Health Economics, Policy and Law (2024), **19**, 174–191
doi:10.1017/S1744133123000087

HEALTH ECONOMICS, POLICY and LAW

ARTICLE

A review of heath economic evaluation practice in the Netherlands: are we moving forward?

Andrea Gabrio

Department of Methodology and Statistics, Faculty of Health Medicine and Life Science, Maastricht University, Maastricht, the Netherlands
Corresponding author. Email: a.gabrio@maastrichtuniversity.nl

(Received 18 October 2021; revised 27 October 2022; accepted 6 May 2023; first published online 6 June 2023)

Abstract

Economic evaluations have been increasingly conducted in different countries to aid national decision-making bodies in resource allocation problems based on current and prospective evidence on costs and effects data for a set of competing health care interventions. In 2016, the Dutch National Health Care Institute issued new guidelines that aggregated and updated previous recommendations on key elements for conducting economic evaluation. However, the impact on standard practice after the introduction of the guidelines in terms of design, methodology and reporting choices, is still uncertain. To assess this impact, we examine and compare key analysis components of economic evaluations conducted in the Netherlands before (2010–2015) and after (2016–2020) the introduction of the recent guidelines. We specifically focus on two aspects of the analysis that are crucial in determining the plausibility of the results: statistical methodology and missing data handling. Our review shows how, over the last period, many components of economic evaluations have changed in accordance with the new recommendations towards more transparent and advanced analytic approaches. However, potential limitations are identified in terms of the use of less advanced statistical software together with rarely satisfactory information to support the choice of missing data methods, especially in sensitivity analysis.

Keywords: Cost-effectiveness; economic evaluation; statistical methods; the Netherlands

1. Introduction

Health economics is a relatively new discipline focused on the evaluation of relevant health and cost evidence in order to systematically and rigorously inform decisions about the allocation of limited resources across a pool of alternative health care interventions within a given health care system. The first official adoption of health economics within a national public health care system is attributed to the Australian government (Australian Pharmaceutical Benefits Advisory Committee, 1992) in the early 1990s, and later followed by other public authorities in other countries (Hjelmgren *et al.*, 2001). Although the purpose of health economic evaluations remains the same across different jurisdictions, the presence of geographical and socio-cultural differences imposes national decision-making committees to define their own requirements and guidelines for economic evaluations (ISPOR, 2017).

In the Netherlands, the Dutch National Health Care Institute (Zorginstituut Nederland or ZIN) is the body in charge of issuing recommendations and guidance on good practice in health economic evaluations, not just for pharmaceutical products, but also in relation to other fields of application such as medical devices, long-term care and forensics. In 2016, ZIN issued an update on the guidance for health economic evaluations (Zorginstituut Nederland, 2016), which aggregated into a single document and revised three separately published guidelines for

pharmacoeconomics evaluation (Postma and Krabbe, 2006), outcomes research (Delwel, 2008) and costing manual (Hakkaart-van Roijen *et al.*, 2016). The novel aspects and future policy direction introduced by these guidelines have already been object of discussion, particularly with respect to the potential impact and concerns associated with their implementation in standard health economics practice in the Netherlands (Versteegh *et al.*, 2016a; Garattini and Padula, 2017). Given the importance covered by these guidelines, an assessment of their impact on economic evaluation practice is desirable.

The objective of this paper was to review the evolution of health economic evaluation practice in the Netherlands before and after the introduction of the ZIN's 2016 guidelines. Based on some key components within the health economics framework addressed by the new guidelines, we specifically focus on reviewing the statistical methods, missing data methods and software implemented by health economists. Given the intrinsic complexity of analysing health economics data, the choice of the analytical approaches to deal with these problems as well as transparent information on their implementation is crucial in determining the degree of confidence that decision-makers should have towards cost-effectiveness results obtained from these studies (Ramsey *et al.*, 2015).

The rest of the article is structured as follows. Section 2 briefly outlines the key elements of the ZIN's 2016 guidelines, with a focus on the changes that were introduced with respect to previous guidance. Section 3 presents the methodology and compares the components of the analysis of the included studies with the recommendations from the 2016 guidelines. Section 4 reviews the analytical methods and software used, while Section 5 focuses on the choice of missing data methods and uses a structured grading scheme to evaluate the studies based on the overall level of missingness information provided. Finally, Section 6 summarises our findings and recommendations for future research.

2. The ZIN 2016 guidelines

The main objective of the guidelines is to ensure the comparability and quality of health economic evaluations in the Netherlands, therefore facilitating the task of the decision-maker regarding the reimbursement of new health care interventions. Following the example of guidelines issued by decision-making bodies in other countries, including the National Institute for Health and Care Excellence (NICE) in the UK (NICE, 2013), the recommended features for economic evaluations are summarised in a typical scenario referred to as 'reference case', although deviations from it are allowed when properly justified.

Based on the structure of the reference case, four essential components of a health economic evaluation are identified: framework, analytic approach, input data and reporting. For the purpose of the review, we only focus on these components in the reference case as the main elements upon which evaluating health economics practice. For a thorough examination of the guidelines and advice that is specific to different types of health economic analyses we refer the interested reader to the original document (Zorginstituut Nederland, 2016) and two recent articles (Versteegh *et al.*, 2016a; Garattini and Padula, 2017).

2.1 Framework of the economic evaluation

The establishment of a framework for health economic evaluation is based on the identification of the perspective of the analysis, that is the definition of the users or actors to which such analysis is aimed at. According to the reference case, the societal perspective should always be adopted, which implies that the analysis should take into account all types of costs and benefits irrespective of who is the bearer/beneficiary (i.e. either the patients, private or public sector). Results from other perspectives may also be presented as additional analyses. The research question behind the analysis should be defined in terms of the PICOT (Patient, Intervention, Control, Outcome and Time) criteria and should involve: a population in the Dutch setting (P); a new

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

health care intervention (I) and standard of care (C) that can be applied in the Netherlands; predefined outcome measures (e.g. clinical, patient-reported); the expected lifetime of the target population (T). It is also recommended to 'scope' the PICOT criteria beforehand with the relevant stakeholders (e.g. patient organisations) to receive their feedback and refine the research question of the analysis (Zorgverzekeraars Nederland, 2015).

2.2 Analytic approach

The number and type of analytic techniques that should be implemented depend on the type of analysis considered. *Cost-effectiveness analysis* (CEA) and *cost-utility analysis* (CUA), respectively based on disease-specific (e.g. prevalence rates) or generic health measures (e.g. *quality-adjusted life years* or QALYs), are the most popular types of analyses. Between the two, CUA is generally the preferred choice as it allows the comparison of the results of the analysis across different disease areas and health conditions. *Discounting* should always be applied when data are analysed over a time horizon exceeding 1 year using a yearly discount rate of 1.5% for effects and 4% for costs. *Uncertainty* surrounding the economic results from the analysis should be assessed to: (1) quantify the impact on cost-effectiveness conclusions and (2) determine if and how much additional research should be conducted to reduce uncertainty around the results. The methods to assess uncertainty vary according to the type of analyses conducted, with a clear distinction between empirical (e.g. CUA alongside a trial) and model-based (e.g. simulation models) study designs.

In empirical analyses, statistical methods, such as regression approaches, should be used to correct for potential sources of bias and derive key cost-effectiveness estimates, including the mean incremental costs and effects between interventions and the ratio between these two quantities, known as the incremental cost-effectiveness ratio (ICER). Uncertainty around these estimates should be quantified with appropriate analytical methods that allow the computation of confidence intervals and standardised graphical tools such as cost-effectiveness planes or CEP (Black, 1990) and cost-effectiveness acceptability curves or CEAC (Van Hout *et al.*, 1994). Adequate statistical methods should also be used to quantify the impact of missing data uncertainty on the results, with multiple imputation or MI (Van Buuren, 2018) being the recommended approach. In model-based analyses, patient-level simulation methods should be used to evaluate cost-effectiveness results over a long-enough (typically lifetime) horizon and modelling assumptions should be varied across different scenarios to assess the robustness of the study conclusions, an exercise typically known as probabilistic sensitivity analysis or PSA (Claxton *et al.*, 2005). Value of information or VOI analysis (Claxton and Sculpher, 2006) can be used to assess the uncertainty around the model parameters associated, usually via appropriate quantitative measures, such has the expected value of perfect information or EVPI, and how it affects the choice of making an immediate decision or deferring the final decision of cost-effectiveness in favour of first collecting more evidence.

2.3 Input data

Input data on *clinical effectiveness* can be either directly collected from a study (empirical analyses) or retrieved from a systematic review of the literature (model-based analyses), preferably using evidence from randomised studies and head-to-head comparisons. All relevant costs should be identified and valuated, including those related to the health care system (direct and indirect medical costs), patient and family (e.g. travel, informal care), other sectors (e.g. volunteering) and productivity losses (e.g. due to absenteeism) in accordance with the guidance in the reference costing manual (Hakkaart-van Roijen *et al.*, 2016). Quality of life data should be collected by means of validated, generic quality-of-life self-reported questionnaires, with the EQ-5D-5L questionnaire (Janssen *et al.*, 2013) in combination with Dutch reference values (Versteegh *et al.*, 2016b) being the reference approach to derive patient-specific quality of life utility scores.

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

2.4 Reporting

Information related to input data should be reported in a transparent way. This includes, but is not limited to, details of studies used to retrieve effectiveness data (e.g. patient characteristics), prices and volumes of all cost components, questionnaires or valuation methods for quality of life data. For empirical studies, *missing data information* should be clearly reported in terms of amount and differences between individuals who completed and those who failed to complete the study. Results from the analysis should be reported separately for the base-case scenario and all additional uncertainty analyses in terms of total and incremental costs/effects, ICER estimates both via tabular form and graphical tools. Results of PSA (model-based) or bootstrapping (empirical) should be presented graphically via CEP and CEAC graphs, while results of VOI analysis should be presented using different reference values of the ICER.

3. Methods

We performed a bibliographic search in June 2021 using search engines of two online full-text journal repositories: (1) PubMed and (2) Zorginstituut. These sources were chosen to maximise the number of studies that could be accessed given the scoping nature of the review and the lack of a search strategy based on a pre-defined and rigid approach typical of systematic reviews. Articles were considered eligible for the review only if they were cost-effectiveness or cost-utility analyses targeting a Dutch population. To allow the inclusion of a reasonable amount of studies, the key words used in the search strategy were (cost-effectiveness OR cost-utility OR economic evaluation), and we targeted studies published between January 2016 and April 2021. However, to ensure the feasibility of the review, we excluded any published studies that were not written in English as well as any qualitative, phase I, methodological, evidence-synthesis, pilot and feasibility studies. A total of 4319 articles were identified, of which 3672 were either duplicates or satisfied at least one of the exclusion criteria and were thus discarded. After abstract review, 647 articles were considered, of which 190 fulfilled the eligibility criteria and were included in the analysis. We report a more detailed description of the inclusion and exclusion criteria used, data extraction and analysis strategy used in the Appendix, while the full list of reviewed studies is reported in the online Appendix.

3.1 Review

We present and compare the articles reviewed between two separate periods (2010–2015 and 2016–2020) to assess changes in standard health economics practice after the introduction of the ZIN's 2016 guidelines. We summarised key results in terms of the type of analysis and analytic approaches implemented. With regards to empirical analyses, we looked in detail at the statistical methods and software used, while also reviewing and evaluating the strategies implemented to handle missing data. Table 1 reports information about the reviewed studies, separately between by the two time periods, and compares it to the 2016 guidelines for each of the four key analysis components described in the reference case.

Out of the 190 studies, about half were published between 2010 and 2015 (96) and between 2016 and 2020 (94), with also comparable numbers in terms of empirical (86 vs 80) as well as model-based analyses (10 vs 14). In the Appendix, we give a visual representation of the sample size distribution based on the 166 empirical studies included in the review. Three major changes in health economics practice are observed between the two periods. First, there is a sensible increase in the proportion of studies adopting a societal perspective for the base-case analysis and a health care perspective in supplementary analyses (from 23 to 40%). Second, we observe an increase in the proportion of studies performing CUAs in the base-case analysis (from 31 to 44%) and a decrease in the number of base-case CEAs (from 30 to 17%). Third, there is an

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

Table 1. Descriptive information of the reviewed studies for the periods 2010–2015 and 2016–2020

Component	2010–2015 ($n = 96$) *n* (%)	2016–2020 ($n = 94$) *n* (%)
Perspective		
Societal	49 (51%)	41 (44%)
Health care/third party	24 (26%)	15 (16%)
Societal and health care	21 (23%)	37 (40%)
Unclear	2 (2%)	0
Analysis		
CUA	30 (31%)	41 (44%)
CEA	29 (30%)	16 (17%)
CUA and CEA	37 (39%)	37 (39%)
Design		
Empirical	86 (90%)	80 (83%)
Model-based	10 (10%)	14 (17%)
Type of technologies		
Pharmaceutical	11 (12%)	18 (20%)
Medical devices	10 (11%)	3 (3%)
Health care/public programmes	23 (25%)	12 (13%)
Medical/surgical procedures	9 (9%)	20 (21%)
Complex interventions	23 (25%)	30 (31%)
Other	11 (12%)	27 (15%)
Horizon		
<1 year*	25 (30%)	23 (29%)
1 year*	46 (53%)	41 (51%)
>1 year*	16 (17%)	15 (20%)
Lifetime**	7 (70%)	6 (43%)
Discounting (horizon >1 year)		
Relevant	26 (27%)	25 (27%)
4% costs and 1.5% effects	15 (58%)	17 (68%)
Costs		
Societal	51 (53%)	64 (68%)
Productivity losses (societal)		
Friction	20 (39%)	38 (59%)
Human capital	5 (10%)	6 (9%)
Friction and human capital	7 (14%)	8 (13%)
Unclear	14 (37%)	20 (19%)
Quality of life (CUA)		
EQ-5D-5L	3 (4%)	9 (12%)

(*Continued*)

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

Table 1. (*Continued.*)

Component	2010–2015 ($n = 96$) *n* (%)	2016–2020 ($n = 94$) *n* (%)
EQ-5D-3L	16 (24%)	26 (33%)
EQ-5D (unclear version)	34 (51%)	22 (28%)
Other	14 (21%)	21 (27%)
Uncertainty analysis		
CEP and CEAC	60 (63%)	62 (66%)
CEP	24 (25%)	14 (15%)
CEAC	9 (9%)	14 (15%)
None	3 (3%)	4 (4%)
Tornado diagram**	3 (30%)	7 (50%)
Value of information		
EVPI**	1 (10%)	1 (7%)

For each component of the economic evaluation, the approaches implemented are summarised and compared with the recommended approach from the 2016 guidelines (highlighted in bold). The asterisks * and ** denote components for which proportions are calculated out of the total number of empirical and model-based analyses, respectively.

uptake in the number of studies using societal costs (from 53 to 68%) and an increase in the proportion of CUAs which provide clear information on the EQ-5D questionnaires, for both 5L (from 4 to 12%) and 3L (from 24 to 33%) versions. Further noticeable changes include: an increase in the proportion of studies following the recent guidelines in regards to the choice of the discount rates for future effects and costs (from 58 to 68%) as well as the use of the friction method to calculate productivity losses (from 39 to 59%). In terms of types of technologies investigated, a considerable increase in the number of medical/surgical procedures (from 9 to 20%) and a decrease in the number of medical devices (from 11 to 3%) and health care/public programmes (from 25 to 13%) is observed. Finally, we observe that only one study within each period conducted VOI analysis and provided an estimate of EVPI. We note how these descriptive results should however be considered with care given the limited number of model-based studies included in the review which may lead to underrepresentation of specific components of the studies (e.g. pharmaceuticals typically assessed over a lifetime horizon within a model-based analysis).

4. Analytic approaches

In this section we explore in more detail the information provided by the reviewed studies in relation to the type of analytical approaches used to perform the economic evaluation and assess uncertainty. We also review information concerning the specific software program used as it may provide insights on practitioners' preferences of implementation. We specifically focus on the choice of the statistical approaches as it represents a crucial element in any health economic evaluation to determine the validity and reliability of cost-effectiveness conclusions.

4.1 Statistical methods

According to ZIN's 2016 guidelines and current literature, for empirical analyses, bootstrapping is the recommended approach to deal with non-normal distributions and quantify the level of uncertainty around the incremental mean cost and effect estimates (Campbell and Torgerson,

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

1999). Regression techniques are also important in order to obtain adjusted estimates and to control for potential imbalances in some baseline variables between treatment groups (Manca *et al.*, 2005a; Van Asselt *et al.*, 2009). Almost all reviewed empirical analyses used bootstrapping (95%), although the number of replications varied largely across the studies, with the most popular choices being 5000 (55%) followed by 2000 (29%). Studies showed even more variability in the choice of the methods used in combination with bootstrapping. Figure 1 shows the type of statistical techniques implemented among the 166 empirical analyses in our review.

Seven general classes of statistical approaches were identified, among which unadjusted methods were the most popular choice across both time periods. Regression-based adjustment methods were also widely used either in the form of: univariate regression adjustment (Manca *et al.*, 2005a); bivariate regression adjustment via Seemingly Unrelated Regression or SUR (Zellner and Huang, 1962); linear mixed modelling to account for clustering effects, e.g. in cluster randomised trials (Rice and Jones, 1997; Manca *et al.*, 2005b). Finally, delta adjustment (Vickers and Altman, 2001; Van Asselt *et al.*, 2009) or simulation methods were only rarely adopted.

A clear change in the type of statistical methods used between the two periods is denoted by a strong decrease (from 64 to 39) in the number of unadjusted analyses (red bars) in 2016–2020 compared to the earlier period, which is compensated by a rise in the number of adjusted analyses using either SUR (from 2 to 17) or linear mixed effects model (LMM) (from 4 to 10) methods (blue bars). Although these methods are not explicitly mentioned in the 2016 guidelines, the need to perform regression adjustment was clearly indicated as an important component in empirical analyses and both LMMs and SURs are widely used methods to achieve this goal (Willan *et al.*, 2004). No considerable changes between the two periods are observed with regards to the methods used for the calculation of bootstrapped confidence intervals, although only 53 studies (32%) provided information on the methods used. Among those providing such information, 29 (55%) applied bias-adjusted and accelerated methods (Efron and Tibshirani, 1994) and 24 (45%) applied standard percentile methods. No considerable differences between the two periods are observed for model-based analyses with Monte Carlo simulation methods (Briggs, 1999) being the approach used by all 24 model-based analyses (see Table 1) included in our review, almost exclusively in the form of Markov models (88%).

4.2 Software

We looked at the different type and combination of software programs used as an indication of the implementation preferences of analysts for health economic evaluations. Although in principle the choice of software should have no impact on the quality of the statistical methods implemented, it has been highlighted how use of simpler software (e.g. spreadsheet calculators such as Excel) may become increasingly cumbersome for matching more realistic and therefore complex modelling requirements (Baio and Heath, 2017; Incerti *et al.*, 2019). Indeed, although these tools may be sufficient for relatively simple analyses, when the complexity of the analysis demands for the development of more realistic and reproducible models, modern programming languages (e.g. R) can facilitate the implementation of the modelling task.

Since no considerable differences were observed when comparing software use over time, we present the results across the whole period, but divide them by type of analysis (empirical and model-based). Figure 2 shows a heatmap of the type of software used among the 166 empirical studies included in the review. Software programs are distinguished into 'main' and 'additional' categories according to the order (i.e. first mentioned) or tasks (i.e. base-case vs secondary analyses) for which they were used.

The most popular software was SPSS, chosen by 87 (52%) of the studies, either in the base-case (33%) or secondary (19%) analyses, often used in combination with Excel or by itself. When either STATA (26%) or R (13%) was used in the base-case analysis, SPSS was still the most popular choice in secondary analyses. Other combinations of software were less frequently chosen,

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

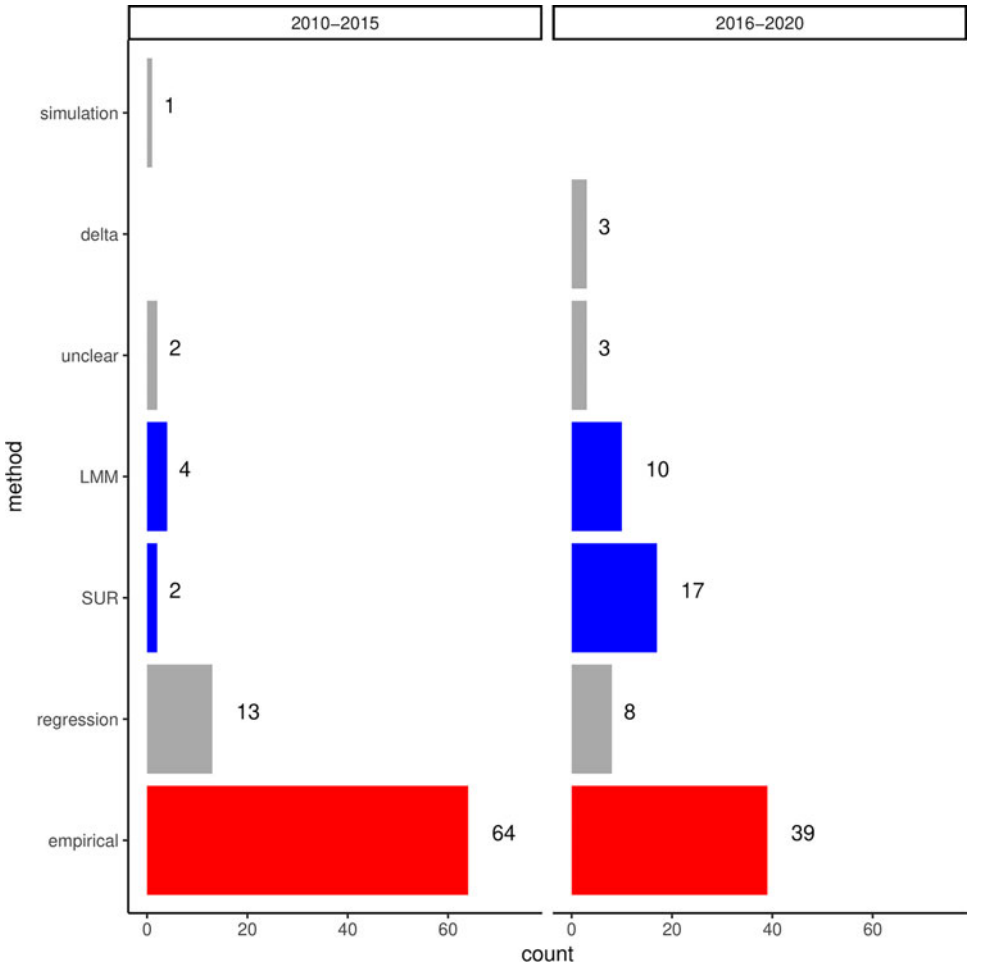

Figure 1. Bar chart of the number of empirical studies grouped by statistical methods implemented. Results are distinguished by the time period (2010–2015 and 2016–2020) and grouped using the following method's classes: simulation, delta method, unclear, LMM, SUR, regression and empirical.

even though 38 (23%) of the studies were unclear about the software implemented. Among the 24 model-based analyses, 14 (58%) did not provide any information in regards to the choice of software, while Excel alone was the most frequent software choice in 9 (38%) studies, followed by TreeAge with 2 (8%), and R, Delphi and SPSS with 1 (all <5%).

5. Missing data methods

The choice of the missing data methods can have a large impact on cost-effectiveness results and should be made in accordance with reasonable assumptions about the reasons behind the missing values. Since it is never possible to check assumptions about unobserved data, unless the amount of missing data is negligible (e.g. <5%), a principled approach to handle missingness is typically recommended. This amounts to perform the analysis under a benchmark missing data assumption (base-case analysis), and then assess the robustness of the base-case results to alternative assumptions using different methods [sensitivity analysis (SA)]. It is important that both base-case and sensitivity analyses rely on methods that are based on 'plausible' missingness

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

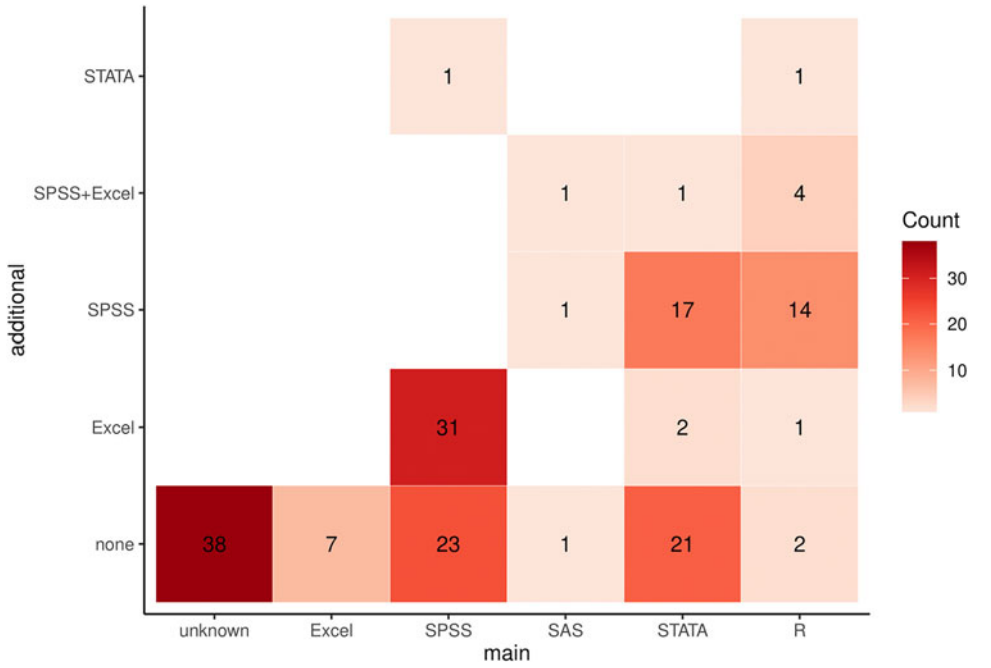

Figure 2. Heatmap of the type of software programs used among the empirical analyses. Software use is distinguished between main and additional analyses, defined according to the associated order or tasks that was specified in the information obtained from the studies. For each pairwise (main-additional) software combination, darker-coloured squares are associated with higher frequencies of use compared to lighter-coloured squares.

assumptions to ensure that the impact of missing data uncertainty is adequately quantified (Molenberghs and Kenward, 2007).

By their own nature missing data represent a crucial problem in empirical analyses but are less relevant in the context of model-based analyses. Within the second class of models, the long-term extrapolation of outcome data (e.g. survival beyond observed time horizon) represents a similar problem and is often accomplished through parametric or non-parametric methods. However, for the purpose of this review, we will exclusively focus on standard missing data methodology implemented in empirical analyses which represents the majority of the reviewed analyses.

5.1 Base-case and SA

We initially planned to report missing data information separately by effects and costs but, after reviewing the analyses, we noticed that only a small number of studies provided this level of detail. In the following, we will therefore provide results under the assumption that the same approaches were used to handle both missing effects and costs. In the Appendix, we report information about the number of studies that reported information about missingness rates as well as summary statistics (Table A1) and histograms (Figure A2) of the observed rates by type of outcome and time period.

Figure 3 shows, for both periods, a bubble plot for each combination of missing data methods implemented in the base-case and SA for empirical analyses, where the size of the bubbles indicates the frequency of use for each pairwise combination.

Across both periods limited changes are observed in terms of order of choice for missing data methods, with MI being the most popular base-case analysis, followed by complete case analysis (CCA), as the most popular SA choice. However, two noticeable variations in the frequency of these methods are observed between the two periods. First, the proportion of studies using MI

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

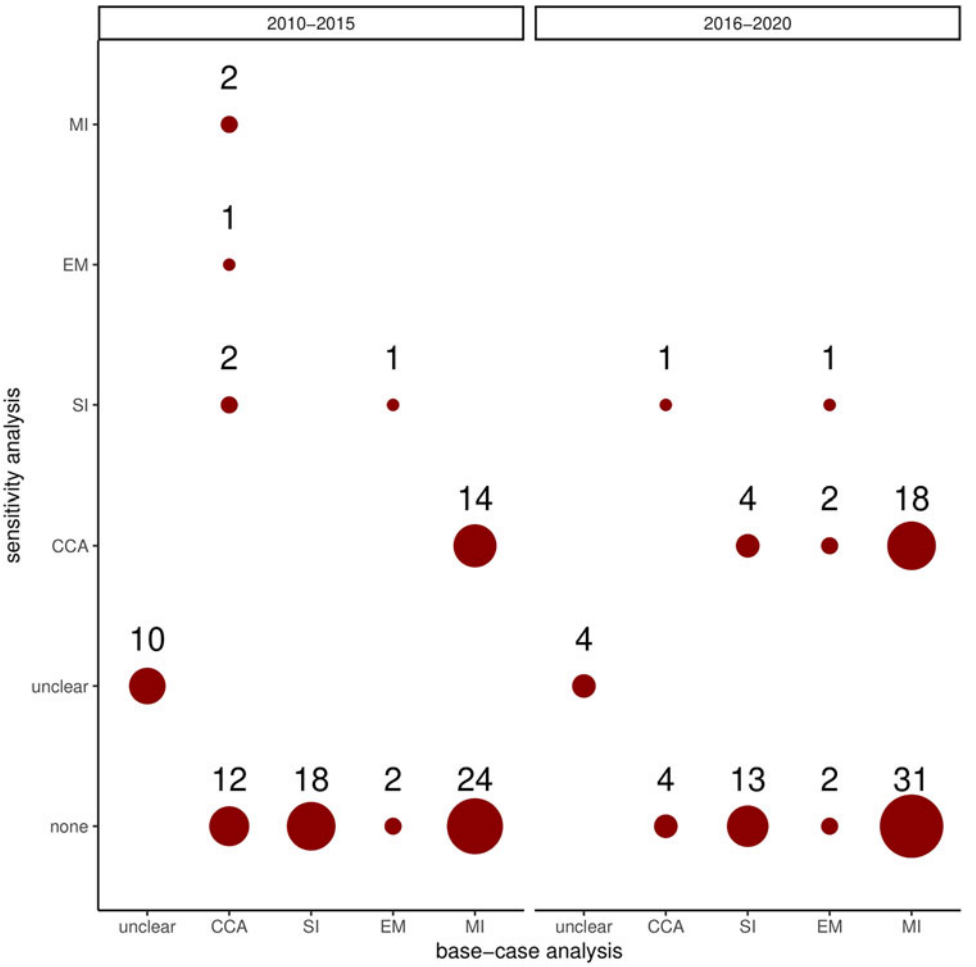

Figure 3. Bubble plot of the type of missing data methods used among the empirical studies. Results are distinguished by the time period (2010–2015 and 2016–2020) and grouped into six categories: no method (none); unclear (unclear); complete case analysis (CCA); single imputation (SI); expectation-maximisation (EM); multiple imputation (MI). Methods are distinguished according to whether they were used in the base-case or SA with the size of each bubble representing the frequency of use for each combination of base-case and SA missing data method.

in the base-case analysis has considerably increased over time (from 28 to 39%), which is compensated by a decrease in the proportion of less advanced methods such as CCA (from 14 to 5%) and single imputation (SI) (from 21 to 16%). Second, the number of studies not clearly reporting the methods has also considerably decreased (from 12 to 5%). The observed trend between the two periods may be the result of the specific recommendations from the 2016 guidelines in regards to the 'optimal' missing data strategy, resulting in a more frequent adoption of MI techniques and, at the same time, a less frequent use of CCA in the base-case analysis. However, in contrast to these guidelines, a large number of studies still does not perform any SA to missing data assumptions (about 65% in 2010–2015 and 63% in 2016–2020).

5.2 Quality of missing data information

We finally review the quality of the overall missing data information reported by the studies. We specifically rely on the quality evaluation scheme (QES), a structured reporting and analysis

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

system that embeds key guidelines for missing data handling for health economic evaluations (Gabrio *et al*., 2017). Detailed information about the rationale and structure of the scheme are provided in Gabrio *et al*. (2017), while here we only provide a concise explanation for clarity.

First, a numeric score is created to reflect the amount and type of information provided on three components characterising the missing data problem: description (e.g. number and pattern of missing data), method (e.g. type of method and detail of implementation) and limitations (e.g. limitations of assumptions). Each component is assigned a score weight (using a ratio 3:2:1) according to its importance, and then summed up to obtain an overall score for each study, ranging from 0 (no information) to 12 (full information). Next, grades are created by grouping the scores into ordered categories from A (highest score) to E (lowest scores). Finally, studies are also grouped by type of missingness method into five ordered classes, reflecting the strength of the underneath assumptions: unknown (UNK); SI; CCA; multiple imputation/expectation maximisation (MI/EM) and SA. We note that SA represents the less restrictive method as it requires studies to justify the assumptions explored in both base-case and SA based on the available information. Figure 4 shows a graphical representation of the quality scores (expressed in grades) in combination with the strength of assumptions (expressed by type of method) for each of the 80 empirical studies in the period 2016–2020. We specifically focus on studies in the later period as we want to assess current missing data practice (after the introduction of the 2016 guidelines).

Most of the studies lie in the middle and lower parts of the plot, and are associated with a limited (grades D and E) or sufficient (grade C) quality of information. However, only a few of these studies rely on very strong and unjustified missing data assumptions (red dots in the bottom-down part), while the majority provides either adequate justifications or uses methods associated with weak assumptions (green dots in the middle part). Only 11 (14%) studies are associated with both high-quality scores and less restrictive missingness assumptions (blue dots in the top-right part). No study was associated with either full information (grade A) or adequate justifications for the assumptions explored in base-case and SA.

6. Discussion

The objective of this paper was to review and compare the practice of conducting economic evaluation in the Netherlands before and after the introduction of the ZIN's 2016 guidelines. We focused on the type of analytic approaches and software used to conduct the analysis, while also examining the missing data methods and critically appraise the studies based on the overall information provided on missingness. It is important to highlight how, given the limited number of model-based studies examined in our review (13% of the total number of studies), any conclusions drawn about these studies is subject to considerable uncertainty and should be interpreted with caution.

6.1 Descriptive review

Descriptive information extracted from the reviewed studies (Table 1) provides some first insights about changes in practice in the years following the publication of the guidelines. First, a clear trend is observed towards an increase in the adoption of a societal and health care perspective and of CUA as the reference base-case analysis approach. Second, a similar increment is observed in the use of recommended instruments for the collection and valuation of health economic outcomes, such as EQ-5D-5L for QALYs and friction method for costs. Most of these changes are in accordance with the 2016 guidelines, which are likely to have played a role in influencing analysts and practitioners towards a clearer and more standardised way to report health economic results. Weaker trends are observed with respect to other analysis components, such as a stable reporting of cost-effectiveness results in terms of graphical tools (e.g. CEP and CEAC) and a limited adherence to the new guidelines in terms of VOI analysis or time horizon.

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

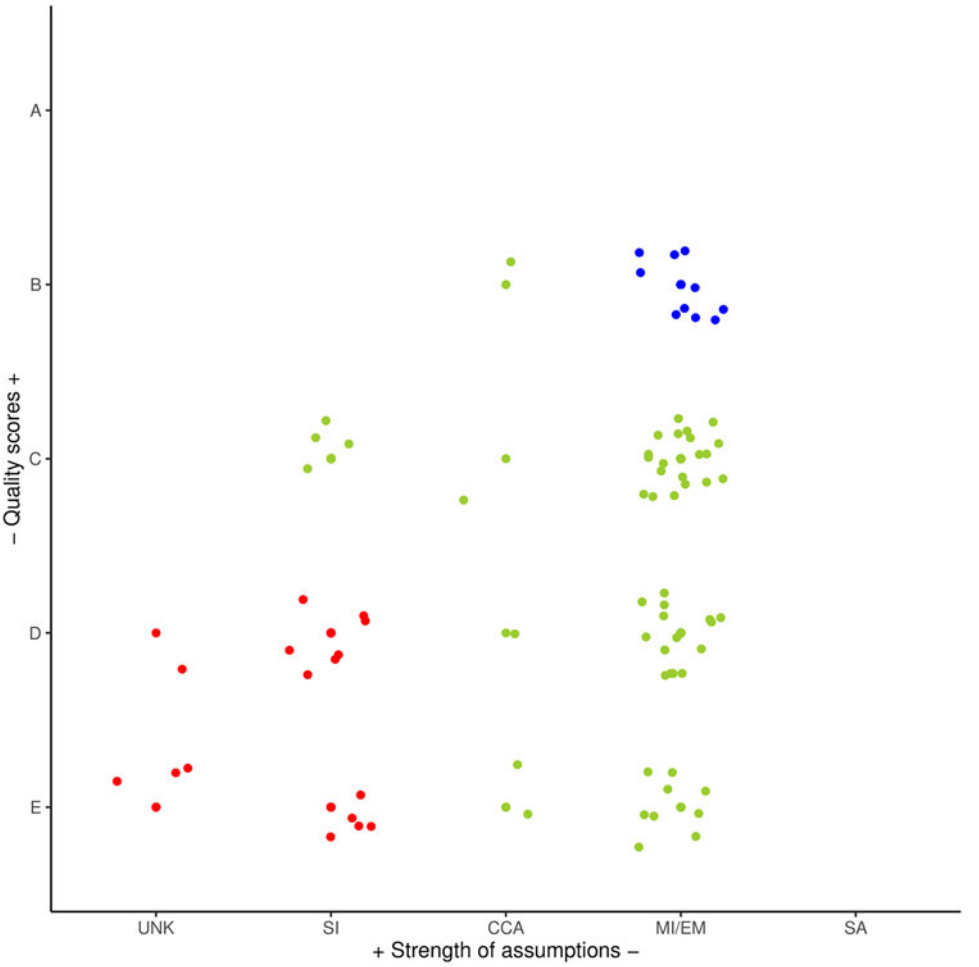

Figure 4. Jitter scatterplot for the joint assessment of the quality of missing data assumptions and information provided by empirical studies in the period 2016–2020. The strength of missing data assumptions is represented in terms of the type of methods used to handle missing values: unknown (UNK), single imputation (SI), complete case analysis (CCA), expectation-maximisation or multiple imputation (MI), sensitivity analysis (SA). The quality of the missing data information to support the method's assumptions is measured using the scores based on the QES and graded into the ordered categories: E, D, C, B, A.

6.2 Health economic analysis

When looking at the type of statistical methods used to perform the analysis, an important shift occurs between the two periods towards the use of methods that allow for regression adjustment, with a considerable uptake in the use of SURs and LMMs in the context of empirical analyses (Figure 1). These techniques are strongly supported by the 2016 guidelines in that they allow us to correct for potential bias due to confounding effects, deal with clustered data and formally take into account the correlation between costs and effects (Willan *et al.*, 2004). Bootstrapping remains the most popular methods to quantify uncertainty around parameter estimates across both periods. However, the health economic analysis framework requires that the level of complexity of the analysis model is reflected in the way uncertainty surrounding the estimates is generated. For example, if clustered data are handled by means of LMMs, then clustered bootstrap methods should be used to properly generate resampling draws. Among all reviewed studies,

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

we identified 13 cluster randomised trials of which only 10 took into account clustering at the analysis stage and, among these, only one study implemented clustered bootstrap methods.

We believe that these inconsistencies are due to either limited familiarity of practitioners with advanced statistical methods or potential limitations of the software used to conduct the analysis. This seems to be supported by the fact that a considerable amount of studies still rely on non-statistical software (e.g. Excel), or a combination of these and user-friendly statistical software (e.g. SPSS) to perform the analysis (Figure 2). Although this does not represent an issue *per se*, it may become problematic and potentially lead to difficult-to-spot errors when performing complex analyses without the use of more advanced and flexible software programs, such as R or STATA (Incerti *et al.*, 2019).

6.3 Missing data

The transition between the two time periods reveals an increase in the use of MI techniques in the base-case analysis together with a decrease in the overall use of CCA (Figure 3). This change is in line with the 2016 guidelines which warns about the inherent limitations and potential bias of simple methods (e.g. CCA) when compared to MI as the potential reference method to handle missing values. Nevertheless, improvements are still needed given that many studies (more than 6%) performed the analysis under a single missing data assumption. This is not ideal since by definition missing data assumptions can never be checked, making the results obtained under a specific method (i.e. assumption) potentially biased. For example, MI is often implemented under a missing at random or MAR assumption (i.e. missingness only depends on observed data). However, there is no way to test if MAR is appropriate and it is always possible that missingness depends on some unobserved quantities, corresponding to a so-called missing not at random assumption (Rubin, 2004). This is why SA has a crucial role in assessing the robustness of the results to a range of plausible departures from the benchmark missing data assumption of the base-case analysis (Daniels and Hogan, 2008). In principle, the choice of the assumptions to explore should be justified in light of the available information. However, in all reviewed studies, few reasonable justifications were provided to support the choice of the alternative methods used in SA. This is reflected by the relatively small number of studies providing full information about the missing data problem at hand (Figure 4).

7. Conclusions

Given the complexity of the health economics framework, the implementation of simple but likely inadequate analytic approaches may lead to imprecise cost-effectiveness results. This is a potentially serious issue for bodies such as ZIN in the Netherlands that use these evaluations in their decision making, thus possibly leading to incorrect policy decisions about the cost-effectiveness of new health care interventions. Our review shows, over time, a change in common practice with respect to different analysis components in accordance with the recent ZIN's 2016 guidelines. This is an encouraging movement towards the standardised use of more suitable and robust analytic methods in terms of both statistical, uncertainty and missing data analysis. Improvements are however still needed, particularly in the choice of adequate statistical techniques to deal with the complexity of the data analysed and in the assessment of the impact of alternative missing data assumptions on the results in SA.

It is important to highlight how the limited number of studies and the scoping nature of the review do not allow for in-depth causal speculations about the reasons behind the trends observed. For example, it is possible that changes between the two time periods are in part due to the influence of international trends in health economic evaluations rather than by only the 2016 guidelines. In addition, we chose the cut-off year of 2016 as a reasonable guess for assessing the impact following the publication of the guidelines but it is very likely that

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

some studies published after 2015 were not affected by the guidelines, which makes the task of disentangling the impact of the guidelines on current practice more difficult.

Regardless, the creation of unified guidelines for health economics practice represents a first key step in order to set common standards for assessing the quality of health economics evidence through a scientific and transparent evaluation process. Our review shows how, over time and especially in the years after the publication of the 2016 guidelines, standard practice in the Netherlands has moved towards the implementation of methods more consistent with those recommended by the ZIN. Finally, as the evidence from past experience from other countries suggests, e.g. such as NICE in the UK, the update and revision over time of the guidelines represents a powerful tool that can be used to reflect how changes in international trends and national policies can be translated into recommendations that can shape health economics practice.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/S1744133123000087

Conflict of interest. The author declares none.

References

Australian Pharmaceutical Benefits Advisory Committee (1992) Guidelines for the pharmaceutical industry on preparation of submissions to the pharmaceutical benefits advisory committee, including submissions involving economic analyses.

Baio G and Heath A (2017). When simple becomes complicated: why excel should lose its place at the top table.

Black WC (1990) The CE-plane: a graphic representation of cost-effectiveness. *Medical Decision Making* **10**, 212–214.

Briggs A (1999) Handling uncertainty in economic evaluation. *BMJ* **319**, 120.

Campbell MK and Torgerson DJ (1999) Bootstrapping: estimating confidence intervals for cost-effectiveness ratios. *QJM* **92**, 177–182.

Claxton KP and Sculpher MJ (2006) Using value of information analysis to prioritise health research. *PharmacoEconomics* **24**, 1055–1068.

Claxton K, Sculpher M, McCabe C, Briggs A, Akehurst R, Buxton M, Brazier J and O'Hagan T (2005) Probabilistic sensitivity analysis for nice technology assessment: not an optional extra. *Health Economics* **14**, 339–347.

Daniels MJ and Hogan JW (2008) *Missing Data in Longitudinal Studies: Strategies for Bayesian Modeling and Sensitivity Analysis*. Boca Raton, FL: CRC Press.

Delwel G (2008) *Guidance for outcomes research for the assessment of the cost-effectiveness of in-patient medicines*. Diemen, The Netherlands: Health Care Insurance Board. Report No.: 270: 28113706.

Efron B and Tibshirani RJ (1994) *An introduction to the Bootstrap*. Boca Raton, FL: CRC Press.

Gabrio A, Mason AJ and Baio G (2017) Handling missing data in within-trial cost-effectiveness analysis: a review with future recommendations. *PharmacoEconomics* **1**, 79–97.

Garattini L and Padula A (2017) Dutch guidelines for economic evaluation: 'from good to better' in theory but further away from pharmaceuticals in practice? *Journal of the Royal Society of Medicine* **110**, 98–103.

Garattini L and Padula A (2017) Dutch guidelines for economic evaluation:'from good to better'in theory but further away from pharmaceuticals in practice?. *Journal of the Royal Society of Medicine* **110**(3), 98–103.

Hakkaart-van Roijen L, van der Linden N, Bouwmans C, Kanters T, Tan SS and Kostenhandleiding S (2016) Methodologie van kostenonderzoek en referentieprijzen voor economische evaluaties in de gezondheidszorg. *Zorginstituut Nederland*, 1–120. Available from: www.zorginstituutnederland.nl/publicaties/publicatie/2016/02/29/richtlijn-voor-het-uitvoeren-vaneconomische-evaluaties-in-de-gezondheidszorg.

Hjelmgren J, Berggren F and Andersson F (2001) Health economic guidelines-similarities, differences and some implications. *Value in Health* **4**, 225–250.

Incerti D, Thom H, Baio G and Jansen JP (2019) R you still using Excel? The advantages of modern software tools for health technology assessment. *Value in Health* **22**, 575–579.

ISPOR (2017) Pharmacoeconomic guidelines around the World. Available at https://tools.ispor.org/peguidelines/ (accessed: 2021-07-23).

Janssen M, Pickard AS, Golicki D, Gudex C, Niewada M, Scalone L, Swinburn P and Busschbach J (2013) Measurement properties of the EQ-5D-5L compared to the EQ-5D-3L across eight patient groups: a multi-country study. *Quality of Life Research* **22**, 1717–1727.

Manca A, Hawkins N and Sculpher MJ (2005a) Estimating mean QALYs in trial based cost-effectiveness analysis: the importance of controlling for baseline utility. *Health Economics* **14**, 487–496.

Manca A, Rice N, Sculpher MJ and Briggs AH (2005b) Assessing generalisability by location in trial-based cost-effectiveness analysis: the use of multilevel models. *Health Economics* **14**, 471–485.

Molenberghs G and Kenward M (2007). *Missing Data in Clinical Studies*, vol. **61**. Chichester, West Sussex, England: John Wiley & Sons.
NICE (2013) Guide to the methods of technology appraisal 2013.
Postma MJ and Krabbe PFM (2006) Farmaco-economisch onderzoek: doelmatigheid van geneesmiddelen. *Geneesmiddelen Bulletin, Jaargang* **40**, 133–140.
Ramsey SD, Willke RJ, Glick H, Reed SD, Augustovski F, Jonsson B, Briggs A and Sullivan SD (2015) Cost-effectiveness analysis alongside clinical trials II – an ISPOR good research practices task force report. *Value in Health* **18**, 161–172.
Rice N and Jones A (1997) Multilevel models and health economics. *Health Economics* **6**, 561–575.
Rubin DB (2004) *Multiple Imputation for Nonresponse in Surveys*, vol. **81**. John Wiley & Sons.
Van Asselt AD, Van Mastrigt GA, Dirksen CD, Arntz A, Severens JL and Kessels AG (2009) How to deal with cost differences at baseline. *PharmacoEconomics* **27**, 519–528.
Van Buuren S (2018) *Flexible Imputation of Missing Data*. Boca Raton, FL: CRC Press.
Van Hout BA, Al MJ, Gordon GS and Rutten FF (1994) Costs, effects and C/E-ratios alongside a clinical trial. *Health Economics* **3**, 309–319.
Versteegh M, Knies S and Brouwer W (2016a) From good to better: new Dutch guidelines for economic evaluations in healthcare. *PharmacoEconomics* **34**, 1071–1074.
Versteegh MM, Vermeulen KM, Evers SM, De Wit GA, Prenger R and Stolk EA (2016b) Dutch tariff for the five-level version of EQ-5D. *Value in Health* **19**, 343–352.
Vickers AJ and Altman DG (2001) Analysing controlled trials with baseline and follow up measurements. *BMJ* **323**, 1123–1124.
Willan AR, Briggs AH and Hoch JS (2004) Regression methods for covariate adjustment and subgroup analysis for non-censored cost-effectiveness data. *Health Economics* **13**, 461–475.
Zellner A and Huang DS (1962) Further properties of efficient estimators for seemingly unrelated regression equations. *International Economic Review* **3**, 300–313.
Zorginstituut Nederland (2016) Guideline for economic evaluations in healthcare. Retrieved from https://english.zorginstituutnederland.nl/publications/reports/2016/06/16/guideline-foreconomic-evaluations-in-healthcare.
Zorgverzekeraars Nederland (2015) Beoordeling stand van de wetenschap en praktijk. https://www.zorginstituutnederland.nl/binaries/zinl/documenten/rapport/2015/01/15/beoordelingstand-van-de-wetenschap-en-raktijk/Beoordeling+stand+van+de+wetenschap+en+praktijk.pdf. Accessed 18-04-2019.

Appendix

A.1 Literature review methods

A.1.1 Inclusion and exclusion criteria

The eligibility of each study was assessed first by title and abstract and then by full text. We included studies that were identified as cost-effectiveness or cost-utility analyses, regardless of whether they were within-trial or model-based studies. We excluded review or qualitative studies, phase I trials, pilot studies and feasibility studies. We imposed the language limitation that only studies written in English could be reviewed.

A.1.2 Data extraction

A data extraction form was developed to record information from each included study. The information extracted included journal, year of publication, study design, type of intervention, type of economic evaluations (i.e. cost-effectiveness or cost-utility analysis, within-trial or model-based analysis), perspective, time horizon, data collection methods, the number and proportion of missing values, statistical analysis methods, missing data methods and type of software used. The data extraction tables were created using Microsoft Excel (version 16.42) and all the extracted information was summarised using R (version 4.0.3). Supplementary materials for each study were checked, and information of interest was recorded to supplement the original articles.

A.1.3 Analysis

The extracted information was compared across studies. First, we investigated the type of statistical approach used to estimate the mean incremental costs and effectiveness between treatment groups and the methods used to quantify the level of uncertainty around the estimates, separately by type of economic evaluation (i.e. empirical vs model-based analyses). Second, for the empirical analyses, we recorded the different types of software used to conduct the main analysis (main task) and to perform further tasks (additional task). Third, for empirical analyses, we assessed the extent of missing data, distinguishing between the type of missing data methods used. We focused on the proportion of missingness reported and grouped missing data methods into selected categories according to the types of missingness methods used across the reviewed studies.

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

A.1.4 Sample size distribution of reviewed studies

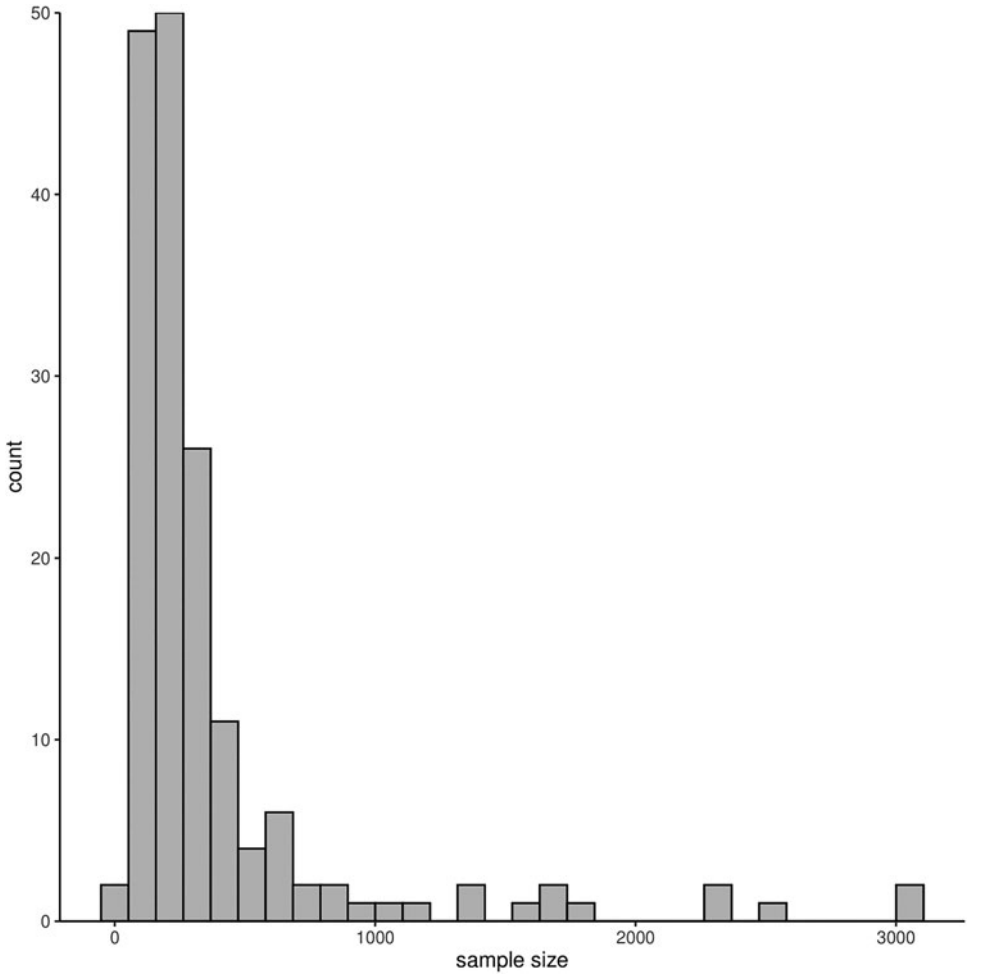

Figure A1. Histogram of the sample size distribution of the 166 empirical studies in included in the review.

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

A.1.5 Missing data rates of reviewed studies

Table A1. Number (proportion) of studies reporting information about missingness effect and cost rates as well as mean (standard deviation) of the observed missingness effect and cost rates based on the information reported by the studies, separately displayed by the two time periods included in the review (2020–2015 and 2016–2020) and across both periods

	2010–2015	2016–2020	Total
Studies	*n* (%)	*n* (%)	*n* (%)
Effects	50 (52%)	68 (72%)	118 (62%)
Costs	48 (50%)	63 (67%)	111 (58%)
Rates	Mean (SD)	Mean (SD)	Mean (SD)
Effects	0.31 (0.18)	0.27 (0.17)	0.29 (0.17)
Costs	0.33 (0.19)	0.27 (0.17)	0.29 (0.18)

https://doi.org/10.1017/S1744133123000087 Published online by Cambridge University Press

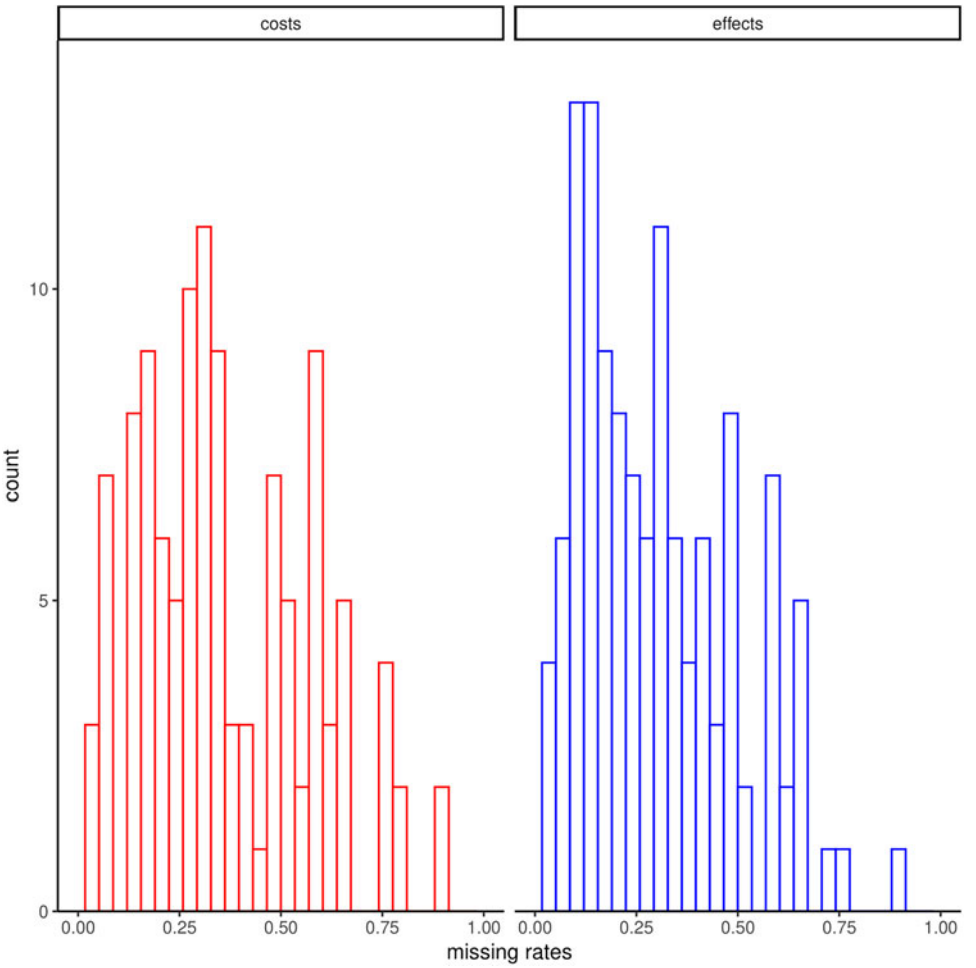

Figure A2. Histograms of the distributions of missingness rates for effects (blue bars) and costs (red bars) among the empirical studies which provided the information.

Cite this article: Gabrio A (2024). A review of heath economic evaluation practice in the Netherlands: are we moving forward? *Health Economics, Policy and Law* **19**, 174–191. https://doi.org/10.1017/S1744133123000087