NBER WORKING PAPER SERIES

INDIVIDUAL RESULTS MAY VARY: ELEMENTARY ANALYTICS OF INEQUALITY-PROBABILITY BOUNDS, WITH APPLICATIONS TO HEALTH-OUTCOME TREATMENT EFFECTS

John Mullahy

Working Paper 23603
http://www.nber.org/papers/w23603

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2017

For helpful discussions and comments, and with the standard disclaimer, I would like to thank Chris Adams, Bruce Hansen, M.-J. Lee, David Madden, Chuck Manski, Joao Santos Silva, Justin Sydnor, Dave Vanness, and participants in presentations at the Barcelona Conference on the Economics of Addiction and Health Inequality, the York Centre for Health Economics, ISEG-Lisbon, Resources for the Future, University College Dublin, the University of Chicago, the University of Coimbra, and the University of Wisconsin-Madison. Parts of this work were supported by the RWJF Health & Society Scholars Program, RWJF Evidence for Action Grant 73336, and NICHD grant P2CHD047873 to the Center for Demography and Ecology, all at UW-Madison, and by the UCD Geary Institute during a sabbatical visit. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2017 by John Mullahy. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Individual Results May Vary: Elementary Analytics of Inequality-Probability Bounds, with Applications to Health-Outcome Treatment Effects
John Mullahy
NBER Working Paper No. 23603
July 2017
JEL No. I1

ABSTRACT

While many results from the treatment-effect and related literatures are familiar and have been applied productively in health economics evaluations, other potentially useful results from those literatures have had little influence on health economics practice. With the intent of demonstrating the value and use of some such results in health economics applications, this paper focuses on one particular class of parameters that describe probabilities that one outcome is larger or smaller than other outcomes, namely inequality probabilities. While the properties of such parameters have been explored in the technical literature, they have scarcely been considered in informing practical questions in health evaluations. This paper discusses how such probabilities can be used informatively, and describes how they might be identified or bounded given standard sampling assumptions and information only on marginal distributions of outcomes. Graphical and algebraic exposition reveals the logic supporting these results, as well as their empirical implementation, to be quite straightforward. Applications to health outcome evaluations are presented and discussed throughout.

John Mullahy
University of Wisconsin-Madison
Dept. of Population Health Sciences
787 WARF, 610 N. Walnut Street
Madison, WI 53726
and NBER
jmullahy@facstaff.wisc.edu

1. Introduction

Nivolumab Treatment for Non-Small-Cell Lung Cancer

Nivolumab—a biological product marketed by Bristol-Myers Squibb (BMS) in the U.S. as Opdivo—has several FDA-approved indications, one being previously treated advanced non-small-cell lung cancer (NSCLC). For treatment of NSCLC, two primary outcomes pre-specified[1] in a pivotal, phase-III randomized trial of nivolumab versus docetaxel (chemotherapy) were overall survival time and one-year overall survival rate. In summarizing that study Borghaei et al., 2015, report:

> Overall survival was significantly longer with nivolumab than with docetaxel… At the time of the interim analysis (minimum followup for overall survival, 13.2 months), the median overall survival was 12.2 months (95% confidence interval [CI], 9.7 to 15.0) with nivolumab and 9.4 months (95% CI, 8.1 to 10.7) with docetaxel, representing a 27% lower risk of death with nivolumab (hazard ratio, 0.73; 96% CI, 0.59 to 0.89; P = 0.002). The overall survival rate at 1 year was 51% (95% CI, 45 to 56) with nivolumab and 39% (95% CI, 33 to 45) with docetaxel.

Figure 1 depicts the data from which these results are computed.[2] The two reported estimates of overall survival are shown: a difference in median survival time of 2.8 months in panel (a); and a difference in twelve-month survival probability of .12 in panel (b).

A BMS direct-to-consumer advertising campaign prominent in the U.S. in 2017 has boasted that Opdivo treatment for NSCLC offers "a chance to live longer."[3]

A Chance to Live Longer?

What might the Borghaei et al. results say about the "chance to live longer" pitched in the Opdivo ads? In claiming "a chance to live longer" at least two questions arise logically. The first is: "a chance to live longer," compared to what? The second is: "a chance to live longer,"

[1] ClinicalTrials.gov study NCT01673867.

[2] The smoothed empirical distribution functions depicted in figure 1 are approximated from the data depicted in figure 1A in Borghaei et al., 2015, as survival curves.

[3] http://www.opdivo.com/advanced-nsclc, accessed April 30 2017. BMS applied on January 31, 2017, for a U.S. trademark for "A CHANCE TO LIVE LONGER" (U.S. Patent and Trademark Office, Serial No. 87319390). At the time this draft was completed, the status of that application was "Under Examination."

measured how? In light of the Borghaei et al. results, "compared to what" might be answered reasonably as "treatment with docetaxel" or "treatment with other relevant comparators."

Of greater interest in this paper, however, are questions in line with the second one: measured how? Given the outcomes studied in Borghaei et al., one reasonable measure of "a chance to live longer" might be a difference in median survival times between the two treatments: subjects had a 50 percent chance to live longer than 12.2 months with nivolumab treatment compared with a 50 percent chance to live longer than 9.4 months with docetaxel. Alternatively, "a chance to live longer" might reasonably be characterized in terms of twelve-month survival probabilities: patients treated with nivolumab showed a twelve-percent greater chance to live at least twelve months longer than did patients treated with docetaxel.[4]

Letting t_{niv} and t_{doc} denote a patient's survival time with nivolumab or docetaxel treatment, then "living longer" amounts essentially to $t_{niv} > t_{doc}$. While the 2.8-month difference in median survival times or the .12 difference in twelve-month survival times hints at a relationship like this, the notion of "a chance to live longer" is something different. "A chance to live longer" might reasonably be translated as the probability a patient treated with nivolumab lives longer—not some particular amount longer,[5] just longer by some unspecified amount—than had they otherwise been treated with docetaxel, i.e. $\Pr\left(t_{niv} > t_{doc}\right)$.[6] If a patient asks: "If I'm treated with nivolumab, what is the chance that I'll live longer than if I'm treated with docetaxel?", then one number that answers this question is $\Pr\left(t_{niv} > t_{doc}\right)$.

[4] Either of these characterizations is of the sort that might be advanced as the basis of FDA marketing-approval applications or of other more general efficacy or effectiveness claims.

[5] Cost-effectiveness questions would often balance "how many weeks longer" against cost across comparators. If such outcome and cost measures are converted into univariate net-benefit (NB) measures then the choice between comparators amounts to knowing whether or not $NB_j > NB_k$.

[6] DTC advertising for Keytruda (pembrolizumab; Merck), whose indications for NSCLC are similar to Opdivo's, uses the catchphrase "a chance for a longer life." (https://www.keytruda.com/non-small-cell-lung-cancer/, accessed April 30, 2017). Entresto (sacubitril/valsartan; Novartis), a treatment for chronic heart failure, is promoted in DTC ads to "help increase your chances of more tomorrows" (http://www.entresto.com/info/about-entresto.jsp, accessed April 30, 2017).

Inequality probabilities like this are the main focus of this paper. In comparing two outcomes in a population exhibiting outcome heterogeneity, questions about the chance or probability that one outcome exceeds the other may be natural to pose. How one might analyze such questions is the main purpose of this paper.

Why Might Inequality Probabilities Be of Interest?

Let y_0 and y_1 be two outcomes of interest (e.g. $y_0 = t_{doc}$ and $y_1 = t_{niv}$). The inequality probability[7] $\Pr\left(y_1 > y_0\right)$ provides an intuitive characterization of the extent to which one outcome is stochastically larger than another. This can be appreciated from its definition,

$$\Pr\left(y_1 > y_0\right) = \int_{-\infty}^{\infty}\int_{y_0}^{\infty} f\left(y_0, y_1\right) dy_1\, dy_0\,, \tag{1}$$

wherein $f\left(y_0, y_1\right)$ is the joint probability density of y_0 and y_1. $\Pr\left(y_1 > y_0\right)$ is sometimes referred to as "fraction who benefit" (Huang et al. 2016; see also Aakvik et al., 2005). Unlike familiar criteria based on population expected benefit, $E\left[y_1 - y_0\right]$, measures like $\Pr\left(y_1 > y_0\right)$ are relevant indicators in voting (e.g. median voter, majority rule, etc.), strict-Pareto, and other social choice contexts (e.g. Coate, 2000, Gerber and Lewis, 2004, Jacob and Lundin, 2005, and Pauly, 1989; also see Heckman et al., 1997, for general perspectives).

Inequality probabilities also play a central role in stochastic settings where the benefit associated with a choice depends on the ordering among but not the magnitudes of competing outcomes, for instance a payoff (V) from choosing the winner in an M-participant competition (e.g. a horserace, a basketball game, or an exclusive therapeutic-category formulary listing). In such cases y_j might measure speed, score, therapeutic cost-effectiveness, etc. In such a competition the realized benefit from selecting competitor j is

$$B_j = V \times \prod_{k \neq j} 1\left(y_j > y_k\right), \tag{2}$$

[7] The term "inequality probability" is used in this paper to refer to parameters $\Pr\left(u > v\right)$ or $\Pr\left(u \geq v\right)$ for arbitrary and possibly jointly distributed variables of interest, u and v.

with corresponding expected benefit[8] (using standard "$\wedge$" notation for "and"):

$$E\left[B_j\right] = V \times \Pr\left(\left(y_j > y_1\right) \wedge \ldots \wedge \left(y_j > y_{j-1}\right) \wedge \left(y_j > y_{j+1}\right) \wedge \ldots \wedge \left(y_j > y_M\right)\right), \quad (3)$$

or, in the two-outcome case,

$$E\left[B_j\right] = V \times \Pr\left(y_j > y_k\right). \quad (4)$$

Finally, reconsider the nivolumab example. If it is of interest[9] to know the difference $\Pr\left(t_{niv} \geq 12\right) - \Pr\left(t_{doc} \geq 12\right)$, then $\Pr\left(t_{niv} \geq t'\right) - \Pr\left(t_{doc} \geq t'\right)$ may also be of interest for other $t' \neq 12$ or over all possible t'.[10] Yet these are different considerations than those involving $\Pr\left(t_{niv} \geq t_{doc}\right)$ whose definition in (1) embeds consideration of all values of $\left(t_{doc}, t_{niv}\right)$. Only by reference to a particular decision criterion might it be determined which such parameters should be of interest.

Summary Outcome Measures Used in Evaluations

Asking different questions about relationships between two outcomes leads logically to different ways to characterize and summarize statistically such outcomes in heterogeneous populations. In essence the previous discussion posed questions about whether one outcome (say y_1) is larger than another (say y_0), and focused on a particular metric of comparison, $\Pr\left(y_1 > y_0\right)$. Whether the outcomes of interest are survival times or perhaps other outcome-

[8] When the y_j are random utilities associated with different choice prospects, quantities like the probability in (3) are familiar from the multinomial discrete-choice literature.

[9] Presumably this quantity is of interest since it is one of the study's primary endpoints; see ClinicalTrials.gov study NCT01673867 and U.S. FDA, 2015. Why a particular value of t' is privileged merits consideration. Whether for parsimony, for convenience, due to biostatistical or regulatory convention, or for other reasons is often not obvious. While such choice should ideally square with decisionmakers' loss functions, it is rarely made explicit that it does; see Manski, 1998, 2007.

[10] If it converges the integral of $\Pr\left(t_{niv} \geq t'\right) - \Pr\left(t_{doc} \geq t'\right)$ over t'—characterizing second-order stochastic dominance— equals $E\left[t_{niv}\right] - E\left[t_{doc}\right]$.

relevant metrics—better, lower, greater, faster, clearer, easier, safer, longer-acting, cheaper, etc.—the same basic ideas apply. Of interest is the probability that y_1 is "better" than y_0, not "how much better" it might be.

Of course other evaluation-oriented metrics are encountered commonly in empirical health research. Letting $F_j(y)$ denote the population marginal distribution for outcome y_j and $V(\ldots)$ denote some statistical functional defined on $F_j(y)$(e.g. moment, quantile, probability, etc.), empirical investigations focus typically[11] on $V(F_0(y))$ and $V(F_1(y))$ as the summary measures to be estimated, and on some contrast between them—most typically, their difference—as the basis of a treatment-effect, comparative-effectiveness, or other claim. For example, the two primary outcomes in the Borghaei et al. study correspond to $V(F_j(y))=\text{med}(F_j(y))$ and $V(F_j(y))=F_j(y)$.

While the specification of $V(\ldots)$ and its estimation from sample data are broadly important considerations, this paper's specific concern is how observed data on the marginal distributions of two or more outcomes can be used to at least partially inform decisionmakers about inequality probabilities $\Pr(y_1 > y_0)$ and related parameters. When outcomes are observed jointly such an exercise is straightforward; the challenge in knowing $\Pr(y_1 > y_0)$ is when, for whatever reason, y_0 and y_1 are not observed together at the subject level.[12]

This Paper

Relationships to Existing Literature

This paper's focus intersects several broad themes that have been well developed in

[11] Stochastic dominance comparisons are an obvious exception to this form of comparison, as are measures involving any features of the joint distribution of y_0 and y_1.

[12] See Imbens and Wooldridge, 2009, p. 17, and Abbring and Heckman, 2007, p. 5151, for views on why decisionmakers might or mightn't want to "bother" identifying features of joint distributions.

the literature: treatment-effect estimation and heterogeneous treatment effects[13]; decisionmaking criteria in stochastic environments[14]; and point versus interval identification of treatment effects.[15] These broader literatures are not surveyed here although references to specific work are made when useful. The work most closely related to this paper includes that of Heckman and coauthors[16], a series of studies by Fan and coauthors[17], as well as studies by Adams, 2013[18], Basu and Thariani, 2016, Firpo and Ridder, 2008, Lee, 2000, and Manski, 1997. This paper's main results and a discussion of their applicability to a range of policy questions were discussed in a much earlier working paper by the author (Mullahy, 2005).

Motivation and Plan

The paper is motivated mainly by the observation that there are important and potentially useful results on inequality probabilities of the sort examined here that—while established in the technical literature—have thusfar had little impact on health economics research.[19] In particular, this paper attempts to exposit (relying often on simple graphical depictions) the elementary features of such results and then extend and apply them to contexts of interest in health economics.

Until section 4 the paper's results are largely not new; indeed, the paper's main results on inequality probabilities presented in section 3 are just tailored applications of Fréchet-

[13] Angrist, 2004; Athey and Imbens, 2006; Basu et al., 2007; Bitler et al., 2006; Borah et al., 2011; Chan and Hamilton, 2006; Hauck et al., 2000; Horwitz et al., 1996; Huang et al., 2016; Koenker and Bilias, 2001; Kravitz et al., 2004; Vanness and Mullahy, 2012; Willke et al., 2012. Imbens and Wooldridge, 2009, provide an comprehensive overview of many of these issues.

[14] Gerber and Lewis, 2004; Grandmont, 1978; Jacob and Lundin, 2005; Stinnett and Mullahy, 1998.

[15] Manski, 1999, 2007.

[16] Aakvik et al., 2005, Abbring and Heckman, 2007, Carneiro et al., 2001, Heckman, 2001, and especially Heckman et al., 1997.

[17] Fan et al., 2014, 2017; Fan and Park, 2010, 2012.

[18] As this draft was being completed the author was made aware of the paper by Adams, 2013, whose approach and examples overlap with some of this paper's.

[19] Exceptions include Adams, 2013, Cameron et al., 2004, and Huang et al., 2016.

Boole probability bounds.[20] Yet their discussion in technical literatures distant from health economics may have hindered their application in health economics and elsewhere. Describing, extending, and implementing these results in health economics contexts are the goals of this paper; at a minimum it is hoped that the paper provides a useful practitioner's guide.

The plan is as follows. Section 2 describes the main assumptions and notation. Section 3 presents the results on probability bounds. Section 4 extends the main results in several directions and offers examples. Section 5 considers the application of the main results to cost-effectiveness analysis. Section 6 discusses bounds when more than two outcomes are of interest. Section 7 considers empirical implementation: sampling, estimation, and inference. Section 8 summarizes.

2. Definitions, Assumptions, and Notation

The setup here is familiar in the treatment-effect literature. M+1 outcomes of interest, $\left(y_0, y_1, \ldots\right)$, are jointly distributed in the population according to $F\left(y_0, y_1, \ldots\right)$ with corresponding joint probability density denoted $f\left(y_0, y_1, \ldots\right)$.[21] $F\left(y_0, y_1, \ldots\right)$ might be interpreted as representing a population heterogeneous in outcomes or as a joint distribution of random variables.[22]

For now assume that there are two outcomes of interest, $\left(y_0, y_1\right)$, although more-general cases are considered in section 5. Unless noted otherwise $\left(y_0, y_1\right)$ are assumed to be continuously distributed. To be consistent with the technical definition of distribution

[20] The main results here involve set or interval identification, or probability bounds, of the sort studied and advocated forcefully by Manski. While there may be increasing receptivity by analysts of set identification, point identification is still the standard in many contexts (e.g. FDA regulation).

[21] This notation is informal; formally, $F\left(c_0, c_1, \ldots\right) = \Pr\left(\left(y_0 \le c_0\right) \wedge \left(y_1 \le c_1\right) \wedge \ldots\right)$.

[22] Outcomes are denoted in lower-case to keep notation concise. Distinctions between random variables and realizations should be clear from context.

functions the focus will on $\Pr\left(y_1 \geq y_0\right)$ instead of $\Pr\left(y_1 > y_0\right)$ although these are essentially the same with continuously distributed outcomes.[23] The particular y_j measures may be ratio-scale, interval-scale, ordinal, or any measure for which strict or weak inequality provides a meaningful comparison.

The population marginal distribution functions for the y_j are denoted $F_j\left(y\right) = \Pr\left(y_j \leq y\right)$, j=0,1, for all y in their respective supports $S_j = \left[L_j, U_j\right]$. Of course the $F_j\left(y\right)$ are related to $F\left(y_0, y_1\right)$ via $F_j\left(y\right) = \int_{y_j \leq y} \int_{S_k} F\left(y_0, y_1\right)\, dy_k dy_j$, $j \neq k$. Until section 7 "conditional on **x**" can be assumed tacitly if appropriate, but will not be made explicit unless useful; the role of covariates **x** is revisited in section 7. Moreover until section 7 the discussion is concerned only with population distributions and identification; considerations of sampling, estimation, and inference are deferred until then. Define the subject-level difference $\Delta_{01} = y_0 - y_1$.[24] In the population Δ_{01} is often considered a treatment effect but in general is just some contrast of interest. Understanding Δ_{01} is challenging when only one of the y_j is observable. Define the population distribution of Δ_{01} as $F_{\Delta_{01}}\left(c\right) = \Pr\left(y_0 - y_1 \leq c\right) = \Pr\left(y_1 + c \geq y_0\right)$. Of interest in most of what follows is $c = 0$, or $\Pr\left(y_1 \geq y_0\right)$. $\Pr\left(y_1 \geq y_0\right)$ is thus one feature of the treatment-effect distribution.

3. Main Results: Bounds on Inequality Probabilities

Revisiting the Nivolumab vs. Docetaxel Example

To motivate the general results discussed below, consider again the nivolumab vs.

[23] For discrete outcomes the difference between weak and strict inequality will matter; see below.

[24] Subject-indexing subscripts are suppressed unless useful for clarity. Note that the 0 and 1 subscripts are reversed from what is typical in the literature. Economists often consider such contrasts in a Rubin counterfactual framework, but they are also relevant in other contexts where information about the jointness properties of $F\left(y_0, y_1\right)$ is absent.

docetaxel twelve-month-survival results discussed in section 1. In a population, t_{niv} and t_{doc} will in general be jointly distributed even if at the subject level only one of them is observable. With reference to figure 2 wherein roman numerals denote the four subspaces with origin $\left(t_{doc},t_{niv}\right)=\left(12,12\right)$, the reported result on twelve-month survival, $\Pr\left(t_{niv}\geq 12\right)-\Pr\left(t_{doc}\geq 12\right)$ =.12, can be obtained as[25]

$$\begin{aligned}\Pr\left(t_{niv}\geq 12\right)-\Pr\left(t_{doc}\geq 12\right) &= \Pr\left(\left(t_{doc},t_{niv}\right)\in \mathrm{I}\cup\mathrm{II}\right)-\Pr\left(\left(t_{doc},t_{niv}\right)\in \mathrm{I}\cup\mathrm{IV}\right)\\ &= \Pr\left(\left(t_{doc},t_{niv}\right)\in \mathrm{II}\right)-\Pr\left(\left(t_{doc},t_{niv}\right)\in \mathrm{IV}\right)\\ &= .12\end{aligned} \qquad (5)$$

Now suppose outcomes are binary with $q_j=1\left(t_j\geq 12\right)$, $j\in\left\{doc,niv\right\}$, being indicators of twelve-month survival under the two treatments. The general joint and marginal probability structure is shown in panel (a) of table 1. Note that for the strict inequality event $q_{niv}>q_{doc}$

$$\Pr\left(q_{niv}>q_{doc}\right) \;=\; \Pr\left(q_{doc}=0 \wedge q_{niv}=1\right) \;=\; \pi_{01} \;=\; \Pr\left(\left(q_{doc},q_{niv}\right)\in \mathrm{II}\right). \qquad (6)$$

Bounding π_{01} is straightforward using Fréchet-Boole probability bounds. The best bounds on π_{01} knowable from the marginals π_j are

$$\max\left\{0,\pi_1-\pi_0\right\} \;\leq\; \pi_{01} \;\leq\; \min\left\{1-\pi_0,\pi_1\right\}. \qquad (7)$$

The lower bound, $\pi_1-\pi_0$, is $\Pr\left(\left(t_{doc},t_{niv}\right)\in \mathrm{II}\right)-\Pr\left(\left(t_{doc},t_{niv}\right)\in \mathrm{IV}\right)$, coinciding with (5). Applying this result to the nivolumab example one finds $.12 \leq \pi_{01} \leq .51$, i.e. notwithstanding sampling error $\Pr\left(q_{niv}>q_{doc}\right)$ is at least .12 but not greater than .51; see panel (b) of table 1.

General Results: Bounding $Pr\left(y_1 \geq y_0\right)$ using Fréchet-Boole Probability Bounds

For arbitrary, jointly distributed variables $\left(z_a,z_b\right)$ and corresponding sets Z_a and Z_b,

[25] At this point these estimates are treated as if population parameters. This example's empirical properties are considered in section 7.

the Fréchet-Boole lower bound ("FLB") on the joint probability of the events $z_j \in Z_j$ is:

$$\Pr\left(z_a \in Z_a \ \wedge \ z_b \in Z_b\right) \ \geq \ \max\left\{\left(\Pr\left(z_a \in Z_a\right)+\Pr\left(z_b \in Z_b\right)-1\right),0\right\}, \tag{8}$$

which is informative if $\Pr\left(z_a \in Z_a\right)+\Pr\left(z_b \in Z_b\right)>1$. For disjunctions ("or", symbolized "$\vee$"),

$$\Pr\left(z_a \in Z_a \ \vee \ z_b \in Z_b\right) \ \geq \ \max\left\{\Pr\left(z_a \in Z_a\right),\Pr\left(z_b \in Z_b\right)\right\}. \tag{9}$$

The corresponding upper bounds ("FUB") are

$$\Pr\left(z_a \in Z_a \ \wedge \ z_b \in Z_b\right) \ \leq \ \min\left\{\Pr\left(z_a \in Z_a\right),\Pr\left(z_b \in Z_b\right)\right\}, \tag{10}$$

which is informative so long as either of the $\Pr\left(z_j \in Z_j\right)$ is less than one. For disjunctions,

$$\Pr\left(z_a \in Z_a \ \vee \ z_b \in Z_b\right) \ \leq \ \min\left\{\Pr\left(z_a \in Z_a\right)+\Pr\left(z_b \in Z_b\right),1\right\}, \tag{11}$$

which is informative if the sum of the $\Pr\left(z_j \in Z_j\right)$ is less than one.

For arbitrary y', consider the events $y_0 \leq y'$ and $y_1 > y'$. Applying (8) gives

$$\begin{aligned}\Pr\left(y_0 \leq y' \ \wedge \ y_1 > y'\right) \ &\geq \ \max\left\{\Pr\left(y_0 \leq y'\right)+\Pr\left(y_1 > y'\right)-1,0\right\} \\ &= \ \max\left\{F_0\left(y'\right)+\left(1-F_1\left(y'\right)\right)-1,0\right\} \\ &= \ \max\left\{F_0\left(y'\right)-F_1\left(y'\right),0\right\}.\end{aligned} \tag{12}$$

This result is illustrated in figure 3(a) depicting $\left(y_0,y_1\right)$-space and illustrative isodensity contours of $f\left(y_0,y_1\right)$ drawn using $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim \mathrm{BVN}\left(\begin{bmatrix} 4.1 \\ 5.1 \end{bmatrix},\begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$. The arbitrary y' is indicated on both axes. Let $P\left(J\right)$ denote $\Pr\left(\left(y_0,y_1\right)\in J\right)$, where J is any of the six subspaces I'_A, I'_B, ... , IV' whose common origin is indicated in the figure at $y_0 = y_1 = y'$. Then:

$$\Pr\left(y_0 \leq y'\right) \ = \ P\left(II'\right)+P\left(III'_A\right)+P\left(III'_B\right) \tag{13}$$

$$\Pr\left(y_1 > y'\right) = P\left(II'\right) + P\left(I_A^{'}\right) + P\left(I_B^{'}\right) \tag{14}$$

$$\Pr\left(y_0 \leq y'\right) + \Pr\left(y_1 > y'\right) - 1 = P\left(II'\right) - P\left(IV'\right) \tag{15}$$

$$\Pr\left(y_0 \leq y' \wedge y_1 > y'\right) = P\left(II'\right) \tag{16}$$

If it exceeds zero, (15) is the FLB on $\Pr\left(y_0 \leq y' \wedge y_1 > y'\right)$; for $P\left(IV'\right) > 0$ this is smaller than the true probability in (16). Thus,

$$\begin{aligned} \Pr\left(y_1 \geq y_0\right) &= P\left(I_B^{'}\right) + P\left(II'\right) + P\left(III_A^{'}\right) \\ &\geq P\left(II'\right) \\ &\geq P\left(II'\right) - P\left(IV'\right) \\ &= \left(P\left(II'\right) + P\left(III_A^{'}\right) + P\left(III_B^{'}\right)\right) - \left(P\left(III_A^{'}\right) + P\left(III_B^{'}\right) + P\left(IV'\right)\right) \\ &= F_0\left(y'\right) - F_1\left(y'\right), \end{aligned} \tag{17}$$

wherein line three, $P\left(II'\right) - P\left(IV'\right)$, is the FLB from (15). Thus, using only the marginals $F_j\left(y\right)$ potentially informative lower bounds on $\Pr\left(y_0 \leq y' \wedge y_1 > y'\right)$ and thus $\Pr\left(y_1 \geq y_0\right)$ are obtained. Analogously, if informative the FUB on $\Pr\left(y_0 \leq y' \vee y_1 > y'\right)$ follows from (11) as $F_0\left(y'\right) + \left(1 - F_1\left(y'\right)\right)$, seen by noting that $P\left(I'\right) + P\left(II'\right) + P\left(III'\right) \geq P\left(I_B^{'}\right) + P\left(II'\right) + P\left(III_A^{'}\right) = \Pr\left(y_1 \geq y_0\right)$ with reference to figure 3(a), then applying (11) for the event $\left(y_0 \leq y' \vee y_1 > y'\right)$.

Best Bounds on Pr $\left(y_1 \geq y_0\right)$

Since $\Pr\left(y_1 \geq y_0\right) \geq \Pr\left(y_1 > y_0\right) \geq \Pr\left(y_0 \leq y' \wedge y_1 > y'\right)$ a nonzero FLB on $\Pr\left(y_0 \leq y' \wedge y_1 > y'\right)$ is partially informative as a lower bound on $\Pr\left(y_1 \geq y_0\right)$. Since y' is arbitrary, however, such a bound will generally not be a sharp or best possible such bound on $\Pr\left(y_1 \geq y_0\right)$. Intuitively from (17), a best lower bound on $\Pr\left(y_1 \geq y_0\right)$ is defined by determining the value(s) of y such that the difference between $F_0\left(y\right)$ and $F_1\left(y\right)$ is maximized.

To show this and some of its implications, the paper by Fan and Park, 2010, (henceforth FP) is especially useful, particularly since its results help structure empirical investigations as discussed in section 7.[26] Rearranging the expressions of FP's Lemma 2.1 and eq. (2), and defining S as the common support of $F_0(y)$ and $F_1(y)$,[27] FP show (in this paper's notation) that for arbitrary c:

$$\sup_{y\in S}\max\left\{F_0(y)-F_1(y-c),0\right\} \;\le\; \Pr\left(y_1 \ge y_0 + c\right) \;\le\; \inf_{y\in S}\min\left\{1+F_0(y)-F_1(y-c),1\right\}, \quad (18)$$

which FP note are sharp bounds on $\Pr\left(y_1 \ge y_0\right)$. Of particular interest here is c=0, giving

$$\sup_{y\in S}\max\left\{F_0(y)-F_1(y),0\right\} \;\le\; \Pr\left(y_1 \ge y_0\right) \;\le\; \inf_{y\in S}\min\left\{1+F_0(y)-F_1(y),1\right\}. \quad (19)$$

That is, the greatest lower bound on the inequality probabilities $\Pr\left(y_1 \ge y_0\right)$ identifiable from the marginals $F_j(y)$ is the maximum over all y in S of the difference (if positive) between $F_0(y)$ and $F_1(y)$. The corresponding smallest upper bound is $1+F_0(y)-F_1(y)$ if this quantity is less than one. In essence, these best bounds are determined by searching over $y\in S$ to determine where the FLB and FUB are greatest and smallest, respectively. These results are the foundation for what follows.

At this point some additional notation will be useful. For $j,k\in\left\{0,1,\ldots\right\}$, $j\ne k$:

$$\delta_{jk}(y) = F_j(y)-F_k(y) \quad (20)$$

$$D_{jk} = \max_{y\in S}\left\{\delta_{jk}(y),0\right\} \quad (21)$$

$$Y_{jk} = \arg\max_{y\in S}\left(\delta_{jk}(y)\right) \text{ if } D_{jk}>0, \;\text{ undefined if } D_{jk}=0 \quad (22)$$

[26] The 0 and 1 subscripts are reversed from those in FP's exposition. FP credit a line of earlier research upon which their work is based, with Makarov, 1982, and Williamson and Downs, 1990, figuring prominently. The work by Adams, 2013, Fan and Park, 2012, Fan et al., 2017, Firpo and Ridder, 2008, Lee, 2000, and Manski, 1997, is also noteworthy here. Applications of related ideas in health research are considered by Adams, 2013, Basu and Thariani, 2016, and Huang et al., 2016.

[27] $S=\left[\min\left\{L_0,L_1\right\},\max\left\{U_0,U_1\right\}\right]$. FP discuss technical considerations involved in defining the relevant supports, i.e. the domains of the sup and inf in (19).

That is, so long as δ_{jk} is positive Y_{jk} is the set of values of y at which δ_{jk} is greatest, while D_{jk} is that maximal value of δ_{jk}. The D_{jk} are known familiarly as Kolmogorov's distance or Kolmogorov's D statistics, which are the basis of some nonparametric tests for equality of two marginal distributions.[28] While in general Y_{jk} is set- or interval-valued, it is assumed for now that, if defined, it is a unique value to simplify notation and analysis; most important results go through whether or not uniqueness holds (FP discuss the role of uniqueness).

To visualize the result in (19), consider figure 3(b) in which the six subspaces $I_A, I_B, \ldots, IV$ have common origin $y_0 = y_1 = Y_{01}$. Here the difference $P(II) - P(IV)$ (red reference lines) is at least $P(II') - P(IV')$ (blue reference lines). Since $P(II) - P(IV) = F_0(Y_{01}) - F_1(Y_{01})$, $P(II) - P(IV)$ corresponds to the FP characterization of the best lower bound on $\Pr(y_1 \geq y_0)$. Thus as in (17):

$$\begin{aligned} \Pr(y_1 \geq y_0) &= P(I_B) + P(II) + P(III_A) \\ &\geq P(II) \\ &\geq P(II) - P(IV) \\ &= \left(P(II) + P(III_A) + P(III_B)\right) - \left(P(III_A) + P(III_B) + P(IV)\right) \\ &= F_0(Y_{01}) - F_1(Y_{01}) \\ &= D_{01}. \end{aligned} \tag{23}$$

Whether or not the Y_{jk} are defined it follows that $\Pr(y_1 \geq y_0) \geq D_{01}$, i.e. when $D_{01} = 0$ and Y_{10} is undefined, the most that can be said is that $\Pr(y_1 \geq y_0) \geq 0$, i.e. the FLB is not informative. Analogous arguments establish that, if it is informative, the best FUB is $1 - D_{10}$.

To summarize: if informative, the best possible bounds available from the $F_j(y)$ are

[28] See Darling, 1957, and Mann and Whitney, 1947. D_{jk} metrics arise in other contexts; for instance they correspond to stop-loss distance of degree one in the insurance literature (Denuit et al., 2002).

$$D_{jk} \leq \Pr\left(y_k \geq y_j\right) \leq 1 - D_{kj}. \tag{24}$$

For example, in the example depicted in figure 3 $Y_{01} = 4.6$, Y_{10} is undefined, $D_{01} = .38$, and $D_{10} = 0$. With respect to the nivolumab example, the "chance to live longer, " $\Pr\left(t_{niv} \geq t_{doc}\right)$, is at least .17 but not greater than .96, sampling considerations notwithstanding.

4. Examples, Extensions, and Related Results

Two Numerical Examples

Two numerical examples are pictured in figure 4. Panel (a) shows two $N\left(\mu_j, \sigma_j^2\right)$ marginal distributions, where $F_0\left(y\right)$is $N\left(0,4\right)$ and $F_1\left(y\right)$ is $N\left(.5,1\right)$. This yields $Y_{01} = -.73$, $Y_{10} = 2.07$, $D_{01} = .36 - .11 = .25$, and $D_{10} = .94 - .85 = .09$. Panel (b) shows results for exponential marginal distributions, where $F_0\left(y\right)$ is $\text{Exp}\left(5\right)$ and $F_1\left(y\right)$ is $\text{Exp}\left(1\right)$. These assumptions result in $Y_{01} = .40$, Y_{10} undefined, $D_{01} = .87 - .33 = .54$, and $D_{10} = 0$.[29]

Zero- and First-Order Stochastic Dominance

Consider first the case of zero-order stochastic dominance (ZSD; Castagnoli 1984).

[29] If the $F_j\left(y\right)$ are $N\left(\mu_j, \sigma_j^2\right)$ then Y_{01} and Y_{10} are given by the quadratic formula with $a = \sigma_1^2 - \sigma_0^2$, $b = -2\left(\sigma_1^2\mu_0 - \sigma_0^2\mu_1\right)$, and $c = \sigma_1^2\mu_0^2 - \sigma_0^2\mu_1^2 - 2\sigma_0^2\sigma_1^2 \ln\left(\sqrt{\sigma_1^2 / \sigma_0^2}\right)$ if $\sigma_0^2 \neq \sigma_1^2$, with roots $Y_{01} < Y_{10}$ if $\sigma_0^2 > \sigma_1^2$ and $Y_{01} > Y_{10}$ if $\sigma_0^2 < \sigma_1^2$. If $\sigma_0^2 = \sigma_1^2$, then one of Y_{01} or Y_{10} is given by $.5\left(\mu_0 + \mu_1\right)$ (Y_{01} if $\mu_0 < \mu_1$; Y_{10} if $\mu_0 > \mu_1$). See figure 5, panels (a) and (b). If the $F_j\left(y\right)$ are exponential with $F_j\left(y\right) = 1 - \exp\left(-\theta_j y\right)$, then $D_{01} = \exp\left(-\theta_1 Y_{01}\right) - \exp\left(-\theta_0 Y_{01}\right)$, $D_{10} = 0$, $Y_{01} = \ln\left(\theta_1 / \theta_0\right) / \left(\theta_1 - \theta_0\right)$, and Y_{10} is undefined if $\theta_0 > \theta_1$; the subscripts are reversed if $\theta_1 > \theta_0$. While parametric distributions may be helpful for illustrative and modeling purposes, applications often consider nonparametric empirical distributions. Estimating the D_{jk} nonparametrically is discussed in section 7.

$F_1(y)$ zero-order dominates $F_0(y)$, denoted $F_1 \succ_0 F_0$, if $U_0 < L_1$, i.e. if the entire probability mass of $F_1(y)$ sits above that of $F_0(y)$ on the real line (see figure 6). A noteworthy feature of ZSD is that $\Pr(y_1 \geq y_0) = 1$, i.e. regardless of a population member's outcome in $F_0(y)$, that outcome will be less than their outcome in $F_1(y)$.[30] Note that $D_{01} = 1$ for any Y_{01} in $[U_0, L_1]$ so $\Pr(y_1 \geq y_0)$ is point-identified as $\Pr(y_1 \geq y_0) = 1$, i.e. the FLB on $\Pr(y_1 \geq y_0)$ at Y_{01} is maximally informative. With first-order dominance $F_1 \succ_1 F_0$, Y_{01} is defined, Y_{10} is undefined, $D_{01} > 0$, and $D_{10} = 0$.

Informativeness of the D_{jk} Bounds

To see how closely the D_{jk}-based bounds correspond to the true inequality probabilities suppose $y_0, y_1 \sim BVN(\mu_0, \mu_1; 1, 1, \rho)$. The entries in table 2 are D_{01} and the true $\Pr(y_1 \geq y_0)$ for selected $\mu_1 - \mu_0$ and ρ (the probabilities depend only on the differences $\mu_1 - \mu_0$). When $\mu_1 - \mu_0$ is large and ρ is negative, the D_{01}-based bounds are relatively close to $\Pr(y_1 \geq y_0)$, but with positive ρ these bounds are quite conservative relative to the true $\Pr(y_1 \geq y_0)$. Such results are intuitive: for given marginals, negative correlation tends to situate more joint probability mass in quadrants II and IV than does positive correlation (e.g., contrast figures 7(a) and 7(b)).

$Pr(y_1 \geq y_0)$ under Independence

Gastwirth, 1975, considers situations where y_0 and y_1 are statistically independent. Here, $\Pr(y_1 \geq y_0)$ is identified given the marginals: $\Pr(y_1 \geq y_0) = \int_{-\infty}^{\infty} f_0(y_0) \int_{y_0}^{\infty} f_1(y_1) dy_1 dy_0$.

[30] If y is net benefit, then a policy shifting $F_0(y)$ to $F_1(y)$ yields a Pareto improvement.

Consider the exponential case in figure 4(b). The FLB on $\Pr\left(y_1 \geq y_0\right)$ for any dependence structure is $D_{01} = .54$, whereas $\Pr\left(y_1 \geq y_0\right)$ under independence is $\theta_0 / \left(\theta_0 + \theta_1\right) = .83$.

Relationships to Permutation Distributions

Let $\mathbf{y}_{0,N} = \left[y_{0,n}\right]$ and $\mathbf{y}_{1,N} = \left[y_{1,n}\right]$ denote N-vectors describing outcomes for a sample or a finite population of size 2N. Let $P\left(\mathbf{y}_{0,N}\right)$ be the $N \times N!$ matrix containing the N! permutations of the elements of $\mathbf{y}_{0,N}$; let $\mathbf{C} = \left[\mathbf{y}_{1,N} - P\left(\mathbf{y}_{0,N}\right)_c\right]$ be the $N \times N!$ matrix whose c-th column is the difference between $\mathbf{y}_{1,N}$ and the c-th column of $P\left(\mathbf{y}_{0,N}\right)$; and let $\mathbf{d} = \left[\frac{1}{N}\sum_{n=1}^{N} 1\left(\mathbf{C}_{n,c} > 0\right)\right]$ be the $1 \times N!$ vector describing the fraction of elements in each of $\mathbf{C}$'s columns for which $y_{1,n} > y_{0,n(c)}$. Then the smallest and largest elements of $\mathbf{d}$ are D_{01} and $1 - D_{10}$, respectively. These relationships are discussed by Heckman et al., 1997, who suggest that when N is large summary statistics like deciles of the sample marginal distributions might be used to approximate the permutation relationships.

Alternative Characterizations of Δ_{01} and Transformations

Beyond $\Delta_{01} = y_0 - y_1$, other contrasts may be of interest, for instance $t\left(y_0\right) - t\left(y_1\right)$ where $t\left(\ldots\right)$ is a monotone-increasing transformation. The previous results apply here: $\Pr\left(t\left(y_1\right) \geq t\left(y_0\right)\right) = \Pr\left(y_1 \geq y_0\right)$, $Y_{jk,t} = t\left(Y_{jk}\right)$, $D_{01,t} = F_{0,t}\left(Y_{01,t}\right) - F_{1,t}\left(Y_{01,t}\right) = F_0\left(Y_{01}\right) - F_1\left(Y_{01}\right)$, etc., using obvious notation. For $y_j > 0$ contrasts might involve ratios, $\Pr\left(\left(y_0/y_1\right) \leq c\right)$ or proportional differences $\Pr\left(\left(\left(y_0 - y_1\right)/y_0\right) \leq c\right)$.[31] Non-inferiority assessments may concern

[31] See Imbens and Wooldridge, 2009, and Lee and Kobayshi, 2001, and Lee, 2005, for conceptual considerations, and Langley et al., 2014 for a related application.

probabilities like $\Pr\left(y_0 - y_1 \leq c\right)$ for nonzero c.[32] So long as c and/or $t\left(\ldots\right)$ are known all these cases can be subsumed by specifying $\Delta = \tau_0\left(y_0\right) - \tau_1\left(y_1\right)$ and considering $\Pr\left(\tau_1\left(y_1\right) \geq \tau_0\left(y_0\right)\right)$. For example, in the proportional-difference example $\tau_0\left(y_0\right) = \left(1-c\right)y_0$ and $\tau_1\left(y_1\right) = y_1$. The previous results go through directly if the respective $F_j\left(y\right)$ reference the distributions of the transformed measures obtained, e.g., by standard change-of-variable methods.

Discrete and Ordinal Outcomes

The main results on identifying bounds on inequality probabilities apply also when population outcome measures are integer-valued (e.g. count-data; see Cameron et al., 2004), discrete-ordinal, or categorical measures (e.g. health-status scores or indexes, Likert scales).[33] One important consideration in such cases is whether the parameter of interest is $\Pr\left(y_1 \geq y_0\right)$ or $\Pr\left(y_1 > y_0\right)$ since in the population a nonzero probability of ties, i.e. of the event $y_0 = y_1$, is relevant.[34] The approach described in section 3 that identifies the Y_{jk} and D_{jk} is applicable here, but the quantity whose bounds are identified as such is $\Pr\left(y_1 > y_0\right)$, not $\Pr\left(y_1 \geq y_0\right)$.[35]

The 2009 study by Volpp et al. on the effects of financial incentives on smoking cessation and related outcomes offers an instructive example. One outcome of interest in that study is a five-point Likert scale measure of subjects' self-assessed health; the distributions of their sample data are pictured in figure 8(a). Treatment effects using this measure are assessed by Volpp et al. by examining differences between treatment and control separately at

[32] See U.S. Food and Drug Administration, 2016.

[33] Huang et al., 2016, consider a discrete functional disability measure as their main outcome. Also see Allison and Foster, 2004, for some related perspectives on discrete ordinal outcomes.

[34] In empirical applications consideration of ties is relevant not only when the data are naturally discrete but also when data that are in principle continuously distributed are measured coarsely.

[35] For the Y_{jk} to be (potentially) unique when outcomes are discrete or categorical, the domain of the argmax in (22) should be redefined as the set $\left\{y \middle| \Pr\left(y_0 = y \vee y_1 = y\right) > 0\right\}$.

the five Likert scale points (see their figure 2). In these data Y_{01} occurs at the "Very Good" category with a resulting $D_{01} = .03$. This result can be imagined by reference to figure 8(b) which depicts the sample space for these data; $D_{01} = .03$ corresponds to the probability mass of the red dots minus that of the black dots.

Spreading or Rectangularizing Distributions

Figure 9(a) illustrates a case where $F_0(y)$ is $N(0,1)$ and $F_1(y)$ is $N(0,4)$ giving $Y_{01} = 1.36$, $Y_{10} = -1.36$, and $D_{01} = D_{10} = .16$. Assume now that some intervention replaces $F_1(y)$ with $F_2(y)$, which is $N(0,16)$, resulting in $Y_{02} = 1.72$, $Y_{20} = -1.72$, and $D_{02} = D_{20} = .29$. Spreading $F_1(y)$ relative to $F_0(y)$ in the sense of increasing $|F_0(y) - F_1(y)|$ for all y (e.g. in increase in σ_1^2 when $\mu_0 = \mu_1$) increases the D_{jk} and thus gives tighter bounds on $\Pr(y_1 \geq y_0)$. Conversely, rectangularizing one distribution results in the limit in a degenerate distribution for which $Y_{01} = Y_{10}$ so that $D_{10} = 1 - D_{01}$ and $\Pr(y_1 \geq y_0)$ is point-identified: $1 - D_{10} = D_{01} \geq \Pr(y_1 \geq y_0) \geq D_{01}$. For example, figure 9(b) shows a case where $F_1(y)$ is degenerate $N(1,0)$ and $F_0(y)$ is $N(0,4)$. This gives $Y_{01} = Y_{10} = 1$, $D_{01} = .69$, and $D_{10} = .31$ so that $\Pr(y_1 \geq y_0)$ is point-identified at .69.

5. Inequality Probabilities and Cost-Effectiveness Analysis

Inequality probabilities may usefully inform some questions in cost-effectiveness analysis (CEA). Much applied CEA involves consideration of mean incremental costs and outcomes, and focuses on uncertainties arising from sampling variation. This is often true whether the evaluation strategy is based on incremental cost-effectiveness ratios (ICERs), cost-effectiveness acceptability curves (CEACs; Fenwick et al., 2004, and Willan, 2001), or some other approach. The ideas discussed in this paper permit alternative perspectives on stochastic CEA wherein the main focus is on underlying population heterogeneity of costs and outcomes instead of sampling variation.[36]

Suppose the y_j are defined as net health benefit ("h"; Stinnett and Mullahy, 1998),

$$y_j = h_j = e_j - \left(c_j / \lambda\right), \tag{25}$$

where e_j and c_j denote the health outcomes and costs arising from intervention j (T_j) in some population, and λ represents a population-constant standard like social marginal willingness to pay for e (e.g. dollars per QALY). For instance, in a social choice setting where population members vote self-interestedly for one intervention to be applied uniformly, $\Pr\left(h_1 \geq h_0\right)$ signals the likelihood that T_1 would be the intervention adopted. $\Pr\left(h_1 \geq h_0\right)$ is also one characterization of "the probability of cost-effectiveness" (Willan, 2001).

Define the subject-level outcomes $\mathbf{q} = \left[e_0, e_1, c_0, c_1\right]$, and for a given λ let

$$\begin{aligned} \Pr_\lambda\left(h_1 \geq h_0\right) &= \Pr\left(e_1 - \left(c_1/\lambda\right) \ \geq \ e_0 - \left(c_0/\lambda\right)\right) \\ &= \Pr\left(\left(e_1 - e_0\right) \geq \left(c_1 - c_0\right)/\lambda\right) \\ &= \Pr\left(r \leq \lambda\right), \end{aligned} \tag{26}$$

where $r = \left(c_1 - c_0\right) / \left(e_1 - e_0\right)$. For given $\lambda > 0$ $\Pr_\lambda\left(h_1 \geq h_0\right)$ is increasing in e_1 and c_0 and decreasing in e_0 and c_1, while the relationship between $\Pr_\lambda\left(h_1 \geq h_0\right)$ and λ may be nonmonotonic. Note too that the relationship between $\Pr_\lambda\left(h_1 \geq h_0\right)$ and λ is essentially that of an incremental CEAC: as λ varies it tells the probability that intervention 1 becomes more or less acceptable relative to intervention 0. Defined in terms of underlying random variables, however, this CEAC differs from that of more-familiar[37] CEACs that have been considered.

In data-rich contexts wherein all elements of **q** are jointly observable—i.e., when the full joint probability structure of $F_{\mathbf{q}}\left(\ldots\right)$ is available—$\Pr_\lambda\left(h_1 \geq h_0\right)$ can be point-identified. Yet

[36] This is sometimes cast as 2nd- vs. 1st-order uncertainty; see Vanness and Mullahy, 2012.

[37] That is, criteria using $\left(\mu_{c_1} - \mu_{c_0}\right) / \left(\mu_{e_1} - \mu_{e_0}\right)$ and analog estimates $\left(\hat{\mu}_{c_1} - \hat{\mu}_{c_0}\right) / \left(\hat{\mu}_{e_1} - \hat{\mu}_{e_0}\right)$.

in many settings only joint marginal distributions $F_j(e,c)$ and, therefore, marginal distributions $F_{j,\lambda}(h)$ are available. This would be the case, e.g., in a two-arm trial where both outcome and cost data from the each T_j are available at the subject level (van Hout et al., 1994), or when (e_0,c_0) and (e_1,c_1) are observed in separate datasets.[38] When only the joint marginals $F_j(e,c)$ are available $\Pr_\lambda(h_1 \geq h_0)$ cannot generally be point-identified unless (e_0,c_0) is statistically independent of (e_1,c_1).

Yet in light of the results in section 3, it may be possible to obtain informative bounds on $\Pr_\lambda(h_1 \geq h_0)$ when only the marginals $F_{j,\lambda}(h)$ are available. As an illustrative example assume that $\mathbf{q} \sim \mathrm{MVN}(\boldsymbol{\mu}_\mathbf{q},\mathbf{V}_\mathbf{q})$. Let $\boldsymbol{\mu}_\mathbf{q} = \left[\mu_{e_0},\ \mu_{e_0}+5,\ \mu_{c_0},\ \mu_{c_0}+10\right]$, and let $\mathbf{V}_\mathbf{q}$ be defined to have all diagonal elements equal 1 and all off-diagonal elements equal .5. Then for a given λ $h_1 - h_0 \sim N\left(5-(10/\lambda), 1+(1/\lambda)\right)$. The resulting true probabilities $\Pr_\lambda(h_1 \geq h_0) = \Pr_\lambda(h_1 - h_0 \geq 0)$ and corresponding FLB based on the marginals $F_{0,\lambda}(h)$ and $F_{1,\lambda}(h)$ are plotted in figure 10 for values of $\lambda \in (0,10]$. In this case the FLB is seen to be informative, at least for values of $\lambda > 2$.

6. Inequality Probabilities with More than Two Outcomes

Three or More Competing Univariate Outcomes

While most attention in the evaluation literature is on contrasts between two outcomes, in some cases more than two outcomes are of interest. For instance, Nissen et al., 2016, compare in a three-arm trial the cardiovascular safety profiles of celecoxib, ibuprofen, and naproxen for patients with osteoarthritis or rheumatoid arthritis, while marketing[39] for

[38] Indeed, much as in the mainstream treatment-effect literature one reason that means-based CEA (ICERs, CEACs, etc.) may be popular is that mean differences in outcomes and costs correspond to differences in their marginal means under suitable sampling schemes.

[39] https://www.victoza.com/consider-using-victoza-/compared-with-januvia----byetta-.html, accessed May 10, 2017.

Victoza (liraglutide; Novo Nordisk), a treatment for type 2 diabetes, compares its therapeutic properties with those of Januvia (sitagliptin; Merck) and Byetta (exenatide; AstraZeneca).

Expanding the discussion of section 3, one consideration might be the probability that one treatment (say y_1) results in a better outcome than either of the others (say y_0 and y_2), i.e. $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$.[40] With three outcomes the earlier results can be extended to obtain potentially informative bounds on $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$. Specifically, Fréchet-Boole inequalities can themselves be used recursively to bound the bounds on $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$, the latter being unknowable given only information on the marginals $F_j\left(y\right)$.

To this end, define D_{21} using (21). Let $\Pr\left(y_1 \geq y_0\right)$ and $\Pr\left(y_1 \geq y_2\right)$ correspond, respectively, to $\Pr\left(z_a \in Z_a\right)$ and $\Pr\left(z_b \in Z_b\right)$ in (8), and note that $\Pr\left(y_1 \geq y_k\right) \geq D_{k1}$ for k=0,2. Using D_{01} and D_{21}, it follows that a lower bound on the lower bound on $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$—and, therefore, a lower bound on $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$ itself—is $\max\left\{D_{01}+D_{21}-1,0\right\}$, i.e.

$$\max\left\{D_{01}+D_{21}-1,0\right\} \leq \max\left\{\Pr\left(y_1 \geq y_0\right)+\Pr\left(y_1 \geq y_2\right)-1,0\right\} \leq \Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right), \quad (27)$$

which is informative if $D_{01}+D_{21}>1$. The corresponding approach to obtaining an upper bound on the upper bound on $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$ uses

$$\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right) \leq \min\left\{\Pr\left(y_1 \geq y_0\right),\Pr\left(y_1 \geq y_2\right)\right\} \leq \min\left\{1-D_{10},1-D_{12}\right\}, \quad (28)$$

which is informative if either or both of the D_{1k} exceed zero.

For example, suppose $\left[y_0,y_1,y_2\right] \sim \text{TVN}\left(\boldsymbol{\mu},\mathbf{V}\right)$ with $\boldsymbol{\mu}=\left[\mu_0,\mu_1,\mu_2\right]=\left[1,3,0\right]$ and covariance $\mathbf{V}=\begin{bmatrix} 1 & \rho & 2\rho \\ \rho & 1 & 2\rho \\ 2\rho & 2\rho & 4 \end{bmatrix}$. The $F_j\left(y\right)$ and corresponding $Y_{01}=2$ and $Y_{21}=1.58$ are

[40] Such questions might be of interest when all the outcomes are observed in the same sample or when—as discussed below in section 7—observations from different datasets are used.

depicted in figure 11(b), showing D_{01}=.68 and D_{21}=.71. The lower bound on the population FLB on $\Pr\left(y_1 \geq y_0 \wedge y_1 \geq y_2\right)$ obtained from D_{01} and D_{21} is thus .39=.68+.71-1. Table 3 compares this with the true probabilities and the population FLBs for $\rho \in \left\{-.25, 0, .25\right\}$. The lower bound on the FLB based on D_{01} and D_{21} is conservative, albeit still informative. The corresponding upper bound in (28) is minimally informative, .9996, resulting from D_{10}=0 and D_{12}=.0004.

Competing Multivariate Outcomes

In some evaluations the outcomes of interest are multivariate. A prominent example is that of co-primary ("and", "all") outcomes in clinical studies.[41] In regulatory settings co-primary outcomes may involve "use of two or more endpoints for which demonstration of an effect on each is needed to support regulatory approval" (U.S. FDA, 2017). One characterization of "effect" might be that all outcomes under one treatment are not smaller than those under the comparator, i.e. $\mathbf{y}_1 \geq \mathbf{y}_0$ for P-vectors $\mathbf{y}_j$. Analogous considerations arise in healthcare quality measurement contexts where all-or-nothing indicators of quality may be of interest (Nolan and Berwick, 2006).

To formalize these ideas, suppose the P-dimensional outcomes are $\mathbf{y}_j = \left[y_{j,1}, \ldots, y_{j,P}\right]$, j=0,1. Of concern may be the probability $\Pr\left(\mathbf{y}_1 \geq \mathbf{y}_0\right)$ where ≥ is element-by-element. For instance, with M=P=2, the parameter of interest is $\Pr\left(y_{1,1} \geq y_{0,1} \wedge y_{1,2} \geq y_{0,2}\right)$. Two approaches might be considered.

For the first, assume M=P=2 and that only the four univariate marginals $F_{j,m}\left(y\right)$, j=0,1, p=1,2, are available. Using the recursive-bounding idea in (27) and (28), and letting $D_{jk,p}$ denote quantities akin to (21), the lower and upper $D_{jk,p}-$based bounds on

[41] Atkinson, 2003, discusses the ideas of union ("or") and intersection ("and") outcomes.

$\Pr\left(y_{1,1} \geq y_{0,1} \wedge y_{1,2} \geq y_{0,2}\right)$ are

$$\max\left\{D_{01,1} + D_{01,2} - 1, 0\right\} \leq \Pr\left(y_{1,1} \geq y_{0,1} \wedge y_{1,2} \geq y_{0,2}\right) \leq \min\left\{1 - D_{10,1}, 1 - D_{10,2}\right\}. \quad (29)$$

The second approach assumes that M=2 P-dimensional joint marginals $F_j\left(\mathbf{y}\right)$, j=0,1, are available.[42] Using results from Rüschendorf, 2004 (see also Kotz and Seeger, 1993), the joint probability of the events $\mathbf{y}_0 \leq \mathbf{y}'$ and $\mathbf{y}_1 > \mathbf{y}'$ for arbitrary $\mathbf{y}'$ is bounded as follows:

$$\max\left\{F_0\left(\mathbf{y}'\right) - F_1\left(\mathbf{y}'\right), 0\right\} \leq \Pr\left(\mathbf{y}_0 \leq \mathbf{y}' \wedge \mathbf{y}_1 > \mathbf{y}'\right) \leq \min\left\{1 + F_0\left(\mathbf{y}'\right) - F_1\left(\mathbf{y}'\right), 1\right\} \quad (30)$$

Obtaining the best bounds on $\Pr\left(\mathbf{y}_1 \geq \mathbf{y}_0\right)$ in this case follows in a manner analogous to (24) except that determining the particular $\mathbf{Y}_{jk}$ (the vector analog of Y_{jk} in (22)) at which $F_0\left(\mathbf{y}\right)$ and $F_1\left(\mathbf{y}\right)$ are evaluated to identify the best bounds may entail additional computational considerations.[43]

For illustration, consider M=2 co-primary outcomes $\mathbf{y}_j$ where the $y_{j,p}$ are binary, the joint marginals are known, and $\mathbf{y}' = \mathbf{0}$ in (30).[44] Here the best bounds on $\Pr\left(\mathbf{y}_1 > \mathbf{y}_0\right)$ are (refer to (7)):

$$\begin{aligned} \max\left\{\Pr\left(\mathbf{y}_1 = \mathbf{1}\right) - \Pr\left(\mathbf{y}_0 = \mathbf{1}\right), 0\right\} &\leq \Pr\left(\mathbf{y}_0 = \mathbf{0} \wedge \mathbf{y}_1 = \mathbf{1}\right) = \Pr\left(\mathbf{y}_1 > \mathbf{y}_0\right) \\ &\leq \min\left\{1 - \Pr\left(\mathbf{y}_0 = \mathbf{1}\right), \Pr\left(\mathbf{y}_1 = \mathbf{1}\right)\right\} \end{aligned} \quad (31)$$

This idea also covers the weak inequality case, $\Pr\left(\mathbf{y}_1 \geq \mathbf{y}_0\right)$, albeit with messier probability

[42] The assumption that the joint marginals are identifiable would often be a reasonable one.

[43] In a closely related context Andrews, 1997, discusses how a grid or hypercube search over $\mathbf{y}'$ can be confined to the observed sample values of **y**—as these define the steps in the empirical joint distribution—thus simplifying estimation. Note that the M elements of $\mathbf{Y}_{jk}$ will generally not be the M scalar values that would obtain from applying (22) with reference to the M univariate marginals.

[44] For example, Langley et al., 2014, consider two binary co-primary endpoints in a study of secukinumab versus etanercept in the treatment of plaque psoriasis.

algebra. Moreover, using (9) and (11) the recursive-bounding idea in (27) and (28) can be used to bound composite ("or", "any") outcome probabilities (U.S. FDA, 2017), e.g. $\Pr\left(y_{1,1} > y_{0,1} \vee y_{1,2} > y_{0,2}\right)$.

7. Sampling, Estimation, and Inference

This section considers empirical implementation of the univariate-outcome results described in section 3. In what follows the empirical marginal distributions of the observed outcomes $y_{j,n}$, given sample sizes N_j, are defined as $F_{j,N_j}\left(y\right) = \frac{1}{N_j}\sum_{n=1}^{N_j} 1\left(y_{j,n} \leq y\right)$.

Sampling

The sampling assumptions are standard ones. FP state: "observations on the outcome of participants in the treatment group identify the distribution of the potential outcome with treatment, and observations on the outcome of participants in the control group identify the distribution of the potential outcome without treatment."[45] In essence, a random sample containing information on the true $\left(y_0, y_1\right)$, or more generally $\left(y_0, y_1, \mathbf{x}\right)$, is drawn from the population. Then for each subject the information on either y_0 or y_1 is deleted at random, resulting in samples of size N_j of observations on y_j. More generally, the FP results apply with unconfounded conditioning on $\mathbf{x}$—i.e. selection on observables only—if covariates are relevant.[46] These assumptions are standard and point to what matters being consistent estimates of the $F_j\left(y\right)$ in the sense of convergence in distribution: $F_{j,N_j}\left(y\right) \rightarrow F_j\left(y\right)$ as $N_j \rightarrow \infty$

[45] All the standard reasons to scrutinize the validity of such assumptions in light of the processes that may actually generate the observed data are applicable here. See Adams, 2013, Chan and Hamilton, 2006, Fan et al., 2017, Imbens and Wooldridge, 2009, and Manski, 1996.

[46] See FP, pages 932 and 944-945, and Imbens and Wooldridge, 2009, section 2.2. In an unconfounded regression context with $y = \alpha t + m\left(\mathbf{x}\right) + u$, $E\left[u \middle| t, \mathbf{x}\right] = 0$, and $t \in \left\{0,1\right\}$, an analyst might imagine empirical bounds analysis using as "outcomes" the estimated adjusted or semi-residuals $\hat{r} = y - \widehat{m}\left(\mathbf{x}\right)$ from the two subsamples defined by the binary treatment indicator. Consideration of the properties of such an approach is left for future exploration.

for all y in S_j (Hansen, 2017, section 6.7). Technical considerations aside, sampling schemes that identify criteria like $V(F_1(y)) - V(F_0(y))$ (e.g. differences in means or medians) would generally suffice for purposes at hand.

Additional Sampling Considerations

Censoring of Empirical Outcome Distributions

In applications left (e.g. Tobit-type) or right (e.g. survival times) censoring may be relevant. Censoring of either or both of the empirical marginal distributions may or may not affect the magnitudes of the FLB or FUB depending on where censoring occurs relative to the uncensored data's Y_{jk}. Informative bounds on $\Pr(y_1 \geq y_0)$ may still be defined from censored samples regardless of the degree of censoring so long as some outcome data are uncensored. Consider the study by Lee et al., 2016, comparing naltrexone and usual treatment for opioid relapse. The study's primary outcome is relapse-free survival time. The outcome data (derived from approximating the data in Lee et al.'s figure 2) are depicted in figure 12. While these data are right-censored at 24 weeks, it can be determined that D_{01} is at least .29 based on a provisional Y_{01} at 15 weeks.[47]

Marginal Distributions Observed in or Estimated from Different Datasets or Samples

Nothing about the results discussed above demands that the data on y_0 and y_1 be obtained from the same sample or dataset. All that is required is that the respective empirical marginal distributions converge to the corresponding population marginals of $F(y_0, y_1)$, as above. If the marginal distributions of the two outcomes observed in different datasets (e.g.

[47] When either or both of the $F_{j,N_j}(y)$ are censored, point identification of $E[y_1] - E[y_0]$ is generally not possible. Depending on where censoring occurs this is also true for differences between marginal quantiles although informative bounds may be available if one of the marginial quantiles is observed. For instance, while it is not possible to identify $\text{med}(F_1(y)) - \text{med}(F_0(y))$ in the Lee et al. example, it is evident from Lee et al.'s figure 2 that this difference is at least 13 weeks.

repeated cross-sections, synthetic panels, separate trials, etc.) are truly representative of the same population—characterized by time, place, and all other observable and unobservable characteristics—then the previous analysis is applicable without modification.[48]

Estimation

Estimation of D_{01} and D_{10} requires an algorithm that computes the difference between empirical distribution functions across their common support. In Stata, this is straightforward using the `ksmirnov` procedure.[49] With the data on $y_{0,n}$ and $y_{1,n}$ stacked into a single variable (say $y=\left[y_n\right]$) having N_0+N_1 observations, and a second variable (say $g=\left[g_n\right]$) defined as the binary indicator of group membership, e.g. $g_n=1\left(n>N_0\right)$, then the Stata command is simply:

```
ksmirnov y, by(g)
```

`ksmirnov` returns the scalar stored results `r(D_1)` and `r(D_2)` whose absolute values are, respectively, the estimates of D_{01} and D_{10}. To illustrate, 500 observations are drawn from the $N\left(0,2\right)$ and $N\left(.5,1\right)$ distributions depicted in figure 4. The `ksmirnov` estimates are shown in exhibit 1. From `r(D_1)` and `r(D_2)`, the estimates of D_{01} and D_{10} are .224 and .12, corresponding to their respective population counterparts .25 and .09 shown in figure 4.

Inference

The emphases to this point in the paper have been identification of probability bounds based on D_{jk} and estimation of such bounds. Considerations of inference might involve at least two questions (see Imbens and Manski, 2004, and Tamer, 2010). First, what purpose is

[48] The assumption that the two samples are drawn from the same population is a strong one. For clinical trials inclusion criteria, study sites, etc., would all be relevant considerations; for population surveys or administrative data, sampling frames, exclusion criteria, etc., would be relevant.

[49] R has a procedure, `ks.test`, that appears to provide output similar to that of Stata's `ksmirnov`.

served by conducting inference about bounds? Second, which parameters are of interest for conducting inference? Assuming useful purposes exist then at least two types of inference may be relevant: inference about the D_{jk}-based bounds per se, and inference specifically about $\Pr\left(y_1 \geq y_0\right)$.

For the first type, FP provide large-sample results. Since the $F_{j,N_j}\left(y\right)$ are averages of independent Bernoulli variates (Hansen, 2017, section 13.2), FP's proposition 3.1 gives

$$\sqrt{N}\left(D_{jk,N} - D_{jk}\right) \rightarrow N\left(0,\sigma^2_{jk}\right) \tag{32}$$

where

$$\sigma^2_{jk} = F_0\left(Y_{jk}\right)\left(1-F_0\left(Y_{jk}\right)\right) + F_1\left(Y_{jk}\right)\left(1-F_1\left(Y_{jk}\right)\right), \tag{33}$$

assuming equal sample sizes in the two groups (this is easily relaxed) and that various regularity conditions[50] are met. Confidence intervals built on these large-sample results must also respect the 0-1 probability bounds. For the data in figure 1, using (33) to compute 95% (±2 s.e.) CIs around the estimated D_{01} and $1-D_{10}$ bounds whose point estimates are .17 and .96, respectively, results in respective CIs of [.10, .24] and [.90, 1]. FP also discuss bootstrap-based inference.

For the second type, inference may be undertaken to understand sampling variation in the estimates of $\Pr\left(y_1 \geq y_0\right)$. `ksmirnov` gives p-values for testing directional hypotheses that one of y_0 or y_1 is stochastically smaller than the other (see exhibit 1). These p-values are computed as $p_{jk,N} = \exp\left(-\frac{2N_0N_1}{N_0+N_1}D^2_{jk,N}\right)$ for the null that y_j is not stochastically smaller than

[50] FP's results use the assumption (which they suggest can be relaxed) that the Y_{jk} are unique. FP also discuss bootstrap inference; see also Abrevaya, 2000, and Abadie, 2002. A sampling exercise suggests that even a naive bootstrap—with computation of each replicate's estimate of D_{jk} around the original sample's value of Y_{jk}—reproduces closely both the population (known $F_j\left(y\right)$) and analog ($F_{j,N_j}\left(y\right)$"plugged in") versions of the (33). These results are available on request.

y_k; the $p_{jk,N}$ depend only on the $D_{jk,N}$, not on the particular values of the $F_{j,N_j}(y)$.[51]

8. Summary

This paper has proposed the utility in health economics evaluations of some results on inequality probabilities from the treatment-effect literature that have gone largely unnoticed or unused in health applications. In comparing outcomes y_j across a population, which metric(s) are used for comparison is at the decisionmaker's discretion. While standard contrasts like ATEs are informative for some questions, other perspectives may be more relevant in some decisionmaking contexts. Questions regarding inequality probabilities are natural to consider in a range of decisionmaking settings. While point identification of such parameters is challenging, the paper has shown how inequality probabilities can be informatively bounded using information on the marginal outcome distributions. Of course, estimating the relevant marginal outcome distributions from the data at hand may itself be challenging for all the standard reasons.

Whether decisionmakers are comfortable relying on bounds is a consideration whose relevance and importance have been emphasized by Manski. Entrenched approaches to evaluation in regulatory (e.g. FDA) and other contexts may be challenging to budge. Yet superior decisions will be made if evaluations that inform them are anchored to criteria that reflect what actually matters to decisionmakers[52] rather than to criteria that happen to be biostatistically convenient or time-honored. True value-based policymaking and healthcare delivery demand no less.

[51] See Darling, 1957. The two directional tests ksmirnov reports are against null hypotheses that y_0 is not stochastically smaller than y_1 and that y_1 is not stochastically smaller than y_0.

[52] See Lynn et al., 2015, for a compelling discussion.

Acknowledgments

For helpful discussions and comments, and with the standard disclaimer, I would like to thank Chris Adams, Bruce Hansen, M.-J. Lee, David Madden, Chuck Manski, Joao Santos Silva, Justin Sydnor, Dave Vanness, and participants in presentations at the Barcelona Conference on the Economics of Addiction and Health Inequality, the York Centre for Health Economics, ISEG-Lisbon, Resources for the Future, University College Dublin, the University of Chicago, the University of Coimbra, and the University of Wisconsin-Madison. Parts of this work were supported by the RWJF Health & Society Scholars Program, RWJF Evidence for Action Grant 73336, and NICHD grant P2CHD047873 to the Center for Demography and Ecology, all at UW-Madison, and by the UCD Geary Institute during a sabbatical visit.

References

Aakvik, A. et al. 2005. "Estimating Treatment Effects for Discrete Outcomes when Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs." *Journal of Econometrics* 125: 15-51.

Abadie, A. 2002. "Bootstrap Tests for Distributional Treatment Effects in Instrumental Variable Models." *JASA* 97: 284-292.

Abbring, J.H. and J.J. Heckman. 2007. "Econometric Evaluation of Social Programs, Part III: Distributional Treatment Effects, Dynamic Treatment Effects, Dynamic Discrete Choice, and General Equilibrium Policy Evaluation." Chapter 72 in J.J. Heckman and E.E. Leamer, Eds., *Handbook of Econometrics*, Vol. 6B. Amsterdam: Elsevier.

Abrevaya, J. 2000. "Testing for a Treatment Effect in a Heterogeneous Population: A Modified Sign-Test Statistic and a Leapfrog Statistic." *Journal of Applied Statistics* 27: 679-687.

Adams, C.P. 2013. "Median Survival, Hazard Ratios, and Bounding Drug Effectiveness." Working Paper, U.S. Federal Trade Commission.

Allison, R.A. and J.E. Foster. 2004. "Measuring Health Inequality Using Qualitative Data." *Journal of Health Economics* 23: 505-524.

Andrews, D.W.K. 1997. "A Conditional Kolmogorov Test." *Econometrica* 65: 1097-1128.

Angrist, J.D. 2004. "Treatment Effect Heterogeneity in Theory and Practice." *Economic Journal* 114: C52-C83.

Athey, S. and G.W. Imbens. 2006. "Identification and Inference in Nonlinear Difference-in-Differences Models." *Econometrica* 74: 431-497.

Atkinson, A.B. 2003. "Multidimensional Deprivation: Contrasting Social Welfare and Counting Approaches." *Journal of Economic Inequality* 1: 51-65.

Basu, A. et al. 2007. "Use of Instrumental Variables in the Presence of Heterogeneity and Self-Selection: An Application to Treatments of Breast Cancer Patients." *Health Economics* 16: 1133-1157.

Basu, A. and R. Thariani. 2016. "Jointness Box (JB)-Area: A Novel Metric to Contemplate Potential Value of Individualized Care from Traditional Trials Data." Working Paper, Univ. of Washington.

Bitler, M.P. et al. 2006. "What Mean Impacts Miss: Distributional Effects of Welfare Reform Experiments." *American Economic Review* 96: 988-1012.

Borah, B.J. et al. 2011. "Assessing the Impact of High Deductible Health Plans on Health-Care Utilization and Cost: A Changes-in-Changes Approach." *Health Economics* 20: 1025-1042.

Borghaei, H. et al. 2015. "Nivolumab versus Docetaxel in Advanced Nonsquamous Non-Small-Cell Lung Cancer." *NEJM* 373: 1627-1639.

Cameron, A.C. et al. 2004. "Modelling the Differences in Counted Outcomes using Bivariate Copula Models with Application to Mismeasured Counts." *Econometrics Journal* 7: 566-584.

Carneiro, P. et al. 2001. "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policies." *Swedish Economic Policy Review* 8: 273-301.

Castagnoli, E. 1984. "Some Remarks on Stochastic Dominance." *Revista di Matematica per le Scienze Economiche e Sociali* 7: 15-28.

Chan, T.Y. and B.H. Hamilton. 2006. "Learning, Private Information, and the Economic Evaluation of Randomized Experiments." *Journal of Political Economy* 114: 997-1040.

Coate, S. 2000. "An Efficiency Approach to the Evaluation of Policy Changes." *Economic Journal* 110: 437-455.

Darling, D.A. 1957. "The Kolmogorov-Smirnov, Cramér-von Mises Tests." *Annals of Mathematical Statistics* 28: 823-838.

Denuit, M. et al. 2002. "Measuring the Impact of Dependence between Claims Occurrences." Insurance: Mathematics and Economics 30: 1-19.

Fan, Y. et al. 2014. "Identifying Treatment Effects under Data Combination." *Econometrica* 82: 811-822.

Fan, Y. et al. 2017. "Partial Identification of Functionals of the Joint Distribution of 'Potential Outcomes'." *Journal of Econometrics* 197: 42-59.

Fan, Y. and S.S. Park. 2010. "Sharp Bounds on the Distribution of Treatment Effects and Their Statistical Inference." *Econometric Theory* 26: 931-951.

Fan, Y. and S.S. Park. 2012. "Confidence Intervals for the Quantile of Treatment Effects in Randomized Experiments." *Journal of Econometrics* 167: 330-344.

Fenwick, E. et al. 2004. "Cost-Effectiveness Acceptability Curves—Facts, Fallacies and Frequently Asked Questions." *Health Economics* 13: 405-415.

Firpo, S. and G. Ridder. 2008. "Bounds on Functionals of the Distribution of Treatment Effects." Univ. of Southern California, IEPR Working Paper 08.09.

Gastwirth, J.L. 1975. "Statistical Measures of Earnings Differentials." *The American Statistician* 29: 32-35.

Gerber, E.R. and J.B. Lewis. 2004. "Beyond the Median: Voter Preferences, District Heterogeneity, and Political Representation." *Journal of Political Economy* 112: 1364-1383.

Hansen, B.E. 2017. *Econometrics* (online textbook). University of Wisconsin-Madison, Dept. of Economics.

Hauck, W.W. et al. 2000. Generalized Treatment Effects for Clinical Trials." *Statistics in Medicine* 19: 887-899.

Heckman, J.J. et al. 1997. "Making the Most out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts." *Review of Economic Studies* 64: 487-535.

Heckman, J.J. 2001. "Accounting for Heterogeneity, Diversity and General Equilibrium in Evaluating Social Programmes." *Economic Journal* 111: F654-F699.

Horwitz, R.I. et al. 1996. "Can Treatment That Is Helpful on Average Be Harmful to Some Patients? A Study of the Conflicting Information Needs of Clinical Inquiry and Drug Regulation." *Journal of Clinical Epidemiology* 49: 395-400.

Huang, E.J. et al. 2016. "Inequality in Treatment Benefits: Can We Determine If a New Treatment Benefits the Many or the Few?" *Biostatistics* (E-pub. ahead of print).

Imbens, G.W. and C.F. Manski. 2004. "Confidence Intervals for Partially Identified Parameters." *Econometrica* 72: 1845-1857.

Imbens, G.W. and J.M. Wooldridge. 2009. "Recent Developments in the Econometrics of Program Evaluation." *Journal of Economic Literature* 47: 5-86.

Jacob, J. and D. Lundin. 2005. "A Median Voter Model of Health Insurance with Ex Post Moral Hazard." *Journal of Health Economics* 24: 407-426.

Koenker, R. and Y. Bilias. 2001. "Quantile Regression for Duration Data: A Reappraisal of the Pennsylvania Reemployment Bonus Experiments." *Empirical Economics* 26: 199-220.

Kotz, S. and J.P. Seeger. 1993. "Lower Bounds on Multivariate Distributions with Preassigned Marginals." *Stochastic Inequalities* 22 (IMS Lecture Notes—Monograph Series): 211-218.

Kravitz, R.L. et al. 2004. "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages." *Milbank Quarterly* 82: 661-687.

Langley, R.G. et al. 2014. "Secukinumab in Plaque Psoriasis—Results of Two Phase 3 Trials."

NEJM 371: 326-338.

Lee, J.D. et al. 2016. "Extended-Release Naltrexone to Prevent Opioid Relapse in Criminal Justice Offenders." *NEJM* 374: 1232-1242.

Lee, M.-J. 2000. "Median Treatment Effects in Randomized Trials." *JRSS-B* 62: 595-604.

Lee, M.-J. and S. Kobayashi. 2001. "Proportional Treatment Effects for Count Response Panel Data: Effects of Binary Exercise on Health Care Demand." *Health Economics* 10: 411-428.

Lynn, J. et al. 2015. "Value-Based Payments Require Valuing What Matters to Patients." *JAMA* 314: 1445-1446.

Makarov, G.D. 1982. "Estimates for the Distribution Function of a Sum of Two Random Variables When the Marginal Distributions are Fixed." *Theory of Probability & Its Applications* 26: 803-806.

Mann, H.B. and D.R. Whitney. 1947. "On a Test of Whether One of Two Random Variables is Stochastically Larger than the Other." *Annals of Mathematical Statistics* 18: 50-60.

Manski, C.F. 1988. *Analog Estimation Methods in Econometrics.* London: Chapman and Hall.

Manski, C.F. 1996. "Learning about Treatment Effects from Experiments with Random Assignment of Treatments." *Journal of Human Resources* 31: 709-733.

Manski, C.F. 1997. "Monotone Treatment Effect." *Econometrica* 65: 1311-1334.

Manski, C.F. 2007. *Identification for Prediction and Decision.* Harvard University Press.

Mullahy, J. 2005. "Individual Results May Vary." Working Paper, Presented at the Conference on the Economics of Addiction and Health Inequality, Barcelona (May 2005).

Nissen, S.E. et al. 2016. "Cardiovascular Safety of Celecoxib, Naproxen, or Ibuprofen for Arthritis." *NEJM* 375: 2519-2529.

Nolan, T. and D.M. Berwick. 2006. "All-or-None Measurement Raises the Bar on Performance." *JAMA* 295: 1168-1170.

Pauly, M.V. 1989. "Positive Political Economy of Medicare, Past and Future." in M.V. Pauly et al., Eds., *Lessons from the First Twenty Years of Medicare—Research Implications for Public and Private Sector Policy*. Philadelphia: University of Pennsylvania Press.

Reck, M. et al. 2016. "Pembrolizumab versus Chemotherapy for PD-L1-Positive Non-Small-Cell Lung Cancer." *NEJM* 375: 1823-1833.

Rüschendorf, L. 2004. "Comparison of Multivariate Risks and Positive Dependence." *Journal of Applied Probability* 41: 391-406.

Stinnett, A.A. and J. Mullahy. 1998. "Net Health Benefits: A New Framework for the Analysis of Uncertainty in Cost-Effectiveness Analysis." *Medical Decision Making* 18: S68-S80.

Tamer, E. 2010. "Partial Identification in Econometrics." *Annual Review of Economics* 2: 167-195.

U.S. Food and Drug Administration. 2015. *Clinical Trial Endpoints for the Approval of Non-Small Cell Lung Cancer Drugs and Biologics—Guidance for Industry.* USDHHS/FDA CBER/CDER.

U.S. Food and Drug Administration. 2016. *Non-Inferiority Trials to Establish Effectiveness—Guidance for Industry.* USDHHS/FDA CBER/CDER.

U.S. Food and Drug Administration. 2017. *Multiple Endpoints in Clinical Trials—Guidance for Industry*. USDHHS/FDA CBER/CDER.

van Hout, B.A. et al. 1994. "Costs, Effects and C/E-Ratios alongside a Clinical Trial." *Health Economics* 3: 309-319.

Vanness, D.J. and J. Mullahy. 2012. "Moving beyond Mean-Based Evaluation of Health Care." Chapter 52 in A.M. Jones, Ed., *Elgar Companion to Health Economics*, 2nd Edition*.* Cheltenham: Edward Elgar.

Volpp, K.G. et al. 2009. "A Randomized, Controlled Trial of Financial Incentives for Smoking Cessation." *NEJM* 360: 699-709.

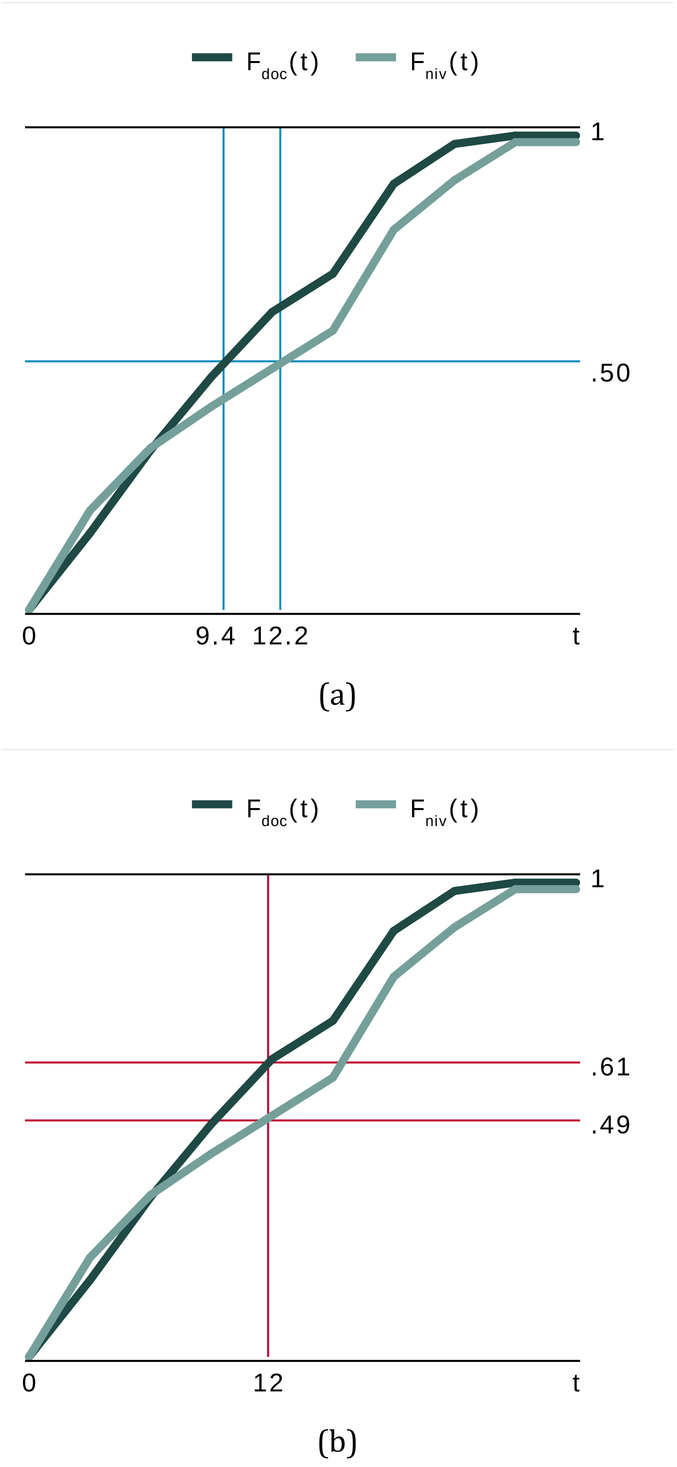
Willan, A.R. 2001. "On the Probability of Cost-Effectiveness Using Data from Randomized Clinical Trials." *BMC Medical Research Methodology* 1:8.

Willke, R.J. et al. 2012. "From Concepts, Theory, and Evidence of Heterogeneity of Treatment Effects to Methodological Approaches: A Primer." *BMC Medical Research Methodology* 12: 185.

Williamson, R.C. and T. Downs. 1990. "Probabilistic Arithmetic. I. Numerical Methods for Calculating Convolutions and Dependency Bounds." *International Journal of Approximate Reasoning* 4: 89-158.

Figure 1

Survival Time Distributions: Nivolumab=$F_{niv}\left(t\right)$ versus Docetaxel=$F_{doc}\left(t\right)$—
Panel (a): Median Survival Times; Panel (b): Twelve-Month Survival Probabilities



(a)

(b)

Figure 2

Survival Time, Nivolumab versus Docetaxel: Sample Space for Continuous and Binary Outcomes

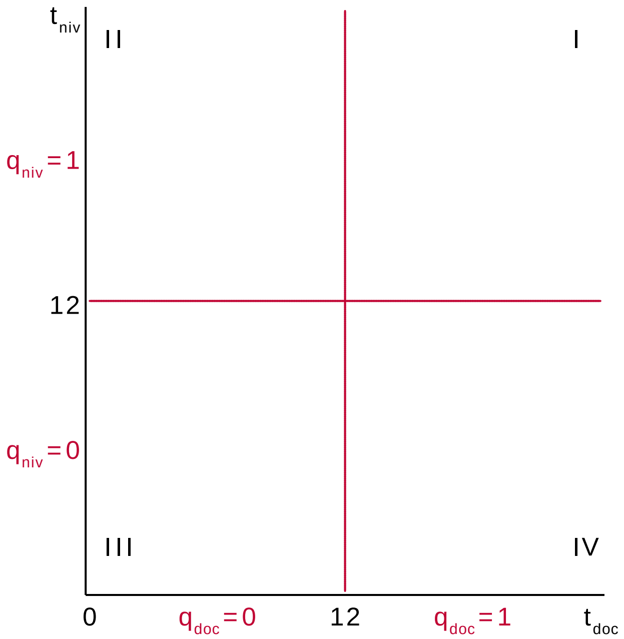
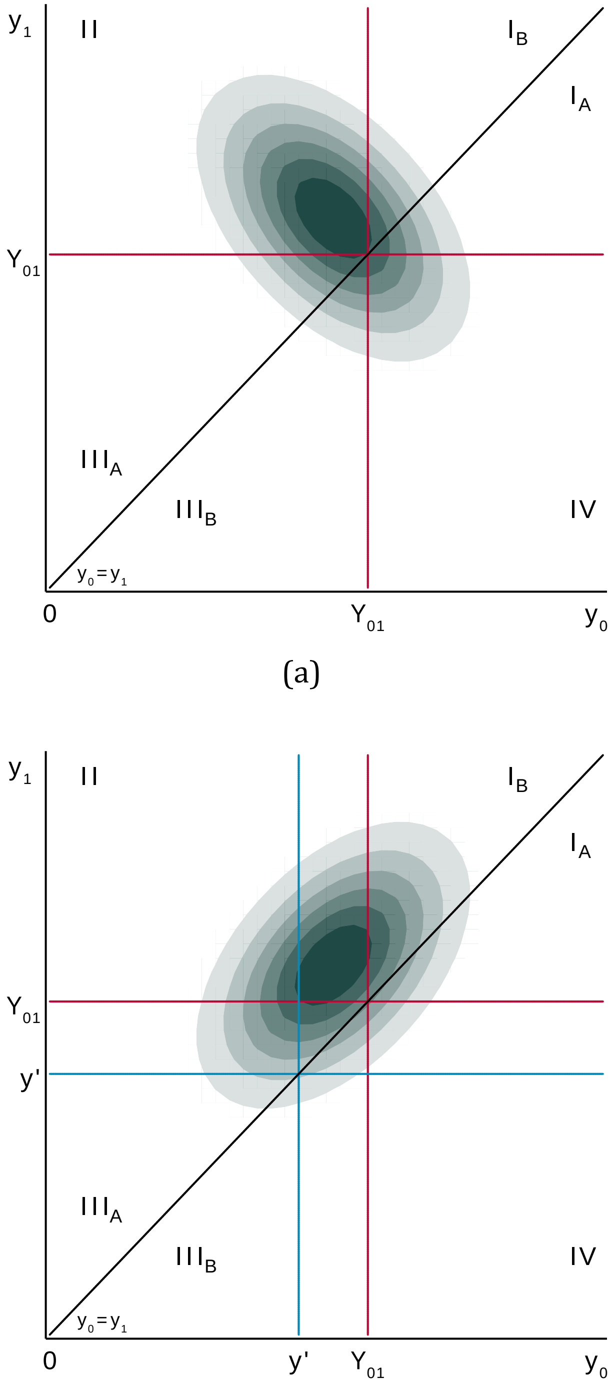


Figure 3

FLB based on Marginals $F_j(y)$ from Joint Distribution $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim \mathrm{BVN}\left(\begin{bmatrix} 4.1 \\ 5.1 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$ —

Panel (a): FLB for Arbitrary y=y'; Panel (b): Best FLB at $y = Y_{01}$

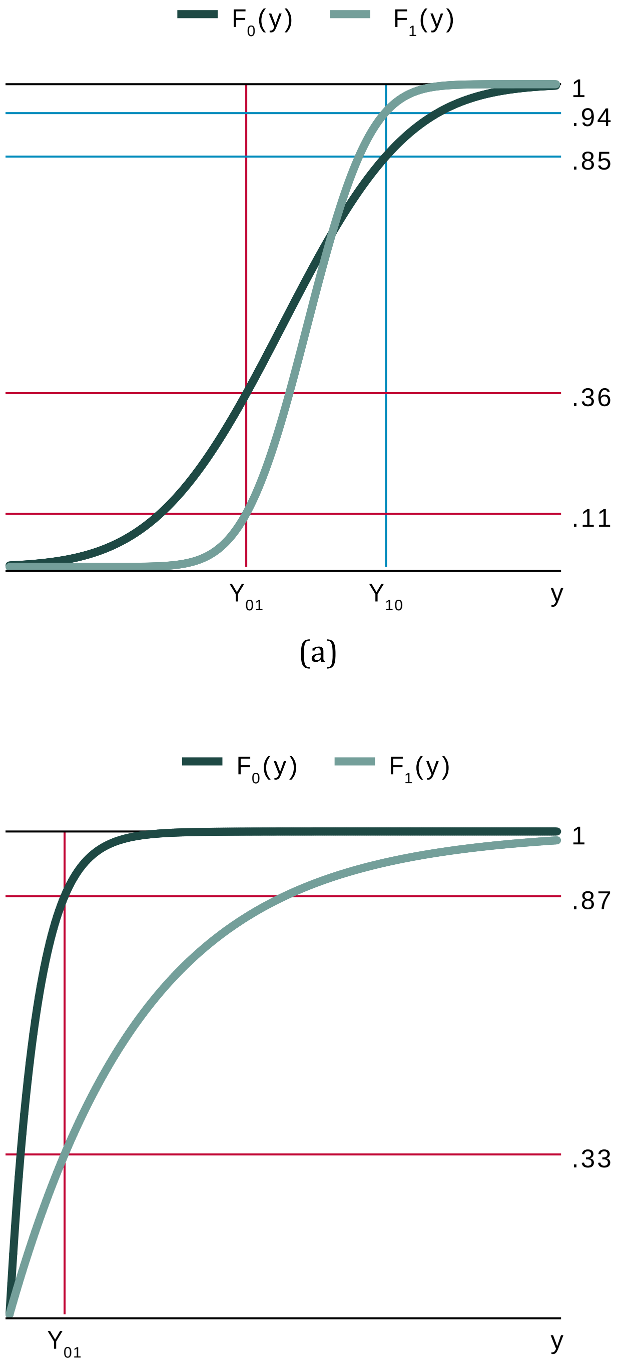


(a)

(b)

Figure 4

Numerical Examples: Computing Y_{jk} and D_{jk} —
Panel (a): $F_0\left(y\right)=N\left(0,4\right)$, $F_1\left(y\right)=N\left(.5,1\right)$; Panel (b): $F_0\left(y\right)=\text{Exp}\left(5\right)$, $F_1\left(y\right)=\text{Exp}\left(1\right)$.



(a)

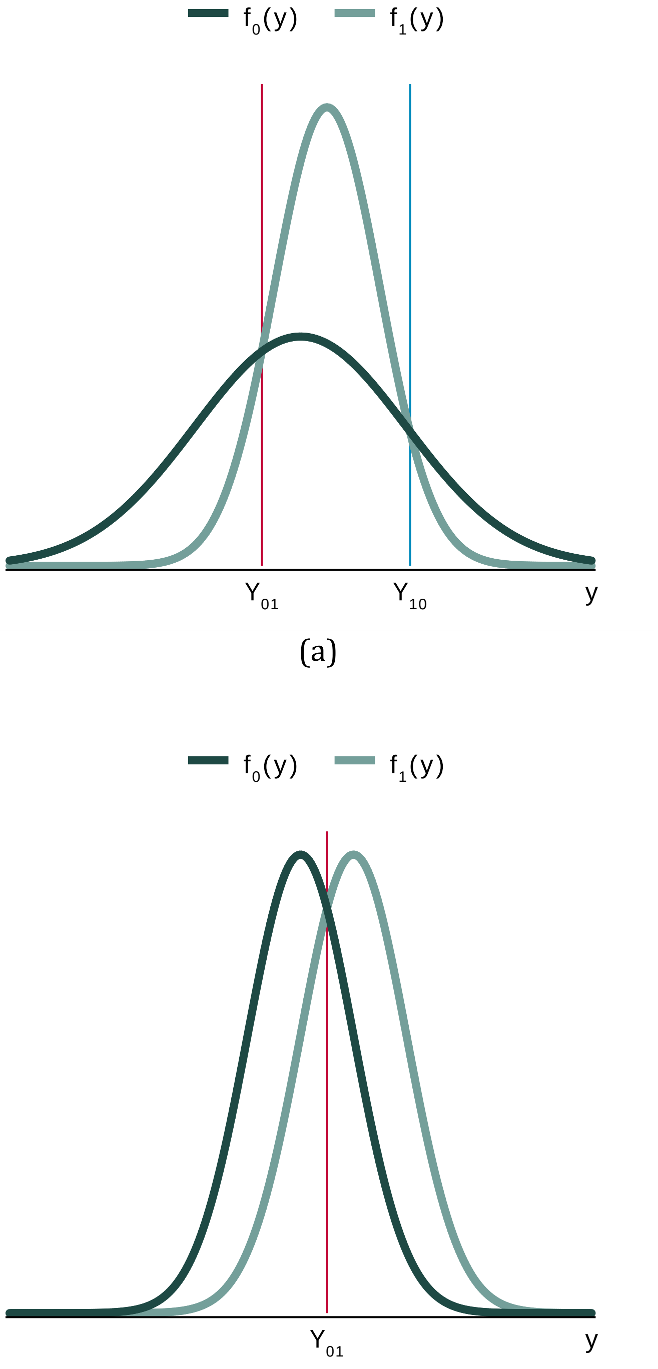
(b)

Figure 5

Numerical Examples: Computing Y_{jk} —

Panel (a): $\sigma_0^2 \neq \sigma_1^2$, Both Y_{01} and Y_{10} Defined;

Panel (b): $\sigma_0^2 = \sigma_1^2$, Only One of Y_{01} or Y_{10} Defined



(a)

(b)

Figure 6

Illustration of Zero-Order Stochastic Dominance, $Y_{01} = \left[U_0, L_1\right]$, and $D_{01} = 1$, with $F_1\left(y\right) \succ_0 F_0\left(y\right)$

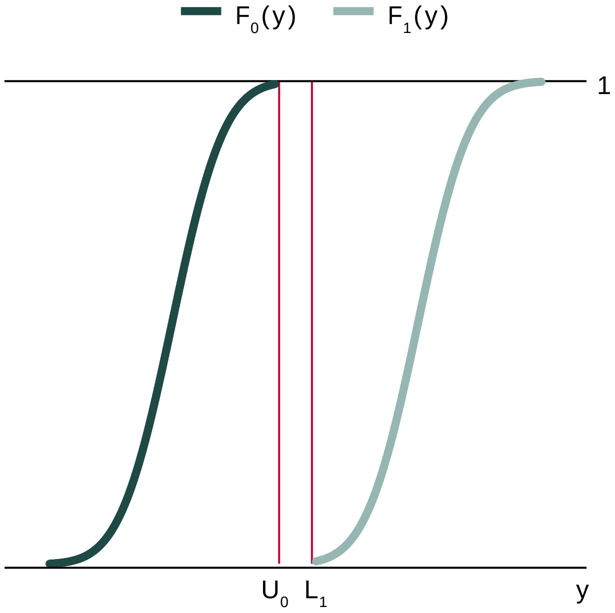
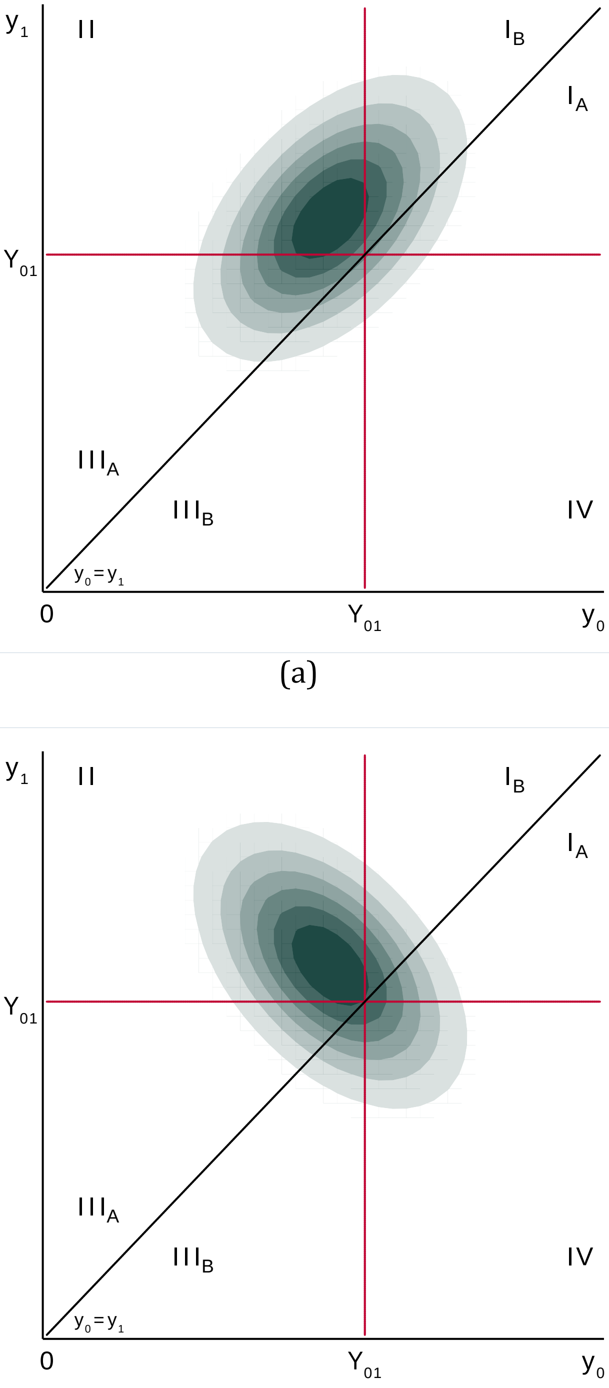


Figure 7

FLB based on Marginals $F_j\left(y\right)$ from Joint Distribution $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim BVN\left(\begin{bmatrix} 4.1 \\ 5.1 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right)$ —

Panel (a): $\rho = .5$; Panel (b): $\rho = -.5$

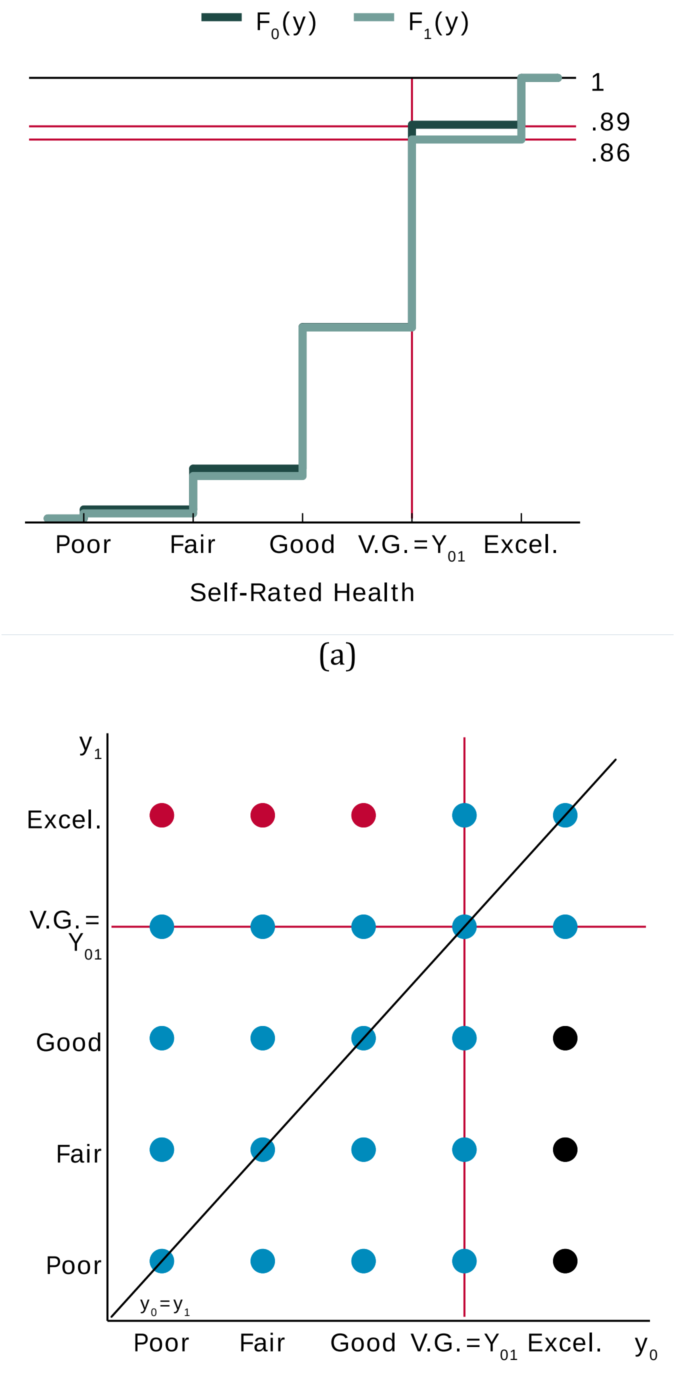


(a)

(b)

Figure 8

Volpp et al., 2009: Self-Rated Health Status Results, Computation of Y_{01} and D_{01} — Panel (a): Control=$F_0\left(y\right)$, Intervention=$F_1\left(y\right)$; Panel (b): Sample Space

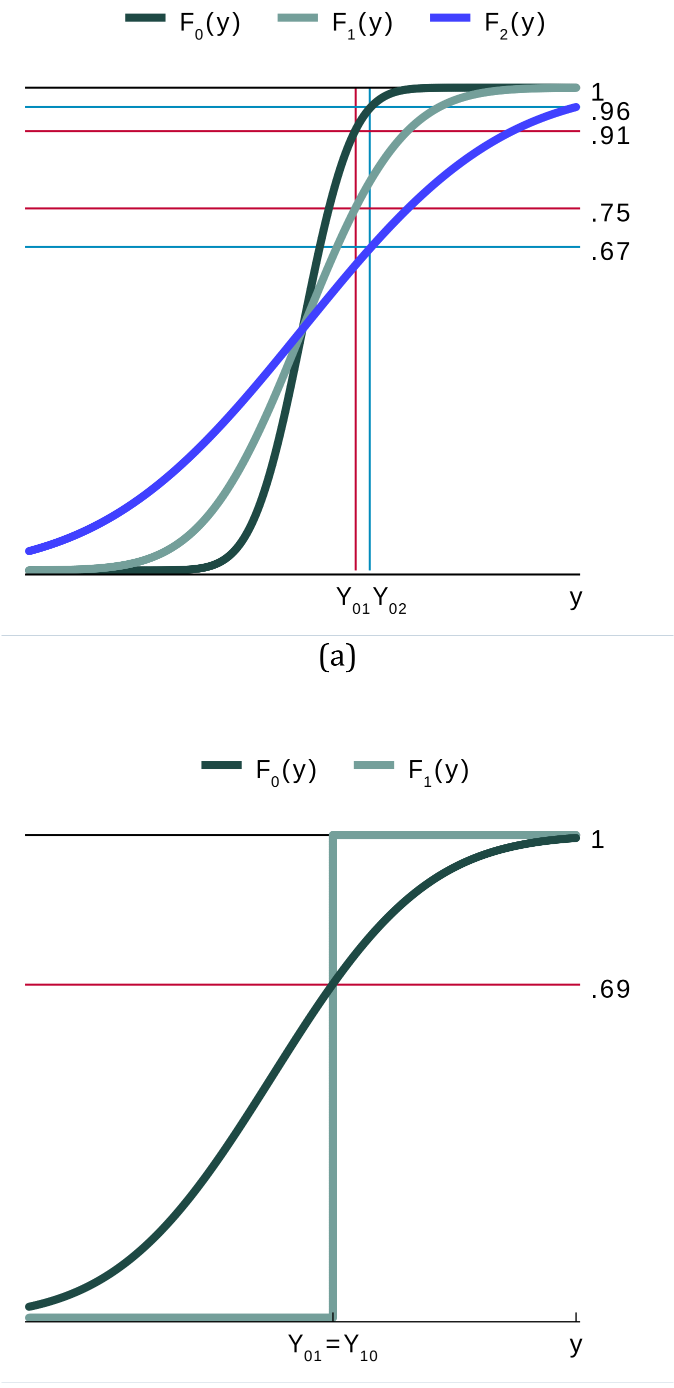


(a)

(b)

Figure 9

Numerical Examples: Spreading and Rectangularizing —
Panel (a): $F_0\left(y\right)=N\left(0,1\right)$, $F_1\left(y\right)=N\left(0,4\right)$, $F_2\left(y\right)=N\left(0,16\right)$;
Panel (b): $F_0\left(y\right)=N\left(0,4\right)$, $F_1\left(y\right)=N\left(1,0\right)$



(a)

(b)

Figure 10

Net Health Benefit: True $\Pr_{\lambda}\left(h_1 \geq h_0\right)$ and FLB based on Marginals $F_{0,\lambda}\left(h\right)$ and $F_{1,\lambda}\left(h\right)$

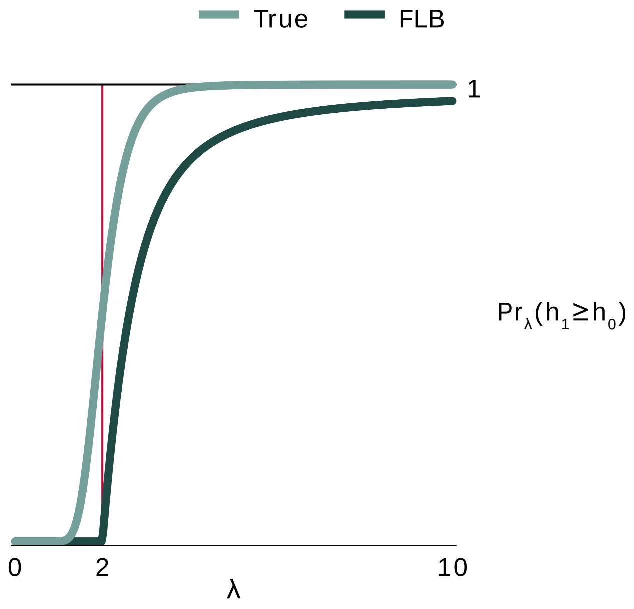


Figure 11

Demonstrating Fréchet-Boole Bounds with Three Outcomes:
Computation of D_{01} and D_{21} with $\left[y_0, y_1, y_2\right] \sim \mathrm{TVN}\left(\boldsymbol{\mu}, \mathbf{V}\right)$ (Parameters Defined in Text)

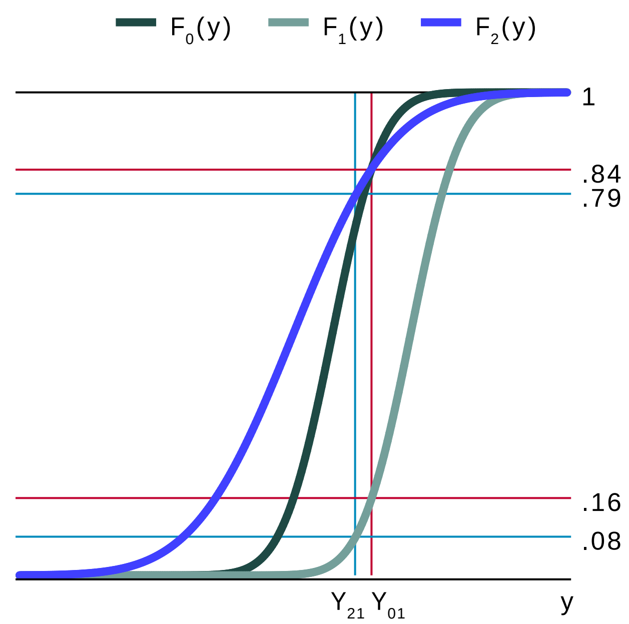


Figure 12

Lee et al., 2016, Relapse-Free Survival Time Results with Censoring:
Usual Treatment= $F_0\left(y\right)$, Naltrexone= $F_1\left(y\right)$, and Computation of Y_{01} and D_{01}

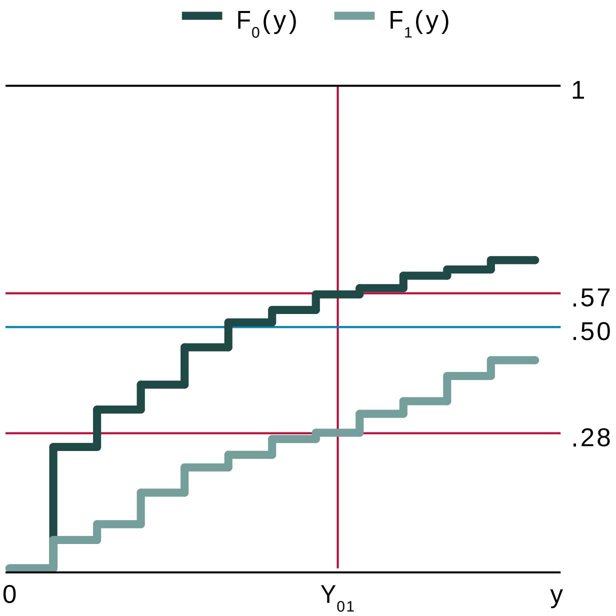


Table 1

Binary Outcomes Bounds on $\Pr\left(y_1 \geq y_0\right)$—
Panel (a): General Case; Panel (b): Nivolumab Example Probability Bounds

		y_0		Marginal Total
		0	1	
y_1	0	π_{00}	π_{10}	$1-\pi_1$
	1	π_{01}	π_{11}	π_1
Marginal Total		$1-\pi_0$	π_0	1

(a)

Twelve-Month Survival		Docetaxel (q_{doc})		Marginal Total
		Died	Survived	
Nivolumab (q_{niv})	Died	π_{00}	π_{10}	.49
	Survived	$.12 \leq \pi_{01} \leq .51$	π_{11}	.51
Marginal Total		.61	.39	1

(b)

Table 2

D_{01} and $\Pr\left(y_1 \geq y_0\right)$ for Alternative Mean and Correlation Structures;

$\left(y_0, y_1\right) \sim \text{BVN}\left(\mu_0, \mu_1; 1, 1, \rho\right)$

		$\Pr\left(y_1 \geq y_0\right)$ for $\rho =$				
$\mu_1 - \mu_0$	D_{01}	-.9	-.5	0	.5	.9
.5	.20	.60	.61	.64	.69	.87
1	.38	.70	.72	.76	.84	.99
2	.68	.85	.88	.92	.98	>.999

Table 3

Three-Outcome Case: Fréchet Bounds and Bounds Based on D_{01} and D_{21}

$F(y_0, y_1, y_2) = \text{MVN}(\boldsymbol{\mu}, \mathbf{V})$		ρ		
		-.25	0	.25
Population Parameters	$\Pr(y_1 > y_0)$	.90	.92	.95
	$\Pr(y_1 > y_2)$	.89	.91	.93
	$\Pr(y_1 > y_0 \wedge y_1 > y_2)$	.81	.85	.89
	FLB on $\Pr(y_1 > y_0 \wedge y_1 > y_2)$	.79	.83	.88
FLB on FLB using D_{01}, D_{21}		.39		

Exhibit 1

Using Stata's `ksmirnov` to Estimate D_{01} and D_{10}, with $y_0 \sim N(0,2)$ and $y_1 \sim N(.5,1)$

```
. by g: sum y

-------------------------------------------------------------------------
-----> g = 0

    Variable |       Obs        Mean    Std. Dev.     Min        Max
-------------+--------------------------------------------------------
           y |       500    .1139187    1.982952  -6.232434   6.283308

-------------------------------------------------------------------------
-----> g = 1

    Variable |       Obs        Mean    Std. Dev.     Min        Max
-------------+--------------------------------------------------------
           y |       500    .4553814    1.035908  -2.470682   3.612597


. ksmirnov y, by(g)

Two-sample Kolmogorov-Smirnov test for equality of distribution functions

 Smaller group       D       P-value  Corrected
 ----------------------------------------------
 0:                  0.2240    0.000
 1:                 -0.1200    0.001
 Combined K-S:       0.2240    0.000      0.000

. disp r(D_1)
.224

. disp r(D_2)
-.12
```