with the American and European Brain Injury Consortia and wish to thank Marja van Gemerden for administrative assistance. Grant support was provided by NIH NS 42691.

Appendix

Details of the prognostic models

The probability of a poor outcome (mortality or unfavorable outcome according to the GOS) 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 mortality = −3.267 − (0.0198 × age) + (0.000528 × age^2) + (1.126 × motor score 1 or 2) + (0.918 × motor score 3) + (0.494 × motor score 4) + (0.364 × one pupil reacts) + (0.648 × no pupil reacts) + (0.745 × hypotension) + (0.377 × hypoxia) + (0.808 × CT classification III) + (1.354 × CT classification IV) + (0.860 × CT classification V or VI) + (0.694 × traumatic subarachnoid haemorrhage).

LP for unfavorable outcome = −2.842 + (0.00106 × age) + (0.000391 × age^2) + (1.690 × motor score 1 or 2) + (1.275 × motor score 3) + (0.675 × motor score 4) + (0.440 × one pupil reacts) + (0.938 × no pupil reacts) + (0.740 × hypotension) + (0.449 × hypoxia) + (0.637 × CT classification III) + (0.628 × CT classification IV) + (0.619 × CT classification V or VI) + (0.657 × traumatic subarachnoid haemorrhage).

Coding of the predictors:
Age and age^2: age in years;
All other predictors: 1 if true and 0 if false.

References

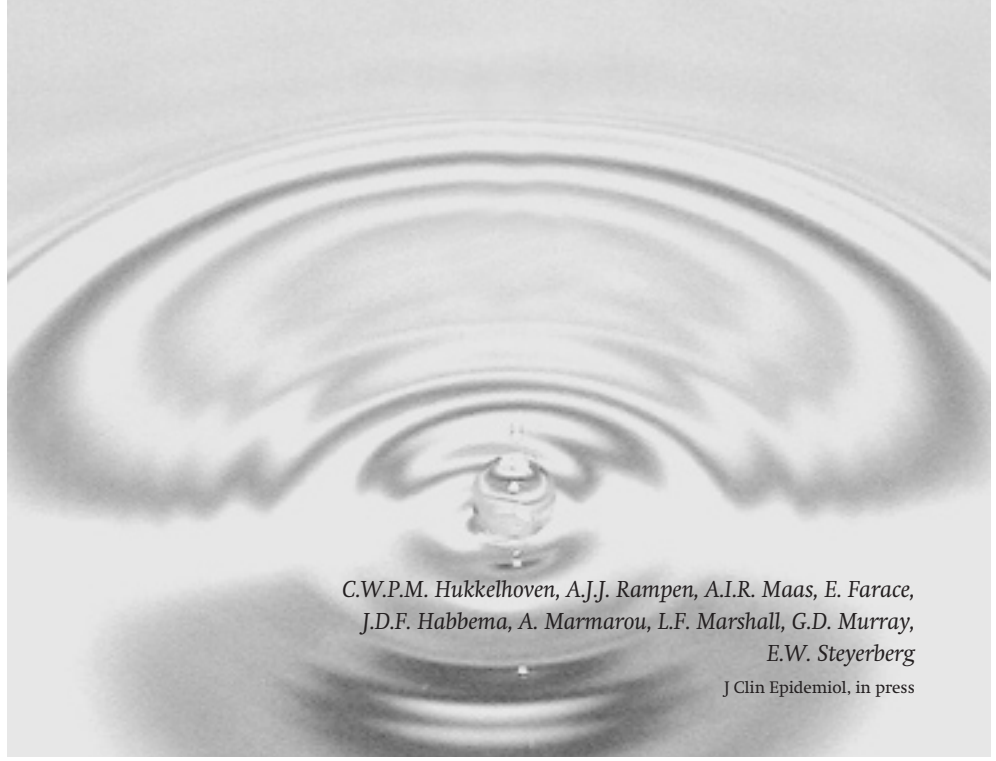
1. Sosin DM, Sniezek JE, Thurman DJ. Incidence of mild and moderate brain injury in the United States, 1991. Brain Inj 1996;10(1):47-54.
2. Machado SG, Murray GD, Teasdale GM. Evaluation of designs for clinical trials of neuroprotective agents in head injury. European Brain Injury Consortium. J Neurotrauma 1999;16(12):1131-8.
3. Jennett B, Teasdale GM, Braakman R, Minderhoud J, Knill-Jones R. Predicting outcome in individual patients after severe head injury. Lancet 1976;1(7968):1031-4.
4. Choi SC, Ward JD, Becker DP. Chart for outcome prediction in severe head injury. J Neurosurg 1983;59(2):294-7.
5. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head-injured patients. J Neurosurg 1991;75(2):251-5.
6. 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;66(1):20-5.
7. Braakman R, Gelpke GJ, Habbema JDF, Maas AIR, Minderhoud JM. Systematic selection of prognostic features in patients with severe head injury. Neurosurgery 1980;6(4):362-70.
8. 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.
9. Stablein DM, Miller JD, Choi SC, Becker DP. Statistical methods for determining prognosis in severe head injury. Neurosurgery 1980;6(3):243-8.
10. Lannoo E, Van Rietvelde F, Colardyn F, et al. Early predictors of mortality and morbidity after severe closed head injury. J Neurotrauma 2000;17(5):403-14.
11. Young B, Rapp RP, Norton JA, Haack D, Tibbs PA, Bean JR. Early prediction of outcome in head-injured patients. J Neurosurg 1981;54(3):300-3.
12. Narayan RK, Greenberg RP, Miller JD, 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;54(6):751-62.
13. 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(3):267-79.
14. Schaan M, Jaksche H, Boszczyk B. Predictors of outcome in head injury: proposal of a new scaling system. J Trauma 2002;52(4):667-74.
15. Barlow P, Teasdale GM, Jennett B, Murray LS, Duff C, Murray GD. Computer-assisted prediction of outcome of severely head-injured patients. Journal of Microcomputer Applications 1984;7:271-277.
16. Andrews PJ, Sleeman DH, Statham PF, 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;97(2):326-36.
17. Buechler CM, Blostein PA, Koestner A, Hurt K, Schaars M, McKernan J. Variation among trauma centers' calculation of Glasgow Coma Scale score: results of a national survey. J Trauma 1998;45(3):429-32.
18. Healey C, Osler TM, Rogers FB, et al. Improving the Glasgow Coma Scale score: motor score alone is a better predictor. J Trauma 2003;54(4):671-8; discussion 678-80.
19. Meredith W, Rutledge R, Fakhry SM, Emery S, Kromhout-Schiro S. The conundrum of the Glasgow Coma Scale in intubated patients: a linear regression prediction of the Glasgow verbal score from the Glasgow eye and motor scores. J Trauma 1998;44(5):839-44; discussion 844-5.

20. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(6):515-24.
21. Hukkelhoven CWPM, Eijkemans MJ, Steyerberg EW. Predicting survival using simple clinical variables: a case study in traumatic brain injury. J Neurol Neurosurg Psychiatry 2000;68(3):396-7.
22. Jennett B, Bond M. Assessment of outcome after severe brain damage. A practical scale. Lancet 1975;1:480-484.
23. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. J Neurosurg 2002;97(3):549-57.
24. Marshall LF, Maas AIR, Marshall SB, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 1998;89(4):519-25.
25. Murray GD, Teasdale GM, Braakman R, et al. The European Brain Injury Consortium Survey of head injuries. Acta Neurochir (Wien) 1999;141:223-236.
26. Marshall LF, Becker DP, Bowers SA, et al. The National Traumatic Coma Data Bank. Part 1: Design, purpose, goals, and results. J Neurosurg 1983;59(2):276-84.
27. Teasdale GM, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974;2(7872):81-4.
28. Hall ED, Yonkers PA, Andrus PK, Cox JW, Anderson DK. Biochemistry and pharmacology of lipid antioxidants in acute brain and spinal cord injury. J Neurotrauma 1992;9 Suppl 2:S425-42.
29. Foulkes MA, Eisenberg HM, Jane JA, Marmarou A, Marshall LF, and the Traumatic Coma Data Bank Research Group. The Traumatic Coma Data Bank: design, methods, and baseline characteristics. J Neurosurg 1991;75:S8-S13.
30. Bullock R, Chesnut R, Clifton G, et al. Management and prognosis of severe traumatic brain injury. Part 1: Guidelines for the management of severe traumatic brain injury. J Neurotrauma 2000;17(6/7):451-553.
31. Chesnut R, Ghajar J, Maas AIR, et al. Management and prognosis of severe traumatic brain injury. Part 2: early indicators of prognosis in severe traumatic brain injury. J Neurotrauma 2000;17(6/7):557-627.
32. Marshall LF, Bowers Marshall S, Klauber MR, et al. A new classification of head injury based on computerized tomography. J Neurosurg 1991;75:S14-S20.
33. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
34. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227 - 1237.
35. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
36. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis: Springer-Verlag New York, Inc., 2001.
37. Hukkelhoven CWPM, Steyerberg EW, Rampen AJJ, et al. Patient age and outcome following severe traumatic brain injury: an analysis of 5600 patients. J Neurosurg 2003;99(4):666-73.
38. Efron B, Tibshirani R. An Introduction to the Bootstrap: Chapman and Hall, New York, 1993.
39. Moons KG, Harrell FE Jr., Steyerberg EW. Should scoring rules be based on odds ratios or regression coefficients? J Clin Epidemiol 2002;55(10):1054-5.
40. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med 2000;19(24):3401-15.

41. Lemeshow S, Le Gall JR. Modeling the severity of illness of ICU patients. A systems update. JAMA 1994;272(13):1049-55.
42. Schreiber MA, Aoki N, Scott BG, Beck JR. Determinants of mortality in patients with severe blunt head injury. Arch Surg 2002;137(3):285-90.
43. Baltas I, Gerogiannis N, Sakellariou P, Matamis D, Prassas A, Fylaktakis M. Outcome in severely head injured patients with and without multiple trauma. J Neurosurg Sci 1998;42(2):85-8.
44. Califf RM, Woodlief LH, Harrell FE, Jr., et al. Selection of thrombolytic therapy for individual patients: development of a clinical model. GUSTO-I Investigators. Am Heart J 1997;133(6):630-9.
45. Livingston BM, Mackenzie SJ, MacKirdy FN, Howie JC. Should the pre-sedation Glasgow Coma Scale value be used when calculating Acute Physiology and Chronic Health Evaluation scores for sedated patients? Scottish Intensive Care Society Audit Group. Crit Care Med 2000;28(2):389-94.
46. Moskopp D, Stahle C, Wassmann H. Problems of the Glasgow Coma Scale with early intubated patients. Neurosurg Rev 1995;18(4):253-7.
47. 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;92(1):1-6.

7

Only few prognostic models were valid to classify patients with traumatic brain injury



C.W.P.M. Hukkelhoven, A.J.J. Rampen, A.I.R. Maas, E. Farace, J.D.F. Habbema, A. Marmarou, L.F. Marshall, G.D. Murray, E.W. Steyerberg

J Clin Epidemiol, in press

Abstract

Objective
Various prognostic models have been developed to predict outcome after traumatic brain injury (TBI). We aimed to determine the validity of six models that used baseline clinical and CT characteristics to predict mortality or unfavorable outcome at six months or later after severe or moderate TBI.

Study design and setting
The validity was studied in two selected series of TBI patients enrolled in clinical trials (Tirilazad trials: n=2269; International Selfotel trial: n=409) and in two unselected series of patients consecutively admitted to participating centers (EBIC survey: n=796; Traumatic Coma Data Bank: n=746). Validity was indicated by discriminative ability (AUC) and calibration (Hosmer-Lemeshow goodness-of-fit test).

Results
The models varied in number of predictors (four to seven) and in development technique (two prediction trees and four logistic regression models). Discriminative ability varied widely (AUC: 0.61-0.89), but calibration was poor for most models. Better discrimination was observed for logistic regression models and for models including more predictors. Further, discrimination was better when tested on unselected series that contained more heterogeneous populations.

Conclusion
Our findings emphasize the need for external validation of prognostic models. The satisfactory discrimination indicates that logistic regression models, developed on large samples, can be used for classifying TBI patients according to prognostic risk.

Introduction

Traumatic brain injury (TBI) carries a high mortality and is a major cause of life long disability in a predominantly young population. Many studies have investigated the value of baseline clinical and CT characteristics for early prediction of long-term outcome. International guidelines contain a section dedicated to prognosis[1,2]. This section summarizes existing knowledge on the predictive value of age, Glasgow Coma Scale (GCS), pupils, hypoxia and hypotension as well as individual CT characteristics (status of basal cisterns, shift, and traumatic subarachnoid hemorrhage). A number of studies in TBI have focused on the development of prognostic models, with which the risk for mortality or morbidity after TBI can be assessed using multiple patient characteristics. Early attempts showed that risk prediction was difficult on admission, but more recent studies have shown better results[3-10].

Risk assessment is important for various purposes: to inform relatives on realistic expectations, to support clinical decision-making and resource allocation, or to classify patients according to prognostic risk. The latter may be useful to compare different patient series, to study treatment results over time or to stratify patients for randomized clinical trials (RCTs)[11]. The different aims of models pose different requirements to performance. For purposes of classification, a high discriminative ability is most relevant, while reliable support for clinical decision-making requires well-calibrated predictions. Whatever the purpose, external validation of models is essential to support general applicability[12]. When simply tested on data that were used for model development, the apparent performance may be excellent, but when tested on new patients, or on a different patient series, the performance may be considerably poorer[13]. Surprisingly, only few of the published models in the field of TBI have been subjected to extensive validation, assessing both internal and external validity[5,6,10]. Internal validity (or reproducibility) indicates how a model performs in patients similar to those used for model development. External validity (or generalizability) evaluates how a model performs for new patients, e.g. more recent patients or patients from different centers[12].

The aim of the present study was to identify models that use baseline characteristics to predict outcome after moderate (Glasgow Coma Scale 9 to 12) or severe (Glasgow Coma Scale 3 to 8) TBI[14], and to determine the validity of predictions from these models on external data sets.

Materials and methods

Models

We conducted a systematic literature search in PubMed for prognostic models in TBI published after 1990; keywords for the search were 'head injury', 'traumatic brain injury', 'prognosis', 'prediction', 'models', 'Glasgow Outcome Scale', 'mortality' and 'unfavorable outcome'. We first checked all abstracts for studies that used baseline characteristics to predict long term (≥ 6 months) mortality or unfavorable outcome[15] after severe or moderate TBI. We then performed a detailed study of selected papers, and checked cross-references for other potentially relevant publications.

Validation populations

For validation, we used two series of patients included in randomized controlled trials (RCTs) in TBI: the Tirilazad studies (N=2269) and the International Selfotel trial (N=427). Further, we used two relatively unselected series of patients with severe or moderate TBI consecutively admitted to participating centers: the European Brain Injury Consortium (EBIC) survey (N=796) and the Traumatic Coma Data Bank (TCDB) (N=746).

The Tirilazad trials were two phase III RCTs, investigating efficacy and safety of the drug Tirilazad Mesylate in TBI. The International trial (N=1120) was conducted from 1992 until 1994 in Europe, Israel and Australia and the North-American trial (N=1149) from 1991 until 1994 in the USA and Canada[16,17]. The trials included patients between 15 and 65 years of age, with severe (GCS 3-8) or moderate (GCS 9-12) TBI. Patients with an absent motor score or penetrating head injury were excluded. For the present study we considered patients from both trials as originating from one population. Data on six-month mortality were present for all 2269 patients, data on six-month unfavorable outcome for 2149 patients[17].
The International Selfotel trial was a phase III RCT, investigating the competitive NMDA-glutamate antagonist Selfotel[18]. The trial was conducted in Europe, Canada, Australia and Argentina between 1994 and 1995. Enrollment criteria were: age 16 - 65 years, severe (GCS 3-8) TBI, presence of abnormalities on the CT scan, at least one reactive pupil and non penetrating TBI. Data on six-months outcome were available for 409 patients.

The EBIC survey included patients admitted to 104 centers in 12 European countries between February and April 1995[19]. In this series, data were prospectively collected in all adult (> 16 years) patients suffering a moderate or severe TBI, consecutively admitted to participating centers. For the present study we selected patients from centers routinely reporting six-months outcome data (N=796).
The TCDB included patients with severe (GCS 3- 8) TBI admitted to four centers in the USA between 1984 and 1987[20]. For our analyses, we selected all adult (age > 15 years) patients admitted with non-penetrating TBI (N=746). In the TCDB the six-month GOS was assessed with a broad time window. We therefore chose to limit our analysis to mortality, since deviations in time windows for outcome assessment are less relevant for mortality.

For several patients, values of predictors were missing (4.9% of the required values, Table 3). These values were statistically estimated, based on the regression models including the other

predictors, and subsequently imputed[21,22]. The imputation method (function 'transcan' from the Hmisc library for S-plus) transformed continuous and categorical predictor variables to have maximum correlation with the best linear combination of the other variables. Subsequently, missing values were imputed with best guess expected values of the transformed variables, and transformed back to the original scale. Such imputation is recommended as more efficient than dropping cases. This procedure assumes that values are 'missing at random', i.e. random after conditioning on the other predictors[21,23]. After imputation, all values of predictors were present for the development of the prognostic models.

Performance of models

We calculated the predicted outcome (mortality or unfavorable outcome) for each population according to the prognostic equations, with statistical imputation of missing values for predictors[24]. The performance of the models was assessed with respect to discrimination and calibration. Discrimination is the model's ability to separate patients with different outcomes. To quantify the discrimination we calculated the area under the receiver-operating characteristic curve (AUC)[25]. A model with an AUC of 0.50 has no discriminative power, while an AUC of 1.0 reflects perfect discrimination. Ninety-five percent confidence intervals around this AUC were calculated by AUC +/- 1.96*SE, where SE was half the standard error of the Somer's D, as estimated by the rcorr.sens function in S-plus (version 2000, Insightful Corporation, Seattle, WA).
Calibration is the ability of a model to produce unbiased estimates of the probability of outcome, e.g. if patients with certain characteristics are predicted to have a 10% chance of mortality, the actually observed mortality should also be 10%. Calibration was assessed graphically by plotting observed outcome against the predicted probability (0-100%). Calibration was further tested with the Hosmer-Lemeshow goodness-of-fit test, which assesses agreement between predicted and observed risks. In our analyses, predicted risk was grouped into deciles for the regression models. For the tree models we used the predicted groups for the comparison of observed outcome versus predicted risks.

In a secondary analysis, we refitted the original models on the Tirilazad population, updating the regression coefficient (weights) of the predictors to best fit this population. This was done in order to better assess the 'pure' predictive value of the different combinations of risk factors, excluding influences of differences in the development population (e.g. selected or unselected) or statistical methodology (e.g. use of bootstrapping in model development). For prediction trees we refitted two variants: one with the tree structure as published before, and another with the risk factors considered in the tree as main effects in logistic regression models. Internal validity of the refitted models was assessed with bootstrapping procedures for which we used a standard algorithm[25-28]. Bootstrapping involved drawing samples of patients with replacement from the development sample. 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. Regression models were estimated in 200 bootstrap samples. These models were each evaluated on the original sample. The average difference in AUC was determined to indicate the optimism in the initially estimated discriminative ability[25]. In addition, we reduced the regression coefficients to provide better predictions for future patients[21,28]. The reduction was based on the mean of the slope of the linear predictor (shrinkage factor)[27,29]. The external validity of these models was evaluated on the other three series, with calculation of discriminative ability and calibration.

Table 1 Overview of previous prognostic models, predicting GOS or mortality after TBI. End of data collection maximally 15 years ago (from 1988 onwards)

Author, year of publication	No. of patients, used for development	Patient selection	Period and place data collection	Outcome	Model specification	Included characteristics	Validation	Performance
Choi, 1991[4]	555	Severe TBI	1976 – 1989, Virgina, US	GOS at 1 year	Tree	Age, pupillary reactivity,motor response, intracerebral lesion. Characteristics were measured at admission	Not done	Accuracy rate: 77.7%
Benzer, 1991[30]	421	'Severely injured' (no exact criterion)	1981 – 1990, Innsbruck, Austria	Mortality at 21 days	Contingency table analysis	Reaction to acoustic stimuli, reaction to pain, body posture, eye opening, pupil size, pupillary reactivity, position and movement of the eye balls, oral automatisms. Characteristics were measured at first examination after trauma	Not done	No performance measure presented prediction: everybody with score 0-1 will die
Signorini, 1999[6]	372	Moderate or severe TBI, or mild TBI with ISS > 15	1989-1991, Edinburgh, UK	Survival at 1 year	Logistic regression analysis	Age, GCS, ISS, pupillary reactivity, haematoma on CT scan. Characteristics were measured at admission	External on 520 more recent patients from the same center	Apparent AUC = 0.90, Accuracy rate = 90%; external AUC = 0.84, Accuracy rate = 85%
Signorini, 1999[31]	110	Moderate or severe TBI, or mild TBI with ISS > 15	1989 – 1991, Edinburgh, UK	Survival at 1 year	Logistic regression analysis	Age, GCS, ISS, pupillary reactivity, haematoma on CT scan, ICP at different time points during first 72 hours	External on 140 patients from the same center, 1991- 1996	Apparent AUC = 0.90 Error rate=11.6% External performance was poorer (no exact numbers were given)
Andrews, 2002[10]	124 (Different numbers for different trees)	Severe or moderate TBI, or mild TBI with ISS > 15	1989 – 1991, Edinburgh, UK	Unfavorable outcome at 1 year	Several trees	Pupillary reactivity, age, GCS, type of accident, referral, isolation head injury, CPP grade, hypotension, hypocarbia, sex, type of haematoma, evacuation haematoma. Time of measurement not mentioned	Cross-validation in 10 subsets of 1/10th of the patients	Accuracy rate: 60-96% (dependent of tree and development or validation set)
Combes, 1996[5]	132	Severe TBI	1989 – 1992, Grenoble, France	Unfavorable outcome at 48 hours	Logistic regression analysis	Motor score, age and hypoxia. Characteristics were measured at admission	External on 66 patients from the same center, 1989-1992	External accuracy rate: 73% external AUC=0.87
Hukkelhoven, 2004[32]	2269	Severe or moderate TBI	1992 – 1994, US, Canada, Europe, Australia	Mortality and unfavorable outcome at 6 months	Logistic regression analysis	Age, gender, cause of injury, motor score, pupillary reactivity, hypotension, hypoxia, CT classification and tSAH. Characteristics were measured at admission	External on the two international surveys (EBIC: n=796 TCDB: n=746)	Internal AUC = 0.80 – 0.81; External AUC = 0.83 – 0.89
Sakellaropoulos, 1999[7]	525	TBI	1994 – 1996 Patras, Greece	GOS (D,V,SD,MD,GR) at 24 hours	Two bayesian networks	Age, GCS, mean arterial pressure, delay, concomitant injuries, modified diffuse injury scale. Characteristics were measured at admission	External on 75 patients from the same center	Accuracy rate: 81% (network 1), 69% (network 2)
Schaan, 2002[9]	554	Isolated TBI	1989-1999, Murnau, Germany	Mortality and unfavorable outcome at various time points (mean: 5.7 Months)	Mann-Whitney U test	Pupillary reactivity, hemiparesis, brain stem signs, contusion , type of haematoma, midline shift, cerebral edema, basal cisterns. Characteristics were measured within the first 24 hrs after admission	Not done	-
Schreiber, 2002[8]	213	Abbreviated Injury Scale ≤ 5 and blunt TBI	1994-2000, Houston, US	Survival at discharge	Logistic regression analysis	Age, GCS, pupillary reactivity, hypotension, ICP, midline shift. Characteristics were measured within the first few hours after admission	Not done	AUC = 0.81 (0.74 – 0.87)

Results

Models

We identified 10 papers published after 1990 that described prognostic models (Table 1). Four of these studies used early end points (24 hours, 48 hours, 21 days, discharge)[5,7,8,30]. Two studies included data obtained during the clinical course after admission in the model[9,31]. Consequently, only four studies provided prediction models that used baseline clinical and/or CT characteristics obtained on admission to predict long term outcome[4,6,10,32]. These included two papers that presented prediction trees[4,10] and two presenting logistic regression models[6,32]. One of these describes two models that were previously developed by us on the Tirilazad studies. The tree developed by Choi et al.[4] and the models developed by us predicted both mortality and unfavorable outcome. Thus, the four studies described six models in total.

The tree developed by Choi et al.[4] selected four characteristics from 23 candidate variables, where the optimum number of splits of the tree (and thus the selected characteristics) were statistically determined in the dataset using recursive partitioning[33]. Signorini et al.[6] developed a logistic regression model and used a forward selection algorithm ($p<0.05$), with eight candidate variables. Andrews et al.[10] developed a tree, using 10-fold cross-validation to determine the optimal tree size using recursive partitioning[33]. The number of candidate variables was not mentioned. For the logistic regression models by Hukkelhoven et al.[32], a backward stepwise selection procedure ($p<0.20$) was used to select 7 predictors from 9 candidate variables. More information about the models is shown in Table 1.

All six prediction models included age, pupillary reactivity and either the GCS or the Motor Score (Table 2). Some models included cause of injury, hypotension, hypoxia, injury severity score (ISS), and various CT scan characteristics (Table 2). The tree models included four predictors[4,10], and the logistic regression models five[6] and seven[32]. The ISS, used in Signorini's model[6], and intracerebral lesion, used in Choi's tree[4], were not available in (some of) the validation populations. Therefore, we considered the mean ISS in the population used to develop Signorini's model[6] in all validation patients. Since Choi's tree used intracerebral lesion to split two tree branches from each other[4,10], we could combine these branches and further validate the remaining seven branches of the tree.

Validation populations

Characteristics of the validation populations are described in Table 3. Although the populations showed comparable distributions for most variables, some differences were noted. The unselected series (EBIC and TCDB) showed more heterogeneity, both in clinical severity (motor score and pupillary abnormalities) as well as in CT characteristics (e.g. mass lesions, obliteration of cisterns). The EBIC patients were generally older (mean 41 years) than the Tirilazad, Selfotel and TCDB patients (32 to 33 years). The occurrence of road traffic accidents as cause of injury was highest in the Selfotel and TCDB patients (80% and 75% respectively, compared to approximately 55% in the other two populations).

Table 2 Predictors included in the models that were validated in the present study

Predictors	Choi [4]	Signorini [6]	Andrews [10]	Hukkelhoven [32]
	Tree*	LR#	Tree*	LR#
Age	X	X	X	X
Cause of injury			X	
GCS		X	X	
Motor Score	X			X
Pupillary reactivity	X	X	X	X
Hypotension				X
Hypoxia				X
CT-classification				X
TSAH				X
Intracerebral lesion	X			
Lesion		X		
Injury Severity Score		X		

* Tree = tree model
\# LR = logistic regression model

Table 3 Patient characteristics and outcome in the Tirilazad and Selfotel trials, the European Brain Injury Consortium (EBIC) series and the American Traumatic Coma Data Bank (TCDB)

Characteristics and outcome	Coding	Tirilazad (n = 2269)			Selfotel (n = 409)			EBIC (n = 796)			TCDB (n = 746)		
		n	%	missing	n	%	missing	n	%	missing	n	%	missing
Age	Mean (SD)	33.2 (13.5)		1	32.3 (13.4)		0	40.9 (20.6)		1	33.1 (16.4)		0
	Range in years	12 – 79			15 – 79			2 – 92			16 - 93		
Sex	Men	1757	77%	0	320	78%	0	594	75%	1	575	77%	0
Traffic accident		1319	58%	0	328	80%	0	421	53%	3	560	75%	0
GCS	3 – 8	1785	79%	0	409	100%	0	283	61%	333	263	86%	440
	9 – 12	483	21%		0	0%		123	27%		29	9%	
	13 – 15	1	0%		0	0%		57	12%		14	5%	
Motor score	No response	17	1%	0	0	0%	0	153	26%	197	189	27%	43
	Extending	286	13%		55	13%		50	8%		106	15%	
	Abnormal flexion	374	16%		91	22%		48	8%		89	13%	
	Normal flexion	659	29%		127	31%		97	16%		142	20%	
	Localizing	850	37%		134	33%		169	28%		156	22%	
	Obeying	83	4%		2	0%		82	14%		21	3%	
Pupil reactivity	Both pupils reacting	1401	70%	281	308	77%	11	482	66%	63	262	63%	328
	One pupil reacting	279	14%		77	19%		65	9%		27	6%	
	No pupil reacting	308	16%		13	3%		186	25%		129	31%	
Hypotension		395	18%	61	58	14%	5	180	23%	6	310	42%	0
Hypoxia		129	21%	263	24	6%	0	213	27%	5	360	48%	0
CT-classification	No abbreviations or diffuse injury	1006	45%	20	154	38%	0	324	41%	14	197	29%	59
	Swelling	426	19%		94	23%		72	9%		145	21%	
	Shift	88	4%		26	6%		19	2%		36	5%	
	Mass lesion	729	32%		135	33%		367	47%		309	45%	
Traumatic subarachnoid hemorrhage on CT		1185	53%	46	317	78%	5	314	41%	37	148	23%	113
Hematoma on CT		1691	75%	9	168	41%	0	367*	47%*	14	NA		
Intracerebral hematoma on CT		419	19%	19	12	3%	0	42#	5%#	NA	NA		
Mortality		500	22%	0	94	23%	0	244	31%	0	326	44%	0
Unfavorable outcome		864	40%	130	177	43%	0	388	49%	0	NA		

GCS = Glasgow Coma Score; NA = Not available

* In the EBIC population we gave hematoma on CT the same value as CT-classification class V + VI

\# Because this variable was only designated to a certain subgroup, the missing values were not imputed, but an average risk was estimated, contrary to other variables

With respect to outcome measures, both trial populations showed lower percentages for mortality and unfavorable outcome (22% and 23% for mortality and 40% and 43% for unfavorable outcome in the Tirilazad and Selfotel populations respectively), compared to 31% (EBIC) and 44% (TCDB) for mortality and 49% (EBIC) for unfavorable outcome.

External validation

The models selected from the literature showed a wide variability in discrimination in most series (Table 4). Signorini et al. reported an apparent AUC of 0.90, while we found an AUC of 0.61 to 0.85 at external validation of this model. Discriminative performance was not only dependent on the model, but also on the validation population, with better results in the more heterogeneous series of EBIC and TCDB patients, and poorer in the selected trial populations (Table 4).

Performance was also associated with the number of risk factors in the model. The seven predictor models developed earlier by us showed a maximum AUC of 0.89, the five predictor model of Signorini et al. a maximum AUC of 0.85 and the four predictor models of Andrews et al. and Choi et al. a maximum AUC of 0.75 and 0.82 respectively.

The calibration of four of the six models was poor, as demonstrated by the low p-values of the Hosmer-Lemershow goodness of fit test ($p < 0.01$). Figure 1 shows the poor external calibration of these models in more detail, demonstrating that calibration curves often differed considerably from the dotted line representing perfect agreement between predictions and observations. For example, the calibration curve for mortality as predicted by the four-predictor tree[4] was considerably below the dotted ideal line for the Tirilazad, Selfotel and EBIC patients. This implies that most predicted mortalities were systematically too high, e.g. Tirilazad patients with a predicted mortality of 60% actually had an observed mortality of around 35% (Figure 1). The calibration curves of the prediction trees showed several twists, resulting from the inherent structure of the tree, which divides patients into only seven categories. The seven-predictor models showed good calibration in the EBIC and Selfotel series, but performed more poorly in the TCDB patients.

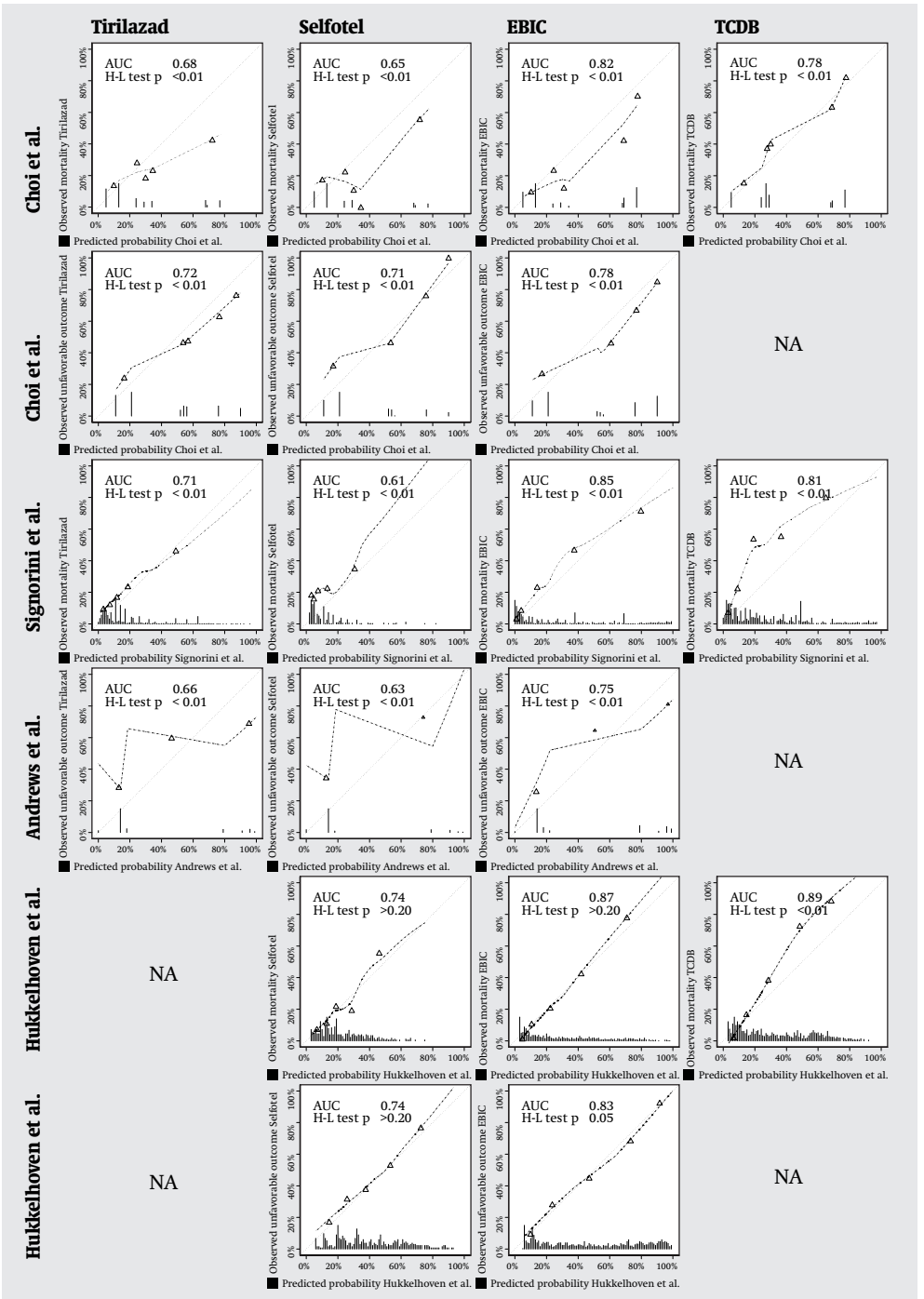
Most refitted models discriminated somewhat better than the original models (Table 5). Several similarities with the original models were however observed; discriminative abilities were mainly dependent on the number of risk factors in the model and on the validation population. Compared to the original models, calibration was considerably better for most refitted models, except for one of the prediction trees[4]. For the prediction trees discrimination was better when refitting the individual risk factors from the tree than when refitting the original tree structure. Remarkably, the refitted models for mortality still showed poor calibration on external validation on the TCDB series.

– continues on page 147 –

Table 4 Performance of previously developed prognostic models. The models were externally validated in the Tirilazad and Selfotel trials, the European Brain Injury Consortium (EBIC) series and the American Traumatic Coma Data Bank (TCDB)

Evaluated outcome	Model	Tirilazad (n=2269)			Selfotel (n=409)			EBIC (n=796)			TCDB (n=746)		
		Discrimination		Calibration	Discrimination		Calibration	Discrimination		Calibration	Discrimination		Calibration
		AUC*	95% CI#	(p-value‡)	AUC*	95% CI#	(p-value‡)	AUC*	95% CI#	(p-value‡)	AUC*	95% CI#	(p-value‡)
Mortality	Choi – tree	0.68	0.65-0.70	< 0.01	0.65	0.58-0.71	< 0.01	0.82	0.79-0.85	< 0.01	0.78	0.74-0.81	< 0.01
	Signorini – LR	0.71	0.69-0.74	< 0.01	0.61	0.54-0.68	< 0.01	0.85	0.82-0.88	< 0.01	0.81	0.78-0.84	< 0.01
	Hukkelhoven – LR	NA		NA	0.74	0.68-0.80	0.49	0.87	0.84-0.89	0.42	0.89	0.86-0.91	< 0.01
Unfavorable outcome	Choi – tree	0.72	0.70-0.74	< 0.01	0.71	0.66-0.76	< 0.01	0.78	0.75-0.81	< 0.01	NA		NA
	Andrews- tree	0.66	0.63-0.68	< 0.01	0.63	0.58-0.68	< 0.01	0.75	0.71-0.78	< 0.01	NA		NA
	Hukkelhoven – LR	NA		NA	0.74	0.69-0.79	0.95	0.83	0.80-0.86	0.05	NA		NA

* AUC = Area under the receiver operating curve
\# 95% CI = 95% Confidence Interval of the AUC
‡ Hosmer-Lemeshow goodness-of-fit test, low p-values indicate poor goodness-of-fit
Tree = a logistic regression model was constructed with the tree structure
LR = a logistic regression model was constructed with the risk factors included as main effects



H-L test p = p-value of the Hosmer-Lemeshow goodness-of-fit test
AUC = Area under the ROC-curve
NA = Not applicable

Figure 1 Validation of the previously developed prognostic models in the pooled Tirilazad population (n=2269), the Selfotel population (n=427), the European Brain Injury Consortium (n=796) and the Traumatic Coma Data Bank (n=746).

The tree by Choi et al.[4] predicts both mortality and unfavorable outcome, as indicated in the first and second rows in the figure respectively. Signorini's model predicts mortality, Andrews' model[10] predicts unfavorable outcome. Hukkelhoven et al.[32] developed two models on the Tirilazad data; one predicting mortality (fifth row) and one predicting unfavorable outcome (sixth row).
The *dashed smooth curves* reflect the relationship between observed frequency and predicted probability of poor outcome. Perfect calibration is represented by the *straight dotted line* through the origin. *Triangles* indicate the incidence of poor outcome in quintiles of patients with similar predicted probabilities. *Spikes* at the bottom of each graph represent the distribution of predicted probabilities.

Table 5 Performance of the previously developed prognostic models after refitting on the Tirilazad patients. External validation of the refitted models was performed on the patients included in the Selfotel trial, the European Brain Injury Consortium (EBIC) survey and the North American survey of the Traumatic Coma Data Bank (TCDB)

Evaluated outcome	Model	Tirilazad (n=2269)			Selfotel (n=409)			EBIC (n=796)			TCDB (n=746)		
		Discrimination		Calibration	Discrimination		Calibration	Discrimination		Calibration	Discrimination		Calibration
		AUC*	95% CI#	(p-value‡)	AUC*	95% CI#	(p-value‡)	AUC*	95% CI#	(p-value‡)	AUC*	95% CI#	(p-value‡)
Mortality	Choi – tree	0.69	0.66-0.71	< 0.01	0.65	0.58-0.71	< 0.01	0.81	0.78-0.84	< 0.01	0.72	0.68-0.76	< 0.01
	Choi – LR	0.72	0.70-0.75	0.43	0.67	0.60-0.74	0.04	0.84	0.82-0.87	0.41	0.82	0.78-0.85	< 0.01
	Signorini - LR	0.72	0.70-0.75	0.75	0.65	0.58-0.72	0.16	0.86	0.83-0.88	0.05	0.83	0.80-0.86	< 0.01
	Hukkelhoven – LR	0.78	0.76-0.80	0.06	0.74	0.68-0.80	0.49	0.87	0.84-0.89	0.44	0.89	0.86-0.91	< 0.01
Unfavorable outcome	Choi – tree	0.72	0.70-0.75	< 0.01	0.71	0.66-0.75	< 0.01	0.78	0.75-0.81	< 0.01	-		-
	Choi – LR	0.77	0.75-0.79	0.78	0.72	0.67-0.77	< 0.01	0.80	0.77-0.83	< 0.01	-		-
	Andrews – tree	0.68	0.66-0.70	0.05	0.66	0.61-0.70	0.08	0.76	0.73-0.79	< 0.01	-		-
	Andrews – LR	0.77	0.75-0.79	0.03	0.72	0.67-0.76	0.61	0.79	0.76-0.82	< 0.01	-		-
	Hukkelhoven – LR	0.80	0.78-0.82	0.02	0.74	0.69-0.79	0.95	0.83	0.80-0.86	0.05	-		-

* AUC = Area under the receiver operating curve (in the Tirilazad data with optimism-correction by bootstrapping),
\# 95% CI = 95% Confidence Interval of the AUC
‡ Hosmer-Lemeshow goodness-of-fit test, low p-values indicate poor goodness-of-fit
Tree = a logistic regression model was constructed with the tree structure
LR = a logistic regression model was constructed with the risk factors included as main effects

Discussion

The external validity varied considerably for previously developed models that aim to predict long term outcome after severe or moderate TBI with baseline characteristics available at admission. We found only a moderate discriminative ability for prediction trees described by Choi et al.[4] and by Andrews et al.[10] (AUC around 0.70), reasonable discrimination for the five-predictor logistic regression model described by Signorini et al.[6] (AUC around 0.75) and better discrimination in the seven-predictor models described earlier by us[32] (AUC around 0.80). Calibration of most models was relatively poor, but improved when models were refitted on the same data set with the same, modern, statistical techniques.

It was difficult to directly compare the performance of models as reported by the authors with the results of our validation studies. Only the model by Signorini et al. reported discriminative ability, with an apparent AUC of 0.90[6] as compared to values of 0.61 – 0.85 in our validation study. In the past, performance of published models has often been expressed in the accuracy rate, which is the proportion of patients with a certain outcome that was predicted correctly. The tree by Choi et al. had an accuracy rate of 78% in the development population[4] and the tree by Andrews et al. had accuracy rates of 96% (development set) and 60% (training set)[10]. We do not consider accuracy rates for validation purposes appropriate. For instance, this rate is greatly influenced by the outcome distribution of the population. In a representative TBI population with an average mortality of 20%, the accuracy rate will already be 80% if all patients would be labeled as survivors.

External validation of prognostic models is essential to assess generalizability, and for fair comparisons of alternative models. Models often perform less well at external validation[12,13] and this has several reasons. First, prognostic modeling on small samples provides limited power to identify predictive variables and to quantify a model with sufficient precision. Therefore, models based on small samples may often have biased and imprecise regression coefficients[34]. Larger samples are required for reliable statistical modeling[25,35]. The present study confirms this: the model developed on the largest series discriminated best in new patients[32], while the model developed on the smallest sample (121 patients) from a single center discriminated poorly[10]. Further, a clear increase in performance was achieved by refitting the models that were developed from single centers with the large, multi-center Tirilazad patient population.

In general, a small development population may induce "overfitting". Overfitting is the phenomenon that a model predicts outcome well in the development population, but tends to predict too extreme probabilities for new patients. Indeed, overfitting was noted before for the logistic regression model developed by Signorini et al.[6,36]. The risk of overfitting can best be limited by increasing the sample sizes for predictive modeling studies. This can be achieved by multi-center collaborations, which has the advantage that generalizability may also be increased[12]. Alternatively, more advanced statistical approaches may be useful, e.g. bootstrapping techniques to determine a shrinkage factor[25,27,28]. Shrinkage of regression coefficients improves the calibration of predictions in new patients. Bootstrapping is a relatively new technique, which was only used in the development of the seven predictor models[32]. We recommend that such techniques should be more generally applied[12,25]. The need for such modeling techniques is well recognized in other fields, such as genetic profiling in cancer research, where the number of patients is often small in relation to the number of candidate predictors[37,38].

Discrimination was clearly related to the number of predictors included in the model, with better discrimination in models using seven predictors than in the four- or five-predictor models. This is caused by the inclusion of additional and statistically independent information in the models, which permits a more detailed differentiation of risks between patients. We note that the seven predictors included in our previously developed models are well-known from previous research[2] and are readily available in clinical practice. The number of predictors that can potentially increase discriminative abilities of a model is however not infinite, and including larger numbers of predictors may increase the risk of overfitting. As a rule of thumb, the number of predictors should be less that 1/10 of the number of events (e.g. number of patients with a particular outcome)[25,35]. We further observed a higher discrimination of all models at validation in the unselected series (EBIC and TCDB) as compared to results obtained on validation on the more selected trial population (Tirilazad and Selfotel populations). This is explained by the greater heterogeneity of the unselected series, which include both more severe and less severely injured patients (e.g. the extremes). This permits better discrimination between patients at low and high risks. Consequently, the interpretation of the results of external validation studies requires insight into the characteristics of the validation population. If a validation population is more heterogeneous than the development population, the expected discriminative ability is higher than that of the development population.

Calibration of most models was relatively poor (Figure 1). The substantial differences between predicted and observed outcomes may have been caused by several factors, including influences of regional trauma organization and referral policy on composition of populations, center differences in therapeutic approaches or changes in treatment over time. From a methodological point of view, the poor calibration points to the desirability for updating a prognostic model according to specific patient characteristics[13,39]. This was further confirmed by the improvement in performance following refitting (Table 4 and 5). No improvement was observed in calibration of the refitted models for the TCDB population, which suggests that a different factor overrides the benefits of refitting. As the TCDB data set is the oldest data set included in our analysis plausible explanations include improvements in trauma organization and standards of care. From a clinical perspective the poor calibration suggests a significant limitation for application of most models to support clinical decision-making and resource allocation.

We found no advantage of the tree models over logistic regression models. In fact, the refitted tree structures[4,10] performed more poorly than refitted logistic regression models containing the same predictors (Table 5). As found in earlier methodological studies, the selection and number of predictor variables was however more important than the particular methodology applied[40]. Further improvement in performance may hence come from including more (and more powerful) predictor variables in future prediction models.

Several limitations of the present study should be acknowledged. First, we only selected prognostic models based on admission clinical and CT characteristics, as it was our primary intent to focus on the baseline situation, without influences of subsequent clinical treatment. Conclusions can therefore only be drawn concerning models using baseline characteristics; it is likely that prognostic models including information from later time periods will perform better. Secondly, some models were developed for predicting twelve-months outcome, while in the validation populations outcome was determined at six months post injury. This may have biased results in favor of the seven predictor model, which was developed on the six-month GOS[32]. However the GOS is considered fairly stable at six months post injury[41] and we therefore do not consider this a significant confounding factor. The injury severity score (ISS), included in the model by Signorini et al.[6] and intracerebral lesion, included in the model by Choi et al.[4] were not used in the validations. It is conceivable that the performance of these models could have been better if it had been possible to include these risk factors, although the improvement would likely have been small[8,42]. Another limitation of our study was that calibration was assessed with standard statistical procedures, which group patients with similar predictions. Therefore, no information was available about the performance of specific clinical groups, e.g. how well the model predicted the outcome for individuals who are male, 75 years old, with one reactive pupil and involved in traffic accidents. Further, two of the four validation populations were patients from RCTs. Although such populations are somewhat selected, they had the advantage to be relatively large, to originate from various Western countries, and to contain prospectively collected and carefully verified data. Three of the four validation populations used were collected between 1992 and 1995. Although these populations cannot directly be considered old or outdated, some factors affecting external validation may have changed since this data collection. We therefore advise to validate prognostic models repeatedly on new patient series.

We conclude that models developed with baseline characteristics available on admission may provide satisfactory discrimination, but often suffer from poor calibration. This implies that these models can best be applied for discriminative purposes, such as ranking or classifying patients. Generally the clinical severity of patients with TBI is classified into severe/moderate/mild according to the GCS. This however provides only a very rough classification, and we contend that the classification can be much more detailed if based on a prognostic model, that includes more characteristics. Caution should be applied when using prognostic models for supporting clinical decision-making and resource allocation, for which good calibration is essential. The better calibration, observed in the refitted models, confirms the desirability for updating a model on more recent patient populations.

Acknowledgements

The authors gratefully acknowledge the significant amount of work performed by all investigators originally involved in the data collection of the Tirilazad studies, the International Selfotel study, the EBIC Core Data Survey and the Traumatic Coma Data Bank; without this extensive work, the current validation studies could not have been performed. The authors express their gratitude to Marja van Gemerden for secretarial and administrative assistance in preparation of the manuscript. Grant support was provided by NIH NS042691-01A1.

References

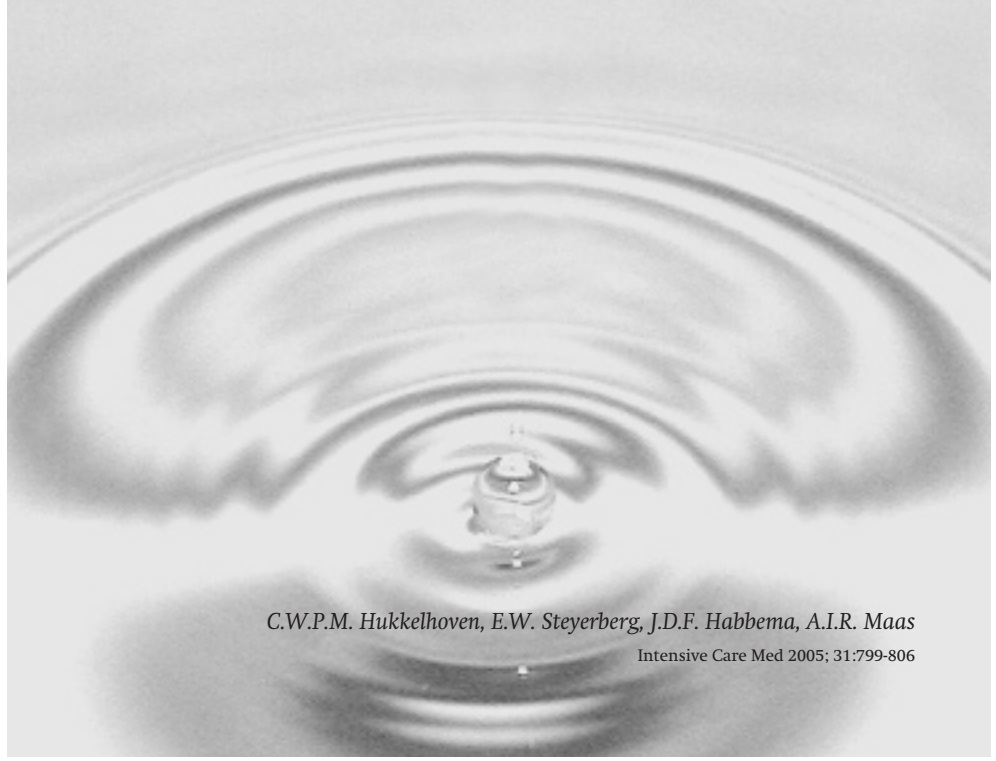
1. Bullock R, Chesnut R, Clifton G, et al. Management and prognosis of severe traumatic brain injury. Part 1: Guidelines for the management of severe traumatic brain injury. J Neurotrauma 2000;17(6/7):451-553.
2. Chesnut R, Ghajar J, Maas AIR, et al. Management and prognosis of severe traumatic brain injury. Part 2: early indicators of prognosis in severe traumatic brain injury. J Neurotrauma 2000;17(6/7):557-627.
3. Braakman R, Gelpke GJ, Habbema JDF, Maas AIR, Minderhoud JM. Systematic selection of prognostic features in patients with severe head injury. Neurosurgery 1980;6(4):362-70.
4. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head-injured patients. J Neurosurg 1991;75(2):251-5.
5. Combes P, Fauvage B, Colonna M, Passagia JG, Chirossel JP, Jacquot C. Severe head injuries: an outcome prediction and survival analysis. Intensive Care Med 1996;22(12):1391-5.
6. 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;66(1):20-5.
7. Sakellaropoulos GC, Nikiforidis GC. Development of a Bayesian Network for the prognosis of head injuries using graphical model selection techniques. Methods Inf Med 1999;38(1):37-42.
8. Schreiber MA, Aoki N, Scott BG, Beck JR. Determinants of mortality in patients with severe blunt head injury. Arch Surg 2002;137(3):285-90.
9. Schaan M, Jaksche H, Boszczyk B. Predictors of outcome in head injury: proposal of a new scaling system. J Trauma 2002;52(4):667-74.
10. Andrews PJ, Sleeman DH, Statham PF, 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;97(2):326-36.
11. 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;16(12):1131-8.
12. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(6):515-24.
13. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(4):453-73.
14. Teasdale GM, Jennett B. Assessment of coma and impaired consciousness. A practical scale. Lancet 1974;2(7872):81-4.
15. Jennett B, Bond M. Assessment of outcome after severe brain damage. A practical scale. Lancet 1975;1:480-484.
16. Marshall LF, Maas AIR, Marshall SB, et al. A multicenter trial on the efficacy of using tirilazad mesylate in cases of head injury. J Neurosurg 1998;89(4):519-25.
17. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. J Neurosurg 2002;97(3):549-57.
18. Morris GF, Bullock R, Marshall SB, Marmarou A, Maas AIR, Marshall LF. Failure of the competitive N-methyl-D-aspartate antagonist Selfotel (CGS 19755) in the treatment of severe head injury: results of two phase III clinical trials. The Selfotel Investigators. J Neurosurg 1999;91(5):737-43.

19. Murray GD, Teasdale GM, Braakman R, et al. The European Brain Injury Consortium Survey of head injuries. Acta Neurochir (Wien) 1999;141:223-236.
20. Marshall LF, Becker DP, Bowers SA, et al. The National Traumatic Coma Data Bank. Part 1: Design, purpose, goals, and results. J Neurosurg 1983;59(2):276-84.
21. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
22. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227 - 1237.
23. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J Clin Epidemiol 2003;56(1):28-37.
24. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. Am J Epidemiol 1995;142(12):1255-64.
25. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis: Springer-Verlag New York, Inc., 2001.
26. Efron B, Tibshirani R. An Introduction to the Bootstrap: Chapman and Hall, New York, 1993.
27. Van Houwelingen JC, Le Cessie S. Predictive value of statistical models. Stat Med 1990;9(11):1303-25.
28. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
29. 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;54(8):774-81.
30. Benzer A, Mitterschiffthaler G, Marosi M, et al. Prediction of non-survival after trauma: Innsbruck Coma Scale. Lancet 1991;338(8773):977-8.
31. 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;66(1):26-31.
32. Hukkelhoven CWPM, Steyerberg EW, Habbema JDF, et al. Outcome after severe or moderate traumatic brain injury: development and validation of a prognostic score based on admission characteristics. Accepted for publication in J Neurotrauma.
33. Breiman L, Freidman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Monterey, Calif., U.S.A.: Wadsworth, Inc., 1984.
34. Steyerberg EW, Eijkemans MJ, Habbema JDF. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52(10):935-42.
35. Steyerberg EW, Eijkemans MJ, Harrell FE, Jr., Habbema JDF. Prognostic modeling with logistic regression analysis: in search of a sensible strategy in small data sets. Med Decis Making 2001;21(1):45-56.
36. Hukkelhoven CWPM, Eijkemans MJ, Steyerberg EW. Predicting survival using simple clinical variables: a case study in traumatic brain injury. J Neurol Neurosurg Psychiatry 2000;68(3):396-7.
37. Ransohoff DF. Rules of evidence for cancer molecular-marker discovery and validation. Nat Rev Cancer 2004;4(4):309-14.
38. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst 2003;95(1):14-8.

39. van Houwelingen HC. Validation, calibration, revision and combination of prognostic survival models. Stat Med 2000;19(24):3401-15.
40. Titterington DM, Murray GD, Murray LS, et al. Comparison of discrimination techniques applied to a complex data set of head injured patients. Journal of the Royal Statistical Society, Series A 1981;144:145-175.
41. Choi SC, Barnes TY, Bullock R, Germanson TA, Marmarou A, Young HF. Temporal profile of outcomes in severe head injury. J Neurosurg 1994;81(2):169-73.
42. Baltas I, Gerogiannis N, Sakellariou P, Matamis D, Prassas A, Fylaktakis M. Outcome in severely head injured patients with and without multiple trauma. J Neurosurg Sci 1998;42(2):85-8.

8

Admission of patients with severe and moderate traumatic brain injury to specialized ICU facilities: a search for triage criteria



C.W.P.M. Hukkelhoven, E.W. Steyerberg, J.D.F. Habbema, A.I.R. Maas

Intensive Care Med 2005; 31:799-806

Abstract

Objective
To investigate whether triage for direct admission of patients with traumatic brain injury to a trauma center is facilitated by predicting the risk of potentially removable lesions or raised intracranial pressure (ICP).

Design and setting
Cohort study in a level I university trauma center.

Patients and participants
A prospective cohort of primarily (n=200) and secondarily (n=75) referred patients with moderate or severe traumatic brain injury.

Measurements and results
Predictive characteristics for the risk of surgically removable lesions and the risk of raised ICP (repeatedly ≥ 20 mm Hg) were identified and included in prognostic models. These models were validated internally with bootstrapping techniques and externally on a historic sample (n=205) regarding discriminative ability (AUC). Among the cohort patients, 67% had raised ICP and 54% had surgically removable lesions. Both outcomes occurred more frequently in patients secondarily referred, but the incidence in patients primarily referred was also high (62% and 33% respectively). No strong predictors of raised ICP were identified. Age and pupillary reactivity were significant predictors of surgically removable lesions. The models discriminated reasonably for surgically removable lesions (AUC=0.78 at development and AUC=0.67 at external validation) but not for raised ICP (AUC=0.59 at development and AUC=0.50 at external validation).

Conclusions
It is difficult to accurately identify patients in need of specialized intensive care using baseline characteristics. The high incidence of both outcomes in patients primarily referred support direct admission of more and particularly older patients with severe or moderate brain trauma to level I trauma centers.

Introduction

Traumatic brain injury (TBI) is one of the most important causes of death in young adults[1]. Survivors are often confronted with severe limitations in their daily life. In patients with severe or moderate TBI therapeutic principles aim at early detection and evacuation of mass lesions and treatment of raised intracranial pressure (ICP). Raised ICP is reportedly the leading cause of inhospital deaths after TBI[2]. Invasive monitoring of ICP and surgical treatment of mass lesions requires admission to a level I trauma center with specialized intensive care and neurosurgical facilities.

In many European countries capacity problems in level I trauma centers prohibit direct admission of all patients with severe or moderate TBI, and many patients are secondarily referred. The European Brain Injury Consortium (EBIC) data survey reported a national proportion of 35% - 75% secondarily referred patients[3]. In European countries participating in a large trial in TBI 24% - 57% of the patients were secondarily referred[4]. Secondary referral may delay the initiation of appropriate therapy, increase the risk of adverse events and systemic insults during inter-hospital transport, and be disadvantageous for the recovery of the patient[5]. On the other hand, some patients admitted primarily to the level I trauma center may have less severe injuries than initially suspected and do not require specialized facilities. Currently these patients occupy scarce beds and absorb care and medical resources. A more efficient triage may be aided by early identification of patients in need of specialized intensive care.

The objective of our study was to analyze differences in baseline characteristics between primarily and secondarily referred patients with TBI and to investigate the feasibility for predicting the need of specialized intensive care, i.e. the risk of (a) potentially operable lesions or (b) raised intractranial pressure (ICP) within 7 days after TBI with simple baseline characteristics. To estimate these risks more precisely we developed prognostic models.

Patients and methods

Patients

We collected data from 275 patients ('cohort') with severe TBI (Glasgow Coma Scale, GCS, 3-8) or moderate TBI (GCS 9-12) admitted to the trauma center of Erasmus Medical Center, Rotterdam, The Netherlands, between 1 January 2000 and 31 March 2003. Erasmus Medical Center functions as a primary care facility for Rotterdam and its immediate surroundings (population 500.000) and as tertiary care facility for a wide region in the southwest of The Netherlands (population 1.5000.000 – 2.000.000). Consequently the cohort represents a mix of primarily referred (directly admitted) and secondarily referred (from other centers) patients (Table 1). For validation of the models developed we selected 205 patients with severe or moderate TBI who had been enrolled in randomized clinical trials at Erasmus Medical Center between 1989 and 1997 (historic sample). Data from treatment and placebo groups were combined for the present analysis, as none of these studies had shown any significant difference between treatment groups.

Table 1. Patient characteristics and outcome measures in patients with traumatic brain injury (n = 275)

Clinical characteristics	**Directly admitted** (n=200)		**Secondary referral** (n=75)	
Age, mean ± SD (years)	42	±20	52	±21
Male gender	156	78%	54	72%
*Cause of injury**				
Traffic accidents	110	55%	29	39%
Domestic/Falls	34	17%	24	32%
Other#	56	28%	21	28%
Missing	0		1	
*Motor score^**				
No response	42	22%	3	4%
Extension	20	11%	5	7%
Abnormal flexion	24	13%	5	7%
Flexion withdrawal	38	20%	7	10%
Localizing	37	20%	21	29%
Obeying commands	27	14%	32	44%
Missing	12^		2	
*Eye score^**				
No reaction	118	62%	21	29%
To pain	20	11%	9	13%
To speech/spontaneous	52	27%	42	58%
Missing	10^		3	
*Verbal score^**				
None	104	57%	18	25%
Incomprehensible	36	20%	13	18%
Inappropriate/confused/oriented	44	24%	42	58%
Missing	16		2	
*GCS^**				
3-8	125	68%	26	36%
≥ 9	59	32%	46	64%
Missing	16		3	
Pupillary reactivity				
Both pupils reacted	127	66%	47	69%
One pupil reacted	17	9%	9	13%
No pupil reacted	47	25%	12	18%
Missing	9		7	

Table 1 (continued)

Pupillary size				
Both wide	27	14%	8	11%
One wide	25	13%	12	16%
Both normal	148	74%	55	73%
Missing	0		0	
*Hypoxia**				
Yes or suspected	72	38%	10	14%
Missing	11		3	
*Hypotension**				
Yes or suspected	26	14%	3	4%
Missing	11		3	
*Hypothermia**				
Yes or suspected	43	25%	6	9%
Missing	28		11	
Injury Severity Score*, mean ± SD	40	±14	32	±10
Time to definitive treatment, mean ± SD (min)	86	±175	477	±732
Outcome measures	**Directly admitted** (n=200)		**Secondary referral** (n=75)	
Raised intracranial pressure (ICP)‡				
Yes	63	62%	23	70%
Missing	99		42	
*Surgically removable lesion**				
Yes	66	33%	49	66%
Missing	1		1	

* $p < 0.05$
\# Includes accidents at work, during sports, falls under the influence of alcohol , assaults, and other
^ Score at or, if missing, before admission to the neurosurgical unit
‡ ICP ≥ 20 mm Hg

Predictors and outcome

As potential predictors for the need of specialized intensive care we considered age, gender, cause of injury, motor score, hypotension, hypoxia, pupillary response, and Injury Severity Score (ISS). These patient characteristics have been previously identified as important predictors of poor outcome[6-11] and are easy to assess with high interobserver agreement. Other potentially important predictors such as the duration and degree of hypotension, duration of hypoxia, time from injury to referral, and time to definitive therapy were not considered for analysis since these characteristics would not be available at the site of injury. Age and ISS were included as continuous variables. Cause of injury was categorized into traffic accidents, falls, and other causes, including work related injuries, sports injuries, and assaults. Hypotension and hypoxia were defined by systolic blood pressure below 90 mmHg and pO_2 below 60 mmHg or if supported by strong clinical suspicion. The pupils were defined as wide if their size was 5 mm or larger. All characteristics were measured before or on admission to the neurosurgical unit.

The two outcome measures were: (a) development of a large hematoma, potentially eligible for surgery (volume hematoma ≥ 25 ml and/or size hematoma > 1 cm and/or mass effect) within the first 7 days after the injury and (b) occurrence of raised ICP (repeatedly ≥ 20 mm Hg) within the first 7 days after the injury. If monitored, the ICP was assessed each hour. To determine raised ICP two persons (C.H., F.S.) evaluated all hourly measured ICP-values ('eye-ball' assessment), while blinded to the values of the potential predictors of a patient. ICP values were weighted according to their deviance of the limit of 20 mmHg, for example, three consecutive values of 21 mmHg were not considered as raised ICP, while two values of 40 mmHg were. In the case of any doubt the two evaluators conferred to reach consensus. The historic sample contained daily ICP measurements. If any of these daily measurements was 20 mmHg or higher, the patient was categorized as having raised ICP.

Since the ICP was monitored only for a limited number of the patients (134 patients of the cohort and 180 patients of the historic sample) we examined whether radiological signs of raised ICP, i.e. midline shift of 5 mm or more and compressed or absent basal cisterns on computed tomography within the first 7 days, could be used as an alternative outcome measure for monitored raised ICP. Unfortunately, a poor correlation was observed between documented raised ICP and radiological signs of raised ICP (only 65% of the cohort patients and 46% of the historic patients had matching outcomes). We therefore limited our further analysis to patients in whom ICP had been monitored invasively.

Values of measured predictors were missing in some patients (3.1% in the development sample, 1.5% in the historic sample). These values were estimated statistically based on regression models including the other predictors, and subsequently imputed[12,13]. Such imputation is recommended as more efficient than dropping cases[12]. Thus, all patients contributed to the development of the prognostic models. In the historic sample the ISS was not assessed. We followed a conservative approach by assigning the median ISS of the cohort as the ISS for all historic patients. We chose a 7-day window for the outcome measures as the vast majority of problems related to the development of operable lesions and raised ICP occurs within this time period. Age was quantified in years, ISS in points; all other predictors were coded as '1' if true and '0' if false.

Development and validation of the prognostic models

From the group of nine potential predictors we selected those that had a p-value < 0.50 in a backward stepwise logistic regression procedure[14,15]. Subsequently, we developed a prognostic model for each outcome measure using logistic regression analysis. Internal validity was assessed with standard bootstrapping procedures[15,16]. Bootstrapping involved drawing samples with replacement from the development sample. Each sample can be seen as if one is repeating the data collection with the same number of patients and under identical circumstances as the original. Multivariable prediction models were estimated in 200 bootstrap samples and each evaluated on the original sample. The average difference in performance indicates the optimism (overfitting)[12]. This is the phenomenon that a model predicted outcome well in the development sample, but tends to predict too extreme probabilities in new patients. Subsequently the coefficients were corrected (decreased) to provide better predictions for future patients[12,14]. External validity, that is, whether the models perform well for patients from another setting, was assessed on the historic sample.

Performance

Performance of the models was assessed with regard to calibration and discrimination. Calibration was assessed graphically by plotting observed outcome against the predicted probability. A smooth, non-parametric calibration line was created with the lowess algorithm[15]. Calibration was tested with the Hosmer-Lemeshow goodness-of-fit test. Discrimination is defined as the model's ability to differentiate patients with different outcomes. To quantify the discrimination we used the area under the receiver operating characteristic curve (AUC), which considers pairs of sensitivity and specificity for consecutive cutoff points of the predicted probabilities from a model (0-100%). The AUC indicates among all possible pairs of patients with different outcomes, the likelihood that those with higher predictive risk indeed are more likely to have a poor outcome (raised ICP or surgically removable lesions). The higher the AUC, the better the model discriminates. A model with an AUC of 0.50 has no discriminative power at all, while an AUC of 1.0 reflects perfect discrimination.

The robustness of the prognostic models was examined by performing sensitivity analyses. We repeated the analyses for those patients with a direct (primary) referral to the Erasmus Medical Center. Calculations were performed using SAS version 6.12 (SAS Institute INC., Cary, NC, USA) and S-plus version 2000 (Insightful Corporation, Seattle, Wash., USA).

Table 2. Associations between predictors and monitored raised ICP in patients with moderate or severe traumatic brain injury (n=134)

Predictors	Raised ICP	OR uni (95% CI*)	OR multi (95% CI*)
Age#			
26 years	na	- §	- §
61 years	na	1.2 (0.6 – 2.3)	1.2 (0.6 – 2.4)
Gender			
Male	69 (64%)	- §	-
Female	17 (65%)	1.1 (0.4 – 2.6)	-
Cause of injury			
Traffic accident	49 (64%)	1.0 (0.4 – 2.3)	-
Domestic/falls	15 (63%)	0.9 (0.3 – 2.7)	-
Other causes	22 (65%)	- §	-
Motor score			
No reaction/extension	23 (79%)	0.8 (0.3 – 1.8)	0.7 (0.3 – 1.8)
Abnormal flexion/flexion withdrawal	27 (66%)	0.9 (0.4 – 2.2)	1.0 (0.4 – 2.4)
Localizing/obeying	31 (65%)	- §	- §
Pupillary reactivity			
Both pupils reacted	50 (63%)	- §	- §
One pupil reacted	10 (63%)	1.0 (0.3 – 3.1)	0.8 (0.2 – 2.8)
No pupil reacted	24 (69%)	1.3 (0.6 – 3.1)	1.5 (0.5 – 4.2)
Pupillary size			
None wide	55 (63%)	- §	- §
One wide	10 (71%)	1.5 (0.6 – 3.8)	1.6 (0.6 – 4.8)
Both wide	11 (61%)	0.9 (0.3 – 2.7)	0.9 (0.2 – 3.2)
Hypoxia			
No	55 (65%)	- §	-
Yes or suspected	26 (63%)	0.9 (0.4 – 2.0)	-
Hypotension			
No	74 (65%)	- §	- §
Yes or suspected	9 (56%)	0.7 (0.2 – 1.9)	0.8 (0.2 – 2.3)
ISS‡			
26	na	- §	- §
45	na	0.9 (0.6 – 1.5)	1.0 (0.6 – 1.7)

* If this interval does not include the value 1, the factor has a statistically significant effect on the outcome
\# 21 years is the 25th percentile and 42 years the 75th percentile, including age as a continuous linear term
‡ 26 is the 25th percentile and 45 the 75th percentile, including ISS as a continuous linear term
§ Reference
OR = odds ratio, uni = univariable logistic regression analysis, multi = multivariable logistic regression analysis, na = not applicable since percentiles were used for describing the association between age and outcome or ISS and outcome, CI = confidence interval*

Table 3. Associations between predictors and surgically removable lesions in patients with moderate or severe traumatic brain injury (n=275)

Predictors	Removable lesions	OR uni (95% CI*)	OR multi (95% CI*)
Age#			
26 years	na	- §	- §
61 years	na	3.4 (2.2 – 5.4)	3.3 (2.0 – 5.6)
Gender			
Male	85 (41%)	- §	-
Female	30 (48%)	1.3 (0.8 – 2.4)	-
Cause of injury			
Traffic accident	45 (33%)	0.6 (0.3 – 1.1)	0.7 (0.4 – 1.3)
Domestic/falls	35 (61%)	2.0 (1.0 – 4.1)	1.1 (0.5 – 2.4)
Other causes	34 (44%)	- §	- §
Motor score			
No reaction/extension	28 (40%)	0.7 (0.4 – 1.3)	-
Abnormal flexion/flexion withdrawal	26 (35%)	0.6 (0.3 – 1.1)	-
Localizing/obeying	55 (47%)	- §	-
Pupillary reactivity			
Both pupils reacted	62 (36%)	- §	- §
One or no pupil reacted	46 (55%)	2.1 (1.3 – 3.6)	2.2 (1.1 – 4.5)
Pupillary size			
None wide	79 (39%)	- §	- §
One or both wide	36 (50%)	1.5 (0.9 – 2.7)	1.6 (0.8 – 3.5)
Hypoxia			
No	77 (43%)	- §	-
Yes or suspected	31 (38%)	0.8 (0.4 – 1.3)	-
Hypotension			
No	99 (43%)	- §	- §
Yes or suspected	9 (31%)	0.6 (0.2 – 1.3)	1.2 (0.4 – 3.0)
ISS‡			
26	na	- §	- §
45	na	0.5 (0.3 – 0.7)	0.5 (0.3 – 0.7)

* If this interval does not include the value 1, the factor has a statistically significant effect on the outcome
\# 21 years is the 25th percentile and 42 years the 75th percentile, including age as a continuous linear term
‡ 26 is the 25th percentile and 45 the 75th percentile, including ISS as a continuous linear term
§ Reference
OR = odds ratio, uni = univariable logistic regression analysis, multi = multivariable logistic regression analysis, na = not applicable since percentiles were used for describing the association between age and outcome or ISS and outcome, CI = confidence interval*

Results

Patient characteristics and outcome

The characteristics and outcome of the cohort patients are presented in Table 1, differentiated into primary (73%) and secondary (27%) referrals. Most patients were male, both in the primarily (78%) and the secondarily referred group (72%). The distribution of many baseline characteristics differed between the two groups. In general, secondarily referred patients were older (52 vs. 42 years), were more frequently injured in the domestic setting or by a fall (32% vs. 17%), and had initially less severe clinical characteristics (higher GCS, fewer secondary insults).
If monitored, raised ICP was observed frequently, both in directly admitted (62%) and secondarily referred patients (70%). Potentially operable lesions were observed in 42% of the patients in the cohort, more frequently in those secondary referred (66% vs. 33% in the directly admitted patients). When including only ICP monitored patients, the need for specialized intensive care (i.e. the presence of one or both outcome measures) was present in 97% of patients secondarily referred and in 68% of patients directly admitted.

The historic sample was more homogeneous with respect to the baseline clinical characteristics the historic sample was more homogeneous, with fewer patients in the motor score categories 'no response' and 'obeying command' (8% and 3% respectively). A large proportion of the patients in the historic sample (84%) was classified as having a severe TBI. Gender and the occurrence of hypotension were similar to the cohort study. Raised ICP was noted in 54% of the ICP monitored patients in the historic sample (88% monitored) and potentially operable lesions were present in 49%.

Predictors

Both univariable and multivariable analyses showed small effects of most potential predictors on the two outcome measures. No strong predictors of raised ICP were identified, and only age and pupillary reactivity were statistically significant (p-value < 0.05) predictors of potentially operable lesions (Table 2 and 3). Multivariable analyses with a high p-value (p-value < 0.50) included age, motor score, pupillary reactivity, pupillary size, hypotension, and ISS as potential predictors of raised ICP and age, cause of injury, pupillary reactivity, pupillary size, hypotension and ISS as potential predictors of surgically removable lesions. Details of the developed prognostic models are presented in the **Appendix**.

Table 4. Discrimination and calibration of the prediction rules. The rules were developed in patients at the Erasmus Medical Center in 2000 – 2003 (cohort, n = 275) and externally validated in patients there in 1989 – 1997 (historic sample, n = 205)

Outcome measures	Cohort (n=275)		Historic sample (n=205)	
	Discrimination AUC (95% CI)	Calibration P-value *	Discrimination AUC (95% CI)	Calibration P-value *
Monitored raised ICP	0.59 (0.48 – 0.69)	0.42	0.50 (0.41 – 0.58)	0.18
Surgically removable lesions	0.78 (0.72 – 0.83)	0.66	0.67 (0.60 – 0.75)	0.01

* Hosmer-Lemeshow goodness-of-fit test, low p-values indicate poor goodness-of-fit
AUC = area under the receiver operating curve, CI = confidence interval

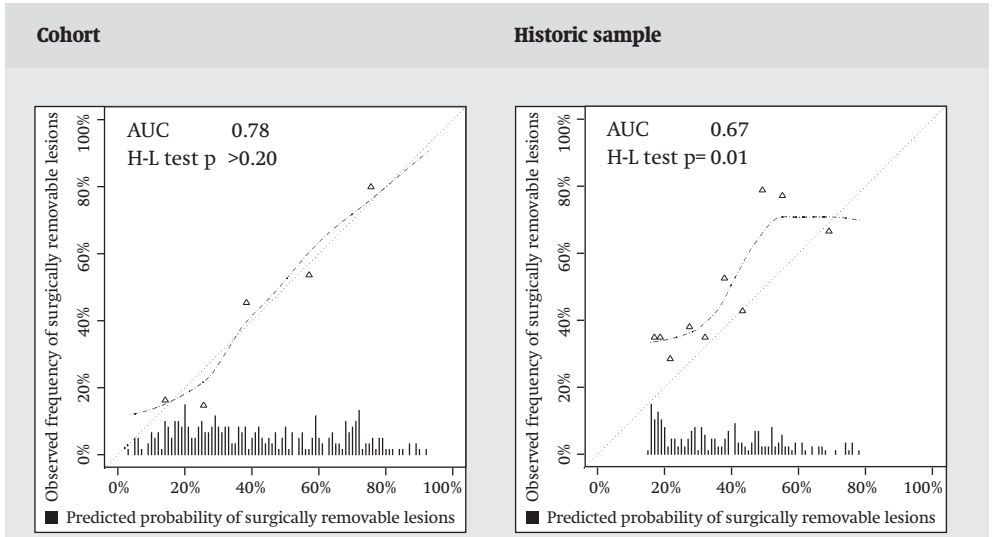
Table 5. Classification of patients into three risk categories, according to their predicted risk of surgically removable lesions, as estimated by the prognostic model (n=275)

Risk of surgically removable lesions	Direct admission (n=199)		Secondary referral (n=74)	
	n	%	n	%
< 0.40	122	61	20	27
0.40 – < 0.60	36	18	23	31
≥ 0.60	41	21	31	42

Performance of the models

Patients with and without raised ICP could not be distinguished; AUC was 0.59 in the cohort with the 95% confidence interval (CI) of the AUC including the value 0.50, and AUC was 0.50 in the historic sample (Table 4). Surgically removable lesions were predicted with reasonable discrimination (AUC = 0.78, 95% CI 0.72 – 0.83 in the cohort; AUC = 0.67, 95% CI 0.60 – 0.75 in the historic sample, a decrease of 0.11), also if the model included only those patients who were primarily referred to the trauma center (AUC = 0.73, 95% CI 0.66 – 0.81). When the model excluded ISS, the decrease in performance was only slightly smaller (AUC = 0.74 in the cohort; AUC = 0.67 in the historic sample, a decrease of 0.07). Discriminative ability of the model predicting surgically removable lesions is also demonstrated in Table 5; 82% of the directly admitted and 69% of the secondarily referred cohort patients were categorized in one of the two extreme categories. Calibration of the model predicting surgically removable lesions was satisfactory in the cohort, but poor in the historic sample (p-value = 0.01, Hosmer-Lemeshow goodness-of-fit test). Figure 1 shows discrimination and calibration of the model predicting surgically removable lesions in more detail.

Figure 1. Validation of the model predicting surgically removable lesions in the cohort (n = 275) and the historic sample (n = 205). *Dashed smooth curves* reflect the relationship between observed and predicted probability of surgically removable lesions; *straight dotted line* through the origin perfect calibration; *triangles* incidence of poor outcome in deciles of patients with similar predicted probabilities; *spikes* at the bottom of each graph distribution of predicted probabilities



Discussion

Various studies have reported prognostic models of baseline characteristics for predicting outcome after TBI[6-10,17-20], but to our knowledge no previous studies have attempted to predict the need of specialized intensive care for patients with TBI. We observed that several baseline characteristics were associated with this need, with age and pupillary reactivity showing statistical significance when predicting surgically removable lesions. Our results, however, indicate that it is not possible to predict the need for specialized intensive care in patients with severe or moderate TBI with sufficient confidence to use such predictions for refining triage criteria.

Several factors may have contributed to the poor discriminative ability of the model predicting raised ICP. We only investigated a threshold value of 20 mm Hg for raised ICP by eyeball assessment, and the relatively small sample size of the development population (134 patients with monitored ICP) may have hampered the identification of predictors and the assessment of the regression coefficients of the model. The relatively poor calibration of the model predicting surgically removable lesions at external validation may be explained by the selection of patients in the historic sample according to more rigorous criteria, as well as by the different (earlier) period of data collection. Limited calibration during external validation is, however, observed often for prognostic models and indicates that the model may require adjustment for specific circumstances before it can be used in new populations[21,22].

Several limitations of our study should be acknowledged. First, we only collected data from patients who were admitted to the trauma center of the Erasmus Medical Center. Consequently, we do not know how many other patients may have needed specialized intensive care. This may have caused some selection bias. However, since the local trauma policy is that all patients having sustained severe or moderate traumatic injury within the Rotterdam area are primarily referred to the level I trauma center of the Erasmus Medical Center, this subpopulation of patients referred may be considered representative of the overall population of severe and moderate TBI. Second, the models might be improved by including other patient characteristics or information obtained at later time periods. Further, predictors were measured at or, occasionally, before baseline. Before a model like this can actually be implemented for optimizing the initial triage of TBI patients, its validity needs to be confirmed with pre-hospital data.

Although we were not able to predict the need of intensive care after TBI, we observed that the rates of both outcome measures in patients primarily referred were high: 33% of the patients had potentially operable lesions and in 62% had raised ICP among those with ICP monitored. Even if we would consider raised ICP to be absent in patients not monitored, the rate of one or both outcomes occurring in the population studied was 48%. This high incidence indicates that the triage criteria, according to which all patients with a Revised Trauma Score[23] lower than 11 and/or a GCS below 12 are referred, are relatively specific, i.e., relatively successful in selecting patients at risk for intracranial complications. On the other hand, the triage criteria are less sensitive since the proportion of secondarily referred patients, who almost all needed specialized intensive care, was substantial (27%).

Secondary referral may delay initiation of appropriate monitoring and surgical intervention and incurs the additional risk of adverse events during interhospital transport[24-26]. Studies on adult and pediatric populations have shown initial treatment at a local hospital with subsequent transfer to a trauma center to be associated with a doubling of mortality[27]. A general approach to prevent secondary referral is to get the right patient to the right facility at the right time. Definitions of the right facility and the right time have been proposed in the international guidelines[28] as a center with 24 h computed tomography service and neurosurgical facilities and expertise in the treatment of TBI, which should be reached as soon as possible. Defining the right patient, however, is complex. The high incidence of both outcomes in patients primarily referred, together with the substantial proportion of patients secondarily referred may support more liberal triage criteria, for example, admitting more patients with moderate TBI directly to a level I trauma center. Since secondarily referred patients were generally older, and age was identified as a significant predictor, especially older patients may be suitable for such a policy. The final answer to whether such a more liberal admission policy is appropriate can be answered definitively only following a population-based study. Alternatively, the risk of secondary referral may be reduced by using a specialist retrieval team[29].

Expanding the triage criteria of the trauma centers to older, moderately injured TBI patients would have considerable consequences for local trauma organization and capacity of tertiary referral centers. The majority of patients primarily admitted to our center originate from the immediate surroundings of Rotterdam. If we extrapolate the expected number of admissions to the wider region for tertiary referral, a two- to threefold increase in the number of admitted patients might be expected. Nevertheless, such a liberal admission policy, if cost effective, may be considered preferable to the current approach in which patients are only secondarily referred after intracranial complications have developed and the neurological condition deteriorated. The negative consequences on capacity may in part be compensated by early transferral to a regional hospital of those patients in whom intracranial complications do not develop after a couple of days. Such a policy may limit the risks of additional interhospital transport in the acute phase, offer facilities for prompt initiation of specialized management, and is in accordance with the concept of concentration of care within trauma centers.

Acknowledgements

The authors express their gratitude to all of the study participants whose cooperation made this report possible. The authors thank Frans Slieker for evaluating the hourly measured ICP values of the cohort patients, together with the first author. The authors also thank Marja van Gemerden for her administrative support and Tineke Landman and other coworkers of the neurological trauma center of the Erasmus Medical Center for their assistance during the data collection.

Appendix

Details of the prognostic models.

The probability of a poor outcome (raised ICP or surgically removable lesions) is calculated as $1/(1 + \exp^{-LP})$.
The linear predictor (LP) takes the form of LP=intercept + regression coefficients × **predictor** values.

LP# for monitored raised ICP = **0.55 + (0.00089** × age) - (**0.062** × motor score equal to 'no response' or 'extension') + (**0.00091** × motor score equal to 'abnormal flexion' or 'flexion withdrawal') - (**0.049** × one pupil reacts) + (0.076 × no pupil reacts) + (**0.098** × one pupil wide) - (0.033 × both pupils wide) - (0.059 × hypotension) - (0.00013 × injury severity score).

LP# for radiological signs of raised ICP = **- 0.64 + (0.025** × age) + (**0.22** × motor score equal to 'no reponse' or 'extension') + (**0.48** × motor score equal to 'abnormal flexion' or 'flexion withdrawal') + (**0.52** × one pupil reacts) + (1.20 × no pupil reacts) + (1.0068 × one pupil wide) + (1.17 × both pupils wide) + (0.50 × hypotension) - (0.027 × injury severity score).

LP# for surgically removable mass lesions = **- 0.60 + (0.031** × age) - (0.33 × traffic accident) + (0.062 × accident at home or fall) + (0.72 × one or none pupil react) + (0.44 × one or both pupils wide) - (0.036 × injury severity score).

Coding of the predictors:
Age in years; ISS in points; All other predictors: 1 if true and 0 if false.

References

1. Ghajar J. Traumatic brain injury. **Lancet 2000;356(9233):923-9.**
2. Marshall L, Gautille T, Klauber M, et al. **The outcome of severe closed head injury. J Neurosurg** 1991;75:S28-S36.
3. Murray GD, Teasdale GM, Braakman R, et al. **The European Brain Injury Consortium Survey of** head injuries. Acta Neurochir (Wien) 1999;141:223-236.
4. Hukkelhoven CWPM, Steyerberg EW, Farace E, Habbema JDF, Marshall LF, Maas AIR. Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. J Neurosurg 2002;97(3):549-57.
5. Seelig JM, Becker DP, Miller JD, Greenberg RP, Ward JD, Choi SC. Traumatic acute subdural hematoma: major mortality reduction in comatose patients treated within four hours. N Engl J Med 1981;304(25):1511-8.
6. Choi SC, Muizelaar JP, Barnes TY, Marmarou A, Brooks DM, Young HF. Prediction tree for severely head-injured patients. **J Neurosurg 1991;75(2):251-5.**
7. Andrews PJ, Sleeman DH, Statham PF, 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;97(2):326-36.
8. Combes P, Fauvage B, Colonna M, Passagia JG, Chirossel JP, Jacquot C. Severe head injuries: an outcome prediction and survival analysis. Intensive Care Med 1996;22(12):1391-5.
9. 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;66(1):20-5.**
10. Braakman R, Gelpke GJ, Habbema JDF, Maas AIR, Minderhoud JM. **Systematic selection of** prognostic features in patients with severe head injury. Neurosurgery 1980;6(4):362-70.
11. Narayan RK, Kishore PR, Becker DP, et al. Intracranial pressure: to monitor or not to monitor? A review of our experience with severe head injury. J Neurosurg 1982;56(5):650-9.
12. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
13. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227 - 1237.
14. Steyerberg EW, Eijkemans MJC, Harrell FE, Jr., Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
15. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis: Springer-Verlag New York, Inc., 2001.
16. Efron B, Tibshirani R. An Introduction to the Bootstrap: Chapman and Hall, New York, 1993.
17. Sakallaropoulos G, Nikiforidis G. Development of a Bayesian Network for the prognosis of head injuries using graphical model selection techniques. Methods Inf Med 1999;38:37-42.
18. Schreiber MA, Aoki N, Scott BG, Beck JR. Determinants of mortality in patients with severe blunt head injury. Arch Surg 2002;137(3):285-90.
19. Schaan M, Jaksche H, Boszczyk B. Predictors of outcome in head injury: proposal of a new scaling system. **J Trauma 2002;52(4):667-74.**
20. Hukkelhoven CWPM, Steyerberg EW, Habbema JDF, et al. **Predicting outcome after traumatic** brain injury: development and validation of a prognostic score based on admission characteristics. J Neurotrauma, accepted for publication.
21. van Houwelingen HC. **Validation, calibration, revision and combination of prognostic survival** models. Stat Med 2000;19(24):3401-15.
22. Altman DG, Royston P. What do we mean by validating a prognostic model? Stat Med 2000;19(4):453-73.
23. Champion HR, Sacco WJ, Copes WS, Gann DS, Gennarelli TA, Flanagan ME. A revision of the Trauma Score. J Trauma 1989;29(5):623-9.
24. Andrews PJ, Piper IR, Dearden NM, Miller JD. Secondary insults during intrahospital transport of head-injured patients. **Lancet 1990;335(8685):327-30.**
25. Bleeker JK, Rutten FL, van Leeuwen FL, Jansen YG. [The quality of ambulance transportation between regional hospitals and a central hospital] De kwaliteit van het ambulancevervoer tussen streekziekenhuizen en een centrumziekenhuis. Ned Tijdschr Geneeskd 1993;137(22):1091-5.
26. Gentleman D. Causes and effects of systemic complications among severely head injured patients transferred to a neurosurgical unit. Int Surg 1992;77(4):297-302.
27. Smith JS, Jr., Martin LF, Young WW, Macioce DP. Do trauma centers improve outcome over non-trauma centers: the evaluation of regional trauma care using discharge abstract data and patient management categories. **J Trauma 1990;30(12):1533-8.**
28. Bullock R, Chesnut R, Clifton G, et al. **Management and prognosis of severe traumatic brain** injury. Part 1: Guidelines for the management of severe traumatic brain injury. **J Neurotrauma** 2000;17(6/7):451-553.
29. Spoedeisende hulpverlening: haastige spoed niet overal goed. Geneeskundige Inspectie van de Volksgezondheid, 2004.

9

General discussion

This thesis addresses three research questions regarding prognosis for patients with severe or moderate TBI. The focus is on developing and validating prognostic models that use baseline patient characteristics to predict long-term outcome and the need for specialized intensive care. In this chapter the findings of our studies are first discussed in the light of the research questions as formulated in Chapter 1. Furthermore, implications of the results for clinical practice are presented. Finally, conclusions are drawn and suggestions for further research are given.

Study findings

1. *Methodological developments in prognostic modeling in TBI.*

The review of 26 prognostic modeling studies highlighted important deficiencies in the methods used to develop and validate prognostic models in TBI. Many previously developed models were limited by old, small and relatively homogeneous study populations for their development and by the way predictors were chosen. Furthermore, rather crude statistical methods were used and patients with missing predictor values were often omitted from the study population. Additionally, models were seldom validated on more recent patients from the same place, and never on patients from another place.
Based on the observed deficiencies we proposed guidelines to develop and validate prognostic models. The guidelines concern five subjects, i.e. study population, predictors, outcome, model development and model validation. We hope that our guidelines contribute to improve the quality of prognostic modeling and validation in future TBI studies.

2. *Construction and validation of prognostic models that predict long-term outcome for patients with severe or moderate TBI*

We first examined the characteristics of the study population (the Tirilazad cohort). The cohort was relatively heterogeneous, which was partly explained by regional differences. Remarkably, outcome was better in patients treated in the United States than those treated in Europe or other countries. We found that age should be included in a prognostic model as a continuous predictor, and that various combinations of CT characteristics could well predict outcome. We developed two logistic regression models that predict six-month mortality and unfavorable outcome. Both models included the same seven predictors, measured at baseline. The models provided high discrimination between patients with good and poor six-month outcome (AUC 0.78 to 0.80 in the development series and AUC 0.83 to 0.89 at external validation). The models showed satisfactory calibration in most patient series. We compared the performance of our seven-predictor models with several other models, including prediction trees and logistic regression models. Better discrimination was observed for the logistic regression models and for the models including more predictors.
Overall, we consider our seven-predictor models suitable for clinical practice, especially to classify patients according to prognostic risk. Carefulness is however required, because an adjustment for possible poor calibration may be necessary.

3. *Predict the need of specialized intensive care.*

We developed two prognostic models, one predicting the risk of surgically removable lesions and one predicting the risk of raised ICP, using baseline characteristics. Both outcome measures are indicators of the need of specialized intensive care. The models could reasonably discriminate patients with and without surgically removable lesions (AUC = 0.78 at development and AUC = 0.67 at external validation in a historic sample of 205 patients from the Erasmus MC), but could – unfortunately – not discriminate patients with and without raised ICP. Overall, the models were not considered suitable for implementation in clinical practice.

Methodological considerations

Below, we discuss the different steps we have taken to construct and validate the models and we relate these steps to the methodological guidelines for model development and validation (Table 1), as proposed in Chapter 2.

Models predicting six-month mortality and unfavorable outcome

Study population
We intended to use the Tirilazad patients for model development since we regarded it as close to ideal: with more than 2200 – consecutively admitted – patients it is the largest available data set on severe and moderate TBI. The cohort originated from multiple centers, i.e. 76 centers from 17 countries in North America, Europe, Australia and Israel. In Chapter 3 we observed that the Tirilazad cohort was relatively heterogeneous. Thus, the first two guidelines for model development were met (Table 1). A limitation of the series, however, was that data were collected more than 10 years ago (from 1991 to 1994). Although no major diagnostic or therapeutic changes have taken place since that time, the validity of the models in current clinical practice requires further study.

In the studies described in this thesis treatment and placebo-groups were combined, since in neither trial a statistically significant difference between the treatment and placebo group was shown for the primary outcome (GOS at 6 months after the injury). Nevertheless, we realize that some (even negative) treatment effect may be present. In the North American Tirilazad trial mortality was significantly higher in treatment group. However, this effect was considerably less than the prognostic effects of the predictors. Moreover, the Tirilazad patients were randomized between the treatment and the placebo group, which guaranteed balance with respect to observed and unobserved characteristics. Consequently, we considered it legitimate to combine treatment and placebo groups in addressing our research questions.

Table 1. Guidelines for developing and validating prognostic models in TBI

Guidelines	Model for outcome	Model for need of care
Study population		
• Large, well-defined cohort	+	-
• Heterogeneous	+/-	+/-
• Representative for current clinical practice	+/-	+
Predictors		
• Plausible	+	+
• Precisely defined	+	+
• Readily available or easily obtainable	+	+
Outcome		
• Relevant for clinical practice	+	+ and +/-*
• Precisely defined	+ and +/-#	+
• Measurable with minor observer variability	+/-	+
Model development		
• Use of appropriate statistical techniques	+	+
• Use of sensible performance measures, evaluating calibration and discrimination	+	+
• Presentation in a readily applicable format	+	na
Model validation		
• Internal validation	+	+
• External validation	+	-

* + for the model predicting surgically removable lesions, +/- for the model predicting raised ICP
\# + for the model predicting mortality, +/- for the model predicting unfavorable outcome;
na = not applicable

Predictors

After identifying the association between age and outcome, we studied the literature to determine which predictors might be eligible for inclusion in the models. Nine predictors were considered, i.e. age, gender, cause of injury, motor score, hypotension, hypoxia, pupillary reactivity, the Marshall CT-classification and the presence of traumatic subarachnoid hemorrhage (tSAH).

All nine candidate predictors were previously identified as important[1,2] and are thus plausible. Additionally, they are precisely defined and can be obtained easily and reliably within the first four hours after the injury. Consequently, the candidate predictors fulfill the guidelines for predictors, as proposed in Chapter 2 (Table 1).

The observed value of the predictors may be affected by several factors, such as treatment. For instance, the motor score may become untestable by paralysis and – to a lesser extent – sedation. Deep sedation may also affect pupillary reactivity. The influence of such treatments on the prognostic value of these predictors needs further study in more recent data sets.

Some potential predictors were not considered for inclusion in the models. For example, the verbal and the eye score of the GCS were dropped because they are not testable for patients in coma. Consequently, their discriminating ability will be limited and only relevant for patients with moderate TBI. In Chapter 3 we identified a difference in outcome between patients in North America and patients in Europe, which could not be explained by adjustment for a large set of potential confounders. Confounders included clinical characteristics, management and referral policy. Consequently, one might suggest that 'continent' (as a proxy for one or more unknown underlying characteristics) is a potential predictor for outcome. Nevertheless, we did not consider this suggestion since no plausible explanation could be found and the difference may be caused by coincidence. Furthermore, including 'continent' as predictor may have resulted in a lack of 'face-validity' of the model, thus hampering implementation of the models in clinical practice. The observed outcome difference between continents requires further thorough study.

In Chapter 5 we observed that a combination of several individual CT characteristics discriminates better than the Marshall CT-classification. In our models, however, we included only the Marshall CT-classification and tSAH, because the models were developed before the predictive ability of various combinations of CT characteristics had been explored. Although the Marshall CT-classification is well known, which may enhance implementation of the model in clinical practice, a new model in which the Marshall CT-classification is replaced by several individual CT characteristics may perform better. This subject requires further study.

In this thesis we focused on characteristics that were previously identified as important predictors. It may however be possible that other – less generally accepted – characteristics add prognostic value. For instance, blood parameters, such as glucose or hemoglobin, are easy to assess shortly after admission. Also, extracranial injuries may be important in outcome prediction[3-6]. Their potential predictive power needs to be examined further.

Outcome

Only few patients with an unfavorable outcome are expected to become favorable after six months post-injury[7]. Both outcome measures (mortality and unfavorable outcome) are relevant for clinical practice, since they assess overall outcome after TBI (Table 1).

The developed models predicted six-month mortality and six-month unfavorable outcome. Mortality was derived from the presence of a date of death and could be assessed very accurately with a time-window of maximally 24 hours. For unfavorable outcome a much broader time-window of two months was used. Moreover, if patients had no GOS measurements within this time-window (12%), GOS was estimated according to a weighting algorithm, based on values at other points in time. This algorithm was developed specifically for our studies (see appendix of Chapter 3). Patients with missing GOS, for whom the GOS could also not reasonably be imputed (5%), were omitted from the study population. Omission of these patients may have caused selection bias, because patients with a good outcome may have been more likely to withdraw from follow-up. However, this bias would have been larger if all patients with missing outcome were simply deleted from the data set. Also imputing missing outcome values with the last known outcome value (Last Value Carried Forward), a generally applied approach for missing outcome values, is expected to create more bias than our algorithm, as outcomes usually

improve over time. Consequently, we consider mortality as precisely defined and unfavorable outcome as defined with satisfactory precision (Table 1).

The third proposed guideline for outcome, i.e. measurable with a low degree of observer variability, is partially fulfilled for unfavorable outcome. For the GOS some degree of inter-observer variability has been observed[8], although this may be diminished by using a structured interview during GOS assessment[9].

Model development

In Chapter 6 of this thesis, we developed prognostic models to predict mortality and unfavorable outcome for TBI patients, using logistic regression analysis. Out of nine potential predictors, we selected seven predictors (age, motor score, pupillary reactivity, hypotension, hypoxia, Marshall CT classification and tSAH), which were included in a multivariable logistic regression model. Predictors were selected with backward stepwise selection, using a p-value of 0.20. Such a liberal p-value balances simplicity of the model and predictive information and it may be preferable to the usually applied p-value of 0.05[10,11]. Bootstrapping techniques were used for internal validation of the model[10].

Most predictors were included in such a way that the relation with outcome was adequately described. For instance, age was included as a quadratic and linear term, as identified in Chapter 4. The other predictors were dichotomous or categorical. For motor score some categories were combined to increase numbers per class, although it may have caused loss of information. Further loss of information may have been caused by the fact that we included motor score as 'dummy'-variables, thus ignoring its ordinal scale. However, before motor score can be included as an ordinal scale, further study is needed, e.g. on the size of the steps between the different scores. For the Marshall CT-classification the categories 'mass lesion evacuated' and 'mass lesion non-evacuated' were combined, since the distinction between these is artificial.

For several patients, values of some predictors were missing (5% of the required values). These values were statistically estimated with logistic regression, including the other predictors, and subsequently imputed in a single imputation procedure[10,12]. Such imputation is considered valid for patient series with a relatively small proportion of missing values. For data sets with a larger proportion of missing values multiple imputation is recommended[13].

Overall, we consider the statistical techniques used to develop our models satisfactory (Table 1).

The second and third guideline for model development were 'the use of sensible performance measures' and 'presentation in a readily applicable format'. These were well accomplished (Table 1). A score chart was presented to facilitate the quick estimation of the predicted probability (see Chapter 6). In this chart, each clinical characteristic is assigned a score. These scores can be easily added into a sum score that, through the logistic formula, corresponds with the predicted probability of mortality or unfavorable outcome.

To develop prognostic models for TBI patients various statistical techniques have previously been used, such as logistic regression analysis, recursive partitioning and neural networks. In Chapter 7 we observed a slightly better performance for the logistic regression models, compared

to the models developed with recursive partitioning, while both considered the same predictors. Chapter 7 also showed that the performance of a model depended more on the selection and number of predictor variables than on the applied analysis technique. The same phenomenon has been observed earlier by others[14]. Various studies observed that the performance of various statistical techniques is similar when the same set of predictors is considered [15-17]. Therefore, it is not plausible that further improvement in model performance may come from other qualified statistical techniques. Further improvement should come from inclusion of more – both independent and more powerful – predictors.

Model validation

The methods used for validation largely met the proposed guidelines (Table 1). The models were validated internally by bootstrapping techniques and externally on three different patient series. These validation populations were all relatively large and were collected at different time points in different parts of the Western world. Validity was assessed regarding calibration and discrimination.

Performance was good with respect to discrimination. Calibration according to the Hosmer-Lemeshow statistic was sometimes disappointing. However, in large samples, such as the Tirilazad cohort, even small disagreements between observed frequencies and predicted probabilities might result in a significant Hosmer-Lemeshow statistic.

Models predicting the need for specialized intensive care

The models predicting the risk of surgically removable lesions and the risk of raised ICP (ICP ≥ 20 mmHg) fulfilled only part of the proposed guidelines for model development and validation (Table 1). Although the study population was recent and relatively heterogeneous (as shown in Table 1 of Chapter 8), the population was also relatively small and originated from one center only.

Age, motor score, pupillary reactivity, hypotension, hypoxia, gender, cause of injury and injury severity score (ISS) were considered as candidate predictors. The first five were already discussed in the paragraph above. Gender, cause of injury and ISS complied with the guidelines; they were all plausible predictors, precisely defined and readily available.

The outcome measures were both easy to define and could be obtained with a low degree of intra- and inter-observer variation. Raised ICP, however, has as disadvantage that we may debate whether the used threshold value of 20 mmHg is appropriate.

Model development occurred with reliable statistical techniques. The predictors age and ISS were included as continuous linear characteristics. Values of missing predictors were statistically estimated with logistic regression, including the other predictors, and subsequently imputed[12,18].

The models were internally validated with bootstrapping techniques. External validation was also performed, but on a small and rather old dataset from the same center. To be relevant for clinical practice, both models needed to perform well, since both outcomes are indicators for the need for specialized intensive care. Unfortunately, the model predicting raised ICP performed poorly, both on the development and the validation population. Consequently, we decided to create no format to facilitate applicability of the model in clinical practice.

Practical considerations

This paragraph discusses strengths and limitations regarding generalizability of our models to predict six-month outcome, and the applicability in clinical practice.

Population

Patients with TBI constitute a heterogeneous population, including patients with widely varying severity of the injury, patients with different pathophysiology, e.g. focal versus diffuse injuries, and also patients in whom basic management may differ between centers. The models developed in this thesis are suitable for patients with severe or moderate closed TBI. Consequently, extrapolation of the models to patients with penetrating TBI is not advised, since this type of injury is pathofysiologically different from a closed TBI.

Furthermore, the cohorts used in this thesis originated from various centers all over the world. All centers are well-known trauma units with specialized intensive care and neurosurgical facilities and the management techniques in such units are generally of high professional level. Consequently, performance of the models in centers with less highly qualified personnel and instruments may be poorer. It is always advisable to validate the models before implementation in clinical practice, and especially so for such centers.

Outcome

Both mortality and unfavorable outcome were derived from the GOS. The GOS, however, has been criticized for being insensitive, especially in patients with more favorable outcomes, an insensitivity which was even further increased by the dichotomization applied in this thesis. We therefore realize that the dichotomized GOS may not catch relatively small differences in morbidity. Our models are especially suitable for providing an overall indication of outcome. More sensitive predictions may be provided with a model that predicts, for instance, the Barthel Index[19] or the Disability Rating Scale[20], both used to quantify the degree of functional outcome, or the EuroQoL, used for assessing health-related quality of life[21].

Application of the models in clinical practice

Before a prognostic model can be used in clinical practice, it needs to be validated to support generalizability. A minimum of 200 patients has been suggested for such validation studies[22]. If validity is confirmed the model can be safely applied, although monitoring is still needed.

At validation, however, also a structural difference in calibration may be observed. We found this for instance at validation of our models in the International Tirilazad patients. A structural difference in calibration implies a difference in the average predicted outcome between patient series, but similar odds ratios for the predictors in the development and validation patients. This is in agreement with the more general idea that biological associations between predictors and outcome do not change much over time or between centers. For example, it is plausible that age

will affect outcome after TBI similarly in patient series collected at different places or different time periods. If such a structural difference in calibration is observed, the model can simply be adjusted by re-estimating the model intercept[23,24]. This was done in Chapter 6 and with re-estimation of the model intercept; calibration improved in the International Tirilazad patients.

At validation, also a difference in the odds ratios for the predictors in the model and the odds ratios in the validation population may be observed. This was for instance found at validation of our model for mortality in the TCDB cohort. Such differences in odds ratios may, for instance, be caused by differences in definitions of the data in the development and validation patients, or by subgroup effects (e.g. a new treatment which is effective in part of the population) that influence the individual regression coefficient of a predictor. Then, more complex strategies may be needed, such as re-estimating the individual model parameters (e.g. fitting of the regression coefficients of a pre-specified set of predictors) and model revision (assessment of the relevance of predictors plus estimation of their parameters)[23,24]. Such strategies are, however, only recommended if the quality and size of the validation data are substantial. Furthermore, attention should be paid to the modeling guidelines as proposed in Chapter 2.

In some circumstances, however, validation of the models is not feasible. For instance, a small center may not be able to collect the required minimum of 200 patients needed for reliable validation. In that case, we should be more conservative in the application of the model[25]. For instance, if the aim of the model is to classify patients according to prognostic risk or to compare patient series and treatment results over time and place, application of the model may be valid. If the aim of the model is towards treatment limiting decisions, however, application may not be advised.

Application of the models in clinical research

The proposed seven-predictor models may serve various purposes in research, including case-mix adjustment in comparisons between centers, and over time within centers. An attractive role is in improving the efficiency of randomized clinical trials in TBI. A large-scale NIH sponsored project currently addresses this issue. The developed models can be used for covariate adjustment of the treatment effects in a RCT. This will increase the interpretability of the treatment effect (corrected for any imbalance at randomization, and conditional on predictive effects) as well as the statistical power of tests for the treatment effect. Hence, smaller sample sizes can be used for the same power as an analysis which ignores predictive effects[26-28]. Another application of the models is in an analysis of treatment effect according to a 'sliding dichotomy'[29]. Here, the dichotomization of the GOS is determined by prognosis, rather than defined the same point for all patients. The sliding dichotomy may also lead to better interpretation and to more statistical power for treatment effects in TBI patients.

Conclusions and recommendations

Conclusions and recommendations with regard to methods developing prognostic models

A sensible development and validation strategy for prognostic models in TBI should pay attention to:

- The size, heterogeneity, origin and time of data collection of the study population
- The plausibility, definition and availability of the predictors
- The definition, clinical relevance and observer variability of the predicted outcome measure
- The procedures used for model development and presentation. These include type of model, handling of missing predictor values, selection and coding of predictors, validity of performance measures, and applicability of the model presentation
- The procedures used for model validation. These include internal and external validation in new patients.

Conclusions and recommendations with regard to predicting six-month mortality and unfavorable outcome after severe or moderate TBI

- An older age is continuously associated with a worsening outcome after TBI. Hence, it is disadvantageous to define the effect of age on outcome in a discrete manner when aiming to estimate prognosis of adjust for confounding variables.
- To predict outcome after TBI the often-used Marshall CT-classification might be replaced by a combination of individual CT characteristics, including the status of basal cisterns, midline shift, traumatic subarachnoid or intraventricular hemorrhage, and the presence of different types of mass lesions.
- Possible improvements of the developed models may come from less generally accepted predictors. Among them are blood parameters, extracranial injuries and region of treatment, where the latter may be a proxy for one or more unknown underlying characteristics.
- The developed prognostic models can be used reliably to classify patients according to prognostic risk.
- The developed prognostic models can serve important roles in research, especially in increasing the efficiency of randomized clinical trials.
- When implementing the models in clinical practice, it is recommended to repeatedly assess the validity of predictions, and perform updating if needed.

Conclusions and recommendations with regard to the need of specialized intensive care after severe or moderate TBI

- It is difficult to accurately identify patients in need of specialized intensive care using baseline characteristics. Current guidelines, used for triage of TBI patients, are relatively specific in selecting patients in need of such care. On the other hand, these guidelines are less sensitive, and direct admission of more, and especially older, patients with severe or moderate TBI to level I trauma centers may be supported.

References

1. Chesnut R, Ghajar J, Maas AIR, et al. **Management and prognosis of severe traumatic brain** injury. Part 2: early indicators of prognosis in severe traumatic brain injury. **J Neurotrauma** 2000;17:557-627.
2. Bullock R, Chesnut R, Clifton G, et al. **Management and prognosis of severe traumatic brain** injury. Part 1: Guidleines for the management of severe traumatic brain injury. J Neurotrauma 2000;17:451-553.
3. Gibson RM, Stephenson GC. Aggressive management of severe closed head trauma: time for reappraisal. **Lancet 1989;2(8659):369-71.**
4. Andrews PJ, Sleeman DH, Statham PF, 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;97(2):326-36.
5. Sakellaropoulos GC, Nikiforidis GC. Development of a Bayesian Network for the prognosis of head injuries using graphical model selection techniques. Methods Inf Med 1999;38(1):37-42.
6. 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;66(1):20-5.
7. Choi SC, Barnes TY, Bullock R, Germanson TA, Marmarou A, Young HF. Temporal profile of outcomes in severe head injury. **J Neurosurg 1994;81(2):169-73.**
8. Maas AIR, Braakman R, Schouten HJ, Minderhoud JM, van Zomeren AH. **Agreement between** physicians on assessment of outcome following severe head injury. J Neurosurg 1983;58(3):321-5.
9. Wilson JT, Pettigrew LE, Teasdale GM. Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J Neurotrauma 1998;15(8):573-85.
10. Harrell FE, Jr. Regression modeling strategies: with applications to linear models, logistic regression and survival analysis. New York: Springer, 2001.
11. Steyerberg EW, Eijkemans MJC, Harrell FE, Jr., Habbema JDF. Prognostic modelling with logistic regression analysis: a comparison of selection and estimation methods in small data sets. Stat Med 2000;19(8):1059-79.
12. Little R. Regression with missing X's: a review. J Am Stat Assoc 1992;87:1227-1237.
13. Rubin D. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley and Sons, 1987.
14. Titterington DM, Murray GD, Murray LS, et al. Comparison of discrimination techniques applied to a compex data set of head injured patients. J Roy Stat Soc, Series A 1981;144:145-175.
15. Ennis M, Hinton G, Naylor D, Revow M, Tibshirani R. A comparison of statistical learning methods on the Gusto database. Stat Med 1998;17(21):2501-8.
16. Clermont G, Angus DC, DiRusso SM, Griffin M, Linde-Zwirble WT. Predicting hospital mortality for patients in the intensive care unit: a comparison of artificial neural networks with logistic regression models. Crit Care Med 2001;29(2):291-6.
17. Borque A, Sanz G, Allepuz C, Plaza L, Gil P, Rioja LA. **The use of neural networks and logistic** regression analysis for predicting pathological stage in men undergoing radical prostatectomy: a population based study. J Urol 2001;166(5):1672-8.

18. Harrell FE, Jr., Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15(4):361-87.
19. Collin C, Wade DT, Davies S, Horne V. The Barthel ADL Index: a reliability study. Int Disabil Stud 1988;10(2):61-3.
20. Rappaport M, Hall KM, Hopkins K, Belleza T, Cope DN. Disability rating scale for severe head trauma: coma to community. Arch Phys Med Rehabil 1982;63(3):118-23.
21. EuroQol--a new facility for the measurement of health-related quality of life. The EuroQol Group. Health Policy 1990;16(3):199-208.
22. Vergouwe Y, Steyerberg EW, Eijkemans MJC, Habbema JDF. Substantial effective sample sizes were required for external validation studies of predictive logistic regression models. J Clin Epidemiol 2005;58(5):475-83.
23. van Houwelingen HC. **Validation, calibration, revision and combination of prognostic survival** models. **Stat Med 2000;19(24):3401-15.**
24. Steyerberg EW, Borsboom GJ, van Houwelingen HC, Eijkemans MJC, Habbema JDF. **Validation** and updating of predictive logistic regression models: a study on sample size and shrinkage. Stat Med 2004;23(16):2567-86.
25. Justice AC, Covinsky KE, Berlin JA. Assessing the generalizability of prognostic information. Ann Intern Med 1999;130(6):515-24.
26. Steyerberg EW, Bossuyt PM, Lee KL. Clinical trials in acute myocardial infarction: should we adjust for baseline characteristics? Am Heart J 2000;139(5):745-51.
27. Hauck WW, Anderson S, Marcus SM. Should we adjust for covariates in nonlinear regression analyses of randomized trials? Control Clin Trials 1998;19(3):249-56.
28. Hernandez AV, Steyerberg EW, Habbema JDF. Covariate adjustment in randomized controlled trials with dichotomous outcomes increases statistical power and reduces sample size requirements. **J Clin Epidemiol 2004;57(5):454-60.**
29. Murray GD, Barer D, Choi SC, et al. **Design and analysis of phase III trials with ordered outcome** scales: the concept of the sliding dichotomy. J Neurotrauma 2005;22(5):511-7.

Summary

This thesis describes studies on prognosis after severe or moderate traumatic brain injury (TBI). In **Chapter 1**, the clinical problem of TBI is discussed. TBI is generally defined as an injury to the brain caused by an external physical force, such as a traffic accident, a fall or a gunshot. TBI is an important public health care problem in the western world. It is one of the most common causes of death in young adults and it can affect people's lives enormously.

The focus of this thesis is on developing and validating prognostic models: statistical models that combine individual patient characteristics to predict the probability of a particular outcome or disease state. The objectives of this thesis were: (1) to study methodological developments in prognostic modeling in TBI; (2) to develop and validate prognostic models that predict long-term outcome for patients with severe or moderate TBI an (3) to predict the need of specialized intensive care to aid a more efficient triage of patients.

Methodological developments in prognostic modeling in TBI

In **Chapter 2**, we systematically review 26 previously developed TBI models with the purpose to gain insight into methodological developments in prognostic modeling in TBI. We observed several shortcomings. For instance, many models were developed on old and relatively small patient series, originating from one single place or region. This makes the generalizability of these models to new patients or patients from another region questionable. Furthermore, rather crude statistical methods were used and the handling of missing values was often not reported or – if reported – patients with missing values were simply omitted from the study population. Before a prognostic model can reliably be applied in clinical practice, the performance of the model in new patients, e.g. more recent patients or patients from another region, needs to be studied ('external validition'). However, the 26 models were seldom validated on more recent patients from the same place, and never on patients from another place. In this chapter, we propose guidelines to develop and validate future prognostic models in TBI (the first research question). These guidelines include five subjects, i.e. study population, predictors (characteristics that predict the outcome), outcome, model development and model validation.

Prediction of long-term outcome for patients with severe or moderate TBI

Accurate prediction of long-term outcome at baseline (in our studies: within 4 hours after injury) is important for several purposes. It may support clinical decision-making and provide realistic and evidence-based expectations to relatives (counseling) and caregivers. Outcome predictions may also be applied to classify patients according to prognostic risk, which may be useful to compare outcome between patient series from different centers or to study treatment results over time. Furthermore, the design and analysis of randomized clinical trials (RCTs) may be improved; prognostic models may be used for defining enrollment criteria and for risk-stratification, such that covariate-adjusted treatment effects are estimated.

Chapter 3 till 7 of this thesis describe the development and validation of prognostic models that predict long-term outcome after a severe or moderate TBI. Long-term outcome is often evaluated with the 'Glasgow Outcome Scale' (GOS), a five level classification scale that assesses overall outcome. The five categories are: 'good recovery' (resumption to normal life, even though there may be deficits), 'moderate disability' (disabled, but independent), 'severe disability' (conscious, but disabled), 'vegetative state' (awake, but not aware) and 'death'. The GOS is often dichotomized into two groups: favorable and unfavorable outcome, with favorable outcome including 'good recovery' and 'moderate disability', and unfavorable outcome including the remaining three categories. Our models predict death and unfavorable outcome at six months after the injury. After six months, most patients are stable with respect to outcome category.

In **Chapter 3**, we describe the patient characteristics of our main study population: the Tirilazad cohort. This cohort consists of 2269 patients with severe and moderate TBI, who were included in the International and North American Tirilazad trials. Despite the strict enrolment criteria, used to select patients for the trials, the cohort was rather heterogeneous, containing both many severely and many less severely injured patients. The heterogeneity in patient characteristics was associated with region. Furthermore, we also observed regional differences in case management. The variation in patient characteristics and case management may partly be explained by regional differences, such as differences in demography and culture or variation in local policies regarding trauma. Remarkably, outcome was better in patients treated in the United States than those treated in Europe or other countries. This difference could not be explained by differences in demographical and clinical patient characteristics or differences in case management.

Next, we describe the association between important predictors and long-term outcome (mortality and unfavorable outcome). In **Chapter 4** we examine the optimal way to include age in a prognostic model and in **Chapter 5** we consider Computed Tomography (CT) characteristics.

In **Chapter 4**, we perform a meta-analysis to study the association between age and outcome (mortality and unfavorable outcome) and observed that an older age is continuously associated with a worsening outcome after TBI. Hence, it is disadvantageous to define the effect of age on outcome in a discrete manner (using a cut-off point) when we aim to estimate prognosis or adjust for confounding variables. The association between age and both mortality and unfavorable outcome could be described adequately by a linear trem or – statistically even better – by a linear and a squared term. These descriptions fit the association well and are also simple to apply in clinical practice.

In **Chapter 5** we study the association between different combinations of CT characteristics and mortality. First, we examined the Marshall CT-classification – which groups patients with TBI according to multiple CT characteristics, i.e. presence of intracranial abnormalities, presence of mass lesions, CT signs of raised intracranial pressure (status of basal cisterns, midline shift) and planned evacuation of mass lesions –, since this classification is often used for prognostic purposes. Although the Marshall CT-classification performed reasonably, performance could be improved by rearranging the underlying individual CT characteristics of the Marshall CT-classification. Best predictions may be obtained by replacing the Marshall CT-classification by an alternative set of CT characteristics, including the states of basal cisterns, midline shift, traumatic subarachnoid or intraventricular hemorrhage, and the presence of different types of mass lesions.

In **Chapters 6 and 7**, we address research objective 2. **We assess the performance of prognostic** models with regard to discrimination and calibration. Discrimination is the ability to distinguish a patient with a good outcome from a patient with a poor outcome. Perfect discrimination is reached if patients who die have predicted probabilities of mortality close to 100%, while patients who survive have predictions close to 0%. To quantify the discrimination we calculated the area under the receiver operating characteristic curve (AUC). A model with an AUC of 0.50 has no discriminative power, while an AUC of 1.0 reflects perfect discrimination. Calibration refers to the agreement between the observed outcome frequencies (e.g. the proportion of TBI patients who die) in the data and the predicted probabilities that patients have the outcome (e.g. the predicted probability that TBI patients die) of the model. For instance, if a group of patients are predicted to have a 10% chance of mortality, then approximately 10 out of 100 patients should actually die.

In **Chapter 6**, two prognostic models that predict mortality and unfavorable outcome are developed. Both models included the same seven predictors: i.e. age, motor score, pupillary reactivity, hypotension, hypoxia, the Marshall CT-classification and the presence of a traumatic subarachnoid haemorrhage. The models provided high discrimination between patients with good and poor six-month outcome (AUC=0.78 for the model predicting mortality and AUC=0.80 for the model predicting unfavorable outcome). Calibration of both models was satisfactory.
Before the models can reliably be used in clinical practice, they have to be validated in other clinical settings. Such external validation studies were performed in **Chapter 6 and Chapter 7**, using three multi-center cohorts; the EBIC (n=796), TCDB (n=746) and Selfotel cohort (n=427). Discrimination of the models was satisfactory in the Selfotel cohort (AUC=0.74 for both models) and even better in the other cohorts (AUC 0.83 to 0.89). Calibration was satisfactory in all cohorts, except in the TCDB cohort: in this historic cohort predicted probabilities were lower than observed frequencies. The poor calibration may be explained by the improvement of treatment standards, including trauma organization and critical care management, since the TCDB data collection (1984-1988). Overall, we consider the developed seven-**predictor models suitable** for clinical practice. Caution is however required, because recalibration may be necessary. To facilitate implementation of the models in clinical practice, we developed score charts. With these charts the physician can calculate the risk of a poor long-term outcome using baseline characteristics.

In **Chapter 7**, we also assess the performance of four other models that use baseline clinical and CT characteristics to predict long-term mortality or unfavorable outcome after severe or moderate TBI. All models were externally validated on the EBIC, TCDB and Selfotel cohorts. The seven-predictor models had the highest discriminative abilities and best calibration. In a secondary analysis, we refitted the other models on the Tirilazad population, updating the regression coefficients (weights) of the predictors to best fit this population. This was done to better assess the 'pure' predictive value of the different combinations of predictors, excluding the influences of differences in study population or statistical methodology. External validation of the refitted models on the EBIC, TCDB and Selfotel cohort showed that performance was better for logistic regression models than for classification trees, and for models with more predictors than models with only a few predictors.

Need of specialized intensive care

Nowadays a high proportion of the severe and moderate TBI patients are first transported to a general hospital and later to a level I trauma center (secondary referral). Secondary referral, however, may delay the institution of appropriate therapy and increase the risk of adverse events and systemic insults during inter-hospital transport.

In **Chapter 8**, we study whether a more efficient triage may be aided by early identification of patients in need of specialized intensive care (research objective 3). These patients have a high risk of developing surgically removable lesions or raised intracranial pressure (ICP). In a prospective cohort of 275 patients admitted to the neurosurgical unit of the ErasmusMC, we developed two prognostic models; one predicting the risk of surgically removable lesions and one predicting the risk of raised ICP, using baseline characteristics. We observed, however, that the models could only reasonably discriminate patients with and without surgically removable lesions (AUC=0.78 at development and AUC=0.67 at external validation in a historic sample of 205 patients from the Erasmus MC), and could not discriminate patients with and without raised ICP. Therefore, these models are not considered suitable for implementation in clinical practice.
In this chapter we also compare patients primarily (73%) and secondarily referred (27%) to the neurosurgical unit. We observe that patients secondarily referred were older, more frequently injured in the domestic setting or a fall, had initially less severe clinical characteristics and a high incidence of both outcomes (66% surgically removable lesions and 70% raised ICP, if monitored). The high proportion of secondarily referred patients and the high incidence of both outcomes in this group of patients may support direct admission of more, and especially older, patients with severe or moderate TBI to specialized neurosurgical units.

This thesis ends with a general discussion of the findings of the presented studies (**Chapter 9**). Conclusions and recommendations are presented. With respect to predicting long-term outcome, we conclude that the developed models can be used to classify TBI patients according to prognostic risk. The models can serve important roles in research, especially in increasing the efficiency of randomized clinical trials. When implementing the models in clinical practice, it is recommended to repeatedly assess the validity of predictors, and perform updating if needed. Furthermore, we conclude that it is difficult to accurately identify patients in need of specialized intensive care. Current criteria, used for triage of TBI patients, are relatively specific but also less sensitive in selecting patients in need of such care. The latter may support direct admission of more patients with severe or moderate TBI to specialized neurosurgical units. For future model development and validation, we recommend the guidelines presented in **Chapter 2**, such that more accurate and evidence-based prognostic estimates become available for physicians treating individual patients with TBI.

Samenvatting

Dit proefschrift beschrijft een aantal studies op het gebied van prognose na matig ernstig of ernstig traumatisch hersenletsel (THL). In **hoofdstuk 1** wordt het klinische probleem van traumatisch hersenletsel besproken. Traumatisch hersenletsel wordt gedefinieerd als elk hersenletsel dat is ontstaan door een oorzaak van buitenaf, zoals een ongeval, een val of een schotwond. THL vormt een belangrijk volksgezondheidsprobleem in de Westerse wereld; het is één van de meest voorkomende doodsoorzaken bij jong volwassenen en het kan het leven en het functioneren van jonge mensen enorm beïnvloeden.

De nadruk van dit proefschrift ligt op de ontwikkeling en validatie van prognostische modellen; statistische modellen waarin individuele patiëntkenmerken worden gecombineerd om de kans op een bepaalde uitkomst of ziekte status te kunnen voorspellen. De doelstellingen betroffen: (1) het beschrijven van methodologische ontwikkelingen ten aanzien van eerder ontwikkelde prognostische modellen voor THL patiënten; (2) de ontwikkeling en validatie van nieuwe prognostische modellen die de lange termijn gevolgen voorspellen voor patiënten met matig ernstig of ernstig traumatisch hersenletsel en (3) het voorspellen van de behoefte van een THL patiënt aan behandeling in een gespecialiseerd traumacentrum om zo de triage criteria (al dan niet transporteren naar een gespecialiseerd trauma centrum) te kunnen verbeteren.

Methodologische ontwikkelingen ten aanzien van eerder ontwikkelde prognostische modellen

In **hoofdstuk 2** wordt een overzicht (systematic review) gegeven van eerder gebruikte methodes voor de ontwikkeling en validatie van prognostische modellen voor THL patiënten. Hiertoe is de methodologie van 26 eerder ontwikkelde modellen systematisch beoordeeld. Wij constateerden verschillende methodologische beperkingen. Zo zijn veel modellen ontwikkeld op gegevens van een oud en relatief klein cohort met patiënten afkomstig uit één enkele plaats of gebied. Hierdoor is de validiteit van deze modellen voor nieuwe patiënten of voor patiënten uit een ander gebied twijfelachtig. Verder werd meestal niet aangegeven hoe is omgegaan met missende waardes of werden patiënten met missende waardes simpelweg verwijderd uit het cohort. Voordat een prognostisch model gebruikt kan worden in de klinische praktijk, dient de prestatie van het model te worden bestudeerd in nieuwe patiënten (externe validatie), bijvoorbeeld recentere patiënten of patiënten uit een ander gebied. Echter, uit de review bleek dat de eerder ontwikkelde modellen slechts zelden zijn gevalideerd op recentere patiënten uit hetzelfde gebied en nooit op patiënten uit een ander gebied. In **hoofdstuk 2** worden richtlijnen voorgesteld ten aanzien van de ontwikkeling en validatie van toekomstige prognostische modellen. Deze richtlijnen hebben betrekking op vijf onderwerpen, te weten: ontwikkelpopulatie, predictoren (patiëntkarakteristieken die de uitkomst voorspellen), voorspelde uitkomst, model ontwikkeling en model validatie.

Het voorspellen van de lange termijn gevolgen

Het accuraat kunnen voorspellen van de lange termijn uitkomst is van belang voor verschillende toepassingen, waaronder het informeren van familieleden en zorgverleners en het classificeren van patiënten op basis van hun prognostisch risico. Een dergelijke classificatie is bruikbaar

voor verschillende doeleinden, zoals het vergelijken van de uitkomst tussen verscheidene patiëntenseries of voor het bestuderen van behandelingsresultaten door de tijd heen, waarbij wordt gestratificeerd voor prognostisch risico. Daarnaast kan de opzet en analyse van gerandomiseerde klinische trials worden verbeterd; prognostische modellen kunnen worden gebruikt voor de verfijning van in- en exclusiecriteria, voor risico-stratificatie, en voor covariaat correctie in de statistische analyse.

De **hoofdstukken 3 tot en met 7** van dit proefschrift beschrijven de ontwikkeling en validatie van modellen voor het voorspellen van de lange termijn gevolgen na een traumatisch hersenletsel. Vaak worden de lange termijn gevolgen geëvalueerd met behulp van de 'Glasgow Outcome Scale' (GOS). Deze schaal kent vijf categorieën: 'goed herstel' (terugkeer naar het gewone leven, ondanks eventuele tekorten), 'matige handicap' (gehandicapt, maar onafhankelijk), 'ernstige handicap' (bewust, maar gehandicapt en afhankelijk), 'vegetatieve status' (geen contact met de omgeving) en 'overlijden'. De GOS wordt vaak gedichotomiseerd, waarbij de eerste twee categorieën worden samengevoegd tot 'gunstige uitkomst' en de laatste drie tot 'ongunstige uitkomst'. De modellen die door ons zijn ontwikkeld, voorspellen het overlijden en de 'ongunstige uitkomst' na zes maanden herstel. Na deze herstelperiode is de toestand van de meeste THL patiënten gestabiliseerd.

In **hoofdstuk 3** worden de patiëntkarakteristieken van ons belangrijkste studiepopulatie, het Tirilazad cohort, beschreven. Dit cohort bestaat uit 2269 patiënten met matig tot ernstig THL, geselecteerd voor de Noord-Amerikaanse en Internationale Tirilazad trials. De gebruikte selectiecriteria en het voorgeschreven behandelprotocol waren relatief streng. Desondanks bestond het cohort uit een relatief heterogene patiëntenpopulatie, met zowel zeer ernstig als minder ernstige patiënten. Ook werden de patiënten aanzienlijk verschillend behandeld. De verschillen in patiëntkarakteristieken en behandeling konden deels worden verklaard door regionale verschillen in demografische opbouw, cultuur en patiëntenmanagement. Opvallend is dat we ook een verschil in de uitkomst vonden: patiënten die werden behandeld in de Verenigde Staten hadden een betere uitkomst dan de patiënten behandeld in Europa. Dit uitkomstverschil kon niet worden verklaard door verschillen in demografische of klinische patiëntkarakteristieken en ook niet door verschillen in behandeling.

In **hoofdstuk 4 en 5** wordt de associatie tussen enkele belangrijke predictoren en de lange termijn gevolgen na traumatisch hersenletsel bestudeerd. De predictoren betreffen leeftijd (**hoofdstuk 4**) en afwijkingen op de CT-scan (**hoofdstuk 5**).

In **hoofdstuk 4** beschouwen we verschillende studies (meta-analyse), en vinden we een continue associatie tussen de leeftijd en de 6-maands uitkomst (mortaliteit en 'ongunstige uitkomst'): hoe ouder de patiënt, hoe slechter de uitkomst. Door leeftijd als een continue lineaire term of als een continue lineaire plus een kwadratische term op te nemen, wordt de associatie op een optimale manier beschreven. Optimaal houdt in dat de geobserveerde data goed worden benaderd, maar dat de termen tevens makkelijk toepasbaar zijn in de klinische praktijk.

Hoofdstuk 5 heeft als doel het bestuderen van de associatie tussen verschillende combinaties van afwijkingen op de CT-scan en de mortaliteit. Allereerst onderzochten we de Marshall CT-classificatie – een classificatie die patiënten met hersenletsel groepeert aan de hand van verschillende CT afwijkingen, zoals de aanwezigheid van grote hematomen (bloedingen > 25 cc) – aangezien deze classificatie vaker wordt gebruikt voor prognostische doeleinden. Alhoewel de Marshall CT-classificatie een redelijk goed onderscheid maakte tussen patiënten die al dan niet overleden, kon de discriminatie worden verbeterd door de individuele CT-afwijkingen uit de Marshall CT-classificatie een andere weging te geven. De mortaliteit van patiënten met matig tot ernstig THL werd het best voorspeld door een combinatie van de volgende afwijkingen op de CT-scan: compressie van de basale cisternen, aanwezigheid van een 'midline shift', aanwezigheid van een subarachnoïdale en/of intraventriculaire bloedingen en verschillende typen grote hematomen.

Hoofdstuk 6 en 7 behandelt doelstelling 2. Vanaf **hoofdstuk 6** worden prognostische modellen ontwikkeld en gevalideerd. De kwaliteit van deze modellen bepalen we aan de hand van het discriminerend vermogen en de calibratie. Discriminatie heeft betrekking op het vermogen van het model om een patiënt met een goede uitkomst (bijvoorbeeld een patiënt die overleeft) te onderscheiden van een patiënt met een slechte uitkomst (een patiënt die overlijdt). In het ideale geval liggen de voorspelde kansen voor sterfte voor patiënten die daadwerkelijk overlijden dicht bij de 100% en voor patiënten die overleven dicht bij de 0%. Het discriminerend vermogen wordt over het algemeen gekwantificeerd met de 'area under the receiver operating charateristics curve', de AUC. Een model met een AUC van 0.50 heeft geen discriminerend vermogen, terwijl een AUC van 1.00 perfect discrimineert. Calibratie vergelijkt de geobserveerde uitkomst frequentie met de gemiddelde voorspelde kans op die uitkomst per categorie voorspelde kansen. Bijvoorbeeld, van een groep patiënten met gemiddeld voorspelde kans van 10% op overlijden zou – bij een model met goede calibratie – ongeveer 10 van de 100 patiënten daadwerkelijk moeten overlijden.

In **hoofdstuk 6** worden twee prognostische modellen ontwikkeld: het eerste model voorspelt de mortaliteit en het tweede de 'ongunstige uitkomst' na zes maanden herstel. Beide modellen gebruiken dezelfde zeven predictoren, te weten: leeftijd, motor score, pupilreactiviteit, hypotensie, de Marshall CT-classificatie en de aanwezigheid van een traumatische subarachnoïdale bloeding. De predictoren zijn allen voor of vlak na binnenkomst in het traumacentrum bepaald. Beide modellen presteren goed: AUC=0.78 voor het model dat de mortaliteit voorspelt en AUC=0.80 voor het model dat de ongunstige uitkomst voorspelt. De calibratie is voldoende. Hoofdstuk 6 en 7 beschrijft de externe validatie van de ontwikkelde modellen in drie cohorten: het EBIC (n=796), TCDB (n=746) en Selfotel cohort (n=427). Het discriminerend vermogen van de modellen was voldoende in het Selfotel cohort (AUC=0.74 voor beide modellen) en goed in de andere cohorten (AUC=0.83 tot 0.89). De calibratie was voldoende in alle cohorten, behalve in het TCDB cohort: het voorspelde risico op een slechte uitkomst was voor deze patiënten lager dan het werkelijke risico. Dit laatste zou verklaard kunnen worden door de periode van dataverzameling voor het TCDB cohort, namelijk tussen 1984 en 1988. Sinds die tijd is de behandeling van THL patiënten verbeterd, hetgeen het risico van een slechte uitkomst verlaagt.
Concluderend beschouwen we de ontwikkelde zeven-predictor modellen als toepasbaar in de klinische praktijk. Om deze toepasbaarheid te vergemakkelijken zijn scorekaarten ontwikkeld,

waarmee de arts het risico op een slechte uitkomst kan berekenen. Voorzichtigheid is echter geboden, aangezien hercalibreren van de modellen nodig kan zijn voor een nieuwe populatie.

In **hoofdstuk 7** vergeleken we tevens de prestatie van onze zeven-predictor modellen met dat van vier andere eerder ontwikkelde modellen (logistische modellen en beslisbomen) die ook, met behulp van patiëntkarakteristieken die voor of vlak na aankomst in het traumacentrum zijn bepaald, de lange termijn gevolgen na een matig tot ernstig traumatisch hersenletsel voorspellen. Alle modellen werden extern gevalideerd in het EBIC, TCDB en Selfotel cohort. De zeven-predictor modellen presteerden het best, zowel wat betreft de discriminatie als de calibratie. In een tweede analyse werden de vier eerder ontwikkelde modellen opnieuw geschat op het Tirilazad cohort. Bij deze nieuwe schatting werden aan de predictoren uit het desbetreffende model andere wegingsfactoren toegekend op basis van het Tirilazad cohort. Het doel hiervan was het bepalen van de 'werkelijke' voorspellende waarde van de verschillende combinaties van predictoren, zonder verstorende invloed van verschillen in ontwikkelpopulatie of de gebruikte statistische methoden. Externe validatie van de opnieuw geschatte modellen op het EBIC, TCDB en Selfotel cohort liet zien dat logistische regressiemodellen beter presteerden dan classificatiebomen ('trees'). Modellen met meer predictoren presteerden eveneens beter dan modellen met slechts enkele predictoren.

Behoefte aan behandeling in een gespecialiseerd traumacentrum

Tegenwoordig worden relatief veel patiënten met een matig tot ernstig hersenletsel eerst naar een perifeer ziekenhuis getransporteerd en pas later naar een neurochirurgisch traumacentrum (secundaire verwijzing). Secundaire verwijzing kan echter de start van een adequate behandeling vertragen en het risico op bijwerkingen en systemische letsels tijdens het transport tussen ziekenhuizen verhogen.

In **hoofdstuk 8** wordt bestudeerd of de triage efficiënter kan worden door een vroegtijdige identificatie van patiënten die behoefte hebben aan behandeling in een neurochirurgisch traumacentrum (doelstelling 3 van dit proefschrift). Deze patiënten hebben een hoog risico op de ontwikkeling van een operabele bloeding of een verhoogde intracraniële druk (ICP). De onderzoekspopulatie bestond uit 275 THL patiënten die zijn opgenomen in het gespecialiseerde traumacentrum van het ErasmusMC. Patiëntgegevens werden prospectief verzameld. We ontwikkelden twee prognostische modellen: het eerste model voorspelt het risico op operabele bloedingen en het tweede model het risico op verhoogde intracraniële druk. Het eerste model kon patiënten met en zonder operabele bloeding redelijk goed onderscheiden (AUC = 0. 78 en AUC = 0.67 tijdens externe validatie op een cohort van 205 THL patiënten die lang geleden in het traumacentrum van het ErasmusMC zijn opgenomen). Het tweede model kon echter geen onderscheid maken tussen patiënten met en zonder verhoogde intracraniële druk. Om de triage te kunnen verbeteren, dienen beide modellen goed te presteren. Aangezien dit slechts voor één model niet het geval is, vinden we de modellen niet geschikt voor implementatie in de klinische praktijk.

In **hoofdstuk 8** vergelijken we tevens de karakteristieken van de direct (73%) en secundair (27%) verwezen patiënten. De secundair verwezen patiënten zijn in het algemeen iets ouder, vaker gewond geraakt in een huiselijke omgeving of door een val en hebben bij aanvang iets minder ernstig letsel. De incidentie van beide uitkomstmaten is echter hoog: 66% heeft een operabele bloeding en 70% een verhoogde intracraniële druk (indien druk bepaald). De hoge proportie van secundair verwezen patiënten en de hoge incidentie van beide uitkomsten in deze groep patiënten zou directe verwijzing van meer, en speciaal oudere, patiënten met matig tot ernstig hersenletsel naar een neurochirurgisch traumacentrum kunnen ondersteunen.

Dit proefschrift eindigt met een algemene discussie van de resultaten en het geven van conclusies en aanbevelingen voor toekomstig onderzoek (**hoofdstuk 9**). Met betrekking tot het voorspellen van de lange termijn gevolgen wordt geconcludeerd dat de ontwikkelde zeven-predictor modellen kunnen worden gebruikt voor het classificeren van THL patiënten op basis van hun prognostisch risico. De modellen kunnen bijvoorbeeld worden toegepast in wetenschappelijk onderzoek om de opzet en analyse van gerandomiseerde klinische trials te verbeteren. Bij implementatie van de modellen in de klinische praktijk wordt geadviseerd om regelmatig de validiteit te testen, en – indien nodig – de modellen aan te passen. Het voorspellen van de behoefte aan behandeling in een gespecialiseerd traumacentrum blijkt moeilijk te zijn. De huidige criteria voor de triage van THL patiënten zijn relatief specifiek. Ze zijn echter minder sensitief, hetgeen de directe verwijzing van meer THL patiënten zou kunnen ondersteunen. Voor toekomstig onderzoek ten aanzien van de ontwikkeling en validatie van prognostische modellen, wordt aanbevolen om de richtlijn uit **hoofdstuk 2** te volgen, zodat meer nauwkeurige, door degelijk empirisch onderzoek ondersteunde, prognostische schattingen beschikbaar komen voor artsen en hun patiënten met THL .

Dankwoord

Het zit er bijna op. Jaren van data ordenen en analyseren, nieuwe technieken leren, patiëntgegevens verzamelen, overleg, theepauzes, bij 30∞C in fleecejack op het werk zitten vanwege de goed werkende airco en schrijven, schrijven en nog meer schrijven. Uiteindelijk heeft al dat werk geresulteerd in dit 'boekje'. Dit proefschrift had nooit tot stand kunnen komen zonder de bijdrage van een groot aantal mensen die hebben meegedacht, uitgelegd en geadviseerd. Aan iedereen die heeft meegeholpen: bedankt!

Een aantal mensen wil ik specifiek noemen. In de eerste plaats wil ik mijn co-promotoren, Ewout Steyerberg en Andrew Maas, en mijn promotor, Dik Habbema, bedanken. Ewout, jij was als dagelijks begeleider mijn belangrijkste aanspreekpunt en vraagbaak. Bedankt voor je scherpe inzichten en je kritische maar altijd waardevolle commentaar. Ik heb ongelooflijk veel van je geleerd! Andrew, je creativiteit en enthousiasme voor de wetenschap zijn groot. Dank voor het geduld dat je hebt gehad met deze niet-clinicus die telkens wilde weten hoe het ook alweer zat met die CT-scan en die Glasgow Coma Scale. Dik, ook al had je de rol van 'persoon-op-de-achtergrond', toch heb jij op een aantal momenten een belangrijke rol gespeeld. Bedankt voor je vriendelijke en doortastende begeleiding.

Naast Ewout, Andrew en Dik wil ik ook de collega's van MGZ bedanken voor de bereidheid om met mij mee te denken over allerlei inhoudelijke, methodologische en praktische zaken, maar vooral ook voor de leuke contacten. In het bijzonder denk ik aan mijn kamergenoten Claudine Hunault, Yvonne Vergouwe en Laetitia Veerbeek. Ik kijk met veel plezier terug! Bij René Eijkemans, Caspar Looman en Gerard Bosboom kon ik altijd binnenlopen voor een doortimmerd statistisch advies. Ook bij de afdeling automatisering en het secretariaat van MGZ en de afdeling neurochirurgie kon ik altijd terecht. Dank jullie wel!

Anneke Rampen heeft als afstudeerstudent meegewerkt aan het onderzoek. Anneke, dank voor je bijdrage. De medewerkers van het neurochirurgische traumacentrum van het Erasmus MC, en Tineke Landman in het bijzonder, wil ik bedanken voor hun inspanningen bij de dataverzameling.

I am grateful to my co-authors for the time and effort they have put in reading the manuscripts, and for their useful suggestions. Furthermore, I would like to thank Rodrigo Labouriau for the stimulating discussions. Rodrigo, it was nice to see that a Brazilian man living in Denmark liked a typical Dutch 'stoofpotje' with gingerbread so much!
Prof.dr. C.J.J. Avezaat heeft het hele manuscript van constructief en nuttig commentaar voorzien. Dank hiervoor!

Het CBO, waaronder Kitty Rosenbrand en Teus van Barneveld, wil ik bedanken voor de mogelijkheid die ze me gaven om de laatste loodjes van dit proefschrift sneller af te ronden. Jan Wille heeft met behulp van een teiltje water, een pipet en vooral heel veel geduld de foto op de voorkant van dit proefschrift gemaakt en Albert Epping heeft met veel aandacht de lay-out van dit proefschrift verzorgd. Dank jullie wel!

Tot slot enkele woorden voor mijn (schoon)familie en vrienden, waaronder mijn paranimfen Dirk en Liesbeth: dank jullie wel voor de belangstelling en voor al die leuke momenten. Papa, dit is een mooie gelegenheid om je te bedanken voor je onvoorwaardelijke steun op alle gebieden. Helaas kan mama mijn promotie niet meer meemaken, maar in gedachten zal zij er ook bij zijn. Carl, bedankt voor al het goeds dat we samen hebben. Dat er nog maar veel moois mag volgen!

Curriculum vitae

Chantal Hukkelhoven werd geboren op 29 mei 1971 te Sittard. Zij behaalde in 1989 haar VWO-diploma aan het Bisschoppelijk College Dr. Edith Stein te Echt. Aansluitend begon zij de studie Voeding van de Mens (tegenwoordig: Voeding en Gezondheid) aan Wageningen Universiteit. In 1995 rondde zij deze studie af met doctoraalonderzoeken in de voeding en epidemiologie en met het volgen van het pre-doctoraal gedeelte onderwijskunde. Vervolgens werkte zij enkele jaren als statistisch analist bij Organon te Oss. In oktober 1998 kwam zij als wetenschappelijk onderzoeker in dienst bij het Centrum voor Klinische Besliskunde, afdeling Maatschappelijke GezondheidsZorg van de Erasmus Universiteit Rotterdam (tegenwoordig: Erasmus MC, Universitair Medisch Centrum Rotterdam). Tijdens deze periode was zij betrokken bij verschillende projecten van de afdeling Neurochirurgie en voerde zij de onderzoeken beschreven in dit proefschrift uit. Sinds 2003 werkt zij als adviseur richtlijnen/indicatoren bij het Kwaliteitsinstituut voor de Gezondheidszorg CBO.