

Original Articles

Development and Validation of a Grading System for the Quality of Cost-Effectiveness Studies

CHIUN-FANG CHIOU, PhD,* JOEL W. HAY, PhD,† JOEL F. WALLACE, PHARM.D, MBA,*
BERNARD S. BLOOM, PhD,‡ PETER J. NEUMANN, ScD,§ SEAN D. SULLIVAN, PhD,¶ HSING-TING YU, MPH,*
EMMETT B. KEELER, PhD,|| JAMES M. HENNING, MS,# AND JOSHUA J. OFMAN, MD, MSHS***

PURPOSE. To provide a practical quantitative tool for appraising the quality of cost-effectiveness (CE) studies.

METHODS. A committee comprised of health economists selected a set of criteria for the instrument from an item pool. Data collected with a conjoint analysis survey on 120 international health economists were used to estimate weights for each criterion with a random effects regression model. To validate the grading system, a survey was sent to 60 individuals with health economics expertise. Participants first rated the quality of three CE studies on a visual analogue scale, and then evaluated each study using the grading system. Spearman rho and Wilcoxon tests were used to detect convergent validity and analysis of covariance (ANCOVA) for discriminant validity. Agreement between the global rating by experts and the grading system was also examined.

RESULTS. Sixteen criteria were selected. Their coefficient estimates ranged from 1.2 to 8.9, with a sum of 93.5 on a 100-point scale. The only insignificant criterion was "use of subgroup analyses." Both convergent validity and discriminant validity of the grading system were shown by the results of the Spearman rho (correlation coefficient = 0.78, $P < 0.0001$), Wilcoxon test ($P = 0.53$), and ANCOVA ($F_{3,146} = 5.97$, $P = 0.001$). The grading system had good agreement with global rating by experts.

CONCLUSIONS. The instrument appears to be simple, internally consistent, and valid for measuring the perceived quality of CE studies. Applicability for use in clinical and resource allocation decision-making deserves further study.

Key words: Cost-effectiveness; quality assessment; conjoint analysis; convergent validity. (Med Care 2003;41:32-44)

*From Zynx Health Inc., Cedars-Sinai Health System, Beverly Hills, California.

†From the Department of Pharmaceutical Economics and Policy, University of Southern California, Los Angeles, California.

‡From the Department of Medicine, Institute on Aging, and Leonard Davis Institute for Health Economics, University of Pennsylvania, Philadelphia, Pennsylvania.

§From the School of Public Health, Harvard University, Boston, Massachusetts.

¶From the Department of Pharmacy, University of Washington, Seattle, Washington.

||From RAND, Santa Monica, California.

#From TAP Pharmaceutical Products, Inc., Lake Forest, Illinois.

**From the Department of Medicine, and Health Service Research, Cedars-Sinai Health System, Los Angeles, California.

Supported by TAP Pharmaceutical Products, Inc., Lake Forest, Illinois.

Address correspondence and reprint requests to: Chiun-Fang Chiou, PhD, Zynx Health, Inc., 9100 Wilshire Blvd, East Tower, Suite 655, Beverly Hills, CA 90212. E-mail: cchiou@cerner.com

Received October 1, 2001; initial decision January 14, 2002; accepted June 28, 2002.

derive the needed information from the qualitative

instruments.

Our objective was to develop and validate an easy-to-use, weighted, grading system to appraise the quality of health economic evaluations (ie, general cost-effectiveness studies including cost-minimization, cost-benefit, cost-effectiveness, and cost-utility analyses). The instrument can assist decision-makers in separating high quality from low quality studies, enable users to more accurately determine which economic analyses should be considered in systematic reviews and in decision-making, and provide a simple mechanism for editor/reviewers of medical journals to grade the quality of submitted papers.

Materials and Methods

Literature Review

To identify checklists and guidelines used to evaluate the quality of economic analyses a comprehensive search of the literature was performed using the MEDLINE, HEALTHSTAR, and COCHRANE computerized bibliographic databases, an Internet search, and screening of textbooks on economic evaluation studies. The checklists and guidelines selected were written in English language, published after 1990, and developed for "cost minimization," "cost-effectiveness," or "cost-utility" analyses. A data-base of checklists and guidelines for economic evaluations was then developed. A steering committee comprising health economists and editors of health economics journals was convened. The steering committee consisted of five experts in the field of health economics (Dr. Hay, Dr. Bloom, Dr. Neumann, Dr. Sullivan, Dr. Kee-ler, with Dr. Hay serving as Chair), and three study investigators (Dr. Ofman, Dr. Chiu, and Dr. Wallace). The development of the quality assessment tool was limited to the following four major types of health economic analyses: "cost-minimization," "cost-effectiveness," and "cost-utility" analyses.

Criterion Selection

An initial item pool was compiled based on all identified checklists and guidelines (Table 1). The steering committee first reviewed each criterion

The allocation of limited medical resources is one of the most important issues for employers, insurers, and governments. Rigorously performed economic evaluations are considered useful tools in evaluating the clinical and financial consequences of competing health care interventions, and are frequently relied upon by governmental authorities and decision-makers to determine strategies for using scarce health care resources more efficiently.¹

Several countries have incorporated the results of economic evaluations into the process of health policy decision-making.^{2,3} Consequently, the amount of health economic literature published in the past decade has grown dramatically.⁴ Despite the increase in the quantity and reach of economic analyses, the quality of the performance and reporting in the published health economic literature remains less than optimal.⁵⁻⁸

Efforts to improve the quality of health economic evaluations include those by the *British Medical Journal* (BMJ),⁹ Canadian Collaborative Workshop for Pharmacoeconomics, and the US Public Health Service Panel on Cost-Effectiveness in Health and Medicine. They and others have aided development of generic and disease-specific guidelines, checklists, and recommendations for the acceptable methods for conducting and reporting economic analyses, and the systematic application of these methods throughout the peer review process.^{3,9-14} These efforts were intended to help researchers improve their performance of such analyses and assist users to evaluate the quality of health economic studies.

Although some of the current checklists and appraisal systems have a more general objective of specifying the reporting required in health economic evaluations, others suffer from a number of limitations in serving as appraisal tools for the quality of health economic evaluations. First, none have been formally validated as containing items that are related to both internal and external validity of economic evaluations. Second, these instruments are qualitative, most contain subjective and open-ended items, and none provide a score to enable the simple comparison among studies. Finally, these checklists and appraisal criteria assume that each criterion shares an equal weight or level of importance. Thus, it is unclear if current instruments have the capability to discriminate between health economic analyses of high and low quality, and whether users of economic literature without specific expertise are able to

TABLE 1. Summary of Existing Guidelines, Checklists, and Recommendations for Health Economic Studies*

Criterion/Source	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	SUM
Objective			✓	✓					✓	✓	✓	✓				✓				7
Perspective			✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	✓	✓			✓	14
Study design			✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓		✓				12
Analysis		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	18
Data collection	✓		✓	✓				✓		✓	✓				✓	✓			✓	9
Time horizon								✓		✓			✓			✓				4
Cost/resources		✓		✓	✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓	✓	15
Outcome measures		✓	✓	✓		✓		✓	✓	✓	✓		✓	✓	✓	✓	✓	✓	✓	15
Discounting			✓	✓	✓			✓			✓	✓		✓	✓	✓	✓	✓	✓	12
Transparency				✓				✓	✓				✓						✓	5
Cost-effectiveness ratio	✓		✓						✓					✓		✓				5
Discussion		✓	✓								✓		✓	✓		✓		✓		7
Conclusions	✓		✓							✓	✓		✓	✓		✓	✓	✓		9
Sponsorship			✓	✓				✓	✓											4
Non-specified		✓	✓					✓	✓			✓							✓	6
SUM	3	5	12	10	5	5	3	10	9	9	10	7	9	9	6	12	4	7	7	
Number of criteria [†]	9	15	36	16	8	16	8	24	18	21	40	13	23	28	15	35	10	8	14	

*I, N, P, and R are commonly referred to as the "Canadian guidelines," "Drummond's guidelines," "BMJ guidelines," and "U.S. Panel recommendations," respectively.

[†]Criteria were presented in the format of "yes/no" questions, statements, or recommendations.

A: Problems with the interpretation of pharmacoeconomic analyses: A review of submissions to the Australian Pharmaceutical Benefits Scheme. Hill, et al., 2000.²⁰

B: The revised Canadian Guidelines for the Economic Evaluation of Pharmaceuticals. Glennie, et al., 1999.²¹

C: Evaluating the quality of published pharmacoeconomic evaluations. Sanchez, et al., 1995.²²

D: Emerging standardization in pharmacoeconomics. Mullins, et al., 1998.²³

E: Use of economic evaluation guidelines: 2 years' experience in Canada. Baladi, et al., 1998.²⁴

F: Common errors and controversies in pharmacoeconomic analyses. Byford, et al., 1998.²⁵

G: The Danish approach to standards for economic evaluation methodologies. Alban, et al., 1997.²⁶

H: Canada's new guidelines for the economic evaluation of pharmaceuticals. Menon et al., 1996.²⁷

I: Canadian guidelines for economic evaluation of pharmaceuticals. Torrance, et al., 1996.¹⁷

J: Methodological and conduct principles for pharmacoeconomic research. Pharmaceutical Research and Manufacturers of America. Clemens, et al., 1995.²⁸

K: Evaluation of pharmacoeconomic studies: Utilization of a checklist. Sacristan, et al., 1993.²⁹

L: Guidelines for the clinical and economic evaluation of health care technologies. Guyatt, G., et al. 1986.³⁰

M: Economic analysis of health care technology. A report on principles. Task Force on Principles for Economic Analysis of Health Care Technology, 1995.³¹

N: Critical assessment of economic evaluation. Drummond, et al., 1997.³²

O: The U.K. NHS economic evaluation database. Economic issues in evaluations of health technology. Nixon, et al., 2000.³³

P: Guidelines for authors and peer reviewers of economic submissions to the BMJ. Drummond, et al., 1996.¹²

Q: Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice.

A. Are the results of the study valid? Evidence-Based Medicine Working Group. Drummond, et al., 1997.³⁴

R: Recommendations of the Panel on Cost-Effectiveness in Health and Medicine. Weinstein, et al., 1996.¹³

S: Pharmacoeconomic Models in Disease Management. A Guide for the Novice or the Perplexed. Milne, 1998.³⁵

and added new ones to the item pool if the committee agreed on importance. Each criterion in the item pool was then subjected to a vote by the steering committee using a 9-point Likert scale (1

= "not important at all" to 9 = "extremely important"). Because a score of 5 implies that a reviewer feels "indifferent" about an item, items with an average score equal to or less than 5.5 were

several blocks of 10 in an effort to reduce the response burden (ie, each respondent only rated 10 scenarios). The survey instructions included a description of the survey, the method for rating each scenario, and two example scenarios and their ratings. An answer sheet was provided to respondents along with ten scenarios. For each scenario (a hypothetical economic study with a combination of attributes), respondents were asked to rate the quality of the study on a visual analogue scale (VAS) anchored with 0 denoting "Extremely Poor" quality and 100 denoting "Excellent" quality. Participants were also instructed to complete questions about their experience in health economics, their occupational setting, background information, and their perceived value of a grading system for the quality of economic studies. We also solicited comments related to the selected or missing criteria.

The conjoint analysis survey was pilot-tested on six individuals working in the health economics field. The questionnaire was modified based on comments from the pilot group and the steering committee. The modified questionnaire was then mailed to 120 members, randomly selected from the membership files of the International Health Economics Association and Society for Medical Decision Making. Scenarios used in Phase I (conjoint analysis) Survey available from corresponding author upon request.

Phase I Statistical Analysis

Consistency of ratings provided by survey participants was checked before analysis. Data with inconsistent or illogical ratings were excluded from the analysis. Each respondent rated 10 scenarios, and because the ratings may not be independent, we used a random effects general least squares (GLS) regression to estimate the equation for predicting the VAS (global quality) scores. Our a priori hypothesis was that all the criteria would have positive coefficients. Two models were estimated. The first model used respondents' characteristics in addition to 16 binary variables regarding whether a scenario meets the criteria as the independent variables. Because none of the characteristics variables was a statistically significant predictor in the model, the second model only used the 16 criteria variables as the independent variables. The nonstandardized coefficients were then multiplied by an adjustment factor (ie, 100

removed from the item pool. The committee's ratings on the remaining items were assembled anonymously and distributed to the members for review. After the final ratings were discussed, a determination was made by consensus about which items to retain. Although no predefined cut-points were assigned for inclusion, the items retained generally had an average score equal to or greater than 7.0 on the 9-point scale and were confirmed by most members as important. Several criteria were reworded based on suggestions from the steering committee after the criterion selection was finalized. In the operationalization of each criterion—for simplicity and potential ease of use—it was decided that each criterion should have a "yes/no" format.

Survey Design and Statistical Analysis

The survey was conducted in two phases. The objective of phase I was to collect data to develop weights for each criterion. The objective of phase II was to collect data to validate the quantitative, weighted instrument against the global quality assessment of experts reviewing actual economic articles.

Phase I Survey

Conjoint analysis was used to estimate weights for the criteria in the instrument. Conjoint analysis is a technique frequently used to assign utility values to individual attributes of a product by rating profiles/scenarios of attributes in relation to the product rather than rating the individual attributes separately. In the present study the instrument was considered the "product" and the criteria were considered the "attributes." The utility values generated from the conjoint analysis were used to obtain the estimated weights. We used the orthogonal main effect plan ("fractional factorial design") function of SPSS Conjoint 8.0 (SPSS Inc., Chicago, IL) to generate a number of scenarios of hypothetical health economic studies. An orthogonal design reduces the number of scenarios requiring presentation to respondents to a reasonable level, while still being able to infer utilities (weights) for all possible combinations of attributes, and ensure the absence of multicollinearity between criteria. Scenarios were generated including various combinations of criteria, and were divided into

divided by sum of the nonstandardized coefficients) to obtain weights for each criterion. Thus, the weighted quality score of a health economic analysis would equal the sum of weights for each criterion that the analysis meets and would be interpreted on an interval scale between 0 and 100 (0 means "extremely poor" and 100 means "excellent").

Phase II Survey

To determine the validity of the quantitative grading system for the quality of health economic studies, the expert panel (including both clinicians and nonclinicians) evaluated the global quality of each of three health economics articles that were assigned to them using the VAS global score. After assigning each study a global score, they were instructed to appraise the article using the grading system.

The steering committee believed it was important that "experts" had sufficient understanding in the clinical area of the article being appraised. Thus, health economics articles were selected from six clinical categories: cardiology, gastroenterology, oncology, infectious disease, colorectal cancer screening, and diseases of the central nervous system. Articles were identified through a search of the MEDLINE computerized bibliographic database using the MeSH headings "cost-benefit analysis" and "health care cost" for the years 1990 to 1998. From the group of articles in each clinical category, we identified nine that were believed qualitatively to represent articles of poor, fair, and good quality. Therefore, a total of 54 (6×9) articles were distributed to experts participating in the study. Sixty experts (10 for each category) were chosen based on their published areas of expertise and the panel for each clinical category was comprised of both clinician and nonclinician experts in health economics. The clinical experts must hold an advance degree in medicine or pharmacy. The nonclinical experts must hold a doctoral degree.

The validation survey was pilot-tested with five of the six individuals who participated in the pilot-test in Phase I. After the survey was modified based on comments from the five individuals, the questionnaire along with three articles were mailed to each of the 60 experts (30 clinicians and 30 nonclinicians).

Phase II Statistical Analysis

Assuming that the global quality score provided by the experts in health economics indicates the perceived quality of a paper, we tested several hypotheses related to the validity of the grading system: (1) weighted scores from the checklist would be highly correlated with the global quality scores (indicating convergent validity); (2) articles with lower quality (based on the global quality score) would have lower weighted scores, and those of higher quality would have higher weighted scores, thus indicating discriminant validity; (3) that there would be strong agreement between the experts' global rating and the grading system (indicating the validity of using the grading systems to represent experts' global rating); and (4) a weighted scoring method would predict the experts' global rating more effectively than an unweighted scoring method (ie, the number of criteria met divided by the total number of criteria in the checklist and multiplied it by 100).

We used the Spearman rho test to examine the correlation between the weighted scores and the global quality scores, and Wilcoxon signed-rank test to determine whether there was a significant difference in the sample distribution and median between the weighted scores and the global quality scores. To examine discriminate validity, a univariate analysis of covariance (ANCOVA) on weighted quality scores was used to detect whether papers with lower or higher quality (represented by the global quality score) have lower or higher weighted scores. The experts' global scores were categorized into four quality categories (group 1 with scores between 0 and 25, group 2 between 25.1 and 50, group 3 between 50.1 and 75, and group 4 between 75.1 and 100). In the ANCOVA, the quality category was the independent variable and expert respondent characteristics were covariates in the model. Expert characteristics included: years in health economics field, the primary working environment, the perceived value of the grading system (1-5 scale, 1 = "not valuable at all" and 5 = "extremely valuable"), and whether they would recommend the instrument to others.

We applied the approach suggested by Bland and Altman to examine the agreement between the experts' global rating and the grading system.^{15,16} The differences between scores measured by the two methods (denoted as $D1$) and the mean (d), and the standard deviation (SD) of the

Results

The literature review resulted in the identification of 19 guidelines and checklists with a total of 357 items (Table 1). The number of items included in each guideline/checklist had a range of 8 to 40, whereas the questions were grouped in 3 to 12 categories. All but one of the 19 guidelines and checklists contained questions in the "analysis" category whereas only four of them contained questions in the "time horizon" and "sponsorship" categories. After removing duplicates, an initial item pool of 151 criteria remained. After rating the individual criterion, 52 (34%) had an average score > 5.5 (on a scale of 1-9) and were believed to be important by at least one member. They were retained for the second round of selection. After members reviewed the final ratings and exchanged their opinions, the steering committee selected a final set of 16 criteria for further testing (Table 2). Based on these 16 criteria, the orthogonal main effect plan was generated containing 60 scenarios of hypothetical health economic studies composed of different combinations of criteria. For example, scenario 1 represented a study that only met criteria 2, 6, 9, and 13, whereas scenario 2 met criteria 1, 2, 7, 8, 9, 14, 15, and 16. The 60 scenarios were divided into 6 blocks of 10.

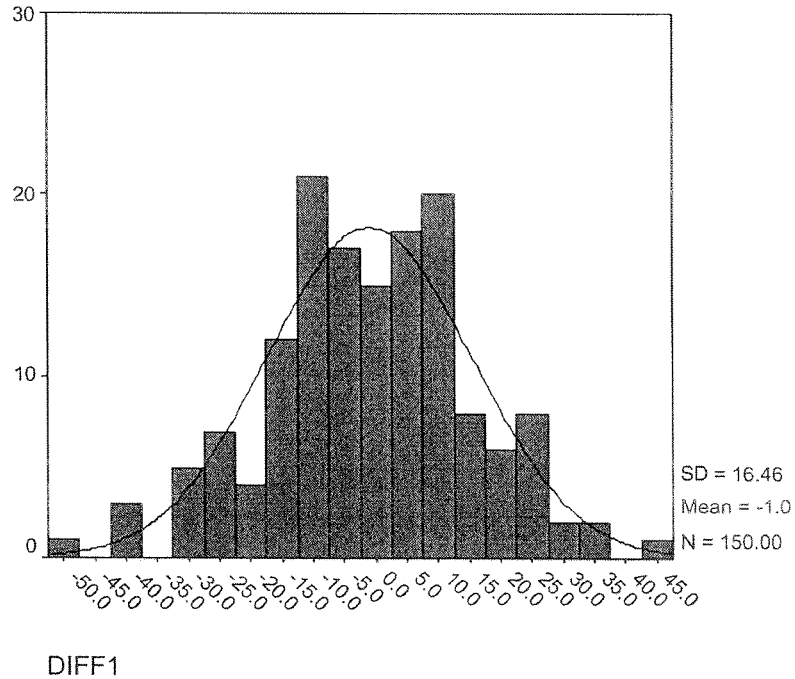
In phase 1, 106 of the 120 (88%) respondents returned the survey. Of the 106 responses, 7 contained illogical ratings, 1 was missing all 10 visual analog scores, and 2 contained 9 of the 10 visual analog scores. Therefore, a total of 978 observations from 98 respondents were used to estimate weights for criteria in the checklist. The participants worked in the field of health economics for an average of 11.4 (± 7.2) years, mostly in an academic setting (77.4%). Approximately 40% rated the value of the grading system as greater than 3 on a scale of 1 to 5 (1 "not valuable at all" and 5 "extremely valuable") whereas 30% rated it as equal to 3, with the mean 3.2 (± 1.1). Before seeing the final instrument, nearly half (49%) of the respondents indicated that they would use a grading system or recommend one to others compared with 28% of the respondents indicating they would not (Table 3).

The random effects GLS regression was based on 978 observations and resulted in nonstandardized coefficients ranging from 1.20 to 8.86 (Table 4). Coefficients for all criteria had a positive sign and were statistically significant ($P < 0.01$) with the exception of criterion 4 ($P = 0.19$). The R^2

differences were first calculated. Provided the differences follow a normal distribution and are independent of the magnitudes of the averages of scores, 95% of differences should lie between the "95% limits of agreement" (ie, $d - 1.96$ SD and $d + 1.96$ SD). If this occurs, it implies that the degree of agreement between the two methods is high. If the differences do not follow a normal distribution and are not independent of the magnitudes of the averages, it implies the agreement is low. The range of the "95% limits of agreement" may also indicate the degree of agreement (ie, the greater the range the lower the agreement). We used a histogram of differences between global quality scores and weighted scores to determine the normality of the distribution of differences (Fig. 1). A scatter plot of differences (y-axis) against averages of weighted scores and global quality scores was produced to determine whether the differences are independent of the averages of scores measured by the two methods (Fig. 2).

Differences between scores measured by the grading system (ie, weighted score) and experts' global quality scores (denoted as D1) were compared with those between scores measured by the unweighted scoring method (ie, unweighted scores) and the global quality scores (denoted as D2) for each of the four quality categories. A t test was used to examine whether the difference between D1 and D2 was significant. It was expected that the difference between D1 and D2 would more likely be significant for papers that only meet few criteria (ie, with poor quality) than those that meet most criteria (ie, with excellent quality). As more criteria are met, the weights for each criterion have less impact on the final score. Multivariate regression analysis identified factors that influenced the predictive validity of the grading system. The dependent variable was the absolute difference between the experts' global score and the score estimated with the grading system while independent variables were the experts' characteristics. Because the impact of "experience" may not be linear, a quadratic for years of experience was added to the regression as another explanatory variable. In addition, the "perceived value" was kept as categorical because it was more informative than making it binary. Analyses were performed with STATA 6.0. (STATA Corporation, College Station, TX) or SPSS (SPSS, Chicago, IL), and the scatter plot was completed with Microsoft Excel 2000 (Microsoft Corporation, Redmond, WA).

FIG. 1. Histogram of differences between global score and weighted score.



within, between, and overall was 0.53, 0.05, and 0.37, respectively. The sum of the nonstandardized coefficients was 93.51, thus the adjustment factor was 1.07.

In phase II, 50 of the 60 (83%) surveys were returned. The validation group worked in the field of health economics for an average of 9.3 (± 7.8) years with the majority working in an academic setting (80%). Fifty six percent rated the perceived

value of the grading system as greater than on the 5-point scale whereas 30% equal to 3. The mean was 3.7 (± 0.9). Sixty four percent of respondents reported that they would use the grading system or recommend it to others (Table 3).

The Spearman correlation coefficient for global scores and weighted scores was 0.783 ($P = 0.000$). The p-value for the Wilcoxon signed rank was 0.51 indicating that there were no significant differ-

FIG. 2. Difference against average for scores measured by two methods (n = 150).

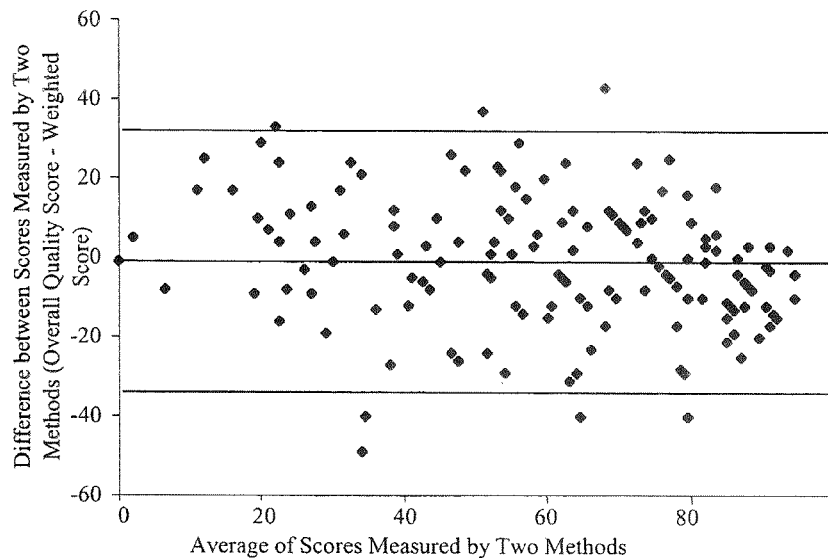


TABLE 2. Criteria Selected for Grading System

1	Was the study objective presented in a clear, specific, and measurable manner?
2	Were the perspective of the analysis (societal, third-party payer, etc.) and reasons for its selection stated?
3	Were variable estimates used in the analysis from the best available source (i.e. Randomized Control Trial—Best, Expert Opinion—Worst)?
4	If estimates came from a subgroup analysis, were the groups prespecified at the beginning of the study? Was uncertainty handled by: 1) statistical analysis to address random events; 2) sensitivity analysis to cover a range of assumptions?
6	Was incremental analysis performed between alternatives for resources and costs?
7	Was the methodology for data abstraction (including value health states and other benefits) stated?
8	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?
9	Was the measurement of costs appropriate and the methodology for the estimation of quantities and unit costs clearly described?
10	Were the primary outcome measure(s) for the economic evaluation clearly stated and were the major short term, long term and negative outcomes included?
11	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?
12	Were the economic model (including structure), study methods and analysis, and the components of the numerator and denominator displayed in a clear transparent manner?
13	Were the choice of economic model, main assumptions and limitations of the study stated and justified? Did the author(s) explicitly discuss direction and magnitude of potential biases?
14	Were the conclusions/recommendations of the study justified and based on the study results?
15	Were there a statement disclosing the source of funding for the study?

The mean difference between global scores and weighted scores (D1) was smaller than the mean difference between global scores and unweighted scores (D2) for papers in all quality categories except category 4. However, only the difference between mean D1 (-4.95 ± 19.26) and mean D2 (-7.12 ± 19.88) for the quality category 1 was statistically significant ($P = 0.017$). Among variables that might affect the accuracy of the grading system in predicting experts' global quality scores, only the respondents' attitude about using the grading system or recommending it to others had a statistically significant negative coefficient ($P = 0.049$) (Table 7). However, the R^2 for the OLS regression was only 0.053.

Discussion

Appraising the quality of health economic evaluations is important for researchers, those performing systematic reviews, peer-reviewers and journal editors, regulatory bodies, and decision-makers attempting to utilize health economic information presented in the peer-reviewed literature. The results of our literature review revealed that although many checklists and recommendations exist, they are qualitative, there is significant

ences in either the distribution or the median between the two scores. ANCOVA revealed that quality category 4 (based on the global score) had the highest mean weighted score $85.49 (\pm 12.56)$ followed by quality categories 3, 2, and 1 with mean weighted scores $66.27 (\pm 17.00)$, $42.29 (\pm 20.62)$, and $23.16 (\pm 18.38)$ respectively ($F_{3,146} = 5.97, P = 0.001$) (Table 5). Figure 1 displays the distribution of the difference between the experts' global score and the weighted score estimated with the grading system. It is approximately normally distributed with a mean difference equal to $-0.98 (\pm 16.46)$. According to Bland and Altman,^{15,16} the higher limit of "95% limits of agreement" for the two methods was $31.28 (d + 1.96 SD)$ whereas the lower limit was $-33.24 (d - 1.96 SD)$. The scatter plot of differences (y-axis) against the means of the global scores and weighted scores (x-axis) reveals that more than 95% of differences lay between the "95% limits of agreement" (Fig. 2). Across all the means of scores measured with the two methods, at least 65% of the differences lay within one SD from the mean. Differences thus appear to be independent of the means. The comparison between the weighted and unweighted scoring methods in predicting experts' global quality scores is reported in Table 6.

TABLE 3. Characteristics of Participants

Characteristic	Phase I (n = 106)		Phase II (n = 50)	
	Frequency (n)	%	Frequency (n)	%
Years work in health economics field*				
1 ~ 5	28	26.4	17	34.0
6 ~ 10	32	30.2	20	40.0
11 ~ 15	19	17.9	4	8.0
16 ~ 20	18	17.0	2	4.0
20+	9	8.5	6	12.0
Missing	0	0	1	2.0
Primary work environment				
Academia	82	77.4	40	80.0
Government	4	3.8	2	4.0
Managed care	2	1.9	1	2.0
Clinical practice	1	0.9	3	6.0
Pharmaceutical industry	4	3.8	2	4.0
Contract research	6	5.7	1	2.0
Other	7	6.6	1	2.0
Value of the grading system [†]				
Less than 3	27	25.5	5	10.0
Equal to 3	32	30.2	15	30.0
Greater than 3	42	39.6	28	56.0
Missing	5	4.7	2	4.0
Will use or recommend others to use the grading system?				
Yes	52	49.1	32	64.0
No	30	28.3	9	18.0
Not sure	19	17.9	8	16.0
Missing	5	4.7	1	2.0

*For the sample in Phase I, mean (\pm SD) = 11.4 (\pm 7.2) years; for the sample in Phase II, mean (\pm SD) = 9.3 (\pm 7.8) years.

[†]For the sample in Phase I, mean (\pm SD) = 3.2 (\pm 1.1); for the sample in Phase II, mean (\pm SD) = 3.7 (\pm 0.9).

variability in the proposed criteria, each criterion is assumed to carry equal weight or importance, and there does not appear to have been any formal validation of the existing instruments. The results of our study indicate that a simple, weighted grading system may accurately predict experts' global ratings of the quality of health economic evaluations, and that an instrument with different weights may have more discriminative power than those with equal weights.

In an effort to increase the validity and comprehensiveness of the grading system, criteria were selected and worded based on the results of the literature review, suggestions from the steering committee, and results of the pilot test. Our survey achieved a high response rate in both phases, and most study respondents indicated that there was perceived value to the concept of a grading system, and that they would use it or recommend the instrument to others.

There are several aspects of our study that speak to the internal validity of the findings. The coefficient for each criterion was positive, the coefficients were statistically significant for all but one criterion, and the sum of the coefficients was nearly 100, indicating good internal consistency. Moreover, we found that the experts' global scores and the weighted scores had similar approximate normal distributions and medians, and the discriminative power of the grading system was greatest for the low quality (low score) studies as expected.

Our study findings are based on the assumption that the experts' global quality rating serves as the gold standard. Given this assumption, the high correlation between the global score and the weighted score demonstrated good convergent validity. The grading system also demonstrated good discriminant validity, with the ability to

TABLE 4. Results of Random Effects GLS Regression and Estimated Weights for Criteria

Criteria	Coefficient	Weight
1	6.89	7
2	3.91	4
3	7.36	8
4	1.20	1
5	8.86	9
6	5.79	6
7	4.46	5
8	6.53	7
9	7.31	8
10	6.02	6
11	6.54	7
12	7.11	8
13	6.23	7
14	5.48	6
15	7.05	8
16	2.77	3
Constant	-0.95	100 (Total)
R ²	Within	0.53
	Between	0.05
	Overall	0.37
Wald χ^2 (16)		977.76
Prob > χ^2		0.00

n = 978.

distinguish economic studies of good quality from those of poor quality. Although good construct validity (both convergent and discriminant validity) is a necessary condition for the grading system to be considered valid, the level of "agreement" indicates that the grading system may be a practical tool for use in appraising the quality of studies (as an expert would with a global score). Our findings suggest that there was moderate agree-

TABLE 5. Means and Standard Deviations for Weighted Quality Scores by Quality Categories

Quality Category (n)	Mean	SD
1 (19)	23.16	18.38
2 (35)	42.29	20.62
3 (55)	66.27	17.00
4 (41)	85.49	12.56
Total (150)	60.47	26.92

n = 150. F (3, 146) = 5.97, P = 0.001.

TABLE 6. Comparisons on D1 (Difference between Global Score and Weighted Score) and D2 (Difference between Global Score and Unweighted Score)*

Category	Mean	SD	t	P
Quality Category 1 (Global Score: 0-25; n = 19)	-4.95	19.26	2.62	0.017
D1	-4.95	19.26		
D2	-7.12	19.88		
Quality Category 2 (Global Score: 25.1-50; n = 35)	-1.34	18.48	1.89	0.067
D1	-1.34	18.48		
D2	-2.45	18.70		
Quality Category 3 (Global Score: 50.1-75; n = 55)	0.31	16.25	-1.17	0.247
D1	0.31	16.25		
D2	0.90	16.08		
Quality Category 4 (Global Score: 75.1-100; n = 41)	-0.56	13.64	-1.90	0.065
D1	-0.56	13.64		
D2	0.17	12.97		
All Papers (Global Score: 0-100; n = 150)	-0.98	16.46	0.41	0.683
D1	-0.98	16.46		
D2	-1.10	16.53		
*D1 = Global Score—Weighted Score; D2 = Global Score—Unweighted Score.				

ment between the grading system and the global rating by experts. Moreover, the differences did not appear to depend on the mean values, indicating that the degree of agreement between the grading system and the global rating by experts does not vary significantly when used to evaluate papers of different quality. The finding of good construct validity and moderate agreement is important, because data that seem to be in poor agreement can produce quite high correlations, implying convergent validity.¹⁵

TABLE 7. Results of the Multivariate Regression on the Accuracy of the Grading System in Predicting Global Score

Variable	Coefficient	P
Intercept	13.51	0.000
Years of experience	-0.22	0.165
Years of experience 2	2.34E-03	0.174
Academic	2.94	0.168
Perceived value	0.25	0.731
Use/recommend	-3.77	0.049

R² = 0.053.
n = 150.

Results from the comparison between the weighted and unweighted scoring methods indicate that the grading system predicted the global quality score of papers with lower quality (especially for those in quality category 1) better than the unweighted scoring method. Intuitively this should be expected. Most of the weights of the grading system are approximately 6 (eg, three are 6, four are 7, and four are 8), which happens to be the average weight (100/16) for every criterion in the unweighted scoring method, thus weighted and unweighted scores are more likely to be similar for papers meeting most of the criteria in the instrument. Given that many published studies have been viewed as being of poor quality,⁵⁻⁷ it may be more discriminating to use the grading system rather than a guideline or checklist with unweighted scoring methods, when evaluating health economic analyses.

Although the coefficient for criterion 4 was not significant, criteria 4 was retained in the grading system because "predefining the subgroups when estimates from analyses based on them are used" was emphasized in the Canadian Guidelines¹⁷ and reinforced by the steering committee. The steering committee emphasized the importance of subgroup analyses being prespecified and based on sound clinical rationale rather than attempts to obtain favorable results. The committee also emphasized the empirical data suggesting that industry sponsored work may be biased and more likely to result in favorable findings, and thus believed that item 16 was critical to attaining complete transparency, and that transparency may be related to quality. Interestingly, with a statistically significant coefficient, criterion 16 pertains to the disclosure of the source of funding, and had the smallest weight next to criterion 4. It is unclear whether the respondents simply did not view the funding source as important as other criteria, or they thought that the type/nature of the funding source plays a more critical role to the quality of study than just disclosing the information. Although not indicated in pilot testing or in expert comments, some respondents might have found that the wording of both criteria was unclear, or the concept unfamiliar. Alternatively, the large number of criteria presented at the same time (ie, 16) could have affected the results of the conjoint analysis.

The final grading system contained many items similar to those found in the BMJ referees' checklist,¹² Drummond's criteria,³² and the Canadian

guidelines.¹⁷ However, several new items were added regarding potential biases, the source of funding, subgroup analysis, and validity and reliability of outcomes measures. Some participants in Phase II Survey suggested that users should be allowed to reweigh the grades by removing any criteria that they believed not applicable to the study under review. We instructed users of the grading system to reward the weights of these criteria during the validation process.

There are several limitations present in the study. First, several respondents in both groups suggested using a 3 or 5 point scale rather than the "yes/no" response scale. We believed that it would make the scoring system overly complicated and reduce the potential value of a simple grading system for quality appraisal. However, we recognize that the dichotomous response scale might limit ability of the grading system to differentiate studies on a more granular level, such as those of moderate quality from those of high quality. Further work will be required to determine the incremental value of a more complex response scale. Second, most respondents in the development and validation studies were recruited from academic institutions. Although the use of health economics experts in academia may have resulted in a more reliable gold standard, further studies should test the utility and value of the grading system in a more diverse sample of users of health economic studies. Finally, because of a limited sample size, there is limited power to detect whether the predictive validity of the grading system was dependent upon expert characteristics. In relation to the conjoint analysis, we assumed that the scoring function is additive and preferences for individual criterion are independent. Researchers have suggested that future work applying conjoint analysis in health economics should explore methods beyond the traditional linear additive model.^{18,19} Furthermore, only 60 out of possible 65536 (2^{16}) scenarios were selected for the conjoint analysis survey. Although this provided adequate power to determine the weights, it is unclear whether the results could be sensitive to how the blocks of 10 were chosen, or how the questionnaire was designed. More importantly, it is also not clear to what degree the 60 scenarios represented the evaluation studies commonly seen in the literature. This is an issue that has received little attention in the health economic literature and warrants further investigation.¹⁹

Acknowledgments

The authors thank participants in both surveys who provided us with many constructive comments, and the International Health Economics Association and Society for Medical Decision Making for making their membership lists available for our research. The authors also acknowledge the experience and knowledge shared with us from Miranda Mugford, PhD, and numerous members of the Cochrane Economics Method Group.

References

1. Gerard K, Seymour J, Smoker I. A tool to improve quality of reporting published economic analyses. *Int J Technol Assess Health Care* 2000;16:100-110.
2. Canadian Coordinating Office of Health Technology Assessment. Guidelines for economic evaluation of pharmaceuticals: Canada. 2nd. 1997. Ottawa, CCOHTA Publications.
3. Commonwealth Department of Health HaCS. Guidelines for the Pharmaceutical Industry on Preparation of Submissions to the Pharmaceutical Benefits Advisory Committee. 1995. Canberra: Australian Government Publishing Service.
4. Elixhauser A, Luce BR, Taylor WR, et al. Health care CBA/CBA: an update on the growth and composition of the literature. *Med Care* 1993;31(suppl):J51-149.
5. Gerard K. Cost-utility in practice: a policy maker's guide to the state of the art. *Health Policy* 1992;21:249-279.
6. Udvarhelyi IS, Colditz GA, Rai A, Epstein AM. Cost-effectiveness and cost-benefit analyses in the medical literature. Are the methods being used correctly? *Ann Intern Med* 1992;116:238-244.
7. Adams ME, McCall NT, Gray DT, et al. Economic analysis in randomized control trials. *Med Care* 1992;30:231-243.
8. Neumann PJ, Stone PW, Chapman RH, et al. The quality of reporting in published cost-utility analyses. *Ann Intern Med* 2000;132:964-972.
9. Detsky AS. Guideline for economic analysis of pharmaceutical products: a draft document for Ontario and Canada. *Pharmacoeconomics* 1993;3:354-361.
10. Report from the Canadian Coordinating Office for Health Technology Assessment (CCOHTA). Guidelines for economic evaluation of pharmaceuticals: Canada. *Int J Technol Assess Health Care* 1995;11:796-797.
11. Gold MR, Siegel JE, Russell LB, et al. Cost-effectiveness Analysis in Health and Medicine. New York, NY: Oxford University Press; 1996.

Respondents in both phases revealed some

degree of reservation about the value of the grading system with those in Phase II Survey being more enthusiastic about using it or recommending it to others. We believed that the exercise of grading articles with the grading system might have resulted in a more significant impact on how respondents perceived the value of the grading system than survey question asking about years worked in health economics field. Although the statistical results indicated that the newly devised grading system has some merits, additional efforts may be required to help both "producers" and "consumers" of cost-effectiveness studies derive the optimal value of the grading system. The result may also indicate that readers should not consider the grading system as a surrogate for a detailed clinical and methodological review performed by experts. Practically, health plan, government, and police decision-makers face complex decisions that are impacted by several competing factors: clinical, economic, political, and ethical. The value of high quality health economic evaluations might be outweighed by ethical or political factors in actual decision-making. We also recognize that in some instances specific checklists and appraisal systems may be mandated by organizations. However, we believe that the formal validation of the weighted grading system may contribute to its adoption in diverse decision-making settings.^{20-31,33-35}

Conclusion

The grading system appears to be practical, internally consistent, and valid for measuring experts' perceived quality of health economic studies. While providing a more robust method for appraising studies than current qualitative and un-weighted instruments, the utility and reliability of this instrument in different health care settings deserves further exploration. Most importantly, we must determine whether the application of this tool will actually sway the "selected" use of higher quality health economic information, and whether this more "selective" use will impact clinical and resource allocation decision-making to any meaningful degree. In the meantime, we propose the use of this practical weighted appraisal "tool" to investigators, reviewers and editors needing a mechanism to assess clearly and fairly the quality of health economic evaluations.

12. **Drummond MF, Jefferson TO.** Guidelines for authors and peer reviewers of economic submissions to the BMJ. *BMJ* 1996;313:275–283.
13. **Weinstein MC, Siegel JE, Gold MR, et al.** Recommendations of the Panel on Cost-effectiveness in Health and Medicine. *JAMA* 1996;276:1253–1258.
14. **Neumann PJ, Hermann RC, Berenbaum PA, et al.** Methods of cost-effectiveness analysis in the assessment of new drugs for Alzheimer's disease. *Psychiatr Serv* 1997;48:1440–1444.
15. **Bland JM, Altman DG.** Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1:307–310.
16. **Bland JM, Altman DG.** Comparing two methods of clinical measurement: a personal history. *Int J Epidemiol* 1995;24(Suppl 1):S7–14.
17. **Torrance GW, Blaker D, Detsky A, et al.** Canadian guidelines for economic evaluation of pharmaceuticals. Canadian Collaborative Workshop for Pharmacoeconomics. *Pharmacoeconomics* 1996;9:535–559.
18. **Ratcliffe J, Buxton M.** Patients' preferences regarding the process and outcomes of life-saving technology. An application of conjoint analysis to liver transplantation. *Int J Technol Assess Health Care* 1999;15:340–351.
19. **Jan S, Mooney G, Ryan M, et al.** The use of conjoint analysis to elicit community preferences in public health research: a case study of hospital services in South Australia. *Aust N Z J Public Health* 2000;24:64–70.
20. **Hill SR, Mitchell AS, Henry DA.** Problems with the interpretation of pharmacoeconomic analyses: a review of submissions to the Australian Pharmaceutical Benefits Scheme. *JAMA* 2000;283:2116–2121.
21. **McNee W.** RE: M Thomas. The change of cost: reference-based pricing and the statins 1999;15:535–538. *Can J Cardiol* 1999;15:1287.
22. **Sanchez LA.** Evaluating the quality of published pharmacoeconomic evaluations. *Hosp Pharm* 1995;30:146–148, 151–152.
23. **Mullins CD, Ogilvie S.** Emerging standardization in pharmacoeconomics. *Clin Ther* 1998;20:1194–1202.
24. **Baladi JF, Menon D, Otten N.** Use of economic evaluation guidelines: 2 years' experience in Canada. *Health Econ* 1998;7:221–227.
25. **Byford S, Palmer S.** Common errors and controversies in pharmacoeconomic analyses. *Pharmacoeconomics* 1998;13:659–666.
26. **Alban A, Gyldmark M, Pedersen AV, et al.** The Danish approach to standards for economic evaluation methodologies. *Pharmacoeconomics* 1997;12:627–636.
27. **Menon D, Schubert F, Torrance GW.** Canada's new guidelines for the economic evaluation of pharmaceuticals. *Med Care* 1996;34(12 Suppl):DS77–DS86.
28. **Clemens K, Townsend R, Luscombe F, et al.** Methodological and conduct principles for pharmacoeconomic research. Pharmaceutical Research and Manufacturers of America. *Pharmacoeconomics* 1995;8:169–174.
29. **Sacristan JA, Soto J, Galende I.** Evaluation of pharmacoeconomic studies: utilization of a checklist. *Ann Pharmacother* 1993;27:1126–1133.
30. **Guyatt G, Drummond M, Feeny D, et al.** Guidelines for the clinical and economic evaluation of health care technologies. *Soc Sci Med* 1986;22:393–408.
31. Economic analysis of health care technology. A report on principles. Task Force on Principles for Economic Analysis of Health Care Technology *Ann Intern Med* 1995;123:61–70.
32. **Drummond M, O'Brien B, Stoddart G, et al.** Critical assessment of economic evaluation. Methods for the Economic Evaluation of Health Care Programmes. Oxford: Oxford Medical Publications; 1997:27–51.
33. **Nixon J, Stoykova B, Glanville J, et al.** The U.K. NHS economic evaluation database. Economic issues in evaluations of health technology. *Int J Technol Assess Health Care* 2000;16:731–742.
34. **Drummond MF, Richardson WS, O'Brien BJ, et al.** Users' guides to the medical literature. XIII. How to use an article on economic analysis of clinical practice. A. Are the results of the study valid? Evidence-Based Medicine Working Group *JAMA* 1997;277:1552–1557.
35. **Milne RJ.** Pharmacoeconomic Models in Disease Management. A Guide for the Novice or the Perplexed. *Dis Manage HealthOutcomes* 1998;4:120–135.

