Volume 3 • Number 4 • 2000
VALUE IN HEALTH

Data Collection Methods in Prospective Economic Evaluations: How Accurate Are the Results?

Christopher J. Evans, PhD, Bruce Crawford, MA

MAPI Values, Boston, MA

ABSTRACT

Objectives: Often in economic evaluations a division is made between those studies that have a high level of accuracy versus those that are easily generalized. This interstudy dichotomy is often translated into prospective, randomized controlled trials with high internal validity and observational and modeling studies with a high level of external validity. This article challenges this conventional view and examines intrastudy effects on validity.

Method: A review and summary of the literature was conducted in order to assess the impact that data collection strategies will have on internal validity. Two scenario models were created in order to gain a preliminary understanding of the magnitude of the problem.

Results: Data collection strategies have an impact on the level of internal validity found in an economic evaluation. Comparisons of studies that are prospective in nature is misleading as data collection strategy can lead to different resource and cost estimates even when all other relevant factors are similar. It is possible to shift and improve the level of validity by combining different collection methods.

Conclusions: Instead of viewing internal and external validity as polar opposites, validity should be considered in terms of a continuum within a particular study. The use of proxies to collect resource utilization estimates, the reliance on patient self-reported data, and the method of collecting this type of data all impact the validity of study results. National guidelines for the economic evaluation of agents and devices should consider this issue in more depth, and existing evidence rankings should be adapted to be more appropriate to pharmacoeconomic studies.

***Keywords*:** cost benefit, data collection, economic evaluation, methodology, proxy, resource utilization, self-report.

Introduction

Research in the discipline of pharmacoeconomics has expanded considerably over the past decade [1]. This expansion in the quantity of pharmacoeconomic evaluations has been accompanied, in part, by an increased awareness of some of the methodological issues surrounding this discipline. The greatest interest has been in developing guidelines for the conduct and reporting of pharmacoeconomic evaluations.

These guidelines, for the most part, have focused on what should be included in an economic evaluation rather than on how to conduct a pharmacoeconomic study. For instance, the Dutch guidelines [2], which were developed recently, provide, among other things, guidance on the appropriate type of study, the timing of the evaluation, the relevance of the comparator, the use of sensitivity and modeling analyses, and the reporting of results. Although guidelines of this type are important for providing a common framework for researchers, they may provide a false sense of security insofar as design issues are not adequately addressed.

The internal validity of a particular study, defined as the true reflection of treatment on the outcome of interest, is dependent critically on the design chosen. Researchers and individuals that evaluate pharmacoeconomic studies recognize the trade-off between choosing a prospective evaluation over a retrospective one: The estimates derived from a prospective evaluation in most cases will have a higher degree of precision compared to a retrospective study [3]. Unfortunately, there has been relatively little research conducted into how within a prospective or a retrospective study internal validity can be improved or compromised due to the selection of the method or mode of data col-

Address correspondence to: Christopher J. Evans, PhD, MAPI Values, 15 Court Square, Suite 620, Boston, MA 02108. E-mail: chris.evans@mapivaluesusa.com

 1098-3015/00/$15.00/0 277–286

lection. Such issues in validity and bias may be considerable across studies.

This article examines this issue in depth, and in particular for the case of prospective, clinical, trial-based evaluations. Although there are several methods for collecting data in a pharmacoeconomic trial, we concentrate on three main areas: the use of patient self-reported data, the use of surrogate respondents to collect resource utilization data, and the mode of administration for questionnaires. We briefly describe, in turn, why and how these methods are used and the relative merits and weaknesses of the techniques. We then describe how these techniques lead to different levels of internal validity within a trial and how switching between techniques may alter the level of validity. We conclude with a discussion on the types of studies that are required and how guidelines may need to be altered in order to ensure that the most accurate results are achieved. We focus on the internal and external validity of these data collection processes and on the measurement validity, defined as the ability of a measure to assess what it was designed to assess.

Patient Recall and Self-Reported Data

Researchers who conduct pharmacoeconomic evaluations often use face-to-face, telephone, or self-administered questionnaires to obtain information on health care resource utilization and indirect costs. This occurs when access to hospital records or claims data is limited or the patient is the only practical source of information (e.g., work loss and disability days). This self-reported information is then used in evaluation exercises of health service use. This information, in many instances, may provide the bulk of resource use estimates in any particular evaluation. As such it is crucial to evaluate the validity of these estimates. For instance, there will necessarily be some error present when using self-reported data, and when overreporting and underreporting do not cancel each other out, estimates of actual resource use will be incorrect.

A comprehensive review was undertaken to determine the potential impact that patient self-reported data has on the internal validity of estimates [4,5]. This review noted that there were several areas that researchers should consider before employing self-reported data in an economic evaluation:

- the length of the recall period or the recency of the event
- the salience of the episode, both in terms of its psychological impact on the patient and the length of the episode
- the level of social desirability attached to an episode
- the tendency to recall recurring events as a group
- the demographic characteristics of patients

Although each study is unique in terms of demographic characteristics, the condition evaluated (chronic vs. acute), and the therapies examined, several common problems have emerged. Table 1 provides a list of these issues as well as the anticipated impact on resource utilization estimates. For example, a patient responder may recall an incidence of resource use; however, the actual resource consumption occurred outside of the study period. Backward telescoping occurs when the resource estimate actually occurred within the reference period, but the respondent recalls it as occurring before the reference period. Forward telescoping occurs when resource use that occurred before the reference period is incorrectly remembered as occurring within the reference period.

The perceived desirability of the event also impacts the accuracy of results. Events or behaviors that are seen as desirable may be overreported (such as income). Conversely, if the behavior or the event is viewed as undesirable there may be a tendency to under-report the associated resource use due to embarrassment or a lack of willingness to be candid with an interviewer.

Another problem that arises is that patients, particularly in chronic conditions, may form a ge-

Table 1 Issues in the use of patient self-reported data

Problem	Issue	Typical impact on results
Length of recall period	Backward telescoping	False negative
	Forward telescoping	False positive
Socially desirable event	Acquiescence bias/self-presentation bias	False positive
Socially undesirable event	Interviewer effect	False negative
Frequency of consultation	Generic memory	False negative
Salience of event	Difficulties in temporal sequencing of low-salience events	False negative

neric memory of their health care resource use. That is, patients may recall events as a group and have difficulty in recalling particular instances. This problem will be particularly acute for repeat visits of a chronic nature that have low salience or impact on patients' lives (such as general practitioner, or GP, consultations).

Demographic characteristics also play an important role in determining the internal validity of estimates based on patient self-reported data. It has been noted that there is a tendency for certain factors to influence the ability or the willingness of a patient to recall a particular health care episode [4]. Although there has been a lack of consistency in the findings of studies, on balance the following characteristics are likely to be associated with underreporting of resource use: older age, single marital status, low income, and low education level [6–9].

In general, any study will have a mix of false positives (overreporting) and false negatives (underreporting); however, the use of self-reported data will in most instances lead to a net underreporting of resource utilization in a particular study.

The implications of this for study design are considerable. For example, studies that examine inpatient hospital resource use for an acute condition in a middle-aged population may be able to use a recall period of approximately 6 months. However, a study that wished to examine physician consultations in a chronic condition in an elderly population would need to consider using a much shorter recall period. This is necessitated by the fact that the GP consultations have lower salience than hospitalizations, chronic conditions face the pitfall of generic memory, and elderly patients are more likely to be cognitively challenged.

Further problems arise if researchers desire to disaggregate results by resource category. In this case, each resource category compared between groups will have problems. For instance, GP visits may be underreported, whereas inpatient stays may conceivably be overreported. Although the underreporting and overreporting will partially cancel each other out, if the researcher chooses to focus on one particular category of resource consumption where there is misreporting, a misleading picture of cost differences may occur.

Proxies

It is generally agreed that in outcomes research patients are the most appropriate source of information on their resource use. Patients have direct experience with health care providers and are more likely to remember, in detail, health care encounters. However, in some instances it is necessary to search for a valid alternative source of patients' resource use when the patient is cognitively impaired due to declining health or is too young to provide direct estimates. In addition, under conditions where a patient may be able to complete a questionnaire, but due to ill health or a lack of willingness (e.g., terminal diseases or excess respondent burden) does not comply, the use of a proxy respondent may lead to more complete case records.

There are various techniques for handling missing data, but as a general proposition it is preferable to have an actual estimate from a valid data source. The problem of missing data is substantial, particularly for elderly patients. A large survey of the general population in the United States [10] revealed that 15% of hospitalizations occurred in 5% of the sample that were nonrespondents to the survey. A study conducted in Baltimore of elderly hip fracture patients [11] revealed that of 858 patients identified, 51% could not be interviewed due to cognitive impairment, illness, refusal, or death. Thus, gaining the cooperation of proxies is likely to increase the sample size for a study and reduce the number of missing observations.

Although the use of proxies may increase the sample size of a study and reduce the number of missing items, their use in pharmacoeconomic studies leads to a number of methodological problems. The primary issue is the accuracy of resource estimates derived from proxies. Proxies may not be as sensitive to patient resource use as the patient and they may fail to detect an episode of care. In addition, they may identify resource use when in fact it did not occur. Thus, conceptually, pharmacoeconomic studies that use proxies to collect estimates of resource use will have some level of error associated with them as exact agreement between patient and proxy reports will occur only in limited circumstances (e.g., when the proxy provides a report of an acute episode of care that occurred recently).

Researchers in the field of quality of life and epidemiology [12,13] have theorized that the relationship between the patient and the proxy, living arrangements, the time spent between the patient and the proxy, and how directly observable the activity is will have an impact on the accuracy of estimates between patients and proxies [14]. A strong relationship between patients and proxies, as often found in spouses, will encourage greater agreement. In addition, at least for quality-of-life assessments, if the proxy and the patient live and spend

a substantial amount of time together the level of agreement will also be high. Activities that are directly observable by the surrogate, as many resource use items are, will likely have a high level of agreement when compared to the self-responder.

The level of agreement for pharmacoeconomic evaluations will not be perfect if studies in the field of quality-of-life research are a guide. A review of this issue in the quality of life of patients with chronic diseases found that providers might overestimate patients' feelings of anxiety and depression and underestimate impairment in functional status [15]. Of course, pharmacoeconomic research is different from quality-of-life research, and it is possible that when agreement on quality-of-life scores between provider proxies and patients is low, it might be quite high for resource use, because this is directly observable (to some extent) for providers.

The level of agreement between proxies and patients has not been well established. Recently a review of the literature [16] was conducted that found that when comparisons were made, proxies suffer from the same issues of patient recall and reporting as shown above. Although the use of proxies has been relatively well researched in the fields of epidemiology and quality of life, there is a dearth of evidence available concerning the impact of surrogate respondents on resource utilization estimates. Even answering the seemingly simple question of the direction of the error is problematic. Although the evidence suggests underreporting to be more likely, the studies reviewed found instances of overreporting as well, particularly in the area of hospital days. In general, it seems likely that the salience of any particular event is less for the proxy than for a self-responder. This will lead to a further underreporting of resource use over that which would have been found had the information been based solely on patient self-reported data.

Collection Modes

There are several factors that must be considered when selecting a data collection mode for a pharmacoeconomic trial. Often, the primary consideration is the time and expense associated with a particular mode. Face-to-face interviews are considered the most expensive and time consuming to conduct, while telephone interviews and mail surveys are relatively less expensive.

Beyond time and cost, the mode selected has important implications for study validity [17]. For instance, noncoverage and nonresponse bias are highest for mail (self-administered) questionnaires and for telephone interviews. However, they tend to be better than in-person interviews because they minimize interviewer effects. Mail surveys also have an advantage in that they allow respondents time to reflect on their answers compared to top-of-the-head responses typical in telephone and in-person interviews. Two disadvantages of mail surveys are that they do not allow for probing by interviewers and they cannot be used in populations with visual impairment or low literacy.

The level of accuracy obtained from different data collection modes has not been well investigated for items of health care resource utilization. Weeks et al. [18] validated information obtained from in-person interviews and telephone interviews with medical records and found only minor differences in the two modes in terms of the accuracy of the data. The accuracy of ambulatory care visits was higher for telephone (56.3% exact or partial agreement) compared to in-person interviews (46%). A slightly higher percentage of agreement on hospital stay and condition was achieved for the telephone (53.2%) compared to in-person interviews (49.3%).

Yaffee et al. [7] also found no consistent difference in reporting accuracy among different survey strategies over a 6-month period: monthly telephone, bimonthly telephone, monthly telephone–in-person, and bimonthly in-person. All of the strategies suffered in terms of accuracy when compared to validated data. However, the in-person data collection mode did slightly better in one location of the survey for reporting of utilization and charges.

An important shortcoming of the studies reported above is that they only report the level of accuracy, not the level of under- or overreporting. An early study conducted by the US Department of Health, Education and Welfare [19] examined this issue, comparing two modes to the medical record criterion, the standard health interview survey, and an altered health interview survey with a mail follow-up. This study found that percentage of underreporting fell with follow-up. For self-responders, episodes of hospitalization were underreported by 10% and proxies for adults underreported by 21%. For respondents who received a mail follow-up the percentage of underreporting fell to 6% for both proxies and self-reporters.

It has been suggested that face-to-face and telephone modes are likely to be comparable in terms of validity [20]. Postal surveys are likely to yield a lower level of validity. Mixed modes of data collection in some instances may provide the highest

level of validity. For example, the use of regularly scheduled telephone interviews to supplement the collection of patient diary information may reduce recall bias when a patient completes the entire diary a week before his/her 6-month protocol visit. The telephone interview will collect data for the noncompliers at a more appropriate interval. Alternatively, a postal copy could be sent to a patient, followed by a telephone interview for data collection. This would allow the patient time to reflect on the questions prior to the telephone interview, which is beneficial because the telephone interview typically requires patients to provide top-of-the-head responses without adequate reflection and processing time.

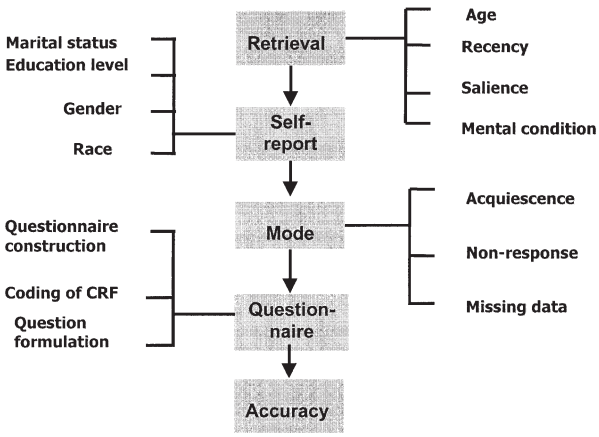

Figure 1 Relationship between retrieval and internal validity. CRF, case report form.

Compromised Validity and Bias

The information provided in the previous section represents the impact that the use of patient recall, proxies, and data collection strategies may have on time, expense, accuracy, and generalizability. These discussions demonstrate that even within a prospective study, different levels of internal validity are achieved depending on the data collection strategy chosen. The use of patient self-reported data lowers the level of internal validity compared to a source-documented case report form (CRF). Consistent with the findings reported in the literature, the use of proxy reports further lowers the validity of estimates. In terms of the mode selected, a mixed-mode approach may yield a higher level of validity compared to the self-reported data and proxies, but it does so by increasing the financial cost of the study.

As a theoretical proposition, it is interesting to note that the internal validity of a prospective study may fall so far, due to the combined effects of recall problems and mode selection, that a retrospective study may in fact achieve a higher level of internal validity. It should be noted that the apparent advantage in this case is not due to any inherent benefit of using a retrospective study, but to the lack of accurate data in a prospective trial. The extent to which this may happen may be dependent on the interaction of a number of factors. Figure 1 represents the relationship between retrieval problems and internal validity.

Given that internal validity or the accuracy of the estimates is impacted by the data collection methods chosen in a particular study, it is worthwhile to investigate the implications of this on decision making. Under most circumstances, within a standard piggyback trial, the problem will be with data accuracy rather than bias. Here bias is defined as misidentifying the pharmacoeconomic profile of a particular agent so that one agent is erroneously favored over another one, when in some cases the reverse situation is true. Assuming that the trial is adequately powered, has relatively strict inclusion and exclusion criteria, and is randomized, demographic characteristics of patients should be similar at baseline between groups. This will, in most cases, control for demographic factors that may impact recall, proxy response, and the ability to complete a particular data collection mode. If differences in baseline factors occur, one will need to examine their impact on resource utilization reporting. Thus, the net effect will be to inaccurately estimate resource use by a similar (although not exact) proportion. In this case, if one agent is shown to dominate another agent, researchers may feel confident that the correct agent was identified even though the cost estimate is probably understated. However, under conditions that are arguably unlikely to arise, it is possible that bias may appear when two interventions are compared within a trial.

Scenario I. Consider the scenario outlined in Table 2, where a surgical intervention (Intervention A) is compared to a strategy of watchful waiting (Intervention B) in the context of a Phase III piggyback study. The following assumptions have been made:

- The cost of the surgical intervention is substantially higher than the cost of the strategy of watchful waiting.
- The cost of follow-up care for the strategy of watchful waiting is substantially higher than the cost associated with the surgical care.

Table 2 Bias within a piggyback trial

	Cost of Intervention A ($)	Cost of follow-up care (actual) ($)	Cost of follow-up care (reported) ($)	Total cost (actual) ($)	Total cost (reported) ($)
Intervention A	6,000	500	250	6,500	6,250
Intervention B	600	10,000	5,000	10,600	5,600

Source: The information provided in this table was provided to the authors as an illustration at the 1999 summer meeting of the HESG.

- The trial is sufficiently large and participants are randomized to treatment (although obviously not blinded). Due to this, the percentage of underreporting of resource use is nearly identical (in this case equal) between trial arms, assumed at 50% in this example.

The results, reported as total costs, show that the watchful waiting strategy is preferred ($5,600 vs. $6,250). However, if the true data were available, then Intervention A would be revealed as having the better cost profile ($6,500 vs. $10,600). Thus, in this situation, bias has occurred within a clinical trial.

DATA 3.5 (TreeAge Software, Inc., Williamstown, MA) was used to generate a univariate sensitivity analysis and threshold value to test the above scenario. When the cost of Intervention A was reduced to between $700 and $1200 (potentially a more realistic scenario) the expected value for Intervention A fell to $1450 (at the upper end of the range). In this case, Therapy A would be identified correctly as having the lower costs and the problem of bias would not occur. The threshold analysis, on a broader range of $0 to $7000, indicated bias becomes an issue at $5350. When the cost of Intervention A is greater than $5350, it would be shown as the more expensive strategy, when in fact it was not. In other words, assuming the follow-up costs reported, Intervention A has to be $4650 more expensive than B in order for there to be a problem with bias.

Although the percentage of underreporting is equal between arms, this assumption should be questioned in practice when there is such a large difference in the cost of follow-up care. The higher cost of follow-up care for the watchful waiting strategy suggests more inpatient consultations, which are more likely to be recalled. In this situation, the reported values for the watchful waiting strategy would actually be much higher (and closer to the true value) than is represented here. Researchers should be sure to examine their data for potential fallacies, such as this one. A threshold analysis here indicates that if the follow-up care reported for Intervention B were greater than $5650 (which would reflect more accurate reporting), then Intervention A would have an expected value lower than Intervention B. This would be the correct identification of the relative cost profiles of the interventions.

Given the extreme nature of the above scenario, users, researchers, and reviewers should have some confidence that pharmacoeconomic studies can be a valuable tool in the identification of preferred therapies, although the actual impact on costs will remain somewhat illusionary. In other words, the more cost-effective treatment will likely be selected, but the magnitude of cost differences will be imprecise. Although the problem of bias is unlikely to arise within a study, it may occur when two studies are compared. To illustrate the issue, Scenario II has been developed.

Scenario II. Pharmaceutical Company A develops Drug A for the treatment of Condition X. As part of its clinical development activities, it conducts piggyback studies as part of two Phase III pivotal trials that compare Drug A to usual care in terms of resource use.

Pharmaceutical Company B develops Drug B for the treatment of Condition X. As part of its clinical development activities it conducts piggyback studies on two Phase III pivotal trials that compare Drug B to usual care in terms of resource use.

The trials conducted by both Companies A and B are identical in every aspect (e.g., inclusion/exclusion criteria, resources collected, end points, and time frame). The only difference in the protocol is in the data collection strategy used to collect information on resource use.

Company A decides to collect resource utilization based on CRFs with source documentation based on all medical records for the inpatient baseline hospitalization and rehospitalization to Month 12. Follow-up visits are recorded for a further 12 months by an in-person interviewer. To ensure that accurate results are achieved, a mail survey is sent to respondents to check and alter any incorrect answers.

Company B decides to collect resource utilization based on patient self-reports at Month 12. An

in-person interview is conducted 12 months later to collect information on additional items of resource use.

Assume that the data in Table 3 are derived. The example provided in Table 3 is meant to be illustrative, rather than a definitive statement as to the magnitude of the problem. As the table demonstrates, the presence of recall and mode effects are more prevalent in the second drug study conducted by Pharmaceutical Company B. In the absence of any recognition of the impact of data collection strategies on internal validity, erroneous conclusions might be reached. Under the stylized situation above, the trials are identical in all seemingly important aspects. Thus, it could be argued that it is appropriate to bridge the results of the study, so that Drug A can be compared to Drug B. If this were the case, then the conclusion would be reached that Drug B has a more favorable profile compared to Drug A. However, if figures were available on the true cost of the treatment (Table 3, column 2), Drug A clearly has the preferred profile. The only marker, in terms of study results, that there is a bias issue is the underestimate of costs in the usual care arm for the second study compared to the first. Again, researchers should look to identify these signs in an effort to identify problematic results.

Conclusion

As the number of economic evaluations has increased it is important to expand our understanding of data collection issues to the impact of study validity. The quantity of economic evaluations has not been matched by an equivalent rise in the quality of evaluations. The first step in ensuring the quality of pharmacoeconomic research is through good study design. The above discussion has highlighted areas where there may be shortcomings in the design of economic evaluations. However, our knowledge about the direction and magnitude of recall, proxy, and mode effects is limited. Researchers should evaluate a priori the anticipated effects of different modes and methods of data collection. This will allow the appropriate development of a data collection strategy. For example, if a study is to be conducted in a terminal disorder, the researcher can anticipate nonresponse as the patient approaches death. To avoid loss of data, they may wish to employ proxy reporting. Since it is known that proxy responses are not that accurate, the researcher may wish to collect information from both the proxy and the patient throughout the study. This way, when the patient will not or cannot complete a questionnaire, the proxy responses could be adjusted for the lack of agreement.

This work has suggested that there may be a considerable gap between obtaining data and deriving a correct answer to any question in pharmacoeconomic research. This article has highlighted the fact that internal validity may suffer substantially in the face of different data collection methods. As internal validity is a necessary precursor to external validity, it is likely that studies believed to have a high level of external validity in fact do not, even when such evaluations are conducted under naturalistic conditions. Given the lack of attention placed on the importance of data collection methods it appears that an assumption has been made by some researchers that data collected—regardless of source or collection methods—yield similar levels of quality. Further research is required, in particular for proxy and mode effects, to determine the impact on the internal validity of results.

Validation studies are necessary in order to endorse the various techniques in the field of pharmacoeconomics. Validation studies have been a fixture of quality-of-life research and epidemiology for decades. However, there have been relatively few of these types of studies conducted in pharmacoeconomics. Prospective validation studies, which compare the different techniques in terms of agreement and error, are necessary so that researchers can evaluate the impact that the data collection methodology and chosen mode have on the results of economic evaluations. This is not an easy task, as some pharmacoeconomic

Table 3 Bias between studies: Scenario for mode and recall effect

	True cost ($)	Recall effect	Proxy effect	Mode (and recall) effect	Estimated cost ($)
Drug A	15,000	0	NA	0.06	14,100
Usual care A	15,500	0	NA	0.06	14,570
Usual care B	16,500	0.10	NA	0.10	13,200
Drug B	16,000	0.10	NA	0.10	12,800

Recall and mode effect based on US Department of Health, Education and Welfare [19].

studies will require several validation studies. For instance, validation studies that are conducted should be designed so as to decompose the mode effect from the recall effect. Where under- and overreporting have been reported in patient recall studies it is not always possible to discriminate between mode and recall effect (Fig. 1).

The creation of preliminary guidelines may also prove useful. Guideline development for economic evaluations has become stagnant. The recent Dutch guidelines offer no improvement over guidelines developed for Ontario in 1994. In the intervening 5 years considerable research has been conducted into the methodology of economic evaluations, yet additional information is excluded from these guidelines. Although it is apparent that adherence to guidelines is often irregular, they are often most useful to individuals least familiar with economic evaluations. In terms of incorporating information on data collection strategies, it would be premature to include information of proxy and mode effects; however, the issue of patient recall is more straightforward and it would be possible, at the least, to give guidance as to the issues involved in using alternative strategies.

Currently, there is no objective way to assess the overall validity of estimates derived from pharmacoeconomic trials. Cook et al. [21] proposed an evidence-ranking system that considered the design of a trial in determining the strength of recommendation that should be made from a particular trial (Table 4). This suggests that prospective, randomized trials provide the highest-quality evidence, whereas information derived from case studies are anecdotal in nature and may not be nearly as useful in informing decision making. As noted above problems arise in pharmacoeconomic studies that suggest that existing evidence rankings are likely to be of little use in pharmacoeconomic research. For instance, Level I in the typology of Cook et al. [21], includes large randomized trials with clear-cut results. Studies that fall into this category receive a high level of recommendation. This would include many Phase III trials of drugs used for registration purposes. However, even though the clinical aspects of the trial may receive an A recommendation, there is little reason to assume that the pharmacoeconomic results expressed as part of a piggyback deserve such a high recommendation. The reason for this is that a piggyback trial resorts to retrospective data collection within the prospective trial and may include proxy and mode effects.

Evidence-ranking systems developed recently are more comprehensive, yet they are still insufficient for pharmacoeconomic purposes (Table 5). Additional work needs to be directed at developing evidence rankings that are applicable to pharmacoeconomic studies. Even though a considerable amount of pharmacoeconomic research is conducted alongside clinical trials, the nature of eco-

Table 4 Levels of evidence for therapy

Level	Type of trial	Grade
I	Randomized trial with low errors	A
II	Randomized trials with high errors	B
III	Nonrandomized concurrent cohort studies	C
IV	Nonrandomized historical cohort studies	C
V	Case series	C

Table 5 Hierarchy of evidence

Experimental	
I	Well-designed randomized controlled trials
	Other types of trials
II-1a	Well-designed controlled trial with pseudo-randomization
II-1b	Well-designed controlled trials with no randomization
Observational studies	
II-2a	Well-designed cohort (prospective study) with concurrent controls
II-2b	Well-designed cohort (prospective study) with historical controls
II-2c	Well-designed cohort (retrospective study) with concurrent controls
II-3	Well-designed epidemiological case control (retrospective) study
III	Large differences from comparisons between times and/or places with and without intervention (in some circumstances these may be equivalent to level II or I)
Expert opinion	
IV	Opinions of respected authorities based on clinical experience; descriptive studies and reports of expert committees

Source: NHS Center for Review and Dissemination (1996) cited in Rittenhouse B. Use of Models in Economic Evaluations of Medicines and Other Health Technologies. London: Office of Health Economics, 1996.

nomic evaluations is such that the strength of any recommendation is based on many more factors. Evidence rankings that are based on principles rooted in epidemiology will prove more useful than rankings developed for clinical research. The NHS Center for Review and Dissemination has proposed a hierarchy of evidence that may provide a framework for pharmacoeconomic research. Table 5 shows that additional categories are appropriate for these ranking systems.

Table 3 suggests that there are also implications for reporting pharmacoeconomic information in league tables. League tables have been developed so that researchers may place their findings in a broader context and make informed decisions about the allocation of health care resources. The usefulness of league tables was reviewed several years ago [22] and it was revealed that decision-makers should exercise caution when interpreting them because of differences in discount rates, the method for estimating utility values, the range of costs and consequences considered, and the choice of comparators between studies. Because data collection procedures are likely to differ between studies, league tables will also contain misordering of conditions and therapies due to the method selected. Inasmuch as inappropriate order occurs on league tables, therapies may be adopted upon erroneous data. Budgetary impact analyses of these therapies will also be misleading in magnitude. Thus, in the long run, this will cause an unexpected burden on the health system. Although this effect is likely to be small, for poorly designed trials and at certain thresholds the inappropriate selection of data collection methods may have important implications for drug adoption and reimbursement.

The authors would like to express appreciation for comments received from individuals who participated in International Society for Pharmacoeconomics and Outcomes Research workshops covering portions of this material and to the discussants and participants at the Health Economists' Study Group meeting in Aberdeen, Scotland, July 14–16, 1999. In particular we would like to acknowledge the discussant at the HESG conference for providing us with the baseline example used in Scenario I.

References

1 Elixhauser A, Halpern M, Schmeir J, Luce B. Health care CBA and CEA from 1991–1996: an updated bibliography. Med Care 1998;36(5):MS1–9.
2 National Health Insurance Council. Draft Report on the Pharmacoeconomic Research Guideline. Amstelveen, The Netherlands: National Health Insurance Council, 1999.
3 Rittenhouse B, O'Brien B. Threats to the validity of pharmacoeconomic analyses based on clinical trial data. In: Spilker B, ed., Quality of Life and Pharmacoeconomics in Clinical Trials (2nd ed.). Philadelphia: Lippincott-Raven Publishers, 1996.
4 Evans C, Crawford B. Patient self-reports in pharmacoeconomic studies: their use and impact on study validity. Pharmacoeconomics 1999;15(3):241–56.
5 Crawford B, Evans C. Self-reported resource utilization data in pharmacoeconomic trials: their use and impact on study validity. Study Methods Workshop presented at the ISPOR Inaugural European Conference. Cologne, Germany, December 10–12, 1998.
6 Weissman J, Levin K, Chasan-Taber S, et al. The validity of self-reported health care utilization by AIDS patients. AIDS 1996;(10):775–83.
7 Yaffe R, Shapiro S, Fuchseberg RR, et al. Medical economics survey-methods study: cost effectiveness of alternative survey strategies. Med Care 1978; 16(8):641–59.
8 Green S, Kaufert J, Corkhill R, et al. The collection of service utilization data: a research note on validity. Soc Sci Med 1979;13A:231–4.
9 Burman M. Health diaries in nursing research and practice. J Nurs Scholarsh 1995;27(2):147–52.
10 Corder L, Woodbury M, Manton K. Proxy response patterns among the aged: effects on estimates of health status and medical care utilization from 1982–1984 long term care surveys. J Clin Epidemiol 1996;49(2):173–82.
11 Magaziner J, Simonsick E, Kashner T, Hebel J. Patient-proxy response comparability on measures of patient health status and functional status. J Clin Epidemiol 1988;41(11):1065–74.
12 Magaziner J. The use of proxy respondents in health studies of the aged. In: Wallace R, Woolson R, eds., The Epidemiologic Study of the Elderly. New York: Oxford University Press, 1992.
13 Clipp E, Elder G. Elderly confidants in geriatric assessment. Compr Gerontol [B] 1987;1:25–40.
14 Magaziner J, Hebel J, Warren J. The use of proxy respondents for aged patients in long-term care settings. Compr Gerontol [B] 1987;1:118–21.
15 Sprangers MA, Aaronson NK. The role of health care providers and significant others in evaluating the quality of life of patients with chronic disease: a review. J Clin Epidemiol 1992;45:743–60.
16 Evans C, Crawford B. Proxy reports in pharmacoeconomic studies: applications and accuracy for resource utilization estimates. Forthcoming in Journal of Research in Pharmaceutical Economics 2000.
17 Aday L. Designing and Conducting Health Surveys (2nd ed.). San Francisco: Jossey Bass, 1996.
18 Weeks M, Kulka R, Lessler J, Whitmore R. Personal versus telephone surveys for collecting household health data at the local level. Am J Public Health 1983;73:1389–94.

19 US Department of Health, Education and Welfare. Comparison of hospitalization in three survey procedures: vital and health statistics. Series 2, No. 8. Washington, DC: National Center for Health Statistics, 1965.
20 Crawford B, Evans C. Data collection methods for resource utilization: choosing the right approach. Value Health 1999;2(3):229.
21 Cook DJ, Guyatt GH, Laupacis A, Sackett DL. Rules of evidence and clinical recommendations on the use of antithrombitic agents. Chest 1992; 102(Suppl 4):S305–S311.
22 Drummond M, Torrance G, Mason J. Cost-effectiveness league tables: more harm than good? Soc Sci Med 1993;37(1):33–40.