International Journal of Technology Assessment in Health Care, **24:3** (2008), 318–325.
Copyright © 2008 Cambridge University Press. Printed in the U.S.A.
doi:10.1017/S0266462308080422

Comparison of three instruments assessing the quality of economic evaluations: A practical exercise on economic evaluations of the surgical treatment of obesity

Sophie Gerkens
Université Catholique de Louvain

Ralph Crott
Cliniques Universitaires Saint-Luc

Irina Cleemput
Belgian Health Care Knowledge Centre (KCE)

Jean-Paul Thissen
Université Catholique de Louvain and *Cliniques Universitaires Saint-Luc*

Marie-Christine Closon
Université Catholique de Louvain

Yves Horsmans
Université Catholique de Louvain and *Cliniques Universitaires Saint-Luc*

Claire Beguin
Cliniques Universitaires Saint-Luc

Objectives: The increasing use of full economic evaluations has led to the development of various instruments to assess their quality. The purpose of this study was to compare the frequently used *British Medical Journal* (BMJ) check-list and two new instruments: the Consensus Health Economic Criteria (CHEC) list and the Quality of Health Economic Studies (QHES) instrument. The analysis was based on a practical exercise on economic evaluations of the surgical treatment of obesity.
Methods: The quality of nine selected studies was assessed independently by two health economists. To compare instruments, the Spearman rank correlation coefficient was calculated for each assessor. Moreover, the test–retest reliability for each instrument was assessed with the intraclass correlation coefficient (ICC) (3,1). Finally, the inter-rater agreement for each instrument was estimated at two levels: comparison of the total score of each article by the ICC(2,1) and comparison of results per item by kappa values.

We thank Maria-Isabel Portet (Researcher, Université catholique de Louvain, School of Public Health, Brussels, Belgium) for useful comments. Financial support: no funding, no sponsors. Logistic support: Université catholique de Louvain, Cliniques universitaires Saint-Luc.

Results: The Spearman's rank correlation coefficient between instruments was usually high (rho $>$ 0.70). Furthermore, test–retest reliability was good for every instruments, that is, 0.98 (95 percent CI, 0.86–0.99) for the BMJ check-list, 0.97 (95 percent CI, 0.73–0.98) for the CHEC list, and 0.95 (95 percent CI, 0.75–0.99) for the QHES instrument. However, inter-rater agreement was poor (kappa $<$ 0.40 for most items and ICC(2,1) $\leq$ 0.5).
Conclusions: The study shows that the results of the quality assessment of economic evaluations are not so much influenced by the instrument used but more by the assessor. Therefore, quality assessments should be performed by at least two independent experts and final scoring based on consensus.

Keywords: Quality assurance, Health care, Cost and cost analysis, Questionnaires, Review, Systematic, Technology assessment

With the scarcity of resources in health care, efficacy and safety become insufficient for a well-informed decision making on resources allocations. In the current environment, the priority becomes to reduce the costs without deteriorating the quality of care, or to improve quality of care at a reasonable cost (10). Consequently, interest in full economic evaluations, that is, studies comparing both costs and outcomes of at least two healthcare programs (10) has increased and numerous countries have now developed specific guidelines for economic evaluations. As a consequence, the use of systematic reviews of economic evaluations to summarize knowledge has been intensified and quality assessment instruments have been developed to evaluate the quality of published economic evaluations. The most frequently used instruments (13) are the Drummond et al. ten-item check-list (10) or the BMJ check-list (9), based on the Drummond check-list.

Jefferson et al. showed that too disparate quality assessment instruments were used (13), illustrating the need for a validated and internationally accepted list. To respond to this need, the Consensus Health Economic Criteria (CHEC) list has recently been developed (11).

On the other hand, neither the CHEC list nor the BMJ check-list were created with the aim of performing a simple comparison between studies in a quantifiable way. Quantitative measures of quality would allow ranking studies according to a quality score. One solution is to apply an equal weight for each item, but this strategy does not allow analysts to take into account the relative importance of each criterion. For this reason, a new instrument has been developed: the Quality of Health Economic Studies (QHES) (5), a grading system in which weightings differ according to the relative importance of each criterion.

The first objective of this study was to compare the BMJ check-list, the CHEC list, and the QHES instrument as quantitative tools to measure the quality of economic evaluations, and to examine the importance of weighting the criteria. The second objective was to assess the test–retest reliability and the inter-rater agreement for each instrument. Finally, problems associated with analyzing the quality of economic evaluations with these instruments were also determined and recommendations were made. The analysis was performed through a systematic review of economic evaluations of the surgical treatment of obesity.

METHODS

Quality Assessment Instruments Description

The BMJ Check-List. The BMJ set up a working party to develop a quality assessment check-list for use by both referees and authors. Drafts of the check-list were transmitted to health economists and journal editors and were debated at the "biannual meeting of the UK Health Economists' Study Group" in January 1996. The final check-list was based on a broad consensus and contains thirty-five items under three headings: study design, data collection, and analysis and interpretation of results (see Table 1). This check-list concentrates on full economic evaluations but could also be used for partial economic evaluations, or report and commentaries on economic evaluations. If items were not applicable to a specific study, a "not appropriate" (NA) response can be stated. The working party admitted that it is not possible to address all the points in the article and that authors can, for example, refer the reader to other published sources. More details about this check-list can be found in the literature (9).

The CHEC List. An initial item pool divided in nineteen categories was first developed by performing a literature search from Medline, Psychlit, Econlit, the Cochrane Library, and the National Health Service Economic Evaluation Databases (NHS EED). The criteria list was then created using the Delphi method. This method made use of a panel of expert on a specific topic to reach a consensus (21). In a first round, international experts were asked to give their opinion on the categories and the items selected from the literature research. Comments and the resulting list were redistributed among experts until a consensus was reached. Three rounds were sufficient to obtain the final criteria list. More details on the method used can be found in the literature (2;11).

The list contains nineteen yes-or-no questions (see Table 2). Authors recommended that, if not enough information was available in the article or in other published material to answer to a question, a "No" response should be

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press

Table 1. The *British Medical Journal* Check-list[a]

Study design
The research question is stated
The economic importance of the research question is stated
The viewpoints of the analysis are clearly stated and justified
The rationale for choosing the alternative programmes or interventions compared is stated
The alternatives being compared are clearly described
The form of economic evaluation used is stated
The choice of form of economic evaluation is justified in relation to the questions addressed
Data collection
The sources of effectiveness estimates used are stated
Details of the design and results of effectiveness study are given (if based on a single study)
Details of the method of synthesis or meta-analysis of estimates are given (if based on an overview of a number of effectiveness studies)
The primary outcome measure(s) for the economic evaluation are clearly stated
Methods to value health states and other benefits are stated
Details of the subjects from whom evaluations were obtained are given
Productivity changes (if included) are reported separately
The relevance of productivity changes to the study question is discussed
Quantities of resources are reported separately from their unit costs
Methods for the estimation of quantities and unit costs are described
Currency and price data are recorded
Details of currency or price adjustments for inflation or currency conversion are given
Details of any model used are given
The choice of model used and the key parameters on which it is based are justified
Analysis and interpretation of results
Time horizon of costs and benefits is stated
The discount rate(s) is stated
The choice of rate(s) is justified
An explanation is given if costs or benefits are not discounted
Details of statistical tests and confidence intervals are given for stochastic data
The approach to sensitivity analysis is given
The choice of variables for sensitivity analysis is justified
The ranges over which the variables are varied are stated
Relevant alternatives are compared
Incremental analysis is reported
Major outcomes are presented in a disaggregated as well as aggregated form
The answer to the study question is given
Conclusion follow from the data reported
Conclusions are accompanied by the appropriate caveats

[a] See reference 9.

stated. A description of the items can be found on the Web (www.beoz.unimaas.nl/chec/). It should be noted that this list was not created to analyze the quality of economic evaluations based on modeling studies. However, in this study, all economic evaluations found in the literature were analyzed with the three lists, including modeling studies. Consequently, the quality score generated by the CHEC list for modeling studies has to be interpreted with caution.

The QHES List. A steering committee comprised of five experts in the field of health economics and three investigators developed a check-list for economic evaluations from a literature search using Medline, Healthstar, and the Cochrane databases. From existing guidelines and check-lists, the committee selected 16 criteria with a "Yes" or "No" format (see Table 3). The selection was made by consensus. Then, weights for each criterion were estimated using a general linear regression (random effects) based on data collected from a conjoint analysis survey on 120 international health economists. More details about the QHES list can be found in the literature (5).

Studies Selection

The different quality assessment instruments were applied to nine economic evaluations of surgical treatment techniques of obesity. More details about the studies and the selection criteria for the studies are described elsewhere (14). While initially only full economic evaluations were selected for review, we also tested the quality assessment instruments on cost-outcome descriptions, that is, studies describing both costs and effects but not presenting an incremental cost-effectiveness ratio (ICER). Five full economic evaluations, of which two included a cost-utility and cost-effectiveness analysis and three included only a cost-utility analysis were included in the quality assessment exercise (4;7;8;19;20). Moreover, four cost-outcome description

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press

Table 2. The Consensus Health Economic Criteria List[a]

Is the study population clearly described?
Are competing alternatives clearly described?
Is a well-defined research question posed in answerable form?
Is the economic study design appropriate to the stated objective?
Is the chosen time horizon appropriate in order to include relevant costs and consequences?
Is the actual perspective chosen appropriate?
Are all important and relevant costs for each alternative identified?
Are all costs measured appropriately in physical units?
Are costs valued appropriately?
Are all important and relevant outcomes for each alternative identified?
Are all outcomes measured appropriately?
Are outcomes valued appropriately?
Is an incremental analysis of costs and outcomes of alternatives performed?
Are all future costs and outcomes discounted appropriately?
Are all important variables, whose values are uncertain, appropriately subjected to sensitivity analysis?
Do the conclusions follow from the data reported?
Does the study discuss the generalizability of the results to other settings and patient/client groups?
Does the article indicate that there is no potential conflict of interest of study researcher(s) and funder(s)?
Are ethical and distributional issues discussed appropriately?

[a] See references 2, 11.

studies were evaluated with the three quality assessment instruments (1;6;16;17).

Quality Assessment of Economic Evaluations

The quality of selected studies was assessed independently by two heath economists (rater 1 and rater 2) using the BMJ check-list, the CHEC list, and the QHES list. Each economist blindly evaluated the quality of studies with the three instruments consecutively. Moreover, rater 1 repeated the analysis 8 weeks later. During the investigation, the guidelines of the instruments were followed and an inventory of problems associated with analyzing the quality of economic evaluations according to these guidelines was made.

Quality scores were then evaluated for each study. In a first stage, a score with an equal weight for each item was

Table 3. The Quality of Health Economic Studies Instrument[a]

Item	Weightings
Was the study objective presented in a clear, specific, and measurable manner?	7
Were the perspective of the analysis (societal, third-party payer, and so on) and reasons for its selection stated?	4
Were variable estimates used in the analysis from the best available source (i.e., Randomized Control Trial-Best, Expert Opinion-Worst)?	8
If estimates came from a subgroup analysis, were the groups prespecified at the beginning of the study?	1
Was uncertainty handled by: (i) statistical analysis to address random events; (ii) sensitivity analysis to cover a range of assumptions?	9
Was incremental analysis performed between alternatives for resources and costs?	6
Was the methodology for data abstraction (including value health states and other benefits) stated?	5
Did the analytic horizon allow time for all relevant and important outcomes? Were benefits and costs that went beyond 1 year discounted (3–5%) and justification given for the discount rate?	7
Was the measurement of costs appropriate and the methodology for the estimation of quantities and unit costs clearly described?	8
Were the primary outcome measure(s) for the economic evaluation clearly stated and were the major short term, long term and negative outcomes included?	6
Were the health outcomes measures/scales valid and reliable? If previously tested valid and reliable measures were not available, was justification given for the measures/scales used?	7
Were the economic model (including structure), study methods and analysis, and the components of the numerator and denominator displayed in a clear transparent manner?	8
Were the choice of economic model, main assumptions and limitations of the study stated and justified?	7
Did the author(s) explicitly discuss direction and magnitude of potential biases?	6
Were the conclusions/recommendations of the study justified and based on the study results?	8
Was there a statement disclosing the source of funding for the study?	3

[a] See reference 5.

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press

calculated as a quantitative proxy for the evaluation's quality. In a second stage, the importance of weighting the criteria was examined. For the QHES instrument, weights determined by Chiou et al. were used (5). For the BMJ and the CHEC list, no weighting exist. Implicit weightings determined by one of the assessor according to the relative importance he confers to each item (subjective assessment) were thus used. In summary, three types of quality scores were obtained: a score without weighting of the criteria, a score with an implicit weighting for the BMJ and the CHEC list, and a score with an explicit weighting determined by Chiou et al. for the QHES instrument.

Statistical Analysis

To compare the instruments, the range and the mean of the quality scores generated by each instrument were calculated. Ranking differences between instruments were then assessed using the Spearman rank correlation coefficient. In this study, we considered a correlation coefficient of $r \geq 0.7$ as high, $0.7 > r \geq 0.5$ as moderate, and $r < 0.5$ as low.

Moreover, test–retest reliability between time 1 and 2 was assessed for each instrument by the rater 1 using model 3 of the six ICCs discussed by Shrout and Fleiss (18), that is, the ICC(3,1) where raters are assumed to be representative of the entire population of raters (Supplementary Figure 1, available at www.journals.cambridge.org/thc).

Finally, for each instrument, the inter-rater agreement at time 1 was estimated at two levels: comparison of the total score of each article by the ICC(2,1) (18), where raters are assumed to be a random subset of all possible raters, and comparison of results per item by kappa values. Kappa values less than 0.40, between 0.40 and 0.74, and between 0.75 and 1 were defined as poor, fair to good, and perfect agreement respectively (12;15). Tests were performed using SAS software version 9.

RESULTS

Instruments Comparison

The comparison of instruments showed that they mainly analyze similar items. Nevertheless, some differences can be highlighted (see Table 4).

First, only the BMJ check-list investigates if the economic importance of the study question is stated and if the choice of alternatives is justified. On the other hand, the BMJ check-list does not assess if the choice of cost and outcome items is appropriate, as it is done in the other two instruments. Finally, this is the only instrument that does not include a question about the presence of conflicts of interest by the authors.

Second, the CHEC list is designed for clinical trial and observational studies. Consequently, there is no item on model characteristics. Moreover, this instrument does not determine whether limitations of the studies are specified.

Table 4. Instruments' Comparisons

	BMJ	CHEC	QHES
Objective	Y	Y	Y
Economic importance of the study question	Y	N	N
Economic study design	Y	Y	N
Description of the population	Y	Y	N
Subgroup analysis	N	N	Y
Perspective	Y	Y	Y
Time horizon and discount rate	Y	Y	Y
Alternative description	Y	Y	N
Alternative choice	Y	N	N
Outcomes choice	N	Y	Y
Outcomes measurement	Y	Y	Y
Outcomes valuation	Y	Y	Y
Costs choice	N	Y	Y
Costs measurement	Y	Y	Y
Costs valuation	Y	Y	N
Details of the model	Y	N	Y
Incremental analysis	Y	Y	Y
Handle of uncertainty	Y	Y	Y
Results presentation	Y	N	N
Appropriateness of the conclusion	Y	Y	Y
Limitations	Y	N	Y
Results generalizability	N	Y	N
Ethical aspect	N	Y	N
Conflict of interest	N	Y	Y

BMJ, British Medical Journal; CHEC, Consensus Health Economic Criteria; QHES, Quality of Health Economic Studies; Y, handled; N, not handled.

On the other hand, it is the only instrument asking whether ethical aspect and generalizability of the results are investigated.

Third, the QHES instrument determines if subgroups analyzed are appropriately defined but does not examine whether details on the population and on the study design are specified, in contrast to both the BMJ check-list and the CHEC list. Finally, this instrument does not investigate if details about price adjustments for inflation or about currency conversion are given.

Inventory of Problems Encountered During the Quality Assessment

It was often difficult to judge from the studies if an item was respected or not because too little information was given in the publication. For example, details on cost calculations were regularly limited and sometimes, only sources were given. In such situation, it is thus important to consult these sources to be able to evaluate the quality of the studies with more accuracy.

Moreover, the BMJ check-list and the QHES instruments were mainly adapted to modeling studies while the CHEC list was conceived for clinical trials and observational studies. Consequently, the item assessing if details of the model were given was, for example, not adapted to clinical trials and

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press

observational studies. Hence, it could be interesting to have an instrument adapted to several study designs with specific subquestions for each design.

It was also often difficult to choose between a "Yes" or "No" response. Some items regrouped various criteria. Consequently, if only one of the criterion was not respected, a "no" response should be stated, even if other criteria were respected. The possibility to use an intermediate value as "partially respected" could thus be interesting. This problem was mostly present with the QHES instrument. For example, one item tested if the time horizon was relevant, if costs and outcomes were discounted and if the discount rate was justified. It would be interesting to test the impact of subdividing this kind of item.

Differences in Quality Scores Between Instruments

With equal weights between items, the quality score of the nine selected studies and the three ratings (rater 1, time 1; rater 1, time 2; rater 2, time 1) varied between 30.8 and 90.0 of 100 points on the BJM check-list, between 15.8 and 84.2 of 100 points on the CHEC list, and between 6.3 and 87.5 of 100 points on the QHES instrument. Means and standard deviations of the studies quality scores for each instrument and the three ratings are detailed on the Web site (Supplementary Table 5, available at www.journals.cambridge.org/thc).

With a weighting between items, the quality score of the nine selected studies and the three ratings varied between 6 and 92 of 100 points on the BJM check-list, between 14 and 89 of 100 points on the CHEC list, and between 22 and 77 of 100 points on the QHES instrument.

Hierarchical ranking of studies

The hierarchical ranking of studies based on their quality score and the Spearman ranking correlation coefficient can be found at the Web site (Supplementary Tables 6 and 7, available at www.journals.cambridge.org/thc). A high Spearman ranking correlation between instruments was found, except for rater 1 in time 1 where the correlation was moderate between the BMJ and the CHEC list.

Moreover, for each instrument, the Spearman ranking correlation coefficients between weighted and not weighted scores were high and ranged from 0.83 to 0.99. Thus, weighting of criteria has little impact on the ranking. On the other hand, the ranking varied according to the assessor, as shown by the inter-rater agreement.

Test–retest Reliability and Inter-rater Agreement

Test–retest reliability in terms of ICC(3,1) was good for all instruments, that is, 0.99 (95 percent CI, 0.86–0.99) for the BMJ check-list, 0.97 (95 percent CI, 0.73–0.98) for the CHEC list, and 0.95 (95 percent CI, 0.75–0.99) for the QHES instrument. However, there was poor inter-rater agreement. For the BMJ check-list, agreement was poor for twenty-seven of thirty-five items (77 percent), fair to good for six of thirty-five items (17 percent), and perfect for only two out of thirty-five items (6 percent). For the CHEC list, agreement was poor for fifteen of nineteen items (79 percent) and perfect for four of nineteen items (21 percent). For the QHES instrument, agreement was poor for ten of sixteen items (63 percent), fair to good for three of sixteen items (19 percent), and perfect for three of sixteen items (19 percent). Overall inter-rater agreement in terms of ICC(2,1) was 0.52 (95 percent CI, 0.21–0.83) for the BMJ check-list, 0.33 (95 percent CI, 0.07–0.71) for the CHEC list, and 0.33 (95 percent CI, 0.02–0.73) for the QHES instrument.

DISCUSSION

Instruments comparisons highlighted the subjective character of the quality assessment. Indeed, results were not influenced by the instruments used but rather by experts who analyzed the studies. As shown in the study, instruments mainly assessed similar items, which could explain the high Spearman rank correlation coefficient.

The poor agreement between experts could be explained by various factors. First, time spent to analyze studies might have an impact on results. One author spent around 1 day per study to assess deeply the quality of the study and returned systematically to referred sources if insufficient details were provided in the basic article. The other expert spent around 2 days to assess the quality of all the studies and based his analysis on the main article only.

Second, the subjectivity of the examinants could also influence the response. A complete respect of criteria was rare and intermediate responses were not authorized. Severe raters could have tendency to state a 0 value if one criterion of the item was not completed, while another rater could have tendency to state a 1 value, considering that on the whole, the criteria were respected.

Third, experience of the rater in economic evaluation could also play a role. One rater has worked in the health economics domain for nearly 20 years, while the experience of the second rater was only 2 years. Thus, it is possible that they consider the quality of the studies differently.

Fourth, the perception and interpretation of the items and the ambiguity of the responses also influenced the results. Items were sometimes large and could be interpreted in various ways. Some items also referred to specific study design and when the design of the study was not appropriate, reaction of raters could differ.

It should also be noted that the BMJ check-list and the CHEC list were created as qualitative instruments and not as scoring instruments. On the other hand, calculating a quality score from these instruments allowed us to easily have an idea of the ranking of these studies according to their quality.

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press

Finally, caution is needed when interpreting our results given that the limited number of studies led to high confidence intervals for the inter-rater agreement. In a previous study, two people analyzed the quality of 30 economic evaluations of health promotion with the QHES instrument and found a better inter-rater reliability (IC95 percent: 0.64–0.91) (3). However, our results plead in favor of doing further research to estimate the overall role of the evaluator in assessing the quality of economic evaluations. To do so, an international study including a higher number of evaluators and in particular a larger sample of studies from various areas should be conducted.

In conclusion, our findings highlight that in practice, results are not so much influenced by the instrument used but more by the assessor. It is thus essential to perform quality analyses of economic evaluations by at least two blinded experts and to base the final scoring on a consensus. Moreover, a clear definition of each item should be given and respected by raters. Experts should also spend a substantial period of time to analyze studies thoroughly and should refer to sources of information when specified in the article if not enough details are provided in the basic study. Finally, in the future, it would be interesting to create a single instrument adapted to each study design and to introduce the possibility to use an intermediate score value.

CONTACT INFORMATION

Sophie Gerkens, MSc, PhD Candidate (sophie.gerkens@uclouvain.be), Health Economist, School of Public Health–Unité de Socioéconomie de la santé (SESA), Université Catholique de Louvain, 30 Clos Chapelle-aux-Champs box 3041, Brussels 1200, Belgium

Ralph Crott, PhD (ralph.crott@uclouvain.be), Health Economist, Department of Medicine, Cliniques Universitaires Saint-Luc, 10, Av. Hippocrate, Brussels B-1200, Belgium

Irina Cleemput, PhD (Irina.Cleemput@kce.fgov.be), Expert Economic Analysis, Department of Research, Belgian Health Care Knowledge Centre (KCE), 62 Rue de la Loi, Brussels B-1040, Belgium

Jean-Paul Thissen, MD, PhD (Thissen@diab.ucl.ac.be), Professor, Department of Endocrinology, Université Catholique de Louvain, 10, av. Hippocrate, Brussels B-1200, Belgium; Chief, Department of Endocrinology, Cliniques Universitaires Saint-Luc, 10, av. Hippocrate, Brussels B-1200, Belgium

Marie-Christine Closon, PhD (closon@sesa.ucl.ac.be), Professor, School of Public Health – Unité de socioéconomie de la santé (SESA), Université Catholique de Louvain, 30 Clos Chapelle-aux-Champs, Box 3041, Brussels B-1200, Belgium

Yves Horsmans, MD, PhD (yves.horsmans@uclouvain.be), Professor, Department of Gastroenterology, Université Catholique de Louvain; Chief, Department of Gastroenterology, Cliniques Universitaires Saint-Luc, 10, Av. Hippocrate, Brussels B-1200, Belgium

Claire Beguin, MD, PhD (claire.beguin@uclouvain.be[]), Chief, Department of Medical Information and Statistics, Cliniques Universitaires Saint-Luc, 10, Av. Hippocrate, Brussels B-1200, Belgium

REFERENCES

1. Agren G, Narbro K, Jonsson E, et al. Cost of in-patient care over 7 years among surgically and conventionally treated obese patients. *Obes Res*. 2002;10:1276-1283.
2. Ament A, Evers S, Goossens M, De Vet H, Van Tulder M. Criteria list for conducting systematic reviews based on economic evaluation studies – the CHEC project. In: Donaldson C, Mugford M, Vale L, eds. *Evidence-based health economics. From effectiveness to efficiency in systematic review*. London: BMJ Books; 2002:99-113.
3. Au F, Prahardhi S, Shiell A. Reliability of two instruments for critical assessment of economic evaluations. *Value Health*. 2007. In press.
4. Chevallier JM, Daoud F, Szwarcensztein K, Volcot MF, Rupprecht MF. Medicoeconomic evaluation of the treatment of morbid obesity by Swedish adjustable gastric banding (SAGB). *Ann Chir*. 2006;131:12-21.
5. Chiou CF, Hay JW, Wallace JF, et al. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care*. 2003;41:32-44.
6. Christou NV, Sampalis JS, Liberman M, et al. Surgery decreases long-term mortality, morbidity, and health care use in morbidly obese patients. *Ann Surg*. 2004;240:416-423.
7. Clegg AJ, Colquitt J, Sidhu MK, et al. The clinical effectiveness and cost-effectiveness of surgery for people with morbid obesity: A systematic review and economic evaluation. *Health Technol Assess*. 2002;6:1-153.
8. Craig BM, Tseng DS. Cost-effectiveness of gastric bypass for severe obesity. *Am J Med*. 2002;113:491-498.
9. Drummond MF, Jefferson TO. Guidelines for authors and peer reviewers of economic submissions to the BMJ. The BMJ Economic Evaluation Working Party. *BMJ*. 1996;313:275-283.
10. Drummond MF, Sculpher MJ, Torrance JW, O'Brien BJ, Stoddart JL. *Methods for the economic evaluation of health care programmes*. Oxford: Oxford University Press; 2005.
11. Evers S, Goossens M, de Vet H, van Tulder M, Ament A. Criteria list for assessment of methodological quality of economic evaluations: Consensus on health economic criteria. *Int J Technol Assess Health Care*. 2005;21:240-245.
12. Fleiss JL. *Statistical methods for rates and proportions*. New York: John Wiley and Sons; 1981.
13. Jefferson T, Demicheli V, Vale L. Quality of systematic reviews of economic evaluations in health care. *JAMA*. 2002;287:2809-2812.
14. Lambert ML, Kohn L, Vinck I, et al. *Pharmacological and surgical treatment of obesity. Residential care for severely obese children in Belgium*. Brussels: Belgian Health Care Knowledge Centre (KCE); 2006. Report 36C.
15. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33:159-174.

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press

16. Martin LF, Tan TL, Horn JR, et al. Comparison of the costs associated with medical and surgical treatment of obesity. *Surgery*. 1995;118:599-606.
17. Nguyen NT, Goldman C, Rosenquist CJ, et al. Laparoscopic versus open gastric bypass: A randomized study of outcomes, quality of life, and costs. *Ann Surg*. 2001;234:279-289.
18. Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*. 1979;86:420-428.
19. van Gemert WG, Adang EM, Kop M, et al. A prospective cost-effectiveness analysis of vertical banded gastroplasty for the treatment of morbid obesity. *Obes Surg*. 1999;9:484-491.
20. van Mastrigt GA, van Dielen FM, Severens JL, Voss GB, Greve JW. One-year cost-effectiveness of surgical treatment of morbid obesity: Vertical banded gastroplasty versus Lap-Band. *Obes Surg*. 2006;16:75-84.
21. Whitman NI. The Delphi technique as an alternative for committee meetings. *J Nurs Educ*. 1990;29:377-379.

https://doi.org/10.1017/S0266462308080422 Published online by Cambridge University Press