**International Journal of Technology Assessment in Health Care**, **23:1** (2007), 96–100.


# Exploration of the difference in results of economic submissions to the National Institute of Clinical Excellence by manufacturers and assessment groups

**Deven Chauhan**
*Office of Health Economics*

**Alec H. Miners**
*London School of Hygiene and Tropical Medicine*

**Alastair J. Fischer**
*St. George's, University of London* and *National Institute for Health and Clinical Excellence*

**Objectives:** A recent study showed that estimates of cost-effectiveness submitted to National Institute for Health and Clinical Excellence (NICE) by manufacturers had significantly lower incremental cost-effectiveness ratios (ICERs) than those submitted by university-based Assessment Groups. This study extends that analysis.
**Methods:** Data were abstracted from relevant NICE documentation for thirty-two of eighty-two possible appraisals.
**Results:** The results from the analysis showed that sources of the difference in ICERs appear to be the effectiveness estimates relating to the comparator technology and the cost estimates relating to the technology under evaluation. That is, manufacturers estimated lower average benefits for the comparator technology and lower costs relating to the technology under evaluation compared with estimates submitted by the Assessment Groups.
**Conclusions:** These findings may be particularly important, given the introduction of the "Single Technology Appraisal." Considerable difficulties were encountered when undertaking this study, highlighting, above all else, the complexity of explaining why results from economic evaluations purporting to answer the same question diverge.

**Keywords:** Cost-effectiveness analysis; Bias (epidemiology); Drug industry; Technology assessment, biomedical

The Technology Appraisal Programme (TAP) of the National Institute for Health and Clinical Excellence (NICE) in the United Kingdom issues guidance on the use of specific healthcare technologies based on evidence relating to clinical and cost-effectiveness. The criterion used in this program to assess cost-effectiveness is the incremental cost-effectiveness ratio (ICER), defined as the difference in costs between two technologies divided by the difference in their benefits (see Equation 1). For a technology that increases both costs and benefits of intervention, the lower the ICER, the more cost-effective a technology, and the higher

We thank Ken Brewer for his advice on the statistical analysis, and Adrian Towse for his comments on the manuscript.

Deven Chauhan is based at the Office of Health Economics, which receives some funding from the Association of the British Pharmaceutical Industry (ABPI). Alec Miners is formerly employed at NICE and is a present member of the Appraisal Committee. Alastair Fischer is based at NICE 80 percent of the time on secondment from St. George's. His position at St. George's is funded by NICE.

priority it should receive in terms of funding, all else being equal.

$$\text{ICER} = \frac{\Delta C}{\Delta E} = \frac{C_B - C_A}{E_B - E_A} \quad \textbf{(1)}$$

where $\Delta C$ = Incremental Costs
$\Delta E$ = Incremental Benefits
$C_A$ = cost of comparator technology
$C_B$ = cost of technology under evaluation
$E_A$ = health effects of comparator technology
$E_B$ = health effects of technology under evaluation

For each technology passing through the TAP, evidence is submitted by the relevant manufacturer, and professional and national patient groups. At the same time, NICE (through the NHS Research and Development Health Programme) commissions an academic center called an Assessment Group (AG) to provide an assessment of the evidence. Guidance to the NHS is formulated with regard to the totality of this evidence base, including information on cost-effectiveness.

Both the AG and, usually, each manufacturer submit an economic evaluation for each appraisal. In 2005, Miners et al. showed that ICERs submitted by manufacturers were significantly more optimistic (or less pessimistic) compared with claims made by the AGs (8). However, no attempt was made to identify which parameters were responsible for this discrepancy. This information would be useful for those critiquing studies and ultimately for decision makers who need to be able to justify preferences for using the results from one economic evaluation over another. At a more strategic level, this information could also help to inform guidelines for assessing and appraising economic evidence.

## METHODS

We denote the costs and effects for the manufacturer with subscript M (for example, $C_{BM}$ is the cost of technology under evaluation reported by M) and for the AG by G (for example, $C_{BG}$). The aim of this study was to compare the degree of similarity between:

- $(C_{BM} - C_{AM})$ and $(C_{BG} - C_{AG})$, **(i)**
- $(E_{BM} - E_{AM})$ and $(E_{BG} - E_{AG})$, **(ii)**
- $C_{BM}$ and $C_{BG}$, **(iii)**
- $C_{AM}$ and $C_{AG}$, **(iv)**
- $E_{BM}$ and $E_{BG}$, and **(v)**
- $E_{AM}$ and $E_{AG}$ **(vi)**

Data on incremental costs, incremental benefits, and individual components of the relevant ICER ($C_{BM}$, $C_{AM}$, $E_{BM}$, and $E_{AM}$; $C_{BG}$, $C_{AG}$, $E_{BG}$, and $E_{AG}$) were collected for Technology Appraisal numbers 1 to 82 using the following sources: the Assessment Report (the report submitted by the AG), each manufacturer's submission, Overview (a summary written by the technology analyst at NICE assigned to the appraisal), final NICE guidance, and electronic versions of the models. Where stated, data relating to base case assumptions were collected. Individual appraisals were excluded from the analysis if either the manufacturer or the AG did not produce a cost-effectiveness estimate. In addition, the health benefits had to be measured in the same units within each appraisal. For example, if the ICER from one source was expressed using QALYs (quality-adjusted life-years), then the ICER from the other source would also have to be expressed using QALYs.

Several appraisals contained more than one ICER. The calculation of multiple ICERs in some appraisals leads to a clustering effect in the statistical analysis. To counteract this effect, a mean value for each of the four components of the ICER for both the Assessment Group and the manufacturer was calculated so that there would only be one pair-wise comparison per appraisal. Data were also collected on the length of the time horizon to see whether this finding differed on average between the manufacturer and AG submissions.

### Analysis

A simple test of whether the $\Delta E_M = E_{BM} - E_{AM}$ (incremental benefits in manufacturer submission) is different from $\Delta E_G = E_{BG} - E_{AG}$ (incremental benefits in the Assessment Report) may be constructed by assuming that $\Delta E_M = \Delta E_G$ as a null hypothesis and using a binomial distribution with $p = .5$ to determine whether the proportion of times that $\Delta E_M$ exceeded $\Delta E_G$ or vice versa was sufficiently unlikely by chance. The reason that the signs of $\Delta E_M$ and $\Delta E_G$, rather than their actual magnitudes, are used, is that the appraisals are measuring magnitudes with means and variances that differ between appraisals. This test was repeated for comparisons (ii) to (vi) above, again using a sign test.

For each of the appraisals, the average of the time horizons for both the Assessment Group and the manufacturer were calculated so that there was only one pair-wise comparison per appraisal. A Wilcoxon sign rank test was then carried out to see whether there was a difference between the time horizon used by the manufacturer and the Assessment Group.

### Scaling Issues

A technical problem that needs consideration arises because the ICER is a ratio. Even if the individual components, denominator and numerator, are different in the manufacturer submission and Assessment Report, the overall ICER might still be the same. This is demonstrated by equation 2 below:

$$\frac{C_{BM} - C_{AM}}{E_{BM} - E_{AM}} = \frac{S.C_{BM} - S.C_{AM}}{S.E_{BM} - S.E_{AM}} = \frac{C_{BG} - C_{AG}}{E_{BG} - E_{AG}} \quad \textbf{(2)}$$

If all the individual components in one submission are multiples of the individual components in the other

submission and S in equation 2 is not equal to 1, then a scaling effect exists. In that case, the difference between the individual components of the Assessment Group's model and their respective components in the manufacturer's model may not be "real." Scaling effects may occur because the average severity of disease among the patients within the model of one submission may be greater than in the other submission or the time horizons used in the two submissions may differ.

Suppose without loss of generality that $S > 1$. Then $C_{BG} - C_{AG}$ will exceed $C_{BM} - C_{AM}$, and for the same reason, $E_{BG} - E_{AG}$ will exceed $E_{BM} - E_{AM}$. Thus the sign of $(C_{BG} - C_{AG}) - (C_{BM} - C_{AM}) = C$, say, will be the same as that of $(E_{BG} - E_{AG}) - (E_{BM} - E_{AM}) = E$, say. The same will be the case if $S < 1$, so a test of whether there is a scaling effect is whether the sign of E and the sign of C are positively correlated, using a binomial distribution with $p = .5$.

In the same way, there may be a scaling effect in only the A components or only the B components of the ICER. That is, possibly $E_{BG} = S.E_{BM}$ and $C_{BG} = S.C_{BM}$ (and there is no scaling effect with the comparator technology) or $E_{AG} = S.E_{AM}$ and $C_{AG} = S.C_{AM}$ (and there is no scaling effect with the technology under evaluation).

The discussion of the scaling effect assumes that the ICER for the manufacturers would otherwise be the same as the ICER for the AG. Because this will not be so, the significance level ($p$ value) estimated on the basis that the ICERs are the same will probably overstate the true $p$ value and, therefore, be conservative.

## RESULTS

Of eighty-two appraisals, fifty were excluded for the following reasons: the Assessment Group alone did not produce an ICER ($n = 24$), there was no manufacturer/the manufacturer did not produce an ICER ($n = 8$), both the Assessment Group and the manufacturer did not produce an ICER ($n = 1$), different measures of health benefit were used ($n = 8$) or the four components of the ICER were not available for analysis from AG, manufacturer, or both ($n = 9$). Table 1 summarizes the results concerning the denominator, numerator, and individual components of the ICER as defined in the statistical analysis.

**Table 1.** Results of Statistical Analysis

| Statistic | Proportion of manufacturer estimates exceeding AG estimates | $p$ value |
|---|---|---|
| $\Delta C_M$ vs. $\Delta C_G$ | 8/32 | .004 |
| $\Delta E_M$ vs. $\Delta E_G^a$ | 21/29 | .012 |
| $C_{BM}$ vs. $C_{BG}$ | 9/32 | .010 |
| $C_{AM}$ vs. $C_{AG}$ | 14/32 | .298 |
| $E_{BM}$ vs. $E_{BG}^a$ | 13/29 | .355 |
| $E_{AM}$ vs. $E_{AG}^a$ | 7/29 | .004 |

[a] Three appraisals were excluded as they were cost-minimization analyses and, therefore, did not include a measure of health benefit.

To make an ICER more favorable (Equation 1), either the incremental effectiveness must increase (ceteris paribus), or the incremental costs must decrease (ceteris paribus). Manufacturers' reported incremental costs ($\Delta C$) were higher than those reported by the AGs in 25 percent of submissions ($p = .004$). Conversely the incremental benefits ($\Delta E$) were higher among the manufacturer submission in 72 percent of submissions ($p = .012$).

Regarding the four individual components of the ICER, a more favorable ICER would result if $C_A$ or $E_B$ were increased or if $C_B$ or $E_A$ were decreased. Statistically significant differences (at the 5 percent level) were recorded for $C_B$ and $E_A$, and both were in the anticipated direction (Table 1). Differences between manufacturer and AG estimates for $C_A$ and $E_B$ were not statistically significant.

## SCALING EFFECTS

Of the three tests of whether a scaling effect operates, presented in Tables 2–4, two are not significant at the 5 percent level and, for the third (scaling effect in the comparator), the $p$ value is .031. If the significance tests in Tables 3 and 4 are independent, the Bonferroni correction for multiple testing would indicate that the critical $p$ value for these two tests taken together would be .025, suggesting that the $p$ value of Table 3 of .031 is not significant. As the tests are probably not independent, the significance of the .031 value is, therefore, indeterminate on this criterion alone. However, as stated above, the test is also likely to undervalue the true $p$ (because there are effects other than scaling present) so the chance that .031 is significant is likely to be low. Additionally, if it existed at all, the scaling effect would be present only within the comparator and it does not apparently exist overall (Table 2). Overall, the evidence for the existence of a scaling effect is, therefore, relatively weak, but is not completely ruled out.

**Table 2.** Analysis of Scaling Effect: Two by two matrix comparing the sign of $\Delta C(M) - \Delta C(G)$ and $\Delta E(M) - \Delta E(G)$[a]

| | | $\Delta E(M) - \Delta E(G)$ | |
|---|---|---|---|
| | | Positive | Negative |
| $\Delta C(M) - \Delta C(G)$ | Positive | 6 | 2 |
| | Negative | 15 | 6 |

[a] Statistical analysis: same sign = 12; different sign = 17 ($p = .87$).

**Table 3.** Analysis of Scaling Effect: Two by Two Matrix Comparing the Sign of $E_{BM} - E_{BG}$ and $C_{BM} - C_{BG}$[a]

| | | $E_{BM} - E_{BG}$ | |
|---|---|---|---|
| | | Positive | Negative |
| $C_{BM} - C_{BG}$ | Positive | 6 | 2 |
| | Negative | 7 | 14 |

[a] Statistical analysis: same sign = 20; different sign = 9 ($p = .031$).

**Table 4.** Analysis of Scaling Effect: Two by Two Matrix Comparing the Sign of $E_{AM} - E_{AG}$ and $C_{AM} - C_{AG}$[a]

| | | $E_{AM} - E_{AG}$ | |
|---|---|---|---|
| | | Positive | Negative |
| $C_{AM} - C_{AG}$ | Positive | 2 | 10 |
| | Negative | 5 | 12 |

[a] Statistical analysis: same sign = 14; different sign = 15 ($p = .64$).

### Time Horizon

When comparing the time horizon data, two more appraisals were excluded, which left a dataset of thirty appraisals. Twelve of the time horizons were greater in the manufacturer submission compared with that of the Assessment Group, nine of the time horizons were greater in the Assessment Report, and in nine of the appraisals the time horizon used was the same. Given a null hypothesis of equality for the length of the time horizons between submission source, and on the basis that only the direction of the difference and not the magnitude matters, no significant difference in time horizon was found ($p = .841$).

## DISCUSSION

Previous research has shown that studies sponsored by industry were more likely to have lower ICERs than nonindustry sources (1;3–5;7). Such differences should theoretically not be so apparent within NICE's TAP, as the manufacturer and AG estimates of the ICER are calculated at the same time, often using the same studies to populate the economic analysis; however, Miners et al. observed that the technology manufacturers submitting to NICE made claims of better cost-effectiveness compared with university-based AGs (8). The purpose of this study was to assess where the differences within the submissions had occurred.

The results from this analysis showed that AGs tended to estimate larger differences in cost ($\Delta C$) between the competing technologies compared with the manufacturers. The analysis also showed that estimates of incremental effectiveness ($\Delta E$) provided by the AGs were more likely to be smaller compared with the manufacturer estimates. These results are consistent with the observation reported by Miners et al., insofar as greater cost differences and smaller treatment effects, both associated with the AGs in this study, would tend to generate larger ICERs.

Further analysis of the data suggested that the differences in incremental costs ($\Delta C$) and effects ($\Delta E$) were the result of two factors. First, discrepancies between the expected costs of the technology under evaluation ($C_B$) provided by both parties were recorded (AGs tended to estimate larger average costs compared with the manufacturers). Second, differences between the estimated effectiveness of the comparator technology ($E_A$) were also recorded (AGs tend to estimate larger expected effectiveness values compared with the manufacturer). No statistically significant differences between the costs of the comparator and the effects of the technology under evaluation between the AGs and the manufacturers were observed.

An original study objective was to delve deeper into the evaluations in an attempt to identify specific reasons for differences in reported ICERs in terms of specific input parameters (e.g., relative risks of disease progression or drug prices). However, because the design and structure of the decision models varied so markedly within each appraisal, and because the methods and results from evaluations were not always well reported, the task was abandoned for all but the time horizon of the evaluations. Thus it is only possible to indicate potential areas where analysts should focus their critical appraisal skills rather than to highlight a particular set of input parameters *per se*. This is an interesting observation insofar as a frequent task for NICE's Technology Appraisal Committees is to explain why results from seemingly similar economic evaluations differ, and in so doing, explain their decision for placing more weight on the results from one evaluation over another in a coherent, logical, and defensible manner. That our attempt at doing this was abandoned highlights the complexity and difficulty of the Committee's task, although better reporting of study methods and results would have helped to reduce this problem.

NICE has recently introduced its "Single Technology Appraisal" (STA) program (9). It differs in several ways from NICE's traditional appraisal approach, but perhaps the main difference, in terms of method, is that only product manufacturers submit evidence. The AG's complete role is currently unclear, but appears to be limited to critically appraising this evidence rather than substantially adding to it. STA may put the burden of proof on one party, which Buxton and Akehurst state may lead to increased rejections (2), but a potential advantage may be that the Committee will no longer need to compare often seemingly incomparable decision models (6). The limited scope for the AG review of manufacturers' submission within the STA process underlines the need for a focused approach. This study shows that systematic differences may exist for $C_B$ and $E_A$ between the manufacturer submissions and the Assessment Report, and, therefore, may provide suggestions for where the AG should focus its analysis in the STA.

There are several limitations with this study. The problem of what we have called the Scaling Effect and the fact that we have only weak evidence for its nonexistence has already been outlined. Additionally, collecting the data was problematic in that value judgments were often required when abstracting the data. For example, the base case scenario was not always clearly identifiable, and appraisals often included multiple treatment comparisons. However, where value judgments were required, they were made with respect to minimizing possible differences between variables.

## POLICY IMPLICATIONS AND CONCLUSIONS

There are two main implications of this study. First, it is suggested that reviewers of economic evaluations pay particular attention to the methods used to estimate technology costs and comparator health effects when attempting to reconcile differences in reported ICERs across different studies. Second, Institutes such as NICE that are responsible for reimbursement decisions are encouraged to fully debate the pros and cons of receiving independent- and/or manufacturer-sponsored economic evaluations as broadly outlined by Buxton and Akehurst (2).

## CONTACT INFORMATION

**Deven Chauhan**, BPharm, MSc (dchuahan@ohe.org), Health Economist, Office of Health Economics, 12 Whitehall, London SW1A 2DY, UK
**Alec H. Miners**, PhD (alec.miners@lshtm.ac.uk), Lecturer, Department of Public Health and Policy, London School of Hygiene and Tropical Medicine, Keppel Street, WC1E 7HT, UK
**Alastair J. Fischer**, PhD (Alastair.Fischer@nice.org.uk), Senior Lecturer, Community Health Sciences Department, St. George's, University of London, Cranmer Tce, London SW17 0RE, UK; Health Economist, Centre for Public Health Excellence, National Institute for Health and Clinical Excellence (NICE), 71 High Holborn, London WC1V 6NA

### REFERENCES

1. Bell CM, Urbach DR, Ray JG, et al. Bias in published cost effectiveness studies: Systematic review. *BMJ*. 2006;332:699-703.
2. Buxton MJ, Akehurst R. How NICE is the UK's fast-track system? *Scrip Magazine*. 2006;152:24-25.
3. Ferner RE, McDowell, SE. How NICE may be outflanked. *BMJ*. 2006;332:1268-1271.
4. Freemantle N, Mason J. Publication bias in clinical trials and economic analyses. *Pharmacoeconomics* 1997;12:10-16.
5. Friedberg M, Saffran B, Stinson TJ, et al. Evaluation of conflict of interest in economic analyses of new drugs used in oncology. *JAMA*. 1999;282:1453-1457.
6. Hill S, Garattini S, van Loenhout J, O'Brien BJ, de Joncheere K. *Technology appraisal programme of the national institute for clinical excellence*. Geneva: WHO; 2003.
7. Lexchin J, Bero LA, DJulbegovic B, Clark O. Pharmaceutical industry sponsorship and research outcome and quality: Systematic review. *BMJ*. 2003;326:1167-1170.
8. Miners AH, Garau M, Fidan D, Fischer AJ. Comparing estimates of cost-effectiveness to the National Institute for Clinical Excellence (NICE) by different organisation: Retrospective study. *BMJ*. 2005;330:65-68.
9. National Institute for Health and Clinical Excellence. *Press release: NICE to issue faster drugs guidance for the NHS*. National Institute for Clinical Excellence, London, 2005. Available at: http://www.nice.org.uk/download.aspx?o=277752. Accessed March, 2006.