HERC
HEALTH ECONOMICS RESEARCH CENTRE
NUFFIELD DEPARTMENT OF POPULATION HEALTH

UNIVERSITY OF OXFORD

HERC Research Paper

The influence of cost-effectiveness and other factors on NICE decisions

Helen Dakin, Nancy Devlin, Yan Feng, Nigel Rice,

Phill O'Neill, David Parkin

For more information contact helen.dakin@dph.ox.ac.uk

HERC Research Paper 05/14

The influence of cost-effectiveness and other factors on NICE decisions

Dakin Helen[1], Devlin Nancy[2], Feng Yan[2], Rice Nigel[3], O'Neill Phill[2], Parkin David[4]

1. Health Economics Research Centre, University of Oxford, UK
2. Office of Health Economics, London, UK
3. Centre for Health Economics and Department of Economics and Related Studies, University of York, UK
4. Department of Primary Care and Public Health Sciences, King's College London, UK

Contact details for corresponding author:

Helen Dakin
Health Economics Research Centre
Nuffield Department of Population Health
Old Road Campus
Headington, Oxford OX3 7LF
Email: helen.dakin@dph.ox.ac.uk
Tel: 01865 289422
Fax: 01865 289271

Key words

Health technology assessment; implicit weights; cost-effectiveness, National Institute for Health and Care Excellence (NICE); logistic regression.

Funding: Funding for this study was provided by a research grant provided by a consortium of British Pharmaceutical companies: Amgen, AstraZeneca, BMS, Celgene UK & Ireland, GSK, Lilly UK, Merck, Merck Serono, Pfizer, Sanofi Aventis, Schering Plough, Takeda UK, and Wyeth.

Conflict-of-interest: None.

ABSTRACT

Background: The National Institute for Health and Care Excellence (NICE) emphasises that cost-effectiveness is not the only consideration in health technology appraisal and is increasingly explicit about other factors considered relevant. Observing NICE decisions and the evidence considered in each appraisal allows us to 'reveal' its implicit weights.

Objectives: This study aims to investigate the influence of cost-effectiveness and other factors on NICE decisions and to investigate whether NICE's decision-making has changed through time.

Methods: We build on and extend the modelling approaches in Devlin and Parkin (2004) and Dakin *et al* (2006). We model NICE's decisions as binary choices: i.e. recommendations *for* or *against* use of a healthcare technology in a specific patient group. Independent variables comprised: the clinical and economic evidence regarding that technology; the characteristics of the patients, disease or treatment; and contextual factors affecting the conduct of health technology appraisal. Data on all NICE decisions published by December 2011 were obtained from HTAinSite [www.htainsite.com].

Results: Cost-effectiveness alone correctly predicted 82% of decisions; few other variables were significant and alternative model specifications led to very small variations in model performance. The odds of a positive NICE recommendation differed significantly between musculoskeletal disease, respiratory disease, cancer and other conditions. The accuracy with which the model predicted NICE recommendations was slightly improved by allowing for end of life criteria, uncertainty, publication date, clinical evidence, only treatment, paediatric population, patient group evidence, appraisal process, orphan status, innovation and use of probabilistic sensitivity analysis, although these variables were not statistically significant. Although there was a non-significant trend towards more recent decisions having a higher chance of a positive recommendation, there is currently no evidence that the threshold has changed over time. The model with highest prediction accuracy suggested that a technology costing £40,000 per quality-adjusted life-year (QALY) would have a 50% chance of NICE rejection (75% at £52,000/QALY; 25% at £27,000/QALY).

Discussion: Past NICE decisions appear to have been based on a higher threshold than the £20,000-£30,000/QALY range that is explicitly stated. However, this finding may reflect consideration of other factors that drive a small number of NICE decisions or cannot be easily quantified.

1. INTRODUCTION

The criteria by which health technology assessment (HTA) agencies make their decisions are of importance to healthcare providers and to patients whose eligibility for healthcare services is established by its recommendations. They may also influence technology firms' investment and production decisions regarding current and potential products. However, although the centralised authorities in 23 European countries generally state their criteria, "there remains a lack of transparency around critical elements, such as how multiple factors or criteria are weighed during committee deliberations" (Stafinski *et al* 2011).

In England and Wales, the National Institute of Health and Care Excellence (NICE) is responsible for providing guidance on which types of healthcare are to be made available by the National Health Service (NHS). The decisions by its appraisal committees have been dominated by new pharmaceutical products, but its remit is much wider, also appraising medical devices and now also public health and social care. Although NICE's overall remit and aims are clearly defined by the legislation that established it, NICE has been allowed to develop its methods and processes over time and they have become increasingly stable and clear. However, with respect to their decision-making criteria, several areas of considerable uncertainty remain.

Rawlins and Culyer (2004) state that NICE's main criterion for decision-making is cost-effectiveness and that the usual measure of cost-effectiveness to be used is the incremental cost-effectiveness ratio (ICER) expressed as the cost per quality-adjusted life-year (QALY) gained. NICE also states that the 'threshold' ICER that determines whether a technology is considered cost-effective is intended to represent the opportunity cost to a fixed-budget NHS in terms of QALYs forgone if the technology is adopted (NICE 2013, McCabe *et al,* 2008). NICE quantifies this 'shadow price of a QALY', but, rather than characterising a single 'threshold', it describes, in loose qualitative terms, ranges that affect the *probability* that a technology will be recommended. Although different documents give slightly different values to these ranges, the most recent and definitive statement by NICE (2013) is that:

1. "Below a most plausible ICER of £20,000 per QALY gained, the decision to recommend the use of a technology is normally based on the cost-effectiveness

estimate and the acceptability of a technology as an effective use of NHS resources".

2. "Above a most plausible ICER of £20,000 per QALY gained, judgements […] will specifically take account of […] the degree of certainty around the ICER […whether] the assessment of the change in health-related quality of life has been inadequately captured [and] the innovative nature of the technology [...whether] the technology meets the criteria for special consideration as a 'life-extending treatment at the end of life […and whether there are] aspects that relate to non-health objectives of the NHS".
3. "As the ICER of an intervention increases in the £20,000 to £30,000 range, the Committee's judgement […] will make explicit reference to the relevant factors listed [above]".
4. "Above a most plausible ICER of £30,000 per QALY gained, the Committee will need to identify an increasingly stronger case […] with regard to the factors listed.

Discussing probabilistic ranges rather than a single threshold enables NICE to have considerable discretion over its decisions and minimises debates about the legitimacy of its approach and disputes about the precise value that such a 'threshold' should take. However, it results in uncertainty about why particular decisions have been made, which is important for assessing NICE's accountability and for predicting what future decisions might be, which in turn may affect future research and development spending on health technologies, amongst other things. Moreover, as noted, NICE considers other decision-making criteria as well as cost-effectiveness. In addition to those detailed above, these include:

1. Severity of underlying illness: more generous consideration is given to the acceptability of an ICER in serious conditions, reflecting society's priorities (Rawlins *et al*, 2010);
2. Stakeholder persuasion: Insights provided by stakeholders (e.g. on the adequacy of measures used in trials to reflect symptoms and quality of life; Rawlins *et al*, 2010; NICE 2008);
4. End of life treatments: the public places special value on treatments that prolong life at the end of life, providing that life is of reasonable quality (Rawlins *et al*, 2010; NICE 2009);

5. Disadvantaged populations: special priority is given to improving the health of the most disadvantaged members of the population (Rawlins et al, 2010; NICE 2008);
6. Children: given methodological challenges in assessing quality of life in children, society would prefer to give ‘the benefit of the doubt’ (Rawlins et al 2010).

The weights attached to these additional criteria are rarely quantified and their importance and impact are therefore even more uncertain than cost-effectiveness. A possible exception is the ‘end of life’ criterion, which is explicitly related to a higher weighting for QALYs at the end of life, although the actual weights to be used are not specified. Culyer (2009) notes:

> “I do not think NICE is very good at weighing qualitative factors explicitly [...] nor is it very good at explaining recommendations of technologies with ICERs above the £20k threshold [...] There is quite a bit of confusion outside NICE (and possibly within it) about the meaning of the threshold range of £20-30k”.

Further, Appleby *et al* (2009) comment that NICE’s statements

> “..mix a precise quantified criterion of a cost per QALY gained range with an imprecise qualitative description of other factors affecting NICE decisions […] the way in which those other factors are combined with the [cost per QALY gained] range in decision-making is unclear”.

This paper aims to assess the impact of NICE’s criteria on its decision-making in practice and thereby estimate the weights implicit within NICE technology appraisal (TA) decisions; these weights could be viewed as analogous to multi-criteria decision analysis (MCDA) weights that are implied by the deliberative process rather than being specified a priori to drive decisions. This could inform NHS patients, health technology industries and the public in England and Wales, and prompt discussion about whether the implicit weights reflect society’s preferences. We investigate empirically the effect of cost-effectiveness evidence and other factors on the likelihood of NICE recommending a technology, using a revealed preference approach to model NICE decision-making. This work builds on and extends earlier studies by Devlin and Parkin (2004) and Dakin *et al* (2006); in particular, the much larger number of decisions now available facilitates exploration of the additional research questions detailed in Section 2.

2. MODELLING NICE DECISION-MAKING

The economic theory underlying this study comprises the *implied values* approach, whereby decisions are assumed to be based on an objective function that can be analysed by examining the outcomes and parameters of the function, to yield the implied relative weights given to those parameters. More specifically, it uses the *revealed preference* method, whereby real-world decisions are examined to estimate the influence of different factors. Within mainstream economics, revealed preference has predominantly been used to analyse the prices paid for similar, but differentiated, goods with respect to the differing levels with which they possess the key characteristics of that class of goods. The weights that are implicitly attached to these characteristics are known as *hedonic prices* (Rosen, 1974) and the principal method by which they are estimated is known as *hedonic regression*. Our approach draws on that theory, but has a somewhat different focus. We are concerned with factors affecting decisions, not prices; our decision-makers are not individual consumers within markets, but agents for public bodies. Furthermore, we assume that the underlying rationale for decisions is not necessarily the maximisation of consumer's utility (although it might be), but an objective function whose maximand is not clearly defined.

Our aim is to explore the role of various decision criteria in decision-making by a public body, comprising multiple decision-making committees, each comprising multiple individuals who are tasked to weigh these up via a ‘deliberative process’ (Culyer 2006, 2009; NICE 2013). Such processes typically involve both scientific evidence (both context-free and context-sensitive) and ‘colloquial’ evidence, which is any other evidence that people use in their decision-making. The appraisal committee’s role is not to ensure that an explicit decision-making formula is correctly applied, but to exercise judgment over whatever evidence is available, including that shared with the committee by its members. With respect to the criteria for decision-making, these are equally non-explicit, involving NICE in a ‘search’ for them based on expert opinion, research and accumulated experience (Culyer *et al*, 2007). This lack of clarity and deliberate non-explicitness makes it all the more important that implicit criteria and value-judgements affecting the spending of public resources are exposed for public scrutiny.

This analytical framework has been used in previous studies. Devlin and Parkin (2004) found that cost-effectiveness was the key driver of NICE decisions; while uncertainty and burden of disease were also significant. Dakin *et al* (2006), characterising NICE decisions as yes, no, or 'restricted', found significant influences on decisions from cost-effectiveness, clinical evidence, technology type, and patient group submissions.

Similar approaches have been used to analyse decisions by other HTA bodies. Linley and Hughes (2012), Mshelia (2013) and Harris *et al* (2008) analysed decisions taken on the use of new medicines by the All Wales Medicines Strategy Group, the Scottish Medicines Consortium and the Australian Pharmaceutical Benefits Advisory Committee (PBAC), respectively. Each found that other criteria significantly affected decisions, in addition to cost-effectiveness. In contrast, Tappenden *et al*, (2007) used a stated preference approach to explore the importance of various decision criteria to individual members of NICE committees; significant variables included the ICER, uncertainty, availability of other therapies, and severity of illness.

The models we use to analyse NICE decision-making predict a binary dependent variable representing whether a technology was recommended. NICE recommendations for whole TAs are commonly characterised as a three-way choice between: 'yes' to all patients and technologies considered within the scope; 'no' to all patients and 'restricted' or 'optimised' (NICE, 2010), which means 'yes' to some patient sub-groups, and no to others. Dakin *et al* (2006) categorised decisions in this manner, although there are important limitations to this approach. First, the clinical evidence, ICERs and other considerations used to inform the decision may be specific to the patient sub-groups for whom the technology is recommended or rejected. 'Restricted' decisions also vary considerably in terms of their implications for patient access to new technologies (O'Neill and Devlin 2010). We therefore sub-divided restricted appraisals into their component 'yes' and 'no' decisions concerning use of a single technology, for a clearly-defined group of patients, to enable us to more precisely link the decision to the corresponding evidence considered by NICE.

We considered several alternative ways to characterise NICE's decision-making process and the way different sets of considerations might affect the final decision outcome: as a sequential, process (Figure 1) and as a single production function (Figure 2).

The sequential model suggests that, for each decision, NICE first considers effectiveness evidence and then, for those technologies that are effective, goes on to consider cost-effectiveness. Other criteria (e.g. 'social values') act as a modifier that may cause NICE to recommend a technology that would otherwise be deemed cost-ineffective by the usual standards. This model is arguably broadly in keeping with NICE's description of the way it takes other criteria into account alongside cost-effectiveness (NICE 2008, 2009, 2013). Qualitative evidence suggests that some committee members adopt a two-step approach to decision-making, with clinical effectiveness considered first, and cost-effectiveness taken into account only if the technology passes the first hurdle (Williams *et al*, 2007).

However, whilst this model is plausible *a priori*, in practice it presents some challenges. We cannot observe the decisions at any point other than the final decision outcome, which prevents us from modelling the third step directly; and the empirical evidence is that a technology is ineffective is likely to perfectly predict rejection by NICE.

An alternative model comprises a simple 'production function' approach (Figure 2). NICE seeks and combines decision inputs, in terms of clinical and economic evidence. The inputs enter a production process, entailing the synthesis and evaluation of that evidence, using NICE's decision-making procedures. Such procedures are influenced by: the composition and organisation of appraisal committees; methods guides that shape the selection of evidence inputs; available (imperfect) information about the opportunity costs in the NHS; and available (imperfect) information on social preferences with respect to the prioritisation of particular patient, disease or treatment characteristics, e.g. as suggested by NICE's Citizen's Council.

All evidence passes through this decision-making process, and the decision output is an observable 'yes' or 'no' in each case. This model suggests a single regression model, where all influences on decisions, including both evidence and decision-making processes, are independent variables. Due to the challenges raised by the sequential model, the model in Figure 2 formed the basis for the econometric modelling reported in this paper.

Specifically, our study aimed to address the following research questions:

1. Does the probability of rejection increase with increasing ICER?
2. Is there empirical support for the sigmoid curve proposed by (Rawlins and Culyer, 2004) showing the increase in risk of rejection with increasing ICER, and the 'inflexion points' at £5,000-£15,000 and £25,000-£35,000/QALY gained?
3. What impact do the other factors identified by NICE have on the probability of NICE rejection? Does NICE take account of factors that they state do not merit special consideration (e.g. orphan drugs for rare conditions, Littlejohns and Rawlins, 2009)?
4. Have NICE's decisions and/or threshold changed over time? For example, NICE statements about its cost-effectiveness threshold have evolved in subtle yet important ways over time: from an initial 'unwritten rule' of £30,000, to the threshold as lying in the £20,000-£30,000 region (NICE, 2005), to an increasing tendency to refer to the threshold as £20,000, with exceptions made above (NICE, 2013). Furthermore, key aspects of NICE TA processes and methods have changed during this period, including: dropping the differential discount rate, thereby increasing the discount rate for costs and lowering that for QALYs (NICE, 2004); introducing the single technology appraisal (STA) process in 2005, in an endeavour to speed up decision-making; and, most recently, introducing an explicit process for weighting QALYs gained at the end of life (NICE, 2009, 2013).

3. DATA

The data for this study were obtained from HTAinSite© (www.htainsite.com) and initially comprised all 240 NICE TAs published by 31st December 2011, with the exception of 11 appraisals that were terminated before any decision was made (Figure 3). The conceptual models outlined in the previous section were used to select a core set of variables for the first regression model from the fields available in HTAinSite (Model 1, Table I). In addition to the ICER, we captured one variable indicating the amount of clinical evidence (Total_pts_in_RCTs) since previous work showed this to be important (Dakin *et al,* 2006) and a variable capturing any temporal trends (Date). We also included one measure of stakeholder involvement (Pt_group_sub), whether the intervention was the only treatment for this population, whether the decision concerned children and a crude measure of disease severity. End of life considerations were not included in Model 1 as such data are only available since 2009. Uncertainty around the ICER and innovation were not included in Model 1 due to difficulties defining variables that consistently capture these issues.

Each of the 229 non-terminated TAs was sub-divided into 1-19 component *decisions*, each representing a NICE decision to either recommend or reject a single technology in a specific patient population. Sub-division of each TA inevitably requires a degree of researcher judgement; our dataset follows that of HTAinSite, which uses a carefully-documented protocol providing a set of principles for making those judgements in a consistent manner. Using this protocol, data were extracted by ≥2 analysts, and differences were referred to an advisory panel to resolve.

However, HTAinSite did not provide all the data required for modelling. A key issue was identification of the 'main' ICER associated with each decision. HTAinSite records *all* ICERs mentioned in the TA documentation. For our analysis, however, stronger value judgements were required to identify the 'main' ICER(s) that drove NICE decisions. We developed a set of principles to guide our selection of the relevant ICERs (Appendix).

4. EMPIRICAL METHODS

We modelled NICE decisions using logistic regression, which assesses the effect of explanatory variables on the log-odds of success, in this case NICE saying 'yes'. Standard errors were adjusted for within-appraisal clustering of decisions, since decisions concerning different drugs or patient populations within the same appraisal are made by the same committee on the same day and are often based on similar or related evidence, so are unlikely to be independent. All statistical analyses were conducted in Stata Version 12 (StataCorp 2011).

For 45% (229/510) of decisions with usable ICERs, we identified ≥2 ICERs that informed NICE's decision-making. For example, some gave separate ICERs for several patient subgroups considered in the same decision or gave equal prominence to two different analyses. Thirty-one decisions gave an ICER range (e.g. stating that the ICER was between X and Y), while others simply said that the ICER was "above A" or "below B". Taking the mean, median or midpoint of the reported ICERs would have made assumptions about how NICE used this information in their decision-making. It would also have prevented us from including decisions with ICERs "above A" and would overestimate the precision of our regression results by ignoring the uncertainty around the ICER. Instead, we used a simulation approach to sample repeatedly from the list of ICERs identified for each decision. For the 198 decisions with 2-40 relevant ICERs, the ICER used in each of 100 iterations[1] was randomly sampled by assigning equal probability to all ICERs. For the 31 cases giving a range or lower/upper limit, ICER values were sampled from a list of all ICERs within our dataset that lay in the relevant range, since ICERs follow an unknown distribution and may approach infinity (Briggs and Fenn 1998). For example, for those decisions for which the Guidance indicated the ICER was "above £30,000 per QALY", we created a list of all ICERs reported in other NICE decisions that were >£30,000/QALY and sampled at random from this list, assigning equal probability to each ICER. For the 281 decisions with one relevant ICER, this single ICER value was used in all 100 datasets. These sampling procedures generated 100 datasets, each with different ICER data for those decisions with >1 relevant ICER.

Regression models were run separately on all 100 datasets and results were combined by implementing Rubin's rule (Carlin *et al* 2008), which averages parameter estimates (e.g. regression coefficients) across multiple imputed datasets and adjusts standard errors to allow for uncertainty around the different ICER values.

The primary measure of model performance comprised the proportion of decisions that were correctly classified, since it is not valid to apply Rubin's rule to measures of model fit or likelihood, such as pseudo-R^2 and Akaike's information criterion (AIC; White, et al 2011). Ideally, the proportion of correctly-predicted outcomes would be based on a validation

[1] ICERs were sampled 100 times to generate 100 datasets to generate robust results capturing the full range ICERs for each decision.

sample independent of the data used to estimate the model (Copas, 1983). Unfortunately, this was not feasible due to the limited number of appraisals available; we therefore rely on a single dataset to both estimate and assess model performance, which may result in overly optimistic results.

The proportion of NICE decisions correctly predicted, together with the specificity (the proportion of rejected decisions predicted as rejected) and sensitivity (the proportion of recommended decisions predicted as recommended), were calculated by assuming that all decisions with ≥50% predicted probability of success would be recommended by NICE. Pseudo-R^2 and AIC calculated from the mean log-likelihood for the best models (averaged across all datasets) are also shown for illustration, although these figures should be interpreted with caution.

Our analyses were primarily exploratory and aimed to identify which factors are most influential and the best way to input each factor. We therefore explored a wide range of model specifications in a series of four stages. In stages B and C, prediction accuracy was compared between models and the model with the highest proportion of decisions correctly classified was taken forward to the next stage.

A) Evaluation of Model 1, which included only the seven variables that we predicted to have most effect on NICE decisions (Table I). This model was compared against Model 5, which included only the ICER.

B) Identification of variables explaining NICE decision-making. We added additional independent variables (Table A1) into Model 1 to assess whether they improved prediction accuracy and/or had a significant effect on NICE decisions and removed variables from Model 1 one at a time to identify which explained NICE decisions. All the variables that improved prediction accuracy when considered individually were then evaluated simultaneously in Model 2. Those variables that were statistically significant in at least one analysis were included in Model 3.

C) Alternative specifications: We then varied the specification of the variables in Model 2 to evaluate the effect that this has on the proportion of decisions that are correctly classified and the statistical significance of this parameterisation (see Appendix). The specification for each variable that had highest prediction accuracy when considered individually was included in Model 4.

D) Sensitivity and subgroup analyses: Conducted on Model 4 (see Appendix).

Methods similar to those described by Devlin and Parkin (2004) were used to estimate the ICER at which there is a 25%, 50% or 75% chance of a positive NICE recommendation. The predicted log-odds of NICE saying 'yes' was calculated for different ICER values by multiplying the vector of estimated coefficients by the vector of mean values for other explanatory variables and the ICER value of interest. Similar figures were estimated for particular types of decisions (e.g. those on cancer) by repeating calculations using values of zero and one for that dummy variable in place of its mean.

Regression analyses included only decisions concerning treatments that are more costly and more effective than their comparator. Decisions for which all relevant ICERs indicated that the technology was either dominated or dominant relative to its comparator were excluded from regression analyses since dominance perfectly predicted NICE recommendations. ICERs in the south-west quadrant of the cost-effectiveness plane (which indicate that treatment is less costly and less effective than its comparator) have the opposite interpretation to those in the north-east quadrant (which indicate that treatment is more effective and more costly) and the two types of ICER data cannot easily be combined without making value judgements about NICE's preferences; we therefore also excluded six decisions for which all ICERs lay in the south-west quadrant. Twenty-two decisions had ICERs in >1 quadrant; these decisions were included in regression analyses in those datasets where a north-east quadrant ICER was sampled and were dropped from regressions in datasets where an ICER from another quadrant was sampled. As result, the number of decisions included in each regression varied between 424 and 432.

5. RESULTS

Our dataset comprised 763 decisions from 229 appraisals (Figure 3). Of these, 253 decisions did not report any usable ICERs and were therefore omitted from regression analyses:

a) 70 decisions were rejected due to lack of clinical evidence; these decisions had significantly fewer patients in RCTs (p<0.001) than other decisions, although 59% (41/70; Table II) were nonetheless supported by one or more RCT.

b) 63 decisions were recommended on clinical grounds (e.g. because all alternative technologies were contraindicated or not tolerated), while 28 decisions were rejected on clinical grounds (e.g. because treatment was "clinically inappropriate" in that patient group). The decisions made on clinical grounds were, on average, published two years earlier than the average decision based on cost-effectiveness (p<0.001), had less RCT evidence (p=0.006) and were more likely to be for children (p<0.001), although the characteristics were otherwise similar (Table II).

c) 174 decisions that appear to have been based on cost-effectiveness did not have available north-east quadrant ICERs. For 39 of these decisions, cost-utility analysis was not undertaken, although another form of economic evaluation was done (e.g. cost-effectiveness analysis calculating the cost per life-year gained). A further 36 decisions made broad references to the committee's judgements about cost-effectiveness but no specific ICERs were quoted or identified; this included statements that the ICER "approaches infinity" or was "likely to be cost-effective". Seventeen decisions were based on cost/QALY ICERs that were not available for analysis (e.g. because they were commercial in confidence, or the guidance document was unavailable). Thirty-three decisions were rejected as treatment was dominated by its comparator, while 31 were recommended as treatment dominated. Six decisions had ICERs in the south-west quadrant, of which one was rejected. The decisions based on cost-effectiveness that lacked available north-east quadrant cost/QALY tended to be published about four years earlier than those included in regression analyses (p<0.001) and were less likely to be STAs (p<0.001) or only treatments (p<0.001).

Among the 510 decisions with available north-east quadrant ICERs, ICERs differed significantly between recommended and rejected decisions (p<0.001; Table III). Exploratory data analysis also demonstrated that the proportion of decisions rejected by NICE increases substantially with ICER (particularly at ~£27,500 and ~£47,500/QALY), although there are numerous exceptions (Figure 4).

5.1. Factors affecting NICE decisions

Model 1 evaluated the impact of the seven variables considered most likely to influence NICE decision-making (Tables III and IV). This model fitted the data well (mean adjusted pseudo-R^2=0.34) and correctly classified 82.5% of NICE decisions (Table II). As expected, the ICER had a significant effect on NICE decisions, with every £1,000 increase in the ICER reducing the odds of NICE recommending the technology by 6.9% (95% CI: 4.3%, 9.4%; p<0.001; Table IV).

However, clinical evidence, having no alternative treatments, paediatric population, patient group submission, disease severity and date had no significant effect on NICE decisions (p≥0.29; Table IV). As hypothesised, there were trends suggesting that decisions concerning children and those with no alternative treatments have a higher chance of being recommended by NICE (Table IV). However, the impact of additional clinical evidence was negligible and treatments for more severe diseases and those supported by patient group submissions had a non-significantly lower chance of being recommended (p=0.53), contrary to our hypothesis. Nonetheless, omitting any variable from the model other than disease severity slightly reduced prediction accuracy, suggesting that these variables may help explain some NICE decisions.

Prediction accuracy was slightly improved by taking account of 12 of the 17 additional variables evaluated in Stage B (Table A1): the appraisal process (STA vs. multiple technology appraisal, MTA); whether the analysis included probabilistic sensitivity analysis (PSA); orphan status; the number of systematic reviews and non-randomised studies considered; the range of ICERs; and certain diseases. Model 2 (Table IV) therefore included these variables, in addition to all variables from Model 1 other than severity (which was omitted to improve prediction accuracy). Model 2 correctly classified 84.67% of NICE decisions which represents a small improvement on Model 1 (Table II).

Model 2 suggested that interventions classed as innovative (p=0.29), those with more systematic reviews (p=0.67) or non-randomised studies (p=0.07) and those with a smaller range of ICERs (p=0.07) were non-significantly more likely to be recommended (Table III). However, contrary to our expectations, decisions with PSA (p=0.13) and those on orphan drugs (p=0.46) were non-significantly less likely to be recommended. Appraisals conducted through the STA process were also found to be 51% (95% CI: -9%, 78%) more likely to be rejected by NICE than MTAs, although this result was not significant (p=0.083).

There were also marked differences in the probability of NICE rejection between diseases. The odds of a positive NICE recommendation were 5.7-fold higher (p=0.007; 95% CI: 1.6, 20.3) for musculoskeletal disease interventions, 3.1-fold higher (p=0.029; 95% CI: 1.1, 8.4) for decisions concerning treatment, prevention or diagnosis of cancer and 71% lower (p=0.037, 95% CI: 7%, 91%) for interventions for respiratory disease. Model 4 gave similar findings.

These findings were largely confirmed by Model 3, which included only statistically significant variables (ICER, musculoskeletal disease, cancer and respiratory disease), although omitting the non-significant variables reduced prediction accuracy to 83.5% and reduced the magnitude of the coefficients for each of the three diseases, such that cancer and musculoskeletal disease had no statistically significant effect at the 5% level (p≥0.103).
The impact of end of life criteria was evaluated in a subset of appraisals published after these criteria were introduced in January 2009 (NICE, 2009). This suggested that decisions meeting the end of life criteria were 3.4-fold more likely (p=0.15, 95% CI: 0.64, 17.9) to be recommended by NICE than those that did not meet the criteria. Within this group of decisions, taking account of end of life criteria improved prediction accuracy from 84.23% with Model 1 to 85.12%. A sensitivity analysis found that allowing for the identity of the committee making NICE recommendations slightly improved prediction accuracy, although there were no statistically significant differences between committees.

However, overall the impact of additional variables on prediction accuracy was very small, with no variable increasing prediction accuracy by more than one percentage point. Indeed, omitting all variables except the ICER correctly classified 82% of NICE decisions (Table III, Model 5). By contrast, omitting the ICER from Model 2 suggests that the other variables in isolation would correctly classify only 73.1% of NICE decisions.

5.2. Relationship between ICER and probability of NICE recommendation

Coefficients from the five models were used to estimate how the probability of NICE rejection varied with ICER, holding all other parameters at mean values (Figure 5, Table III). Model 1 suggested that a treatment with an ICER of £43,356 would have a 50% chance of a positive NICE recommendation, holding all other parameters at mean values. This model also predicted that NICE would recommend 25% of products with an ICER of £62,253 and 75% with an ICER of £27,935. The ICER at which the average product had a 50% chance of rejection decreased as additional variables were taken into account, from £43,949 for Model 5 (which considered only the ICER) to £39,417 for Model 4 (Table III, Figure 5). The interaction between ICER and patient group submission also increased the gradient for Model 4, such that the probability of NICE rejection increases over a narrow range of ICERs.

However, although the choice of model had relatively little effect on the relationship between ICER and recommendation when other variables were held at their mean value, varying the value of other variables often produced substantial shifts in the curve. For example, for Model 4, the ICER at which the probability of NICE saying 'yes' was 50% was £20,356/QALY for respiratory disease, £37,950 for cardiovascular disease, £46,082 for cancer, £49,292 for infectious disease, £55,512 for musculoskeletal disease, and £32,263 for other diseases. For any given ICER point estimate, having uncertainty around the ICER such that the ICER could plausibly be £10,000 higher or lower than the point estimate decreased the 50% point to £43,516/QALY, compared with £48,014 for decisions with only one plausible ICER.

The decisions that were poorly predicted by our models were generally rejected due to substantial uncertainty, or included statements within the guidance suggesting that the committee believed the ICER to be at the top or bottom of the stated range (see Appendix). This is supported by a sensitivity analysis using the minimum ICER for all recommended decisions and the minimum ICER for all rejected decisions correctly classified 93.0% of decisions, which may suggest that around two-thirds of the decisions poorly classified by our model may be due to difficulties identifying the ICER that drove the committee's decision based on secondary data.

5.3. Has NICE's threshold changed over time?

Model 1 suggested that publication date had no significant effect on NICE decisions (p=0.31) and estimated that the odds of a positive NICE recommendation increased by 6% (95% CI: -5%, 19%) per year between 2000 and 2011. Similarly, although inflation will also affect the real value of any ceiling ratio, inflating ICERs to 2011/12 values using the HCHS pay and prices index (Curtis, 2012) reduced prediction accuracy. We examined alternative specifications of publication time to assess the impact on prediction accuracy (Appendix), although no statistically significant temporal trends were observed.

6. DISCUSSION

Implications for understanding how NICE weighs up benefits and costs

Our analyses demonstrate that cost-effectiveness is the principal determinant of most NICE decisions and that the probability of rejection increases significantly with increasing ICER. The finding was robust to extensive sensitivity analyses and modelling approaches.

The relationship between ICER and the probability of NICE rejection appears to follow a sigmoid curve with points of inflexion. However, the data do not appear to support the £5,000-£15,000/QALY and £25,000-£35,000/QALY inflexion points proposed by Rawlins and Culyer (2004). Neither do our results support NICE's stated threshold range. Based on NICE statements, we would expect that: for ICERs under £20,000/QALY, a recommendation would be odds-on; above £30,000/QALY it would be odds-against; and that the odds switch from on to against somewhere in between. We estimate that in practice the ICER at which the probability switches from more-likely-to-accept to more-likely-to-reject is between £39,000 and £44,000: well above the stated £20,000-£30,000 range.

It is informative to compare our estimates with emerging evidence on what the cost-effectiveness threshold *should* be. Although NICE formally subscribes to an opportunity cost definition of the threshold and has advocated research into that (Claxton *et al*, 2013; Appleby et al 2009), it has also advocated research into the social value of a QALY (Baker *et al*, 2010). Our results clearly show that, in practice, NICE often recommends technologies with ICERs that are well above the opportunity cost estimated by Claxton et al (2013), but somewhat closer to the social value of a QALY (Baker *et al*, 2010).

Temporal trends and impact of other factors

Although allowing for temporal trends improved model performance, time had no significant effect on NICE decisions and the relationship we estimate between cost-effectiveness and NICE decisions between 1999 and 2011 is remarkably similar to that reported by Devlin and Parkin (2004) for the years 1999-2002, despite the many changes in NHS budgets, prices and productivity in the intervening seven years. Although the models reported here treat ICERs in nominal terms, inflation must have affected the prices and costs embodied in the ICERs in the appraisals conducted over this 10-year period; yet inflation-adjusting ICERs reduced model performance.

The single factor other than cost-effectiveness that emerged from our analyses as exerting a significant effect on decisions is the type of disease that the technology is intended to prevent, diagnose or treat. NICE rejections were significantly less likely for cancer and musculoskeletal disease, but significantly more likely for respiratory disease. It is unclear whether such trends reflect a causative relationship between disease and NICE decisions (e.g. driven by political priorities, the shadow price of a QALY and/or willingness to pay), or whether it reflects selection of topics or other characteristics of the decisions within each disease area. The finding for cancer was clearest before the End of Life Guidance was introduced, with NICE recommending 75% (49/65) of cancer decisions before January 2009, vs. 46% (24/52) after; however, the end of life guidance may have simply formalised something that NICE was already taking into account.

Other than certain diseases, no variables other than cost-effectiveness significantly predicted NICE decisions. However, the relevance of statistical significance is unclear when the sample includes the whole 'population' of NICE decisions published before 2012. Furthermore, our descriptive analysis suggests that 21% (161/763) of decisions are based on clinical considerations and lack of clinical evidence, without considering cost-effectiveness. It is also possible that NICE took account of other factors that cannot easily be defined or quantified, were not explicitly noted in the Guidance, or were one-off considerations specific to particular decisions. The influence of additional factors not detected in our analysis would have biased upwards our estimate of the ICER at which the probability of rejection is 50%. Furthermore, several factors that NICE says influence its decisions are difficult empirically to define and measure. For example, although severity is said to influence NICE decisions (Rawlins et al, 2009), NICE Guidance does not state whether the condition was considered to be 'severe' and in the absence of a precise definition of 'severity', it is difficult for researchers to judge *ex post* which technologies would be deemed to fall into that category; the measure we used (mean DALY weight across ICD chapters) may not adequately represent the way NICE committees consider severity. 'Innovation' presents a similar challenge, as do other criteria (e.g. disadvantaged populations) that we were not able to explore. NICE's appraisal process is intended to reflect and incorporate multiple criteria, but the effect on decisions of criteria other than cost-effectiveness is not readily detectable; it could therefore be argued that NICE should be more transparent about the criteria being used and the importance attached to these (Devlin and Sussex 2010). However, others would argue that a deliberative process without pre-defined weights is needed to consider the evidence and make complex decisions (Culyer and Lomas, 2006).

Budget impact, population size and media noise might arguably be relevant to understanding and explaining NICE decisions but were not included in our analysis. One argument for excluding budget impact is that NICE is not meant to take that into account. We would have liked to test this hypothesis rather than assuming that it has no impact. However, budget impact estimates are only recorded for whole TAs based on the patient subgroups for which treatment was recommended; estimating the net budget impact for each sub-decision would be a substantial task, beyond the scope of this project.

Although we have explored measures of clinical evidence and uncertainty, this was not entirely satisfactory and remains to be properly captured both conceptually and empirically. Devlin and Parkin (2004) expressed similar reservations regarding the variable they intended to capture the range of the ICERs. We considered, but rejected, the possibility of using confidence intervals or cost-effectiveness acceptability curves estimated using PSA. Although PSA is now more common, modelling this variable would require us to exclude all decisions where PSA was not undertaken. Furthermore, using the probability that treatment is cost-effective at a given ceiling ratio would require value judgements regarding the appropriate ceiling ratio.

Conclusions

Our analysis uses a larger number of decisions than any past analysis of HTA decisions and explores the impact of a wide range of potential predictors. We find that cost-effectiveness is the major driver of NICE decisions and correctly predicts 82% of decisions. No other factors besides the type of condition had a significant effect on NICE decisions, although allowing for clinical evidence, alternative treatments, paediatric population, patient group involvement, publication date, type of process (STA versus MTA), orphan status, innovation and uncertainty improved prediction accuracy somewhat. Our results show that NICE frequently recommends technologies with ICERs considerably higher than its stated £20,000-£30,000/QALY threshold range. However, the analysis relied upon judgements about which ICER(s) were taken into account in each NICE decision and our conclusions are based on the assumption that we have identified the "correct" model. Further work is required to explore the impact of uncertainty, severity, innovation and equity on NICE decisions and to explore the structure of NICE decision-making using sequential models.

ACKNOWLEDGEMENTS

We would like to thank Tony Culyer for his comments on the draft of this manuscript. We are grateful to HTAinSite© for allowing access to the data for this study. In particular, we wish to thank Pam Poppe, Iain Shield, Carmel Guarnieri and Zoe Philips, of Abacus International, for preparing the dataset for our analysis. Funding for this study was provided by a research grant provided by a consortium of British Pharmaceutical companies: Amgen, AstraZeneca, BMS, Celgene UK & Ireland, GSK, Lilly UK, Merck, Merck Serono, Pfizer, Sanofi Aventis, Schering Plough, Takeda UK, and Wyeth. Views expressed in this paper are those of the authors and do not necessarily represent the views of the sponsors of this research.

REFERENCES

Appleby J, Devlin N, Parkin D, Buxton M, Chalkidou K. 2009. Searching for cost-effectiveness thresholds in the NHS. *Soc Sci Med* **91**: 239-245.

Baker R, Bateman I, Donaldson C, Jones-Lee M, Lancsar E, Loomes G, Mason H, Odejar M, Pinto Prades JL, Robinson A, Ryan M, Shackley P, Smith R, Sugden R, Wildman J; SVQ Research Team. 2010. Weighting and valuing quality adjusted life years: preliminary results from the social value of a QALY project. *Health Technology Assessment* **14**:1-162.

Briggs A, Fenn P. 1998. Confidence intervals or surfaces? Uncertainty on the cost-effectiveness plane. *Health Economics* **7**:723-40.

Carlin JB, Galati JC, Royston P. 2008. A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* **8**:49-67.

Claxton K, Martin S, Soares M, Rice, N, Spackman E, Hinde S, Devlin N, Smith PC, Sculpher M. 2013. Methods for the Estimation of the NICE Cost Effectiveness Threshold. CHE Research Paper 81 http://www.york.ac.uk/media/che/documents/papers/researchpapers/CHERP81_methods_estimation_NICE_costeffectiveness_threshold_revised.pdf Accessed 8th November 2013.

Copas JB. 1983. Regression, Prediction and Shrinkage. *J Royal Statist Soc* **B 45** : 311-354.

Curtis L. 2012. *Unit Costs of Health and Social Care 2012*. Canterbury, UK: PSSRU Personal Social Services Research Unit. Available at: www.pssru.ac.uk/project-pages/unit-costs/2012/. Accessed 30th July 2013.

Culyer AJ. 2006 NICE's use of cost-effectiveness as an exemplar of a deliberative process, *Health Economics, Policy and Law* **1**: 299-318.

Culyer, AJ, Lomas, J, 2006. Deliberative processes and evidence-informed decision-making in health care – do they work and how might we know? *Evidence and Policy* **2**: 357- 371.

Culyer A, McCabe C, Briggs A, Claxton K, Buxton M, Akehurst R, Sculpher, M, Brazier J. 2007.Searching for a threshold, not setting one: The role of the National Institute for Health and Clinical Excellence, *Journal of Health Services Research and Policy* **12**: 56 - 58.

Culyer A. 2009. *Deliberative processes in decisions about health care technologies: combing different types of evidence, values, algorithms and people.* OHE Briefing no 48. London: Office of Health Economics.

Dakin HA, Devlin NJ, Odeyemi IA. 2006. "Yes", "No" or "Yes, but"? Multinomial modelling of NICE decision-making. *Health Policy* **77**:352-67.

Devlin N,Parkin D. 2004. Does NICE have a cost-effectiveness threshold and what other factors influence its decisions? A binary choice analysis. *Health Economics* **13**:437-52.

Devlin N , Sussex J. 2011. *Incorporating multiple criteria in HTA: methods and processes*. London: Office of Health Economics.

Harris AH, Hill SR, Chin G, Li JJ, Walkom E. 2008. The Role of Value for Money in Public Insurance Coverage Decisions for Drugs in Australia: A Retrospective Analysis 1994-2004. *Med Decis Making* **28**:713–722.

Hirth RA, Chernew ME, Miller E, Fendrick M, Weissert WG.2000.Willingness to pay for a quality-adjusted life year. *Med Decis Making* **20**:332-42.

Littlejohns P, Rawlins M. 2009. Social value judgements: implementing the citizen's council reports. Chapter 12 in: Littlejohns P, Rawlins M (eds) *Patients, the public and priorities in health care.* Radcliffe.

Linley WG,Hughes DA. 2012. Reimbursement decisions of the All Wales Medicines Strategy Group: influence of policy and clinical and economic factors. *Pharmacoeconomics* **30**:779-94.

McCabe M, Claxton K, Culyer AJ.2008. The NICE cost-effectiveness threshold. What it is and what that means. *Pharmacoeconomics* **26**:733-44.

Mason AR, MF.2009. Public funding of new cancer drugs: Is NICE getting nastier? *European Journal of Cancer* **45**:1188-92.

Mason H, Jones-Less A, Donaldson C. 2009.Modelling the monetary value of a QALY: a new approach based on UK data. *Health Economics* **18**: 933-50.

Mshelia I, White R, Mukke S, 2013. An Investigation into the Key Drivers Influencing the Decision Making of the Scottish Medicines Consortium. *Value in Health* **16:** A264

National Institute for Health and Clinical Excellence. 2010. NICE says "yes" to over 80% of treatments. Available at: www.nice.org.uk/newsroom/news/NICEsaysYes.jsp?domedia=1&mid=E99E7EDB-19B9-E0B5-D4FCD1EB8398B6DD. Accessed 13th September 2013.

National Institute for Health and Clinical Excellence. 2009. Appraising life extending, end of life treatments. Supplementary advice to the Appraisal Committees. Available at: www.nice.org.uk/aboutnice/howwework/devnicetech/endoflifetreatments.jsp?domedia=1&mid=88ACDAE5-19B9-E0B5-D422589714A8EC6D. Accessed 13th September 2013.

National Institute for Health and Clinical Excellence. 2008. *Social value judgements: Principles for the development of NICE guidance.* Second Edition. Available at: http://www.nice.org.uk/media/C18/30/SVJ2PUBLICATION2008.pdf. Accessed 13th November 2010.

National Institute for Health and Clinical Excellence. 2013. *Guide to the methods of technology appraisal.* Available at: http://www.nice.org.uk/media/D45/1E/GuideToMethodsTechnologyAppraisal2013.pdf. Accessed 3rd October 2013.

National Institute for Health and Clinical Excellence. 2005. *Social value judgements: Principles for the development of NICE guidance*. Available at: http://www.nice.org.uk.

National Institute for Health and Clinical Excellence. 2004. *Guide to the methods of technology appraisal.* Available at: http://www.nice.org.uk/niceMedia/pdf/TAP_Methods.pdf. Accessed 18th November 2010.

O'Neill P, Devlin N. 2010. An analysis of NICE's restricted (or 'optimised') decisions. *Pharmacoeconomics* **28** : 987-993.

Rawlins M, Culyer A J.2004. National Institute for Clinical Excellence and its value judgements *British Medical Journal* **329**: 224-227.

Rawlins M, Barnett D, Stevens A. 2010. Pharmacoeconomics: NICE's approach to decision making. *British Journal of Clinical Pharmacology* **70**: 346-349.

Rosen S. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition *The Journal of Political Economy* **82**, 34-55.

StataCorp. 2011. Stata: Release 12. Statistical Software (http://www.stata.com). College Station, TX: StataCorp LP.

Stafinski T, Menon D, Davis C, McCabe C. 2011. Role of centralized review processes for making reimbursement decisions on new health technologies in Europe *ClinicoEconomics and Outcomes Research* **3**: 117–186

Tappenden P, Brazier J, Ratcliffe J, Chilcott J. 2007 A stated preference binary choice experiment to explore NICE decision making. *Pharmacoeconomics* **25**:685-93.

White I. R., Royston P, Wood, A.M. 2011. Multiple imputation using chained equations: Issues and guidance for practice. Stat Med **30**: 377-399.

WHO, World Health Organisation. 2004. *Global Burden of Disease 2004 Update: Disability Weights For Diseases and Conditions*. Available at: http://www.who.int/healthinfo/global_burden_disease/GBD2004_DisabilityWeights.pdf. Accessed 26th July 2013.

Williams I, Bryan S, McIver S. 2007. How should cost-effectiveness analysis be used in health technology coverage decisions? Evidence from the National Institute for Health and Clinical Excellence approach. *J Health Serv Res Policy* **12**: 73-9.

Figure 1. A sequential model of NICE decision-making

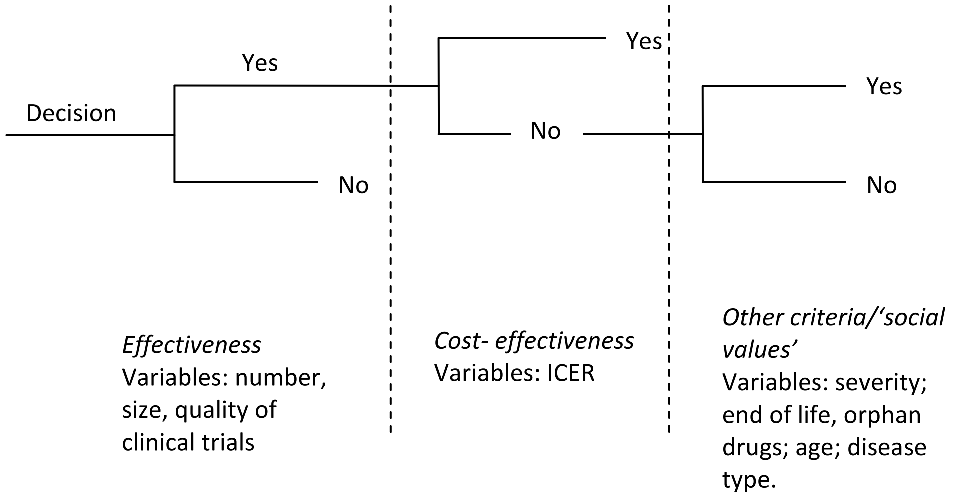

Decision
Yes
No
Yes
No
Yes
No
Effectiveness
Variables: number, size, quality of clinical trials
Cost- effectiveness
Variables: ICER
Other criteria/'social values'
Variables: severity; end of life, orphan drugs; age; disease type.

Figure 2. A production function model of NICE decision-making

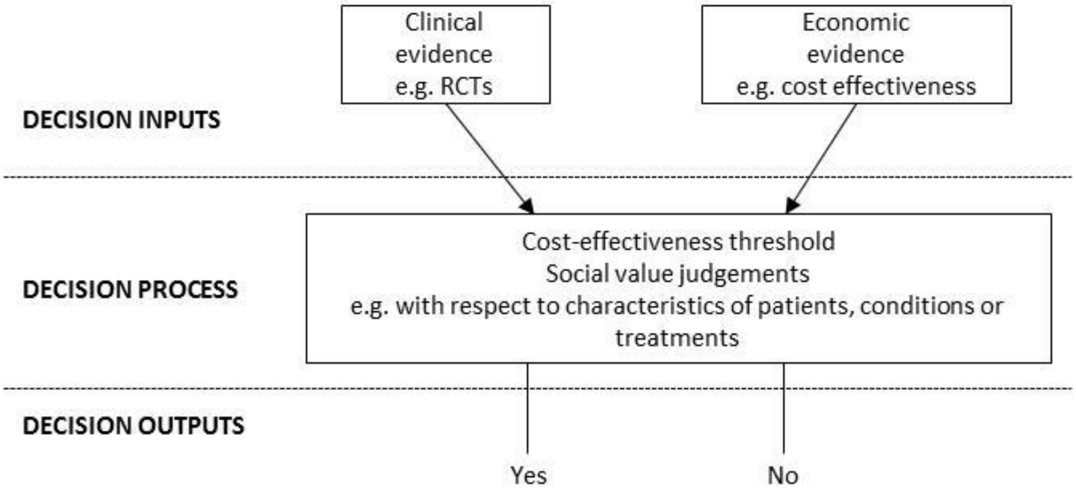

Clinical evidence e.g. RCTs
Economic evidence e.g. cost effectiveness
DECISION INPUTS
DECISION PROCESS
Cost-effectiveness threshold
Social value judgements
e.g. with respect to characteristics of patients, conditions or treatments
DECISION OUTPUTS
Yes
No

Figure 3. Flow diagram of appraisals included in analysis.

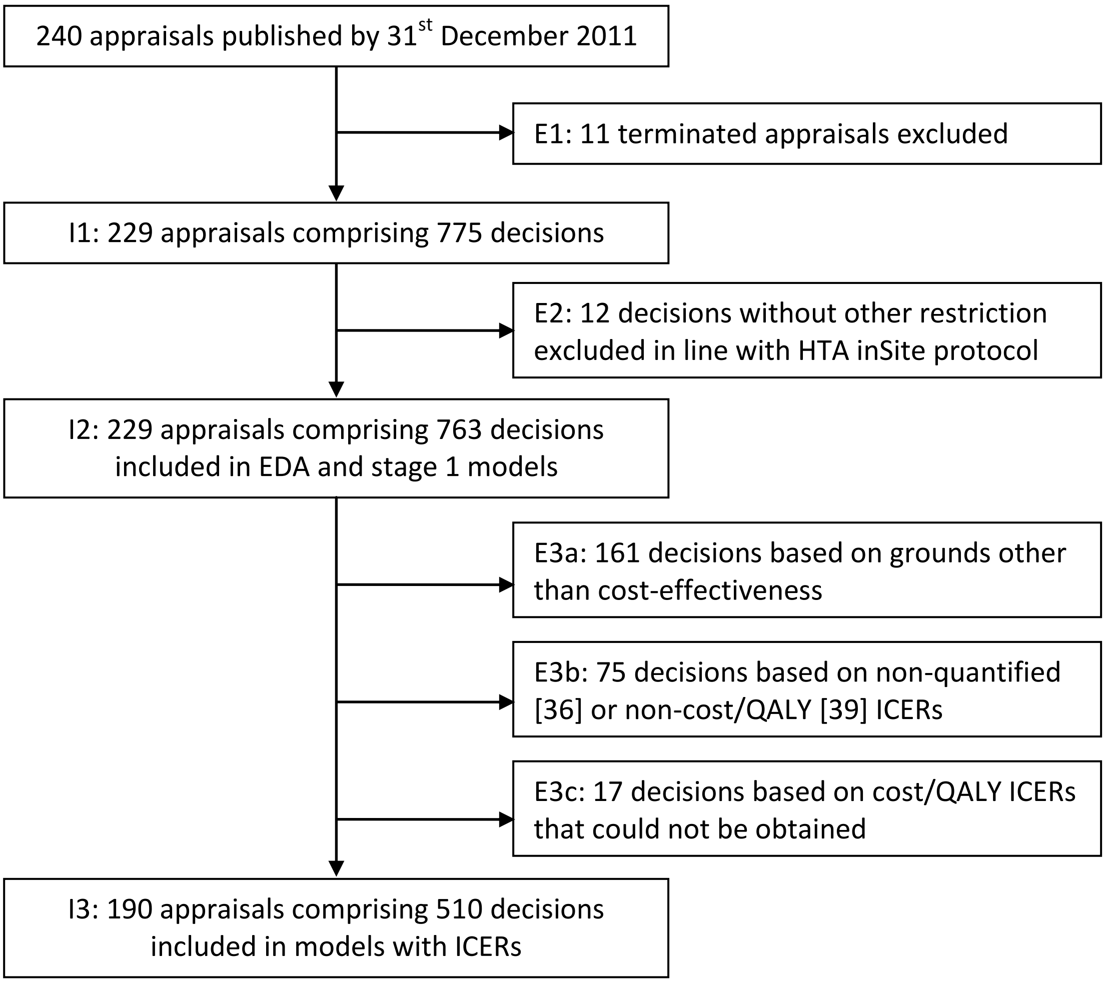
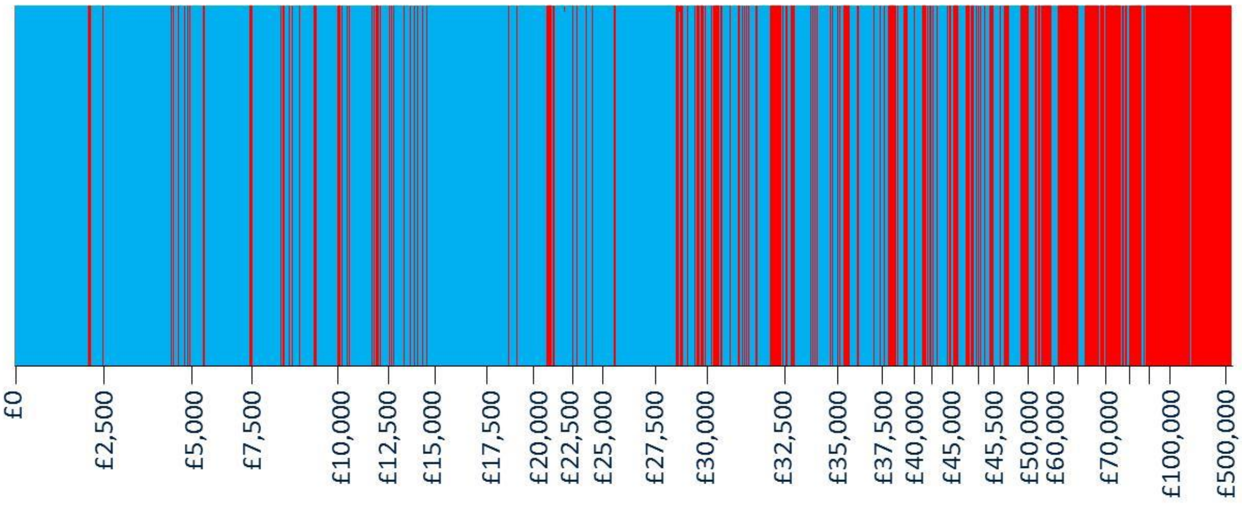

240 appraisals published by 31st December 2011
E1: 11 terminated appraisals excluded
I1: 229 appraisals comprising 775 decisions
E2: 12 decisions without other restriction excluded in line with HTA inSite protocol
I2: 229 appraisals comprising 763 decisions included in EDA and stage 1 models
E3a: 161 decisions based on grounds other than cost-effectiveness
E3b: 75 decisions based on non-quantified [36] or non-cost/QALY [39] ICERs
E3c: 17 decisions based on cost/QALY ICERs that could not be obtained
I3: 190 appraisals comprising 510 decisions included in models with ICERs

Figure 4. Impact of ICER ranking on recommendations.



Notes: Decisions are ranked by ICER, with NICE decisions to 'recommend' shown in blue and to 'reject' shown in red. For clarity, only the first five datasets of randomly-sampled ICERs are shown.

Figure 5. Predicted probability of NICE rejections at different ICER values for Models 1-5, holding all other variables at mean levels

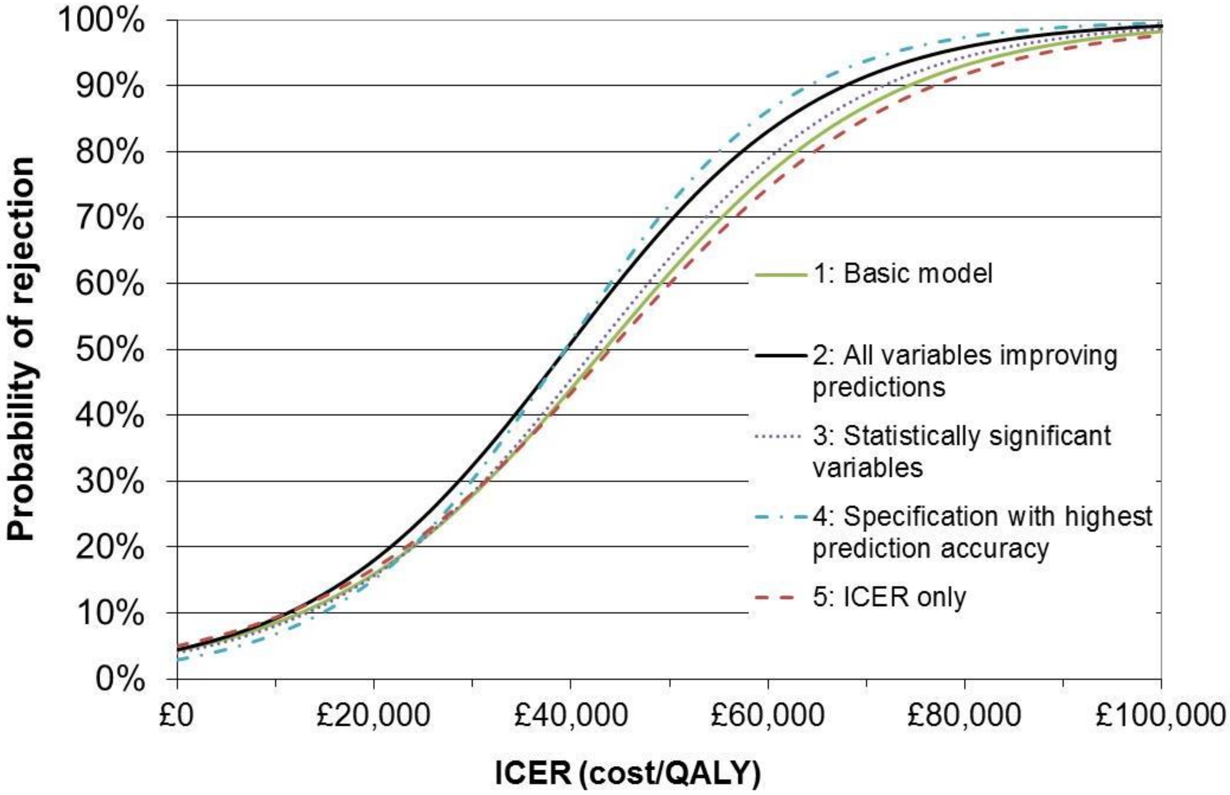


Table I. The core set of variables included in Model 1

Variable name	Coding	Definition	Justification
Dependent variable			
Recommendation	0=Not recommended 1=Recommended	Whether or not NICE recommended the technology for use in the population considered in this decision.*	Main outcome
Independent variables			
ICER	Numeric: £000s/QALY gained	Value of the cost per QALY gained for the technology considered in this decision compared with a comparator that NICE considered relevant to this decision. The ICER(s) most relevant to each decision were extracted for this study (Section 3).	NICE should consider "the broad balance of clinical benefits and costs" and make decisions based on "clinical effectiveness and cost effectiveness" (NICE 2008).
Total_pts_in_RCTs	Numeric: number of patients	Equals number of randomised controlled trials (RCTs) evaluating intervention in this population* (including commercial in confidence trials*) multiplied by mean number of patients in each fully reported RCT.*	"NICE should not recommend an intervention […] if there is […] not enough evidence" (NICE 2008).
Only_treatment	0=Not only treatment 1=Only treatment for this condition	Whether the technology (or all of the technologies considered within the same appraisal) comprises the only treatment available for the condition considered in this decision.*	Hypothesised that NICE is more likely to recommend if no alternatives.
Children	1=Concerns children 0=Does not concern children	Whether the decision concerns use of the treatment in children <18 years. Based on the age groups field in HTAinSite.*	Interventions for children are given 'the benefit of the doubt' due to methodological challenges (Rawlins 2010).
Pt_group_sub	1=Patient group submitted evidence 0=No patient group submission	Whether any patient groups made a submission to NICE in conjunction with the appraisal.*	Proxy for stakeholder involvement.
Date	Numeric (years)	Years elapsed between publication of first NICE appraisal in March 2000 and publication of this appraisal.*	Evaluates whether NICE decision-making is changing.
Severity	Numeric: disutility scale	Mean DALY weight across the diseases considered in the 2004 Global Burden of Disease study that fall into the relevant main disease category (WHO 2004). Severity was modelled in a similar way by Linley & Hughes (2012).	NICE state that they accept higher ICERs for serious conditions (Rawlins, 2010).

* Data taken from HTAinSite (www.htainsite.com).

Table II. Characteristics of included decisions

Variable		All decisions	No due to lack of evidence	Clinical grounds		Based on cost-effectiveness but no available NE quadrant cost/QALY		Included in regression analyses: NE quadrant ICER available	
				No	Yes	No	Yes	No	Yes
Total no. decisions		763	70	28	63	68	106	141	287
Mean ICER (SD)		£28,189 (£52,463)	N/A	N/A	N/A	N/A	N/A	£66,974 (£84,310)	£17,028 (£17,517)
% of ICERs (n/N)	≤£20,000/ QALY	43% (182/428)	N/A	N/A	N/A	N/A	N/A	11% (15/141)	58% (167/287)
	£20-£30,000/ QALY	14% (61/428)	N/A	N/A	N/A	N/A	N/A	8% (11/141)	17% (50/287)
	≥£30,000/ QALY	43% (184/428)	N/A	N/A	N/A	N/A	N/A	81% (114/141)	24% (70/287)
Total_pts_in_RCTs (SD)		3,402 (6,681)	680 (1,474)	1,502 (2,227)	1,965 (2,829)	3,817 (5,048)	2,861 (4,924)	3,108 (5,754)	4,812 (8,938)
% Only_treatment (n/N)		3% (25/763)	6% (4/70)	18% (5/28)	3% (2/63)	0% (0/68)	0% (0/106)	5% (7/141)	2% (7/287)
% Children (n/N)		10% (74/763)	7% (5/70)	14% (4/28)	27% (17/63)	13% (9/68)	11% (12/106)	4% (5/141)	8% (22/287)
% Pt_group_sub (n/N)		96% (730/763)	94% (66/70)	96% (27/28)	97% (61/63)	90% (61/68)	91% (96/106)	99% (139/141)	98% (280/287)
Date (SD): years since Mar 2000		5.8 (3.3)	5.7 (3.1)	3.1 (2.0)	3.9 (2.4)	4.8 (3.3)	3.0 (2.5)	6.3 (3.1)	6.5 (3.1)
Severity (SD): DALY weight		0.241 (0.110)	0.241 (0.119)	0.222 (0.114)	0.230 (0.091)	0.223 (0.106)	0.240 (0.113)	0.259 (0.097)	0.246 (0.114)
% STA (n/N)		19% (144/763)	16% (11/70)	0% (0/28)	6% (4/63)	8% (5/68)	9% (10/106)	34% (48/141)	23% (66/287)
% PSA (n/N)		63% (479/763)	47% (33/70)	29% (8/28)	57% (36/63)	49% (34/68)	33% (35/106)	66% (93/141)	77% (221/287)
% Orphan (n/N)		5% (36/763)	4% (3/70)	0% (0/28)	0% (0/63)	1% (1/68)	1% (1/106)	9% (13/141)	6% (18/287)
No_SRs (SD)		0.9 (2.5)	0.6 (2.8)	2.5 (4.1)	1.2 (3.0)	0.7 (1.7)	1.1 (2.2)	0.7 (1.8)	0.8 (2.5)
No_obs_studies (SD)		2.0 (7.8)	3.6 (11.0)	9.4 (20.6)	1.2 (4.8)	1.4 (6.0)	2.2 (9.1)	1.4 (5.5)	1.5 (5.2)
ICER_range (SD)		£33,641 (£134,021)	N/A	N/A	N/A	N/A	N/A	£97,683 (£235,107)	£9,133 (£17,920)
% Innovative (n/N)		15% (116/763)	9% (6/70)	4% (1/28)	8% (5/63)	17% (11/68)	14% (14/106)	19% (27/141)	18% (52/287)

Table III. Prediction accuracy and model fit for Models 1-5

Model name	% correctly classified	Sensitivity	Specificity	Mean AIC*	Mean adjusted pseudo-R^2*	Cost/QALY at which probability of a NICE recommendation is 50% (25%, 75%)†
1: ICER, Date, Total_pts_in_RCTs, Children, Only_treatment, Pt_group_sub & Severity	82.46%	94.02%	58.90%	-338	0.336	£43,356 (£58,793, £27,936)
2: ICER Total_pts_in_RCTs, Only_treatment, Children, Pt_Group_Sub, Date, STA, Orphan, No_SRs, No_obs_studies, PSA, Cancer, Cardiovascular, Infectious, Musculoskeletal, Respiratory, ICER_range, Innovative (model with best prediction accuracy after Stage B)	84.67%	93.18%	67.35%	-265	0.417	£39,479 (£53,616, £25,358)
3: ICER, Musculoskeletal, Respiratory, Cancer (variables significant in at least one analysis in Stages A&B)	83.50%	93.74%	62.66%	-332	0.362	£42,391 (£57,021, £27,781)
4: ICER Total_RCTs Mean_pts_per_RCT Only_treatmentifICER>30k Children Pt_group_sub ICER*Pt_group_sub [11 dummies for publication year] STA PSA Orphan No_SRs No_obs_studies Cancer Cardiovascular Infectious Musculoskeletal Respiratory ICER_range Innovative (model with best prediction accuracy after Stage C)	87.18%	94.24%	72.80%	-217	0.447	£39,417 (£51,754, £27,047)
5: ICER only	82.00%	93.30%	58.99%	-357	0.332	£43,949 (£60,377, £27,548)

* Mean AIC and pseudo-R^2 are shown for illustration only. Models were estimated separately for each of 100 datasets with ICERs sampled from the list of those relevant to each decision; the log-pseudo-likelihood for the model (LL_M) and for the constant-only model (LL_0) was averaged over the 100 datasets. AIC was calculated manually from the mean log-likelihood as $-2LL_M + 2k$ and adjusted pseudo-R^2 was calculated as $1-(LL_M/LL_0)*((n-1)/(n-k))$, where k=number of model parameters (explanatory variables plus constant) and n=number of decisions.

† The mean values for all other model parameters were multiplied by model coefficients to calculate the predicted log-odds of a positive NICE recommendation at a range of ICER values; the resulting figures were used to identify the ICER at which the probability of NICE saying 'yes' equalled 25%, 50% and 75%.

Table IV. Coefficients from Models 1 and 2

Variable	Odds ratio (95% CI)	
	Model 1	**Model 2**
ICER (£'000s)	0.931 (0.906, 0.957)**	0.925 (0.893, 0.959)**
Total_pts_in_RCTs	1.000 (1.000, 1.000)	1.000 (1.000, 1.000)
Only_treatment (dummy)	2.499 (0.457, 13.667)	4.279 (0.696, 26.297)
Children (dummy)	2.390 (0.312, 18.308)	4.097 (0.384, 43.740)
Pt_group_sub (dummy)	0.962 (0.097, 9.571)	1.119 (0.132, 9.498)
Date (years)	1.062 (0.943, 1.195)	1.134 (0.947, 1.357)
Severity (DALY weights)	0.397 (0.025, 6.362)	-
STA (dummy)	-	0.426 (0.185, 0.975)**
PSA (dummy)	-	0.443 (0.155, 1.271)
Orphan (dummy)	-	0.630 (0.144, 2.759)
No_SRs	-	1.024 (0.928, 1.130)
No_obs_studies	-	1.121 (0.991, 1.268)*
Cancer (dummy)	-	3.063 (1.119, 8.383)**
Cardiovascular (dummy)	-	0.837 (0.291, 2.401)
Infectious (dummy)	-	2.209 (0.359, 13.594)
Musculoskeletal (dummy)	-	5.732 (1.615, 20.343)**
Respiratory (dummy)	-	0.288 (0.089, 0.927)**
ICER_range (£'000s)	-	1.000 (1.000, 1.000)*
Innovative (dummy)	-	1.701 (0.656, 4.411)

* p<0.10; ** p<0.05

Appendix: Modelling strategy and additional coefficients

Principles used to guide our selection of the relevant ICERs:

- Include only cost per QALY gained; alternative cost-effectiveness measures (*e.g.* cost per life-year gained) were excluded;
- Where there were several ICERs reported for alternative comparators, use the ICER relative to the comparator that NICE considered most appropriate; if this is not specified in the Guidance, use the ICER relative to next most effective treatment on the cost-effectiveness frontier.
- Exclude ICERs that the 'consideration of evidence' section of the Guidance specifically indicated that NICE did not 'believe';
- Where the main ICER is a range rather than a point, capture the limits of that range, but do not include the wider range of ICERs that may be generated from full sensitivity analysis.

Table A1 Variables included in Stage B

Variable name	Coding	Definition	Justification
STA	1=STA 0=MTA	Whether the appraisal was conducted via the single technology appraisal (STA) process or the multiple technology appraisal (MTA) process.*	Mason and Drummond (2009) suggested that NICE may be more likely to say no in STAs.
Pharmaceutical	1=Pharmaceutical 0=other technology	Whether the technology was a drug. Based on the HTAinSite product type field HTAinSite.*	May reflect degree of stakeholder involvement.
Orphan	1=orphan drug 0=not an orphan drug	Whether the technology has been granted orphan status by the European Medicines Agency (EMEA).*	“NICE considers that it should evaluate drugs to treat rare conditions, known as ‘orphan drugs’, in the same way as any other treatment” (NICE 2008; Littlejohns and Rawlins, 2009).
No_SRs	Numeric: number of reviews	Number of systematic reviews mentioned in the Guidance and assessment report.*	Additional measure of clinical evidence.
No_obs_studies	Numeric: number of studies	Number of non-randomised studies mentioned in the Guidance and assessment report.*	Additional measure of clinical evidence.
PSA	1=PSA conducted 0= PSA not conducted	Whether the uncertainty around the economic evaluation was quantified using probabilistic sensitivity analysis (PSA).*	Significant predictor of AWMSG decisions (Linley & Hughes 2012).
Broader_perspective	1= considered broader costs 0= NHS only	Whether personal and societal costs were considered in addition to NHS cost (consideration included discussion in the text as well as inclusion in quantitative analyses).*	Reflects consideration of additional costs or savings not captured in the base case ICER.
Disease	Series of 8 dummy variables equal to 1 if concerned that disease	Each decision was classed as one disease category based on the “Main disease category” field within HTAinSite.* Disease categories with less than 20 decisions with ICERs were omitted. As result, decisions were categorised into cancer, cardiovascular, central nervous system, endocrine, infectious disease, mental health, musculoskeletal, respiratory and other.	May reflect variations in clinical need, severity or importance of rule of rescue between diseases, as well as different political priorities.
Innovative	1= classed as innovative 0= classed as non-innovative	Any molecule launched within two years of appraisal AND in an ATC4 class that was created within 5 years of the appraisal. Non-pharmaceutical interventions were classed as non-	For interventions with ICERs above £20,000/QALY, the committee will take account of “innovation that

Variable name	Coding	Definition	Justification
		innovative.	adds demonstrable and distinct substantial benefits that may not have been adequately captured in the measurement of health gain" (NICE 2008).
ICER_range	Numeric: difference between minimum and maximum ICERs	For decisions with more than one north-east quadrant ICER identified as driving the decision, this equalled the difference between the highest and lowest of such ICERs. Range was set to 0 for decisions with only 1 ICER.	For interventions with ICERs above £20,000/QALY, NICE will be "cautious about recommending a technology when they are less certain about the ICERs" (NICE 2008).

* Data taken from HTAinSite (www.htainsite.com).

Exploring the reasons for NICE recommendations for outliers

In general, the decisions that were poorly predicted by one of the five models were also poorly predicted by others. The rationale for the NICE decision was reviewed for 22 decisions where NICE rejected the technology, but models predicted that the probability of a positive recommendation was >0.62 (odds >0.5). In nine of these decisions the modelling was used in deliberations but uncertainty led to a judgement that the "true" ICER was substantially higher, which was not quantified in the guidance. In five instances, although a low ICER was reported, there were other treatment options with lower ICERs. Four further decisions had a wide range of ICER values, and comments in the guidance document implied that the higher ICERs may be plausible. Models predicted that the probability of a positive recommendation was <0.38 (odds <-0.5) in >66 of the 100 datasets for five decisions where NICE recommended the technology. In three of these cases, the committee made non-quantified adjustments to the reported ICER which implied that treatment was cost-effective for the subgroup for which it was recommended. One decision had a wide range of ICERs, although statements in the guidance suggested that the committee believed that the real ICER was in the lower end of this range. The final case was an early appraisal where NICE explicitly stated that an ICER in the range of £34,000 to £43,500 was cost-effective.

Methods of sensitivity analysis

The following analyses varying the specification of variables from the basic model were evaluated in Stage C:

- Replacing the numeric ICER variable with two dummies (CERbetween20and30k and NotCosteffectiveat30kRc) indicating what band the ICER falls into. The hypothesis here is that if NICE based their decisions purely on whether or not the ICER was above or below the threshold, then this model should fit as well as the base case model.
- Linear spline model: adding in two additional ICER terms as well as the ICER variable. One is equal to the ICER if the ICER is above £20,000 and zero if the ICER is below £20,000. Another is equal to the ICER if the ICER is above £30,000 and zero if it is below. This allows for the fact that the wording in the Social Value judgments

document implies that NICE is most sensitive to ICER value if the ICER is in the 20-30k region and insensitive to ICER value below 20k.

- Natural log of the ICER.
- ICER adjusted to allow for inflation to 2011/12 values based on the pay and prices index (Curtis 2012).
- Replacing Total_pts_inRCTs with two variables: total number of RCTs and the mean patient numbers in each reported RCT.
- Replacing the Only_treatment variable with three interaction terms, which were evaluated since this variable is expected to only be taken into account if the ICER is >20k and is expected to be more important if the ICER>30k.
    - onlytreatment20k= Costeffectiveat20kRc* only_treatment (dropped)
    - onlytreatment20_30k= CERbetween20and30k * only_treatment (dropped)
    - onlytreatment30k= NotCosteffectiveat30kRc * only_treatment
- Replacing the children variable with three interaction terms:
        - Children20k= Costeffectiveat20kRc * Children (dropped)
        - Children20_30k= CERbetween20and30k * Children (dropped)
        - Children30k= NotCosteffectiveat30kRc * Children
- Adding an interaction between Pt_group_sub and ICER
- Adding an interaction between Date and ICER interaction term
- Adding a Date squared variable to allow for non-linear effect of date
- Replacing the numeric Date variable with 3 dummies indicating whether the appraisal was:
        - Published between December 2005 and June 2008, while the first edition of the social value judgements document (NICE 2005) was in force.
        - Published after July 2008 when the latest social value judgements document (NICE 2008) was published
- Replacing the numeric Date variable with 11 dummies indicating the year of publication.

- Replacing the numeric Date variable with a dummy indicating whether or not the appraisal was published after (or at the same time as) the first STA appraisal was published.
- Adding an interaction between STA and ICER: explores whether ICERs are interpreted differently if they come from an STA rather than an MTA.

Sensitivity analyses conducted on Model 4:

- Adding in five dummy variables indicating which of the six committees evaluated the decision.
- Adding in five dummy variables indicating committee, in addition to five interactions between committee and ICER.
- Probit model (not logit)
- No clustering
- Random effects analysis to evaluate the impact of clustering by committee as well as clustering by appraisal
- Random effects on appraisal (rather than clustering)
- Fixed effects on appraisal (rather than clustering)
- Replacing the ICER variable and all variables derived from it with each of the following in turn:
    - Mean across al ICERs identified as driving the decision
    - Midpoint between minimum and maximum ICER of those driving the decision
    - Minimum ICER of those driving the decision
    - Maximum ICER of those driving the decision
    - Using the maximum ICER for decisions that were rejected by NICE and the minimum ICER in the list for decisions that were recommended.

Results of Stages C and D

Variable specification was varied within Stage C, with the specification of each variable that had highest prediction accuracy being selected for inclusion within Model 4. This model correctly predicted 87.18% of NICE decisions. This analysis suggested that RCT evidence was best considered as an additive relationship between the total number of trials and the average size (rather than as the product of these two), although neither variable had a significant effect on NICE decisions (Table A2). Each additional RCT increased the odds of a positive NICE recommendation by 1.2% (p=0.54), while increasing the size of the average RCT by one patient decreased the odds by 0.008% (p=0.183). These coefficients are similar to those reported previously (Dakin et al, 2006), although our previous study found the number of RCTs to exert a statistically significant effect. Stage C modelling also suggested that a lack of alternative treatments may only affect NICE decision-making for decisions with ICERs above £30,000/QALY and suggested that prediction accuracy is improved by adding an interaction term between the ICER and patient group submissions, such that patient group submissions have greater impact for decisions with high ICERs.

Table A2: Coefficients for Model 4.

Variable	Variable definition	Odds ratio (95% CI): Model 4
ICER (£'000s)	See Table I	0.858 (0.775, 0.951)**
Total_RCTs	Alternative specification of RCT evidence. Number of randomised controlled trials (RCTs) evaluating intervention in this population* (including commercial in confidence trials*).	1.012 (0.974, 1.052)
Mean_pts_per_RCT	Alternative specification of RCT evidence. Mean number of patients in each fully reported RCT.*	1.000 (1.000, 1.000)
Only_treatment_ifICER>30k	Dummy equal to 1 if the decision has an ICER above £30,000/QALY and has no alternative treatments (zero otherwise)	13.198 (0.945, 184.340)*
Children	See Table I	4.274 (0.325, 56.142)
Pt_group_sub	See Table I	0.403(0.004, 37.486)
ICER*Pt_group_sub	Interaction term: product of ICER and Pt_group_sub	1.067 (0.965, 1.181)
2001-2	Dummy variables indicating the year of guidance publication (base year: 2000-1)	10.117 (0.039, 2616.590)
2002-3		0.352 (0.041, 3.050)
2003-4		0.077(0.008, 0.697)**
2004-5		0.164 (0.008, 3.562)
2005-6		0.172 (0.014, 2.173)
2006-7		0.517 (0.068, 3.907)
2007-8		1.035 (0.119, 9.025)
2008-9		0.369 (0.050, 2.697)
2009-10		0.790 (0.074, 8.407)
2010-11		1.241 (0.139, 11.123)
2011-12		0.358 (0.037, 3.493)
STA	See Table I	0.410 (0.156, 1.083)*
PSA	See Table I	0.611 (0.222, 1.684)
Orphan	See Table I	0.733 (0.147, 3.667)

Variable	Variable definition	Odds ratio (95% CI): Model 4
No_SRs	See Table I	1.103 (0.892, 1.365)
No_obs_studies	See Table I	1.143 (0.981, 1.331)*
Cancer	See Table I	3.417 (1.116, 10.465)**
Cardiovascular	See Table I	1.658 (0.434, 6.335)
Infectious	See Table I	4.532 (0.582, 35.306)
Musculoskeletal	See Table I	7.889 (1.509, 41.247)**
Respiratory	See Table I	0.347 (0.103, 1.172)*
ICER_range (£'000s)	See Table I	1.000 (1.000, 1.000)**
Innovation	See Table I	1.965 (0.687 , 5.616)

* p<0.1; ** p<0.05

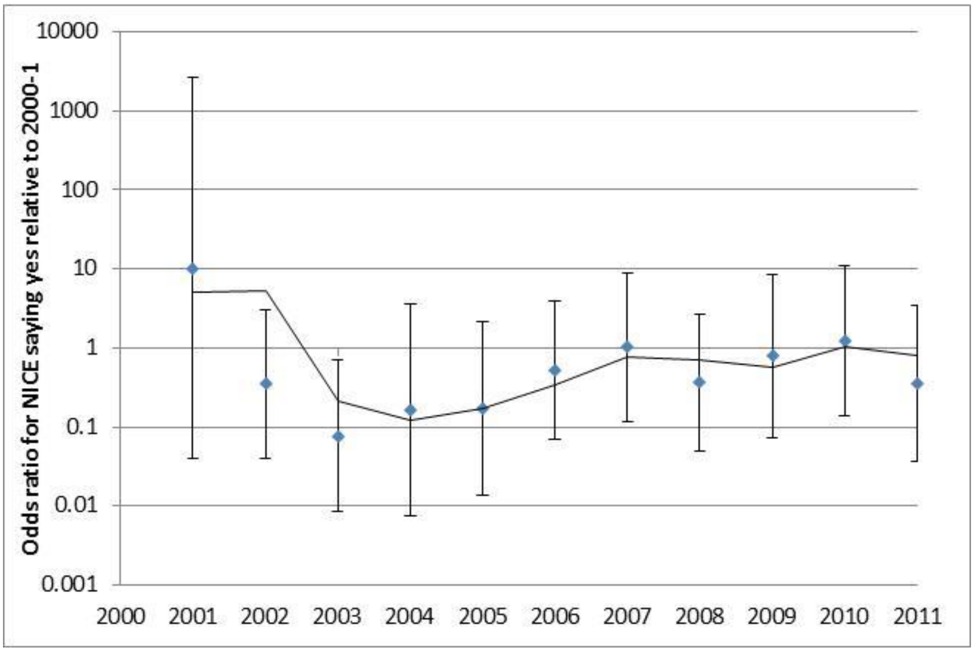
As a sensitivity analysis (Stage D), we also evaluated the impact of committee on NICE decisions, by categorising appraisals into six categories based on the chairperson of the committee that made the recommendation. This suggested that adding committee variables into Model 4 improved prediction accuracy, although there were no statistically significant differences between committees.

Using the mean of the relevant ICERs or the midpoint between the highest and lowest ICERs for those decisions with more than two relevant ICERs rather than using the simulation approach increased the proportion of decisions correctly classified by Model 4. This may suggest that when faced with several equally plausible ICER values, NICE (or individual committee members) base decisions on the mean or midpoint of the available ICERs. Although the illustration of the probabilistic threshold presented by Rawlins (2004) suggested that NICE consider ICERs on a logarithmic scale, taking the natural logarithm of the ICER reduced prediction accuracy, which may suggest that the NICE committees consider ICERs on a natural scale.

Replacing the ICER variable with dummy variables suggested that decisions with ICERs above £30,000/QALY (p<0.001) and those with ICERs between £20,000 and £30,000/QALY (p=0.003) were significantly less likely to be recommended than those with ICERs below £20,000/QALY. However, replacing the numeric ICER variable with dummies reduced prediction accuracy for Model 2, which suggests that although the magnitude of the ICER does affect the odds of NICE rejection, fixed thresholds of £20,000 and £30,000/QALY explain a large proportion of NICE decision-making.

Allowing for non-linear effects of date by including a publication date squared variable reduced prediction accuracy. However, replacing the publication date variable with dummy variables for the year the appraisal was published increased prediction accuracy; this analysis suggested that the chance of NICE saying 'yes' may have decreased between 2000-1 and 2003-4 and risen between 2003-4 and 2011-12, although the odds of NICE saying 'yes' were significantly different from the odds in 2000-12 only in 2003-4 (Table A2, Figure A1). We also investigated whether NICE decision-making changed after NICE published its first Social Value Judgements document in November 2005 or after the those documents and the description of NICE's stated threshold were revised in 2008 (NICE, 2005; 2008). Replacing the date variable with dummy variables suggested that the odds of NICE recommending a treatment were non-significantly lower after June 2008 than before November 2005 (p=0.12) or between November 2005 and June 2008 (p=0.12). A further analysis found that decisions published after (or at the same time as) the first STA appraisal were non-significantly more likely to be recommended than those published earlier (p=0.07). Although we might expect a change in the discount rates recommended by NICE to affect ICERs, we found that neither the odds of NICE decisions nor the coefficient for the ICER changed significantly after the 2004 Methods Guide introducing the new discount rates was published (NICE 2004).

Figure A1. Changes in the odds of NICE recommendation over time


Odds ratio for NICE saying yes relative to 2000-1