

The Quality of Published Health Economic Analyses in Digestive Diseases: A Systematic Review and Quantitative Appraisal

BRENNAN M. R. SPIEGEL,^{*,†,§,||} LAURA E. TARGOWNIK,[¶] FASIHA KANWAL,^{†,§,||} VINCENT DEROSA,^{*} GARETH S. DULAI,^{†,§,||} IAN M. GRALNEK,^{*,†,§,||} and CHIUN-FANG CHIOU[#]

^{*}Division of Gastroenterology, Veteran's Affairs Greater Los Angeles Healthcare System, Los Angeles, California; [†]Division of Digestive Diseases, David Geffen School of Medicine at University of California Los Angeles, Los Angeles, California; [¶]University of Manitoba, Winnipeg, Canada; [§]CURE Digestive Diseases Research Center, Los Angeles, California; ^{||}Center for the Study of Digestive Healthcare Quality and Outcomes, Los Angeles, California; and [#]Zynx Health, a Cerner Company, Beverly Hills, California

Background & Aims: Health economic analyses are increasingly common in the digestive diseases literature and often are cited to frame practice guidelines. Although clinical trials are subjected routinely to critical appraisal, there has been no attempt to appraise the quality of health economic analyses with a validated instrument. We sought to appraise the quality of health economic analyses in digestive diseases, and to identify predictors of study quality. **Methods:** We performed a systematic review to identify digestive disease health economic analyses published since 1980. We assessed these studies using the Quality of Health Economic Studies (QHES), a validated quality-scoring instrument (score range = 0-100; >75 = high quality). We conducted regression analysis to identify predictors of high quality. **Results:** Of 186 identified analyses, 29% were high quality, 71% failed to address potential model biases, 52% failed to disclose conflicts of interest, and 74% failed to describe methods for deriving the model assumptions. Four factors predicted high quality in logistic regression: (1) one or more authors had an advanced degree in health services or a related field (odds ratio for high quality, 5.0; 95% confidence interval, 2.6-9.3); (2) the study used decision-analysis software package (odds ratio, 2.4; 95% confidence interval, 1.2-4.7); (3) the study was federally funded (odds ratio, 2.2; 95% confidence interval, 1.2-4.1); and (4) the study cited the National Panel on Cost Effectiveness guidelines (odds ratio, 2.1; 95% confidence interval, 1.1-4.2). **Conclusions:** Less than one third of health economic analyses in digestive diseases meet criteria for high quality. Study quality is limited by factors that potentially can be remedied. These data may be used to focus the attention of journal editors and peer reviewers to ensure the future high quality of health economic analyses in digestive diseases.

Health economic analyses are increasingly common in the published digestive diseases literature. Although fewer than 20 economic analyses were published between 1980 and 1995, over 175 have been published

since 1995. This rapid increase is driven by several factors including: (1) the influx of new medical therapies in need of pharmacoeconomic appraisal; (2) the rapid development of novel yet incompletely tested technologies; (3) the increasing pressure to defend the widespread use of expensive endoscopic procedures in an era of resource constraint¹; and (4) the development and dissemination of guidelines for the conduct of health economic analyses in digestive diseases.^{2,3} In light of these factors, health economic analyses are now cited routinely in framing policy statements, consensus guidelines, and professional society technical reviews.⁴⁻⁶

Despite the emphasis placed on developing and publishing health economic analyses in digestive diseases, there have been few attempts to systematically review and grade the quality of this expanding body of literature. Although randomized clinical trials routinely are subjected to systematic review, quantitative assessment, and critical appraisal, a similar emphasis has not been placed on health economic analyses in digestive diseases. This is problematic because most reviewers and consumers of health economic analyses lack formal training to rate the quality of these reports objectively. For example, one survey of medical decision makers found that most regard health economic analyses to be black boxes, despite relying on the analyses to formulate health care policy.⁷ In light of this lack of connection between perceived importance and degree of critical appraisal, it is worthwhile to systematically examine the quality of health economic analyses being published in digestive diseases.

There is reason to believe that health economic analyses in digestive diseases are not of uniform quality and

Abbreviation used in this paper: QHES, Quality of Health Economic Studies.

© 2004 by the American Gastroenterological Association
0016-5085/04/\$30.00
doi:10.1053/j.gastro.2004.04.020

Table 1. Search Strategy for Systematic Review

Group	Search terms	Significance of grouping
1	<i>Alimentary Pharmacology and Therapeutics, American Journal of Gastroenterology, Canadian Journal of Gastroenterology, Digestion, Digestive Diseases and Sciences, Diseases of the Colon and Rectum, Endoscopy, Gastroenterology, Gastrointestinal Endoscopy, Gut, Helicobacter, Hepatology, Inflammatory Bowel Diseases, Journal of Clinical Gastroenterology, Journal of Gastroenterology, Journal of Gastroenterology and Hepatology, Journal of Hepatology, Liver, Liver Transplantation, Pancreas, Scandanavian Journal of Gastroenterology</i>	Targeted digestive diseases journals
2	<i>American Journal of Medicine, Annals of Internal Medicine, Archives of Internal Medicine, British Medical Journal, Journal of Internal Medicine, Lancet, Mayo Clinic Proceedings, Medical Care, Journal of the American Medical Association, New England Journal of Medicine</i>	Targeted internal medicine journals
3	Barrett* (MeSH) OR esophag* (tw) OR oesophag* (tw) OR GERD (MeSH) OR GORD (tw) OR gastroesophag* (tw) OR gastro-oesophag* (tw) OR pancre* (tw) OR peptic ulcer (MeSH) OR gastric (tw) OR pylori (tw) OR hepat* (tw) OR liver (tw) OR cirrhosis (tw) OR endoscop* (tw) OR varice* (tw) OR ulcer (tw) OR biliary (tw) OR ERCP (tw) OR colon* (tw) OR colon cancer (MeSH) OR colon polyp (tw) OR diverticul* (tw) OR dyspepsia (MeSH and tw) OR bowel (tw) OR intestin* OR (gastrointestinal) (tw)	Broad key words to focus search on digestive diseases
4	cost-effectiveness (tw and pt) OR cost effectiveness (tw) OR cost-utility (tw) OR cost utility (tw) OR decision analysis (tw and pt) OR economic analysis (tw and pt) OR economic model (tw) OR decision model (tw and pt) OR Markov (tw)	Included study types
5	(Letter (pt) OR editorial (pt) OR review (pt))	Excluded study types

NOTE. *Indicates key word truncation. The 5 search groups were combined as follows: "(1 OR 2) AND 3 AND 4 NOT 5." MeSH, medical subject heading; pt, publication type; tw, text word.

that few meet all standards for high quality. For example, Marshall et al.⁸ showed that only 70% of cost-effectiveness analyses published before 1998 met broad criteria for appropriateness as measured by an nonvalidated subjective assessment. Ofman et al.⁹ found that only 27% of cost-effectiveness analyses in gastroesophageal reflux disease met criteria for high quality as measured by the Quality of Health Economic Studies (QHES) instrument, which is a validated 16-item instrument designed to measure the quality of health economic analyses. These findings suggest that the general quality of health economic analyses in digestive diseases is not high—a concerning realization in light of the increasing reliance on these analyses to formulate policy decisions.

We therefore performed a systematic review to identify and rate the published health economic analyses in digestive diseases. Our general objective was to focus the attention of journal editors, peer reviewers, clinicians, and policy makers on specific data to ensure the future high quality of health economic analyses in digestive diseases. Our specific aims were to measure the quality of these analyses using a validated instrument, and to identify predictors of high study quality.

Materials and Methods

Systematic Review

We performed a systematic review of English-language gastroenterology, endoscopy, hepatology, and internal medi-

cine journals with an impact factor greater than 1.0 to identify health economic analyses in digestive diseases published between January 1980 and January 2004. Table 1 lists the journals subjected to our search and provides the key words and search strings used to perform the systematic review. Three reviewers with experience in health economic analyses independently assessed the relevancy of all titles generated from this initial screen (B.M.R.S., L.E.T., V.D.). We excluded titles for the following reasons: (1) it did not concern a clinical question regarding human subjects, (2) it did not concern a digestive disease, and (3) it was not written in English. We then each independently assessed the relevancy of all abstracts corresponding with the remaining titles. We excluded abstracts for the following reasons: (1) it was not a modeled decision analysis (including cost-effectiveness analysis, cost-utility analysis, or cost-minimization analysis), and (2) it did not compare at least 2 strategies. We then each independently assessed the relevancy of all manuscripts corresponding with the remaining abstracts, and excluded manuscripts that met any of the earlier-described exclusion criteria. All disagreements were settled by consensus between the 3 reviewers.

Quality Scoring

We used the QHES as our outcome measure. The QHES is a validated measure of quality for cost-effectiveness, cost-utility, and cost-minimization analyses (Table 2).^{9,10} The QHES contains 16 items that were selected by a panel of economic experts with experience in health economic analysis. Each item carries a weighted point value that was generated from survey data of a second international panel of health economists. The scale was validated prospectively by using a

Table 2. The QHES Instrument

	Questions	Point	Yes	No
1	Was the study objective presented in a clear, specific, and measurable manner?	7		
2	Were the perspective of the analysis (societal, third-party payer, etc.) and reasons for its selection stated?	4		
3	Were variable estimates used in the analysis from the best available source (i.e., Randomized Control Trial — Best, Expert Opinion — Worst)?	8		
4	If estimates came from a subgroup analysis, were the groups prespecified at the beginning of the study?	1		
5	Was uncertainty handled by: 1) statistical analysis to address random events; 2) sensitivity analysis to cover a range of assumptions? ^a	9		
6	Was incremental analysis performed between alternatives for resources and costs?	6		
7	Was the methodology for data abstraction (including the value of health states and other benefits) stated? ^b	5		
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?	7		
9	Was the measurement of costs appropriate and the methodology for the estimation of quantities and unit costs clearly described?	8		
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?	6		
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?	7		
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?	8		
13	Were the choice of economic model, main assumptions and limitations of the study stated and justified?	7		
14	Did the author(s) explicitly discuss direction and magnitude of potential biases?	6		
15	Were the conclusions/recommendations of the study justified and based on the study results?	8		
16	Was there a statement disclosing the source of funding for the study?	3		
Total points		100		

NOTE. There are 16 dichotomous (yes/no) items in this questionnaire, each weighted by importance as determined by an expert panel of health economists. The quality score is calculated by subtracting points for questions answered with no from 100. Therefore, the highest possible score is 100, and the lowest is 0. Studies with a score exceeding 75 points are considered of high quality.

^aAdequate sensitivity analysis includes 2-way analysis and beyond (e.g. Monte Carlo analysis). Therefore, sensitivity analyses limited to 1-way results are not adequate to receive points for item 5.

^bAt a minimum, studies should describe clearly the databases searched, key words used, dates queried, or prioritization scheme for study types.

third panel of health economists who compared their subjective global assessment of sample studies (using a visual analog scale) with scores obtained by the QHES. Subsequent correlation between subjective assessment and QHES scores was high ($r = 0.78$). Therefore, the QHES has data supporting its content and construct validity.^{9,10} The QHES is scored on a 0 (lowest quality) to 100 (highest quality) scale. This continuous scale also may be dichotomized to include high-quality (75–100 points) and not high-quality studies (<75 points). Alternatively, studies may be grouped by the following quartiles: (1) extremely poor quality (0–24); (2) poor quality (25–49); (3) fair quality (50–74); and (4) high quality (75–100).

Before scoring, each of the reviewers received training in the proper application of the QHES by the developers of the instrument. After this training, the reviewers independently scored a 10% random sample of the full study set and measured interrater agreement with 3 statistics: (1) comparison of mean QHES scores between reviewers using analysis of variance (ANOVA); (2) comparison of dichotomized scoring (i.e., rating >75 vs. <75 on the QHES) between reviewers using a κ statistic; and (3) proportion of studies scored within 10 points between reviewers. Disagreements were settled by consensus and iterative training sets were conducted until the

following thresholds were met: (1) P value for ANOVA across reviewer scores exceeded $P > 0.20$; (2) κ statistic exceeded 0.70, and (3) proportion of studies scored within 10 points exceeded 0.80.

Once adequate agreement was reached on the test set, the reviewer divided the full study set and independently abstracted QHES scores for their assigned studies. The reviewers were blinded to the authors, institution, and source journal. Each manuscript was scored for quality between 0 and 100 using the QHES instrument.

Data Abstraction

We collected data for each manuscript across a range of variables, including funding source, author characteristics, journal characteristics, journal impact factor, presence of editorial, Web of Science citations per year, and reference to National Panel on Cost Effectiveness in Health and Medicine guidelines,¹¹ among others. Table 3 displays the full list of abstracted variables.

Adjusting for Potential Reviewer Bias

Full blinding was not possible because the reviewers are familiar with many authors in the field of digestive diseases

Table 3. Hypothesized Independent Predictors of Study Quality

Predictor	Type of variable
Author characteristics	
Any author with advanced training in health services or related field, including epidemiology, biostatistics, economics, or business administration ^a	Dichotomous
Number of authors on study	Continuous
Journal characteristics	
Journal impact factor	Continuous
Country of origin	Dichotomous
Type of journal	Dichotomous
Study characteristics	
Number of compared strategies in analysis	Continuous
Use of a Markov model	Dichotomous
Use of a Monte Carlo analysis	Dichotomous
Use of a decision-analysis software package	Dichotomous
Citation of National Panel on Cost-Effectiveness in Health and Medicine panel in reference list	Dichotomous
Presence of accompanying editorial	Dichotomous
Rate of Web of Science citations per year	Continuous
Year of publication	Continuous
Type of funding (federal, industry, none)	Categoric
Content area	
Biliary endoscopy	Dichotomous
Viral hepatitis	Dichotomous
Nonviral hepatitis	Dichotomous
<i>Helicobacter pylori</i> and dyspepsia	Dichotomous
Gastroesophageal reflux disease and/or Barrett's esophagus	Dichotomous
Colorectal cancer	Dichotomous
Gastrointestinal bleeding	Dichotomous

^aIn journals that did not provide author degrees in the title line, we used originating departments as a surrogate measure of training. Therefore, any report originating from a department of Health Services, Economics, Public Health, Epidemiology, Biostatistics, or Business received credit when author degrees were not available.

health economics. We therefore devised a bias variable to test the influence of this potential confounder. To measure bias, we reviewed each of the blinded manuscripts to determine whether we could identify one or more of the authors despite the blinding. We then removed the blind. If we correctly identified the author of a study then we assessed a yes for the bias variable. If we could not correctly identify the author then we assessed a no for the bias variable. We then conducted a *t* test to compare the mean QHES score between biased studies and unbiased studies.

Descriptive Analyses

We conducted the following descriptive analyses to measure the quality of the sample studies: (1) mean quality score of all studies; (2) percentage of studies in each quality quartile; (3) mean quality score by year of publication; (4) mean quality score for studies published before vs. after 1996 (the year of publication of the National Panel on Cost-Effectiveness in Health and Medicine guidelines); and (5) frequency of positive endorsement for each of the 16 QHES criteria.

Regression Analyses

Variable selection. We conducted a linear regression analysis to identify predictors of study quality. Before conducting the analysis, we first performed collinearity testing to measure the relationship among independent predictors listed in Table 3. The purpose of this analysis was to identify

potentially redundant variables and, in doing so, to help plan the subsequent multivariable analysis.

We created potential interaction terms based on a priori hypotheses and tested their impact by comparing full and reduced models. We then conducted single coefficient and batch F-tests to evaluate the effect of individual and groups of predictors (including posited interaction terms) on the full model results. This process was guided by the results of the correlation matrix and regression coefficients in the full model. Based on the univariate analyses, correlation matrix, and a priori hypotheses, we identified a subset of variables to test in the final model selection process.

Model selection. We designed the final regression model to balance parsimony with goodness of fit. Toward that end, we performed a best-subset analysis, plotting the number of parameters by R^2 . The point of diminishing returns served as a reference to help identify the optimal number of variables in the final model. We performed an additional best-subset analysis to rank-order the 15 best models stratified by the number of parameters. We then ranked these models qualitatively using a combination of parsimony, C_p statistic, and content. We assessed the highest-ranked model for multicollinearity with the variance inflation factor statistic. We assessed normality with a normal quantile residual plot, and assessed equal variance with additional residual plots. We identified potentially influential points by measuring the Cook's distances, and measured the effect of these points by

conducting robust regression analysis. We report the final linear regression model results in terms of the regression coefficients, standard errors, and *P* values. We then performed a separate logistic regression analysis to measure the ability of each variable to predict high-quality studies (QHES > 75). We reported the odds ratios with 95% confidence intervals and *P* values for each independent predictor.

Results

Interrater Agreement

The mean scores (\pm SD) for the 10% random sample training set for the 3 reviewers were: 60.8 ± 14 , 56.0 ± 16 , and 59.5 ± 15 (*P* value for ANOVA, 0.58). The agreement for dichotomous scoring was high ($\kappa = 0.8$). The 3 reviewers scored 85% of the studies in the test set within 10 points. Therefore, the reviewers achieved excellent interrater agreement.

Descriptive Analyses

Table 4 shows the results of the descriptive analyses. There were 186 studies included in the final dataset (135 from gastrointestinal journals, 51 from general internal medicine journals). The mean QHES score was 63 ± 18 and 29% of the studies met the criteria for high quality (QHES > 75). Studies published after 1996 (the year the National Panel on Cost-Effectiveness in Health and Medicine guideline was published) were of significantly higher quality than those published before 1996 (*P* < 0.001).

Among all studies, 81% specified a clear and measurable objective and 91% provided conclusions and recommendations that were appropriately based on the study results. In contrast, fewer than half (48%) specified whether there was a source of funding for the study, only 29% explicitly addressed potential sources of bias in the model estimates and structure, and fewer than one-fifth (17%) described the perspective of the analysis and the reasons for its selection. Fifty-five percent of the studies used valid and reliable outcome measures, 26% provided details regarding the method of data abstraction for the base-case probability estimates, and 52% performed adequate sensitivity analyses (defined as 2-way analyses or beyond).

Regression Analyses

Diagnostic statistics revealed no deviation from the normality or equal variance assumptions, no evidence of severe outliers, no evidence of severe multicollinearity, and no evidence of bias. The final linear regression model (using continuous QHES score as the dependent variable) included 5 factors that independently predicted high

Table 4. Results of Descriptive Analyses

Variable	Point result	Spread
Summary statistics		
Mean	63.3	18.4 (SD)
Median	64	23.5 (IQR)
Score range	Quality quartile	Percentage of studies
Breakdown by QHES quartiles		
75–100	High	29% (N = 54)
50–74	Fair	52% (N = 97)
25–49	Poor	16% (N = 30)
0–24	Extremely poor	3% (N = 5)
Year range	N	Mean score
Mean QHES score by year of publication		
Before 1990	8	55.1
1990–1993	9	54.1
1994–1996	14	48.4
1997–1999	65	63.0
2000–2004	90	67.6
Before 1996	31	50.7 \pm 17.8
After 1996	155	65.5 \pm 18.0
		<i>P</i> value \leq 0.001
QHES criterion	Frequency	Percentage
Frequency of endorsement by QHES criterion		
QHES 1	150	81%
QHES 2	32	17%
QHES 3	153	82%
QHES 4	173	93%
QHES 5	97	52%
QHES 6	126	68%
QHES 7	48	26%
QHES 8	117	63%
QHES 9	119	64%
QHES 10	136	73%
QHES 11	102	55%
QHES 12	141	76%
QHES 13	136	73%
QHES 14	54	29%
QHES 15	169	91%
QHES 16	89	48%

quality at the $\alpha = 0.05$ level in order of significance ($R^2 = 0.39$): (1) at least one investigator had an advanced degree in health services or a related field (*P* \leq 0.0001); (2) a decision-analysis software package was used to perform the study (*P* = 0.007); (3) the study cited National Panel on Cost-Effectiveness in Health and Medicine guidelines (*P* = 0.01); (4) the study received federal funding (*P* = 0.02); and (5) a high impact factor of the source journal (every 5-point increase in impact factor increased the QHES score by 3.6 points; *P* = 0.03).

Table 5 shows the results of the logistic regression analysis (using the dichotomous QHES score as the dependent variable). Studies conducted by investigators with advanced training in health services or a related field had a 5.0 times higher chance of being high quality

Table 5. Results of Final Linear Regression Analysis

Variable	Parameter estimate	Standard error	P value
Author(s) have advanced training in health services	12.4	2.6	<0.0001
Study used decision-analysis software package	7.9	2.9	0.007
Study cited Gold Criteria in reference list	7.6	2.9	0.01
Study received federal funding	6.8	2.8	0.02
Impact factor of source journal	0.73	0.33	0.03
Year of publication	0.73	0.39	0.06
Study published in U.S. journal	3.9	3.0	0.21
Study received industry funding	2.7	3.3	0.41
Study published in subspecialty journal	-2.3	3.8	0.55

than those by investigators without advanced training. Studies that used a decision-analysis software package had a 2.4 times greater chance of being high quality compared with those that did not. Studies that received federal funding had a 2.2 times greater chance of being high quality compared with those that were not federally funded. The odds ratios with 95% confidence intervals for all the predictors in the final logistic model are listed in Table 6.

Influence of Potential Reviewer Bias

The reviewers were able to identify correctly the authors of 23 studies (12.4%) despite blinding. These studies were considered potentially biased given the reviewers' pre-assessment knowledge. There was no significant difference between the mean QHES score in potentially biased studies compared with potentially unbiased studies (66.0 ± 16.3 vs. 63.0 ± 14.8 ; $P = 0.43$). This parameter remained nonsignificant in multivariable regression analysis.

Conclusions

In light of the increasing reliance on health economic analyses to formulate policy decisions in digestive diseases, it is important to systematically rate and describe the quality of this body of literature. Our analysis

reveals that less than one third of health economic analyses in digestive diseases meet criteria for high quality as measured by a validated instrument. Traditional surrogate markers of quality (e.g., presence of an editorial, number of Web of Science citations per year) do not appear to predict the quality of health economic analyses in digestive diseases. In contrast, quality appears highly dependent on author training, source of funding (federally funded studies are of higher quality than non-federally funded), use of computerized software packages, and adherence to national methodology guidelines.

Health economic analyses are hypothetical depictions of clinical practice. The validity of these studies depends on the implicit assumption that they incorporate valid and reliable data reflecting clinical reality.¹² It is therefore concerning that three quarters of published analyses fail to provide details regarding the method of data abstraction for the probability estimates. Without knowing the authors' methods for generating base-case probability estimates, readers cannot know whether the estimates are valid.

Moreover, our study revealed that 71% of published analyses failed to address potential biases in the model assumptions. This shortcoming is of more than academic interest because the results of health economic models often are dependent on slight variations in the underlying assumptions. Health economists recommend guarding against misleading results by systematically biasing the model parameters against the study hypothesis (and in favor of the null hypothesis).^{9-11,13} This form of a *fortiori*¹³ (Latin: a stronger reason) analysis is a method of addressing uncertainty by stacking the cards against one alternative (generally the one intuitively preferred) in favor of another. If the preferred strategy remains favorable despite this bias, then the analyst builds a stronger case in its favor. For example, Ofman et al.,¹⁴ in an analysis comparing the cost effectiveness of traditional step-up therapy vs. an up-front proton pump inhibitor test in gastroesophageal reflux disease, stated that "we biased the model against the proton pump inhibitor test

Table 6. Results of Final Logistic Regression Analysis

Variable	Odds ratio	95% Confidence limits
Author(s) have advanced training in health services	5.0	2.6-9.3
Study used decision-analysis software package	2.4	1.2-4.7
Study received federal funding	2.2	1.2-4.1
Study cited Gold Criteria in reference list	2.1	1.1-4.2
Study published in U.S. journal	1.6	0.8-3.3
Year of publication	1.2	1.0-1.4
Study received industry funding	1.2	0.3-4.6
Impact factor of source journal	1.1	1.0-1.2
Study published in subspecialty journal	0.6	0.3-1.5

strategy by assuming that the ambulatory 24-hour pH monitoring was only 80% sensitive, despite recent reports documenting poorer sensitivity." Similarly, Dubinski et al.,¹⁵ in an analysis comparing the use of initial serodiagnostic screening vs. standard invasive testing in the diagnosis of inflammatory bowel disease, stated that "to bias the model against the serodiagnostic strategies, we assumed that 50% of patients without IBD [inflammatory bowel disease] and a negative serodiagnostic screen would return with persistent symptoms owing to atypical IBS [irritable bowel syndrome] or other causes." Despite the recommendation to follow examples like these, our study revealed that most published analyses in digestive diseases failed to explicitly address inherent biases and did not perform a fortiori arguments. A possible effect of this shortcoming is that the results of many analyses are overstated. Editors and readers should emphasize the importance of carefully defending model inputs and selecting conservative estimates that guard against potentially invalid findings.

These limitations are compounded by the prevalent failure to perform an adequate sensitivity analysis—a procedure designed, in essence, to measure how far a model can be pushed before it breaks apart.¹² In the 1-way sensitivity analysis, the model parameters are manipulated one at a time to determine whether the results change when the value of each parameter is changed. In the 2-way sensitivity analysis, 2 parameters are manipulated conjointly. In the Monte Carlo analysis, all the model parameters are manipulated simultaneously to emulate the natural variation that occurs in clinical practice. All 3 types of sensitivity analysis should be performed to optimally explore the influence of uncertainty.¹¹ These analyses jointly form the most important step in health economic modeling because they acknowledge that the main results of an analysis may not be generalizable to all populations. More importantly, they address the fact that small errors in the individual probability estimates may lead to magnified errors in the final results of the analysis.¹² If a model leads to similar results despite pushing and stretching the parameters in the sensitivity analysis, then the findings are deemed robust and the model is deemed potentially important. Despite the theoretical and practical rationale for conducting a sensitivity analysis, only one half of published models in digestive diseases adequately perform this exercise. Therefore, the results of many analyses may not only be overstated (as described earlier), but also may not be generalizable to all relevant populations. This finding suggests that editors and readers should emphasize the

importance of conducting a thorough sensitivity analysis, including 1-way, 2-way, and Monte Carlo analyses.

Our analysis further revealed that only 48% of published health economic analyses specified whether there was a source of funding or a potential conflict of interest. In light of our result that the type of funding may be an independent predictor of quality, this shortcoming should prompt journal editors to routinely elicit and publish information regarding funding sources. In addition, readers and consumers of these analyses should remain wary of the potential effect (positive or negative) funding may have on study quality.

Our analysis reveals that only 55% of health economic analyses use valid, reliable, or relevant outcomes measures. Categories of suboptimal outcomes include non-validated health-related quality-of-life scores (utilities) derived solely from expert opinion rather than patients or societal samples, unreliable outcomes such as cost per healed ulcer rather than cost per dyspepsia improvement, or clinically irrelevant outcomes such as cost per correct gastroesophageal reflux disease diagnosis rather than cost per heartburn improvement. Because the selected outcome plays a major role in determining model results, it is critical that outcomes are valid, reliable, and clinically relevant. Clearly, it would not be acceptable if half of the randomized trials in digestive diseases used inappropriate outcomes. It is important to ensure that the standards for health economic analyses are the same as for any other type of study design.

Our data indicate that the use of decision-analysis software packages is associated with improved quality compared with studies conducted without decision-analysis software packages. This suggests that the powerful capabilities of these programs have extended the ability of investigators to conduct increasingly complex analyses while improving quality. Moreover, this suggests that the capabilities of this software have not been abused, and instead investigators have exploited the software's power to generate studies of high quality.

Our analysis has several strengths. First, we performed an explicit and reproducible (Table 1) systematic review to identify all published analyses in a wide range of journals. Second, we did not limit our analysis to subspecialty journals, but instead broadened our scope to include analyses published in general internal medicine journals. Third, we relied on a validated and reliable measure of study quality developed by a panel of expert health economists. Fourth, we ensured that each of the study reviewers had training in the proper application of the QHES from the developers of the instrument. Fifth, we attempted to ensure high interrater reliability

through iterative training sets and quantified our agreement using several methods. Sixth, we attempted to minimize reviewer bias by blinding data abstractors to authors, institutions, and source journals. Seventh, we relied on qualitative best subset selection to develop our final regression model instead of an algorithmic stepwise selection process. Last, we attempted to identify independent predictors of quality while adjusting for a wide array of potential confounding variables (Table 3).

Our analysis had several limitations. First, to focus our study on the most widely read journals, we limited our review to journals with an impact factor greater than 1.0. Although this decision may have excluded some health economic analyses, it is unclear whether the small subset of analyses published in journals with an impact factor of less than 1.0 was systematically different than the study set. Second, we divided the data abstraction among 3 separate reviewers instead of requiring each reviewer to rate the entire set of accepted studies. It is possible that systematic differences between the reviewers led to inaccurate results. However, we made extensive efforts to develop and measure consistency among the 3 reviewers, and each reviewer abstracted a random set of studies. Moreover, the reviewers achieved excellent agreement in the preliminary training phase as measured by 3 pre-specified statistics. There is no a priori reason to expect that the high agreement on the 10% test set would dramatically reverse itself on the remaining studies if the reviewers were to duplicate their abstractions on the full study set. Of further note, although duplicate abstractions are recommended in a meta-analysis of randomized controlled studies (in which small abstraction errors can be magnified), it is not mandated for descriptive analysis—of which the current study is a form. Third, although the outcome measure of quality (the QHES) is a valid and reliable measure of internal validity, it does not measure the external validity, or generalizability, of health economic analyses. For example, a study may be internally valid and score highly on the QHES, yet be poorly generalizable as a result of inadequate comparators. In this circumstance the study would be of limited clinical use despite being methodologically sound. It is therefore critical for users of health economic analyses to consider not only the internal validity of these studies, but also their clinical applicability. Fourth, it may be argued that the QHES is not an adequate surrogate for a detailed clinical and methodologic review performed by experts. The notion of assigning a numeric value to an ultimately subjective assessment is arguably fallacious. Nonetheless, the face validity, content validity, construct validity, criterion validity, and reliability of the QHES

have been established.¹⁰ Future work should aim to confirm the validity and reliability of this scale. However, in the absence of data to the contrary, the QHES remains the only validated objective assessment of study quality for health economic analyses.

In conclusion, this analysis reveals that despite the increasing influence of health economic analyses on the clinical practice of digestive diseases, the general quality of these studies is suboptimal. Study quality appears to be predicted by several factors that potentially may be remedied. These data may be used to help journal editors, reviewers, clinicians, and policy makers ensure the future high quality of health economic analyses in digestive diseases.

References

- Chassin MR, Koseoff J, Park RE, Winslow CM, Kahn KL, Merrick NJ, Keeseey J, Fink A, Solomon DH, Brook RH. Does inappropriate use explain geographic variations in the use of health care services? A study of three procedures. *JAMA* 1987;258:2533–2537.
- Provenzale D, Lipscomb J. A reader's guide to economic analysis in the GI literature. *Am J Gastroenterol* 1996;91:2461–2470.
- Provenzale D, Lipscomb J. Cost-effectiveness: definitions and use in the gastroenterology literature. *Am J Gastroenterol* 1996; 8:1488–1493.
- American Gastroenterological Association. AGA technical review: evaluation of dyspepsia. *Gastroenterology* 1998;114:582–595.
- Grace ND. Practice guidelines: diagnosis and treatment of gastrointestinal bleeding secondary to portal hypertension. *Am J Gastroenterol* 1997;92:1081–1091.
- Winawer S, Fletcher R, Rex D, Bond J, Burt R, Ferrucci J, Ganiats T, Levin T, Woolf S, Johnson D, Kirk L, Litin S, Simmang C. Colorectal cancer screening and surveillance: clinical guidelines and rationale—update based on new evidence. *Gastroenterology* 2003;124:544.
- Goetghebeur MM, Rindress D. Towards a European consensus on conducting and reporting health economic evaluations—a report from the ISPOR inaugural European conference. *Value Health* 1999;2:281–287.
- Marshall JK, Cawdron R, Yamamura DLR, Ganguli S, Lad R, O'Brien B. Use and misuse of cost-effectiveness terminology in the gastroenterology literature: a systematic review. *Am J Gastroenterol* 2002;97:172–179.
- Ofman JJ, Sullivan SD, Neumann PJ, Chiou C, Henning J, Wade S, Hay J. Examining the value and quality of health economic analyses: implications of utilizing the QHES. *J Managed Care Pharm* 2003;1:53–61.
- Chiou C, Hay J, Wallace JF, Bloom BS, Neumann PJ, Sullivan SD, Yu HT, Keeler EB, Henning JM, Ofman JJ. Development and validation of a grading system for the quality of cost-effectiveness studies. *Med Care* 2003;41:32–44.
- Gold MR, Siegel JE, Russell LB, Weinstein MG. *Cost-effectiveness in health and medicine*. New York: Oxford University Press, 1996.
- Petitti DB. Estimating probabilities. In: *Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for synthesis in medicine*. New York: Oxford University Press, 2000.
- Defense Economic Analysis Council, Handbook Committee. *Testing alternatives under uncertainty*. In: Edmonds EW, ed. *Economic analysis handbook*. Chapter 2. Monterey, CA: Defense Resources Management Institute Press, 1997. Available at: <http://www.nps.navy.mil/dirmi/chapter2.htm>.
- Ofman JJ, Dorn GH, Fennerty MB, Fass R. The clinical and economic impact of competing management strategies for gastro-

oesophageal reflux disease. *Aliment Pharmacol Ther* 2002;16:261–273.

15. Dubinsky M, Johanson JF, Seidman EG, Ofman JJ. Suspected inflammatory bowel disease—the clinical and economic impact of competing diagnostic strategies. *Am J Gastroenterol* 2002;97:2333–2342.

Received February 29, 2004. Accepted April 15, 2004.

Address requests for reprints to: Brennan M. R. Spiegel, M.D., M.S.H.S., VA Greater Los Angeles Healthcare System, David Geffen

School of Medicine at UCLA, CURE Digestive Diseases Research Center, Center for the Study of Digestive Healthcare Quality and Outcomes, 11301 Wilshire Boulevard, Building 115, Room 215C, Los Angeles, California 90073. e-mail: bspiegel@ucla.edu; fax: (310) 268-4510.

Supported by a National Institutes of Health training grant (DK-07180 to B.M.R.S.), an American Association for the Study of Liver Diseases (AASLD) Advanced Hepatology Fellowship Award (to F.K.), National Institutes of Health K23 Career Development Award (RR-16188 to G.S.D.), and a VA Health Services Research and Development (HSR&D) (IIR 01-191-1 to I.M.G.).

Eck of the Eck Fistula



Nikolai Vladimirovitch Eck (1847–1908) was a Russian military surgeon curious of certain problems posed by physiology. At age 29, he published a one-and-a-half page account of an experiment he conducted in which he diverted the flow of portal vein blood from the liver to the vena cava by constructing fistulas in a series of 8 dogs. His aim was to disprove the prevailing belief that infusion of the liver via the portal vein was essential to life. Despite the fact that only one of his dogs survived, he pronounced his experiment a success. He suggested, incidentally, that a portocaval fistula might also help alleviate ascites. His investigation was halted, never to be resumed, by a call to active duty in the Russian army. Later, Ivan Pavlov showed that a portocaval fistula was not as benign as Eck had claimed. Nevertheless, the Eck fistula was the forerunner of more refined shunts designed to lessen the tendency to hemorrhage from esophageal varices.

Copyright holder unknown. Image obtained from the Clendening History of Medicine Library (<http://clendening.kumc.edu>).

—Contributed by WILLIAM S. HAUBRICH, M.D.
The Scripps Clinic, La Jolla, California