

Utilitarian or Quantile-Welfare Evaluation of Health Policy?

Charles F. Manski

Northwestern University and IPR

John Mullahy

University of Wisconsin-Madison

Version: February 18, 2026

DRAFT

Please do not quote or distribute without permission.

Abstract

This paper considers quantile-welfare evaluation of health policy as an alternative to utilitarian evaluation. Manski (1988) originally proposed and studied maximization of quantile utility as a model of individual decision making under uncertainty, juxtaposing it with maximization of expected utility. That paper's primary motivation was to exploit the fact that maximization of quantile utility requires only an ordinal formalization of utility, not a cardinal one. This paper transfers these ideas from analysis of individual decision making to analysis of social planning. Manski and Mullahy begin by summarizing basic theoretical properties of quantile welfare in general terms rather than related specifically to health policy. They then propose a procedure to nonparametrically bound the quantile welfare of health states using data from binary-choice time-tradeoff (TTO) experiments of the type regularly performed by health economists. After this they assess related econometric considerations concerning measurement, using the EQ-5D framework to structure our discussion.

Acknowledgements. Thanks are owed to Julian Reif and Dave Vanness for helpful suggestions.

1. Introduction

Economic evaluation of health policy has sought to ground analysis in foundations of welfare economics. Economists assume that an actual or hypothetical planner specifies a social welfare function (SWF). Research seeks to characterize the welfare achieved by alternative feasible policies, aiming to find one that maximizes an SWF. In applications, such research may be called cost-benefit analysis by general economists and cost-effectiveness analysis by health economists.

Paul Samuelson placed responsibility for specification of the SWF on society rather than on the economist, writing (Samuelson, 1947, p. 220): “It is a legitimate exercise of economic analysis to examine the consequences of various value judgments, whether or not they are shared by the theorist.” In practice, economists have studied policy choice with what might be called *pragmatic* social welfare functions, these being ones motivated by (Manski, 2024, p. 58): “some combination of conjecture regarding societal values, empirical study of population preferences, and concern for analytical tractability.”

Economists have mainly studied *personalist* SWFs, ones that are increasing functions of the personal welfare (aka *utility*) of the members of a specified population.¹ They have, moreover, commonly presumed utilitarian aggregation of interpersonally comparable cardinal utilities. In practice, utilitarian analysis implies a focus on population mean preferences, which sums individuals’ utilities and divides by population size.

The utilitarian perspective has been especially prevalent in health economics. It has motivated econometric analysis of treatment response to estimate average treatment effects rather than other aspects of distributions of treatment response. In research measuring individual health-related utility in quality-adjusted life years (QALYs), it has motivated calculation of so-called *value sets* for health-related quality

¹ The term *personalist* SWF revises the word *welfarism* originated by Sen (1977). He wrote (p. 1559): “The general approach of making no use of any information about the social states other than that of personal welfares generated in them may be called ‘welfarism.’” Manski (2024) argues that *personalist* SWF expresses Sen’s intended distinction more clearly than the word *welfarism*.

of life (HRQoL), which seek to measure the mean valuations of health states in a population. A subfield of health economics using the EQ-5D framework has developed methods to measure mean preferences through surveys that pose hypothetical choice problems and ask respondents to state their preferences among multiple health states. See Devlin *et al.* (2022).

Although economists have long found the utilitarian perspective congenial, they have recognized that summation of interpersonally comparable cardinal preferences is only one way to aggregate heterogeneous preferences into a personalist SWF. Rawls (1971) notably argued for a different way, aiming to maximize the utility of the worst-off member of society. Welfare economists have often opined that it is objectionable that utilitarian aggregation is sensitive to cardinal strength of preference. It is sometimes argued that mean preferences may be unduly influenced by individuals with extreme values of cardinal utility.

In health economic analysis using the EQ-5D framework, extremely high values of HRQoL are prevented by the convention of defining the value 1 to express “full health,” which cannot be exceeded. The value 0 is routinely used to express the value of death or a health state equivalent to dead, but individuals are not required to view death as worse than any living state of health. Empirical analysis of stated preference experiments regularly finds that some individuals view certain hypothetical states of health to be “worse than dead,” expressing negative values of HRQoL in these states. The standard EQ-5D survey protocol only permits respondents to express a HRQoL that is moderately worse than death, providing no way to express very negative values. Analysts have interpreted this as generating a data censoring problem. They have commonly dealt with it by estimation of a Tobit model, which assumes that utilities in a specified health state are normally distributed. Again see Devlin *et al.* (2022) and also section 4.2 below.

The EQ-5D process using the Tobit model to estimate value sets has been controversial for econometric, psychological, and normative reasons. The econometric concern is that Tobit model estimates are fragile, being sensitive to departures from normality in the actual distribution of cardinal utilities (see section 4.2). The psychological concern is that personal evaluation of their utility in very

poor health states is a challenging cognitive task, suggesting that individuals may not have well-defined utilities for such states. The normative concern is that, even if individuals would express very negative values of health-related utility, society may not want to make policy using an SWF that is “unduly influenced” by these expressions.

In this paper, we consider quantile-welfare evaluation of health policy as an alternative to utilitarian evaluation. Manski (1988) originally proposed and studied maximization of quantile utility as a model of individual decision making under uncertainty, juxtaposing it with maximization of expected utility.² The primary motivation was to exploit the fact that maximization of quantile utility requires only an ordinal formalization of utility, not a cardinal one. It was discovered that quantile utility has related interesting properties not shared by maximization of expected utility.

We transfer these ideas from analysis of individual decision making to analysis of social planning. The substantive context differs. The utility distribution is over heterogeneous states of nature in the individual context and over heterogeneous persons in the social planning context. Nevertheless, the mathematics is the same.

As far as we are aware, research in health economics has not previously studied quantile-welfare evaluation of health policy. However, the literature has on occasion come close when health economists have suggested the use of median rather than mean preferences to evaluate policies. Such suggestions have sometimes been motivated empirically by the well-known insensitivity of the median to the particular shape of the tails of probability distributions, a property loosely called invariance to “outliers.” They have sometimes been motivated conceptually by the heuristic notion that the median is a better summary statistic of “central tendency” than the mean when a distribution is skewed (see, Austin, 2002, p. 336). The median, of course, is a synonym for the 0.5-quantile.

We hope that researchers working on empirical HRQoL questions will find the analysis in this paper useful. The paper provides welfare-theoretic foundations for empirical strategies that applied researchers

² A small literature on the subject has developed since then. Rostek (2010) developed an axiomatic interpretation. Manski and Tetenov (2023) studied maximization of quantile utility with sample data.

may elect to implement. Adopting a particular objective (EU max, quantile max, etc.) leads naturally to corresponding empirical strategies (estimation of mean, median, quantile). Conversely, adopting a particular estimation strategy may imply at least to some degree the particular objective that such estimates serve to inform. We think it desirable to more tightly link specification of policy goals to the empirical strategies used to inform them.

Section 2 summarizes basic theoretical properties of quantile welfare. The discussion is general rather than related specifically to health policy. It does not address the practical issue of measuring utilities in applications. Section 3 proposes a procedure to nonparametrically bound the quantile welfare of health states using data from binary-choice time-tradeoff (TTO) experiments of the type regularly performed by health economists. Section 4 assesses the data collection and econometric methods that have been commonly used by research performed in the EQ-5D framework, which has sought to measure mean quality of life given a QALY representation of health-related utility. Section 5 concludes.

2. Maximization of Quantile Welfare

2.1. Individual Maximization of Quantile Utility

Manski (1988) proposed maximization of quantile utility as a class of criteria for individual decision making under uncertainty. Both quantile and expected utility maximization respect weak stochastic dominance, a basic feature of any reasonable decision rule. Nevertheless, maximization of expected and quantile utility differ in important respects. The most fundamental is that the ranking of actions by expected utility is invariant only to cardinal transformations of the utility function. In contrast, rankings by quantile utility are invariant to ordinal transformations.

This fundamental difference has important consequences. An immediate technical difference is that expected utility is not well-defined when the distribution of utility has unbounded support with fat tails,

but quantile utility is always well-defined. A more subtle substantive consequence is that quantile utility yields a significant generalization of traditional ideas of risk preference.³

2.1.1. The Quantile Utility Perspective on Risk Preference

From the quantile utility perspective, it is reasonable to define one action to be riskier than another if the utility distribution function of the latter crosses that of the former from below. Lemma 2 of Manski (1988) shows that this single-crossing property is equivalent to ‘spreading’ the less risky utility distribution in a natural manner.

The single-crossing criterion for comparing the riskiness of actions is much more general than the accepted characterization of riskiness in expected utility theory. With expected utility, actions are risk comparable only if the single-crossing property and other conditions hold, being: (1) utility is an increasing function of a real-valued outcome, such as income or health, and (2) the actions being compared have outcome distributions with the same mean. Then one action is deemed riskier than another if its outcome distribution is a mean-preserving spread of the other. With quantile utility, the outcomes generating utilities are unrestricted and utility need only be a measurable function of the outcome. Outcome distributions need not have the same means; indeed, the means need not exist.

With the risk comparability of actions defined by the single-crossing property, Proposition 3 of Manski (1988) shows that the risk preference of a quantile utility maximizer increases with the utility distribution quantile that he maximizes. The reasoning underlying the Proposition is simplest when considering two actions, say A and B, whose utility distributions are continuous and strictly increasing. Suppose that the distribution function for A crosses that of B from below at some utility value u^* , where the distributions functions both take some value p^* . Thus, $P[u(A) \leq u^*] = P[u(B) \leq u^*] = p^*$. Then the α -

³ We emphasize that this paper’s quantile utility welfare comparisons are those of Manski (1997), termed ΔD evaluations, that compare quantiles of the marginal distributions of utilities that arise under different actions or policies. In contrast, $D\Delta$ evaluations compute quantiles of the distribution of the difference in utility under alternative policies. ΔD evaluations are common in clinical research where, for instance, differences in marginal medians of time-to-event outcome distributions are often of primary interest (Mullahy, 2021).

quantile of A is larger than that of B for all $\alpha < p^*$, equals that of B for $\alpha = p^*$, and is smaller than that of B for all $\alpha > p^*$. Hence, an individual who maximizes the α -quantile of utility strictly prefers A to B if $\alpha < p^*$, is indifferent if $\alpha = p^*$, and strictly prefers B to A if $\alpha > p^*$.

The quantile-utility characterization of risk preference contrasts sharply with that in expected utility theory. There, individuals with concave utility functions are called risk-averse and ones with convex utility functions are called risk loving. Concavity and convexity are cardinal properties of functions, which are not germane in quantile utility theory.

2.1.2. Sequential Quantile Utility Maximization

The expected utility criterion has a property that initially appears more appealing than quantile utility criteria. Consider actions A and B such that the utility distribution of A stochastically dominates that of B strictly. Then the expected utility of A is strictly larger than that of B, but some quantile utilities of A may equal those of B. Thus, quantile utility maximization may indicate indifference between actions A and B, even though the utility distribution of A stochastically dominates that of B strictly.

To address this issue, Manski (1988) proposed sequential quantile utility maximization, a lexicographic refinement in which the individual establishes a sequence of quantiles for consideration. If one action yields a larger value of the first quantile in the sequence, the individual strictly prefers this action. If both actions yield the same value of the first quantile in the sequence, the individual compares them using the second quantile and strictly prefers one to the other if they differ in the second quantile. If not, the individual considers the third quantile, and so on. Levy and Kroll (1978) proved that one probability distribution stochastically dominates another strictly if and only if all quantiles of the former distribution are weakly larger than those of the latter and some quantile is strictly larger. Hence, the result of sequential quantile utility maximization is that strict stochastic dominance yields strict preference.

2.2. Social Maximization of Quantile Welfare

Whereas decision theory studies an individual who chooses an action in an unknown state of nature, personalist welfare economics studies a planner who chooses a policy for a population of individuals. This is a profound substantive difference, but the mathematics of the two settings are similar. The population of welfare economics is analogous to the state space of decision theory.

The mathematical analogy strengthens when one places a probability distribution on the state space in decision theory and, correspondingly, characterizes the population in welfare economics by a distribution of utility functions. Then the individual and the planner both associate each feasible action with a probability distribution, respectively a distribution of utility across the state space and one across the population. In both settings, the decision maker may rank actions by some functional of the relevant distribution, such as the mean or a quantile.

Whereas decision theory has considered respect for stochastic dominance to be a basic feature of any reasonable decision criterion, personalist welfare economics has taken respect for Pareto dominance to be a *sine qua non*. The latter property is not probabilistic, as it views each member of the population to be a named individual. However, it is usual in welfare economics to suppose that the planner is concerned only with the population distribution of utility, not with the identities of population members. When this condition is imposed, the essence of Pareto dominance is conveyed by stochastic dominance.

Risk preference in decision theory is mathematically analogous to inequality preference in welfare economics. In both cases, the concern is with some measure of the spread of the relevant probability distribution. In utilitarian welfare economics, an inequality-averse planner maximizes the mean of a concave transformation of interpersonally comparable cardinal utility, with strength of inequality aversion being expressed by the degree of concavity. A leading example is optimal income tax theory as initiated by Mirrlees (1971). With quantile-welfare maximization, aversion to inequality is conveyed by the quantile maximized, a lower quantile expressing greater aversion. This is the central idea underpinning value-at-risk metrics in insurance and actuarial research. The limiting case is the SWF proposed by Rawls

(1971), which evaluates social welfare by the utility of the worst-off (smallest quantile) member of the population.

3. Bounding Quantile Welfare with Data from Binary-Choice Time-Tradeoff Experiments

We now turn to measurement of health-related utility using stated preference data. Henceforth, let H be a time-invariant set of potential health states that could be realized in each year $t = 1, \dots, T$, where T is a specified terminal year. Each state in year t is a value $h_t \in H$, $t = 1, \dots, T$. At each t , one value in H will be realized. As normalizations, $h_t = 1$ denotes full health and $h_t = 0$ denotes death in a year before or equal to t . All health state vectors $h \equiv (h_t, t = 1, \dots, T)$ are feasible, except that $h_t = 0 \Rightarrow h_{t+1} = 0$.

Let J denote the population of interest. For each $j \in J$, let $u_j(h)$ denote the person-specific ordinal utility of state vector h to person j . In the analysis of this section, we make no assumptions about the structure of preferences except for two properties. One is that preferences are strict rather than weak; that is, indifference does not occur. The other is that full health in a given year is preferred to other health states, ceteris paribus (see section 4.4). That is, let h be any feasible state vector, and let $h^* = h$ except that it replaces some components of h that do not equal 1 with components that equal 1. Then $u_j(h^*) > u_j(h)$.

The broad definition of health states and utility posed here encompasses the tighter specifications commonly assumed in research on health economics. It has been particularly common to assume that utility has the undiscounted, additive QALY form $u_j(h) = \sum_{t=1}^T q_j(h_t)$, where $q_j(h_t)$ is person j 's quality of life in health state h_t , $q_j(1) = 1$, and $q_j(0) = 0$. Health economists occasionally consider a time-discounted extension of this utility function of the form $[u_j(h) = \sum_{t=1}^T (\delta_j)^t q_j(h_t)]$, where δ_j is the rate of time discount used by j .

In principle, each $j \in J$ may be presented with binary choice experiments that draw pairs of feasible health state vectors and ask the person to choose between them. With sufficient data collection, this reveals $u_j(\cdot)$. If the entire population is surveyed and all respond, it reveals the utility distribution $P[u(\cdot)]$.

Thus far, utility functions are not interpersonally comparable. Defining quantiles requires a reasonable way to make them ordinally comparable. The maintained assumption that full health in a given year is preferred to other health states, *ceteris paribus*, provides a way to proceed.

Let d denote year of death and consider the health state vector $h^{*d} \equiv (1, \dots, 1, 0, \dots, 0)$ in which a person is alive with full health through year d and dies at the end of year d . We normalize ordinal utilities by letting $u_j(h^{*d}) = d$ for all $j \in J$ and $d \in (1, \dots, T)$. Similarly, immediate death at the time of the choice experiment has utility $u_j(h^{*0}) = 0, j \in J$. These normalizations are consistent with the assumption that full health is preferred to other health states, *ceteris paribus*. In particular, it is consistent with the undiscounted QALY form of utility.

Now consider the utility $u_j(h)$ of any health state vector h . We can use the h^{*d} normalization to place $u_j(h)$ in one of $T + 1$ intervals, as follows:

- If TTO experiments reveal that $u_j(h) < u_j(h^{*0})$, then $u_j(h) < 0$.
- If TTO experiments reveal that $u_j(h^{*(d-1)}) < u_j(h) < u_j(h^{*d})$ for d such that $0 < d \leq T$, then $d - 1 < u_j(h) < d$.

It follows that we can, without further assumptions, place the α -quantile of $u(h)$ in one of $T + 1$ intervals.

These are

- $Q_\alpha[u(h)] < 0$ if $P[u(h) < 0] \geq \alpha$.
- $d - 1 < Q_\alpha[u(h)] < d$ if $P[u(h) < d - 1] < \alpha$ and $P[u(h) < d] \geq \alpha, 0 < d \leq T$.

The above derivation supposes that a survey of the entire population enables precise determination of $P[u(h) < d], d = 0, \dots, T$. In practice, researchers may draw a random sample of the population and assume that nonresponse is random. Then empirical frequencies yield consistent estimates of these probabilities.

4. Empirical Considerations in QALY Measurement

Whereas the utilities of health state vectors were considered abstractly in Section 3, we now specialize to the QALY form of health-related utility prevalent in health economics. This section describes econometric issues involved in empirical research aiming to measure the distribution of QALYs in a population, given a specified medical treatment or other health intervention.⁴ Instructive discussions of some issues from the perspective of one regulatory body, the UK National Institute for Health and Care Excellence (NICE), are found in chapter 4 of NICE (2025). After defining basic concepts in Section 4.1, we consider estimation of mean QALYs in Section 4.2 and quantile welfare evaluation in Section 4.3.

4.1. The Anatomy of QALYs

NICE (2025), Chapter 4 describes the nature of a QALY measure in paragraph 4.3.2:

“4.3.2. A QALY combines both quality of life and life expectancy into a single index. In calculating QALYs, each of the health states experienced within the time horizon of the model is given a utility reflecting the health-related quality of life associated with that health state. The time spent in each health state is multiplied by the utility. Deriving the utility for a particular health state usually comprises 2 elements: measuring health-related quality of life in people who are in the relevant health state and valuing it according to preferences for that health state relative to other states (usually perfect health and death).”

Regarding the time horizon of the experienced health states, NICE (2025) writes:

⁴ Many considerations important in empirical practice are not discussed here, including the structures of TTO survey elicitation, sampling designs, sampling execution (including interviewer effects), cognitive burdens and challenges for subjects (e.g. understanding TTO logic). While outside the focus of this paper, one such consideration we view as essential to address concerns potential disconnects between conceptual characterizations of better and worse health states (EuroQol Research Foundation, 2023, page 20) and empirical findings that may not adhere to the monotonicity relationships implied by the conceptual characterizations. Post hoc repair of such empirical findings by imposing monotonicity relationships on such disobedient parameter estimates strikes us as potentially problematic.

“4.2.22. The time horizon for estimating clinical effectiveness and value for money should be long enough to reflect all important differences in costs or outcomes between the technologies being compared.”

“4.2.23. Many technologies have effects on costs and outcomes over a patient's lifetime. In these circumstances, a lifetime time horizon is usually appropriate. A lifetime time horizon is needed when alternative technologies lead to differences in survival or benefits that last for the remainder of a person's life.”

“4.2.25. A time horizon shorter than a patient's lifetime could be justified if there is no differential mortality effect between technologies and the differences in costs and clinical outcomes relate to a relatively short period.”

Recalling from Section 3 that $QALY_j(h) = u_j(h) = \sum_{t=1}^T q_j(h_t)$, the central considerations in measuring QALYs in practice concern measurement of the $q_j(h_t)$ and determination of the relevant time horizon(s) over which the $q_j(h_t)$ are to be learned.

Incremental QALYs between two treatments 0 and 1 that may yield different health-state time patterns are

$$(1) \quad QALY_j(h^1) - QALY_j(h^0) = \sum_{t=1}^T [q_j(h_t^1) - q_j(h_t^0)],$$

where h^0 and h^1 are the health state vectors experienced under treatments 0 and 1. If time to death varies between treatments, using the same time horizon T for both treatments is not problematic because $q_j(h_t^k) = 0$ at any time period at or after death. The recognition that, for a given treatment and individual, health states may vary over time is consistent with the paragraphs from NICE (2025) excerpted above.

In applications, measurement of QALYs typically relies on at least two types of data sources. One (D1) is a clinical (e.g. RCT) or population study that estimates the time patterns of situation-relevant (e.g. treatment-specific) health states h_t^k . The other (D2) is a population-level elicitation of the utilities

associated with the health states.⁵

Much of the empirical QALY literature has focused on estimation of population mean QALYs, $E(QALY)$, with QALY defined as above. One can conceive of $E(QALY)$ using the iterated expectation

$$(2) \quad E(QALY) = E[E(u(h)|h)].$$

The inner expectation fixes the health state h and uses the D2 data to learn its mean across the population distribution of $u_j(h)$. The outer expectation uses the D1 data to learn the mean of the inner expectation with respect to the population distribution of health states $h^k = (h_{jt}^k, \text{ all } j, \text{ all } t)$ that would occur under each treatment k .

Quantile welfare evaluation does not use knowledge of $E(QALY)$. Instead it uses knowledge of the relevant quantile of the population distribution of the QALYs measured using D1 and D2. While there is no corresponding law of iterated quantiles, the idea is similar. That is, variation in measured QALYs across the population arises from distinct two sources. One is individual variation in utility functions and the other is individual variation in the health states generated by a specified treatment.

4.2. TTO Measurement of Health State Utilities and Tobit Estimation of Mean Quality of Life

EuroQoL Foundation (2023) describes EQ-5D-based health-state utilities as follows:

“Each health state can potentially be assigned a summary index score based on societal preference weights for the health state. These weights, sometimes referred to as ‘utilities’, are often used to compute QALYs for use in health economic analyses. Health state index scores generally range from less than 0 (where 0 is the value of a health state equivalent to dead; negative values representing values as worse than dead) to 1 (the value of full

⁵ A third type of data (D3) is sometimes relevant. For instance, a key clinical study may study outcome measures y that differ from the EQ-5D categories h . In this case a data set that allows one to “map” between y and h may be used. See Wailoo et al. (2023).

health), with higher scores indicating higher health utility. The health state preferences often represent national or regional values and can therefore differ between countries/regions.”

Regarding empirical measurement, the range of observed health-state utilities in conventional TTO choice experiments is a bounded interval $[L, U]$, where U is one (“the value of full health”) and $-\infty < L \leq 0$ (“equivalent to dead” or “worse than dead,” but not infinitely worse than dead).

Empirical analysis of QALYs using EQ-5D data has mainly focused on estimation of the mean of $q(h_t)$ in some population of concern. Note that if each health-state utility $q(h_t)$ is bounded in $[L, U]$, then each of the t quality of life summands in the summation $\sum_{t=1}^T q_j(h_t)$ is also bounded in $[L, U]$, so the total QALYs is bounded in $[TL, TU]$.

Research using EQ-5D has generally interpreted the upper and lower bounds U and L differently. It has been presumed that health-related utility cannot logically exceed that of “full health,” so $U = 1$ is a true upper bound on utility. However, there is no similar consensus on how low individuals may perceive health states that are “worse than dead.” Hence, the lowest observed utility L is viewed as an artifact of the TTO measurement protocol, yielding a censored value of the true negative utility of very poor health states.

4.2.1. Estimation of Mean Quality of Life Using a Tobit Model

To cope with the assumed censoring, the UK NICE and other agencies have used TTO data to estimate a Tobit model (Tobin, 1957) of the inner expectation $E(u(h)|h)$, with estimation by the maximum likelihood method (Amemiya, 1973). Indeed, the study protocol for the UK valuation of EQ-5D-5L initiated in 2020 explicitly committed estimation using versions of the Tobit model. See Rowen *et al.* (2023).

The canonical Tobit specification assumes that an underlying latent (partially observed) quality of life variable q^* is distributed $N(xb, v)$. Here x is a covariate vector that includes h and may also include individual demographic attributes, b are the corresponding coefficients, and $v > 0$ is a constant variance

(i.e., the distribution of q^* is homoskedastic). Tobit is implemented using data on the observed x and on the observed counterpart of q^* , $q = L \cdot 1(q^* \leq L) + U \cdot 1(q^* \geq U) + q^* \cdot 1(L < q^* < U)$. Thus, the observed q is interpreted to be a doubly-censored version of q^* , even though U is logically a true upper bound on utility and only L is a censored lower bound.

The fragility of the Tobit model, manifest in the sensitivity of estimates to departures from the homoskedastic normal assumption, has long been known. See, for example, Hurd (1979) and Arabmazar and Schmidt (1983). A question that arises sometimes in estimation of Tobit models is whether the estimands of interest to decision makers are features of the conditional distribution of latent outcomes like $E(q^*|x) = xb$, or features of the conditional distribution of observed outcomes like $E(q|x)$, which does not in general equal xb .⁶ Neither is unambiguously correct or incorrect across all circumstances.

A related issue that should be resolved before implementing a Tobit estimation strategy is whether the bounds L and U represent censoring values or have some different meaning. Doubly-censored Tobit treats both as censoring values, suggesting that it is meaningful to conceive of values of q^* in $(-\infty, L)$ and in $(U, +\infty)$. Considering the lower bound L to express censoring is defensible since some methods of utility elicitation could in principle produce values less than L . It is not clear, however, how to conceive of values q^* in $(U, +\infty)$, given that U is defined as the value of full health. Reconciliation of these issues appears to us essential if one is contemplating Tobit-type estimation strategies.

Note that, if the Tobit model is specified correctly, the conditional mean $E(q^*|x)$ coincides with the conditional .5-quantile of q^* (the conditional median) because $N(xb, v)$ is symmetric around xb . In contrast, $E(q|x)$ does not generally coincide with the conditional median of q .

4.3. Quantile Welfare Evaluation of QALYs

To inform quantile welfare evaluation requires estimation of the decision-relevant quantiles of the

⁶ If a decision maker's interests exclusively concern $E(q|x)$, then alternatives to Tobit like fractional regression (Papke and Wooldridge, 1996) may in some instances be attractive.

distribution of the QALYs $\sum_{t=1}^T q_j(h_t)$, conditional on relevant covariates x . Nonparametric quantile estimation would be ideal. We provide a numerical illustration in Section 4.4.

Limitations of sample size may make parametric estimation desirable in practice. Suppose one considers observed quality of life q measured in the bound $[L, U]$ to be censored values of latent quality of life in $(-\infty, \infty)$, and that one observes individual covariates x . Then it is natural to consider implementing a doubly-censored quantile regression strategy proposed initially by Powell (1986); see also Koenker (2017). This method assumes that the quantile of interest is a linear function of x , but it makes no other substantive distributional assumption. In sharp contrast to the Tobit model, it does not require one to assume that quality of life is normally distributed and homoskedastic.

Importantly, Powell's approach does not require one to conceive of the values $q = L$ or $q = U$ as arising from a censoring process. In general, one observes non-zero point mass of the empirical distribution of q^* at L and U . If one interprets L as a fixed lower bound on the measurement of utility, then all quantiles of P are point identified. If one interprets L as a left-censoring point, then some quantiles of P —specifically those where the probability mass at L is larger than the quantile(s) under consideration—are not point identified. Being agnostic about the nature of L , one can say that the quantiles of P where the probability mass at L is larger than those quantiles are at most L .

4.4. Numerical Illustration

We present an illustrative simulation to showcase key features of evaluation using quantile versus utilitarian welfare in populations with heterogeneous preferences and health outcomes. The details of the simulation are described in the Appendix.

In our exercise we assume that an individual's health state while alive is time-invariant, as has commonly been assumed in EQ-5D TTO experiments. Quality of life in alive health state h is $q_j(h) = q(h; \theta_j)$, i.e., population preference heterogeneity arises from heterogeneity in the health-state

utility parameters θ_j . An individual who realizes health state h will experience total QALYs in that health state over a ten-year horizon of

$$QALY_j(h) = \sum_{t=1}^{10} q_j(h) \cdot s_{jt},$$

where s_{jt} is a binary indicator that j is alive in year t and $s_j = (s_{jt}, t = 1, \dots, 10)$. For example, $s_j = (1, 1, 1, 0, \dots, 0)$ means that person j is alive in years 1 through 3 and dead thereafter.

Value-elicitation exercises using EQ-5D commonly use a titrated TTO method (Devlin *et al.*, 2022) where subjects are offered a sequence of binary discrete choices. There are five dimensions of health. A subject is asked initially if they prefer ten years in full health ($h = 11111$) to ten years in comparator health state h . If the response is yes, then the subject is asked about preference between nine years in full health versus ten years in health state h , and so on.

We think it important to point out that binary choice TTO experiments yield interval rather than point measures of QALYs. For instance, a subject who answers yes to the initial 10-versus-10 question but no to the subsequent 9-versus-10 question reveals that their QALYs for living 9 years in health state h lies in the interval $[9, 10]$ and that $q_j(h)$ lies in the interval $[\frac{9}{10}, 1]$. The conventional EQ-5D practice using TTO data to estimate a Tobit model ignores the interval nature of observations, acting as if the experiments reveal precise QALY values. We do not follow this practice here. Instead, we use the interval measurement of QALYs by TTO data⁷ to estimate lower and upper bounds on the means and quantiles of QALY distributions. Moreover, we consider $[L, U]$ to be true lower and upper bounds on QALYs rather than values at which censoring occurs. See the Appendix for details.

In this exercise we address quantile versus utilitarian evaluation by contrasting estimated nonparametric bounds on quantiles of population distributions of realized QALYs with estimated bounds on their means. To facilitate nonparametric estimation, we simulate a large sample of one million

⁷ Two exceptions are noteworthy. First, the health-state utility of the full-health health state $h = 11111$ is exactly one in a TTO exercise, so QALYs in this health state exactly equals the number of years alive. Second, individuals who die in $t = 1$ have exactly zero QALYs, regardless of the health-state utility of their realized health state.

subjects, yielding sufficient observations for each feasible health state to enable us to consider each feasible state as a separate cell.⁸

We use the EQ-5D-3L health state structure, where health has five dimensions and each dimension has three levels (1, 2, 3), denoting full, moderate, and poor health respectively. Thus, there are $3^5 = 243$ possible health states. Years alive can range from 0 to 10. We specify a distribution $P(h, s)$ of realized health states and survival, described in the Appendix. Persons have utility functions over health states drawn from a specified distribution, described in the Appendix. This generates a population distribution of QALYs conditional on realized health states and years of survival, denoted $P(\text{QALY}|h, s)$.

We simulate by drawing health states (h, s) and utility functions at random and computing the resulting simulated empirical distribution of QALYs for each (h, s) pair. Recognizing that binary choice TTO experiments only bound QALYs to intervals, we obtain lower and upper bounds on mean QALYs $E(\text{QALY}|h, s)$ and on α -quantiles $Q_\alpha(\text{QALY}|h, s)$.

Table 1 summarizes the findings for selected values of h , averaged across the distribution of survival $P(s|h)$. In each case, we present bounds on $E(\text{QALY}|h)$ and $Q_\alpha(\text{QALY}|h)$ for $\alpha = .10, .25, .50, .75, \text{ and } .90$. The columns of the table give a value for h , the number of simulated observations experiencing that health state, and the bounds on means and quantiles of the distribution $P(\text{QALY}|h)$. For brevity we report results for only a selection of the feasible h values. The full results are available on request. See figure 1 for depiction of the bounds on $P(\text{QALY})$ over all simulated h and $P(\text{QALY}|h = 12311)$.

As should be anticipated, the bounds on QALY quantiles rise with the specified quantile. There has been a continuing discussion in health economics regarding measurement of social welfare by mean or median QALYs. Hence, it is of particular interest to compare these bounds. The table shows that, in these

⁸ Nonparametric estimation may not be feasible in practical settings where the sample may contain at most a few thousand subjects. The econometrics literature has studied partial identification and estimation of bounds on parametric models for mean and quantile regressions using interval measurement of outcomes. For parametric mean regression, see Manski and Tamer (2002) Section 4.5 and Beresteanu and Molinari (2008), Section 4. For parametric quantile regression, see Beresteanu and Sasaki (2021). In contrast to Tobit, these approaches parametrize only the regression of interest, not the entire conditional distribution.

simulations, the bound differ but overlap for all but one of the health states shown. The exception is the state $h = 11111$, where we obtain precise values of QALYs. In this state, the median is 10 and the mean is 7.62.

5. Conclusion

We believe that this paper's analysis should be of interest both conceptually and empirically. The conceptual interest may arise in large part because the paper's juxtaposition of utilitarian and quantile welfare may spur decision makers to assess or reassess the goals they are attempting to achieve in the populations whose welfare is of concern. Rethinking goals is vital if utilitarian and quantile welfare maximization lead to different recommendations about which policies to adopt. Unquestioningly treating utilitarian welfare maximization as the only reasonable way to evaluate policies seems to us indefensible, particularly when quantile welfare maximization offers a solidly grounded alternative. If nothing else, we hope to prod decision makers to think carefully about their welfare functions. Of course, if quantile welfare is relevant, the issue of *which* quantile(s) matter must be faced squarely.

Much of the paper's empirical interest arises because the analytical frameworks we propose should be of direct relevance across a variety of real-world evaluation settings. We have focused on evaluations based on EQ-5D and corresponding TTO-based interval utility measures, as these have been prominent in health technology assessment. Our results are applicable much more generally; indeed, they are more straightforward to implement if utilities are point- rather than interval-observed. Given that regression of health outcomes on personal covariates is often of interest in applied work, future research might productively focus on tools that would enable empirical researchers to straightforwardly implement nonparametric (ideally) or parametric versions of quantile interval regression, perhaps extending the work of Beresteanu and Sasaki (2021).⁹

⁹ Confronted with interval outcome measures and interested in pursuing quantile regression, researchers might be tempted to consider analytical shortcuts. Perhaps most obvious would be to compute interval

Appendix: The Simulation Process

A.1. Preference Distribution

In the 3L version of EQ-5D, $h_k \in \{1, 2, 3\}$ for $k = 1, \dots, 5$. In our simulations, we assume that the health-state utilities for individual j are given by

$$q_j(h) = q_j(h_1, \dots, h_5) = (1 + w_j) \cdot \left(\prod_{k \in \{1, \dots, 5\}} u_{jk}^{c_{jk}(h_k - 1)} \right) - w_j$$

Population preference heterogeneity arises from heterogeneity in u , c , and w . Each dimension's associated utility values $u_{jk} \in (0, 1)$ are drawn from independent uniform [.4, .9] distributions for $k = 1, \dots, 5$. Each dimension's parameters $c_{jk} \in (0, 1)$ are drawn from independent uniform [.3, .9] distributions. Potential for “worse than dead” health-state utilities is captured by the w_j , which are drawn from uniform [0, .4] distributions. Note that $q_j(1, 1, 1, 1, 1) = 1$ for all combinations of the u , c , and w . As specified, $q_j(h)$ must reside in the $(-.4, 1]$ interval, with a negative value indicating worse than dead. Thus, $L = -.4$ and $U = 1$ are bounds on feasible utility rather than censoring values.

As noted in the main text, the binary choice TTO data yielded by EQ-5D elicitation methods do not yield knowledge of specific values of $q_j(h)$. They yield intervals, revealed by the choices that subjects make between living for ten years in state h and living for a shorter period in full health. The simulations in the paper mimic this structure by assuming that the point realizations of the $q_j(h)$ are unknown but are observed to reside in known intervals.¹⁰ For simplicity we use the same bracketing method for worse-

midpoints and treat them as data. Beresteanu and Sasaki (2021) warn against this approach. An alternative would be consider a fully parametric strategy, estimating interval regression models akin to homoskedastic-normal Tobit models using maximum likelihood (e.g. Stata's *intreg* procedure). The resulting estimates of conditional distributions could be used to derive those distributions' conditional quantiles. This strategy is defensible if the homoskedastic-normal assumption correctly describes the data-generating process, although it effectively treats the upper bound as a right-censoring point. However, failure of a homoskedastic-normal assumption would render such estimates questionable.

¹⁰ There are two exceptions in which precise QALY values are known. One occurs when $h = 11111$ (“full health”), in which case utility is known to equal one for all combinations of u , c , and w . Hence, total

than-dead health states $q_j(h) < 0$.¹¹

A.2. Health State Distribution

Let the 5-dimension population distribution of h be $P(h)$ and suppose $P(h)$ has a multivariate ordered probit structure. For computational simplicity, $P(h)$ is assumed to be independent across its five dimensions. For each dimension, let the ordered probit cut points be defined such that $P(h_{jk}=1) > P(h_{jk}=2) > P(h_{jk}=3)$. Specifically $P(h_{jk}=1) = \Phi(t_{j1})$, $P(h_{jk}=2) = \Phi(t_{j2}) - \Phi(t_{j1})$, and $P(h_{jk}=3) = 1 - \Phi(t_{j2})$, where the t_{j1} are draws from uniform[0, 1] distributions and the t_{j2} are draws from uniform[t_{j1} , 2] distributions. Realized health states are drawn from these ordered probit distributions.

A.3. Survival Distribution

Mortality probabilities at each $t = 1, \dots, 10$ are specified as .01 times the sum of the health-state indicators; that is, mortality probability increases with poorer health-state realizations. As such, the mortality probabilities at each t range from .05 for health state $h = 11111$ to .15 for health state $h = 33333$. The mortality draws from uniform distributions are independent over the 10 time periods and across observations. If individual j dies at time t , then they will be dead at all subsequent times.

QALYs are known to equal one times the number of periods survived, which is specified in the choice experiment. The other exception occurs if an individual survives for no time periods, i.e., dies at $t = 1$. Then their total QALYs are known to equal zero regardless of the living health-state utility.

¹¹ EQ-5D value-elicitation exercises often treat negative (worse than dead) values differently by using methods like Composite-TTO, which pose further choice experiments specifying lengths of life beyond ten years when subjects reveal that they view some health states as worse than death. See Janssen et al. (2013).

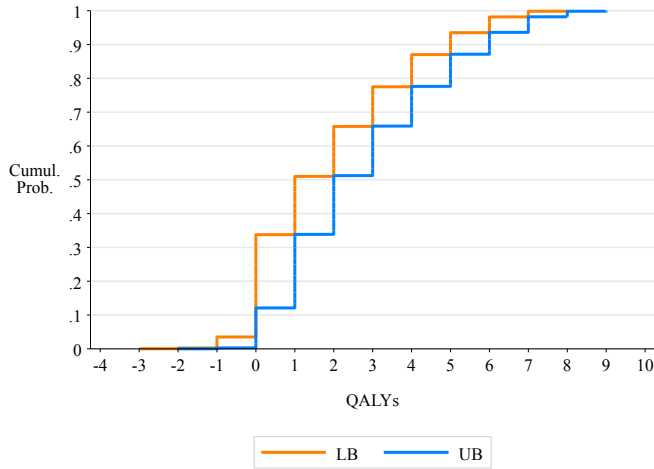
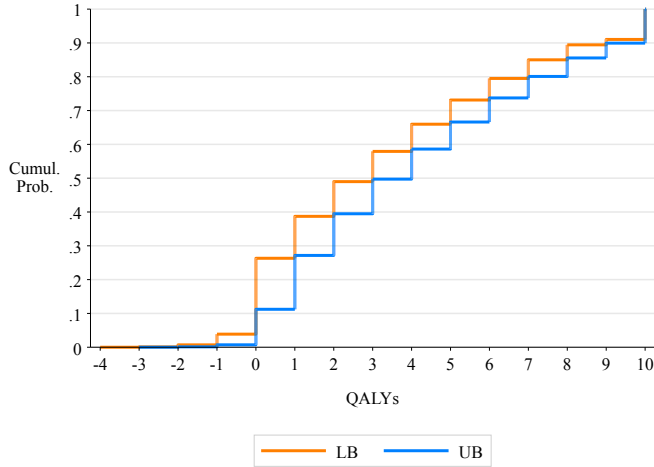
A.4. Simulation Details

The simulations set the number of observations equal to 1 million. In the simulations reported here, all 243 health states are realized with positive frequency. Stata's Mata programming language is used to generate the simulated data. This code is available on request.

Table 1: Summary of Simulations

h	N. Obs.	Quantiles										Means	
		.10		.25		.50		.75		.90		L	U
		L	U	L	U	L	U	L	U	L	U		
11111	150070	2	2	5	5	10	10	10	10	10	10	7.62	7.62
11113	28270	0	1	1	2	3	4	5	6	7	8	3.14	4.06
11221	11096	0	1	1	2	3	4	5	6	6	7	3.10	4.02
11312	7743	0	0	0	1	2	3	3	4	5	6	1.93	2.84
12311	7631	0	0	0	1	1	2	3	4	5	6	1.90	2.80
31311	5367	0	0	0	1	1	1	2	3	4	5	1.06	1.96
22211	3066	0	1	0	1	2	3	3	4	4	5	1.86	2.78
32112	2179	0	0	0	1	1	2	2	3	3	4	1.03	1.93
31122	2119	0	0	0	1	0	1	2	3	3	4	0.96	1.86
31212	2077	0	0	0	1	1	2	2	3	3	4	0.96	1.87
22113	1974	0	0	0	1	1	2	2	3	3	4	0.98	1.87
33112	1481	-1	0	0	0	0	1	1	2	2	3	0.35	1.25
32113	1451	-1	0	0	0	0	1	1	2	2	3	0.39	1.28
11332	1421	-1	0	0	0	0	1	1	2	2	3	0.41	1.30
33113	992	-1	0	-1	0	0	1	0	1	1	2	-0.06	0.83
12222	763	0	0	0	1	1	2	2	3	3	4	0.99	1.89
21322	582	-1	0	0	0	0	1	1	2	2	3	0.37	1.24
22213	562	-1	0	0	0	0	1	1	2	2	3	0.37	1.26
23123	403	-1	0	-1	0	0	1	0	1	2	3	0.01	0.93
23231	391	-1	0	-1	0	0	1	0	1	1	2	-0.12	0.77
21323	382	-1	0	-1	0	0	1	0	1	1	2	0.02	0.92
13323	287	-2	-1	-1	0	0	0	0	1	0	1	-0.43	0.44
32313	272	-1	0	-1	0	0	0	0	1	1	2	-0.35	0.54
33331	203	-2	-1	-1	0	0	0	0	1	0	1	-0.52	0.31
22232	135	-1	0	-1	0	0	1	0	1	1	2	-0.13	0.70
33222	94	-1	0	-1	0	0	0	0	1	1	2	-0.32	0.54
32323	59	-2	-1	-1	0	-1	0	0	1	0	1	-0.69	0.20

Figure 1: Lower and Upper Bounds on Simulated QALY Distributions
(Top Panel—All Realized Health States;
Bottom Panel—Realized Health State $h = 12311$)



References

- Amemiya, T. (1973), "Regression Analysis When the Dependent Variable is Truncated Normal," *Econometrica*, 41, 997-1016.
- Arabmazur, A. and P. Schmidt (1983), "An Investigation of the Robustness of the Tobit Estimator to Non-Normality," *Econometrica*, 50, 1055-1063.
- Austin, P. (2002), "A Comparison of Methods for Analyzing Health-Related Quality-of-Life Measures," *Value in Health*, 5, 329-337.
- Beresteanu, A. and F. Molinari (2008), "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763-814.
- Beresteanu, A. and Y. Sasaki (2021), "Quantile Regression with Interval Data," *Econometric Reviews*, 40, 562-583.
- Devlin, N., B. Roudijk, and K. Ludwig, Editors (2022), *Value Sets for EQ-5D-5L*, Springer.
- EuroQol Research Foundation (2023), *EQ-5D-5L User Guide*, Rotterdam. <https://euroqol.org/wp-content/uploads/2023/11/EQ-5D-5LUserguide-23-07.pdf>
- Hurd, M. (1979), "Estimation in Truncated Samples When There is Heteroscedasticity," *Journal of Econometrics*, 11, 247-258.
- Janssen, B.M.F., M. Oppe, M.M. Versteegh, and E.A. Stolk (2013), "Introducing the Composite Time Trade-off: A Test of Feasibility and Face Validity." *European Journal of Health Economics* 14 (Suppl 1): 5–13. <https://doi.org/10.1007/s10198-013-0503-2>
- Koenker, R. (2017), "Quantile Regression: 40 Years On," *Annual Review of Economics*, 9, 155-176.
- Levy, H. and Y. Kroll (1978), "Ordering Uncertain Options with Borrowing and Lending," *Journal of Finance*, 33, 553-574.
- Manski, C. (1988), Ordinal Utility Models of Decision Making under Uncertainty, *Theory and Decision*, 25, 79-104.
- Manski, C. (1997), "Monotone Treatment Response," *Econometrica*, 65, 1311-1334.
- Manski, C. (2024), *Discourse on Social Planning under Uncertainty*, Cambridge: Cambridge University Press.
- Manski, C. and E. Tamer (2002), "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519-546.
- Manski, C. and A. Tetenov (2023), "Statistical Decision Theory Respecting Stochastic Dominance," *Japanese Economic Review*, 74, 447-469.
- Mirrlees J. (1971), "An Exploration in the Theory of Optimal Income Taxation," *Review of Economic Studies*, 38, 175-208.

Mullahy, J. (2021), "Discovering Treatment Effectiveness via Median Treatment Effects—Applications to COVID-19 Clinical Trials," *Health Economics* 30, 1050-1069.

National Institute for Health and Care Excellence (NICE) (2025), *NICE Health Technology Evaluations: The Manual*, NICE Process and Methods. Published 31 January 2022, Last Updated 14 July 2025. <https://www.nice.org.uk/process/pmg36>

Papke, L and J. Wooldridge (1996), "Econometric Methods for Fractional Response with an Application to 401(K) Plan Participation Rates," *Journal of Applied Econometrics*, 11, 619-632.

Powell, J. (1986), "Censored Regression Quantiles," *Journal of Econometrics*, 32, 143-155.

Rawls, J. (1971), *A Theory of Justice*, Cambridge: Harvard University Press.

Rostek, M. (2010), "Quantile Maximization in Decision Theory," *Review of Economic Studies*, 77, 339-371.

Rowen, D., C. Mukuria, N. Bray, J. Carlton, S. Cooper, L. Longworth, D. Meads, C. O'Neill, and Y. Yang (2023), "UK Valuation of EQ-5D-5L, a Generic Measure of Health-Related Quality of Life: A Study Protocol," *Value in Health*, 26, 1625-1635.

Samuelson, P. (1947), *Foundations of Economic Analysis*, Cambridge, MA: Harvard University Press.

Sen, A. (1977), "On Weights and Measures: Informational Constraints in Social Welfare Analysis," *Econometrica*, 45, 1539-1572.

Tobin, J. (1957), "Estimation of Relationships for Limited Dependent Variables," *Econometrica*, 26, 24-36.