<https://helda.helsinki.fi>

Validation of a decision support model for investigation and improvement in stroke thrombolysis

Keshtkaran, Mahsa

2016-08-16

Keshtkaran , M , Churilov , L , Hearne , J , Abbasi , B & Meretoja , A 2016 , ' Validation of a decision support model for investigation and improvement in stroke thrombolysis ' , European Journal of Operational Research , vol. 253 , no. 1 , pp. 154-169 . <https://doi.org/10.1016/j.ejor.2016.02.035>

<http://hdl.handle.net/10138/223994>

<https://doi.org/10.1016/j.ejor.2016.02.035>

publishedVersion

Downloaded from Helda, University of Helsinki institutional repository.

This is an electronic reprint of the original article.

This reprint may differ from the original in pagination and typographic detail.

Please cite the original version.

Decision Support

Validation of a decision support model for investigation and improvement in stroke thrombolysis

Mahsa Keshtkaran^{a,b,*}, Leonid Churilov^{b,a}, John Hearne^a, Babak Abbasi^c, Atte Meretoja^{d,e}^a School of Mathematical and Geospatial Sciences, RMIT University, Melbourne, Australia^b The Florey Institute of Neuroscience and Mental Health, University of Melbourne, Melbourne, Australia^c School of Business IT and Logistics, RMIT University, Melbourne, Australia^d Department of Medicine, The Royal Melbourne Hospital, University of Melbourne, Melbourne, Australia.^e Head and Neck Center, Helsinki University Central Hospital, Helsinki, Finland

ARTICLE INFO

Article history:

Received 20 October 2014

Accepted 25 February 2016

Available online 2 March 2016

Keywords:

OR in health services

OR/MS model validation

ABSTRACT

Validation of Operations Research/Management Science (OR/MS) decision support models is usually performed with the aim of providing the decision makers with sufficient confidence to utilize the model's recommendations to support their decision-making. OR/MS models for investigation and improvement provide a particularly challenging task as far as validation is concerned. Their complex nature often relies on a wide variety of data sources and empirical estimates used as parameters in the model, as well as on multiple conceptual and computational modeling techniques. In this paper, we performed an extensive literature review of validation techniques for healthcare models in OR/MS. Despite calls for systematic approaches to validation of complex OR/MS decision support models; we identified a clear lack of appropriate application of validation techniques reported in published healthcare models. The "Save a minute – save a day" model for evaluation of long-term benefits of faster access to thrombolysis therapy in acute ischemic stroke is used as a case to demonstrate how multiple aspects of data validity, conceptual model validity, computerized verification, and operational model validity can be systematically addressed when developing a complex OR/MS decision support model for investigation and improvement in health services.

© 2016 Elsevier B.V. All rights reserved.

1. Introduction

Operations Research historically developed as a scientific approach to analyzing problems and making decisions through the use of OR/MS models. It is distinct from other disciplines by applying scientific principles in the context of practical management decision making. An important aspect of scientific approach is *validation*. Schlesinger et al. (1979), p. 3) define validation as the "substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model". It has become a focus of sustained research efforts in the OR/MS community, with a number of research papers providing guidelines to assist with the process of validation of Decision Support (DS) models (Balci, 1989, 1994; Gass, 1977, 1983, 1993; Oral & Kettani, 1993; Sargent, 1984, 2013; Schellenberger, 1974). For such models, validation is performed with the aim of providing the decision makers with sufficient confidence

to utilize the model's recommendations as a support for their decision-making (Quade, 1980).

Based on the above broad understanding of validation, the concept of an OR/MS model validation is tightly coupled with the concept of an *intended use of the model*, and the intended use could vary significantly depending on the decision-making requirements. The potential scope of the intended use for the OR/MS models is substantial and there is no single validation methodology suitable for different kinds of DS models (Balci, 1994; Gass, 1977, 1983; Miller, Butler, & Bramall, 1976; Oberkampf & Roy, 2010; Oral & Kettani, 1993; Pidd, 2010; Sargent, 2013; Schellenberger, 1974; Whittner & Balci, 1989).

Modeling for *investigation and improvement*, where models are supporting "investigations that are relatively unique, which may involve system design, system improvement or just an attempt to gain understanding of a very complex situation" (Pidd, 2010, p. 18), provides a particularly challenging environment for model validation. This type of modeling could be employed to develop a DS model for a new real-world system which may not even physically exist or may be in the process of being designed. In these cases there is very limited amount of empirical data available on the

* Corresponding author at: Melbourne Brain Centre, 245 Burgundy Street, Heidelberg, Victoria 3084, Australia. Tel.: +61 3 90357048.

E-mail address: mahsa.keshtkaran@unimelb.edu.au (M. Keshtkaran).

model behavior, or even none at all. Due to the inability to conduct an empirical “output-based” validation, validation for this type of model is specifically focused on critically testing all the model inputs, assumptions, and parameters to provide a solid understanding of the limitations for the model’s use for the intended purpose (Pidd, 2010). In addition, validation of DS models for investigation and improvement in health systems and services is particularly complicated. This is due to the wide variety of often disparate data sources and published empirical estimates that originate from the clinical literature for the important model input parameters. Furthermore there is a reliance on multiple conceptual and computational modeling techniques used at the model-building phase.

Despite the broad consensus about the general importance of DS model validation presented in the OR/MS literature (Balci, 1994; Forrester & Senge, 1978; Lane, 1995; Sargent, 2013), the relevant literature where specific cases of an in-depth validation of DS models for understanding and improvement are discussed, in particular, in the context of health systems and services applications, is rare. Thus, there is a *research gap* in the reported knowledge about *practical aspects of how to validate a DS model for investigation and improvement in the context of health systems and services operations*.

The objective of this paper is to demonstrate how an in-depth validation of a complex DS model for investigation and improvement in the context of health systems and services operations can be conducted. This is achieved by demonstrating how multiple aspects of *data validity*, *conceptual model validity*, *computational verification*, and *operational validity* were systematically investigated when building and testing a “Save a minute – save a day” DS model for investigation and improvement. This model presents a world first study that explicitly quantifies the long-term patient benefits of earlier access to stroke thrombolysis. The non-technical overview of the model was presented for a clinical audience in the flagship journal *Stroke* in early 2014 (Meretoja et al., 2014) and has attracted considerable interest and publicity over the short period since its publication.

To the best of our knowledge, this contribution is novel as there are no publications in OR/MS literature specifically targeting the identified research gap. The examples of the comprehensive discussions of the model validity in the OR/MS literature are mostly confined to the domain of simulation modeling (Altinel & Ulaş, 1996; Dittus, Klein, DeBrota, Dame, & Fitzgerald, 1996; Repede & Bernardo, 1994), while other types of DS models, in particular, those relying on multiple modeling techniques, disparate data sources, diverse model inputs, remain largely unexplored.

The discussion in this paper is organized as follows: Section 2 is dedicated to the review of the general approaches to model validation proposed in the OR/MS literature as well as to a systematic review of specific examples of DS model validation in Healthcare OR/MS literature. In Section 3, the “Save a minute – save a day” DS model used to investigate the long-term benefits of early access to acute stroke thrombolysis is discussed; Section 4 provides an in-depth discussion of the validation process of the DS model presented in Section 3; conclusions are presented in Section 5.

2. Validation of DS models: the evidence from the OR/MS literature

2.1. General approaches to model validation

Schlesinger et al. (1979), p. 3) define validation as the “substantiation that a model within its domain of applicability possesses a satisfactory range of accuracy consistent with the intended application of the model”. In models for investigation and improvement, in particular, the accuracy needed is usually obtained by careful examination of the model parameters and assumptions used to

build the conceptual model, since often there is no data from the real-world system to be compared with that of the model output.

A number of notions discussed in the OR/MS literature, such as *validation*, *verification*, *acceptability*, and *credibility* are mainly used to address the confidence of model developers/users in using a DS model for its proposed application (Pidd, 2003). The links between these concepts are strong and differences are often subtle: for example, the terms *validation* and *verification* are frequently used together, where *verification* is understood as the process of ensuring that a computerized model has implemented accurately to represent a conceptual model (Sargent, 2013; Schlesinger et al., 1979). *Acceptability* “usually refers to the entire study, which includes the model and is also clearly a reflection of the relationship between the modeler(s) and the user or client” (Pidd, 2003, p. 298). Finally, Robinson (2002) describes *credibility* as the confidence of the model users and clients in using a model and its results. In simulation models, this refers to as the confidence in using the derived information from a simulation model to make decisions (Schruben, 1980). In this paper we use the terms *validity* and *validation* to refer to this broad spectrum of related notions and corresponding techniques and tests that are used for assessing DS models.

We base our subsequent review of the DS models’ validation on the following four distinct categories of validity and validation activities: *data validity* (Balci, 1989; Gass, 1977; Oral & Kettani, 1993; Sargent, 2013), *conceptual model validity* (Balci & Nance, 1985; Gass, 1983; Oral & Kettani, 1993; Sargent, 2013), *computational verification* (Adrian, Branstad, & Cherniavsky, 1982; Chattergy & Pooch, 1977; Deutsch, 1981; Dunn, 1987; Gass, 1983; Myers, 1978; Sargent, 2013; Schlesinger et al., 1979; Whitner & Balci, 1989), and *operational validity* (Boehm, Brown, & Lipow, 1976; Gass, 1977; Sargent, 2013).

2.1.1. Data validation

The purpose of data validation is to ensure that “the data necessary for model building, model evaluation and testing, and conducting the model experiments to solve the problem are adequate and correct” (Sargent, 2013, p. 14). Similarly, Balci (1989) describes data validation as assuring that both input data and model parameters have the required *accuracy*, *completeness*, *impartiality* and *appropriateness* for the proposed objectives of the model. Historically, empirical disciplines such as health and social sciences emphasize examining the *accuracy*, *consistency*, and *completeness* of the study data (Rothman, Greenland, & Lash, 2008). Oral and Kettani (1993) proposes a number of goals for a data validation procedure – these include ensuring data *sufficiency*, *accuracy*, *appropriateness*, *availability*, *maintainability*, *reliability*, as well as the *feasible cost* of data collection and manipulation.

For the OR/MS models, Gass (1983) distinguishes between raw data and structured data, i.e. the raw data that has undergone some types of manipulation. Three desirable properties for raw data validation are then recommended: *accuracy*, defined as “the ability to correctly identify, obtain, and measure what is desired”, *impartiality*, i.e. “the assurance that the data are recorded correctly”, and *representativeness*, namely “the assurance that the universe from which any sample data are drawn is properly identified and the sample was random” (Gass, 1983, p. 612). For the structured data validation, the emphasis should be placed on the auditing of every step of data manipulation before the data are used as a part of the DS model (Gass, 1983).

Although there is a broad consensus in the literature as to what constitutes data validity, the recommendations in the literature as to how to perform data validation for a DS model are much less frequent. Empirical disciplines, including health sciences, emphasize the procedures of obtaining data and empirical estimates from accredited data sources and published research (Biau, Kernéis, & Porcher, 2008; Ellenberg, 1994). There is a strong emphasis on

sampling procedures and sample size estimations to ensure that the precision and applicability of the outputs are adequate for the intended purpose of the study (Biau et al., 2008). The recommendations also include screening the data for any unspecified outliers and missing values which might have been developed during the sampling process, or while raw data are being transformed to any type of structured data (Balci, 1989; Sargent, 2013). According to Liu, Cheng, and Wu (2002) outliers either relate to measurement errors or *phenomena of interest*. Two methods that can be applied here are the *outlier identification* and *outlier accommodation* (Lin & Brown, 2006). In the *outlier identification* the goal is to detect the outliers and decide whether they should be accepted or rejected (Hawkins, 1980), while in *outlier accommodation*, the researchers try “to develop some robust estimates that are insensitive to the existence of outliers” (Lin & Brown, 2006).

When following these procedures, additional emphasis is placed on the importance of proper documentation of both raw input and parameters data, as well as all data modifications (Balci, 1989; Gass, 1983; Sargent, 2013).

2.1.2. Conceptual model validity

Compared to the discussions on data validity, various aspects of conceptual model validity are explored in the literature in greater detail. Sargent (2013) defines the goals of the conceptual model validation as to ensure that the assumptions and theories used to build the model are correct, as well as that there is a “reasonable” logical, mathematical and causal relationship in place for the intended use of the model. Gass (1983) describes three main groups of assumptions that should be examined to achieve conceptual model validity: *mathematical assumptions* about the model structure; *content assumptions* that are used to define terms and variables of the model, and *causal assumptions* that reflect the hypothesized relationships between terms and variables. In addition, to ensure logical and mathematical validity, Gass (1983) suggests to check the accuracy of the mathematical and numerical calculations, to check the accuracy of the logical flow of data and relevant results, and to ensure that none of the essential variables in the model or their relationships have been neglected. The role of the conceptual model validation, according to Oral and Kettani (1993), is in examining the “*appropriateness of the process of obtaining and using mental data bases*”. Following a similar line of thought, Balci and Nance (1985, p. 16) refer to the “*formulated problem*” verification as “*substantiation that the formulated problem contains the actual problem in its entirety and is sufficiently well structured to permit the derivation of a sufficiently credible solution*”.

As far as the specific procedures to ensure the conceptual model validity are concerned, Balci (1994) and Sargent (1986) suggest the application of the graphical models to provide better understanding of the conceptual model and its specifications. “Face validation” or “expert opinion” is another validation technique suggested by different authors (Balci, 1994; Hermann, 1967; Oral & Kettani, 1993). Sargent (2013) refers to this as systematic investigation of the subjective opinions of individuals working on the model in order to examine whether the model and its behavior are logical. Similarly, Gass (1983, p. 611) points out to the question of whether “*the initial impression of the model's realism is positive when reviewed by decision makers who know the system being modeled*.” Finally, “structured walkthrough” – i.e. the process of explaining the model by the model developer to a peer group is used to obtain the accuracy of the conceptual model for its intended use (Balci, 1994; Sargent, 2013).

The second group of tests includes the techniques to verify the logical behavior of the model and all of its sub-models (Balci, 1989; Gass, 1983; Oral & Kettani, 1993; Schellenberger, 1974). Such techniques are “tracing” where the logical behavior of a model entity is checked to verify its correctness and accuracy (Balci, 1994; Sargent,

2013), the “degeneracy test” by verifying that inputs and internal parameters have reasonable values (Gass, 1983; Sargent, 2013); and the “data relationship correctness” test, to ensure that all data in the model have the “*proper values regarding relationships that occur within a type of data, and between and among different types of data*” (Sargent, 2013, p. 16). It is also suggested to apply appropriate mathematical and statistical methods to test the main theories and assumptions of the model (Balci, 1994; Gass, 1983; Schellenberger, 1974).

2.1.3. Computational verification

The purpose of the computational model verification is to check the logic of the computer program and to ensure that all the numerical and data procedures based on the conceptual model have been implemented correctly (Gass, 1983). Not surprisingly, this topic has attracted major attention in the computer and computational science literature (Adrion et al., 1982; Chattergy & Pooch, 1977; Deutsch, 1981; Dunn, 1987; Myers, 1978; Sargent, 2013; Schlesinger et al., 1979; Whitner & Balci, 1989).

Two main approaches to computational model verification, as described by Fairly (1976), are *static* and *dynamic* testing. Static testing is aimed to verify the correctness of the computer code of the computational model; while in dynamic testing the computer code is executed under different scenarios, and the outcomes are used to identify whether the code and its execution are correct. For more details on validation, verification, and testing (VV&T) techniques of the computational models see Balci (1994) and Balci and Nance (1985). Different techniques for computer model verification include *debugging*, *walkthrough* and *execution tracing* (Balci, 1994). Debugging is the process of locating the errors, correcting them and checking the computer program of the computational model to confirm the code correctness (Whitner & Balci, 1989). Although debugging is usually a long and non-trivial task, it is an inevitable part of the computational model verification process (Dunn, 1987). Similarly, the term *walkthrough* refers to “an effort to locate the flaws in the design and/or source code” by the model development team (Whitner & Balci, 1989, p. 7). Different authors refer to this as “structured walkthrough” (Adrion et al., 1982; Deutsch, 1981; Myers, 1978; Sargent, 2013), with Yourdon (1979) identifying seven different roles for this task which usually can be performed by a group of three members. Finally, *execution tracing* can be used with debugging to help the model builder with isolating the identified errors in the code script and is described as “locating model defects by watching the line-by-line execution activity of the model” (Whitner & Balci, 1989, p. 22).

2.1.4. Operational validity

Operational validity refers to the accuracy of the model's outputs being sufficient for the model's intended use (Boehm et al., 1976). Gass (1983) sees the role of operational validity as that of justifying the use of the model based on the observed and expected errors of the model. Similarly, Sargent (2013) defines the operational validity of a DS model as the degree to which the model's outputs satisfy the accuracy requirements based on the intended use of the model and its applicability.

Specific techniques and tests employed to examine the operational validity of the model include model output analysis, robustness analysis, comparison to the results produced by other models, and tests to validate an appropriate application of the model (Boehm et al., 1976; Gass, 1983; Pidd, 2010; Sargent, 2013).

For output analysis, different types of visual (graphs) and analytical techniques are usually employed to verify the accuracy of the model's output (Balci, 1994; Gass, 1983; Sargent, 2001). Gass (1983) and Boehm et al. (1976) advocate the use of *robustness* test through checking the model's behavior while changing parameters and inputs of the model. Whitner and Balci (1989), Balci (1994)

Table 1

Summary of a systematic review of specific examples of decision support model validation in health operations research/management science literature.

Year	Publication title	Number of articles from key-word search
1971–2014	Operations Research	24
1971–2014	Management Science	49
1980–2014	Journal of Operations Management	79
1977–2014	European Journal of Operational Research	293
1973–2014	Omega	96
1974–2014	Computers and Operations Research	72
1971–2014	The Journal of the Operational Research Society	96
1984–2014	Annals of Operations Research	92
2012–2014	Operations Research for Health Care	9
1985–2014	Decision Support Systems	153
Total number		963

and Sargent (2013), refer to *extreme conditions test* for testing the credibility of the model structure and output for any extreme value of the internal parameters. Finally Myers, Sandler, and Badgett (2011) suggest application of *boundary analysis* techniques which “tests the activity of the model using test cases on the boundaries of input equivalence partitions. Test cases are generated just within, on top of, and just outside of the partition boundaries.”

Wherever possible, it is important that the results of the developed DS model are compared to the results of other previously validated models. Within the context of simulation models, this comparison can be made between two validated simulation models (Sargent, 2013). Finally, as suggested by Pidd (2010) and (Gass, 1983), the decisions made based on the model outputs should be verified in terms of the intended use of the model.

In summary, the importance of model validation for DS models, as well as the specific methods and techniques for model validation, have been extensively addressed in the OR/MS literature. At the same time, most of the effort in OR/MS model validation literature has been limited to broad descriptive articles and only little is reported on the specific cases of in-depth validation of individual non-simulation DS models in OR/MS literature.

2.2. Systematic review of specific examples of DS model validation in health OR/MS literature

To illustrate the limited extent of in-depth model validation studies of DS reported in health OR/MS literature, in March 2014 we conducted a systematic search of online journal archives of the following ten most popular OR/MS and DS journals: *Operations Research*, *Management Science*, *Journal of Operations Management*, *European Journal of Operational Research*, *Omega*, *Computers and Operations Research*, *Journal of the Operational Research Society*, *Annals of Operations Research*, *Operations Research for Health Care*, and *Decision Support Systems* (Table 1). The scope of this search has been explicitly limited to the studies that did not use simulation as the main modeling methodology. The reasons for such a choice are three-fold: firstly, as in our study we focus on the validation of the model that is statistical/analytical in nature, where simulation is used mainly for conducting probabilistic sensitivity analyses, it is more relevant to review the literature focusing on validation of similar “non-simulation” models; secondly, the domain of simulation modeling within OR/MS is well known for its careful attention to detailed model validation (Balci, 1989; Sargent, 2001, 2013; Whitner & Balci, 1989), often focusing on highly simulation-specific issues (e.g. visual walkthroughs, etc) that could be less relevant to “non-simulation” focused studies; and thirdly, as discussed by Brailsford and Vissers (2011), the domain of health care simulation modeling applications is expanding by the rate of up to 30 papers per day and conducting a comprehensive review of such

a body of literature deserves its own special focus (such as in, e.g., (Fone et al., 2003) and (Karnon & Afzali, 2014)) and would have been impossible within the scope of this study.

The search was conducted in two stages. At the first stage, we searched the “health*” & “model” & “valid*” string search criteria in the full text of the online archives of the mentioned journals, initially identifying 963 articles. These articles were screened to ensure that only non-simulation studies that have reported on in-depth validation of a DS model in the context of health systems and services applications are included. This resulted in inclusion of 72 articles for further study.

At the second stage, the “valid”, “validated”, “validity” and “validation” strings were used as the search criteria to search the contents of the remaining 72 articles for the description of the validation tasks undertaken in the paper. Four broad validation categories described in the previous subsection (*data validation*, *conceptual model validation*, *computational model verification*, and *operational validation*), formed a classification system and each identified paper was classified as belonging to one or more of these validation categories based on the reported validation activities performed in that paper.

As the result, we identified seven studies that reported on performing some data validation only, 25 studies reporting elements of conceptual model validation only, and 19 articles reporting some operational validation as the only validation approaches applied or discussed in the study. Four studies reported some aspects of both data and conceptual model validation, twelve studies included elements of both conceptual and operational validation, and further two studies reported addressing data as well as operational validation issues. Finally, only three articles in the reviewed non-simulation health OR/MS literature addressed some aspects of data, conceptual model, and operational validation (Blake & Carter, 2002; Junglas, Abraham, & Ives, 2009; Zanakakis, Alvarez, & Li, 2007) (Fig. 1). In the study by Blake and Carter (2002) authors reported on using the linear goal programming models for strategic resource allocation in health services. Zanakakis et al. (2007) investigated the effect of epidemic HIV/AIDS socio-economic determinants across different countries; and in the article by Junglas et al. (2009) the authors employed both qualitative and quantitative techniques to report on medical staff decisions on using mobile technologies in healthcare centers.

The true extent to which various studies have reported validation issues of the DS models varied broadly, with the majority of the studies only claiming the *fact of the application* of one or more validation techniques or tests without providing sufficiently detailed information regarding the validation procedures or their results. One of the rare exceptions to this trend is the study by Blake and Carter (2002) in which the authors have provided an extensive section on theoretical, data and predictive validity of

The process of identifying specific cases of DS model validation in Health OR/MS literature

Fig. 1. The process of identifying specific cases of DS model validation in Health OR/MS literature.

the model using a three-phase validation approach proposed by Schellenberger (1974). Another example is a study by Mason, Denton, Shah, & Smith, 2014 who employed a Markov decision process model to identify the optimal timing of blood pressure and cholesterol treatment for diabetes patients. The authors devoted a separate section to model validation, where data validation and comparison of model outputs to other validated models are discussed and a brief report on the validation procedures and outcomes are provided.

In summary, despite both well-recognized need for appropriate validation of DS models in the OR/MS literature and the relative abundance of health OR/MS models reported in the literature, only very few health non-simulation OR/MS publications could be identified that not only *mention the fact* of performing one or more DS model validation activities as a part of the reported study, but also *systematically discuss both the process and the results* of the undertaken DS model validation activities.

3. "Save a minute – save a day" model for evaluation of long-term benefits of early access to acute stroke thrombolysis

In this study, the model for evaluation of long-term benefits of faster access to thrombolysis therapy in acute ischemic stroke is used as a case to demonstrate how multiple aspects of data validity, conceptual model validity, computational verification, and operational validity can be systematically addressed when developing a complex decision support model for investigation and improvement in the healthcare context.

"Save a minute – save a day" model is an example of a complex OR/MS model in stroke care systems domain. Various OR/MS models addressing a number of interventions in stroke care systems have been proposed by OR/MS researchers and practitioners (Churilov & Donnan, 2012). These applications target, in particular,

the areas of stroke care operations improvement (Bayer, Petsoulas, Cox, Honeyman, & Barlow, 2010; Churilov et al., 2013; Cordeaux, Hughes, & Elder, 2011; Lahr, van der Zee, Luijckx, Vroomen, & Buskens, 2013; Lahr, van der Zee, Vroomen, Luijckx, & Buskens, 2013; Mar, Arrospeide, & Comas, 2010; Monks, Pitt, Stein, & James, 2012; Pitt et al., 2012), economic analysis (Bayer et al., 2010; Davidson, Husberg, Janzon, Oldgren, & Levin, 2013; Ghijben, Lancsar, & Zavarsek, 2014; Mar et al., 2010), public policy (Bayer et al., 2010; Churilov et al., 2013; Davidson, et al., 2013; Ghijben et al., 2014; Lahr et al., 2013; Monks, Pitt, Stein, & James, 2014; Yang, Chen, Chitkara, & Xu, 2014), and clinical decision making (Ghijben et al., 2014; Mar et al., 2010; Monks et al., 2012; Pitt et al., 2012; Yang et al., 2014).

Intravenous thrombolysis (tPA) is an acute medical therapy shown to improve patient outcomes in ischemic stroke (Jauch et al., 2013). The treatment is more effective when given early after symptom onset (Lees et al., 2010). Despite the accepted health benefits of faster treatment for the stroke patients, there was no measure to quantify the link between reductions in treatment delays and patient lifetime benefits. In the "Save a minute – save a day" study Meretoja et al., 2014 developed a DS model to estimate the effect of faster access to treatment on patient lifetime outcomes. The model demonstrates that for an average patient, each saved minute of stroke-onset-to-treatment time granted extra 1.8 disability-adjusted days of life. The model was computationally implemented using Excel 2010 worksheets (Microsoft Corp, Redmont, WA, USA), and STATA IC version 12 (StataCorp, College Station, TX, USA). The results of the study were reported for the clinical audience in the leading journal of the field, Stroke (Meretoja et al., 2014).

The "Save a minute – save a day" model is a complex model that relies on a wide variety of data sources and empirical estimates for important parameters originating from clinical literature

that are used as model inputs. Aimed at understanding the long-term effects of hypothetical changes in the complex stroke care process, this model can be naturally classified as belonging to the third class of *Pidd's (2010)* model use taxonomy, i.e. it should be considered as a model for *investigation and improvement*.

3.1. Model inputs

The following data were used as inputs for model building:

1. **An observational cohort of consecutive tPA patients:** based on a combined sample of 2258 patients retrieved from the Helsinki Stroke Thrombolysis Registry (*Meretoja et al., 2012*) and the Safe Implementation of Treatments in Stroke (SITS)-Australia database (*Simpson et al., 2010*), used to derive demographic and outcome distributions in real-life tPA use. This cohort provided distributions of age, sex, and stroke severity measured on the National Institutes of Health Stroke Scale (NIHSS) (*Lyden et al., 1994*), onset-to-treatment (OTT) times; and post-stroke disability level at 3 months, measured by modified Rankin Scale (mRS) (*Rankin, 1957*). mRS categorizes the functional disability into seven broad groups ordered from mRS 0 (No symptoms at all), to mRS 6 (Dead) (*Hong & Saver, 2009*).
2. **Published pooled analysis of tPA randomized controlled trials that estimates the tPA treatment effect over time:** The effect is graphically summarized as odds ratios with corresponding confidence intervals for tPA treatment compared to placebo of obtaining favorable outcome (mRS 0–1) and for mortality (mRS 6) vs. onset-to-treatment time (*Lees et al., 2010*). Updated results of these analyses were recently published in a paper by *Emberson et al. (2014)* and are used later in this paper to validate the model outputs;
3. **General Australian population life expectancy age- and sex- specific data obtained from Australian Bureau of Statistics:** These data are used to model long-term survival of patients at various mRS categories compared with the general population (*Australian Bureau of Statistics, 2013*);
4. **Parameters necessary to translate the 3-month mRS outcome data into a long-term metric of Disability-adjusted Life Years (DALYs) lost:** The parameters include age-weighting modulation factor (k), age-weighting function (β), adjustment constant (C), discounting rate (r), mRS specific disability weights (DWs), and mRS specific annual risk of death (*Murray, 1996; World Health Organization, 2014*).

3.2. Model-building process

Using these data, the model was built in the following stages (*Fig. 2*):

Stage 1: Generating patient-specific probabilities of achieving specific mRS categories at the cohort observed onset-to-treatment time (OTT) t_0 - We randomly selected 80 percent of the combined observational cohort and created seven separate binary logistic regression models (one for each individual mRS category as the dependent variable and age, baseline NIHSS, and OTT as independent variables). Using these regression models we then generated the patient-specific probability distributions for each mRS category.

The choices of age and baseline stroke severity as input parameters reflect an evidence-based understanding of these variables being strong prognostic factors for the functional outcome after stroke (*Jauch et al., 2013*). Using a simple normalization scaling practice, we ensured that the sum of patient-specific probabilities for each mRS category is equal to one, and then we used the estimated probabilities to generate patient-specific probability distributions

of achieving a given mRS category at the observed OTTs t_0 .

Stage 2: Changing probabilities of achieving mRS 0–1 and mRS 6 over time - The pooled analysis of thrombolysis randomized controlled trials (*Lees et al., 2010*) provides an estimation of the treatment delays on the treatment effect of thrombolysis compared to placebo. This relationship between the odds ratio of achieving mRS 0–1 and mRS 6 as a function of OTT between 60 and 360 minutes has been reported by *Lees et al. (2010)* in graphical format; without providing any analytical expressions for the odds ratio curves. Therefore, in this paper we derived relevant analytical expressions for odds ratios of mRS 0–1 and mRS 6 using the best fit (based on adjusted R^2 criterion) and used these equations to estimate the odds ratios for mRS 0–1 and mRS 6 as a function of OTT for any value of OTT between 0 and 270 (the currently accepted evidence-based upper time limit for tPA treatment) minutes (*Jauch et al., 2013; Lees et al., 2010*).

The odds ratio(t) reported by *Lees et al. (2010)* presents the ratio of odds of achieving mRS 0–1 (or, respectively, mRS 6), by the patients treated with tPA at a time point t and the odds of achieving the same outcome by the patients not treated by tPA. To estimate the probabilities of mRS 0–1 and mRS 6 at any given time, we used the following formula:

$$p(t) = 1 / \{1 + (\text{odds ratio}(t_0) / \text{odds ratio}(t)) * [(1 - p(t_0)) / p(t_0)]\}$$

where $p(t_0)$ represents the probability of mRS 0–1 (or, respectively, mRS 6) at OTT= t_0 time, and $\text{odds ratio}(t)$ represents the fitted value of odds ratios for mRS 0–1 (mRS 6) at a given time.

Stage 3: Estimating probabilities of achieving a specific mRS at any time - Since the graphs presented in *Lees et al. (2010)* only report on the odds ratios for mRS 0–1 and mRS 6, we used the patient-specific probabilities of achieving mRS 0–1 and mRS 6 at any given time t obtained at Stage 2 to estimate the probability distributions for each individual mRS category (both for treated and non-treated patients). These probabilities were estimated assuming that the ratios of probabilities for achieving individual mRS categories remain identical to those obtained in Step 1 from the logistic regression models based on the observed cohort data.

Stage 4: Expected Disability-adjusted Life Years (DALY) lost - Disability adjusted life-years lost (DALYs) metric developed by the World Health Organization (WHO) was used to translate the 3-month mRS outcome data into a meaningful long-term metric. This metric expresses the total amount of optimal life-years lost due to both premature mortality and living with disability and consists of two components: years of life lost due to premature death (YLL) and years of life lost due to disability (YLD) (*Hong & Saver, 2010*).

YLL is calculated as the difference between general population life expectancy of a person at a given age and sex, that is, life expectancy of a person without stroke, and age- and sex-matched life expectancy of a stroke patient in a certain mRS category. The long-term annual risk of death after stroke was taken from published literature for mRS categories 0–5 times that of the general population for each mRS level (*Hong & Saver, 2010*).

The following equation was used to estimate YLL:

$$YLL[r, K] = KCe^{rA} / (r + \beta)^2 e^{-(r+\beta)(L+A)} [-(r + \beta)(L + A) - 1] - e^{-(r+\beta)A} [-(r + \beta)A - 1] + [(1 - K)/r](1 - e^{-rL})$$

where K indicates age-weighting modulation factor ($K = 1$ or 0); β is the parameter from age weighting function ($\beta = 0.04$); r is the discount rate ($r = 0.03$ or 0); C is a constant ($C = 0.1658$); A is the age of death, and L is the life expectancy of general population at the age of stroke.

YLD is calculated by multiplying the life expectancy of a stroke patient by a disability weight and, therefore, demonstrates how much the value of life has diminished in years lived after stroke (*Hong & Saver, 2009*). The following equation was used to estimate

Fig. 2. Overview of the model development and validation - Computational model verification was performed in all stages of the model development.

YLD:

$$YLD[r, K] = DKCe^{rAs} / (r + \beta)^2 [e^{-(r+\beta)(Ld+As)}] [-(r + \beta)(Ld+As) - 1] - e^{-(r+\beta)As} [-(r + \beta)As - 1] + [(1 - K)/r](1 - e^{-rLd})$$

where K , β , r and C are the same parameters as in YLL formula; A is age of onset of disability; L is duration of disability; and D is disability weight.

Having estimated the values for YLL and YLD , DALYs are then derived from the formula $DALYs[r, K] = YLL[r, K] + YLD[r, K]$. In "Save a minute- save a day" model, DALYs lost for each patient in the observational cohort at OTT has been initially estimated, and then we have modeled the DALYs lost with regard to the treatment delays for each patient (Hong & Saver, 2010).

3.3. Model results

3.3.1. Observational cohort characteristics

The observational cohort of 2258 tPA patients was investigated to observe data characteristics. Median age was 70 (interquartile range (IQR), 60–78), onset-to-treatment time 125 (92–162), NIHSS 9 (6–15), and 1247 (55 percent) male stroke patient. At 3 months, 252 (11.2 percent) patients were dead and 850 (37.6 percent) had an excellent outcome of mRS 0–1 (Meretoja et al., 2014).

3.3.2. Effect of treating faster in the whole cohort and individual patients

Running the observations with real-life OTT in the whole cohort, each minute of OTT time saved resulted in on average extra

1.8 days of DALY (median 1.7, IQR, 1.1–2.3, SD 0.8, range 0.1–4.6 days of DALY). On average a 70 years old female patient gained extra 0.6 (sever stroke) and 0.9 (mild stroke) days, a 50 y.o. patient gained extra 3.5 and 2.7 days respectively, and a 30 y.o. patient gained extra 4.3 and 3.3 days respectively for each minute saved. The younger patients and females slightly gained more benefit from faster treatment as a result of their longer lifetime (Meretoja et al., 2014).

3.3.3. Robustness analysis

Two approaches of one-way analysis and probabilistic analysis were applied to evaluate the model robustness. We performed the one-way analysis by setting the value of each model input to their upper and lower 95 percent confidence interval. The probabilistic analysis is performed by "formulating uncertainty in the model inputs by a joint probability distribution and then analysing the induced uncertainty in outputs" as described by Oakley and O'Hagan (2004). In this study, we have performed the probabilistic analysis by sampling based on an underlying distribution from the feasible space of the mRS 0–1 and mRS 6 odds ratio(t) equations estimated by the 95 percent confidence interval limits provided by Lees et al. (2010). Both tests demonstrated that the results were robust overall.

In one-way analysis by varying the odds ratio of mRS 0–1 to the lower and upper 95 percent CI respectively, a patient benefits on average extra 0.84 and 2.75 days for each minute saved. These values are changing between 1.41 and 2.08 days when varying the odds ratio of mRS 6 to the lower and upper 95 percent CI. With respect to the effect of age and NIHSS on outcome, a patient benefits vary from 1.65 to 1.90 days for age, and 1.75 to 1.81 days for NIHSS with respectively changing the lower and upper 95 percent CI. For the probabilistic analyses we estimated a 95 percent prediction interval from 0.9 to 2.7 days for each minute saved. For all of these robustness analyses a point estimate was 1.78 days for each minute saved (Meretoja et al., 2014).

3.3.4. Intended use of the model

This model was developed to provide better understanding of the effect of faster tPA treatment on patient lifetime outcomes. The intention was that the improved understanding of the benefits of earlier thrombolysis expressed in an easier-to-understand manner as days of disability-free life saved per minute of earlier access to tPA provided by this model would lead to increased awareness of public and health service providers of the importance of faster treatment for stroke patients.

Appropriate validation of the "Save a minute – save a day" model presented a particular challenge as it needed to encapsulate the validation of a variety of datasets and parameters used as model inputs, as well as detailed validation from the conceptual, computational and operational perspectives. To achieve this, in the next section we demonstrate how multiple aspects of data validity, conceptual model validity, computerized verification, and operational validity can be systematically addressed based on the outcomes of the literature review conducted in the previous section.

4. In-depth validation of the "Save a minute – save a day" model

The aim of this section is to provide insight into validation of the "Save a minute – save a day" model using the four broad validation categories discussed earlier in the paper. These four categories with regard to different stages of the model development have been presented in the Fig. 2.

It is important to note that, as discussed in Section 2.1, while many authors repeatedly addressed different categories of model

validity and validation process since early 1970s, it was Gass (1977, 1983) and Sargent (1979, 2013) who comprehensively summarized these categories. In the context of simulation model development, Sargent (1979, 2013) proposed a "development process" (2013, p.14) based on four categories of validation activities as well as a generic structure for model validation documentation (2013, p.22). Gass (1983) clearly outlined different categories of model validity and provided in-depth discussion of specific technical steps involved in decision model validation. In this paper we utilize both Sargent's (2013) approach and that of Gass (1983) in the context of non-simulation modeling, while employing multiple methods and techniques proposed by various authors as described in Tables 2–5.

4.1. Data validity

Both input data (e.g. stroke thrombolysis cohort and general population life expectancy data) and previously published data in the form of various parameters estimators (e.g. pooled analysis of tPA, annual risk of death and disability weights) were used to build the "Save a minute – save a day" model.

4.1.1. Validation of input data

Observational cohort data: The observational cohort was extracted from two databases: the Helsinki Stroke Thrombolysis Registry (Meretoja et al., 2012), and the Safe Implementation of Treatments in Stroke (SITS-Australia) database (Simpson et al., 2010). Both registries are valid in terms of the representativeness as they represent consecutive patients with prospectively collected data on age, sex, stroke severity, OTT of ischemic stroke patients for Finland and Australia. They were created and maintained in accordance to the best practice guidelines for clinical registries and were approved by the relevant institutional authorities.

Helsinki Stroke Thrombolysis Registry contains the information about all cases of acute stroke thrombolysis is given to patients at the Helsinki University Central Hospital. The data used in this study was generated during the period between March 1998 and December 2011 and included relevant patient information for 1998 patients treated with tPA (Meretoja et al., 2012). Similarly, SITS-Australia contains the information about the cases of acute stroke thrombolysis administered for in various centers in Australia (Simpson et al., 2010). The data from the SITS-MOST Australian dataset used in this study was generated between December 2002 and December 2008 and included relevant patient information on 704 patients from 14 treating centers.

Based on a pre-specified inclusion criteria, in both datasets, the patients with OTT greater than 4½ hours ($n = 150$ in Helsinki database), and those with deviations from standard treatment procedure ($n = 56$ in Helsinki database) and missing value on OTT, stroke severity or mRS outcome ($n = 65$ in Helsinki database and $n = 173$ in SITS database) were excluded. Each dataset was searched separately for any outliers before inclusion to the study. Eventually, we included 1727 patients from Helsinki registry dataset and 531 patients from SITS-Australia dataset which lead to a comparatively large sample size of 2258 patients representing two potentially different demographic groups for the study which was also searched for any outliers generated as a result of combining two different datasets and no outliers were found. The original databases as well as all the subsequent datasets used to develop the model were stored and documented on a password protected computer.

Having validated the observational cohort in terms of the representativeness, outliers, missing values and documentation, we then used the individual patient data to generate the patient-specific probabilities of achieving specified mRS categories as described in Stage 1 of the model development.

General population life expectancy: As described at stage 4 of the model development, general population life expectancies were

Table 2

Validation tests and techniques utilized for data validation of different components of the “Save a minute – save a day” model.

Validation task	Why we performed the validation task	How we performed the validation task	The conclusions/results of the validation task
Representativeness of the dataset			
	Based on: (Biau et al., 2008; Ellenberg, 1994)		
Observational cohort data	To ensure that demographics of the observed data have similar distributions to that of the published literature.	We obtained consecutive prospective data from two registries in Finland and Australia. We compared patient demographics to those in published literature.	The results of the study can potentially be generalized for a larger group of patients.
General population life expectancy	To ensure that the data source used to estimate the life expectancies is trustworthy.	We used the latest official data for age and sex specific life-expectancies of the residents of Victoria, Australia, obtained from ABS. We compared these with similar life-expectancy data from Finland, observing minimal differences (the average life expectancy at birth for men is 79.2 years in Australia versus 76.8 years in Finland and for women 83.8 years versus 83.3 years).	Increased confidence in the accuracy of the estimations for the general population life expectancy.
Pooled analysis of tPA effect over time	To ensure that parameters used to model the effect of tPA treatment over time were obtained from a trustworthy data source.	The parameters used to model the effect of tPA treatment over time were obtained from the pooled analysis of individual patient data of tPA randomized controlled trials by Lees et al. This study includes eight major randomized placebo-controlled trials of tPA for acute stroke (86 percent of the total number of patients in all trials for tPA treatment), which is the most representative of the current state of the knowledge in the field of stroke thrombolysis.	Increased accuracy of the parameters used to model the effect of tPA treatment changing with OTTs.
Parameters to calculate DALYs lost	To ensure that the data source used to obtain parameters for estimating DALYs lost is trustworthy.	We used the parameters obtained from WHO-GBDP, which was undertaken jointly with the World Bank, and Harvard School of Public Health.	Increased accuracy of the parameters used to estimate the expected DALYs.
Proper documentation of the data components			
	Based on: (Balci, 1989; Gass, 1983; Sargent, 2013)		
Observational cohort data	To enable the study replicability.	Both the original and replicated data were dated and stored on a password-protected computer.	Both original and replicated data can be retrieved when needed.
General population life expectancy	To enable the study replicability.	Both the original and replicated data were dated and stored on a password-protected computer.	Both original and replicated data can be retrieved when needed.
Searching for outliers			
	Based on: (Balci, 1989; Sargent, 2013)		
Observational cohort data	The existence of the outliers in the dataset can affect the accuracy of the results.	Each dataset from Finland and Australia as well as the combined dataset from these two registries were searched for any outliers.	Neither improbable nor impossible outliers were found in the observed dataset.
General population life expectancy	The existence of the outliers in the dataset can affect the accuracy of the results.	The dataset was searched for any outliers.	Neither improbable nor impossible outliers were found in the observed dataset.
Searching for missing values			
	Based on: (Balci, 1989; Sargent, 2013)		
Observational cohort data	Any data with missing values on parameters used to build the conceptual model could not be included in the observational cohort.	Each dataset from Finland and Australia as well as the combined dataset were searched for any missing value on age, sex, stroke severity, and OTT. We tried to retrieve data from the source documentation where missing, if not found it was excluded from study.	Under the assumption of missingness-at-random, we included those patients from the observational cohort without missing values on age, sex, stroke severity, and OTT to build the logistic regression model.
General population life expectancy	Any data with missing values on parameters used to build the conceptual model could not be included in the general population life expectancy dataset.	The dataset was searched for any missing value on the parameters needed to estimate the life expectancy of the general population. No missing values were found in this dataset.	We could include patients of all ages to estimate the life expectancy of the general population as we had life-expectancies for everyone.

used in the model to calculate DALYs. These included age and sex specific life-expectancies of the males and females resident in the State of Victoria, Australia for the period of 2007–2009. Being produced by the main government body in charge of the official statistics for the purposes of the analysis of life expectancy, this dataset was assumed to be valid. It contained the mortality rates for a group of 100,000 hypothetical newborn babies throughout their entire life, and all the necessary information to calculate the life expectancy for the mentioned group, such as the number of persons surviving to exact age of x , the proportion of persons dying between exact age x and exact age $x+1$ (mortality rate), the number of person years lived within the age interval x to $x+1$; and life expectancy at exact age x (Australian Bureau of Statistics). After ensuring that there are no outliers or missing values in the data, the full dataset along with its explanatory notes was documented in an Excel file for any retrieval purposes in the future.

4.1.2. Validation of parameters data

Pooled analysis of tPA effect over time- As described at stage 3 of the model development, pooled analysis of tPA treatment effect over time by Lees et al. (2010) was used to derive how the effect of tPA treatment varies with delays for onset-to-treatment time. To achieve this, odds ratios by onset-to-treatment time for mRS 0–1 and mRS 6 of tPA compared with placebo published in their analysis was used as parameters to build the model. The parameters obtained from this meta-analysis are most representative of the current state of the knowledge in the area of stroke thrombolysis as the study includes eight major randomized placebo-controlled trials of intravenous recombinant tissue plasminogen activator (rt-PA) for acute stroke (Lees et al., 2010).

Parameters to calculate expected DALYs- As described at stage 4 of the model development different parameters were derived from previously validated and published literature to transfer the 3-month mRS outcome into a long-term metric. The values of K , β , C , and r were originally determined by World Health Organization (WHO) as a result of Global Burden of Disease Project (GBDP) undertaken jointly by the World Bank, and Harvard School of Public Health in 1992 (Murray, 1996; World Health Organization, 2014). WHO-GBDP also developed DWs for chronic post-stroke states, using the person trade-off (PTO) method where healthcare professionals judge health conditions from a broad public health point of view, ensuring equity across different health states (Murray, 1996; Nord, 1992). Hong and Saver (2009) then formed an international panel of 9 experts to use the trade-off procedure combined with a Delphi process to estimate DWs for each of the seven mRS grades (0.000, 0.053, 0.228, 0.353, 0.691, and 0.998 for mRS categories 0–5, respectively).

Lastly, the long-term annual risk of death after stroke is expressed as disability-linked mortality hazard ratios for premature annual mortality for mRS categories from 0 to 5 which were adopted from previously validated published literature (1.53, 1.52, 2.17, 3.18, 4.55, and 6.55 times that of the general population for mRS categories 0–5, respectively) (Hong & Saver, 2010). These publications however did not provide the CIs around DWs and long-term annual risk of death after stroke that could be used for model validation purposes.

The summary of different validation methods and techniques used to obtain data validity for “Save a minute – save a day” model has been provided in Table 2.

4.2. Conceptual model validity

The conceptual model of the effect of OTT delays on mRS probabilities was mainly developed at stage 2 and stage 3 of the model development. Its validation consisted of the validation of

the model's assumptions and its logical and mathematical structure. Different methods and validation tests used to validate the conceptual model were: degeneracy test, data relationship correctness test, mathematical and statistical methods, tracing and structured walkthrough.

4.2.1. Validation of the model assumptions

Four main assumptions were formulated and validated by the experts as follows:

1. The upper time limit to receive tPA treatment for the stroke patients was set to 270 minutes. As a result, any patient with OTT beyond 270 minutes was regarded as being treated with placebo (Meretoja et al., 2014). This value is the evidence-based upper time limit adopted by majority of international stroke clinical guidelines (Jauch et al., 2013). The appropriate selection of this parameter was validated using what is known as the degeneracy test.
2. For any given patient not treated with tPA, the odds of achieving a specific mRS outcome remain constant over the specified time period.
3. The relative ratios of probabilities of achieving mRS0 and mRS 1, as well as the relative ratios of achieving mRS categories 2–5 at any time point t (both for treated and non-treated patients), are assumed to be identical to those at the observed time t_0 .
4. The probability of achieving mRS 0–1 for a non-treated patient is equal to that of a probability for a treated patient at $t^*_{mRS\ 0-1}=350$ minutes; and the probability of achieving mRS6 for a non-treated patient is equal to that of a probability for a treated patient at $t^*_{mRS6}=165$ minutes: Based on the definition for odds ratios, the probability for a treated patient to achieve a given mRS category is equal to that of a placebo patient when the corresponding odds ratio=1. However, as no analytical expressions for the relationships were depicted in the mRS 0–1 and mRS 6 graphs presented by Lees et al. (2010), we derived relevant analytic expressions for the presented mRS 0–1 and mRS 6 curves, and validated the equations using the best fit R^2 criterion. The role of these equations is to express the odds ratios for mRS 0–1 and mRS 6 as a function of OTT between 0 and 270 minutes.

All the four mentioned assumptions were validated by formally obtaining the opinion of the clinical expert- the approach presented in the validation literature as walkthrough validation technique (Sargent, 1996).

4.2.2. Validation of model structure/formulation

The logical structure of the conceptual model was validated through checking the mRS probability distributions, change over time formulation, numerical relationships in the model, and DALYs mathematical formulation.

4.2.3. The mRS probability distributions

We randomly selected 80 percent of the combined observational cohort and, having validated the standard assumptions underlying the use of logistic regression (such as independence of individual observations, appropriate distributional assumptions, collinearity, and model fit), constructed independent binary logistic regression models for each individual mRS category with age, baseline NIHSS, and OTT as independent variables (refer to stage 1 of the model development). The statistical validity of the mRS prediction model was evaluated in the remaining 20 percent of the observational cohort, with no significant difference between predicted and observed mRS categories being identified ($\chi^2 p$ -value = 0.51).

Having generated the probability distributions for individual mRS categories we assured that the sum of the probabilities for each patient is equal to one; thus validating the normalization scaling procedure performed at Stage 1 of the model development.

4.2.4. Numerical relationships in the model

By selecting a large enough original cohort sample size we ensured that the relationship between age, NIHSS, and mRS at a given point of OTT (based on the cohort data), is no worse than the precision of the relationship between mRS and time (based on meta-analysis data).

4.2.5. Changes over time formulation

As previously stated in stage 2 of the model development, to model how the onset-to-treatment time affects the probability of mRS 0–1 and mRS 6 for a specific patient, we derived relevant independent analytic expressions for mRS 0–1 and mRS 6 curves and validated the resulting equations using the best fit R^2 criterion. For both curves the corresponding values were $R^2 = 0.999$. These equations were used to express the odds ratios for mRS 0–1 and mRS 6 as a function of OTT between 0 and 270 minutes.

4.2.6. DALYs mathematical formulation

As described at Stage 4 of the model development, DALY is a measure consisting of two components: years of life lost due to premature death (YLL) and years of life lost due to disability (YLD). Data relationship correctness technique was employed to validate different parts of DALYs' formulation. For instance, if we compare the values for DALYs between a female and a male patient with identical age and mRS category, the DALYs for a female patient is expected to be greater than that of the male patient.

Each of the three mentioned components of the conceptual model structure were validated by tracing the formulation separately by different members of the model development team and the results were compared to identify and resolve inconsistencies.

Also, the structured walkthrough validation technique was employed to ensure that the logical behavior of the conceptual model is aligned with the clinical practice by explaining the model assumptions, parameters and formulation to a clinician.

The summary of different validation methods and techniques used to obtain conceptual validity of the “Save a minute – save a day” model has been provided in Table 3.

4.3. Computational verification

By computational model verification, the modeler ensures that the computer programs and codes to build the computer model of the conceptual model have been used and implemented correctly. For the present model, this consisted of two main stages: (1) computations verification in Excel; and (2) code scripts verification in STATA. Different techniques used for computational verification were debugging, walkthrough and execution tracing.

4.3.1. Computations verification in excel

The observational cohort data, mRS category-specific life expectancies, and DALYs lost for each combination of age and sex of stroke patients were all stored in Excel worksheets accessible to STATA software through a set of STATA codes. After running the model in STATA, the model outcomes were also exported and stored in a separate Excel worksheet for data processing purposes. The walkthrough verification technique was employed by the analytical expert to verify the correct storage and execution of the computations in the Excel worksheets.

4.3.2. Code scripts verification in STATA

The conceptual model implementation and outcome analysis were mainly executed in STATA, through a set of codes developed within the software to run the model as well as to link STATA to Excel worksheets. Debugging and execution tracing verification techniques (Whitner & Balci, 1989) were employed to verify the codes in STATA.

The summary of different verification methods and techniques used for computational verification of the “Save a minute – save a day” model has been provided in Table 4.

4.4. Operational validity

To achieve operational validity, the outputs of the DS model are verified to obtain the accuracy needed for the intended use of the model. Since the “Save a minute – save a day” model presents the first attempt to investigate the effect of faster access to thrombolysis treatment, there was no data available in a real-life system to be used for specifying a clear range of the values of the DALYs per unit of onset-to-treatment time. However, other studies in stroke literature have addressed the issues of a plausible range of expected DALYs in the absence vs the presence of treatment, thus providing an acceptable range for the model outputs over the full range of plausible onset-to-treatment time. In this scenario, different techniques used to validate the operational model are output analysis, robustness analysis, comparison to the results produced by other known models, and tests to validate an appropriate application of the model.

4.4.1. Validation of the model output

The accuracy of the expected DALYs (as the final output of the model) was validated by output analysis, robustness analysis, and comparison to the results produced by other models as described below:

Different graphs and summary statistical measures (i.e. mean, median, 95 percent CIs) were generated to validate the model outputs (Meretoja et al., 2014). The updated results of the tPA randomized controlled trials used to estimate the tPA treatment effect over time published by Emberson et al. (2014) were used to validate the model outputs and the results were consistent with the earlier results obtained from the meta-analysis by Lees et al. (2010).

Also, DALYs gained per tPA treated patient for this study were compared to the results of a long-term utility of tPA (DALY/QALY gains) from other studies (Meretoja et al., 2014).

Additionally, two approaches were applied to verify the robustness of the model: one-way analysis and probabilistic analysis. In the one-way analysis, by varying each model input to their upper and lower 95 percent CI we evaluated the model robustness with regard to those uncertainties in the inputs. To account for potential uncertainties in the cohort data, we performed a series of one-way analyses by systematically and sequentially substituting the upper and lower 95 percent CIs values for the regression coefficients for the age and NIHSS generated by the binary logistic regression models. To account for potential uncertainties of the pooled analysis by Lees et al. (2010), we modified the equations of $odds\ ratio(t)$ for mRS 0–1 and mRS6 in the mRS probability distribution formula to sequentially reflect the upper and lower 95 percent confidence limits for these two mRS categories as reported by Lees et al. (2010).

Probabilistic robustness analyses were performed by sampling according to an underlying Normal distribution from the feasible space of the mRS0–1 and mRS6 $odds\ ratio(t)$ curves bounded by the 95 percent confidence limits reported by Lees et al. (2010) and reflecting various potential time effects based on the pooled analysis, through a set of 1000 independent runs. The resulting

Table 3

Validation tests and techniques utilized for conceptual model validation of different components of the “Save a minute – save a day” model.

Validation task	Why we performed the validation task	How we performed the validation task	The conclusions/results of the validation task
Degeneracy test			
	Based on: (Gass, 1983; Sargent, 2013)		
Limitation of upper treatment time to 270 minutes	An appropriate selection of the internal parameters directly affects the accuracy of the logical behaviour of the conceptual model.	Based on published international stroke clinical guidelines we observed that the vast majority of these studies implemented the 270 minutes time window, although a few still used the old 180 minutes, which was based on guidelines up to year 2008.	Increased credibility of the model as the vast majority of the users of the model outputs will consider the time window appropriate.
Data relationship correctness			
	Based on: (Sargent, 2013)		
DALYs mathematical formulation	To ensure that there is a logical relationship between DALYs' values of female and male patients.	The values for DALYs between a female and a male patient with similar characteristics were compared.	Increased credibility of DALYs formulation used for male and female patients.
Numerical relationships in the model	To ensure that there is a logical relationship between different data components used to build the conceptual model.	We selected a large enough original cohort sample size to ensure that the precision of the regression coefficient estimates describing the relationship between age, NIHSS, and mRS at a given point of OTT (based on the cohort data) is no worse than the precision of the relationship between mRS and time (based on the meta-analysis data).	This increased the overall precision of the estimates leading to the increased validity of the model.
Tracing			
	Based on: (Balci, 1994; Sargent, 2013)		
Change over time formulation	To ensure that the logical behaviour of the formulation is correct and the required accuracy obtained.	The process of building and selecting the equations to formulate the effect of OTTs on probability distributions was performed separately by three members of the model development team and the results were compared for any inconsistency.	The equations used to formulate the effect of OTTs on probability distributions were verified and the equations were corrected where necessary.
mRS probability distributions	To ensure that the logical behaviour of the prediction model is correct and the required accuracy obtained.	The statistical process of constructing the prediction model was performed separately by three members of the model development team and the results were compared for any inconsistency.	We confirmed that there is no significant difference between predicted and observed mRS categories to be used for building the conceptual model.
DALYs mathematical formulation	To ensure that the logical behaviour of the DALYs formulation is correct and the required accuracy obtained.	The logical behavior of the DALYs formulations was reviewed separately by three members of the model development team and the results were compared for any inconsistency.	DALYs formulations were verified and the equations were corrected where necessary.
Mathematical and statistical validation methods			
	Based on: (Balci, 1994; Gass, 1983; Schellenberger, 1974)		
Change over time formulation	To ensure that the equations used to build the conceptual model are accurate enough.	We derived the relevant analytical expressions for mRS 0–1 and mRS 6 curves by selecting 5 point estimates in the lines and selected the best fit among the resulting equations using the R^2 criterion (achieving R^2 of 0.999).	Correct representation of the relationship between OR and time is achieved.
	To ensure that the probability equations used to build the conceptual model are logically correct.	We ensured that the sum of probabilities for each patient generated in Stage 1 of the model is equal to one.	Increased credibility and accuracy of the probability equations.
mRS probability distributions	To ensure that the mRS probability distributions used to build the conceptual model are logically correct.	We randomly selected 80 percent of the combined observational cohort and constructed binary logistic regression models for each individual mRS category. The statistical validity of the mRS prediction model was evaluated in the remaining 20 percent of the observational cohort.	We ensured that the regression model constructed based on the 80 percent of the observational cohort was reflecting the nature of the relationships in the remaining 20 percent, therefore being valid for the full observational cohort.
Structured walkthrough and face validity			
	Based on: (Balci, 1994; Hermann, 1967; Oral & Kettani, 1993; Sargent, 2013)		
Conceptual model building	To ensure that the conceptual model is accurate enough for its intended use.	The logic of the model structure, assumptions, and parameters were explained step by step to a clinician who asked questions and challenged the choices, leading to significant iterative model changes.	The logic of the model follows true clinical practice.

Table 4

Validation tests and techniques utilized for computerized verification of different components of the “Save a minute – save a day” model.

Validation task	Why we performed the validation task	How we performed the validation task	The conclusions/results of the validation task
Debugging	Based on: (Dunn, 1987; Whitner & Balci, 1989)		
Code scripts in stata	To confirm the correctness of the codes used to build the computational model.	The code script in Stata was screened to locate and correct the potential errors.	Typing and logic errors were identified and removed.
Walkthrough	Based on: (Adrian et al., 1982; Deutsch, 1981; Myers, 1978; Yourdon, 1979)		
Computations in excel	To ensure that all the computations used to build the computational model are correct.	The analytical expert verified the correct storage and execution of the computations in the Excel worksheets.	Storage and execution of relevant computations of the model were verified and revised in Excel worksheets.
Execution tracing	Based on: (Whitner & Balci, 1989)		
Code scripts in stata	To confirm the correctness of the codes used to build the computational model.	Defects of the code script in Stata were located and corrected by line-by-line execution of the code by an analytical expert.	Typing and logic errors were identified and removed.

probability profiles for all mRS categories were then used to estimate DALYs gained or lost if either the whole cohort or an individual patient would have been treated faster or slower.

4.4.2. Validation of the model application

In order to ensure operational validity, the model developers and users should formulate their understanding of the intended application of the model and its boundaries before employing the model and its results as a decision support tool. For our model, these included the following considerations:

1. Intended model use in different population demographics: since the study dataset is based on two separate populations, the characteristics of the Helsinki and SITS-Australia cohort, as well as the mixed cohort were provided for comparison (Meretoja et al., 2014). In addition, we developed a histogram of onset-to-treatment time distributions for each of the two cohorts separately to be considered before generalizing the results of the study (Meretoja et al., 2014).
2. Intended model use in different patient groups: the findings of the study demonstrates that patients with different gender, age and NIHSS benefit differently in terms of disability-free life over their full life-time (Meretoja et al., 2014). Therefore, caution should be exercised if the model were to be used as a decision support tool to understand the long term effects of earlier tPA treatment for specific patient groups.
3. Actual model use for increased public awareness: The “Save a minute– save a day” model was developed to provide better understanding of the effect of faster tPA treatment on patient lifetime outcomes. Ideally, this is supposed to directly lead to an increased awareness of public policy decision makers, stroke campaigns, and stroke care system providers of the importance of faster treatment for stroke patients. The original publication of the “Save a minute – save a day” model lead to a significant media exposure including sources like Bloomberg (Gale, 2014), The Times (Whipple, 2014), Reuters (Seaman, 2014), and Herald Sun (2014). American Heart and Stroke associations produced an infographics encapsulating the findings for the consumers (American Heart Association/American Stroke Association, 2014). The model's findings are also used by the Australian National Stroke Foundation and Victorian Stroke Telemedicine Initiative (State of Victoria, Australia) to advocate for wider use of stroke thrombolysis telemedicine in remote rural areas (Bladin & Cadilhac, 2014). Overall, current levels of the actual model use are quite consistent with the original modeling expectations.

The summary of different validation methods and techniques used to obtain operational validity of the “Save a minute – save a day” model has been provided in Table 5.

5. Discussion

Validation of DS models is performed with the aim of providing the decision makers with sufficient confidence to utilize the model's recommendations as a support for the decision-making activities. It is therefore important to reflect on the relationship between the process of model validation, modeling purpose, and adopted modeling methodology to fully appreciate the intricacies of both why and how model validation is conducted.

As far as the modeling purpose is concerned, it is instructive to consider a taxonomy of modeling purposes developed by Pidd (2010), which relies on four broad types of modeling purposes: modeling for *decision automation*, for *routine decision support*, for *investigation and improvement* (discussed in this paper), and for *providing insights*. Modeling for *decision automation* refers to the model use that is “frequent and routine, with, in general no need to prepare the model for each use” (Pidd, 2010, p. 16). With this modeling purpose, there is typically very little tolerance for any inaccuracy, since decisions made based on the model outputs on a continuous basis in a less supervised environment. Validation-wise, all the model assumptions, parameters, outputs, and their relationships should be examined critically before using the model in a decision making context (i.e. during the model-building stage). In modeling for *routine decision support*, where models are “used to assist, but not replace, people making routine, repeated decisions” (Pidd, 2010, p. 17), validation, in particular, is likely to be focused on assuring that combination of a decision proposed by a model and that made by a decision-maker will lead to a better overall decision outcome. As far as the models to *provide insights* are concerned, where models are “not would-be representations of the real-world, but are rather attempts to understand and represent how different stakeholders and interest groups see that world” (Pidd, 2010, p. 19), validation of these, often qualitative, rather than mathematical models is quite challenging and sometimes presents a point of controversy to OR/MS modelers (Pidd, 2010). Finally, in modeling for investigation and improvement, as discussed earlier in this paper and is the case for the “Save a minute- Save a day” model, a DS model can be employed to gain understanding of a new, possibly hypothetical at the stage of the model-building, system. In such models there is usually very limited or no data on the model behavior to be used for “output-based” validation of the model, and therefore model validation is based on critically testing all the model inputs, assumptions, parameters, and comparing the model outputs with the results of other similar studies (e.g. in our

Table 5

Validation tests and techniques utilized for operational validation of different components of the “Save a minute – save a day” model.

Validation task	Why we performed the validation task	How we performed the validation task	The conclusions/results of the validation task
Output analysis	Based on: (Balci, 1994; Gass, 1983; Sargent, 2001)		
Model output	To identify any unusual behavior of the model and pin-pointing errors that would not have been identified solely through summary statistics.	We created multiple graphical representations of individual patients and time series to validate the model outputs.	We found errors in the outputs as a result of either incorrect logic or implementation of the model which were subsequently corrected.
Robustness	Based on: (Balci, 1994; Boehm et al., 1976; Gass, 1983; Myers et al., 2011; Sargent, 2013; Whitner & Balci, 1989)		
Model output	To check the model's behaviour while changing the parameters and inputs of the model.	Two approaches of the robustness were used: one-way analysis and probabilistic robustness analysis.	Credibility of the outputs was increased by providing the users with estimates of uncertainty.
Comparison of the model outputs	Based on: (Sargent, 2013)		
Model output	To ensure that the model's outputs are accurate enough for the intended use of the model.	DALYs gained per tPA treated patient for this study was compared to the results of a long-term utility of tPA (DALY/QALY gains) from other studies (DALYs gained per tPA treated patient for the “Save a minute – save a day” model is 0.72 compared to 0.75 of QALYs gained for two other similar studies.)	Credibility of the outputs was increased by providing the comparison to other relevant studies.
	To ensure that the model's outputs are accurate enough for the intended use of the model.	The updated results of tPA randomized controlled trials used to estimate the tPA effect treatment over time published by (Emberson et al. (2014)) were used to generate the model outputs, and the results were compared to the previous results obtained by the model.	Increased credibility of the model's outputs.
Intended use of the model	Based on: (Gass, 1983; Pidd, 2010)		
Model application	To verify the decisions made based on the model outputs.	We discussed the limitations and boundaries of application of the model in the original article discussion section.	Users of the decision support model will understand the limitations and will not overgeneralize or use the model outside of its intended use.

case the comparison of the long-term utility of potentially delayed tPA in the “Save a minute – save a day” model and other studies that report long-term differences between tPA and control groups).

With regard to the relationship between the process of model validation and adopted modeling methodology, some validation steps are likely to be generic and do not depend on the adopted modeling methodology (e.g. data relationship correctness, tracing, walkthrough, or historical data validation), while other could be more specific to an adopted modeling method (e.g. workflow animation or event validity could be more appropriate for a simulation model than for a, say, optimization one). The applicability and relative importance of the validation tasks discussed in this paper are highly dependent on both the validation needs of a specific model type (according to, e.g., Pidd's (2010) taxonomy) and the specifics of an adopted modeling methodology. Consider, for example, an optimization model presented by Blake and Carter (2002), that can be classified as a routine decision support model (Pidd, 2010) that is used to assist with decisions about the number, type and prices of the services delivered by the acute care hospitals. Since the presented model is a normative cost and volume model, the authors have paid less attention to comparing the model outputs with that of the “output-based” empirical data, and mainly focused on other validation aspects such as theoretical, data and predictive validity. The authors also test the accuracy of the outputs through a three-stage process focusing on face validity, Turing test and historical data validation techniques to ensure that model results are accurate enough for its intended use, i.e. to be applied by clinical management within the hospitals. Having considered the validation reported by Blake and Carter (2002) through

the prism of the integrated validation framework suggested in our study, two things become apparent. Firstly, the proposed validation framework is generic enough in that it can provide useful insights as to the types of validation undertaken in other studies that differ in both the modeling purpose (routine decision support vs investigation and improvement) and modeling methodology employed (optimization modeling vs analytical/statistical modeling). Secondly, the proposed validation framework is specific enough to highlight the elements of validation that were not adequately performed or reported, such as, e.g. the lack of the computerized verification in Blake and Carter (2002) study.

6. Conclusions

Despite the broad consensus about the general importance of DS model validation present in the OR/MS literature, specific cases of an in-depth validation of DS models for understanding and improvement in the context of health systems and services applications, are rarely presented and discussed.

In this paper we set out to contribute to the OR/MS literature by presenting an in-depth case of *validation* of a “Save a minute – save a day” DS model for investigation and improvement that was designed and used to investigate the long-term benefits of faster access to thrombolysis therapy in acute ischemic stroke. This contribution is novel as comprehensive discussions of the model validity in the OR/MS literature are mostly confined to the domain of simulation modeling while other types of DS models, in particular, those relying on disparate data sources, diverse model inputs, and multiple modeling techniques, remain largely unexplored.

To achieve our research objective, we systematically investigated multiple aspects of *data validity*, *conceptual model validity*, *computational verification*, and *operational validity* when building and testing a complex DS model for investigation and improvement.

The key findings in the paper are as follows:

We empirically confirmed the scarcity of the in-depth model validation activities reported by systematically reviewing the health systems OR/MS literature. Excluding the studies that are using simulation as the only modeling method, remarkably few health OR/MS publications go beyond only mentioning DS model validation activities, but also systematically discuss both the process and the results of the undertaken validation activities.

We also presented an in-depth case of validation of a DS model for investigation and improvement intended to provide better understanding of the effect of faster tPA treatment on patient life-time outcomes. In doing so, we have demonstrated how validation of DS models can be conducted based on generic aspects of model validity (i.e. data validity, conceptual model validity, computerized verification, and operational validity). Our comprehensive approach to validation was based on a wide variety of methodological papers published in OR literature over the last three decades from Boehm et al. (1976) to Sargent (2013). The validation process and relevant literature sources have been comprehensively documented in Tables 2–5, addressing the applied validation task, an intention behind each validation task (i.e. Why), the process of performing the validation task (i.e. How), and the obtained conclusions/results using the case of the “Save a minute – save a day” model. We have argued that, although used in the context of health systems modeling in this paper, the nature of the validation of DS models for investigation and improvement is not necessarily specific to health systems context, and such validation can be conducted for various DS models used in the course of OR/MS research and practice.

Considering how important DS model validation is for OR/MS research and practice, it is easy to argue that as the researchers and practitioners of the discipline of *scientific decision making*, the OR/MS community has a lot to gain by paying more in-depth attention to the issues of appropriate conduct and reporting of model validation.

Acknowledgment

The Florey Institute of Neuroscience and Mental Health acknowledges the strong support from the Victorian Government and in particular the funding from the Operational Infrastructure Support Grant.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ejor.2016.02.035.

References

Adrian, W. R., Branstad, M. A., & Cherniavsky, J. C. (1982). Validation, verification, and testing of computer software. *ACM Computing Surveys (CSUR)*, 14, 159–192.

Altinel, I. K., & Ulaş, E. (1996). Simulation modeling for emergency bed requirement planning. *Annals of Operations Research*, 67, 183–210.

American Heart Association/American Stroke Association. Saving a minute may save a day. Retrieved 17 July, (2014), from http://www.strokeassociation.org/jdc/groups/stroke-public/@wcm/@hcm/@sta/documents/downloadable/ucm_462740.pdf.

Australian Bureau of Statistics. (2013). Retrieved 15 March 2013, from <http://www.abs.gov.au/ausstats>.

Balci, O. (1989). How to assess the acceptability and credibility of simulation results. In E. A. MacNair, K. J. Musselman, & P. Heidelberger (Eds.), *Proceedings of the 1989 winter simulation conference* (pp. 62–71). New York, USA: ACM.

Balci, O. (1994). Validation, verification, and testing techniques throughout the life cycle of a simulation study. *Annals of operations research*, 53, 121–173.

Balci, O., & Nance, R. E. (1985). Formulated problem verification as an explicit requirement of model credibility. *Simulation*, 45, 76–86.

Bayer, S., Petsoulas, C., Cox, B., Honeyman, A., & Barlow, J. (2010). Facilitating stroke care planning through simulation modelling. *Health informatics journal*, 16, 129–143.

Biau, D. J., Kernéis, S., & Porcher, R. (2008). Statistics in brief: The importance of sample size in the planning and interpretation of medical research. *Clinical orthopaedics and related research*, 466, 2282–2288.

Bladin, C. F., & Cadilhac, D. A. (2014). Effect of telestroke on emergent stroke care and stroke outcomes. *Stroke*, 45, 1876–1880.

Blake, J. T., & Carter, M. W. (2002). A goal programming approach to strategic resource allocation in acute care hospitals. *European Journal of Operational Research*, 140, 541–561.

Boehm, B. W., Brown, J. R., & Lipow, M. (1976). Quantitative evaluation of software quality. In *Proceedings of the 2nd international conference on software engineering* (pp. 592–605). Los Alamitos, USA: IEEE Computer Society Press.

Brailsford, S., & Vissers, J. (2011). OR in healthcare: a European perspective. *European Journal of Operational Research*, 212, 223–234.

Chattergy, R., & Pooch, U. W. (1977). Integrated design and verification of simulation programs. *Computer*, 10, 40–45.

Churilov, L., & Donnan, G. A. (2012). Operations Research for stroke care systems: an opportunity for the science of better to do much better. *Operations Research for Health Care*, 1, 6–15.

Churilov, L., Fridriksdottir, A., Keshtkaran, M., Mosley, I., Flitman, A., & Dewey, H. M. (2013). Decision support in pre-hospital stroke care operations: a case of using simulation to improve eligibility of acute stroke patients for thrombolysis treatment. *Computers & Operations Research*, 40, 2208–2218.

Cordeaux, C., Hughes, A., & Elder, M. (2011). Simulating the impact of change: implementing best practice in stroke care. *London Journal of Primary Care*, 4(1), 33–37.

Davidson, T., Husberg, M., Janzon, M., Oldgren, J., & Levin, L.-Å. (2013). Cost-effectiveness of dabigatran compared with warfarin for patients with atrial fibrillation in Sweden. *European Heart Journal*, 34, 177–183.

Deutsch, M. S. (1981). *Software verification and validation: Realistic project approaches*. Prentice Hall PTR.

Dittus, R. S., Klein, R. W., DeBrot, D. J., Dame, M. A., & Fitzgerald, J. F. (1996). Medical resident work schedules: design and evaluation by stimulation modeling. *Management Science*, 42, 891–906.

Dunn, R. H. (1987). The quest for software reliability. *Handbook of software quality assurance*, 1, 342–384.

Ellenberg, J. H. (1994). Selection bias in observational and experimental studies. *Statistics in Medicine*, 13(5–7), 557–567.

Embersen, J., Lees, K. R., Lyden, P., Blackwell, L., Albers, G., Bluhmki, E., Brott, T., Cohen, G., Davis, S., & Donnan, G. (2014). Effect of treatment delay, age, and stroke severity on the effects of intravenous thrombolysis with alteplase for acute ischaemic stroke: a meta-analysis of individual patient data from randomised trials. *The Lancet*, 384(9958), 1929–1935.

Fairley, R. (1976). Dynamic testing of simulation software. In *Proceedings of the 1976 Summer Computer Simulation Conference* (pp. 40–46).

Fone, D., Hollinghurst, S., Temple, M., Round, A., Lester, N., Weightman, A., Roberts, K., Coyle, E., Bevan, G., & Palmer, S. (2003). Systematic review of the use and value of computer simulation modelling in population health and health care delivery. *Journal of Public Health*, 25, 325–335.

Forrester, J. W., & Senge, P. M. (1978). *Tests for building confidence in system dynamics models*. System Dynamics Group, Sloan School of Management, Massachusetts Institute of Technology.

Gale, J. (2014). Stroke patients lose month for each 15-minute delay. Retrieved 12 September, 2014, from <http://www.bloomberg.com/news/2014-03-13/stroke-patients-lose-month-for-each-15-minute-delay.html>.

Gass, S. I. (1977). Evaluation of complex models. *Computers & Operations Research*, 4, 27–35.

Gass, S. I. (1983). Decision-aiding models: Validation, assessment, and related issues for policy analysis. *Operations Research*, 31, 603–631.

Gass, S. I. (1993). Model accreditation: a rationale and process for determining a numerical rating. *European Journal of Operational Research*, 66, 250–258.

Ghijben, P., Lancsar, E., & Zavarsek, S. (2014). Preferences for oral anticoagulants in atrial fibrillation: a best–best discrete choice experiment. *Pharmacoeconomics*, 32, 1115–1127.

Hawkins, D. M. (1980). *Identification of outliers*: Vol. 11. Springer.

Hermann, C. F. (1967). Validation problems in games and simulations with special reference to models of international politics. *Behavioral Science*, 12, 216–231.

Hong, K.-S., & Saver, J. L. (2009). Quantifying the value of stroke disability outcomes WHO global burden of disease project disability weights for each level of the modified rankin scale. *Stroke*, 40, 3828–3833.

Hong, K.-S., & Saver, J. L. (2010). Years of disability-adjusted life gained as a result of thrombolytic therapy for acute ischemic stroke. *Stroke*, 41, 471–477.

Jauch, E. C., Saver, J. L., Adams, H. P., Bruno, A., Demaerschalk, B. M., Khatri, P., McMullan, P. W., Qureshi, A. I., Rosenfield, K., & Scott, P. A. (2013). Guidelines for the early management of patients with acute ischemic stroke: a guideline for healthcare professionals from the American heart association/American stroke association. *Stroke*, 44, 870–947.

Junglas, I., Abraham, C., & Ives, B. (2009). Mobile technology at the frontlines of patient care: understanding fit and human drives in utilization decisions and performance. *Decision Support Systems*, 46, 634–647.

Karnon, J., & Afzali, H. H. A. (2014). When to use discrete event simulation (DES) for the economic evaluation of health technologies? A review and critique of the costs and benefits of DES. *Pharmacoeconomics*, 32, 547–558.

- Lahr, M. M., van der Zee, D.-J., Luijckx, G.-J., Vroomen, P. C., & Buskens, E. (2013). A simulation-based approach for improving utilization of thrombolysis in acute brain infarction. *Medical Care*, 51, 1101–1105.
- Lahr, M. M., van der Zee, D.-J., Vroomen, P. C., Luijckx, G.-J., & Buskens, E. (2013). Thrombolysis in acute ischemic stroke: a simulation study to improve pre- and in-hospital delays in community hospitals. *PLoS one*, 8(11), e79049.
- Lane, D. C. (1995). *The folding star: A comparative reframing and extension of validity concepts in system dynamics*. London, UK: London School of Economic.
- Lees, K. R., Bluhmki, E., Von Kummer, R., Brodt, T. G., Toni, D., Grotta, J. C., Albers, G. W., Kaste, M., Marler, J. R., & Hamilton, S. A. (2010). Time to treatment with intravenous alteplase and outcome in stroke: an updated pooled analysis of ECASS, ATLANTIS, NINDS, and EPITHET trials. *The Lancet*, 375, 1695–1703.
- Lin, S., & Brown, D. E. (2006). An outlier-based data association method for linking criminal incidents. *Decision Support Systems*, 41, 604–615.
- Liu, X., Cheng, G., & Wu, J. X. (2002). Analyzing outliers cautiously. *IEEE Transactions on Knowledge and Data Engineering*, 14, 432–437.
- Lyden, P., Brodt, T., Tilley, B., Welch, K., Mascha, E., Levine, S., Haley, E., Grotta, J., & Marler, J. (1994). Improved reliability of the NIH stroke scale using video training. NINDS TPA stroke study group. *Stroke*, 25, 2220–2226.
- Mar, J., Arrospe, A., & Comas, M. (2010). Budget impact analysis of thrombolysis for stroke in Spain: a discrete event simulation model. *Value in Health*, 13, 69–76.
- Mason, J. E., Denton, B. T., Shah, N. D., & Smith, S. A. (2014). Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients. *European Journal of Operational Research*, 233, 727–738.
- Meretoja, A., Keshtkaran, M., Saver, J., Tatlisumak, T., Parsons, M., Kaste, M., Davis, S., Donnan, G., & Churilov, L. (2014). Stroke thrombolysis save a minute, save a day. *Stroke*, 45, 1053–1058.
- Meretoja, A., Strbian, D., Mustanoja, S., Tatlisumak, T., Lindsberg, P. J., & Kaste, M. (2012). Reducing in-hospital delay to 20 minutes in stroke thrombolysis. *Neurology*, 79, 306–313.
- Miller, D. R., Butler, G., & Bramall, L. (1976). Validation of ecological systems models. *Journal of Environmental Management*, 4, 383–401.
- Monks, T., Pitt, M., Stein, K., & James, M. (2012). Maximizing the population benefit from thrombolysis in acute ischemic stroke: a modeling study of in-hospital delays. *Stroke*, 43, 2706–2711.
- Monks, T., Pitt, M., Stein, K., & James, M. A. (2014). Hyperacute stroke care and NHS England's business plan. *BMJ: British Medical Journal*, 348, 1–2.
- Murray, C. J. (1996). Rethinking DALYs. *The global burden of disease*, 1, 1–98.
- Myers, G. J. (1978). A controlled experiment in program testing and code walk-throughs/inspections. *Communications of the ACM*, 21, 760–768.
- Myers, G. J., Sandler, C., & Badgett, T. (2011). *The art of software testing*. John Wiley & Sons.
- Nord, E. (1992). Methods for quality adjustment of life years. *Social Science and Medicine*, 34, 559–569.
- Oakley, J. E., & O'Hagan, A. (2004). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66, 751–769.
- Oberkampf, W. L., & Roy, C. J. (2010). *Verification and validation in scientific computing*. Cambridge University Press.
- Oral, M., & Kettani, O. (1993). The facets of the modeling and validation process in operations research. *European Journal of Operational Research*, 66, 216–234.
- Pidd, M. (2003). *Tools for thinking*. Wiley Chichester.
- Pidd, M. (2010). Why modelling and model use matter. *Journal of the Operational Research Society*, 61, 14–24.
- Pitt, M., Monks, T., Agarwal, P., Worthington, D., Ford, G. A., Lees, K. R., Stein, K., & James, M. A. (2012). Will delays in treatment jeopardize the population benefit from extending the time window for stroke thrombolysis? *Stroke*, 43, 2992–2997.
- Quade, E. S. (1980). Pitfalls in formulation and modeling. *Pitfalls of Analysis*, 8, 23–43.
- Rankin, J. (1957). Cerebral vascular accidents in patients over the age of 60. II. Prognosis. *Scottish Medical Journal*, 2, 200–215.
- Repede, J. F., & Bernardo, J. J. (1994). Developing and validating a decision support system for locating emergency medical vehicles in Louisville, Kentucky. *European Journal of Operational Research*, 75, 567–581.
- Robinson, S. (2002). General concepts of quality for discrete-event simulation. *European Journal of Operational Research*, 138, 103–117.
- Rothman, K. J., Greenland, S., & Lash, T. L. (2008). *Modern epidemiology*. Wolters Kluwer Health.
- Sargent, R. G. (1984). A tutorial on verification and validation of simulation models. In S. Sheppard, U. Pooch, & D. Pegden (Eds.), *Proceedings of the 16th conference on winter simulation* (pp. 114–121). Piscataway, USA: IEEE Press.
- Sargent, R. G. (1986). The use of graphical models in model validation. In J. Wilson, J. Henriksen, & S. Roberts (Eds.), *Proceedings of the 18th Conference on Winter Simulation* (pp. 237–241). Washington D.C., USA: ACM.
- Sargent, R. G. (1996). Verifying and validating simulation models. In J. M. Charnes, D. J. Morrice, D. T. Brunner, & J. J. Swain (Eds.), *Proceedings of the 28th Conference on Winter Simulation* (pp. 55–64). New Orleans, USA: IEEE Computer Society.
- Sargent, R. G. (2001). Some approaches and paradigms for verifying and validating simulation models. In B. A. Peters, J. S. Smith, D. J. Medeiros, & M. W. Rohrer (Eds.), *Proceedings of the 2001 Winter Simulation Conference: Vol. 1* (pp. 106–114). IEEE.
- Sargent, R. G. (2013). Verification and validation of simulation models. *Journal of Simulation*, 7, 12–24.
- Schellenberger, R. E. (1974). Criteria for assessing model validity for managerial purposes. *Decision Sciences*, 5, 644–653.
- Schlesinger, S., Crosbie, R. E., Gagné, R. E., Innis, G. S., Lalwani, C., Loch, J., Sylvester, R. J., Wright, R. D., Kheir, N., & Bartos, D. (1979). Terminology for model credibility. *Simulation*, 32, 103–104.
- Schruben, L. W. (1980). Establishing the credibility of simulations. *Simulation*, 34, 101–105.
- Seaman, A.M. (2014). Shorter stroke treatment delays tied to more healthy days. Retrieved 12 September, 2014, from <http://www.reuters.com/article/2014/03/19/us-shorter-stroke-idUSBREA2120Z20140319>.
- Simpson, M. A., Dewey, H. M., Churilov, L., Ahmed, N., Bladin, C. F., Schultz, D., Markus, R., Sturm, J. W., Levi, C. R., & Blacker, D. J. (2010). Thrombolysis for acute stroke in Australia: outcomes from the safe implementation of thrombolysis in stroke registry (2002–2008). *Medical Journal of Australia*, 193, 439–443.
- Sun, Herald (2014). Faster treatment, better future for stroke victims. Retrieved 15 September, 2014, from <http://www.heraldsun.com.au/news/victoria/faster-treatment-better-future-for-stroke-victims/story-fni0ft3-1226986227218>.
- Whipple, T. (2014). Why every second counts after a stroke. Retrieved 12 September, 2014, from <http://www.thetimes.co.uk/tto/health/news/article4032754.ece>.
- Whitner, R. B., & Balci, O. (1989). Guidelines for selecting and using simulation model verification techniques. In *Proceedings of the 21st conference on winter simulation* (pp. 559–568). ACM.
- World Health Organization. (2014). Health statistics and information systems. Retrieved June 17, 2014, from http://www.who.int/healthinfo/global_burden_disease/estimates/en/index2.html.
- Yang, J., Chen, L., Chitkara, N., & Xu, Q. (2014). A Markov model to compare the long-term effect of aspirin, clopidogrel and clopidogrel plus aspirin on prevention of recurrent ischemic stroke due to intracranial artery stenosis. *Neurology India*, 62(1), 48–52.
- Yourdon, E. (1979). *Structured walkthroughs*. Prentice Hall PTR.
- Zanakis, S. H., Alvarez, C., & Li, V. (2007). Socio-economic determinants of HIV/AIDS pandemic and nations efficiencies. *European Journal of Operational Research*, 176, 1811–1838.