Unremarked or Unperformed? Systematic Review on Reporting of Validation Efforts of Health Economic Decision Models in Seasonal Influenza and Early Breast Cancer

Pieter T. de Boer¹ · Geert W. J. Frederix² · Talitha L. Feenstra³,⁴ · Pepijn Vemer¹,³

Abstract

Background  Transparent reporting of validation efforts of health economic models give stakeholders better insight into the credibility of model outcomes. In this study we reviewed recently published studies on seasonal influenza and early breast cancer in order to gain insight into the reporting of model validation efforts in the overall health economic literature.

Methods  A literature search was performed in Pubmed and Embase to retrieve health economic modelling studies published between 2008 and 2014. Reporting on model validation was evaluated by checking for the word validation, and by using AdViSHE (Assessment of the Validation Status of Health Economic decision models), a tool containing a structured list of relevant items for validation. Additionally, we contacted corresponding authors to ask whether more validation efforts were performed other than those reported in the manuscripts.

Results  A total of 53 studies on seasonal influenza and 41 studies on early breast cancer were included in our review. The word validation was used in 16 studies (30 %) on seasonal influenza and 23 studies (56 %) on early breast cancer; however, in a minority of studies, this referred to a model validation technique. Fifty-seven percent of seasonal influenza studies and 71 % of early breast cancer studies reported one or more validation techniques. Cross-validation of study outcomes was found most often. A limited number of studies reported on model validation efforts, although good examples were identified. Author comments indicated that more validation techniques were performed than those reported in the manuscripts.

Conclusions  Although validation is deemed important by many researchers, this is not reflected in the reporting habits of health economic modelling studies. Systematic reporting of validation efforts would be desirable to further enhance decision makers’ confidence in health economic models and their outcomes.

Electronic supplementary material  The online version of this article (doi:10.1007/s40273-016-0410-3) contains supplementary material, which is available to authorized users.

© The Author(s) 2016. This article is published with open access at Springerlink.com
Key Points for Decision Makers

All stakeholders have a vested interest in a high validation status of health economic models since they play an important role in the economic evaluation of therapeutic interventions. Transparent reporting of validation efforts and their outcomes will allow stakeholders to make their own judgment of a model’s validation status.

Only a limited number of studies reported on validation efforts, although good examples were identified. To further increase transparency, more explicit and structured attention to the reporting of validation efforts by authors and journals seems worthwhile.

1 Introduction

Health economic decision analytic models play an important role in the economic evaluation of therapeutic interventions [1]. Since policy decisions are influenced by the results of such models, all stakeholders have a vested interest in a high validation status of these models. Transparent reporting of validation efforts and their outcomes will give the stakeholder better insight into the model’s credibility (is the model scientifically sound?), salience (is the model applicable within the context?) and legitimacy (are all stakeholder concerns, values and views included properly?) [2, 3]. Proper information regarding these aspects allows stakeholders to make their own judgement of the models’ validation status.

Several systematic reviews of health economic evaluations in different disease areas indicated that little was reported on model validation [4–10]; however, most of these reviews were not focused on the general quality of modelling aspects and contained little details on model validation performances. Only one review, focusing on interventions on cardiovascular diseases, provided a clear overview on which part of the included studies reported on model validation tests distinguishing model validation techniques according to the International Society for Pharmacoeconomics and Outcomes Research–Society for Medical Decision Making (ISPOR–SMDM) guidelines [9]. However, modelling evaluation processes might vary between different disease areas, therefore more studies assessing model validation efforts are needed.

In this study we aimed to systematically review the reporting of validation efforts of recently published health economic decision models, explicitly distinguishing between different validation techniques. For this purpose, we chose two example diseases, namely seasonal influenza (SI) and early breast cancer (EBC). These two diseases are well-defined and by choosing both a communicable disease, which is often modelled using dynamic models [11], and a non-communicable disease, which is often modelled using static models, we expected to cover a wide range of model types [1]. This should provide a good overview of the current standard in the reporting of validation efforts in the health economic literature.

Since validation is an integral part of the modelling process (see, for example, Fig. 2 in Sargent [12]), low reporting of validation efforts does not have to mean that they were not performed. For instance, impromptu checking of bits of computer code while coding may not always be reported. In order to gain insight into the discrepancy between the performance and reporting of validation efforts, we also reached out to the corresponding author of each of the included papers in this review for comments.

2 Methods

2.1 Search Strategy and Study Selection

We searched the PubMed and Embase databases to identify studies focusing on the health economic evaluations of SI and EBC. The full search strings for both diseases can be found in Appendix 1 and contained free-text searching terms as well as exploded (Medical Subject Heading [MeSH]) terms. For both disease areas, the health economic evaluations had to meet the following criteria: (1) published in peer reviewed journals from January 2008 to December 2014; (2) presented results of costs as well as health effects; and (3) used a computer simulation model to generate these results. We also screened reference lists of selected articles. Review papers, meta-analyses, letters and non-full-text such as abstracts were excluded, and we restricted our selection to the English language. For SI, studies focusing only on pandemic influenza were excluded, as well as studies analyzing interventions against multiple infectious diseases without showing separated results for SI. For EBC, we excluded studies on metastatic breast cancer, breast cancer screening and diagnostic systems to stage breast cancer.

2.2 Study Characteristics

General characteristics of the studies extracted included year of publication, country, income level of the country according to the classification of the World Bank [13], funding source, type of intervention, type of evaluation and model type. For SI, studies that incorporated disease
transmission dynamics (i.e. from carrier/infected to a susceptible individual), or used discrete event simulation were categorised as dynamic models. In all other cases, the model was categorised as static.

2.3 Reporting of Model Validation Efforts

In this study, we defined validation as the act of evaluating whether a model is a proper and sufficient representation of the system it is intended to represent in view of an application, where ‘proper’ was defined as “the model is in accordance with what is known about the system” and ‘sufficient’ was defined as “the results can serve as a solid basis for decision making” [14]. We first searched the publication’s text and appendices for the word validation and its conjugate forms (valid*, verif*). When present, we reported the context in which the word was used. Then, the reporting of model validation efforts were systematically assessed, using the outline presented in the validation-assessment tool AdViSHE (Assessment of the Validation Status of Health Economic decision models) [15]. This tool was designed to provide model users with structured information regarding the validation status of health economic decision models, and therefore enables systematic extraction of the reporting of model validation efforts. An added advantage of this tool is that it explicitly presents clear definitions of validation techniques since there is little, if any, consensus on terminology in the validation literature [16]. An abbreviated form of the AdViSHE tool is shown in Table 1 and includes five validation categories, i.e. validation of the conceptual model (A, the theories and assumptions underlying the model concepts, and the model’s structure and causal relationships), input data (B, available input data and data transformations), computerised model (C, implemented software program, including code, mathematical calculations and implementation of the conceptual model), and operational model (D, behaviour of the model outcomes). Remaining validation techniques, such as, for instance, double programming, are assigned to category E. Assessment of studies on validation reporting was performed by two of the authors (PdB and PV for influenza, and PdB and GF for EBC) separately. After comparing results, differences between the two authors were resolved in a consensus meeting. Examples of each validation technique found were collected and presented.

2.4 Comments from Authors

The corresponding authors of the included studies were contacted by email in August 2015, followed by a reminder in November 2015. In this email, we explained the aim of our study, provided details of the corresponding author's

### Table 1 Validation aspects included in the AdViSHE validation status assessment tool [15]

<table>
<thead>
<tr>
<th>Validation categories</th>
<th>Subcategory</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(A) Conceptual model</td>
<td>A1 Face validity</td>
<td>Have experts been asked to judge the appropriateness of the conceptual model?</td>
</tr>
<tr>
<td></td>
<td>A2 Cross validity</td>
<td>Has this model been compared with other conceptual models found in the literature or clinical textbooks?</td>
</tr>
<tr>
<td>(B) Input data</td>
<td>B1 Face validity</td>
<td>Have experts been asked to judge the appropriateness of the input data?</td>
</tr>
<tr>
<td></td>
<td>B2 Model fit</td>
<td>When input parameters are based on regression models, have statistical tests been performed?</td>
</tr>
<tr>
<td>(C) Computerised model</td>
<td>C1 External review</td>
<td>Has the computerised model been examined by modelling experts?</td>
</tr>
<tr>
<td></td>
<td>C2 Extreme value testing</td>
<td>Has the model been run for specific, extreme sets of parameter values in order to detect any coding errors?</td>
</tr>
<tr>
<td></td>
<td>C3 Testing of traces</td>
<td>Have patients been tracked through the model to determine whether its logic is correct?</td>
</tr>
<tr>
<td></td>
<td>C4 Unit testing</td>
<td>Have individual submodules of the computerised model been tested?</td>
</tr>
<tr>
<td>(D) Operational model</td>
<td>D1 Face validity</td>
<td>Have experts been asked to judge the appropriateness of the model outcomes?</td>
</tr>
<tr>
<td></td>
<td>D2 Cross validity</td>
<td>Have the model outcomes been compared with the outcomes of other models that address similar problems?</td>
</tr>
<tr>
<td></td>
<td>D3 Alternative input</td>
<td>Have the model outcomes been compared with the outcomes obtained when using alternative input data?</td>
</tr>
<tr>
<td></td>
<td>D4 Empirical data</td>
<td>Have the model outcomes been compared with empirical data?</td>
</tr>
<tr>
<td></td>
<td>(A) Comparison against the data sources on which the model is based (dependent validation)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(B) Comparison against a data source that was not used to build the model (independent validation)</td>
<td></td>
</tr>
<tr>
<td>(E) Other techniques</td>
<td>–</td>
<td>Have any other validation techniques been performed?</td>
</tr>
</tbody>
</table>
paper that was included in our review, and enquired whether authors had performed validation efforts other than those reported in their manuscript. To provide help on validation techniques that could have been performed, we attached the AdViSHE tool to this email. Any answers from the authors were reported.

3 Results

3.1 Study Selection

The searches resulted in 53 SI studies [17–70] and 41 EBC studies [71–111] that were eligible for inclusion in our review. More details on the study selection process are shown in Fig. 1a, b.

3.2 Study Characteristics

General characteristics of the included studies are shown in Table 2. For both disease areas, most studies were performed in countries in North America, followed by countries in Europe and Asia. Studies were predominantly performed for high-income countries, although we also found studies for middle-income countries such as China, Taiwan and Argentina (SI), and China, Brazil, Colombia and Iran (EBC). We found no studies for low-income countries.

All SI studies analyzed pharmaceutical interventions, i.e. influenza vaccines or antiviral drugs. One study additionally assessed non-pharmaceutical mitigation strategies, including ventilation, face masks, hand washing and ultraviolet irradiation [21]. For EBC, antineoplastic drugs were predominantly studied, although we also included four studies on radiation or surgery treatments [72, 84, 100, 111].

SI was analyzed using static models in 43 (81 %) of the studies, mostly a decision tree model (36, 68 %) or a state transition model with Markov properties [‘Markov model’] (6, 11 %). One study used a multicohort model, in which cohorts of different ages were simultaneously followed over their lifetimes [64]. A total of eight studies (15 %) used a dynamic transmission model, with all compartmental models using an SIR structure [17, 21, 25, 28, 29, 50, 51, 54]. Such models divide the population between susceptible (S), infected (I) and recovered (R), and include a (often age-stratified) mixing pattern between different groups. One study did not elaborate on model structure [57], and another study called the model ‘spreadsheet based’, with no further information on model type or structure [70]. Thirty-seven EBC studies (90 %) used a Markov model [71–83, 85–95, 97–102, 104–107, 109–...
one study used a decision tree [96], and one study used a semi-Markov model [84]. Two studies did not provide information on model type, but either defined the model as a 'decision analytic model' [108] or stated that 'the model took a state-transition approach' [103].

Within the study selection process, multiple studies had the same first author. For SI, one author conducted ten studies [32–42], two authors conducted three studies [49–51, 67–69] and three authors conducted two studies [23, 24, 55, 56, 60, 61], while for EBC, one author conducted three studies [93–95] and three authors conducted two studies [80, 81, 103, 104, 109, 110].

3.3 Model Validation

3.3.1 Free-Text Search

For SI, the word ‘validation’ or its conjugates was found in 16 studies (30 %). The context in which ‘validation’ was used diverged widely. Three studies did not use validation in a model validation context [27, 52, 53]; two studies mentioned that the evidence level of some input data was low and not validated [47, 57]; one study stated that picking a starting date for the simulation between two influenza seasons would be useful “to demonstrate model validity”, but did not specify how this was the case [48]; two studies used the word ‘validation’ in a context that might be linked to a validation technique, but did not provide information on which parts of the model were validated, by whom, or which techniques were used [44, 64]; and three studies stated that a previously validated model was used, without stating whether the model would be valid for the new purpose [56, 60, 61]. Consulting the prior publications these studies were based on did not provide further clarification on validation efforts performed. In seven studies (13 %), the word ‘validation’ was used in such a way that we were able to link this directly to a validation technique [17, 20, 25, 30, 48, 62, 64]; these are discussed in the next paragraph.

For EBC, 23 studies (56 %) reported the word ‘validation’ or its conjugate forms, also in various contexts. In four studies we found ‘validation’ was not related to validation techniques of the health economic model [73, 78, 101, 106]. A fifth study debated “[t]he validity of the assumption of differences in effectiveness with letrozole and anastrozole” [91], while a sixth study mentioned that “[n]o relevant cost and/or utility data were identified against which the model’s outputs could be validated” [88]. In two studies by the same first author it was stated that “[t]he validity of the model is presented using cost effectiveness-acceptance curves” [93, 95]. A total of 16 (39 %) of the included EBC studies used the word ‘validation’ in a context linked to a model validation technique, which will be addressed below [71, 72, 74, 76, 77, 80, 82, 84, 85, 87, 88, 92, 100, 104, 109, 110].

3.3.2 Validation Techniques

We identified 30 studies (57 %) on SI that reported one or more validation techniques, and 28 (68 %) EBC studies. Two or more validation techniques were found in five

<p>| Table 2 General statistics of the studies included in the review |
|----------------------------------------|----------|----------|</p>
<table>
<thead>
<tr>
<th>Study characteristic</th>
<th>Seasonal influenza</th>
<th>Early breast cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total studies included</td>
<td>53</td>
<td>41</td>
</tr>
<tr>
<td>Region</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Europe</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>North America</td>
<td>34</td>
<td>16</td>
</tr>
<tr>
<td>Asia</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>South America</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Australia</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Income level</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>Middle</td>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>Low</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Funding</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Public health sources</td>
<td>28</td>
<td>17</td>
</tr>
<tr>
<td>Industry</td>
<td>15</td>
<td>17</td>
</tr>
<tr>
<td>No external funding</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Not stated</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>Type of intervention</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharmaceutical-related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Vaccine</td>
<td>45</td>
<td>–</td>
</tr>
<tr>
<td>Drug</td>
<td>8</td>
<td>15</td>
</tr>
<tr>
<td>Non-pharmaceutical related</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Personal protection</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Radiation</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Surgical intervention</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>Type of evaluation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost-benefit</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Cost-effectiveness</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>Cost-utility</td>
<td>41</td>
<td>37</td>
</tr>
<tr>
<td>Model type</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Dynamic</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Compartmental model</td>
<td>15</td>
<td>1</td>
</tr>
<tr>
<td>Semi-Markov</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Static</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Decision tree</td>
<td>36</td>
<td>68</td>
</tr>
<tr>
<td>Multicohort</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Markov</td>
<td>6</td>
<td>37</td>
</tr>
<tr>
<td>Not stated</td>
<td>2</td>
<td>4</td>
</tr>
</tbody>
</table>

* Multiple interventions might be studied in one study
studies (9%) of SI and 15 studies (37%) on EBC. Figure 2 shows the model validation performance stratified by (sub)category.

Five studies reported on validation of the conceptual model (category A). Four EBC studies reported on face validity of (part of) the conceptual model [74, 85, 88, 92] (A1). For example, Au et al. [74] reported that “validation of [treatment] strategies were achieved by consensus of a Canadian panel of breast cancer oncologists”, including the identities of the concerned oncologists. Hall et al. [85] reported that “[t]he structure of the model was developed by consensus between clinical experts, health economists and medical statisticians”, and reported the background as well as selection procedure. This study also performed cross-validation of the conceptual model (A2) by performing a systematic review to identify all previously published models of the same intervention and subsequently comparing the conceptual model. We also found cross-validation of the conceptual model in one other study on EBC [88] and one study on SI [47].

Reporting on validation of input data (category B) was found in nine studies. Face validity (B1) was described in two SI studies [20, 62]. For example, Tarride et al. [62] mentioned that “144 Canadian physicians were surveyed to validate [complication rate] estimates for children aged 2-5 years old”. Testing of the model fit (B2) was addressed in one study on SI [30] and six studies on EBC [77, 81, 92, 94, 98, 107]. Jit et al. [30] performed multiple linear regressions to estimate the proportion of hospitalisations that was caused by influenza and provided details on the goodness of fit (R²) of the model. They also indicated that different regression models to estimate the proportion of healthcare attendances related to influenza gave similar outcomes, which provided internal validation of this input into the health economic model. Purmonen et al. [98] used different parametric survival models (Weibull, exponential, log-logistic) to optimally fit the trace of the curves; however, how they decided on optimal fit was not reported.

Validation of the computerised model (category C) was not found in any of the reviewed articles.

A total of 56 studies reported on validation of the operational model (category D). Cross-validation of the results (D2) was the validation technique found most often in SI (57%) as well as EBC (51%). For example, Chit et al. [22] performed an extensive comparison of the number of influenza cases and various other clinical outcomes with data published in a model from the Centers for Disease Control and Prevention. Next to cross-validation, validation of model outcomes to empirical data (D4A) was often found, namely in 20 studies. Three studies on SI reported on independent validation (D4B) by validating the predicted incidence of a specified influenza-related clinical outcomes.

Fig. 2 Model validation performances of 53 studies on seasonal influenza, as reported in publications, using the classification of the AdViSHE tool. Numbers above the bars represent absolute numbers

![Advishe Classification](https://example.com/advishe_classification.png)
event against data from a national surveillance system or a national registration agency [17, 25, 48]. Seventeen EBC studies performed dependent validation (D4A) by comparing the results with data of the clinical trial the study was based on. Two different EBC studies performed validation against independent data sources, namely against the ‘Adjuvant! Online’ prediction tool [84, 100], an online tool used in the US to help oncologists estimate the risks of mortality and side effects, given different clinical and treatment scenarios [112]. Face validity testing of the results (D1) was not reported in either of the disease areas, and validation by using alternative input data (D3) was found in one study [47]. This study used all parameters of a similar study analyzing the same intervention in the same country to compare the results.

Finally, two studies reported validation techniques that were not categorised in the AdViSHE tool (Category E) [64, 81]. Van Bellinghen et al. [64] performed double programming by programming one cohort in another software package, and Delea et al. [81] conducted an extensive comparison of the input data compared with the input data of other models, which might be regarded as cross-validation of the input data.

3.4 Comments from Authors

We reached out twice to the corresponding authors to ask whether, in practice, more validation efforts were performed than those reported in the manuscript. We were able to reach 77/94 corresponding authors, and of these we received a total of ten responses. Three responding authors informed us that they were not able to answer the enquiry due to various reasons, such as no access to project files anymore, study was conducted a long time ago, or current workload was too high. Comments from the remaining seven responding authors are shown below.

A first author response of an SI study included that, additionally to what had been reported in the manuscript, the complete model was double programmed (category E) using a different software package, and results were compared. When modifications were completed, the affected modules were checked for face validity by another programmer (C1) and run against test data to ensure consistency (D3). A second author response of an SI study reported that face validity of the conceptual model (A1), input data (B1) and model outcomes (D1) were assessed internally and externally by two different panels of independent researchers. Moreover, tests such as likelihood ratio testing, Akaike information criterion (AIC), Bayesian information criterion (BIC) and goodness of fit (B2) were performed to select the optimum regression model to determine the attributable fraction of influenza within surveillance data of ‘influenza-like illness’. The authors indicated that all input data have been varied outside their ranges to detect coding errors (C2), and mentioned that this was not reported in the manuscript since this was considered a natural part of model development. A third author response of a SI study indicated that face validity of the conceptual model (A1), input data (B1) and outcomes (D1) were tested within the team containing clinical experts of different fields. Moreover, the entire model was double programmed (E). Additionally, the authors indicated that validation of the computerised model (C) was not reported as this was considered standard procedure.

Concerning EBC, a first author response mentioned that the majority of the validation techniques that are found on the AdViSHE tool were performed. The model was constructed to the standards of the National Institute for Health and Care Excellence (NICE) guideline on technical appraisal, which, according to the author, in turn implies a certain standard of validation. A second author response indicated that not all validation performances were reported due to a very restrictive word limit of the journal and the audience of the journal being mainly clinical. However, the structure of the model was reviewed by clinical and health economic experts within and outside the team (A1). In addition, a quality control on input data (B1) and model programming (C1) was performed by health economists not involved in the original model design. Additionally, they attempted to perform cross-validity of the model outcomes (D2) but were unable to do so due to a lack of suitable comparison studies. A third author responded that face validity of model structure (A1), input parameters (B1) and code checking (C1) were performed, and that links between different submodules were also tested (C4). A third author indicated that no other additional validation efforts than those described in the manuscript were performed.

4 Discussion

In this study, we assessed the reporting of model validation efforts in the disease areas of SI and EBC within the period 2008 to 2014. Overall, reporting of model validation efforts was found to be limited. Reviewing the papers systematically using the AdViSHE tool, demonstrated that 57% of the studies on SI and 71% of the EBC models performed at least one validation technique; however, only 9 and 37% of studies on SI and EBC, respectively, performed two or more validation techniques. A limited number of author’s responses to our enquiry on model validation efforts performed, indicated that, in practice, considerably more validation techniques might be used than those reported in the manuscripts, provided these few responders are representative of the majority who did not reply to our request for additional information.
The most performed validation technique was cross-validation of the model outcomes. A first explanation for this might be that many general guidelines for writing scientific papers state that the discussion section should include a comparison of the study outcomes with the existing literature (e.g., Hall [113]). Moreover, Eddy et al. [114] specifically name cross-validation as one of the five types of validation. Few reports were identified regarding validation of the conceptual model, and no reports regarding validation of the computerised model. As indicated in two author responses, validation techniques such as code checking and extreme value testing might be regarded as implicit in the model development process and were therefore not reported [12]. This might also partly explain why face validity of the input data and results was not often reported. Moreover, the peer-review process before publication might be regarded by some authors as a way of testing face validity. Another reason why little is reported on validation of the conceptual model might be that many studies of SI used a basic decision-tree model; however, even in case of such simple models, validation remains important. Conceptual model validation in that case might possibly be even more important since the choice of such a simple structure should be justified. A final explanation might be that the word count or the (clinical) audience of the journal might restrict authors on reporting of validation efforts.

In addition to simply describing the conduct of validation, it may be useful for model users to report what was done with the outcomes of the validation techniques. For instance, did the authors make any changes to (parts of) the model when faced with the validation outcomes? Such outcomes may emphasize the importance of validation. Unfortunately, none of the studies included in this review reported this aspect of model validation.

The main difference between SI and EBC was found in validation of the model outcomes by using empirical data. Dependent validation of the model outcomes was found in several studies of EBC but not in studies on SI. This may be due to the nature of SI, which, as a communicable disease, requires complex transmission dynamics and should therefore be studied on a population level rather than on a cohort level. For such dynamic SI models, understanding disease transmission dynamics is complex and the level of indirect protection caused by herd immunity is dependent on vaccine uptake levels. This complicates direct validation of model outcomes using randomised clinical trials of influenza vaccines, further enhanced by the variation of influenza activity by season or nation, and that the vaccine might not match the prevalent circulating strain. Independent validation of model outcomes to incidence data of national healthcare registries might therefore be more suitable compared with randomised clinical trial data, although the quality of influenza monitoring systems should then be taken into account. Such monitoring systems might not always be available in the studied countries.

We feel that simply mentioning that the model was previously validated does not guarantee a high validation status as new validation efforts are necessary when a model is used in a different setting or with different data (a ‘new application’). On the other hand, indicating that some particular input data was not validated due to a lack of suitable data sources was found to be useful for the reader as they can distinguish which parts of the model include high uncertainty [1]. Reporting that the model is validated according to the guidelines of ISPOR–SMDM Task Force on Good Research Practices–Modelling Studies [115], as was reported by Van Bellinghen et al. [64], or to the standards of the NICE, as was communicated through author comments, is insufficient. Although such guidelines give guidance on how model validation should be performed [115], these guidelines are general in nature and it is not clear which parts of validation have been performed, how and by whom. Therefore, simply following these guidelines does not guarantee that a model has a high enough validation status for its purpose, nor that model users can assess the validation status themselves.

Although a probabilistic sensitivity analysis is the most important technique to demonstrate the uncertainty around the model outcomes, using cost effectiveness acceptability curves to demonstrate validity of the results [93, 95] does not evaluate how accurately the model simulates what occurs in reality. The model that was applied by the author presenting with ten papers in our review was very similar in nine of these studies; however, no cross-validation of the conceptual model, or additional testing for variations, was described in any of these papers. For instance, the model type and structure were similar in most of the studies but were adapted because of a different target group, perspective or vaccination strategy. Explanations on these deviations or validation of their outcomes were lacking.

A positive example of reporting of model validation in the eyes of the authors was the study of Campbell et al. [77]. In this study, the underlying probability of a first recurrent breast cancer was based on a regression-based survival model that was externally validated against two online prediction tools: Nottingham Prognostic Index and Adjuvant! Online. A separate paper was devoted to the estimation and validation of this model [116]. Moreover, the web appendix contained an extensive report on external validation of the model used to estimate health-related quality-of-life during and after receiving chemotherapy.

Our finding that reporting of validation activities is limited was confirmed by other studies. Carrasco et al. [10] assessed the validation of dynamic transmission models
evaluating the epidemiology of pandemic influenza, and found that 16% of the compartmental models and 22% of the agent-based models reported on validation. As validation of model outcomes might be more difficult in cases where pandemic influenza is studied, reporting validation efforts of conceptual models or model inputs might be relatively more valuable. A study by Haji Ali Afzali et al. [9] reviewed the validation performances of 81 studies on therapeutic interventions for cardiovascular diseases, and found that 73% of the papers reported some form of model performance. The most executed form of validity was cross-validation (55%), similar to this study. Reporting of face validity (7%), internal validity (12%) and external validity (16%) was low. Although the review process was carried out using a different checklist and was therefore not completely comparable, these results at least support our findings that cross-validation is the most reported validation procedure, and reporting of other validation efforts is rare. Moreover, it demonstrates that limited validation documentation is not restricted to the disease areas of SI and EBC.

A strong point of this study is that we systematically assessed model validation performances. We looked at two disease areas, thereby covering a wider range of models than previous studies. Moreover, compared with previous studies analyzing reporting of validation efforts, we judged validation not only by technique but also by model aspect: conceptual model, input data, computerised model and the model outcomes. This made our findings more specific on which model aspects are generally validated and which are not. A final strong point is that we provided authors an opportunity to comment on whether more validation techniques were performed than those reported in the paper, which gave insight into the difference between performance and reporting of validation.

A limitation of our study was that for most studies we could only evaluate published validation efforts, rather than actual validation efforts undertaken. Thus, these models may have seemed less well-validated to the reader than they actually were. Although the responses of contacted authors of non-reported validation efforts were helpful, the response rate was low. Moreover, authors who performed more validation efforts might have been more aware of the importance of model validation and therefore more eager to respond to our enquiry. On the other hand, the poor response rate might also indicate that authors do not record which model validation tests were performed at the time the analysis was carried out. This was illustrated by three authors responses indicating that they were not able to provide additional information because the analysis was performed many years ago or because their current workload was too high. Next, we included ten papers on SI from the same first author, which might have had an effect on the total model validation performance within SI. Finally, although our search algorithm was extensive, we still may have missed publications that were not included in PubMed or Embase. However, the main focus of the current review was to give insight into present practice with regard to reporting of model validation, rather than a complete comprehensive overview of model-based publications in the fields of SI and EBC.

The main implication of our findings is that readers have no structured insight into the validation status of health economic models, which makes it difficult for them to evaluate the credibility of the model and its outcomes. In order to prevent making wrong decisions due to improper model validation status, readers might therefore be forced to perform validity checks themselves, which is highly inefficient. To date, we are not aware of any studies that have looked into the impact of the model’s validation status on the correctness of the model outcomes; however, we are aware of a case in The Netherlands in which the validation status of the health economic model can have a decisive effect on the reimbursement status of a drug. In this case, a vaccine against human papillomavirus was rejected for reimbursement because of a lack of model transparency and non-face-valid model inputs and model outcomes [117, 118].

Based on our results, we have several recommendations. First, better attention should be given to validation efforts in scientific publications. A more systematic use of model reporting guidelines might be useful [119, 120], possibly aided by reporting tools specifically aimed at validation efforts, such as AdViSHE [15]. In order to circumvent space limitations, inclusion of a small summary on model validation techniques in the Methods and Result sections would be desirable, in combination with a full model validation report in online appendices. Moreover, validation is important for all published health economic models, even if the model was validated for an earlier purpose. In addition, the choice of validation techniques reported deserves more attention and should be less guided by general publication guidelines, which now seem to imply undue attention for cross-validation only. Finally, it will be interesting to see whether the reporting of validation efforts will improve in time. A similar publication in a few years’ time will be very welcome.

5 Conclusions

Although validation is deemed important by many researchers, this is not reflected in the reporting habits of health economic modelling studies. A limited number of studies reported on model validation efforts, although good examples were identified. This lack of transparency might

△ Adis
reduce the credibility of study outcomes and hamper decision makers in interpreting and translating these study outcomes to policy decisions. Since authors have indicated that much more is undertaken than is reported, there is room for quick improvement of the reporting practices if stakeholders such as journals, editors and policy makers start explicitly requesting the results of the validation efforts. Therefore, systematic reporting of validation efforts would be desirable to further enhance decision makers' confidence in health economic models and their outcomes.

Acknowledgments The authors would like to thank Maiwenn Al, Isaac Corro Ramos, George van Voorn, and the AdViSHE Study Group for their work with the final two authors on model validation, which made this paper possible. They would also like to thank Richard Pitman for providing comments on the discussion of the manuscript, as well as the following corresponding authors for responding to our enquiry of providing additional information on model validation efforts: Adriaan Blommaert, Oliver Damm, Peter Hall, Michael Lux, Mattias Neyt, Richard Pitman, and Sorrel Wolowacz.

Author contributions Pieter T. de Boer designed and performed the literature search; Pieter T. De Boer and Pepijn Vemer both read and scored the influenza papers included in the full text analysis; Pieter T. De Boer and Geert W.J. Frederix both read and scored the early breast cancer papers included in the full text analysis; and Pieter T. de Boer drafted the manuscript, which was revised critically by Geert W.J. Frederix, Talitha L. Feenstra, and Pepijn Vemer. All authors read and approved the final text of the manuscript. A selection of the results has been presented at the ISPOR European Conference 2015 in Milan.

Compliance with Ethical Standards

Conflicts of interest Pieter T. de Boer and Geert W.J. Frederix have no conflicts of interest. Talitha L. Feenstra was project leader of the study developing AdViSHE. Pepijn Vemer was the first author of the paper presenting AdViSHE.

Funding No funding was received for performing this study.

Open Access This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (http://creativecommons.org/licenses/by-nc/4.0/), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References


