BMJ Open

Assessment of quality of data submitted for NICE technology appraisals over two decades

Leeza Osipenko ,[1,2] Saba Ajwat Ul-Hasan,[1,2] Debra Winberg,[3] Kseniia Prudyus,[1] Marina Kousta,[4] Artemis Rizoglou,[1] Isabella Rustignoli,[1] Laurens van der Maas[1]

To cite: Osipenko L, Ul-Hasan SA, Winberg D, *et al*. Assessment of quality of data submitted for NICE technology appraisals over two decades. *BMJ Open* 2024;**14**:e074341. doi:10.1136/bmjopen-2023-074341

► Prepublication history and additional supplemental material for this paper are available online. To view these files, please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2023-074341).

Received 04 April 2023
Accepted 17 November 2023

[1]The London School of Economics and Political Science, London, UK
[2]Consilium Scientific, London, UK
[3]Tulane University School of Public Health and Tropical Medicine, New Orleans, Louisiana, USA
[4]King's College London, London, UK

Correspondence to
Dr Leeza Osipenko;
l.osipenko@lse.ac.uk

ABSTRACT

Background The National Institute for Health and Care Excellence (NICE) pioneered the Health Technology Assessment (HTA) processes and methodologies. Technology appraisals (TAs) focus on pharmaceutical products and clinical and economic data, which are presented by the product manufacturers to the NICE appraisal committee for decision-making. Uncertainty in data reduces the chance of a positive outcome from the HTA process or requires a higher discount.

Objective To investigate the quality of clinical data (comparator, quality of life (QoL), randomised controlled trials (RCTs) and overall quality of evidence) submitted by the manufacturers to NICE.

Design This retrospective evaluation analysed active TAs published between 2000 and 2019 (up to TA600).

Methods For all TAs, we extracted data from the Assessment Group and Evidence Review Group reports and Final Appraisal Determinations on (1) the quality of submitted RCTs and (2) the overall quality of evidence submitted for decision-making. For single TAs, we also extracted data and its critique on QoL and comparators. Each category was scored for quality and analysed using descriptive statistics.

Results 409 TAs were analysed (multiple technology appraisals (MTA)=104, single technology appraisal (STA)=305). In two-thirds of TAs, the overall quality of evidence was either poor (n=224, 55%) or unacceptable (n=41, 10%). In 39% (n=119) of the STAs, the quality of comparative evidence was considered poor, and in 17% (n=51) unacceptable. In 44% (n=135) of STAs, the quality of QoL data was considered poor, 15% (n=47) unacceptable, 33% (n=102) acceptable and 7% (n=21) as good. Over 20 years of longitudinal analysis did not show improvements in the quality of evidence submitted to NICE.

Conclusion We found that the primary components of clinical evidence influencing NICE's decision-making framework were of poor quality. It is essential to continue to generate robust clinical data for premarket and postmarket introduction of medicines into clinical practice to ensure they deliver benefits to patients.

STRENGTHS AND LIMITATIONS OF THIS STUDY

- ⇒ We analysed all active technology appraisals which were available on the NICE website in 2020 and which were published between 2000 and 2019.
- ⇒ We assessed critiques of key types of evidence submitted by the manufacturers to NICE as part of the technology appraisal process: comparator data, quality of life data and quality of randomised controlled trials.
- ⇒ Our analyses focused on the key elements but did not provide a full insight into all evidence parameters used in the Health Technology Assessment decision-making.
- ⇒ Our scoring system for grading the quality of evidence is subjective.
- ⇒ We make all raw data available in the public domain for review and re-analysis.

INTRODUCTION

Since the 1990s, evidence-based healthcare has been a pillar of regulations aimed at identifying successful practices and transitioning away from judgements made on the basis of belief or existing behaviour and toward a substantial emphasis on scientific research and knowledge.[1 2] The term 'evidence-based' has gradually evolved into 'evidence-informed' as the evidence-based medical approach has expanded beyond clinical care and into the larger arena of health policy.[3]

The International Network of Agencies for Health Technology Assessment[4] defines Health Technology Assessment (HTA) as a multidisciplinary process that uses explicit methods to determine the value of a health technology at different points in its lifecycle. The purpose is to inform decision-making in order to promote an equitable, efficient and high-quality health system. In England and Wales, the National Institute for Health and Care Excellence (NICE), a pioneer and a world leader in HTA,[5 6] advises the NHS on the clinical and cost-effectiveness of both new and existing technologies.[7] Between its establishment in 1999 and September 2020, NICE has considered 829 products for the technology appraisals (TAs) process across various disease categories.

The HTA process is predicated on the existence of robust evidence. Low data quality leads to uncertainty in the comparative clinical effectiveness and economic model outputs, and uncertainty reduces the chance of a positive outcome from the HTA process.[6] Walton *et al*[8] cited immature, limited or inappropriate clinical data (22.4%, n=15/67), which had been inadequately analysed or modelled as one of the three reasons for a negative preliminary decision.

NICE processes and methods

NICE describes the processes and methods it follows when carrying out TAs in its methods and process guides.[9] The two processes that NICE has adopted to appraise health technologies are multiple technology appraisal (MTA) and single technology appraisal (STA).[9] Before 2006, all technologies evaluated as part of NICE's TA programme were only considered through the MTA process.[2] The MTA approach is intended to evaluate multiple technologies that share one or more criteria.[10] For the MTA process, an independent academic team called the 'assessment group' (AG) is tasked with the responsibility of evaluating available evidence and providing a detailed report on the clinical and cost-effectiveness of the technologies (or indications) for use by the appraisal committee. The assessment report draws on evidence submitted by the manufacturer as well as the AG's systematic review of the literature, which provides an 'independent synthesis' of the existing data. Importantly, the AG has no pecuniary stake in the outcome of the analysis.[2 11]

The STA process was created in 2006 to evaluate a single product, device or technology for a single indication and usually involves new drugs or indications.[10] An independent academic team, the 'evidence review group' (ERG), undertakes 'a technical examination of the manufacturer's evidence submission'.[2] Even though the ERG may 'identify gaps in the evidence base,' NICE suggests that the manufacturer takes the main responsibility for evidence gathering and does not perform further analysis.[2] If the submitted evidence by the manufacturer is either inadequate or the decision problem is not properly defined, the AG/ERG seek further evidence from the manufacturer. The ERG/AG may undertake additional analyses, so-called exploratory analyses, to examine uncertainties around the company's model and its implications for decision-making.[12]

The appraisal committee is an independent body of specialists, who make decisions based on the ERG/AG reports and advice from consultees, clinical, NHS commissioning and patient experts. The appraisal committee examines and evaluates the evidence to determine if the technology should be considered a clinically beneficial and cost-effective use of NHS resources or if it should be approved for certain patient groups or used solely in research settings.[13] The appraisal committee present its final judgement in the form of a final appraisal determination (FAD) document,[8 10] which, among other information, details how it has assessed the submitted evidence, as well as the significant concerns raised by experts. Furthermore, if new data become available that is likely to alter the current recommendations, NICE will then update the published guidance.[13]

NICE commissions the Decision Support Unit, managed by the University of Sheffield, to provide guidance and support on programme evaluation, education, the development of advanced methodological approaches and economic evaluation to undertake assessments and appraisals of new and established health technologies. The Decision Support Unit also aids the independent advisory bodies of NICE in advanced analytical, methodological and other ad hoc approaches to assess these health technologies.[14]

Evidence requirements

Clinical practice varies across different countries. Often clinical trials do not deliver information on relative clinical effectiveness, which would be reflective of a given setting for a reimbursement decision. The NICE method of comparator selection emphasises inclusivity throughout the analytical stage and use in clinical practice in the NHS.[5]

NICE provides evidence-based guidance and traditionally has asserted that when analysing comparative clinical effectiveness, 'different types of study design can be ranked according to a hierarchy describing their relative validity,' with head-to-head randomised controlled trials (RCTs) ranked first and evidence originating from other study types serving as supplementary evidence.[15 16] However, in recent years, the breadth of NICE's evaluation has expanded significantly, as has its evidence base. NICE aims to expand the quantitative and qualitative knowledge base to encompass a broader range of factors that influence health and its distribution.[3] As a result, NICE considers registry data, national statistics, surveys, clinical practice recommendations, expert opinions and additional knowledge from manufacturers in addition to RCTs and observational research.[15 17]

Indirect comparisons

In manufacturers' submissions, NICE is frequently not presented with direct comparisons versus conventional NHS practices and recommends using standard methods for indirect comparisons on aggregate data,[9 16] with the central premise that the distribution of effect-modifying factors is consistent across trials.[18] These strategies employ individual patient data from a subset of trials to construct population-adjusted indirect comparisons between treatments in a specified target group.

Utility scores

As of 2004, NICE has developed the following recommendations for deriving utility values; utility values should be calculated from patients or caregivers, with health state valuation performed by the wider population using a 'standardised and validated generic choice-based instrument (time trade-off or standard gamble)', with the EuroQol 5-dimensions (EQ-5D) being the predominant mode of utility evaluation.[6 11 19] The rationale for suggesting a single verified instrument is that diverse measuring techniques may yield inconsistent findings. This measurement discrepancy might result in inaccurate decision-making, which can have a detrimental effect on population health. The 2013 methods guide[16] includes a checklist that specifically inquires whether the EQ-5D is employed. Even though the EuroQol Group developed a more sensitive five-level descriptive response system (EQ-5D-5L) than EQ-5D-3L, NICE did not recommend using the EQ-5D-5L value set in 2013, nor does it recommend using EQ-5D-5L in its latest update in 2022.[9]

NICE permits the use of instruments other than the EQ-5D. For example, in paediatric groups, the Child Health Utility 9D (CHU9D), the Assessment of Quality of Life 6-Dimension (AQoL-6D), and the adolescent and 16-dimensional assessment of HRQoL (16D) can be used. However, if there is a deficiency of child-specific data from generic preference-based instruments, a variety of practical alternatives are possible, including the use of adult utility data, the algorithmic modification of adult utility data to account for known demographic disparities and the mapping of disease-specific HRQoL tools.[6] In the 2022 update, NICE continues not to propose particular measures of HRQoL in the paediatric group. If data from a children's HRQoL instrument are used to calculate utility values, just a description of how this was done is necessary.

NICE accepts disease-specific instruments if data indicate that the EQ-5D is inappropriate.[9] In addition, when directly collected EQ-5D data are not available from the clinical trial or their quality is not acceptable, mapping to EQ-5D from other instruments is considered to be an appropriate methodological approach[9]; this converts data from other means of assessing the quality of life (QoL) to EQ-5D data and eliminates conventional uncertainty from valuation techniques.[11 19]

Aim of the study

The study aims to systematically review all active TAs published between 2000 and 2019 of NICE's work (up to TA600), investigate the critical issues with the clinical evidence submitted by the manufacturers and assess its quality for decision-making.

METHODS

Selection of technology appraisals

For this study, all active technology appraisals up to TA600 were searched on the NICE website. Active TAs refer to TAs (as of 2020) with a positive or a negative recommendation that has not been replaced. TAs were excluded if they were replaced through review, withdrawal, termination or where background documentation was unavailable on the NICE website. Original and updated TAs were assessed.

Active TAs were allocated to six researchers to extract data from publicly available documents on the NICE website. These documents included:

- FAD document
- AG report or ERG report

One designated researcher completed a freedom of information request for missing documents (n=6) and later acquired these documents through the UK Government web archive.

An Excel template listing significant appraisal features was used for data extraction.[20] Direct quotations were copied and pasted from the NICE documents onto the data extraction sheet, with no restriction on the number count to prevent loss of information. For each STA, the extracted data included evidence on three key components: (1) comparator data, (2) QoL data and (3) overall submitted evidence, which included the committee's and ERGs' conclusion on the manufacturer's submission and grading of the submitted RCTs by the ERG (if available). For MTAs, the only evidence on overall submitted evidence was extracted from the documents. We analysed oncology TAs separately to determine the quality of evidence for this group of therapeutic agents.

Data extraction

Information on TAs was extracted from the FAD and the AR/ERG report for each TA. Table 1 shows the data extracted from each TA and categories which were graded. Further information on the scoring criteria can be found in table 2.

Scoring quality of evidence

The information extracted from the public NICE documents for each category was used to score the quality of comparative clinical evidence, quality of QoL evidence, quality of submitted evidence, overall quality of evidence submitted for decision-making and quality of RCTs submitted to NICE. The grading was done across five categories using a standardised scoring guide (2=good, 1=acceptable, 0=poor, −1=unacceptable). Details of the scoring criteria are presented in table 2. If the documents did not contain the relevant information for a particular data point, not applicable (N/A) was awarded to that column on the datasheet.

Cross-validation

After data were extracted and the scores were assigned to all TAs, each researcher submitted their dataset to the project lead, who blinded the scoring and reissued the files for validation back to researchers making sure they received a different dataset. Each researcher reviewed the data, assigned the scores de novo and resubmitted the dataset to the project lead, who then unblinded the original values and matched these to the original values. If the values matched, this resulted in the final score for the given parameter; if the scores did not match, the original researcher and the validator discussed each issue to agree on the final score. If disparity in the grading remained,

Table 1 Summary of data collection

Category	
Technical appraisal information	▶ TA ID ▶ Sponsor ▶ Year of publication ▶ Process (STA/MTA) ▶ Technology ▶ Technology type ▶ Indication
Data	1. Comparator data – Comparator data summary from NICE documents. – Comparator data issues identified by the committee/ERG – Indirect comparison (yes/no) – Direct comparison (yes/no) – Randomised controlled trial presented (yes/no) 2. Quality of life data – QoL data summary from NICE documents – QoL data issues identified by the committee/ERG – List of QoL instruments used – QoL collected in the trial (yes/no) – QoL collected in dedicated study/ literature (yes/no) – Mapping used for utility values derivation (yes/no) 3. Overall submitted evidence – Critique of submitted evidence summary from NICE documentation. – ERG/AR AG proprietary analysis (yes/no)
Categories graded	▶ Quality of comparative evidence ▶ Quality of QoL evidence ▶ Quality of submitted evidence grade ▶ Overall quality of evidence for decision-making ▶ Quality of RCTs submitted

AG, assessment group; ERG, evidence review group; ID, Identification; MTA, multiple technology appraisal; NICE, National Institute for Health and Care Excellence; QoL, quality of life; RCTs, randomized controlled trials; STA, single technology appraisal; TA, technology appraisal.

the project lead reviewed the extracted data and made the judgement on the final score considering arguments from both the researcher and the validator.

Data analysis

Several steps were taken to ensure data quality. First, all six cross-validated Excel sheets created by the researchers were merged into one data frame using Python. The data frame was inspected for inconsistencies in column names and column entries, which were resolved programmatically. Afterwards, the file was imported to compare each category— active (not in research) non-oncology pharmaceutical STA; active (not in research) MTA non-oncology pharmaceutical appraisals; active (not in research) oncology STA and MTA appraisals; active in research (including Cancer Drugs Fund (CDF)) STA and MTA appraisals; active (not in research) STA and MTA non-pharmaceutical appraisals; and active (not in research) STA and MTA appraisals which received negative recommendations—to determine whether there were missing TAs.

Descriptive statistics were calculated, and outputs were plotted on a graph longitudinally. The analysed parameters were separated into two types: those with yes/no entries and those with a grade between −1 and 2 as entries. A function was created to present descriptive statistics for the yes/no columns and the grade columns, respectively. In turn, descriptive statistics for all columns of either the yes/no or grade type were created in absolute and relative numbers. Finally, a longitudinal analysis was carried out to plot a graph with information about the yes/no columns and grades contained in the merged data frame.

Patient and public involvement

None

RESULTS

Publicly available technology appraisals up to TA600 were identified. TAs are replaced over time, and the productivity of NICE has risen yearly since its inception. Therefore, while reviewing the outputs, it must be considered that there are significantly more TAs in later years than in previous years. For example, 27 TAs were completed between 2001 and 2005, 57 TAs between 2006 and 2010, 102 TAs between 2011 and 2015 and 223 TAs between 2016 and 2019.

Of the 600 TAs, 66 TAs were terminated, and 125 TAs were replaced and updated, receiving a new TA number and making the previous one unavailable. Four hundred and nine active TAs were selected for the analysis consisting of 305 STAs and 104 MTAs. The 409 active TAs comprised 25 non-pharmaceutical products (14 medical devices, six other therapeutic therapies, five surgical procedures) and 384 pharmaceutical products. The full list of TAs included in the analysis is accessible in the public repository[20]. There were no missing TAs, but 11 of them had specific documents such as FAD, AG or ERG reports missing; a freedom of information request was used to obtain these missing documents.

Based on analysis of the comments from the ERGs and AGs, across all examined TAs, over half of the RCTs presented in the manufacturers' submission were deemed to have either poor (n=166, 41%) or unacceptable (n=40, 10%) quality, while just under a half were considered to be of acceptable (n=173, 42%) or good (n=30, 7%) quality.

In more than 50% of TAs (n=226, 55%), the overall quality of evidence presented in the manufacturers' submission was scored by the researchers as poor, and 11% (n=43) were considered to have unacceptable quality. In one-third of TAs the researchers scored the quality of

Table 2 Scoring criteria

Grade	Category	Definition	Detailed description with examples from FADs or ERG/AG documents
2	Good	Good-quality data submitted, minimal modelling used	No critical words used, endorsements of data quality. For example: 'the analysis was sound', 'the quality of clinical trials submitted was generally good'
1	Acceptable	Acceptable quality data submitted, some modelling used (mapping or indirect comparison) and extrapolation, but modelling was assessed as robust	No or few critical words used, some concerns might be raised but committee accepted the data. For example: 'MTC was supported by a reasonably sound systematic review process but MTC has certain limitations in conduct and reporting, including'
0	Poor	Data submitted are from indirect sources, extensive modelling/ assumptions used – indirect comparison, mapping, extrapolation and quality of the modelling is questionable either due to data or methodology	Criticism is clearly expressed. For example: 'The manufacturer's use of indirect comparisons is inappropriate. The manufacturer's submission reported very limited data on the comparator trials, and did not undertake a systematic review of these', 'utility studies were missed in this review by failing to search databases such as Medline. The extent to which studies were missed is unknown'
−1	Non-acceptable	Poor quality of evidence is submitted, significant modelling is used, or poor quality of evidence is submitted and quality of modelling is poor	Harsh criticism expressed. For example: 'committee could not use presented evidence for decision-making', 'unacceptable', 'poor quality' are used

AG, assessment group; ERG, evidence review group; FAD, final appraisal determination; MTC, Mixed treatment comparison.

evidence submitted as acceptable (n=133, 33%), and only 7 (2%) TAs were graded as good-quality submissions. The ERGs/AGs conducted proprietary (in-house) analyses on under a third of the analysed TAs (n=120, 29%). The researchers evaluated the overall quality of the evidence presented to the appraisal committee for decision-making based on the above three domains (quality of RCT, quality of evidence in the manufacturers' submission and the propriety analysis) as well as the comparative and QoL data evidence submitted (applicable to STAs only). The analysis showed that in two-thirds of the TAs (n=265), the evidence provided to the appraisal committee for decision-making was either of poor (n=224, 55%) or unacceptable (n=41, 10 %) quality, while in over one-third of TAs, the quality of evidence was acceptable (n=139, 34 %) or good (n=5, 1%). Therefore, as longitudinal analysis over 20 years shows, the overall quality of clinical evidence submitted to NICE did not improve (figure 1).

Analysis of STAs (n=305)

According to the judgement from the ERGs, over half of the RCTs submitted for STAs were either of acceptable (n=149, 49%) or good (n=23, 8%) quality. However, over two-fifths of the RCTs were of either poor (n=111, 36%) or unacceptable (n=22, 7%) quality. When analysing the comparator data in STAs, 101 (33%) conducted more than one type of comparison (direct and indirect). Comparative evidence was scored by the researchers to be either poor in 119 (39%) STAs or unacceptable in 51 (17%) STAs. In 120 (39%) STAs, the quality of comparative evidence was considered acceptable, and 15 (5%) STAs were deemed to have good quality (figure 2). We found no obvious trends for improvement in the quality of comparative evidence over time as demonstrated by longitudinal analysis in Figure 2.

In almost three-quarters of STAs (n=227, 74%), QoL data were collected in the pivotal trials of the investigational products. In a quarter of STAs (n=78, 26%) QoL data were unavailable from the trials. In these appraisals, QoL data were either collected from other sources (dedicated QoL studies/literature; n=73, 24%) or were not collected at all (n=10, 3%). A dedicated study here refers to research that is conducted for the purpose of advancing knowledge of QoL. Mapping was used in 88 (29%) STAs, where QoL data were collected using tools other than EQ-5D. The researchers scored the majority of STAs (n=182, 60%) as either presenting poor (n=135, 44 %) or unacceptable (n=47, 16%) quality of QoL data. While 123 (40%) STAs were scored by the researchers as having presented either acceptable (n=102, 33%) or good (n=21, 7%) quality of QoL data.

The overall quality of the submitted evidence by the manufacturer was poor in the majority of STAs (n=165, 54%) or unacceptable (n=21, 7%). One hundred and fifteen (38%) STAs were scored to have evidence submitted by the manufacturer as acceptable, and 4 (1%) STAs were graded as having good-quality evidence (figure 3). The ERG conducted a proprietary analysis on 94 (31%) STAs. The proprietary analysis did not have much impact on the overall quality of evidence submitted to the appraisal committee for decision-making. For three STAs, the overall quality of evidence was upgraded from unacceptable to poor, and two STAs were downgraded to poor quality from good-quality evidence.

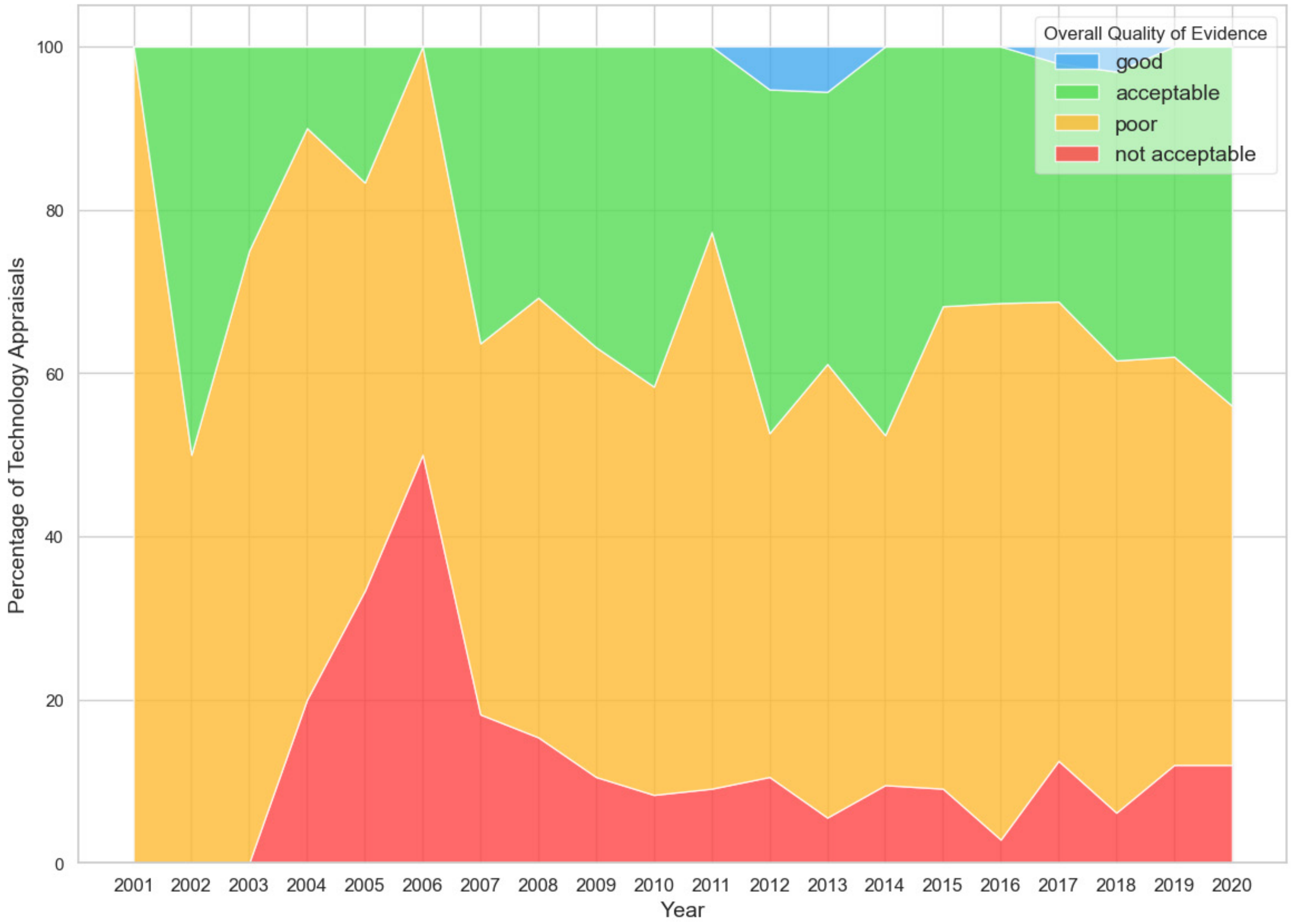

Figure 1 Overall quality of evidence submitted to the appraisal committee.

Analysis of MTAs (n=104)

The AGs judged the majority of the RCTs that provided evidence for MTAs as being of poor (n=55, 53%) or unacceptable (n=18, 17%) quality and slightly more than a quarter of RCTs as acceptable (n=24, 23%) or good quality (n=7, 7%). The project researchers graded the overall quality of evidence provided by the manufacturer to be either poor (n=59, 57%) or unacceptable (n=20, 19%), and 24 (23%) as acceptable or good (n=1, 1%) quality (figure 3).

Analysis of TAs in oncology

We identified 135 TAs with positive recommendations in oncology: 112 STAs and 23 MTAs. Thirty-five TAs were recommended as in research only (including medicines that entered the CDF). Of the recommended oncology TAs, two-thirds (n=90, 67%) of the evidence submitted by the manufacturer was graded as unacceptable (n=73, 54%) or poor (n=17, 13%) quality by the researchers. For 45 oncology TAs, the evidence submitted by the manufacturer was graded as acceptable in 44 (33%) and in 1 (1%) as good quality.

Of the 112 STAs in oncology, 43 (38%) conducted a direct comparison only, and 41 (37%) conducted an indirect comparison only; 28 (25%) STAs relied on both direct and indirect comparisons as the scope listed many comparators. The comparator evidence was graded as poor (n=51, 46%) or unacceptable (n=9, 8%) in over half of oncology STAs (n=60, 54%). While fewer than half of STAs (n=52, 46%) were graded as acceptable (n=46, 41%) or good (n=6, 5%) quality for comparator data.

The vast number of oncology TAs (n=105, 78%) produced evidence from RCTs, and according to the statements made by the ERG, the researchers rated 71 RCTs (68%) as of either poor (n=61, 58%) or unacceptable (n=10, 9.5%) quality. The remainder were rated as acceptable (n=59, 56%) or good (n=5, 4.8%).

More than two-thirds of oncology STAs collected QoL data (n=80, 71%); just over a quarter of STAs (n=30, 27%) gathered QoL data from other sources (dedicated QoL studies/literature), while less than 4% (n=5) of STAs did not submit any QoL data. In 30 (27%) STAs, mapping to EQ-5D from other QoL instruments was used. The majority of QoL data quality was rated as either poor (n=44, 33%) or unacceptable (n=26, 19%), and more than a third of the QoL data quality (n=42, 31%) was

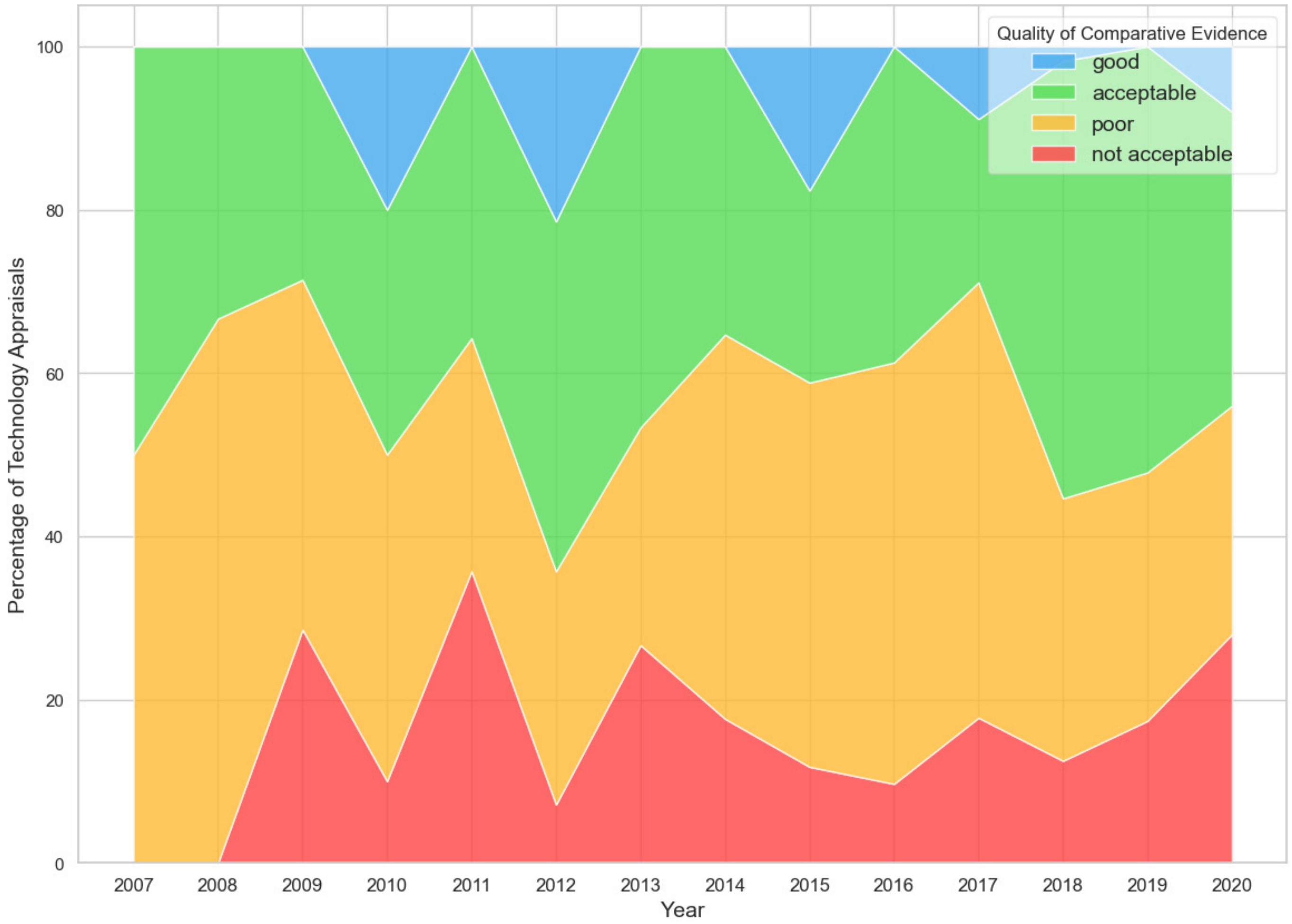

Figure 2 Quality of comparative evidence in the single technology appraisal.

judged as either acceptable (n=36, 27%) or good (n=6, 4%) (figure 4).

DISCUSSION

Our longitudinal analysis that ranged from 2000 to 2019 showed that the quality of evidence submitted to NICE by the manufacturers as part of the technology appraisal process did not show improvements, and in over 50% of the appraisals it was judged as poor. The specific issues were:

1. A lack of clarity on the methodologies applied by the manufacturers while performing systematic reviews and indirect comparisons.
2. Comparator data often did not reflect the UK population and routine treatment pathways. Indirect comparisons were used in 68% (n=207) of STAs to establish the comparative clinical effectiveness of interventions.
3. The QoL data was often of poor or unacceptable quality, even if collected in pivotal trials; clarity in reporting methodology and details by both manufacturers and assessment bodies varied significantly.

Our analysis is consistent with findings from a study conducted by Kaltenthaler *et al*,[21] which examined widespread issues reported by the ERGs in their assessment of manufacturers' submission to NICE in the STA process. Over 10 years ago, Kaltenthaler *et al*[21] found that 'much can be done to improve the quality of manufacturers' submission… including the need for clear and transparent reporting of methods and analyses.' Based on our findings, over the past decade, the situation has not improved. A more recent analysis by Walton *et al*[8] revealed that the appraisal committee judgement was complicated, noting 'immature, limited or inappropriate clinical data that had been inadequately analysed or modelled.' The low quality of the manufacturers' submissions resulted in two significant problems. First, the HTA process was prolonged because of the request for more information or clarification from manufacturers and further assessment from the ERGs.[8 21] Second, the AGs/ERGs and the appraisal committee might not thoroughly investigate the significant concerns, which could affect the decisions taken.[21]

Our analysis revealed that over half of the RCT data used in TAs was of poor or unacceptable quality. We discovered that weak or insufficient evidence from poorly conducted RCTs was often used since it was the only

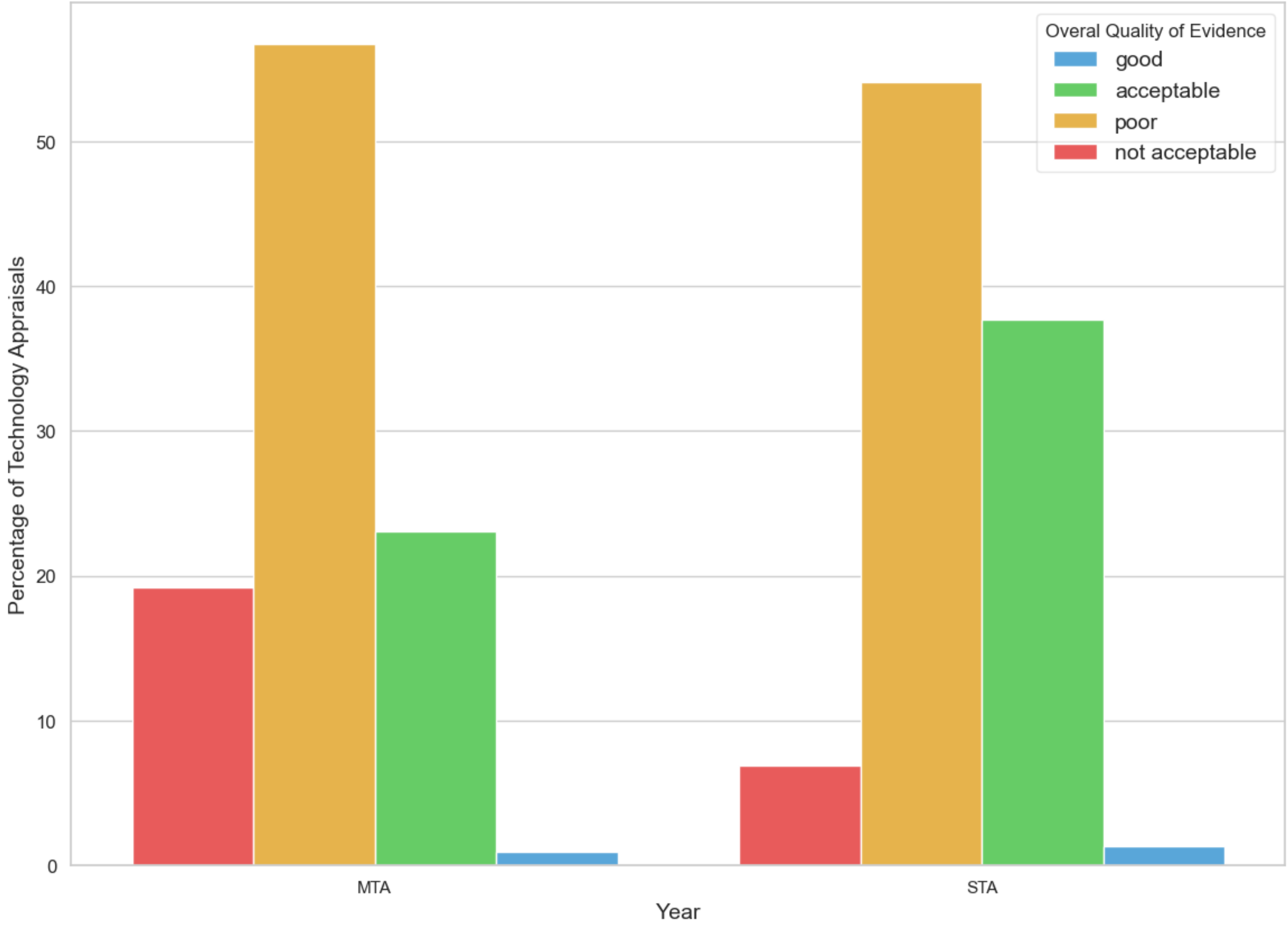

Figure 3 Comparison of overall quality of evidence submissions between multiple technology appraisals and single technology appraisal.

available evidence. When RCTs were executed rigorously and the evidence was comprehensive, the comparators were often unsuitable for decision-making in the NHS context. As a result, measuring the real impact was challenging and complicated the decision-making process for the appraisal committee. Political and institutional demands, combined with rationality, make it difficult for the appraisal committee to avoid making decisions based on inconclusive evidence. A study by Charlton[2] showed that the use of evidence by NICE has evolved, beginning with the admission of non-randomised and indirect study designs, increased frequency and breadth of subgroup analysis and reduced evidence requirements for cancer therapies. NICE still favours RCTs, but it also does not limit quantitative and qualitative evidence, as it recognises that health is influenced by a more significant number of variables and distributions.[3] The new process and methods guide[9] gives further flexibility in using non-randomised evidence for decision-making.

According to our findings, most STAs relied on QoL data collected in pivotal trials rather than literature. However, the QoL data were mostly poor or unacceptable in quality. Earlier research[22 23] has demonstrated that reliable QoL data are scarce for cancer drugs in European HTA processes and that the lack of comprehensive QoL data had little effect on the recommendations.

The exploratory analyses conducted by the ERG seem to alleviate some of the deficiencies and ambiguity in the evidence submitted by the manufacturer and provide enough evidence for the appraisal committee to make a sound judgement on the technologies.[10] Carroll *et al*[12] evaluated the nature and significance of exploratory analyses undertaken by the ERGs on 100 STAs and discovered that the types of exploratory analyses undertaken on the company models were 'fixing errors, addressing violations, addressing matters of judgement and the provision of a new, ERG-preferred base case'. Carroll *et al*[12] concluded that these analyses are 'highly influential in the policy-making and decision-making process,' and it seems that using ERG reports in the STA process is greatly effective, suitable and transparent, although resource and time-intensive. Our study, however, revealed that the AGs/ERGs conducted propriety analyses on fewer than a third of the TAs, but this did not appear to enhance the quality of evidence submitted to the appraisal committee for decision-making.

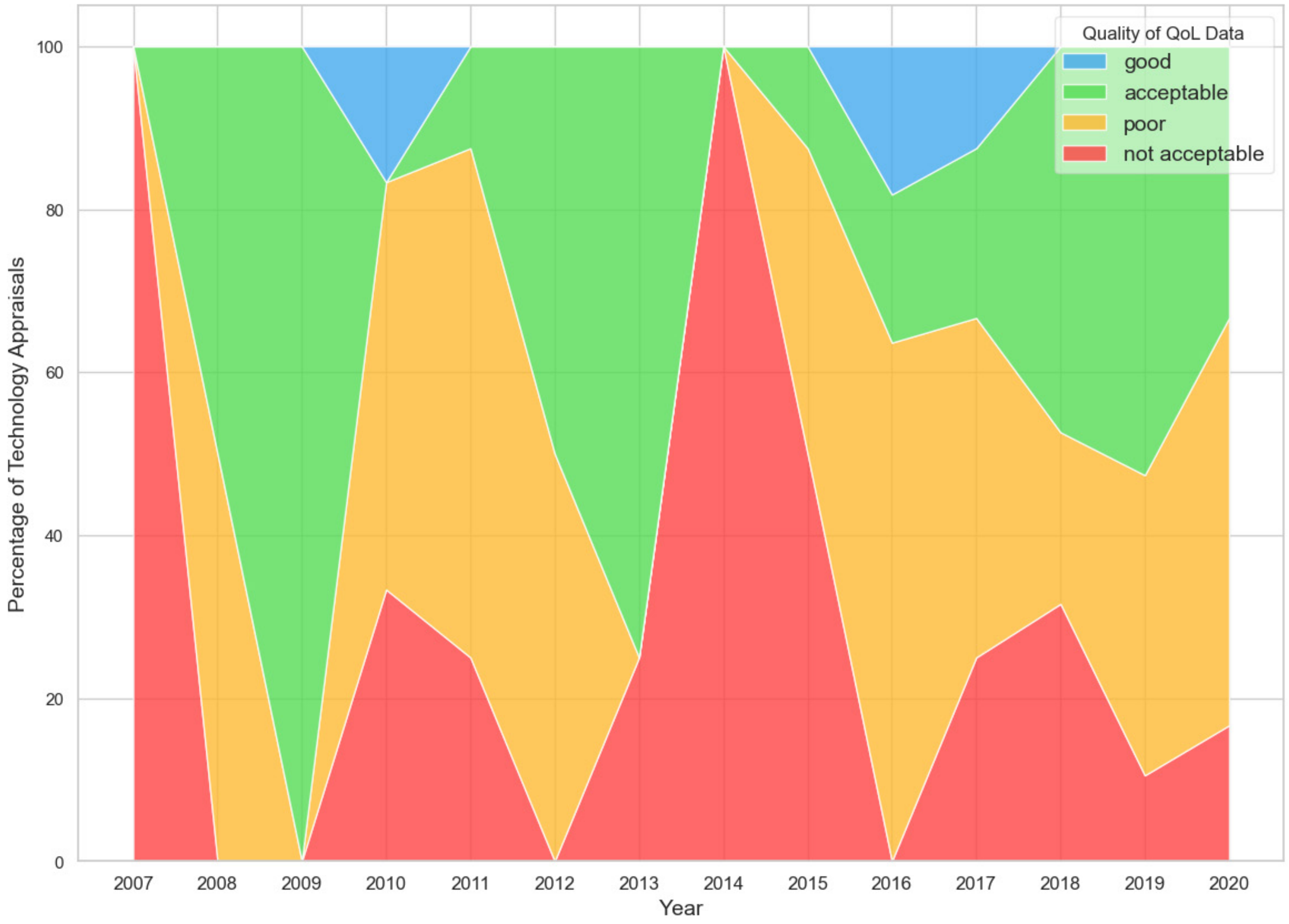

Figure 4 Quality of quality of life (QoL) data presented in oncology technology appraisals.

In January 2022, NICE announced changes to its procedures and methodologies for evaluating health technologies to improve patient access to innovative treatments. The Innovative Medicines Fund (IMF) was launched in June 2022 building on the success of the CDF. The IMF will support faster access to non-cancer drugs and, alongside the CDF, provide a total of £680 million ring-fenced NHS funding for innovative medicines. In addition, like the CDF, it is envisaged that the IMF will assist in collecting evidence for novel technologies when additional data are required to resolve uncertainties in their evidence base. The extra information collected over time will enable NICE to make a final recommendation on the use of the drug in the evaluation process.[24–26] It is important to ensure that data collected via IMF to resolve uncertainties are robust and actually answer the questions which were raised at the initial reimbursement review.

While postmarket data is essential to resolve uncertainties in evidence on novel health technologies and acquire additional knowledge on the use of these products in the real-world setting, we would like to emphasise the importance of the efforts to maximise relevant data generation in the premarket setting. NICE has been offering Scientific Advice to product developers since 2011 encouraging companies to think carefully about their development plans, design high-quality trials and collect robust data for the appraisal process at NICE. With the European HTA bodies, the EUnetHTA 21 initiative is establishing and piloting methods and processes for the joint scientific consultation and the joint clinical assessment,as outlined in the HTA assessment.[27] Such efforts are important for bringing about improved quality of evidence for the technology appraisal processes by HTA bodies.

Numerous studies have raised alarm about lowering the bar for evidence requirements for regulatory approval of medicinal products,[28–30] which leads to many expensive drugs with low or no efficacy entering the market. This problem propagates into the HTA decision-making, requiring more assumptions and modelling steps, which increase uncertainties for the appraisal committee. NICE's objective is 'to consider uncertainty appropriately and manage the risks to patients and the NHS while preventing inappropriate barriers to valuable innovations'.[31] However, in the new methods and process guide, NICE[9] has granted the appraisal committee more flexibility when examining technologies for which it is fundamentally hard to obtain adequate clinical data (eg, paediatric indications or complex cases). These new measures continue to provide further degrees of freedom to manufacturers, and the evidentiary threshold for decision-making at the regulatory and HTA levels continues to decline.[2] The resulting increased freedom and the use of additional powers by the appraisal committee will be the determining factor. As researchers examine the outcomes of evaluations under this new system, we are yet to see if these reforms will have the intended impact.

In our analysis we did not observe decreased quality of evidence in submissions over time, rather it has been found to be consistently poor. It may be that the quality of evidence is decreasing but it requires a different type of analysis to establish this. In our sample we have an uneven distribution of the number of appraisals per year, significantly skewed to later years as we looked only at active TAs. Going into the government archives and examining all TAs that have been replaced might be of interest.

The key limitation of our work is the subjective scoring system for grading the quality of evidence. We provide all data in a public repository so that anyone can review and challenge our scoring conclusions. Furthermore, we examined selected elements of evidence, such as QoL, comparator data and the quality of RCTs submitted by the manufacturers. To gain additional insight into the quality of evidence being generated for reimbursement decisions, it could be of interest to conduct further studies examining other components of manufacturers' submissions.

CONCLUSION

We found that the primary components of clinical evidence (comparative clinical effectiveness, measures of QoL outcomes and overall design of RCTs) that influence patients and are crucial for NICE's decision-making framework are of poor quality. Since the evidence bar continues to be lowered, it is essential to have HTA bodies and payers' input to ensure that the generation of evidence submitted to NICE is strengthened. However, it is essential that stakeholders are aware of this and that organisations put more effort into generating high-quality evidence premarket and postmarket entry. Furthermore, it is important that NICE reverts to issuing recommendations where data needs must be enhanced to ensure that this evidence generation is robust and patient relevant.

Contributors LO contributed to the study design. SAU-H, DW, KP, IR, AR, and MK contributed to acquisition of the data. LvdM performed the data analyses. SUL contributed to interpretation of the data. SAU-H and LO contributed to writing the original version of the manuscript. LO and DW contributed to the critical revision of the manuscript. All authors read and approved the final manuscript. LO is the guarantor. LO accepts full responsibility for the work and conduct of the study, had access to the data, and controlled the decision to publish. LO attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. Transparency: LO (the guarantor) affirms that the manuscript is an honest, accurate, and transparent account of the study being reported; that no important aspects of the study have been omitted; any that any discrepancies from the study as planned have been explained.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Competing interests None declared.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Patient consent for publication Not applicable.

Ethics approval Not applicable.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available in a public, open access repository. All data used for this research are publicly available and can be rechecked by third parties. Data that we extracted for analysis can be accessed through a public repository zenodo.org.

ORCID iD

Leeza Osipenko http://orcid.org/0000-0002-7758-509X

REFERENCES

1 Walker S, Palmer S, Sculpher M. The role of NICE technology appraisal in NHS rationing. *Br Med Bull* 2007;81–82:51–64.
2 Charlton V. Health technology assessment policy under the UK's National Institute for Health and Care Excellence, 1999-2018. *Health Care Anal* 2020;28:193–227.
3 Culyer A, Rawlins M. Evidence and quality, practicalities and judgments: some experience from NICE. *Healthc Q* 2012;15 Spec No:66–9.
4 The International Network of Agencies for Health Technology Assessment (INAHTA). Available: https://www.inahta.org/ [Accessed 08 Mar 2023].
5 Stevens AJ, Longson C. At the center of health care policy making: the use of health technology assessment at NICE. *Med Decis Making* 2013;33:320–4.
6 Montgomery SM, Kusel J. The prevalence of child-specific utilities in NICE appraisals for paediatric indications: rise of the economic orphans? *Expert Rev Pharmacoecon Outcomes Res* 2016;16:347–50.
7 The National Institute for Health and Care Excellence (NICE). Available: https://www.nice.org.uk [Accessed 08 Mar 2023].
8 Walton MJ, O'Connor J, Carroll C, *et al*. A review of issues affecting the efficiency of decision making in the NICE single technology appraisal process. *Pharmacoecon Open* 2019;3:403–10.
9 The National Institute for Health and Care Excellence (NICE) health technology evaluations: the manual process and methods [PMG36]. 2022. Available: https://www.nice.org.uk/process/pmg36/chapter/introduction-to-health-technology-evaluation [Accessed 08 Mar 2023].
10 Kaltenthaler E, Carroll C, Hill-McManus D, *et al*. The use of exploratory analyses within the National Institute for Health and Care Excellence single technology appraisal process: an evaluation and qualitative analysis. *Health Technol Assess* 2016;20:1–48.
11 Tosh JC, Longworth LJ, George E. Utility values in National Institute for Health and Clinical Excellence (NICE) technology appraisals. *Value Health* 2011;14:102–9.
12 Carroll C, Kaltenthaler E, Hill-McManus D, *et al*. The type and impact of evidence review group exploratory analyses in the NICE single technology appraisal process. *Value Health* 2017;20:785–91.
13 The National Institute for Health and Care Excellence (NICE) guide to the processes of technology appraisal. 2018. Available: https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/technology-appraisal-processes-guide-apr-2018.pdf [Accessed 08 Mar 2023].
14 University of Sheffield, NICE decision support unit. Available: https://www.sheffield.ac.uk/nice-dsu [Accessed 08 Mar 2023].
15 Angelis A, Lange A, Kanavos P. Using health technology assessment to assess the value of new medicines: results of a systematic review and expert consultation across eight European countries. *Eur J Health Econ* 2018;19:123–52.
16 The National Institute for Health and Care Excellence (NICE) Guide to the methods of technology appraisal 2013: the manual process and methods [PMG09]. 2013. Available: https://www.nice.org.uk/process/pmg9/chapter/foreword

17 The National Institute for Health and Care Excellence (NICE) real-world evidence framework, Corporate document (ECD9). 2022. Available: https://www.nice.org.uk/corporate/ecd9/chapter/overview [Accessed 08 Mar 2023].

18 Phillippo DM, Ades AE, Dias S, *et al*. Methods for population-adjusted indirect comparisons in health technology appraisal. *Med Decis Making* 2018;38:200–11.

19 Rose M, Rice S, Craig D. Does methodological guidance produce consistency? A review of methodological consistency in breast cancer utility value measurement in NICE single technology appraisals. *Pharmacoecon Open* 2018;2:97–107.

20 Osipenko L. Data for quality of evidence submitted to NICE technology appraisals (2000-2020) [Zenodo Repository]. 2023. Available: https://zenodo.org/record/7674913

21 Kaltenthaler EC, Dickson R, Boland A, *et al*. A qualitative study of manufacturers' submissions to the UK NICE single technology appraisal process. *BMJ Open* 2012;2:e000562.

22 Kleijnen S, Leonardo Alves T, Meijboom K, *et al*. The impact of quality-of-life data in relative effectiveness assessments of new anti-cancer drugs in European countries. *Qual Life Res* 2017;26:2479–88.

23 Shah KK, Mestre-Ferrandiz J, Towse A, *et al*. A review of health technology appraisals: case studies in oncology. *Int J Technol Assess Health Care* 2013;29:101–9.

24 Whitehead V, Catchpole P. Briefing on the innovative medicines fund [The Association of the British Pharmaceutical Industry]. 2022. Available: https://www.abpi.org.uk/publications/briefing-on-the-innovative-medicines-fund/

25 NHS England. Innovative medicine fund. Available: https://www.england.nhs.uk/medicines-2/innovative-medicines-fund/ [Accessed 08 Mar 2023].

26 Roberts M. Innovative medicines fund launched to fast-tract drugs [BBC]. 2022. Available: https://www.bbc.co.uk/news/health-61709542 [Accessed 08 Mar 2023].

27 Hulstaert F, Pouppez C, Primus-de Jong C, *et al*. Gaps in the evidence underpinning high-risk medical devices in Europe at market entry, and potential solutions. *Orphanet J Rare Dis* 2023;18:212.

28 Prasad V. *Malignant: How Bad Policy and Bad Evidence Harm People with Cancer, Edition 1*. Baltimore, Maryland: Johns Hopkins University Press, 2020.

29 Schnog J-J, Samson MJ, Gans ROB, *et al*. An urgent call to raise the bar in oncology. *Br J Cancer* 2021;125:1477–85.

30 Gyawali B, Rome BN, Kesselheim AS. Regulatory and clinical consequences of negative confirmatory trials of accelerated approval cancer drugs: retrospective observational study. *BMJ* 2021;374:n1959.

31 NICE signals commitment to greater flexibility in its evaluation of promising new health technologies and making patient access fairer. 2022. Available: https://www.nice.org.uk/news/article/nice-signals-commitment-to-greater-flexibility-in-its-evaluation-of-promising-new-health-technologies-and-making-patient-access-fairer [Accessed 08 Mar 2023].