Chapter 10

NICE's Cost-Effectiveness Threshold

How We Learned to Stop Worrying and (Almost) Love the £20,000–£30,000/QALY Figure

Gabriele Badano, Stephen John and, Trenholme Junghans

The National Institute for Health and Care Excellence (NICE) is a public body working at arm's length from the Department of Health in England. It is well known for its health technology appraisal (HTA) process, through which it assesses whether new drugs and other health technologies should be used in the National Health Service (NHS). If NICE recommends the use of a certain health technology, then clinical commissioning groups (i.e., local authorities responsible for allocating the NHS budget) are legally bound to fund it. These decisions can be extremely controversial. For example, in November 2015, NICE advised clinical commissioning groups against purchasing Kadcyla, a drug for late-stage terminal breast cancer sufferers. Although Kadcyla offers as much as an extra six months of life, it is highly costly, at around £90,000 for a fourteen-month treatment. This hefty price tag places Kadcyla well beyond the maximum amount of money that according to NICE, the NHS should pay per life year saved (NICE 2015). Many stakeholders expressed anger at this decision, with a representative of the charity Breast Cancer demanding a change to funding arrangements because "people living with incurable cancer don't have time to lose" (Boseley 2015).

What is the process behind this and many other controversial decisions by NICE? To decide whether to recommend a health technology, NICE collects evidence regarding its clinical effectiveness. The Quality-Adjusted Life Year (QALY), which integrates gains in life expectancy with improvements in quality of life, is NICE's measure of choice for determining the health benefits that a course of intervention can provide. The evidence about clinical effectiveness is brought together with evidence about financial costs to calculate the incremental cost-effectiveness ratio (ICER) of the technology

McClimans_9781783488476.indb 151 26-04-2017 20:23:15

(i.e., the additional cost of an additional QALY that the NHS would gain by using such technology compared to the health technology that the NHS currently employs for the same purpose). As a simplified example, imagine that a course of the current standard treatment for a certain form of terminal cancer costs £10,000 and its only effect is that, on average, it extends life by six months, without any improvement in quality of life. A new intervention could replace it, offering an average of twelve months of life extension to each patient at the cost of £15,000 per course of treatment. In this case, the ICER of the new treatment would be £5,000 (£15,000 minus £10,000) per six (twelve minus six) quality-adjusted months, or £10,000/QALY.

A key step in NICE's decision-making process is the comparison between the ICER of the technology under appraisal with a cost-effectiveness threshold of £20,000–£30,000/QALY. Indeed, NICE explains that it is unlikely to reject any technology whose ICER lies below £20,000/QALY. If the ICER falls above £20,000/QALY, NICE's committees must reach beyond cost effectiveness and consider factors including the so-called "equity weightings" (severity of target disease, the innovative nature of the technology, extra priority to be assigned to end-of-life care, a premium to be placed on the treatment of diseases that disproportionately affect children or members of disadvantaged social groups). If the ICER of the technology under appraisal is between £20,000 and £30,000 per QALY, some of these factors must lend support to its use for such technology to be approved by NICE. Above the £30,000/QALY mark, an exceptionally strong case must be built on such factors if decision makers wish to recommend the technology despite its high ICE.[1]

Although NICE is not at liberty to disclose the exact figure, Kadcyla was rejected because of the large gap between its ICER, which falls in the region of £160,000/QALY, and the upper end of NICE's £20,000–£30,000/QALY threshold (NICE 2015). The public anger caused by NICE's use of its threshold to reject drugs is familiar. Generally, economists and ethicists also think that this anger is misguided, at least in one important sense. They argue that given that the resources devoted to health care are finite, the medical needs of our societies are virtually endless, and there is an extremely broad range of beneficial interventions available, there must be some beneficial drugs that health care providers will not purchase. Moreover, no decision to fund a new drug can simply be based on its clinical effectiveness, or in other words, on the fact that it would do good to patients. The clinical benefits that funding that drug would provide to patients must be compared with its "opportunity cost" (i.e., the clinical benefits to other patients that would have to be forgone by diverting the necessary funds from somewhere else in the NHS). One obvious way in which to do this is by comparing treatments in terms of how many QALYs they would generate for the money spent on them and then allocating funds based on a strong concern for cost effectiveness.[2] None of

this is to say, of course, that the specific £20,000–£30,000/QALY threshold used by NICE—and, hence, the Kadcyla decision—is correct, nor is it to say that cost effectiveness is the only relevant consideration in this, or other cases of, just resource allocation. It is, however, to say that there must be something like a cost-effectiveness threshold beyond which the purchase of a drug will become increasingly unlikely even if this means that those without time to lose avoidably lose that time. In other words, the consensus among economists and ethicists is that one possible critique of NICE's £20,000–£30,000/QALY threshold—that it aims to measure something that should not concern resource allocation agencies in the first place—is off the table.

In this paper, we study the history of NICE's threshold as a way of investigating the additional steps necessary for a complete critical assessment of this threshold. We do this from the perspective of both ethics and philosophy of science. More specifically, our aim is twofold. In the first two sections (respectively, "The Threshold: The Theory" and "The Threshold: The Practice"), we engage in a close study of the theory and history of NICE's work. We argue that viewed as an attempt to measure its stated construct, the "opportunity cost" associated with funding a treatment, NICE's threshold—one of the most important measures in British public life, which quite literally determines life-and-death decisions—is deeply flawed. Therefore, NICE's £/QALY threshold badly measures its stated concern with cost effectiveness. But in the third section ("Justifying the Threshold?"), we argue that close attention to the complex institutional and political context of NICE's work suggests a more nuanced understanding of the role of this threshold not only as a measure of the independent quantity that it explicitly sets to target but also as a standard that serves to promote other important goods.

It is commonplace that the adequacy of some measures is related to moral and political ends and therefore that we cannot properly assess these measures without engaging in ethical debates. For example, one could challenge NICE's threshold by questioning the value judgment that NICE should be concerned with cost effectiveness at all. However, reflection on themes in philosophy, sociology, and anthropology of measurement allows us to consider the broader political context of these measures and the roles they might play beyond their stated ends. In turn, we can ask distinctively evaluative questions about those roles. This perspective points toward a way in which our assessment of measures in domains, such as health policy, must be sensitive to moral and political, as well as epistemological, questions. In this paper, we do not aim to say the last word about whether NICE's £20,000–£30,000/QALY threshold is justified. However, we certainly wish to highlight the sheer complexity of assessing this question and the ways in which even apparently flawed measures may do important political work.

THE THRESHOLD: THE THEORY

In philosophy, as well as in other disciplines, much attention has been paid to QALYs as a measure of health benefits.[3] The focus of this chapter is different in that we are interested in NICE's cost-effectiveness threshold quite independently of the choice of QALYs as its health outcome measure. Specifically, we are interested in the threshold's history and how it can be evaluated. To set the stage for our evaluation, this section explores the theory behind NICE's use of a cost-effectiveness threshold and the way in which theorists normally debate these thresholds.

In principle, NICE's £/QALY threshold is a measure of opportunity costs. However, before exploring this understanding, it is worth mentioning that an alternative view exists. According to the social-value view, NICE's threshold should measure the value that the British society at large attaches to one QALY's worth of health gain—a value that is given by the amount of money that the members of society are willing to pay to produce one extra QALY. Although commentators have sometimes used this approach to explain the meaning of NICE's threshold (Smith and Richardson 2005), this explanation suffers from serious flaws. An integral part of the social-value view is that the NHS should pay for all and only those interventions that produce a QALY for a cost that is equal or inferior to what has been found to be society's willingness to pay for it. This proposal suggests that NICE effectively determines the level at which the overall NHS budget should be set simply by setting its threshold. More realistically, the size of the health care budget is understood as an issue for parliamentary debate, to be settled based on a richer set of considerations than willingness to pay (McCabe, Claxton, and Culyer 2008, 735–36).

Indeed, NICE itself acknowledges that setting the overall level of public spending for health care is not its job and formally endorses an opportunity-costs understanding of its cost-effectiveness threshold. For example, in the latest edition of its *Guide to the Methods of Technology Appraisal*, NICE claims that a health technology is to be considered cost effective "if its health benefits are greater than *the opportunity costs* of programmes displaced to fund the new technology, in the context of a fixed NHS budget" (NICE 2013, 14; italics added). Moreover, this view of cost effectiveness is grounded in the acknowledgment that available NHS resources are limited. Every recommendation that NICE makes in favor of a new technology that is costlier than the one adopted so far for the same use requires disinvestment somewhere else in the NHS. In this context, it is not enough that the new technology offers greater health benefits than the currently funded alternative; it is also important that the extra health benefits that would be obtained by commissioning it outweigh its "opportunity costs" (i.e., the health benefits that would

be lost through disinvestment across the NHS). Very roughly, if a new treatment would cost an extra £40,000 to produce an extra QALY, this £40,000 cannot be used for other treatments, which might well produce more than one QALY (albeit, obviously, for different people). Consequently, it is important that the ICER of new technologies falls below the cost of a QALY produced through the least cost-effective intervention currently funded by the NHS. The £/QALY threshold is supposed to measure this cost (McCabe et al. 2008, 737–78).[4]

An interesting implication of this view of cost-effectiveness thresholds is that such thresholds should be frequently updated. In principle, if less cost-effective technologies are displaced over time by more cost-effective ones, NICE's £/QALY threshold should move downward and become more restrictive. Also, any real-term increase (or decrease) in the size of the health care budget should lead to a higher (or a lower) threshold.

The theory behind NICE's cost-effectiveness threshold is premised on the idea that a key goal of health care resource allocation decision makers is to use available funds to improve the aggregate health of the population, measured in terms of QALYs, as much as possible. In the philosophical debate over the ethics of health care resource allocation, there is broad consensus on the importance of this goal and, therefore, on the importance of measuring the opportunity costs of interventions in terms of displaced health benefits. However, there is also broad consensus on the idea that it is legitimate to pursue the goal of QALY maximization across society only under certain constraints, which express the importance of *who* receives the QALYs that are produced. For example, it is often argued that there are cases in which NICE and other health care resource allocation agencies should decide in favor of an intervention that they know will displace more QALYs than it will generate because this intervention will benefit patients who are particularly badly off (e.g., in terms of severity of illness, socioeconomic status, age).[5] These quintessentially distributive concerns are reflected in the equity weightings that we described in the introduction and that are balanced by NICE against the cost effectiveness (or lack thereof) of the technologies under appraisal.

THE THRESHOLD: THE PRACTICE

As explained in the previous section, a £/QALY threshold is meant to measure something rather specific—the cost of a QALY produced through the least cost-effective intervention currently funded by the NHS. We also saw that at least formally, NICE accepts this idea of what the cost-effectiveness threshold is supposed to measure.

In principle, measuring the level at which NICE's threshold should be set would require a complex empirical research project, estimating the cost of a QALY in the various areas of treatment and prevention currently covered by the NHS. When NICE was created in 1999, however, it had no evidence suitable for making such an estimate. Therefore, although the mandate it had received from the Secretary of State required NICE to take the ratio of costs to benefits of heath technologies into account and NICE's committees asked producers to provide £/QALY estimates for such technologies whenever they were available, NICE worked for a while without any formal £/QALY threshold (NICE 2001). Indeed, in the first few years of NICE's life, its representatives were often at pains to point out that no £/QALY threshold was used by the Institute to issue recommendations.[6]

A couple of years after NICE's inception, both external observers and actors from within NICE started looking back at the decisions that it had taken thus far and for which £/QALY estimates were available to establish whether decisions were aligned with the cost effectiveness of the technologies under appraisal. In 2001, James Raftery (2001) found that by restricting the use of technologies to specific subgroups of patients, NICE had kept the cost per QALY of all recommended interventions below £30,000, apart from a single exception. Also in 2001, Sir Michael Rawlins, NICE's then chairman, identified the same upper limit beyond which NICE had been reluctant to issue positive recommendations.[7] Within a few months, it became evident that a single cutoff point was not enough to explain well the decisions that NICE had been making, and in 2002, both £20,000/QALY and £30,000/QALY were highlighted as important figures. Indeed, rejections had been extremely sparse below the £20,000/QALY mark, while the likelihood of rejection appeared to increase beyond it and a positive recommendation became very unlikely above £30,000/QALY (Towse and Pritchard 2002).

The message sent by these early historical analyses of NICE's decision pattern was that NICE had issued recommendations *as if* it had the cost-effectiveness threshold that it lacked the evidence for (although the threshold looked like a "soft" threshold, centered on a range of values, as opposed to a single cutoff point). NICE's reaction to this message is striking: it made the £20,000–£30,000/QALY range into the threshold that NICE committees would be expected to follow in future appraisals. NICE's threshold is supposed to be a measure of an opportunity cost. This supposition means that the level at which the threshold should be placed is a matter of empirical analysis given the cost effectiveness of currently funded health care areas. Identifying a £/QALY threshold by looking at the pattern of past HTA decisions made in absence of any such analysis looks like pulling oneself up by one's own bootstraps. However, this is what NICE did. Michael Rawlins and Anthony Culyer from NICE explained in a 2004 article that a review of past decisions

grounded the threshold that NICE would use in the future. The £20,000–£30,000/QALY threshold was enshrined in the 2004 edition of NICE's *Guide to the Methods of Technology Appraisal* and is still used today (NICE 2004; Rawlins and Culyer 2004). Yet this seems a remarkably haphazard way in which to measure such a socially and politically sensitive value.

Some might object that the process through which NICE identified its threshold looks problematic only if we endorse an overly simplified account of what good measurement involves—an account according to which the very first attempt at measurement must fully satisfy the theory of the object to be measured and what it takes to measure it. Could not NICE's setting of the threshold at £20,000–£30,000/QALY be interpreted as a first and admittedly imperfect iteration of a process of measurement of the relevant opportunity costs that would get better at approximating the relevant theory over time?[8] We are doubtful about this interpretation for at least two reasons.

First, even though the first iteration of a process of measurement does not need to fully satisfy the theory of the object to be measured, it seems plausible to require that it should have at least *some* link to such theory, and this link appears to be missing in NICE's original setting of the threshold. The pattern of NICE's past decisions is simply irrelevant to what the threshold is supposed to measure if past decisions are not grounded in an empirical analysis of the cost of a QALY in different areas of NHS expenditure. NICE has always been open about the fact that its threshold figures are grounded in no such analysis. For example, Rawlins commented during an interview that "[t]he £30,000 emerged. I've always said that it's not locked in some empirical basis. It emerged. And it emerged during the first year or two of the appraisal committee meeting" (Appleby 2016, 161). Second, even if we bracketed the issue of how NICE originally arrived at the £20,000–£30,000/QALY figure, NICE's behavior *after* it endorsed the figure does not fit well with the picture of an iterative process (i.e., one where, iteration after iteration, the threshold is based on a closer approximation of the cost effectiveness of currently funded technologies and therefore better approximates the relevant theory of the object to be measured).[9]

In a 2009 workshop on the threshold, NICE found some reassurance in the outputs of a research project carried out by Peter Smith and other economists of the University of York. Although based on rather limited data, this study was interpreted as suggesting that "NICE is probably not completely out of line in using its current £20-30K per QALY" (NICE 2009).[10] However, four years later, a major study was published that built on this older research project with damning implications for NICE. In a nutshell, Karl Claxton and the other authors of this study aimed to provide an estimate for NICE's threshold based on an empirical analysis of the decrease in the spending of a local commissioning authority that leads to the loss of one QALY's worth of health

gain through displacement. The estimate that Claxton and colleagues came up with for NICE's threshold is slightly lower than £13,000/QALY—less than half of the £30,000/QALY figure and far away from even the lower end of NICE's current range (Claxton et al. 2013)! Obviously, this study has disastrous implications for the ability of the current threshold to indicate when the approval of a new technology does more harm than good in terms of health benefits. Therefore, it seems that unless it had some objection to its reliability, NICE should amend its traditional £20,000–£30,000/QALY threshold in light of the new evidence.

NICE did not voice any objection against the evidence put forward by Claxton and colleagues. Still, it rejected the implication that NICE's threshold should be corrected. In explaining this choice, NICE's chief executive Sir Andrew Dillon appealed to considerations that bore no relation to the theory of opportunity costs that lies behind NICE's attention to the cost effectiveness of new technologies. Indeed, the problem with a £13,000/QALY threshold was that it "would mean the NHS closing the door on most new treatments," therefore failing to provide incentives for the pharmaceutical industry to bring about innovation (Dillon 2015). This revelation is a fitting last chapter in the complicated history of NICE's threshold so far. But how should this history be evaluated? This is the question we wish to tackle in the next section.

JUSTIFYING THE THRESHOLD?

As a measure of opportunity costs, the initial calculation of the £20,000–£30,000/QALY threshold seems bad, and the apparent refusal to change this threshold in light of later evidence even worse: using this number is, in effect, systematically to fund treatments that given NICE's stated aims, it should not be recommending for funding (and, thereby presumably, not leaving sufficient funds to fund drugs that should be funded). However, there is a different way of understanding this case, as reflecting the complex relationship between "measures" and "standards," which while not necessarily endorsing NICE's initial calculation and current apparent intransigence, may complicate our assessment.

A "standard" is a conventional rule, which specifies conditions that must be met before something else can or should happen. Often, standards employ numerical terms, as a way of translating between numerical measures and action. For example, within many UK higher education institutions, achieving 70 percent on an exam is necessary and sufficient for achieving a First-Class degree; this "standard" allows us to translate from numerical measurements of candidates' performance to their final degree classification. When standards are numerical, the relevant numbers can be either "fixed," as

in the example above, or "floating." As an example of the latter, we might adopt a different standard for a First-Class degree: that the candidate achieves whatever mark results in 20 percent of candidates achieving a First. Our numerical standard for a First might then "float": 68 percent one year and 71 percent another. An interesting feature of NICE's work, quite apart from the odd process through which NICE originally arrived at the £20,000–£30,000/QALY figure and the response to the Claxton critique, is that given its stated aims, the £20,000–£30,000/QALY threshold should "float," but instead it is "fixed." In the section "The Threshold: The Theory," we explained that given that the threshold is supposed to reflect opportunity costs, it should change as the NHS budget changes, as new health needs emerge, technologies are introduced, and so on; however, NICE never varies—and seems to lack any mechanism for varying—its threshold.

In the subsection below, "From an Argument for Fixed Standards to the Setting of NICE's Threshold," we will first sketch an argument for thinking that a "fixed" standard may be justifiable and then consider the relevance of this argument for understanding the original setting of NICE's threshold. In the next subsection, "Can Anything Be Said in Favor of NICE's Response to Claxton and colleagues?" we will turn to investigating the relevance of our arguments to NICE's response to the Claxton critique.

From an Argument for Fixed Standards to the Setting of NICE's Threshold

Constantly updating the standard—using a "floating" threshold—would be extremely complex from a technical perspective. However, this complexity is, presumably, only part of the story why NICE prefers a fixed standard. Rather, the preference for a fixed standard seems related to a more fundamental sociological dynamic, explored well by writers such as Ted Porter. This dynamic concerns how bureaucratic needs shape systems of measurement and assessment. As Porter explains, in a "public measurement system," such as the systems used by state bureaucracies (e.g., NICE or the NHS), there are strong forces requiring "standardization" (i.e., that like cases be treated alike or, at least, be seen to be treated alike) and "proper surveillance." As a result, "there is a strong incentive to prefer readily standardizable measures to highly accurate ones, where these ideals are in conflict" (Porter 1994, 391). For example, state-mandated systems for measuring the toxicity of chemicals may often differ from the "most accurate" measures. Indeed, Porter goes further, suggesting that "if an eccentric manufacturer were to invest extra resources to perform a state-of-the-art analysis, this would be viewed by the regulators as a vexing source of inter-laboratory bias, and very likely an effort to get more favorable measures by evading the usual protocol, not as a

welcome improvement in accuracy" (Porter 1994, 391). Of course, Porter is describing systems for measuring and assessing quantities relevant to some standard—rather than what we have called "standards" themselves—but it is easy to apply his model to NICE. A "floating" measure of the relevant cost per QALY would, given NICE's stated aims, be more apt than a "fixed" measure as part of the standard for funding decisions. However, just as using a state-of-the-art toxicity measure would be bureaucratically complex, the same might be true of using a floating standard even though they are both more "accurate."

The actions of a state body can be analyzed and assessed from many viewpoints. Clearly, one central question we can and should ask of state action is whether it abides by basic ethical norms of fairness or accountability. Therefore, to translate key themes of Porter's sociological analysis into the basis for an ethical assessment, we wish to ask, what might justify a fixed standard? There are two reasons to prefer such a standard. First, a fixed standard avoids potential problems of (perceived) diachronic fairness. Use of a floating standard would, presumably, imply that NICE should deny drugs to some patients who would, a year previously, have seen the same drugs approved (or conversely, that NICE should recommend drugs on the basis of their cost effectiveness while equally cost-effective drugs were rejected because of their cost per QALY the previous year). Such decisions might, in fact, be justifiable—given the underlying logic of cost-effectiveness analysis—but they would at least appear (and arguably be) unfair in that they would treat identical health technologies differently.[11] Second, changing the threshold every few months would make it more difficult for outsiders to assess and discuss NICE's decision-making procedures. In turn, this would have a detrimental impact on goods such as transparency and democratic oversight.[12] What is the value of (perceived) fairness, transparency, and democratic oversight? This is an open question, but given NICE's regulatory role, such goods are not to be given up lightly. Therefore, even if NICE's use of a fixed standard eschews accurate measurement of the constantly varying "true" opportunity cost, it is not clear that a more accurate floating standard would, all things considered, be preferable.

This discussion of the justifiability of a fixed threshold highlights a set of goods that are external to the narrow logic of cost-effectiveness analysis but that can be used as a counterbalance to it. In turn, consideration of these goods sheds light on the justifiability of a more fundamental choice that NICE made—that of having an *explicit* threshold at all and, therefore, setting the £20,000–£30,000/QALY figure despite the lack of solid empirical evidence concerning opportunity costs. There are three possible arguments in favor of NICE's setting an explicit standard. To start with, the very existence of an explicit threshold appears to foster transparency and democratic oversight

at an even deeper level than the choice of a fixed over a floating threshold. Furthermore, an explicit threshold ensures coordination across NICE's five health technology appraisal committees. By standardizing the process through which those committees are supposed to issue recommendations about drugs, an explicit threshold reduces the risk of an unfair "committee lottery," in other words, the risk that patients in need of a new drug will have that drug denied simply because it has been evaluated by one NICE HTA committee and not another. Like the diachronic fairness fostered by a threshold that is not constantly recalculated, standardization across NICE's various HTA committees makes it more probable that like cases will be treated alike by decision makers. Finally, an explicit threshold sends a clear message to the pharmaceutical industry. By stating that NICE is unwilling to recommend spending more than a certain amount of money for a QALY, an explicit threshold places pressure on industry to lower the price of many expensive drugs, ultimately benefiting the NHS.[13]

Even if we can justify using a measurement estimate that is both explicit and fixed, it seems important that the estimate be roughly correct. There may be good reasons to use a blood pressure reading of 140/90 as a point for prescribing cardiovascular medications for all patients even if a more complex standard would provide more targeted care; the "benefits" of such a system (in terms of ease of implementation, transparency, ease of communication) might outweigh the "costs" (in terms of occasional over- or underprescription). However, such arguments require that 140/90 tracks something like "actual" increased risk; if it doesn't, ease of use seems irrelevant. Similarly, we might worry that even if there are good reasons for NICE to use an explicit and fixed standard, those arguments are not worth much if that figure wildly underestimates the opportunity costs of new drugs. We will now suggest a way of challenging this thought.

So far, we have used Porter's sociological model to explore one way in which the "constraints" on administrative agencies *may* justify a preference for measures that given those agencies' stated roles, seem peculiar (e.g., a fixed over a floating standard). However, regulatory agencies may serve multiple roles, some of which may differ from their stated roles, and these additional roles may, themselves, be morally or politically valuable. To make this clearer, consider the complex political backdrop to NICE's work. Decisions about drug funding are, of course, highly controversial. There are good reasons for politicians and governments to want such decisions to be removed—at least in part—from their (perceived) sphere of influence. One function of NICE is to serve this end (Gash et al. 2010, 18). From this perspective, it is important that NICE's decisions be seen as impersonal. Political controversy can be avoided if decisions are seen as the result, at least in part, of "objective" number crunching. In the words of Rottenberg and Merry, numeric

representation in governance achieves the political purpose of demonstrating, "adherence to public responsibility and absence of personal or group bias" (2015, 8). It might be added that numbers carry immense symbolic authority as guarantors of objectivity, rigor, and universality and hence may contribute to institutional legitimacy quite independently of their precision and accuracy (Sauder and Espeland 2015, 436). To place these comments in a broader context, we might say that one of NICE's key *political* functions is to ensure that funding decisions are—or are perceived to be—"procedurally objective," in the sense that they are determined by application of impersonal rules rather than individual idiosyncrasies, in Megill's nice phrase, they are "untouched by human hands" (Megill 1994, 10). We can contrast this sense of "objectivity" with the "absolute" sense of "objectivity" at play, for example, when we describe measures as "representing reality as it really is" as in the case at hand, when we ask whether NICE's threshold "really" reflects the "true" opportunity costs of NHS spending (Megill 1994, 1).

Building on these comments and moving once again from a more sociological to a more philosophical level of analysis, we can distinguish two functions of the threshold: one is the "stated aim" (i.e., to ensure that concerns about cost effectiveness are accurately reflected in decisions about resource allocation). As explained above in "The Threshold: The Theory," there is a familiar debate in the philosophical literature over how best to balance the aggregative and maximizing logic of cost effectiveness with more egalitarian or other distributive considerations. From the perspective of such debates, cost effectiveness is a morally relevant consideration, and therefore NICE should revise its threshold to £13,000 and maybe even allow it to "float" around this point. (Strictly speaking, this conclusion might be sensitive to the moral weight we assign to cost effectiveness, but the underlying issue is clear: £30,000/QALY must be rejected!)

However, it is not clear that NICE's only normatively relevant function is to act as a kind of massive central planner (even a central planner whose decisions are to be guided by *more* than cost effectiveness). Rather, we might understand its role differently (i.e., ensuring consistency across cases, placing limits on political pandering to electorally significant groups, allowing for rational planning, stabilizing drug prices, ensuring that decisions can be assessed and criticized, creating a broad democratic debate over the ethics of NHS resource allocation, and so on). From an ethical perspective, these are all potentially important goods, which require only that NICE's recommendations are procedurally objective. The precise numbers specified in these procedures are a bit like the rule that football teams have eleven players on each side. The number eleven is not magical, in that one could play a sport very like football with twelve people, but we need to settle on some number if there is to be fair competition, if teams are to be able to plan strategy, and

so on. What is required is that there be *some* defensible number, not that the number reflects some fact, such as that football is "best played" as eleven a side.

Our claim, then, is that when we think about NICE's work through the prism of "procedural objectivity" and of the politically and morally important goods that this sort of objectivity generates, it seems less important that the number it chose was "true" than that it chose some broadly acceptable number at all and then set this number as a kind of explicit benchmark for itself and others to follow. Clearly, given that there was some kind of apparent implicit agreement on the £20,000–£30,000/QALY figure, choice of this number was not completely unreasonable for this purpose. Admittedly, there is something odd about this approach, in that a number that looks like—and is ostensibly described and justified as—a measure turns out to function more like a convention. We return to this issue below, but note here that any serious attempt to think about measurement and standards in institutional and political contexts, such as NICE, that is not alert to issues of coordination and fairness is likely to overlook morally and politically important concerns.

Can Anything Be Said in Favor of NICE's Response to Claxton et al.?

Once we stress NICE's coordinating role, concerns about the initial choice of the £20,000–£30,000/QALY threshold are less pressing even though at that time, there was no empirical evidence connecting that figure to "true" opportunity costs. Still, you might think that NICE's stated goal—to ensure that money is spent in a cost-effective manner—is of great importance. From this perspective, now that empirical evidence is available, NICE should change its threshold. Doing so may seem compatible with serving its other politically and morally relevant functions: after all, we can just as well coordinate around £13,000/QALY as £20,000–£30,000/QALY.

One might explain NICE's response to the work of Claxton et al. as an instance of a more general familiar sociological phenomenon of bureaucratic inertia: it would be tiresome and costly to change the threshold. Furthermore, it seems there is little political impetus to do so (plausibly, matters would be very different were the report to have suggested a *higher* threshold; there are many patient advocacy groups who would agitate for a higher threshold). Still, important and interesting as these dynamics are, it is hard to see how they might justify at a philosophical level NICE's apparent insouciance in the face of Claxton's critique.

However, the model developed above that viewed NICE as serving many political functions beyond its "official" role provides a more nuanced assessment of NICE's refusal to shift its threshold. Consider again Andrew Dillon's

justification for retaining the higher threshold quoted at the end of the previous section: that a change would disincentivize innovation. His argument seems wrongheaded if we view NICE's work solely as a central planner maximizing under certain distributive constraints the health benefits that NHS interventions can produce. But in our model in which NICE serves many different political functions, we might plausibly say that one function of NICE is to promote pharmaceutical innovation and, hence, contribute to the British economy. If this function is viewed as normatively valuable and if it is true that changes to the threshold would negatively affect innovation and, hence, the economy, then maybe there is some argument for retaining the deeply flawed threshold. Furthermore, it may be possible to argue that if NICE is supposed to view the pharmaceutical industry as a kind of stakeholder in its work, then, given the long-term nature of planning in the pharmaceutical industry, there may be considerations of fairness that count against a rapid change in the threshold's value. In making these remarks, we are not endorsing such arguments. It is unclear that NICE should have the role of promoting innovation and unclear that a lower threshold would stifle—as opposed to incentivize—drug development. What we are suggesting, rather, is that no proper assessment of NICE's work can go forward without proper attention to the purposes behind its measures (or the purposes they have come to serve). Our approach allows us to engage with Dillon's justifications rather than treat them as necessarily irrelevant.

When we have some politically mandated system of measurement or standard that incorporates some numerical value, we can always ask whether that system or standard is fit for purpose. When making such an assessment, we might be willing to sacrifice a certain degree of accuracy for other goods. For example, if we assume that NICE's threshold is intended to capture concerns about cost effectiveness, we can ask whether the £20,000–£30,000/QALY threshold is fit for that purpose. When we realize that it should, but does not, float, we might be willing to tolerate this "inaccuracy" as the "price" for, say, ease of use. However, we also made a second, stronger claim: that NICE's threshold serves multiple functions not related to cost effectiveness but rather to the appearance of fairness, to stabilizing expectations, to facilitating democratic deliberation, and so on. From this broader perspective, "fitness to purpose" is more complicated because what matters is not only that the numerical standard accurately reflects opportunity costs but that there is a standard that remains stable across time and maintaining those goods may require defending this number even when it is "wrong" in the narrower sense. None of these arguments straightforwardly justify NICE's decisions. After all, NICE's scheme turns out not to incorporate a concern that many do think is important: whether drugs are "cost effective." However, any critique of NICE's inertia must start from an assumption about its proper normative

function; there is no point in measuring opportunity costs more or less accurately if those costs are irrelevant to NICE's work. Once this point is made explicit, it is entirely proper to ask whether NICE serves other proper normative functions and if so, whether these functions might, at least in part, justify the (apparently arbitrary) threshold bequeathed by historical accident. To ignore these issues is to endorse one set of value concerns as the only ones that are proper without adequately considering all the alternatives.

CONCLUSION

Many features of the £20,000–£30,000/QALY threshold, such as how it was derived, that it is "fixed," and more, make little sense considering the standard account of NICE's purposes, including the accounts NICE itself gives. Whatever else we can say about these numbers, they are not a good measure of the opportunity costs of funding new technologies. However, many of these puzzling features make sense, and may even be justified, when we rethink NICE's functions not only as a central planner but also as a guarantor of procedural objectivity in a domain of deep conflict. Clearly, this story has implications for our understanding of the work of NICE, similar HTA bodies in other jurisdictions, and heated debates over the "proper" way of measuring and rationing health care interventions more generally. However, it also has a broader implication for measurement in (and maybe beyond) medicine. Any measure of cost effectiveness is, in a trivial sense, value laden because, for example, we need to choose what effects to measure. To measure the effectiveness of a treatment along some dimension is, if only implicitly, to assume that this dimension is of prudential moral or political significance.

These are familiar claims, well covered in the now extensive literature on how to construct measures of health-related quality of life. What is less obvious, but no less important, is that the construction of measures to be used for policy making may be subject to further moral and political considerations that may be in tension with the aim of accurately representing the aspect of reality that these measures are supposed to track. Demands that users of measures be accountable to others for their decisions, for example, may give us reasons to prefer measures that have a certain sort of "inflexibility," for example and thus do not always track what we seek to measure. Furthermore, measures may take on a "life of their own," such that claims that measures are inadequate guides to underlying phenomena may fail to consider the role that these measures play in the complex ecology of policy. For example, when a putative measure becomes a standard or a target, we need not only ask how well it functions as a measure but also what the consequences are of its further uses. Taking account of such concerns is not to say that questions of accuracy

are irrelevant to our assessment of measures. Rather, it is to add a twist to the truisms that measurement is always for a purpose and that adequacy is relative to our purposes: that the purposes of measurement are often multiple and might even be opaque. To make such claims is not to dismiss a measure but rather to open up a new question, about whether those purposes are sufficiently valuable that we should tolerate inaccurate measures.

NOTES

1. For this process in general, see NICE (2008, 17–19) and NICE (2013, 72–74). For the equity weightings, see Rawlins, Barnett, and Stevens (2010).

2. See Bognar and Hirose (2014, 1–6) for a succinct version of these arguments. The first section will cover them in greater detail.

3. For but one recent treatment of this topic from a philosophical perspective, see Hausman (2015).

4. Of course, the reference to the *single* least cost-effective intervention that is currently funded makes sense only on the simplifying assumption that the introduction of the new technology does not have a larger budgetary impact than that intervention currently has. In principle, if the technology under appraisal has a particularly large budgetary impact, the cost-effectiveness threshold should be lowered.

5. For the importance of balancing health maximization and distributive concerns, see Bognar and Hirose (2014, 53–78 and 104–26), Brock (2004), and Daniels and Sabin (2008, 30–34).

6. For example, see the discussion of Michael Rawlins's public statements in Littlejohns (2002, 32).

7. See the references to Rawlins's discussion of the topic at NICE's 2001 annual public meeting in both Littlejohns (2002, 31–32) and Towse and Pritchard (2002, 26–27).

8. See Tal (2013) for an excellent account of why naïve theories of measurement are descriptively and normatively problematic.

9. To use an effective image introduced by Culyer et al. (2007) to outline their idea of NICE's role, we aim to show that NICE has been a poor "threshold-searcher" also after endorsing the £20,000–£30,000/QALY figure in 2004.

10. For more on the research project under discussion, see Martin, Rice, and Smith (2009).

11. That like cases should be treated alike is often proposed as a basic principle of fairness or justice in the allocation of health care resources. For example, see Clark and Weale (2012, 306–307) and Daniels and Sabin (2008, 47–49).

12. To cite but one influential account of fair procedures, transparency is one of the four conditions defining a fair process for the allocation of health care resources according to Daniels and Sabin (2008, 43–66). Also, Daniels and Sabin (2008, 59–60) argue that part of the value of fair procedures in health care resource allocation is that such procedures foster democratic deliberation in society at large.

13. On the other hand, however, there are cases in which a fixed threshold that operates as a standard might just as easily have the opposite effect, namely, encourage industry to come in at the higher end in its pricing if it still comes in within the approved range. This concern is encompassed by the observation of scholars of policy and audit that standards are also susceptible to treatment as targets, which can have perverse effects with respect to the goals that might drive their use in the first place. For this point in general, see Shore and Wright (2015, 425). Bevan and Hood (2006) provide a more focused analysis of the gaming of targets in the NHS.

REFERENCES

Appleby, John. 2016. "What's In and What's Out? The Thorny Issue of the Threshold." In *A Terrible Beauty: A Short History of NICE*, edited by Nicholas Timmins, Michael Rawins, and John Appleby, 154–69. Nonthaburi, Thailand: HITAP.

Bevan, Gwyn, and Christopher Hood. 2006. "What's Measured Is What Matters: Targets and Gaming in the English Public Health Care System." *Public Administration* 84: 517–38.

Bognar, Greg, and Iwao Hirose. 2014. *The Ethics of Health Care Rationing*. New York: Routledge.

Boseley, Sarah. 2015. "Postcode Lottery for Cancer Drug as Nice Rules Kadcyla Too Expensive." *The Guardian*, November 27. Accessed July 9, 2016. https://www.theguardian.com/society/2015/nov/17/postcode-lottery-cancer-drug-nice-rules-kadcyla-too-expensive.

Brock, Daniel. 2004. "Ethical Issues in the Use of Cost Effectiveness Analysis for the Prioritisation of Health Care Resources." In *Public Health, Ethics, and Equity*, edited by Sudhir Anand, Fabienne Peter, and Amartya Sen, 201–23. Oxford and New York: Oxford University Press.

Clark, Sarah, and Albert Weale. 2012. "Social Values in Health Priority Setting: A Conceptual Framework." *Journal of Health Organisation and Management* 26: 293–316.

Claxton, Karl, Stephen Martin, Marta Soares, Nigel Rice, Eldon Spackman, Sebastian Hinde, Nancy Devlin, Peter C. Smith, and Mark Sculpher. 2013. "Methods for the Estimation of the NICE Cost Effectiveness Threshold." *CHE Research Paper 81*.

Culyer, Anthony, Christopher McCabe, Andrew Briggs, Karl Clazton, Martin Buxton, Ron Akehurst, Mark Sculpher, and John Brazier. 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.

Daniels, Norman, and James Sabin. 2008. *Setting Limits Fairly: Learning to Share Resources for Health*. 2nd Edition. Oxford and New York: Oxford University Press.

Dillon, Andrew. 2015. "Carrying NICE over the Threshold." *NICE Blog*, February 19. Accessed July 9, 2016. https://www.nice.org.uk/news/blog/carrying-nice-over-the-threshold.

Gash, Tom, Jill Rutter, Ian Magee, and Nicole Smith. 2010. *Read before Burning: Arm's Length Government for a New Organisation*. London: Institute for Government.

Hausman, Daniel. 2015. *Valuing Health: Well-Being, Freedom, and Suffering*. Oxford: Oxford University Press.

Littlejohns, Peter. 2002. "Does NICE Have a Threshold? A Response." In *Cost Effectiveness Thresholds: Economic and Ethical Issues*, edited by Adrian Towse and Clive Pritchard, 31–37. London: King's Fund.

Martin, Stephen, Nigel Rice, and Peter Smith. 2009. *The Link between Healthcare Spending and Health Outcomes for the New English Primary Care Trusts*. London: Health Foundation.

McCabe, Cristopher, Karl Claxton, and Anthony Culyer. 2008. "The NICE Cost-Effectiveness Threshold: What It Is and What That Means." *Pharmacoeconomics* 26: 733–44.

Megill, Allan. 1994. "Introduction: Four Senses of Objectivity." In *Rethinking Objectivity*, edited by Allan Megill, 1–20. Durham: Duke.

NICE. 2001. *Guide to the Methods of Technology Appraisal 2001*. London: NICE.

NICE. 2004. *Guide to the Methods of Technology Appraisal 2004*. London: NICE.

NICE. 2008. *Social Value Judgements: Principles for the Development of NICE Guidance*. London: NICE.

NICE. 2009. *Threshold Workshop: Report of a Technical Meeting Organised by NICE*. London: NICE.

NICE. 2013. *Guide to the Methods of Technology Appraisal 2013*. London: NICE.

NICE. 2015. "Kadcyla Price Too High for Routine NHS Funding, Says NICE in Final Guidance." Accessed July 9, 2016. https://www.nice.org.uk/news/press-and-media/kadcyla-price-too-high-for-routine-nhs-funding-says-nice-in-final-guidance.

Porter, Theodore. 1994. "Making Things Quantitative." *Science in Context* 7: 389–407.

Raftery, James. 2001. "NICE: Faster Access to Modern Treatments? Analysis of Guidance on Health Technologies." *British Medical Journal* 323: 1300–1303.

Rawlins, Michael, David Barnett, and Andrew Stevens. 2010. "Pharmacoeconomics: NICE's Approach to Decision-Making." *British Journal of Clinical Pharmacology* 70: 346–49.

Rawlins, Michael, and Anthony Culyer. 2004. "National Institute for Clinical Excellence and Its Value Judgments." *British Medical Journal* 329: 224–27.

Rottenberg, Richard, and Sally Engle Merry. 2015. "A World of Indicators: The Making of Governmental Knowledge through Quantification." In *The World of Indicators: The Making of Governmental Knowledge through Quantification*, edited by Richard Rottenberg, Sally E. Merry, Sung-Joon Park, and Johanna Mugler, 1–33. Cambridge: Cambridge University Press.

Sauder, Michael, and Wendy Espeland. 2015. "Comment on 'Audit Culture Revisited: Rankings, Ratings, and the Reassembling of Society.'" *Current Anthropology* 56: 436–37.

Shore, Chris, and Susan Wright. 2015. "Audit Culture Revisited: Rankings, Ratings, and the Reassembling of Society." *Current Anthropology* 56: 421–44.

Smith, Richard, and Jeff Richardson. 2005. "Can We Estimate the 'Social' Value of a QALY? Four Core Issues to Resolve." *Health Policy* 74: 77–84.

Tal, Eran. 2013. "Old and New Problems in Philosophy of Measurement." *Philosophy Compass* 8: 1159–73.

Towse, Adrian, and Clive Pritchard. 2002. "Does NICE Have a Threshold? An External View." In *Cost Effectiveness Thresholds: Economic and Ethical Issues*, edited by Adrian Towse and Clive Pritchard, 25–30. London: King's Fund.