Department of Economics
School of Social Sciences

Does NICE have a cost effectiveness threshold and what other factors influence its decisions? A discrete choice analysis.

By Nancy Devlin and David Parkin

Department of Economics
Discussion Paper Series

No. 03/01

Does NICE have a cost effectiveness threshold and what other factors influence its decisions? A discrete choice analysis.

Nancy Devlin*
City University,
London

David Parkin
City University,
London

Abstract

The decisions made by the National Institute for Clinical Excellence (NICE) give rise to two questions: how is cost effectiveness evidence used to make judgements about the 'value for money' of health technologies? And how are factors other than cost effectiveness taken into account? The aim of this paper is to explore NICE's cost effectiveness threshold(s) and the tradeoffs between cost effectiveness and other factors apparent in its decisions. Discrete choice analysis is used to reveal the preferences of NICE and to consider the consistency of its decisions. For each decision to accept or reject a technology, explanatory variables include: the cost per life year or per QALY gained; uncertainty regarding cost effectiveness; the net cost to the NHS; the burden of disease; the availability (or not) of alternative treatments; and specific factors indicated by NICE. Results support the broad notion of a threshold, where the probability of rejection increases as the cost per QALY increases. Cost effectiveness, together with uncertainty and the burden of disease, explain NICE decisions better than cost effectiveness alone. The results suggest a threshold somewhat higher than NICE's stated 'range of acceptable cost effectiveness' of £20,000 - £30,000 per QALY - although the exact meaning of a 'range' in this context remains unclear.

Keywords: NICE, priority setting, cost effectiveness, equity, cost effectiveness thresholds.

* Corresponding author, address: Department of Economics City University London, **EC1V 0HB, Phone: 020 7040 8518, Fax: 020 7040 8580, e-mail: n.j.devlin@city.ac.uk**

Introduction

The National Institute for Clinical Excellence (NICE) was established in 1999 to address geographic variations in access in the UK ("postcode prescribing") by providing national-level guidance on the effectiveness and cost-effectiveness of new health technologies in the NHS. By May 2002 it had issued guidance on 39 technologies; an ambitious programme of technical appraisals in support of further recommendations is planned for the future. The role of NICE has been strengthened by making implementation of its decisions mandatory in the NHS from 2002.

An issue generating considerable speculation and debate is the weight that NICE attaches to cost-effectiveness evidence in its decisions [1,2] and, in particular, what decision rule it applies to incremental cost effectiveness ratios to judge whether any given technology represents good value for money [3]. Towse (2002) has suggested that the 'threshold' cost per quality adjusted life year gained (CQG) implicit in NICE's decisions is between £20,000 and £30,000; technologies with incremental cost-effectiveness ratios above this level seem more likely, but not certain, to be rejected [4]. Explicit statements made by NICE are contradictory. The NICE guidance on Orlistat for obesity in adults [5] contained a statement that a "sufficient level of cost effectiveness" is "in the range of CQG of between £20,000 and £30,000". Public comments made by the Chairman of NICE suggested that a threshold of £30,000 had emerged from its deliberations; however, NICE's evidence to the Health Select Committee Inquiry maintains that these comments were misinterpreted and that "the Institute does not have such a threshold" [6]. Not only is there is no clear and explicit cost-effectiveness threshold, there is also a lack of clarity over the way in which factors *other* than health gain are taken into account - that is, the tradeoffs that are accepted between efficiency and objectives such as equity.

The aim of this paper is to consider the factors that operate to influence NICE decisions, to explore systematically the influence of each and to establish the

characteristics of the cost-effectiveness threshold, if it exists. We report the results from initial data exploration and from a binary choice model using logistic regression analysis.

A binary choice model of NICE decision making

In the simplest case, illustrated in Figure 1, the threshold is a precise, 'knife-edge' value for a marginal QALY against which evidence from economic evaluation is compared: if the CQG exceeds this it is rejected; if it falls below it is accepted.

In practice, the threshold may be less clearly identified, for three reasons that we will consider in turn. First, decisions to accept or reject new technologies may depend on a wider set of objectives than maximising health gain from the NHS budget. Secondly, the cost effectiveness threshold may be different for investments and disinvestments. Thirdly, the threshold may be affected by the decision maker's response to uncertainty about evidence concerning cost-effectiveness.

The existence of factors other than cost-effectiveness may mean that there is in practice no threshold at all; any new technology has a finite probability of being accepted or rejected, whatever its CQG, if other factors are important enough to outweigh its cost-effectiveness. Alternatively, there may be no single threshold but a lower and an upper threshold, as in Figure 2. Below the lower threshold, low CQG technologies are certain to be accepted; above the upper threshold, high CQG technologies are certain to be rejected. *Within* the range between the two, cost effectiveness may be traded off against other objectives that are seen as relevant to decision making.

What are the factors other than cost effectiveness that may influence NICE decisions? An obvious candidate is equity [1, 7]; less obvious is *which* equity concerns are relevant to NICE's deliberations. The NHS has as one of its central equity principles access to health care irrespective of ability to pay. This is addressed

by the means by which technologies recommended by NICE are funded - general taxation, free at the point of delivery - and is therefore not relevant to its deliberations. Equal access regardless of geography is also an important equity consideration - the avoidance of the 'postcode lottery' (regional variation in access to some technologies) was, as already noted, a principal objective in establishing NICE. This is mainly addressed by weighted population based funding formulae, which aim to ensure that local health organisations have equal resources for equal need. NICE's role is in effect to ensure that this equal availability of resources is translated into equal availability of specific technologies. However, it is again hard to see how it could affect NICE's decisions, as its recommendations pertain to the NHS as a whole. Population characteristics *other* than income and geography, for example age, sex, ethnicity and social class, are possible foci of equity goals. But we do not believe that these are relevant equity criteria for NICE, as it has no mandate for differentiating between population sub-groups (e.g., by weighting QALYs[a]) and, arguably, discrimination legislation precludes its ability to do so on some grounds.

While the relevance of equity goals regarding *population* sub-groups can be largely rejected *a priori*, by contrast, concerns about equity between *patient* groups is likely to be highly relevant to NICE's decision making processes. We consider three ways in which this might be implemented. First, NICE might approach 'orphan' treatments (i.e., a treatment for a disease for which no alternative curative treatment for patients exists) differently from treatments for which there *are* treatment alternatives. Secondly, the 'starting point' in health status terms of particular patient groups - low initial quality of life and short duration of life - may be seen as a relevant and inadequately captured by the measures of health *gain* used in economic evaluation. Thirdly, cost effectiveness *ratios* do not differentiate between the size of the potential group of beneficiaries. A larger patient group, where the total health gain produced from a treatment is larger, may be approached differently from a smaller patient group. In each case, technologies may have a lower probability of being rejected for any given CQG.

A final other possible factor is related neither to efficiency nor equity, but is suggested by the evidence that NICE requires for its technical appraisals on the net budgetary effect of its guidance on the NHS. The role that this evidence does or should have in decision-making is an area of dispute. In one view, NICE makes its decisions on the basis of effectiveness and cost effectiveness evidence, with information on the impact on the NHS budget being used only to operationalise those decisions; that is, to plan how much the NHS budget would need to increase in total to support a new procedure, or what resources will need to be displaced at a local level to implement guidance. Birch and Gafni have argued "…the puzzle here is how recommendations can be made for maximising health gain from a given NHS expenditure where such recommendations require additional resource requirements (and of unknown opportunity cost). If budgetary impact is only important in planning future resource requirements then all interventions with net benefits[b] would be implemented and NICE recommendations would be a prescription for continued expansion of the NHS" [1]. In practice, the establishment of NICE has coincided with a period of unprecedented, planned increases in real NHS budgets, making it impossible to determine cause and effect from casual observation alone. However, as Raftery has noted, NICE has said 'yes' more often than it has said 'no' [2].

The second complicating factor in looking at a possible cost effectiveness threshold is that it may be different for investments and disinvestments. O'Brien *et al* provide evidence that the willingness to accept (WTA) values for relinquishing QALYs, by reducing or removing services, are higher than the willingness to pay (WTP) to obtain QALYs from new services [8][c]. This suggests that the cost effectiveness threshold may be lower at every given CQG for extant, as opposed to new, services, as in Figure 3.

The final complicating factor is that the decision maker's response to uncertainty regarding CQG evidence arising, for example, from sensitivity analysis, may alter the threshold. Figure 4 illustrates this. If NICE is risk averse, the probability of

rejection will be higher for any given base-case CQG for options associated with the possibility of a high CQG under alternative sets of assumptions, compared to options where the base case CQG is relatively robust to changes in assumptions. If NICE is a risk lover, it will be prepared to give the benefit of the doubt and the opposite will apply.

Claxton has made a compelling argument that such uncertainty should not in fact be used to make decisions about whether or not to approve or reject any technology unconditionally [9]. Instead, it should only be used to decide whether or not to seek further evidence to reduce the uncertainty. However, the extent this was accepted and used by NICE at the time it took its decisions is not recorded or known.

In terms of a binary choice model, the response variable is the probability that NICE will reject a given technology. In some cases, the NICE Guidance document involves a simple acceptance or rejection of a technology for "routine use", for example recommending that a drug should be available to all sufferers of a particular condition. However, many NICE Guidance documents specify both clinical groups for whom the technology is recommended and those for whom it is not [2]. In these cases, the guidance actually implies more than one decision: acceptance for some identified groups and rejection for others.

Our exposition of the threshold in the previous section immediately suggests a number of explanatory variables. Incremental cost effectiveness ratios are cited in many of the Guidance documents and in the underlying Technical Appraisal documents. The relevant equity variables can be measured by examining the characteristics of patient subgroups. The hypothesis that NICE is in fact indifferent to the cost impact of its guidance can be tested by including information on the budgetary impact on the NHS. The effect of uncertainty over the cost-effectiveness evidence may be measured by exploring the range of cost effectiveness results reported in the Guidance supporting each decision.

Methods and Data

Data were abstracted from NICE Guidance and Technology Appraisals available at May 2002, covering recommendations on 39 technologies, corresponding to 51 observable yes/no decisions. The abstracted data are available from the authors.

In abstracting the data, we adopted a consistent method, in which we took the information provided at face value and applied identical rules for processing them. We deliberately did not seek any clarification from NICE about whether the *reported* information reflected the information that they *believed* they had taken into account. The data are therefore internally consistent and uncontaminated by *post hoc* rationalisation by NICE. The drawback is that they are only an approximation to the information that was actually taken into account; however, there is no independent source of information about that. We subsequently allowed members of NICE's staff to inspect the abstracted data; however we changed them only in one case, where we had clearly not followed our methods correctly, and in other cases, where interpretation of published data was disputed, we retained our original judgements.

Data

In most cases the data are self-explanatory – however, the impact on the NHS is that of *approving* the technology and is not identical to the cost of the actual NICE decision, which may either have approved *or* rejected the technology. The quantitative variables used in modelling are summarised in Table 1.

A number of issues must be discussed about the data used for modelling. The cost-effectiveness ratio (CER) was the cost per QALY gained (CQG) where it is reported and the cost per Life Year gained (CLYG) where it is not. (The guidance for some decisions reports both.) This implies a one-to-one correspondence between life years gained (LYG) and Quality Adjusted Life Years gained (QALYG). This equivalence will be incorrect if those technologies for which CLYG is reported either have an

additional impact, positive or negative, on existing quality of life or produce life years of less than full quality, or both.

Although the main reason for doing this is to produce sufficient data for analysis, there are some justifications. First, in most cases the guidance for technologies for which only CLYG is reported either explicitly say that there are no quality adjustments to be made or implicitly do so by not mentioning this factor. This implies that NICE believes that the evidence is that LYG and QALYG are in fact the same in those cases. Secondly, an assumption of this analysis is that NICE decision makers respond mainly to data that they are presented with. In the absence of any data on quality adjustments to LYG, which could of course raise or lower the CQG, the lack of such quality adjustments may not be noticed or taken into account. Thirdly, a slighter weaker justification is that we are especially interested in the decisions that are "out-of-order" with respect to CERs, and these apparent anomalies are the same in the two subsets as they are in the data as a whole.

Another problem with the CERs is that they use inconsistent perspectives in costing. Some include patient costs (for example the technical appraisal of beta interferon for multiple sclerosis), some use public sector costs other than those to the NHS (for example, social care costs in the technical appraisal for Trastuzamab), still others restrict their perspective to the NHS. This suggests that the CQG cannot directly be compared between them and, indeed, that a different threshold would apply in each case. Again, however, these were the data which were considered by NICE and there is no evidence that the appraisal committees adjusted these formally or informally to convert them to a common bases for their decision making[d].

The variable UNCERTAINTY is represented in this analysis as the CQG range, as a measure of spread, divided by the base case or mean, as a measure of central tendency. A better measure would be the coefficient of variation of the CER. However, the data are rarely, if ever, presented in the form of a distribution from which a true mean and standard deviation can be calculated.

The equity variable OTHER FACTORS is indicated in only three instances – the treatments considered for Motor Neurone Disease (MND), pancreatic cancer and non-small cell lung cancers. In each case, the Guidance makes particular reference to health status 'starting point' issues in its decision. The clearest indication of factors other than cost-effectiveness influencing its deliberations is provided by the guidance for Riluzole for MND, in which it is noted that "The Committee took account of the severity and relatively short life span of people with ALS and, in particular, as directly reported to it, of the values which patients place on the extension of tracheostomy free survival time" [12]. The Guidance for treatments for pancreatic cancer and non-small lung cell cancer refer to the "extremely poor prognosis" and low survival rates in each case, although CQG evidence is not available in either decision.

The variable BURDEN covers four different, and incompatible, measures of the number affected. These are the number of cases, new cases, treatments and deaths. We have combined these mainly in order to provide sufficient cases for analysis, but again there is a weak argument that NICE decision makers may respond to the magnitudes presented and may not distinguish too finely between the different types of burden.

The variable IMPACT, representing the budgetary implications for the NHS also suffers from inconsistent evidence provided to and by NICE. Some are not incremental (for example, drug costs are 'totals', not taking into account existing spending on that drug where it is already used in the NHS); others take partial account of changes in resource use (for example, where use of one drug or procedure replaces another); and others are based on a fuller account of changes in resource use (for example, changes in hospital use arising from longer term changes in morbidity). We have had to treat these comprehensive estimates of net impact as equivalent to estimates of the direct cost of the technology and to make estimates of the average where only ranges are given. Again, there is only a supporting

argument that magnitudes are what are visible and may enable differences in meaning to be neglected.

A more general point is that abstraction of the data is difficult not only because of the complexity of the evidence in the guidance and technical appraisal reports and the many differences between the data presented in them, but also because it is not clear which data were actually taken into account by NICE. There is in many cases a disparity of some magnitude between the CQG reported in the guidance and that reported in the supporting technical appraisal. Previous analyses of the role of cost effectiveness evidence in NICE decisions have restricted their analysis to the evidence reported in the guidance [2,4]. While it is to be expected that the guidance committee would consider factors other than cost effectiveness in issuing their recommendations, the committee also supplements the independent technical appraisals with confidential manufacturers' evidence as well as incorporating more 'casual' economic evidence and reasoning. In some cases, the cost per QALY in the guidance is the committee's best guess about cost effectiveness, taking into account factors they consider not to be included in the technical appraisal proper.

For example, the guidance for Riluzole for MND cites a CQG for this treatment of £34,000 - £43,500. The technical appraisal has a base case CQG of £58,000, with sensitivity analysis revealing a range of CQG from negative to a considerably higher than base-case CQG. Later revisions to the technical appraisal, in the light of new evidence, suggest the CQG ranges from £16,500 - £20,000. The wide range of results is illustrative of the degree of uncertainty NICE faces in using cost-effectiveness evidence. However, for the purpose of this analysis, the key point to note is that the CQG figures cited in the guidance cannot be located in the evidence in any of the technical appraisals it has provided in support of its decision.

A second example is the cost-effectiveness data for Orlistat for obesity in adults. As stated earlier, the NICE guidance for Orlistat seemed to admit to an approximate threshold of £20,000 to £30,000 per QALY gained and it further implies, by its

favourable decision, that Orlistat meets this. However, the "headline" estimates that it gave for this CER are a much higher independent estimate of £46,000, which is neither endorsed nor rejected, and an explicitly rejected much lower figure based on manufacturers' estimates.

These "headline" figures are taken from the technical appraisal, which itself makes them the "headline". But the technical appraisal also reports, in a less prominent way, some sensitivity analyses around the figures, which are not referred to in the guidance and which support a lower CER. There is however, nothing within those analyses which supports the figure of £20,000-£30,000.

The technical appraisal takes the independent estimate from a Development and Evaluation Committee (DEC) report [13], which again headlined the £46,000 figure, but also contains far more sensitivity analyses than those reported in the technical appraisal. These also support a far lower CER. Of particular note is that the headline figure is based on the DEC method of calculating QALY gains, but the report also reports a much lower set of CER estimates based on EQ-5D utilities, explicitly noting that NICE appears to be preferring estimates based on the EQ-5D. These CERs did not find their way into the technical appraisal and it is unclear whether or not they will have been considered by NICE. However, once again, there is no obvious source in the DEC report for the £20,000 to £30,000 figure.

A further example of the difficulty in interpreting *what* cost effectiveness evidence influenced NICE's decisions is the case of beta interferon. The guidance for this technology indicated that the 'best available evidence' on CQG was considered to be £35,000 to £104,000 (under an assumption of benefit persisting over 20 years) and £120,000- £339,000 (under an assumption of benefit ceasing when treatment stops). Without being party to the decision making process, the independent observer can note only that the range of CQG indicated by NICE to be reliable evidence is £35,000 to £339,000 (with a midpoint of £187,000) although clearly other interpretations are possible.

The conclusion is that there is some uncertainty in many cases about what NICE's conclusions about cost-effectiveness were, the means by which they were derived and indeed what evidence they took into account in deriving them. Nevertheless, we take them at face value.

Modelling

The decisions were initially divided into those for which cost-effectiveness data were available and those for which there was none. The latter were subdivided into acceptance and rejection decisions and these were investigated qualitatively to uncover the apparent reasons for rejection or acceptance. The former were amenable to quantitative analysis of a binary choice model, which was explored using logistic regression analysis.

Several model specifications incorporating different numbers of variables were estimated. However, this was for reasons other than a model building strategy. First, our intended strategy was to estimate a model incorporating all variables, but this could not be done, for reasons explained below, and alternative specifications including fewer variables had to be examined. Secondly, models incorporating fewer variables were in themselves instructive about the impact of the implied decision making criteria. Because we are not estimating a full model, robust standard errors were used in all specifications as a basis for hypothesis testing.

The modelling was constrained by the absence of some items of data. The variable OTHER FACTORS could not be used, because the decision for all but one of the cases for which other factors were recorded was to accept and, for the other, data on cost-effectiveness were not available; OTHER FACTORS therefore had no explanatory power. One decision (27a) was excluded because of missing burden of disease data, but this was not an unusual case and had little effect on the results. The variable IMPACT was also available only for a restricted set of decisions, which

unfortunately included four of the seven rejections; moreover these were amongst the more interesting cases. As a result, we used a basic set of 33 decisions, although we also used a subset of 26 to test the IMPACT variable.

The logistic regression estimates permit calculations of a probability based "threshold". The probability of acceptance or rejection can be calculated for each cost-effectiveness ratio (CER), other variables held constant. There are two possible approaches, both problematic. One is to evaluate a strict marginal effect, assuming all other variables are zero; however, it is unrealistic, for example, to assume that there is no burden of disease. The other is to evaluate at the mean values of the other variables; however, this means that the estimates are highly dependent on their values in the sample, which is not random. From this, the CER at which the probability of acceptance is 0.5, equivalently where the odds ratio is 1, can be calculated.

Given the difficulty noted earlier in selecting, from the CQG evidence reported in NICE guidance, that which most directly influenced decision-making, the sensitivity of the estimated model to choices about the mid-point and range was explored for those technologies in which this issue was present. A special case was chosen, decision 32, because this has the highest CER using the data abstraction rules that we operated, but for which it might be argued that the guidance could be interpreted as giving a much lower estimate. The models were therefore re-estimated using the alternative data for that observation.

Results

Table 2 details the decisions for which there were no cost-effectiveness data, divided into those accepted and those rejected. Although the guidance is in many cases complex, the general conclusion, as might be expected, is that those rejected are those for which there is clear evidence that the technology is *not* effective, for example decision 1, or very unclear evidence that it *is* effective, for example decision

23. The technologies that were accepted without evidence on CLYG or CQG fell into two groups. In five cases, the decision is arguably better characterised as *which* treatment of those considered is most appropriate (for example, which type of prostheses should be used in hip replacement, given their differences in cost and revision rate), rather than whether any treatment *per se* is acceptable value for money relative to other NHS activities. Decisions reflect cost-minimisation or effectiveness-maximisation. In the remaining four cases, cost effectiveness evidence was considered, but in terms other than CLYG or CQG (for example, cost per year of remission).

Table 3 shows those decisions for which there was evidence in terms of either CLYG or CQG. These are ordered from lowest to highest CER, with rejection decisions shown in bold. Those with the three highest CERs (18b, 27b and 32) are rejections. Rejection decision 5 appears to be an outlier, but it is less clear whether it is rejection decisions 30b, 33a and 15b, or the fourteen acceptance decisions within and above them, that are "out of order". A threshold of the type shown in Figure 1 cannot be identified, since there are rejections that have a lower CER ratio than some acceptances. A threshold of the type shown in Figure 2 can be identified, but it would be of very doubtful meaning, since the range would be between a lower threshold of £1,000 to £1,100 and an upper threshold of £47,000 to £50,000, encompassing all decisions except two rejections and two acceptances. There is also no evidence of the alleged £20,000 to £30,000 range; two are rejected below that, five are accepted above it and all but one within it are accepted – and that at the top of the range.

An attempt was made to estimate a logistic regression model including all of the usable variables except OTHER FACTORS, excluded for reasons explained earlier. Unfortunately, this proved not to be estimatable, providing completely determined outcomes along with odds ratios and standard errors that had bizarre signs and magnitudes.

Table 4 shows the results that were obtained from the logistic regression analyses; Tables 5 and 6 show how these affect both the implied NICE "threshold" and which decisions conform to it. Table 5 shows the technologies in ascending order of probability of rejection. The horizontal bars identify the probability-based threshold as the point at which the probability of rejection becomes greater than the probability of acceptance, other things being equal. The CER at which the probability is 0.5 is shown in the first column of Table 6, headed "Central Estimate", using the 'marginal' and 'mean value' methods.

Model 1 includes only cost-effectiveness. The odds ratio is of the expected sign but is not significant at conventional levels ($p = 0.189$), so the model is included only for comparison. It correctly classifies all of the acceptances, but only two out of seven of the rejections. All of the other rejections are 'out of order' because they have rejection probabilities below 0.5.

Model 2 in Table 4 adds UNCERTAINTY to CER as an explanatory variable. Both have odds ratios which are significant and of the expected signs. Rejection decision 5 is no longer an outlier, having the third highest rejection probability. Sensitivity improves, with a small deterioration in specificity: rejection decisions 15b, 30b and 33b remain out of order, and acceptance decision 22 becomes out of order, because it has a rejection probability greater than 0.5. The implied threshold falls by between 15% and 26%.

Although the UNCERTAINTY variable seems to explain well rejection decision 5 despite the low CER and is consistent with statements within the relevant guidance, closer examination shows a less clear picture. The UNCERTAINTY variable measures the degree of variability about the CER, which is certainly relatively high in this case. However, the absolute value of the high estimate of the CER is still well below the implied threshold. The implication is either that NICE's decision took account of the presence of high uncertainty but ignored the significance of this uncertainty, or that the level of uncertainty actually attached to the estimate by NICE

was far higher than that reported. We have no way of judging which of these is closer to the truth.

Model 3 in Table 4 adds the variable BURDEN to Model 2 to assess its incremental impact. An alternative is also to examine a model including CER and BURDEN alone, but the estimated odds ratios for that model were not significant. In Model 3, all are significant and have the expected signs; those for CER and UNCERTAINTY are slightly higher than in Model 2 but have similar standard errors. Specificity improves, with no loss of sensitivity. No acceptance decisions are out of order; acceptance decision 22 now has a probability of rejection below 0.5, because of the very large number of people affected. The out of order rejection decisions remain. The implied threshold reduces further using the marginal method, but is slightly higher using the mean value method.

Model 4 adds the variable OTHER THERAPY. Table 4 shows that the odds ratios are again all significant and have the expected signs. CER, UNCERTAINTY and BURDEN have similar odds ratios to those in Model 3, but the standard errors are slightly higher. Sensitivity improves a lot, but with a small reduction in specificity. Rejection decision 15b is now no longer out of order; however rejection decisions 30b and 33b remain so. In addition, there are now two out of order acceptance decisions, numbers 22 and 34. The implied threshold unambiguously further reduces.

There is evidence of a need to be cautious with Model 4. The odds ratio for OTHER THERAPY is extremely small and its curious effects can be observed in Table 5, where the five decisions affected by this variable (that is, technologies for which no other treatment is available) are shown in italics. Four are given extremely low rejection probabilities, two of which, numbers 20 and 23, were previously at the margin of acceptance and rejection and whose removal from proximity to the threshold will have affected where the threshold is. The other decision is number 32, whose unchanging position as a certainty for rejection is virtually guaranteed by its highly unfavourable CER.

The χ^2 statistic suggests that none of the three models can be rejected. Model 4 is preferred because of its higher Pseudo R^2 and better sensitivity, achieved at the cost of a slightly lower specificity.

The variable IMPACT proved not to be useful, and results are omitted. For example, in adding it to Model 3, it requires a more restricted data set, as explained earlier, that left only three rejections, all of which are fully explained by the CER and UNCERTAINTY variables. Its inclusion had a large effect on the odds ratio of UNCERTAINTY, gave both BURDEN and IMPACT unexpected, though insignificant signs, and produced a completely determined model, with no out of order decisions. A further model adding both OTHER THERAPY and IMPACT to Model 3 was also not properly estimatable, again producing a completely determined model with odd and erroneous odds ratios and standard errors.

Figure 5 shows the relationship between probability of acceptance or rejection and the CER[e]. Observing points horizontally from the vertical axis enables us to assess the threshold, as the point at which probability of acceptance and rejection is equal, along the p=0.5 line. This demonstrates that the inclusion of other factors lowers the threshold, making it more difficult for high CER technologies to be accepted, other things being equal. However, observing points vertically from the horizontal axis suggests that at higher levels of the CER, the inclusion of other factors increases the probability of rejection, dramatically around £40,000 and above. By contrast, at low CER levels, including other factors lowers the probability of rejection.

All of these estimates raise the question of what is meant by a "range" of cost-effectiveness as NICE described it. As noted, it is not possible to estimate this in terms of an upper and lower "threshold". A confidence interval around the estimates can be calculated, based on the standard errors of the coefficients, but this describes the precision of the central estimate, not a range of acceptability. An alternative is to look at a range of acceptability as a range of the probability of

acceptance or rejection. Unfortunately, there is no obvious definition of what range of probability should be used. The remaining columns of Table 6 demonstrate the "threshold" ranges defined by three different probability ranges around the value 0.5 – plus or minus 0.05, 0.05 in each "tail" and the central 50%. However, even these calculations do not seem to capture the indistinct concept of "a range of acceptable cost-effectiveness".

Finally, inclusion of the alternative data for decision 32 had very little impact on the models and the implied threshold; the results are therefore omitted. The model including only CER was improved, and the implied threshold was affected. However, for the other models, the models did not perform as well and the implied threshold was very similar. Because this was such an influential observation, a conclusion could be that the method was robust.

Discussion

The following tentative conclusions can be drawn from the analysis of NICE decisions for which there were cost-effectiveness data. There is support for the idea of a threshold as being probability based rather than a single number. NICE decisions are well explained by the cost-effectiveness evidence, with the effect of uncertainty and of the burden of disease explaining the rejection of some technologies with a relatively low CER and the acceptance of some with a relatively high CER. There remain a few anomalies, which may be the result of the context in which these decisions were taken; essentially, that they were decisions based on comparisons of different indications for particular technologies rather than taking into account a comparison with other services that the NHS provides. The analysis suggests a cost-effectiveness threshold somewhat higher than the £20,000-£30,000 which NICE has publicly identified.

The conclusion that the "threshold" estimates become lower with the inclusion of extra variables is supported. However, the estimate for Model 4 is at the top end of

the range mentioned in the paper, so there is less support for the claim that the threshold is "in the range of £20,000 - £30,000", as the estimates are £35,000 - £40,000.

The modelling results are of course exploratory and are far from definitive. A key problem is the data used, which are less than perfect. However, the decision making that we are trying to model presumably had the same imperfect data. Unless NICE decisions are based upon data which are not made available to the public, which is definitely the case for manufacturers' commercial-in-confidence data, then the modelling is using the correct data, though of course they may not have been interpreted by the NICE appraisal committees in exactly the same way as we have done.

The analysis may have revealed as much about the data upon which decisions have been based as about the decision making itself. NICE should not be criticised for making decisions based on imperfect and missing data; its role is to exercise judgement where that is the case. Of more concern are the widespread inconsistencies in the type of evidence collected and reported, the mismatch between figures reported in the guidance, the technical appraisals and other documents and the obvious existence of key documents and analysis which are not in the public domain. A consequence of this is a reduction in the explicitness and transparency of the decision-making process.

The insights from this work will provide a systematic way of identifying the role of various types of evidence on decision making retrospectively. Inferring the threshold from past decisions usefully serves to promote debate about the threshold and about how NICE should respond to equity concerns. However, the question of what NICE's cost-effectiveness threshold *should* be, as opposed to what it appears to have *been*, is more challenging still and ultimately needs to be supported by stated, rather than implicit, valuations of health outcomes.

Notes

[a] The choice of QALYs (and LYG) as the measure of benefit in cost effectiveness analysis in itself embodies value judgements about the value of health gains between age groups.

[b] This begs the question of what 'net benefit' means in this context. NICE's use of cost effectiveness, rather than cost benefit, analysis means that net benefit can only be determined by invoking a threshold. If the threshold reflects the existing NHS budget (e.g., if it is revealed by the CQG of the last treatment funded from the NHS budget - the 'extra welfarist' position) rather than an externally-determined, stated value per QALY gained (the 'welfarist' approach - see O'Brien et al 2002) this argument becomes invalid.

[c] The asymmetry described by O'Brien *et al* is arguably even more exaggerated in NICE decisions since its remit is to consider cost effectiveness *and* clinical effectiveness [10]. While NICE recommendations in favour of new technologies indicates its WTP for increased effectiveness, it is less clear how it would respond to an option which is cost effective because it is both less costly *and less effective* than current practice. If cost-effectiveness and effectiveness are equally dominant in NICE's preference function, its WTA would be infinitely large i.e., there would be no reduction in costs sufficient to compensate for reduced effectiveness. We do not test for this possibility here, as all options considered by NICE to date sit in the North-East and South East quadrants of the cost effectiveness plane described by O'Brien *et al* i.e. have zero to positive effectiveness.

[d] NICE's guidance to manufacturers and sponsors for preparing evidence [11] now provides clear advice on the perspective to be taken in economic evaluation.

[e]Although this has a shape similar to the theoretical diagram in Figure 2, the interpretation is different. The theoretical curve identifies a threshold range in which there is uncertainty about what will be rejected or accepted. The empirical curve identifies not a range but a single threshold value given a chosen level of probability. The empirical method cannot in any case identify upper and lower threshold values, because probability values of 0 and 1 cannot be observed except at

the limits of precision of calculation. Again, despite their similar shape neither type of figure is related to Cost Effectiveness Acceptability Curves.

References

1. Birch S, Gafni A. On being NICE in the UK: guidelines for technology appraisal for the NHS in England and Wales. *Health Economics* 2002; **11**:185-191.

2. Raftery J. NICE: faster access to modern treatments? Analysis of Guidance on health technologies. *British Medical Journal* 2001;**323**:1300-1303.

3. Devlin N, Towse A.(eds). *Cost effectiveness thresholds: economic and ethical issues.* 2002 London: King's Fund/Office for Health Economics.

4. Towse A. What is NICE's threshold? An external view. Chapter 2 in: Devlin N, Towse, A.(eds) *Cost effectiveness thresholds: economic and ethical issues*. 2002 London: King's Fund/Office for Health Economics.

5. NICE *Guidance on the use of Orlistat for the treatment of obesity in adults.* Technology Appraisal Guidance No. 22. 2002.

6. NICE *Health Select Committee Inquiry: submission of supplementary evidence. Part 2 of 2.* www.nice.org.uk/Docref.asp?d=27652 [31st March 2002]

7. Hutton J, Maynard, A. A NICE challenge for health economics. Editorial. *Health Economics* 2000;**9**:89-93.

8. O'Brien BJ, Gersten K, Willan AR, Faulkner LA. Is there a kink in consumers' threshold value for cost effectiveness in health care? *Health Economics Letters* 2002; **6** (1).

9. Claxton, K. The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. *Journal of Health Economics*, 1999; 18(3), 341-364.

10. Department of Health. *Faster access to modern treatment: how NICE will work*. 1999 London: HMSO.

11. NICE *Guidance for manufacturers and sponsors.* Technology Appraisals Process Series 5. June 2001.

12. NICE *Guidance on the use of Riluzole for the treatment of Motor Neurone Disease.* Technology Appraisal Guidance No. 20. 2001.

13. The Wessex Institute. *Orlistat for the treatment of obesity*. Development and Evaluation Committee report No. 101, 1999.

Acknowledgments

The analysis in this paper was undertaken while Nancy Devlin was employed by the King's Fund. The authors are grateful to Kim Stirling at the King's Fund, whose assistance was invaluable. We benefited from useful comments on an earlier version of this paper presented at a King's Fund seminar and at the Health Economists Study Group, Brunel University, July 2002, and in particular thank the HESG discussant, Alec Miners. The usual disclaimers apply.

Figure 1

The cost-effectiveness threshold as a point

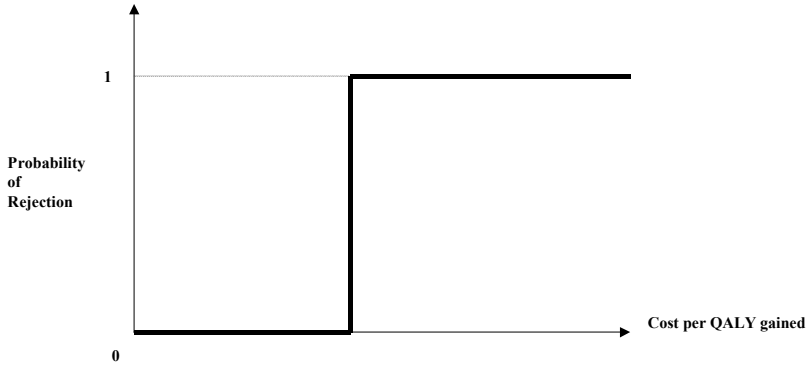

Figure 2

The cost-effectiveness threshold as a range, reflecting tradeoffs against efficiency

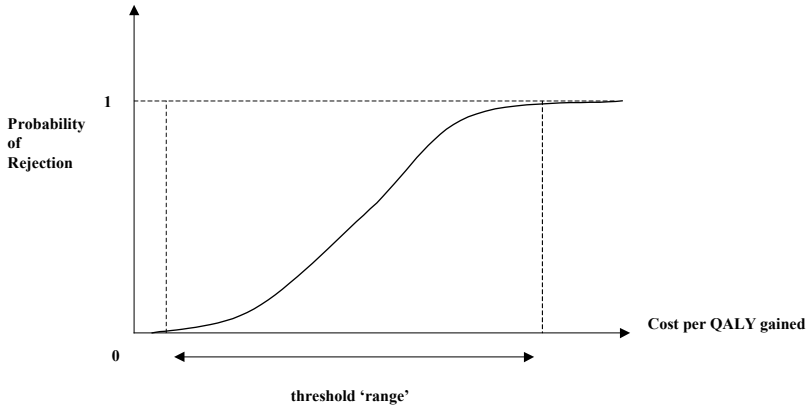

Figure 3

The cost-effectiveness threshold for investments and disinvestments

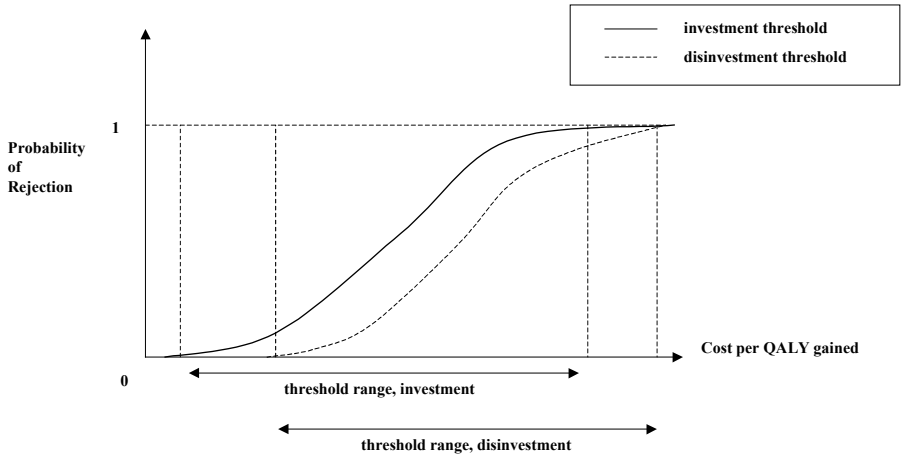

Figure 4

The cost-effectiveness threshold under uncertainty

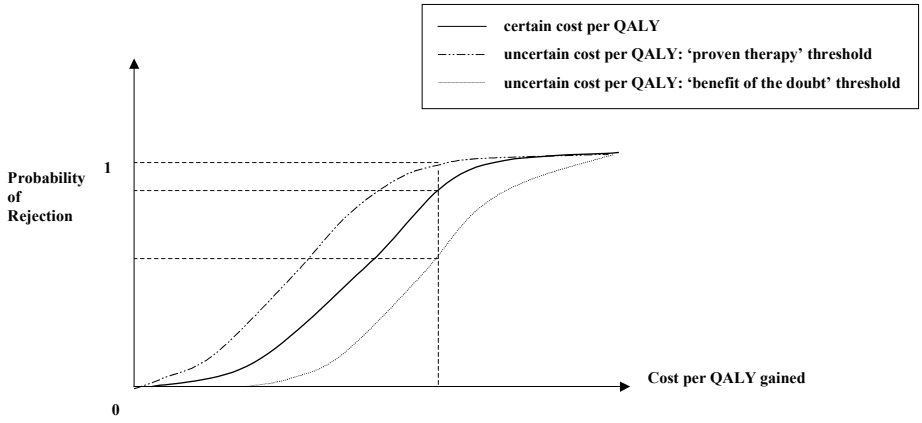

Figure 5. Probabilistic cost effectiveness thresholds for NICE decisions

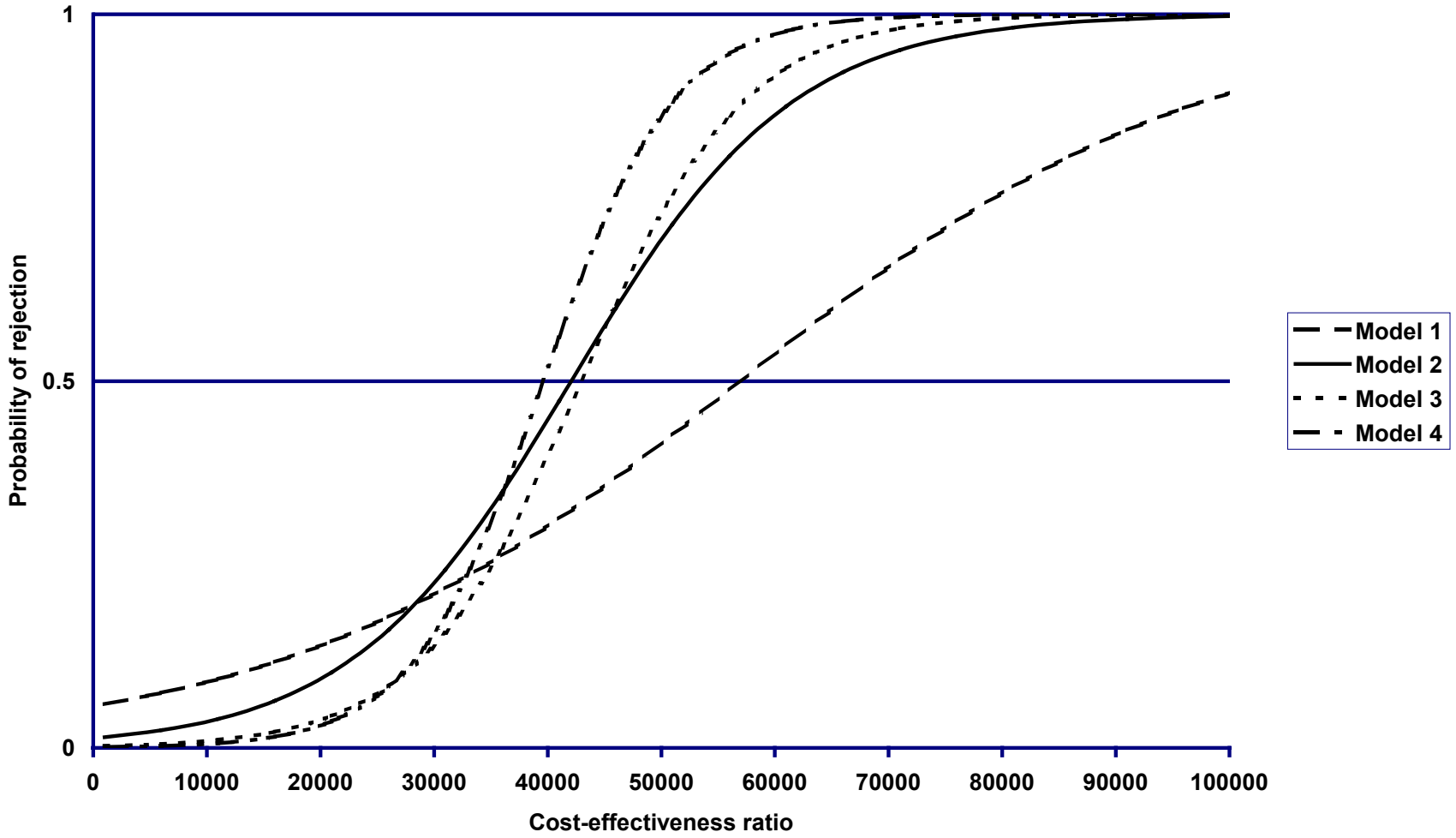

Table 1. Summary of variables

Variable name	Variable construction
Dependent variable:	
DECISION	A binary choice variable which takes the value 0 if the decision is in favour of use, 1 if against.
Independent variables:	
CER	Cost per quality adjusted life year gained or cost per life year gained in £1,000 units. Where a single estimate or a central or base case estimate is provided, this is used. If only a range is given, the mean is used.
UNCERTAINTY	Uncertainty about the cost-effectiveness evidence, calculated as the range of cost effectiveness ratio divided by the mean or base case cost-effectiveness ratio.
OTHER THERAPY	A dummy variable set to 1 if there are no treatment options and 0 if alternative treatments are available as a substitute for the treatment under consideration.
OTHER FACTORS	A dummy variable set to 1 if the guidance report makes specific mention of other variables influencing its recommendation (severity of condition, short duration of life etc), 0 otherwise.
BURDEN	The burden of disease, represented by the number of people affected by the condition to which the option pertains, in 1,000 units.
IMPACT	The impact on NHS budgets of approving the treatment in £1m units. Where the net impact is available this is used, otherwise it is the impact of the technology itself. Where a range is given, the mean is taken.

Table 2. NICE decisions for which there were no cost-effectiveness data

Accepted		Rejected	
Guidance number	*Technology*	*Guidance number*	*Technology*
2	Hip	1	Wisdom
7	Proton pump	16	Cartilage
8	Hearing aid	17	Laparascope colorectal
9	Rosiglitazone	23	Temozolamide (First line)
10	Inhalers	25	Gemcitabine (Other)
21	Pioglatizone	28	Topecetan No
24	Wound care	33	Advanced colorectal 4
29	Fludarabine	37	Rituximab
33	Advanced colorectal 2		

Table 3. NICE decisions ranked by cost-effectiveness ratio (CER)

Guidance Number	*Technology*	*CER*
39	Smoking	£430
28	Topecetan Yes	£1,000
5	**Cytology**	**£1,100**
38	Asthma inhalers	£5,000
3	Taxane Ovarian	£8,271
12	Glycoprotein	£9,250
26_a	Non-small cell lung (First line)	£9,475
13	Methylphenidate	£12,500
25	Gemcitabine (First line)	£12,950
26_b	Non-small cell lung (other)	£14,000
19	Alzheimers	£15,000
30_a	Taxane Breast 2 (Second line)	£15,250
6	Taxane Breast	£15,500
30_b	**Taxane Breast 2 (First line)**	**£19,000**
34	Trastuzumub (monotherapy)	£19,000
15_a	Zanamavir At Risk	£20,400
14	Ribavarin	£20,500
33_a	Advanced colorectal 3	£22,500
31	Sibutramine	£22,500
35	Arthritis juvenile	£22,500
18_a	Laparascope hernia (recurrent)	£25,000
4	Stents	£25,000
11	ICDs	£28,500
33_b	**Advanced colorectal 1**	**£29,000**
36	Arthritis adult	£31,000
23	Temozolamide (Second line)	£35,000
34	Trastuzumub (combination)	£37,500
15_b	**Zanamavir All**	**£38,000**
20	Riluzole	£38,750
22	Orlistat	£46,000
18_b	**Laparascope hernia (primary)**	**£50,000**
27_b	**Cox II (Routine)**	**£150,000**
32	**Beta interferon**	**£187,000**

Notes: The number in column one is the NICE Guidance number corresponding to the technology described in column two. Where a Guidance report contained more than one decision (for example, approval for one sub-group of patients but not for another), these are differentiated in this and following Tables by letter subscripts. Rejections are in bold.

Table 4. Logistic regression analyses of NICE decisions

	Model 1		Model 2		Model 3		Model 4	
CER	1.050495	(.0394156)	1.108109	(.055878)	1.150873	(.0572549)	1.192247	(.0754079)
UNCERTAINTY			2.162435	(.6292503)	2.538829	(.7283039)	2.693232	(.8065491)
BURDEN					.9927932	(.0032699)	.9899055	(.0043667)
OTHER THERAPY							5.59e-07	3.07e-06
N	33		33		33		33	
Log likelihood	-12.955545		-10.544938		-9.5186334		-8.108975	
Pseudo R^2	0.2403		0.3816		0.4418		0.5245	
Pearson χ^2	30.72	(0.2829)	26.66	(0.5901)	23.33	(0.7611)	18.65	(0.9086)
Sensitivity	28.57%		57.14%		57.14%		71.43%	
Specificity	100.00%		96.15%		100.00%		92.31%	
Correctly classified	84.85%		87.88%		90.91%		87.88%	

Notes: Robust standard errors for odds ratios and probability values of the Pearson χ^2 Goodness of fit statistics are shown in parentheses.

Table 5. Ranking of NICE Guidance decisions by cost-effectiveness ratio and probability of rejection

Model 1		
	CER (£)	Prob
39	430	.0580
28	1000	.0596
5	**1100**	**.0599**
38	5000	.0716
3	8271	.0831
12	9250	.0869
26_a	9475	.0878
13	12500	.1004
25	12950	.1025
26_b	14000	.1073
19	15000	.1121
30_a	15250	.1134
6	15500	.1147
30_b	**19000**	**.1333**
34	19000	.1333
15_a	20400	.1415
14	20500	.1421
33_a	22500	.1545
31	22500	.1545
35	22500	.1545
18_a	25000	.1713
4	25000	.1713
11	28500	.1971
33_b	**29000**	**.2011**
36	31000	.2174
23	35000	.2528
34	37500	.2767
15_b	**38000**	**.2817**
20	38750	.2892
22	46000	.3677
18_b	**50000**	**.4146**
27_b	**150000**	**.9899**
32	**187000**	**.9983**

Model 2		
	CER (£)	Prob
28	1000	.0076
38	5000	.0115
39	430	.0132
12	9250	.0255
3	8271	.0276
26_b	14000	.0284
13	12500	.0323
34	19000	.0466
25	12950	.0495
26_a	9475	.0562
30_a	15250	.0711
6	15500	.0736
30_b	**19000**	**.0855**
33_a	22500	.0926
31	22500	.1048
35	22500	.1048
4	25000	.1096
15_a	20400	.1155
33_b	**29000**	**.1200**
11	28500	.1292
19	15000	.1316
14	20500	.1546
36	31000	.1697
23	35000	.2016
34	37500	.2460
15_b	**38000**	**.2557**
18_a	25000	.2972
20	38750	.3095
18_b	**50000**	**.5408**
22	46000	.5882
5	**1100**	**.7996**
27_b	**150000**	**.9999**
32	**187000**	**1**

Model 3		
	CER (£)	Prob
38	5000	.0000
39	430	.0031
28	1000	.0037
12	9250	.0084
36	31000	.0158
13	12500	.0164
26_b	14000	.0186
3	8271	.0200
14	20500	.0273
19	15000	.0285
26_a	9475	.0398
25	12950	.0436
34	19000	.0445
6	15500	.0609
4	25000	.0644
30_b	**19000**	**.0664**
30_a	15250	.0878
33_a	22500	.0923
11	28500	.0955
31	22500	.0965
35	22500	.1210
15_a	20400	.1372
33_b	**29000**	**.1385**
18_a	25000	.2541
23	35000	.3151
34	37500	.3852
15_b	**38000**	**.3907**
22	46000	.4771
20	38750	.4914
18_b	**50000**	**.6394**
5	**1100**	**.8784**
27_b	**150000**	**.9386**
32	**187000**	**1**

Model 4		
	CER (£)	Prob
38	5000	.0000
19	*15000*	*.0000*
14	*20500*	*.0000*
23	*35000*	*.0000*
20	*38750*	*.0000*
39	430	.0014
28	1000	.0022
12	9250	.0051
36	31000	.0084
13	12500	.0125
3	8271	.0163
26	14000	.0164
26	9475	.0329
25	12950	.0422
34	19000	.0502
6	15500	.0602
4	25000	.0662
30_a	15250	.0685
30_b	**19000**	**.1008**
33_a	22500	.1117
31	22500	.1129
11	28500	.1140
35	22500	.1537
15	20400	.1692
33_b	**29000**	**.1954**
18_a	25000	.2872
34	37500	.5774
15_b	**38000**	**.5827**
22	46000	.6013
18_b	**50000**	**.8184**
5	**1100**	**.8763**
27_b	**150000**	**.9168**
32	***187000***	***1***

Notes: Numbers in bold indicate the Guidance number, CER and corresponding probability for technologies rejected by NICE. The bold horizontal bar in each column indicates the technologies and corresponding CERs between which lies a probability of rejection of 0.5.

Table 6. Probabilistic cost effectiveness thresholds for NICE decisions

Model		Central Estimate	Probability 10% range 0.45	0.55	90% range 0.05	0.95	50% range 0.25	0.75
CE only	Marginal	£57,216	£53,116	£61,317	-£2,954	£117,387	£34,766	£79,667
CER + UNCERTAINTY	Marginal	£48,409	£46,454	£50,364	£19,726	£77,091	£37,707	£59,111
(Model 1)	Mean value	£42,268	£40,313	£44,222	£13,585	£70,950	£31,566	£52,969
CER + UNCERTAINTY + BURDEN	Marginal	£40,519	£39,091	£41,947	£19,565	£61,472	£32,700	£48,337
(Model 2)	Mean value	£43,139	£41,711	£44,567	£22,185	£64,093	£35,321	£50,957
CER + UNCERTAINTY + BURDEN	Marginal	£35,380	£34,239	£36,521	£18,635	£52,125	£29,132	£41,628
(Model 3)	Mean value	£40,216	£39,075	£41,358	£23,471	£56,962	£33,969	£46,464