THE INFLUENCE OF COST-EFFECTIVENESS AND OTHER FACTORS ON NICE DECISIONS

HELEN DAKIN[a],*, NANCY DEVLIN[b], YAN FENG[b], NIGEL RICE[c], PHILL O'NEILL[b] and DAVID PARKIN[d]

[a]*Health Economics Research Centre, University of Oxford, Oxford, UK*
[b]*Office of Health Economics, London, UK*
[c]*Centre for Health Economics and Department of Economics and Related Studies, University of York, York, UK*
[d]*Department of Primary Care and Public Health Sciences, King's College London, London, UK*

ABSTRACT

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 but not the weight attached to each.

The objective of this study is to investigate the influence of cost-effectiveness and other factors on NICE decisions and whether NICE's decision-making has changed over time.

We model NICE's decisions as binary choices for or against a health care technology in a specific patient group. Independent variables comprised of the following: clinical and economic evidence; characteristics of patients, disease or treatment; and contextual factors potentially affecting decision-making. Data on all NICE decisions published by December 2011 were obtained from HTAinSite [www.htainsite.com].

Cost-effectiveness alone correctly predicted 82% of decisions; few other variables were significant and alternative model specifications had similar performance. There was no evidence that the threshold has changed significantly over time. The model with highest prediction accuracy suggested that technologies costing £40 000 per quality-adjusted life-year (QALY) have a 50% chance of NICE rejection (75% at £52 000/QALY; 25% at £27 000/QALY).

Past NICE decisions appear to have been based on a higher threshold than £20 000–£30 000/QALY. However, this may reflect consideration of other factors that cannot be easily quantified. © 2014 The Authors. *Health Economics* published by John Wiley & Sons Ltd.

1. INTRODUCTION

1.1. NICE decision-making

Health technology assessment (HTA) agencies' decision-making criteria are important for health care providers, patients' eligibility for health care and technology firms' revenue and investment and production decisions about current and potential products. Stafinski *et al*. (2011) note that although European countries' centralised authorities state their criteria, "there remains a lack of transparency around critical elements, such as how multiple factors or criteria are weighed during committee deliberations".

The National Institute of Health and Care Excellence (NICE) provides guidance on which health care interventions are available from the National Health Service (NHS) in England and Wales. Although NICE's remit and aims are defined by its establishing legislation, NICE was allowed to develop its methods and processes,

*Correspondence to: Health Economics Research Centre, Nuffield Department of Population Health, University of Oxford, Old Road Campus, Headington, Oxford OX3 7LF, UK. E-mail: helen.dakin@dph.ox.ac.uk

which have become increasingly stable and clear. However, areas of considerable uncertainty remain about their decision-making criteria.

Rawlins and Culyer (2004) state that NICE's main decision-making criterion is cost-effectiveness, usually measured by the incremental cost-effectiveness ratio (ICER) expressed as cost per quality-adjusted life-year (QALY) gained. The 'threshold' ICER that determines whether a technology is cost-effective is intended to represent the opportunity cost to a fixed-budget NHS of adopting a technology in terms of QALYs forgone (McCabe *et al*., 2008; NICE, 2013). This 'shadow price of a QALY' is quantified but not as a single 'threshold'. Instead, ranges that affect the *probability* of recommending a technology are described. Although probability is invoked, the actual probabilities are not quantified. However, Rawlins and Culyer (2004) suggested that a curve relating increasing risk of rejection and log-ICER would be a sigmoid, with inflexion points at £5000–£15 000 and £25 000–£35 000/QALY gained. The most recent and definitive statement by NICE (2013) is that technologies costing < £20 000/QALY gained are normally considered cost-effective. Above £20 000/QALY, judgements take account of uncertainty, innovation and whether there are non-health outcomes, end of life considerations or stakeholder views that quality of life gains are inadequately captured; these increase in importance as ICERs increase. Moreover, interventions for children, disadvantaged populations and severe diseases are treated more favourably (NICE, 2008, 2009; Rawlins *et al*., 2010). However, it is explicitly said that some factors are not taken into account, for example, orphan drugs for rare conditions (Littlejohns and Rawlins, 2009).

Appealing to probabilities rather than thresholds gives NICE considerable decision-making discretion and reduces challenges to its approach and to any precise value for the threshold. However, it generates uncertainty about why particular decisions are made, which reduces NICE's accountability and predictability of future decisions, which may affect decisions about health technology research and development. The weights attached to other factors are also rarely quantified, and their impact/importance is therefore even more uncertain than cost-effectiveness.

In this paper, we explore the role of different criteria used in decision-making by a public body comprising multiple committees, each comprising multiple individuals who weigh criteria up via a 'deliberative process' (Culyer, 2006, 2009; NICE, 2013). Such processes typically involve scientific evidence (both context-free and context-sensitive) and 'colloquial' evidence (any other evidence that people use in making decisions). The appraisal committee's role is not to ensure correct application of an explicit decision-making formula but to exercise judgement over available evidence, including that offered by its members. Criteria for decision-making are equally non-explicit, involving a 'search' for them based on expert opinion, research and accumulated experience (Culyer *et al*., 2007). Lack of clarity and deliberate non-explicitness add to the importance of exposing implicit criteria and value judgements affecting public spending to public scrutiny and assessing the extent to which society's preferences are reflected by implied weights.

1.2. Previous research

This analytical framework was used in previous studies. Devlin and Parkin (2004) found that cost-effectiveness was the key driver of NICE decisions, although uncertainty and burden of disease were also significant. Dakin *et al*. (2006) found that decisions were influenced by cost-effectiveness, clinical evidence, technology type, and patient group submissions. Cerri *et al*. (2014) found that in addition to cost-effectiveness, demonstration of statistical superiority of the primary endpoint in clinical trials, the number of pharmaceuticals appraised within the same appraisal and the appraisal year were also important.

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

1.3. Aims

This paper assesses the impact of NICE's criteria on its decision-making in practice and estimates the weights given to different criteria implicit within NICE technology appraisal (TA) decisions. We investigate empirically the effect of different factors on the likelihood of NICE recommending a technology, using a revealed preference approach to model decision-making, building on and extending earlier studies (Devlin and Parkin, 2004; Dakin *et al.*, 2006; Cerri *et al.*, 2014). In particular, the much larger number of decisions now available facilitate exploration of additional research questions and examine changes over time. Statements about thresholds have evolved in subtle yet important ways: initially, an 'unwritten rule' of £30 000; then descriptions as lying between £20 000 and £30 000 (NICE, 2005); finally, references to £20 000 with exceptions (NICE, 2008, 2013). Key aspects of TA processes and methods have also evolved.

We used logistic regression to estimate binary econometric models predicting whether or not a technology was recommended. Recommendations for whole TAs involve three alternatives: 'yes' to all patients and technologies considered; 'no' to all patients; and 'restricted' or 'optimised' (NICE, 2010), meaning 'yes' to some patient subgroups and 'no' to others. Dakin *et al.* (2006) and Cerri *et al.* (2014) categorised recommendations in this manner, but this has important limitations. The evidence considered in restricted recommendations may differ between the patient subgroups for whom the technology is recommended or rejected, and the proportion of patients accessing the technology varies considerably (O'Neill and Devlin, 2010). We therefore divide restricted appraisals into their component yes/no *decisions* about the use of a single technology for a clearly defined group of patients. This enables us more precisely to link the decision to the evidence.

We aimed to address the following:

1. Does the probability of rejection increase with increasing ICER?
2. Are past NICE decisions consistent with NICE's stated threshold range and/or Rawlins and Culyer's sigmoid curve?
3. Do other factors identified by NICE affect the rejection probability? Do factors that are said not to merit special consideration affect it?
4. Have NICE's decision criteria or threshold or both changed over time?

2. METHODS

2.1. Decisions included in the analysis

The data for this study were obtained from HTAinSite© (www.htainsite.com) and initially comprised all 240 NICE TAs published by 31 December 2011, with the exception of 11 appraisals that were terminated before any decision was made (Figure 1).

Each of the 229 non-terminated TAs was subdivided into 1–19 component *decisions*, each representing a NICE decision to either recommend or reject a single technology in a specific patient population. Subdivision of each TA inevitably requires a degree of researcher judgement; our data set follows that of HTAinSite, which uses a carefully documented protocol providing a set of principles for making those judgements consistently (refer to Supporting Information).

Regression analyses focused on decisions with available, quantified cost/QALY estimates lying in the northeast quadrant of the cost-effectiveness plane, where treatment is more effective and more costly than its comparator (refer to Sections 2.2 and 3.1).

2.2. Variables extracted and analysed

The conceptual judgements underpinning NICE appraisals that were outlined in Section 1 were used to select a core set of variables for the first regression model (**model 1**) from the fields available in HTAinSite (Table I). In addition to the ICER, we included one variable indicating the amount of clinical evidence (Total_pts_in_RCTs)
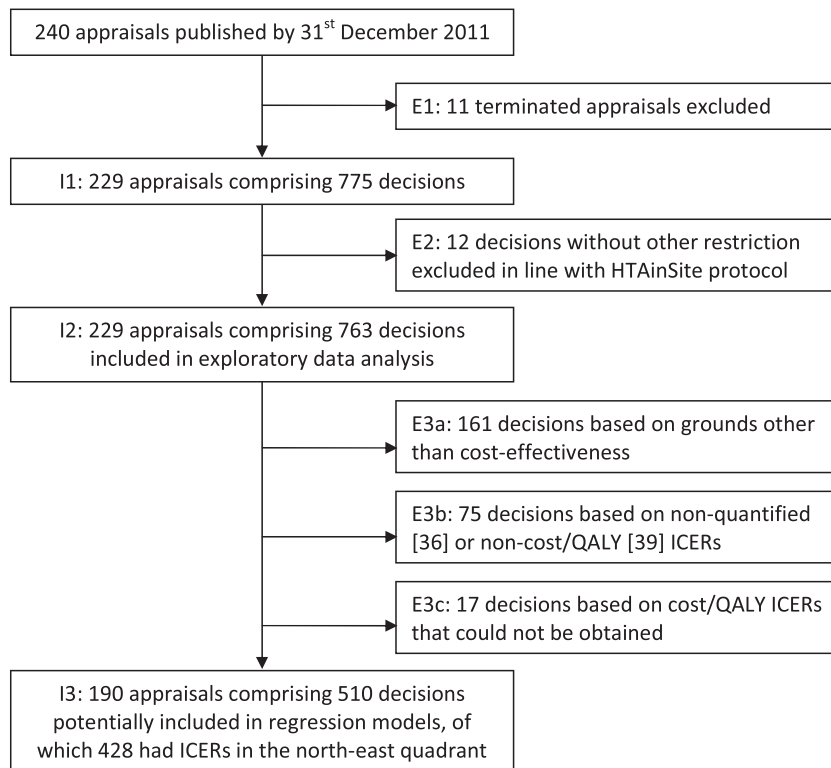
Figure 1. Flow diagram of appraisals included in analysis

because 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 as a result of difficulties defining variables that consistently capture these issues.

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 (refer to Supporting Information).

For 45% (229/510) of decisions with usable ICERs, we identified ≥2 ICERs that informed NICE's decision-making. 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 £X' 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 (100 times) 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 was randomly sampled by assigning equal probability to all ICERs that appeared from the Guidance to have influenced the decision. For the 31 cases giving a range or lower/upper limit, ICER values were sampled from a list of all ICERs within our data set that lay in the relevant range because ICERs follow an unknown distribution and may approach infinity (Briggs and Fenn, 1998). These sampling procedures generated 100 data sets, each with different ICER data for those decisions with >1 relevant ICER.

Table I. Definition of variables included in the analysis

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 included in model 1			
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 (refer to Supporting Information).	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 RCTs evaluating intervention in this population* (including commercial in confidence trials*) multiplied by mean number of patients in each fully reported RCT.* Total_pts_in_RCTs equalled zero for decisions with no RCTs.	"NICE should not recommend an intervention […] if there is […] not enough evidence" (NICE, 2008). "RCTs are […] considered to be most appropriate for measures of relative treatment effect" (NICE, 2013). Total_pts_in_RCTs was used to capture both the number and size of trials informing treatment effects.
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 group field in HTAinSite.*	Interventions for children are given 'the benefit of the doubt' as a result of methodological challenges (Rawlins *et al.*, 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 and Hughes (2012).	NICE state that they accept higher ICERs for serious conditions (Rawlins *et al.*, 2010).
Additional variables explored in stage B and considered for inclusion in model 2			
STA	1 = STA 0 = MTA	Whether the appraisal was conducted via the STA process or the 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.*	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 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.

(*Continues*)

Table I. (Continued)

Variable name	Coding	Definition	Justification
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 PSA.*	Significant predictor of AWMSG decisions (Linley and 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 eight 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 2 years of appraisal AND in an ATC4 class that was created within 5 years of the appraisal. Non-pharmaceutical interventions were classed as non-innovative.	For interventions with ICERs above £20 000/QALY, the committee will take account of "innovation that 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).
ATC4, Anatomical Therapeutic Chemical classification system for drugs; AWMSG, All Wales Medicines Strategy Group; DALY, disability-adjusted life-year; EMEA, European Medicines Agency; ICER, incremental cost-effectiveness ratio; MTA, multiple technology appraisal; NHS, National Health Service; NICE; National Institute for Health and Care Excellence; PSA, probabilistic sensitivity analysis; RCT randomised controlled trial; QALY, quality-adjusted life-year; SR, systematic review; STA, single technology appraisal; WHO, World Health Organisation.

Regression analyses included only decisions concerning treatments that are more costly and more effective than their comparator. Sixty-four decisions for which all relevant ICERs indicated that the technology that was either dominated or dominant relative to its comparator were excluded from regression analyses because dominance perfectly predicted NICE recommendations. ICERs in the south-west quadrant of the cost-effectiveness plane have the opposite interpretation to those in the north-east quadrant, 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 data sets where a north-east quadrant ICER was sampled and were dropped from regressions in data sets where an ICER from another quadrant was sampled. As result, 440 decisions were included in regression analyses on ≥1 data set, although the number of decisions included in each regression varied between 424 and 432 (mean = 428), depending on the number of north-east quadrant ICERs in each data set.

2.3. Modelling approach

We modelled NICE decisions using logistic regression, which relates the probability (P) of a technology being recommended by NICE ($y_i = 1$) to a set of explanatory variables (x_i) as follows:

$$P(y_i = 1|x_i) = \frac{exp(x_i\beta)}{1 + exp(x_i\beta)}$$

where β is a vector of parameters to be estimated and presented as odds ratios: For example, a one unit change in the jth variable, x_j, is associated with the odds ratio, $exp(\beta_j)$.

Standard errors were adjusted for within-appraisal clustering of decisions because 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 they are unlikely to be independent. All statistical analyses were conducted in Stata Version 12.

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

2.4. Model selection

The primary measure of model performance comprised the proportion of decisions that were correctly classified because it is not valid to apply Rubin's rule to the 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 sample independent of the data used to estimate the model (Copas, 1983). Unfortunately, this was not feasible because of the limited number of appraisals available; we therefore rely on a single data set 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 'no' decisions predicted as 'no') and sensitivity (the proportion of 'yes' decisions predicted as 'yes'), was calculated by assuming that all technologies with $\geq 50\%$ 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 data sets) 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 four stages, which were used to determine the specification of models 2–4. 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, a priori, to have the most effect on NICE decisions (Table I).

(B)　Identification of additional variables explaining NICE decision-making. We added additional independent variables 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 those which best 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 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. Alternative specifications (e.g. interactions or squared terms) were considered for each variable (refer to Supporting Information); 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 (refer to Supporting Information).

Models 1–4 were compared against a model including only the ICER (**model 5**). 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. The predicted probability of NICE rejection was plotted against ICER and compared against the sigmoid curve proposed by Rawlins and Culyer (2004).

3. RESULTS

3.1. Exploratory data analysis

Our data set comprised 763 decisions from 229 appraisals (Figure 1). Of these, 253 decisions that did not report any usable ICERs were omitted from regression analyses:

a) Seventy decisions were 'no' as a result of of clinical evidence; these decisions had significantly fewer patients in RCTs ($p < 0.001$) than other decisions (Table II), although 59% (41/70) were nonetheless supported by ≥ 1 RCT.

b) Sixty-three decisions were 'yes' on clinical grounds (e.g. because all alternative technologies were contraindicated or not tolerated), while 28 decisions were 'no' on clinical grounds (e.g. because treatment was 'clinically inappropriate' in that patient group). The decisions made on clinical grounds were, on average, published 2 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) One hundred seventy-four 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/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 'no' as treatment was dominated by its comparator, while 31 were 'yes' as treatment dominated. Six decisions had ICERs in the south-west quadrant, of which one was 'no'. As discussed in Section 2.2, we allowed for decisions with >1 relevant ICER by sampling 100 data sets with different ICER values for these decisions. On average across the 100 data sets, a further 12 decisions were excluded from regression analyses because non-north-east quadrant ICERs were sampled from a set of relevant ICERs spanning >1 quadrant. The 174 decisions based on cost-effectiveness that lacked available north-east quadrant cost/QALY ICERs tended to be published about 4 years earlier than those included in regression analyses ($p < 0.001$) and were less likely to be single technology appraisals (STAs; $p < 0.001$) or only treatments ($p < 0.001$).

Among the 428 decisions with available north-east quadrant ICERs, ICERs differed significantly between 'yes' and 'no' decisions ($p < 0.001$; Table II). Exploratory data analysis also demonstrated that the proportion of technologies rejected by NICE increases substantially with ICER (particularly at ~£27 500 and ~£47 500/ QALY), although there are numerous exceptions (Figure 2).

Table II. Characteristics of included decisions

Variable	All decisions	No as a result of lack of evidence	Clinical grounds		Based on cost-effectiveness but no available NE quadrant cost/QALY		Included in regression analyses: NE quadrant ICER available	
NICE decision			No	Yes	No	Yes	No	Yes
Total no. of 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)
£20 000–£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)	3402 (6681)	680 (1474)	1502 (2227)	1965 (2829)	3817 (5048)	2861 (4924)	3108 (5754)	4812 (8938)
% 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 March 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)	£9133 (£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)

DALY, disability-adjusted life-year; ICER, incremental cost-effectiveness ratio; N/A, not applicable; NE, north-east [quadrant of the cost-effectiveness plane]; NICE, National Institute for Health and Care Excellence; PSA, probabilistic sensitivity analysis; QALY, quality-adjusted life-year; SD, standard deviation
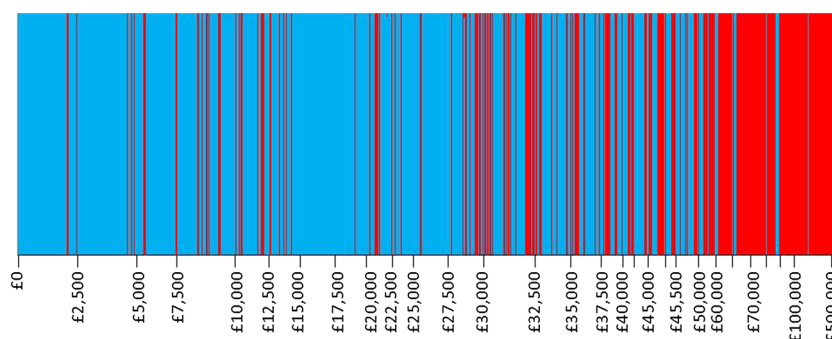
Figure 2. Impact of ICER ranking on recommendations. Decisions are ranked by ICER, with NICE decisions to 'recommend' shown in blue and to 'reject' shown in red. For clarity, only the first five data sets of randomly sampled ICERs are shown

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 and 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 and 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 data sets 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) were averaged over the 100 data sets. AIC was calculated manually from the mean log-likelihood as $-2LL_M + 2k$ and adjusted pseudo-R^2 was calculated as $1 - (LL_M/LL_0) \times ((n-1)/(n-k))$, 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

	Odds ratio (95% CI)	
Variable	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)

$*p < 0.10$;
$**p < 0.05$

3.2. 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 III). As expected, the ICER had a significant effect on NICE decisions, with every £1000 increase in the ICER reducing the odds of NICE recommending the technology by 6.9% (95% confidence interval (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 ($p \geq 0.29$; Table IV). 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.

Taking account of 12 of the 17 additional variables evaluated in stage B slightly improved prediction accuracy. 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: a small improvement on model 1 (Table III).

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 interventions treating, preventing or diagnosing cancer and 71% lower ($p = 0.037$; 95% CI: 7, 91%) for interventions for respiratory disease. Model 4 gave similar findings (Supporting Information).

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 ($p \geq 0.103$).

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 interventions 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. A sensitivity analysis found that allowing for 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 would correctly classify only 73.1% of NICE decisions.

3.3. Relationship between ICER and probability of NICE recommendation

Coefficients from models 1–5 were used to estimate how the probability of NICE rejection varied with ICER, holding all other parameters at mean values (Figure 3, Table III). 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 3). 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 narrower 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 values 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.

The decisions that were poorly predicted by our models were generally rejected because of substantial uncertainty or included statements within the guidance suggesting that the committee believed the ICER to be at/near the top or bottom of the stated range. This is supported by a sensitivity analysis using the minimum ICER for all 'yes' decisions and the maximum ICER for all 'no' decisions, which correctly classified 93.0% of decisions; this may suggest that around two-thirds of the decisions poorly predicted by our model are misclassified because of difficulties identifying the ICER that drove the committee's decision based on secondary data.

3.4. 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 2011/2012 values using the hospital and community health service pay and price index (Curtis, 2012) reduced prediction accuracy. We
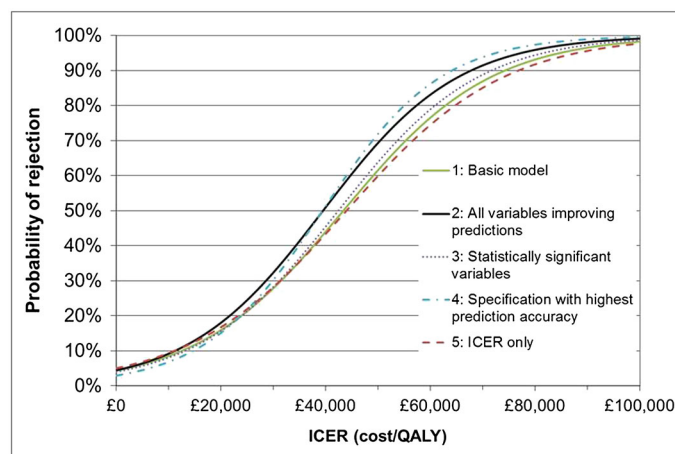

Figure 3. Predicted probability of NICE rejections at different ICER values for models 1–5, holding all other variables at mean levels

examined alternative specifications of publication date to assess the impact on prediction accuracy, although no statistically significant temporal trends were observed.

4. DISCUSSION

4.1. 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 varying 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 support neither the £5000–£15 000/QALY and £25, 000–£35, 000/QALY inflexion points proposed by Rawlins and Culyer (2004) nor NICE's stated threshold range. Based on NICE (2008, 2013) statements, we would expect the following: for ICERs under £20 000/QALY, the probability of rejection would be <0.5; above £30 000/QALY, it would be >0.5; and the switch from odds-on to odds-against would be 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. Claxton *et al.* (2013) and Appleby *et al.* (2009) adopt an opportunity cost definition of the threshold, while Baker *et al.* (2010) provide estimates of the social value of a QALY. While NICE formally subscribes to an opportunity cost definition, our results clearly show that, in practice, NICE often recommends technologies with ICERs that are well above the opportunity cost estimate of Claxton *et al.* (2013) but somewhat closer to the social value of a QALY (Baker *et al.*, 2010).

4.2. 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 7 years. Although the models reported here treat ICERs in nominal terms, inflation must have affected the prices and costs embodied in the ICERs in appraisals conducted over this 10-year period; nevertheless, 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 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 75% (49/65) of cancer decisions before January 2009 being 'yes', versus 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.

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, 2011). However, others argue that a deliberative process without pre-defined weights is needed to consider the evidence and make complex decisions (Culyer and Lomas, 2006).

However, it is 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. In

particular, our descriptive analysis suggests that 21% (161/763) of decisions are based on clinical considerations or lack of clinical evidence without considering cost-effectiveness. The influence of additional factors not evaluated in our analysis that affect decisions with cost-effectiveness evidence would have biased upwards our estimate of the ICER at which the probability of rejection is 50%.

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.*, 2010), 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 disability-adjusted life-year weight across broad disease categories) 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.

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 assume that it has no impact. However, budget impact and population size estimates are only recorded for whole TAs based on the patient subgroups for which treatment was recommended; estimating the net budget impact and population size for each decision would be a substantial task, beyond the scope of this project.

Although we 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 about the variable that they included to capture the ICER range. Amongst the complexities in representing uncertainty around the ICER is that it may be because of the uncertainty around costs, QALYs, or both. QALY uncertainty could arise from many different sources, such as insufficient trial evidence (which might be captured by Total_pts_in_RCTs) or structural model assumptions (e.g. around extrapolation). NICE might, in principle, react differently to these different sources of uncertainty. The way in which uncertainty information is presented to NICE committees might also affect decisions. We considered using a variable estimated from confidence intervals or cost-effectiveness acceptability curves but rejected this because that would require further value judgements and/or exclusion of 126 decisions where probabilistic sensitivity analysis was not undertaken.

4.3. Conclusions

Our main result, based on a larger number of decisions than previous studies (Devlin and Parkin, 2004; Dakin *et al.*, 2006; Cerri *et al.*, 2014), confirm their conclusions that cost-effectiveness is the major driver of NICE decisions. However, although we explored the impact of a wide range of potential predictors, we did not replicate their findings with respect to other important factors. 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 multiple technology appraisal), orphan status, innovation and uncertainty improved prediction accuracy. 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.

CONFLICT OF INTEREST

The authors have no conflict of interest.

REFERENCES

Appleby J, Devlin N, Parkin D, Buxton M, Chalkidou K. 2009. Searching for cost-effectiveness thresholds in the NHS. *Social Science & Medicine* **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**: 723–740.
Carlin JB, Galati JC, Royston P. 2008. A new framework for managing and analyzing multiply imputed data in Stata. *Stata Journal* **8**: 49–67.
Cerri KH, Knapp M, Fernandez JL. 2014. Decision making by NICE: examining the influences of evidence, process and context. *Health Economics, Policy, and Law* **9**(2): 119–141.
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 (Available from: 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. *Journal of the Royal Statistical Society: Series B* **45**: 311–354.
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 A. 2009. Deliberative processes in decisions about health care technologies: combing different types of evidence, values, algorithms and people. OHE Briefing no 48. Office of Health Economics: London.
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.
Curtis L. 2012. Unit Costs of Health and Social Care 2012. PSSRU Personal Social Services Research Unit: Canterbury. (Available from: www.pssru.ac.uk/project-pages/unit-costs/2012/ Accessed 30th July 2013.)
Dakin HA, Devlin NJ, Odeyemi IA. 2006. "Yes", "No" or "Yes, but"? Multinomial modelling of NICE decision-making. *Health Policy* **77**: 352–367.
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**: 437–452.
Devlin N, Sussex J. 2011. Incorporating multiple criteria in HTA: methods and processes. Office of Health Economics: London.
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. *Medical Decision Making* **28**: 713–722.
Linley WG, Hughes DA. 2012. Reimbursement decisions of the All Wales Medicines Strategy Group: influence of policy and clinical and economic factors. *PharmacoEconomics* **30**: 779–794.
Littlejohns P, Rawlins M. 2009. Social value judgements: implementing the citizen's council reports. In Patients, the public and priorities in health care, Chapter 12, Littlejohns P, Rawlins M (eds). Radcliffe: London; 109–124.
Mason AR, Drummond MF. 2009. Public funding of new cancer drugs: Is NICE getting nastier? *European Journal of Cancer* **45**: 1188–1192.
McCabe M, Claxton K, Culyer AJ. 2008. The NICE cost-effectiveness threshold. What it is and what that means. *PharmacoEconomics* **26**: 733–744.

Mshelia I, White R, Mukku 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. 2005. Social value judgements: Principles for the development of NICE guidance. (Available from: http://www.nice.org.uk. Accessed 6th May 2009).

National Institute for Health and Clinical Excellence. 2008. Social value judgements: Principles for the development of NICE guidance (2nd edn). (Available from: http://www.nice.org.uk/media/C18/30/SVJ2PUBLICATION2008.pdf Accessed 13th November 2010).

National Institute for Health and Clinical Excellence. 2009. Appraising life extending, end of life treatments. Supplementary advice to the Appraisal Committees. (Available from: 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. 2010. NICE says "yes" to over 80% of treatments. (Available from: 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. 2013. Guide to the methods of technology appraisal. (Available from: http://www.nice.org.uk/media/D45/1E/GuideToMethodsTechnologyAppraisal2013.pdf Accessed 3rd October 2013).

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

Rawlins M, Culyer AJ. 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.

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**: 685–693.

White IR, Royston P, Wood AM. 2011. Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine* **30**: 377–399.

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

SUPPORTING INFORMATION

Additional supporting information may be found in the online version of this article at the publisher's web site.