Published in final edited form as:

J Health Econ. 2018 September ; 61: 151–162. doi:10.1016/j.jhealeco.2018.06.011.

Individual results may vary: Inequality-probability bounds for some health-outcome treatment effects

John Mullahy^{a,b,c,*}

^aUniversity of Wisconsin-Madison, USA ^bNUI Galway, Ireland ^cNational Bureau of Economic Research, USA

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 of these 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 (“inequality probabilities”). While the properties of such parameters have been explicated in the technical literature, they have scarcely been considered in informing practical questions in health evaluations. This paper shows how such probabilities can be used informatively, and describes how they might be identified or bounded informatively given standard sampling assumptions and information only on marginal distributions of outcomes. The logic of these results and the empirical implementation thereof—sampling, estimation, and inference—are straightforward. Derivations are provided and several health-related applications are presented.

Keywords

Treatment effects; Probability bounds; Inequality probabilities

A chance to live longer

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 nonsquamous non-small-cell lung cancer (NSCLC). For treatment of NSCLC, two primary outcomes pre-specified¹ in a pivotal, phase-III randomized trial of nivolumab versus docetaxel (chemotherapy) were overall survival time and one-year overall survival rate (Borghaei et al., 2015). Fig. 1 depicts the data from which the Borghaei et al. summary results are computed.² The two reported measures of overall survival—both favoring nivolumab over docetaxel—are shown: an estimated difference in median survival times of

*Corresponding author at: Dept. of Population Health Sciences, 787 WARF, 610 Walnut St., University of Wisconsin-Madison, Madison, WI 53726, USA. jmullahy@wisc.edu.

¹ClinicalTrials.gov study NCT01673867.

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

2.8 months (=12.2–9.4) (left); and an estimated difference in twelve-month survival probabilities of 0.12 (=0.51–0.39) (right).

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

Introduction

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 may arise. The first is: “a chance to live longer,” compared to what? The second is: “a chance to live longer,” measured how? In light of the Borghaei et al. results, “compared to what” might reasonably be answered as “treatment with docetaxel” or, more generally, “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-percentage point 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 0.12 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 formally as the probability that a patient treated with nivolumab lives longer—not longer on average, not some particular amount longer,⁵ just simply longer by some unspecified amount—than had they otherwise been treated with docetaxel, i.e. $Pr(t_{niv} > t_{doc})$.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 directly answers this question is $Pr(t_{niv} > t_{doc})$.

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). As of this version’s date the application’s status was still pending.

⁴Either 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.

⁵Cost-effectiveness questions might balance “how many weeks longer” against cost across comparators. If such outcome and cost measures are converted to univariate net-benefit measures then the choice between two alternatives j and k amounts to knowing whether or not $NB_j > NB_k$.

⁶DTC 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. Understanding 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 denote two outcomes of interest (e.g. $y_0 = t_{doc}$ and $y_1 = t_{niv}$). The inequality probability⁷ $Pr(y_1 > y_0)$ provides an intuitive characterization of the extent to which one outcome is stochastically larger than another. This can be appreciated from its definition,

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

wherein $f(y_0, y_1)$ is the joint probability density of y_0 and y_1 . $Pr(y_1 > y_0)$ 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, $E[y_1 - y_0]$, measures like $Pr(y_1 > y_0)$ 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

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

with corresponding expected benefit⁸ (using standard “ $\wedge$ ” notation for “and”):

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

or, in the two-outcome case,

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

⁷The term “inequality probability” is used in this paper to refer to parameters $Pr(u > v)$ or $Pr(u \sim v)$ for arbitrary and possibly jointly distributed variables of interest, u and v .

⁸When 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.

Analogously imagine decisionmakers confronting possible loss from ex ante suboptimal choice of two competing binary treatments j and k given outcome 1 ($y_j > y_k$). Realized loss is zero if the ex post correct choice is made, else is one (of course “ex post correct” is not known in a counterfactual setting). Thus expected loss from choosing treatment j (resp. treatment k) is $1 - \Pr(y_j > y_k)$ (resp. $\Pr(y_j > y_k)$). Ex ante expected loss is smallest by choosing treatment j (resp. treatment k) if $\Pr(y_j > y_k) > .5$ (resp. < 0.5). Knowing $\Pr(y_j > y_k)$ is essential for appropriate treatment choice in this decisionmaking context.

Finally, reconsider the nivolumab example. If it is of interest⁹ to know the difference $\Pr(t_{niv} > 12) - \Pr(t_{doc} > 12)$, then $\Pr(t_{niv} > t') - \Pr(t_{doc} > t')$ may also be of interest for other t' 12 or over all possible t' .¹⁰ Yet these are different considerations than those involving $\Pr(t_{niv} > t_{doc})$ whose definition in (1) embeds consideration of all values of (t_{doc}, t_{niv}) jointly. 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, $\Pr(y_1 > y_0)$. Whether the outcomes of interest are survival times or perhaps other outcome-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 (although see section 4 for discussion of this point).

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 $V(\dots)$ denote some statistical functional defined on $F_j(y)$ (e.g. moment, quantile, probability, etc.), empirical investigations focus typically¹¹ 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 $V(\dots)$ 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. The characteristics of such decisionmakers will, of course, vary across settings and applications; in many contexts they will variously be patients, families, healthcare providers, or other agents confronted with choices between or among “treatments.”

⁹Presumably this quantity is of interest since it is one of the study’s primary endpoints; see [ClinicalTrials.gov](https://clinicaltrials.gov/study/NCT01673867) study [NCT01673867](https://clinicaltrials.gov/study/NCT01673867) and U.S. Food and Drug Administration (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 (1988, 2007).

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

¹¹Stochastic 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 .

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 Of particular importance in this general context, and for many of this paper's results, are randomized clinical trials (RCTs). Under ideal conditions such trials will identify the marginal distributions $F_j(y)$.¹³ The results that follow suggest how this identification may in turn be exploited to partially identify interesting features of the joint distribution of y_0 and y_1 .

This paper

Relationships to existing literature—This paper's focus intersects several broad themes that have been well developed in the literature: treatment-effect estimation and heterogeneous treatment effects¹⁴; decisionmaking criteria in stochastic environments¹⁵; and point versus interval identification of treatment effects.¹⁶ 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 coauthors¹⁷, a series of studies by Fan and coauthors¹⁸, as well as studies by Adams (2013)¹⁹, Basu and Thariani (2016), Firpo and Ridder (2008), Lee (2000), Manski (1997), and 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.²⁰ In particular, this paper attempts to exposit the essential features of such results—relying liberally on graphs that impart the paper's main results perhaps more readily than might the accompanying algebra (even though the algebra itself is straightforward)—and then extend and apply them to contexts of interest in health economics.

The paper's main results on inequality probabilities presented in section 3 are hardly novel, amounting essentially to tailored applications of Boole-Frechet probability bounds.²¹ 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.

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

¹³The main results rely on randomization or effective randomization. As such, the particular setting needn't be an RCT. Other randomized interventions as well as some quasi-randomized settings are also empirically relevant with respect to this paper's arguments. See section 6 for further discussion.

¹⁴Angrist (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 Biliias (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.

¹⁵Gerber and Lewis (2004); Grandmont (1978); Jacob and Lundin (2005); Stinnett and Mullahy (1998).

¹⁶Manski (1988, 2007).

¹⁷Aakvik et al. (2005), Abbring and Heckman (2007), Carneiro et al. (2001), Heckman (2001), and especially Heckman et al. (1997).

¹⁸Fan et al. (2014, 2017); Fan and Park (2010, 2012).

¹⁹As this paper 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.

²⁰Exceptions include Adams (2013), Cameron et al. (2004), and Huang et al. (2016).

²¹The 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).

The plan is as follows. Section 2 describes the main assumptions and notation. Section 3 presents the main results on probability bounds for continuously distributed outcomes. Section 4 extends the main results in several directions. Section 5 considers discrete outcome measures. Section 6 considers empirical implementation: sampling, estimation, and inference. Section 7 summarizes.

Definitions, assumptions, and notation

The setup here is familiar in the treatment-effect literature. M outcomes of interest, $(y_0, y_1, \dots)$, are jointly distributed in the population according to $F(y_0, y_1, \dots)$ with corresponding joint probability density denoted $f(y_0, y_1, \dots)$.²² $F(y_0, y_1, \dots)$ might be interpreted as representing a population heterogeneous in outcomes or as a joint distribution of random variables.²³

For now assume that there are two outcomes of interest, (y_0, y_1) , although more-general cases are considered in Section 4. Unless noted otherwise (y_0, y_1) are assumed to be continuously distributed. To be consistent with the technical definition of distribution functions the focus will be on $Pr(y_1 \geq y_0)$ instead of $Pr(y_1 > y_0)$ although these are essentially the same with continuously distributed outcomes.²⁴ The particular y_j measures may be ratio-scale, interval-scale, or any measure for which strict or weak inequality provides a meaningful comparison.

The population marginal distribution functions for the y_j are denoted $F_j(y) = Pr(y_j \leq y)$, $j = 0, 1$, for all y in their respective supports $S_j = [L_j, U_j]$. Of course the $F_j(y)$ are related to $F(y_0, y_1)$ via $F_j(y) = \int_{y_j \leq y} \int_{S_k} f(y_0, y_1) dy_k dy_j$, $j = k$. Until section 6 “conditional on $\mathbf{x}$ ” can be

assumed tacitly if appropriate, but will not be made explicit unless useful; the role of covariates $\mathbf{x}$ is revisited in section 6. Moreover until section 6 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 $\Delta_{01} = y_0 - y_1$.²⁵ 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 $F_{\Delta_{01}}(c) = Pr(y_0 - y_1 \leq c) = Pr(y_1 + c \geq y_0)$. Of interest in

most of what follows is $c = 0$, or $Pr(y_1 \geq y_0)$. $Pr(y_1 > y_0)$ is thus one feature of what is often called the treatment-effect distribution.

²²This notation is informal; formally, $F(c_0, c_1, \dots) = Pr((y_0 \leq c_0) \wedge (y_1 \leq c_1) \wedge \dots)$.

²³Outcomes are denoted in lower-case to keep notation concise. Distinctions between random variables and realizations should be clear from context.

²⁴For discrete outcomes the difference between weak and strict inequality will matter; see below.

²⁵Subject-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 $F(y_0, y_1)$ is absent.

Main results: bounds on inequality probabilities and related considerations

General results: bounding $\Pr(y_1 \leq y_0)$ using Boole-Fréchet probability bounds

For arbitrary, jointly distributed variables (z_0, z_1) and corresponding sets Z_0 and Z_1 , the Boole-Fréchet lower bound (“BFLB”) on the joint probability of the events $z_j \in Z_j$ is²⁶:

$$\Pr(z_0 \in Z_0 \wedge z_1 \in Z_1) \geq \max\{\Pr(z_0 \in Z_0) + \Pr(z_1 \in Z_1) - 1, 0\}, \quad (5)$$

which is informative if $\Pr(z_0 \in Z_0) + \Pr(z_1 \in Z_1) > 1$. For disjunctions (“or”, symbolized “ $\vee$ ”),

$$\Pr(z_0 \in Z_0 \vee z_1 \in Z_1) \geq \max\{\Pr(z_0 \in Z_0), \Pr(z_1 \in Z_1)\}. \quad (6)$$

The corresponding upper bounds (“BFUB”) are

$$\Pr(z_0 \in Z_0 \wedge z_1 \in Z_1) \leq \min\{\Pr(z_0 \in Z_0), \Pr(z_1 \in Z_1)\}, \quad (7)$$

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

$$\Pr(z_0 \in Z_0 \vee z_1 \in Z_1) \leq \min\{\Pr(z_0 \in Z_0) + \Pr(z_1 \in Z_1), 1\}, \quad (8)$$

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

For arbitrary y' , consider the events $y_0 \leq y'$ and $y_1 > y'$. Applying (5) gives

$$\begin{aligned} \Pr(y_0 \leq y' \wedge y_1 > y') &\geq \max\{\Pr(y_0 \leq y') + \Pr(y_1 > y') - 1, 0\} \\ &= \max\{F_0(y') + (1 - F_1(y')) - 1, 0\} \\ &= \max\{F_0(y') - F_1(y'), 0\}. \end{aligned} \quad (9)$$

This result is illustrated in Fig. 2 (left panel), depicting (y_0, y_1) -space and isodensity contours of $f(y_0, y_1)$ drawn using

$$\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim BVN\left(\begin{bmatrix} 4.1 \\ 5.1 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right).$$

²⁶Boole (1854), Chapter 19.

The arbitrary y' is indicated on both axes. Let $P(J)$ denote $Pr((y_0, y_1) \in J)$, where J is any of the six subspaces $I'_A, \dots, IV'$ in the figure having common “origin” (y', y') , with $I'_A \cup I'_B = I'$ and $III'_A \cup III'_B = III'$. Then:

$$Pr(y_0 \leq y') = P(II') + P(III'_A) + P(III'_B) \quad (10)$$

$$Pr(y_1 > y') = P(II') + P(I'_A) + P(I'_B) \quad (11)$$

$$Pr(y_0 \leq y') + Pr(y_1 > y') - 1 = P(II') - P(IV') \quad (12)$$

$$Pr(y_0 \leq y' \wedge y_1 > y') = P(II') \quad (13)$$

If (12) is positive then it is the BFLB on $Pr(y_0 \leq y' \wedge y_1 > y')$; for $P(IV') > 0$ this is smaller than the true probability in (13). Thus,

$$\begin{aligned} Pr(y_1 \geq y_0) &= P(I'_B) + P(II') + P(III'_A) \\ &\geq P(II') \\ &\geq P(II') - P(IV') \\ &= (P(II') + P(III'_A) + P(III'_B)) - (P(III'_A) + P(III'_B) + P(IV')) \\ &= F_0(y') - F_1(y'), \end{aligned} \quad (14)$$

wherein line three, $P(II') - P(IV')$, is the BFLB from (12). Thus, using only the marginals $F_j(y)$ potentially informative lower bounds on $Pr(y_0 \leq y' \wedge y_1 > y')$ and thus $Pr(y_1 > y_0)$ are obtained. Analogously, if informative the BFUB on $Pr(y_0 \leq y' \vee y_1 > y')$ follows from (8) as $F_0(y') + (1 - F_1(y'))$, seen by noting that $P(I') + P(II') + P(III') \geq P(I'_B) + P(II') + P(III'_A) = Pr(y_1 \geq y_0)$ with reference to Fig. 2, then applying (8) for the event $(y_0 \leq y' \vee y_1 > y')$.

Best bounds on $Pr(y_1 \geq y_0)$

Since $Pr(y_1 \geq y_0) \geq Pr(y_1 > y_0) \geq Pr(y_0 \leq y' \wedge y_1 > y')$ a nonzero BFLB on $Pr(y_0 \leq y' \wedge y_1 > y')$ is partially informative as a lower bound on $Pr(y_1 \geq y_0)$. Since y' is arbitrary, however, such a bound will generally not be a sharp or best possible such bound on $Pr(y_1 \geq y_0)$. Intuitively from (14), a best lower bound on $Pr(y_1 \geq y_0)$ would be defined by determining the value(s) of y such that the difference between $F_0(y)$ and $F_1(y)$ is greatest.

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 6.²⁷ Rearranging the expressions of FP's Lemma 2.1 and Eq. (2), and defining S as the common support of $F_0(y)$ and $F_1(y)$,²⁸ FP show (in this paper's notation) that for arbitrary c :

$$\begin{aligned} \sup_{y \in S} \max\{F_0(y) - F_1(y - c), 0\} &\leq Pr(y_1 \geq y_0 + c) \quad (15) \\ &\leq \inf_{y \in S} \min\{1 + F_0(y) - F_1(y - c), 1\}. \end{aligned}$$

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

$$\begin{aligned} \sup_{y \in S} \max\{F_0(y) - F_1(y), 0\} &\leq Pr(y_1 \geq y_0) \quad (16) \\ &\leq \inf_{y \in S} \min\{1 + F_0(y) - F_1(y), 1\}. \end{aligned}$$

That is, the greatest lower bound on the inequality probabilities $Pr(y_1 \geq y_0)$ identifiable from the marginals $F_j(y)$ is the maximum (sup) over all y of the difference (if positive) between $F_0(y)$ and $F_1(y)$, while the smallest upper bound is the minimum (inf) over all y of $1 + F_0(y) - F_1(y)$ if this is less than one. In essence, these best bounds are found by searching over $y \in S$ to determine where the BFLB and BFUB are greatest and smallest. These results are the foundation for what follows.

At this point some additional notation will be useful. For $j, k \in \{0, 1, \dots\}$, $j \neq k$:

$$\delta_{jk}(y) = F_j(y) - F_k(y) \quad (17)$$

$$D_{jk} = \max_{y \in S} \{\delta_{jk}(y), 0\} \quad (18)$$

$$Y_{jk} = \operatorname{argmax}_{y \in S} (\delta_{jk}(y)) \text{ if } D_{jk} > 0, \text{ undefined if } D_{jk} = 0 \quad (19)$$

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 quantities 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.²⁹ 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

²⁷The 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).

²⁸ $S = [\min\{L_0, L_1\}, \max\{U_0, U_1\}]$. FP discuss technical considerations involved in defining the relevant supports, i.e. the domains of the sup and inf in (16).

²⁹See 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).

important results to follow go through whether or not uniqueness holds (FP discuss the role of uniqueness).

To visualize the result in (16), consider Fig. 2 (right panel) in which the six subspaces $I_a, I_b, \dots, IV$ have common origin $y_0 = y_1 = Y_{01}$ (red reference lines). 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 (14):

$$\begin{aligned} Pr(y_1 \geq y_0) &= P(I_B) + P(II) + P(III_A) \\ &\geq P(II) \\ &\geq P(II) - P(IV) \\ &= (P(II) + P(III_A) + P(III_B)) - (P(III_A) + P(III_B) + P(IV)) \\ &= F_0(Y_{01}) - F_1(Y_{01}) \\ &= D_{01}. \end{aligned} \quad (20)$$

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 BFLB is not informative. Analogous arguments establish that, if it is informative, the best BFUB is $1 - D_{10}$.

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

$$D_{jk} \leq Pr(y_k \geq y_j) \leq 1 - D_{kj}. \quad (21)$$

For example, in the example depicted in the right panel of Fig. 2 $Y_{01} = 4.6$, Y_{10} is undefined, $D_{01} = .38$, and $D_{10} = 0$. With respect to the nivolumab example, the “chance to live longer,” $Pr(t_{niv} \geq t_{doc})$, is at least .17 but not greater than .96, sampling considerations notwithstanding.³⁰

Numerical examples

To make this discussion concrete, two numerical examples are offered in Fig. 3. Its left panel shows two $N(\mu_j, \sigma_j^2)$ marginal distributions, where $F_0(y)$ is $N(0, 4)$ and $F_1(y)$ is $N(.5, 1)$. This yields $Y_{01} = -.73$, $Y_{10} = 2.07$, $D_{01} = .36 - .11 = .25$, and $D_{10} = .94 - .85 = .09$. The right panel shows results for exponential marginals, where $F_0(y)$ is $Exp(5)$ and $F_1(y)$ is $Exp(1)$. These assumptions result in $Y_{01} = .40$, Y_{10} undefined, $D_{01} = .87 - .33 = .54$, and $D_{10} = 0$.³¹

³⁰If additional assumptions are made then point identification of $Pr(y_1 \geq y_0)$ may be possible. For example, if a common effect ($y_1 = \alpha + y_0$) or proportional effect ($y_1 = \alpha y_0$) structure is assumed then $Pr(y_1 \geq y_0)$ is either zero or one depending on the sign/magnitude of α . Alternatively, with perfect negative dependence, i.e. $P(II) + P(IV) = 1$ and $P(II) \times P(IV) > 0$ in Fig. 2's right panel, $Pr(y_1 \geq y_0)$ is point-identified as $.5(1 + D_{01} - D_{10})$. (While negative dependence may be unreasonable to entertain in many contexts, one might nevertheless usefully recall Koenker and Biliias (2001), who remark that “the treatment may make otherwise weak subjects especially robust, and turn the strong to jello.”) Beyond these extreme cases, a referee has raised the interesting question of whether application-plausible assumptions on the jointness structure of $F(y_0, y_1)$ might allow the bounds (21) to be partially narrowed yet still fall short of point identification. Frandsen and Lefgren (2018), explore such ideas by assuming the outcomes are mutually stochastically increasing. See Heckman et al. (1997), Sections 3 and 4 for related discussion.

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 1 are D_{01} and the true $Pr(y_1 \geq y_0)$ for selected $\mu_1 - \mu_0$ and ρ (these 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 the two panels of Fig. 4).

Relationships with stochastic dominance

Consider first the case of zero-order stochastic dominance (ZSD; Castagnoli, 1984). $F_1(y)$ 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 Fig. 5). With ZSD $Pr(y_1 \geq y_0) = 1$, i.e. regardless of a population member's outcome in $F_0(y)$, that outcome will be less than their outcome in $F_1(y)$.³² Note that $D_{01} = 1$ for any Y_{01} in $[U_0, L_1]$ so that $Pr(y_1 \geq y_0)$ is point-identified as $Pr(y_1 \geq y_0) = 1$, i.e. the BFLB 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 $D_{10} = 0$.

Extensions and additional considerations

$Pr(y_1 \geq y_0)$ 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: $Pr(y_1 \geq y_0) = \int_{-\infty}^{\infty} f_0(y_0) \int_{y_0}^{\infty} f_1(y_1) dy_1 dy_0$. For example, consider the exponential case in the right panel of Fig. 4. The BFLB on $Pr(y_1 \geq y_0)$ for any dependence structure between y_0 and y_1 is $D_{01} = .54$, whereas $Pr(y_1 \geq y_0)$ under independence is $\theta_0/(\theta_0 + \theta_1) = .83$.

Relationships to permutation distributions

Let $y_{0,N} = [y_{0,n}]$ and $y_{1,N} = [y_{1,n}]$ denote N -vectors describing outcomes for a sample or a finite population of size $2N$. Let $P(y_{0,N})$ be the $N \times N!$ matrix containing the $N!$ permutations of the elements of $y_{0,N}$; let $C = [y_{1,N} - P(y_{0,N})_c]$ be the $N \times N!$ matrix whose c -th column is the difference between $y_{1,N}$ and the c -th column of $P(y_{0,N})$; and let $d = [\frac{1}{N} \sum_{n=1}^N 1(C_{n,c} > 0)]$ be the $1 \times N!$ vector describing the fraction of elements in each of C 's columns for which $y_{1,n} > y_{0,n(c)}$. Then the smallest and largest elements of d are D_{01} and $1 - D_{10}$, respectively. These relationships are discussed by Heckman et al. (1997).

³¹If $F_j(y)$ is $N(\mu_j, \sigma_j^2)$ then Y_{01} and Y_{10} are given by the quadratic formula with $a = \sigma_1^2 - \sigma_0^2$, $b = -2(\sigma_1^2 \mu_0 - \sigma_0^2 \mu_1)$, and $c = \sigma_1^2 \mu_0^2 - \sigma_0^2 \mu_1^2 - 2\sigma_0^2 \sigma_1^2 \ln(\sqrt{\sigma_1^2/\sigma_0^2})$ 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 $.5(\mu_0 + \mu_1)$ (Y_{01} if $\mu_0 < \mu_1$; Y_{10} if $\mu_0 > \mu_1$). If $F_j(y)$ is exponential, $F_j(y) = 1 - \exp(-\theta_j y)$, then $D_{01} = \exp(-\theta_1 Y_{01}) - \exp(-\theta_0 Y_{01})$, $D_{10} = 0$, $Y_{01} = \ln(\theta_1/\theta_0)/(\theta_1 - \theta_0)$, and Y_{10} is undefined if $\theta_0 > \theta_1$; the subscripts are reversed if $\theta_1 > \theta_0$. While parametric models may be instructive for illustrative and modeling purposes, applications often consider nonparametric empirical distributions. Estimating the D_{jk} nonparametrically is discussed in section 6.

³²If y represents net benefit, then a policy shifting $F_0(y)$ to $F_1(y)$ yields a Pareto improvement.

Alternative characterizations of τ_0 and transformations

Beyond τ_0 , other contrasts may be of interest. If the relevant contrast is $t(y_0) - t(y_1)$ with $t(\dots)$ monotone-increasing, the previous results apply: $Pr(t(y_1) - t(y_0)) = Pr(y_1 - y_0)$, $Y_{jk,t} = t(Y_{jk})$, $D_{01,t} = F_{0,t}(Y_{01,t}) - F_{1,t}(Y_{01,t}) = F_0(Y_{01}) - F_1(Y_{01})$, etc., using obvious notation. For positive y_j contrasts of interest might also involve ratios, $Pr((y_0/y_1) - c)$, or proportional differences, $Pr(((y_0 - y_1)/y_0) - c)$.³³

These ideas can be applied broadly. For instance, the inequality-probability discussion thusfar ignores the magnitudes by which y_1 might exceed y_0 , focusing only on the sign of such exceedance. Yet in some situations the amount by which y_1 might exceed y_0 may be decision-relevant. As an example, the imagined question stated in section 1 might be recast in such circumstances as: “If I’m treated with treatment 1, what is the chance I’ll live at least c months longer than if I’m treated with treatment 0?” Living longer might only be ex ante worthwhile to a patient if, for example, it involves living sufficiently longer—e.g. c months or more—to abide treatment 1’s known side effects. Knowing $Pr(y_1 - y_0)$ does not inform this decision; knowing $Pr(y_1 - y_0 + c)$ is what matters. Transforming to obtain the distribution of $y_1 - c$ and applying the previous results provide bounds on the probability of interest, $Pr(y_1 - y_0 + c)$.³⁴ Non-inferiority margin or point-spread assessments may also concern probabilities like these.

So long as c and/or $t(\dots)$ are known all these cases can be subsumed by specifying $\tau_0(y_0) - \tau_1(y_1)$ and considering $Pr(\tau_1(y_1) - \tau_0(y_0))$. For example, in the proportional-difference case $\tau_0(y_0) = (1 - c)y_0$ and $\tau_1(y_1) = y_1$. The previous results go through if the $F_j(y)$ reference the distributions of the transformed measures.

Spreading or rectangularizing distributions

Fig. 6 (left panel) 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 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 $Pr(y_1 - y_0)$.

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 $Pr(y_1 - y_0)$ is point-identified: $1 - D_{10} = D_{01}$.

$Pr(y_1 - y_0) = D_{01}$. For example, Fig. 6 (right panel) 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 $Pr(y_1 - y_0)$ is point-identified at .69.

Inequality probabilities with more than two outcomes

At the cost of some additional algebra, inequality probabilities bounds involving $M \geq 3$ outcomes, e.g. $Pr((y_1 - y_0) \wedge (y_1 - y_2))$, can be determined using generalized Boole-Fréchet

³³See Imbens and Wooldridge (2009), and Lee and Kobayashi (2001), and Lee (2005), for conceptual considerations, and Langley et al. (2014) for a related application.

³⁴Considerations like these are essentially those of the general literature on quantile treatment effects and, indeed, of FP’s paper.

bounds that extend (5) and (7) to multivariate contexts (Boole, 1854; Rüschendorf, 2004). As such,

$$\begin{aligned} & \max\{Pr(z_0 \in Z_0) + \dots + Pr(z_{M-1} \in Z_{M-1}) - (M-1), 0\} \quad (22) \\ & \leq Pr(z_0 \in Z_0 \wedge \dots \wedge z_{M-1} \in Z_{M-1}) \\ & \leq \min\{Pr(z_0 \in Z_0), \dots, Pr(z_{M-1} \in Z_{M-1})\}. \end{aligned}$$

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. It might be of interest in settings like this to consider the probability that one treatment's safety profile exceeds that of both of the other treatments.

Alternatively, in some evaluations the outcomes of interest under each treatment are multivariate, i.e. y_0 and y_1 are P-vectors. Assume that the $M = 2$ P-dimensional joint marginals $F_j(y)$, $j = 0, 1$, are identified, which may be a reasonable assumption in some applications. Using results from Rüschendorf (2004), and referring to vector inequalities that are element-by-element, the joint probability of the events $y_0 \leq y'$ and $y_1 > y'$ for arbitrary y' is bounded as follows:

$$\begin{aligned} & \max\{F_0(y') - F_1(y'), 0\} \leq Pr(y_0 \leq y' \wedge y_1 > y') \quad (23) \\ & \leq \min\{1 + F_0(y') - F_1(y'), 1\}. \end{aligned}$$

Obtaining the best bounds on $Pr(y_1 \leq y_0)$ follows in a manner analogous to (21) except that determining the particular Y_{jk} (the vector analog of Y_{jk} in (19)) at which the $F_j(y)$ are evaluated to identify the best bounds would entail additional computational considerations.

35

Discrete and ordinal outcomes

The main results on identifying bounds on inequality probabilities apply also when population outcome measures are binary, integer-valued (e.g. count-data; see Cameron et al., 2004), discrete-ordinal, or categorical measures (e.g. health-status scores, Likert scales).³⁵ One important consideration now is whether the parameter of interest is $Pr(y_1 \leq y_0)$ or $Pr(y_1 > y_0)$ since in the population a nonzero probability of ties, i.e. the event $y_0 = y_1$, becomes relevant.³⁷ The approach described in section 3 that identifies the Y_{jk} and D_{jk} is again applicable here, but the quantity whose bounds are identified as such is $Pr(y_1 > y_0)$, not $Pr(y_1 \leq y_0)$.³⁸

³⁵In a related context Andrews (1997), discusses how a grid search over y' can be confined to the sample values of y as these define the steps in the empirical joint distribution, thus simplifying estimation. Note that the M elements of Y_{jk} will generally not be the M scalar values that would obtain from applying (19) with reference to the M univariate marginals.

³⁶Huang 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.

³⁷In 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.

³⁸For the Y_{jk} to be (potentially) unique when outcomes are discrete or categorical, the domain of the argmax in (19) should be redefined as the set $\{y | Pr(y_0 = y \vee y_1 = y) > 0\}$.

Consider again the nivolumab vs. 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 Fig. 7 let roman numerals I, ...,IV denote the four subspaces with origin $(t_{doc}, t_{niv}) = (12, 12)$. Then the reported result on twelve-month survival when the survival times are treated as being continuously distributed can be obtained as³⁹

$$\begin{aligned} Pr(t_{niv} \geq 12) - Pr(t_{doc} \geq 12) &= Pr((t_{doc}, t_{niv}) \in I \cup II) \\ &\quad - Pr((t_{doc}, t_{niv}) \in I \cup IV) \\ &= Pr((t_{doc}, t_{niv}) \in II) - Pr((t_{doc}, t_{niv}) \in IV) \\ &= .12 \end{aligned} \quad (24)$$

Now suppose outcomes are the corresponding binary indicators of 12-month survival $q_j = 1$ ($t_j \geq 12$), $j \in \{doc, niv\}$ under the two treatments. The general joint and marginal probability structure is shown in the top panel of Table 2. For the strict-inequality event

$$\begin{aligned} q_{niv} > q_{doc} \\ Pr(q_{niv} > q_{doc}) &= Pr(q_{doc} = 0 \wedge q_{niv} = 1) = \pi_{01} \\ &= Pr((q_{doc}, q_{niv}) \in II). \end{aligned} \quad (25)$$

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

$$\max\{0, \pi_1 - \pi_0\} \leq \pi_{01} \leq \min\{1 - \pi_0, \pi_1\}. \quad (26)$$

The lower bound, $\pi_1 - \pi_0$, is $Pr((t_{doc}, t_{niv}) \in II) - Pr((t_{doc}, t_{niv}) \in IV)$, coinciding with (24) and corresponding to the general characterization (5); analogously the upper bound, $\min\{1 - \pi_0, \pi_1\}$, follows from (7). Applying this result to the nivolumab example one finds $\pi_{01} = .51$, i.e. notwithstanding sampling error $Pr(q_{niv} > q_{doc})$ is at least .12 but not greater than .51; see the bottom panel of Table 2.

The 2009 study by Volpp et al. on the effects of financial incentives on smoking cessation and related outcomes provides an example based on ordinal outcomes.⁴⁰ 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 Fig. 8 (left panel). Treatment effects using this measure are assessed by examining differences between treatment and control separately at the five Likert scale points (see their Fig. 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 Fig. 8 (right panel) which depicts the sample space; $D_{01} = .03$ corresponds to the

³⁹At this point these estimates are treated as if population parameters. This example's empirical properties are considered in Section 6.

⁴⁰See Scharfstein et al. (2004), for further discussion of the comparison of ordered-outcome distributions using stochastic dominance strategies.

probability mass of the red dots minus that of the black dots. (This study is considered further in Section 6.)

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 $F_{j,N_j}(y) = \frac{1}{N_j} \sum_{n=1}^{N_j} 1(y_{j,n} \leq y)$.

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.”⁴¹ In essence, a random sample containing information on the true (y_0, y_1) , or more generally (y_0, y_1, x) , 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.⁴² These assumptions are standard and point to what matters being consistent estimates of the $F_j(y)$ in the sense of convergence in distribution: $F_{j,N_j}(y) \rightarrow F_j(y)$ as $N_j \rightarrow \infty$ 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.

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 BFLB or BFUB depending on where censoring occurs relative to the uncensored data's Y_{jk} . Informative bounds on $Pr(y_1 > y_0)$ 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 Fig. 2) are depicted in Fig. 9. 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.⁴³

⁴¹All 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).

⁴²See FP, pages 932 and 944–945, and Imbens and Wooldridge (2009), Section 2.2. In an unconfounded regression context with $y = \alpha t + m(x) + u$, $E[u|t, x] = 0$, and $t \in \{0, 1\}$, an analyst might imagine empirical bounds analysis using as “outcomes” the estimated adjusted or semi-residuals $\hat{r} = y - \hat{m}(x)$ from the two sub-samples defined by the binary treatment indicator. Consideration of the properties of such an approach is left for future exploration.

⁴³When 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 marginal quantiles is observed. For instance, while it is not possible to identify $med(F_1(y)) - med(F_0(y))$ in the Lee et al. example, it is evident from Lee et al.'s Fig. 2 that this difference is at least 13 weeks.

Nothing about the above results demands that data on the y_j 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 $F(y_0, y_1)$. If the marginal distributions of the outcomes observed in different datasets (e.g. repeated cross-sections, synthetic panels, separate trials, etc.) are representative of the same population—characterized by time, place, and all other observable and unobservable characteristics—then the previous results are applicable.⁴⁴

Estimation

Estimation of the D_{jk} requires an algorithm that computes the maximum and minimum differences between $F_{0,N_0}(y)$ and $F_{1,N_1}(y)$ across their common support. In Stata, this is straightforward using the *ksmirnov* procedure.⁴⁵ With data on $y_{0,n}$ and $y_{1,n}$ stacked into a single variable (say $y = [y_n]$) having $N_0 + N_1$ observations, and a second binary variable (say $g = [g_n]$) indicating group membership, e.g. $g_n = 1$ ($n > N_0$), 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 simulated from each of the $N(0, 4)$ and $N(.5, 1)$ distributions depicted in left panel of Fig. 3. 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 0.224 and .12, corresponding to their respective population counterparts 0.25 and 0.09 shown in Fig. 3.

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 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 $Pr(y_1 > y_0)$.

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

$$\sqrt{N}(D_{jk,N} - D_{jk}) \rightarrow N(0, \sigma_{jk}^2) \quad (27)$$

where

⁴⁴The 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.

⁴⁵R has a procedure, *ks.test*, that appears to provide output similar to that of Stata's *ksmirnov*.

$$\sigma_{jk}^2 = F_0(Y_{jk})(1 - F_0(Y_{jk})) + F_1(Y_{jk})(1 - F_1(Y_{jk})) \quad (28)$$

assuming equal sample sizes in the two groups (this is easily relaxed) and that various regularity conditions⁴⁶ are met. Confidence intervals built on these large-sample results must also respect the 0–1 probability bounds. For the data in Fig. 1, using (28) to compute 95% 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 $Pr(y_1 > y_0)$. 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 $p_{jk,N} = \exp\left(\frac{2N_0N_1}{N_0 + N_1} D_{jk,N}^2\right)$ for the null that y_j is not stochastically smaller than y_k ; the $p_{jk,N}$ depend only on the $D_{jk,N}$, not on the particular values of the $F_{j,N_j}(y)$.⁴⁷

An example

As an example of the use of these approaches to analyze actual trial data, consider again the SRHS outcome data studied by Volpp et al. (2009), in their smoking-cessation incentive trial. For analytical purposes the SRHS outcome data can be coded in any manner that preserves the {poor,...,excellent} ordering. The information reported in that paper's Fig. 2 indicates 442 subjects in the control group (treatment = 0) and 436 subjects in the financial-incentive intervention group (treatment = 1). The results are summarized in Exhibit 2. The ksmirnov returned results $r(D_1)$ and $r(D_2)$ imply estimated BFLB and BFUB on $Pr(y_1 > y_0)$ of .0336 and 1.0, respectively. These correspond to the result reported earlier in section 5 and depicted in Fig. 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.

⁴⁶FP'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 $F_j(y)$) and analog ($F_{j,N_j}(y)$ "plugged in") versions of (28). These results are available on request.

⁴⁷See 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 .

Questions regarding inequality probabilities are natural to consider in a range of decisionmaking settings.⁴⁸ 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 decisionmakers⁴⁹ rather than to criteria that happen to be biostatistically convenient or time-honored. True value-based policymaking and healthcare delivery demand no less.

References

- Aakvik A, et al., 2005 Estimating treatment effects for discrete outcomes when responses to treatment vary: an application to Norwegian vocational rehabilitation programs. *J. Econom* 125, 15–51.
- Abadie A, 2002 Bootstrap tests for distributional treatment effects in instrumental variable models. *JASA* 97, 284–292.
- Abbring JH, Heckman JJ, Chapter 72 in 2007. *Econometric evaluation of social programs, part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and General equilibrium policy evaluation* In: Heckman JJ, Leamer EE (Eds.), *Handbook of Econometrics*, Vol. 6B Elsevier, Amsterdam.
- Abrevaya J, 2000 testing for a treatment effect in a heterogeneous population: a modified sign-test statistic and a leapfrog statistic. *J. Appl. Stat* 27, 679–687.
- Adams CP, Working Paper 2013 Median Survival, Hazard Ratios, and Bounding Drug Effectiveness. U.S. Federal Trade Commission.
- Allison RA, Foster JE, 2004 Measuring health inequality using qualitative data. *J. Health Econ* 23, 505–524. [PubMed: 15120468]
- Andrews DWK, 1997 A conditional kolmogorov test. *Econometrica* 65, 1097–1128.
- Angrist JD, 2004 Treatment effect heterogeneity in theory and practice. *Econ. J* 114, C52–C83.
- Athey S, Imbens GW, 2006 Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74, 431–497.
- Basu A, Thariani R, Working Paper 2016 Jointness Box (JB)-Area: A Novel Metric to Contemplate Potential Value of Individualized Care from Traditional Trials Data. Univ. of Washington.
- 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 Econ.* 16, 1133–1157. [PubMed: 17910109]
- Bitler MP, et al., 2006 What mean impacts miss: distributional effects of welfare reform experiments. *Am. Econ. Rev* 96, 988–1012.
- Boole G, 1854 *An Investigation of the Laws of Thought, of Which Are Founded the Mathematical Theories of Logic and Probabilities*. Walton and Maberly, London.
- Borah BJ, et al., 2011 Assessing the impact of high deductible health plans on health-care utilization and cost: a changes-in-changes approach. *Health Econ.* 20, 1025–1042. [PubMed: 21630375]

⁴⁸For instance, in the context of cost-effectiveness analysis it may be interesting to explore extensions of this paper's approaches to questions involving the "probability of cost-effectiveness," possibly using net-health-benefit or related outcome measures (Stinnett and Mullahy, 1998; Willan, 2001).

⁴⁹See Lynn et al. (2015), for a compelling discussion.

- Borghaei H, et al., 2015 Nivolumab versus docetaxel in advanced nonsquamous non-small-cell lung cancer. *NEJM* 373,1627–1639. [PubMed: 26412456]
- Cameron AC, et al., 2004 Modelling the differences in counted outcomes using bivariate copula models with application to mismeasured counts. *Econom. J* 7, 566–584.
- Carneiro P, et al., 2001 Removing the veil of ignorance in assessing the distributional impacts of social policies. *Swedish Econ. Policy Rev* 8, 273–301.
- Castagnoli E, 1984 Some remarks on stochastic dominance. *Revista di Matematica per le Scienze Economiche e Sociali* 7, 15–28.
- Chan TY, Hamilton BH, 2006 Learning, private information and the economic evaluation of randomized experiments. *J. Polit. Econ* 114, 997–1040.
- Coate S, 2000 An efficiency approach to the evaluation of policy changes. *Econ. J* 110, 437–455.
- Darling DA, 1957 The kolmogorov-smirnov, cramer-von mises tests. *Ann. Math. Stat* 28, 823–838.
- Denuit M, et al., 2002 Measuring the impact of dependence between claims occurrences. *Insur. Math. Econ* 30, 1–19.
- Fan Y, Park SS, 2010 Sharp bounds on the distribution of treatment effects and their statistical inference. *Econ. Theory* 26, 931–951.
- Fan Y, Park SS, 2012 Confidence intervals for the quantile of treatment effects in randomized experiments. *J. Econom* 167, 330–344.
- 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. *J. Econom* 197, 42–59.
- Firpo S, Ridder G, 2008 Bounds on Functionals of the Distribution of Treatment Effects. Univ. of Southern California, IEPR Working Paper 08.09.
- Frandsen BR, Lefgren LJ, NBER Working Paper 24616 2018. Partial Identification of the Distribution of Treatment Effects With an Application to the Knowledge Is Power Program (KIPP).
- Gastwirth JL, 1975 Statistical measures of earnings differentials. *Am. Stat* 29, 32–35.
- Gerber ER, Lewis JB, 2004 Beyond the median: voter preferences district heterogeneity, and political representation. *J. Polit. Econ* 112, 1364–1383.
- Hansen BE, 2017 *Econometrics (Online Textbook)*. University of Wisconsin-Madison, Dept. of Economics.
- Hauck WW, et al., 2000 Generalized treatment effects for clinical trials. *Stat. Med* 19, 887–899. [PubMed: 10750057]
- Heckman JJ, 2001 Accounting for heterogeneity, diversity and general equilibrium in evaluating social programmes. *Econ. J* 111, F654–F699.
- Heckman JJ, et al., 1997 Making the most out of programme evaluations and social experiments: accounting for heterogeneity in programme impacts. *Rev. Econ. Stud* 64, 487–535.
- Horwitz RI, 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. *J. Clin. Epidemiol* 49, 395–400. [PubMed: 8621989]
- Huang EJ, 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 GW, Manski CF, 2004 Confidence intervals for partially identified parameters. *Econometrica* 72, 1845–1857.
- Imbens GW, Wooldridge JM, 2009 Recent developments in the econometrics of program evaluation. *J. Econ. Lit* 47, 5–86.
- Jacob J, Lundin D, 2005 A median voter model of health insurance with Ex post moral hazard. *J. Health Econ* 24, 407–426. [PubMed: 15721052]
- Koenker R, Biliyas Y, 2001 Quantile regression for duration data: a reappraisal of the Pennsylvania reemployment bonus experiments. *Empir. Econ* 26, 199–220.
- Kravitz RL, et al., 2004 Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q.* 82, 661–687. [PubMed: 15595946]

- Langley RG, et al., 2014 Secukinumab in plaque psoriasis-results of two phase 3 trials. *NEJM* 371,326–338. [PubMed: 25007392]
- Lee M-J, 2000 Median treatment effects in randomized trials. *JRSS-B* 62, 595–604.
- Lee M-J, Kobayashi S, 2001 Proportional treatment effects for count response panel data: effects of binary exercise on health care demand. *Health Econ.* 10, 411–428. [PubMed: 11466803]
- Lee JD, et al., 2016 Extended-release naltrexone to prevent opioid relapse in criminal justice offenders. *NEJM* 374, 1232–1242. [PubMed: 27028913]
- Lynn J, et al., 2015 Value-based payments require valuing what matters to patients. *JAMA* 314, 1445–1446. [PubMed: 26378889]
- Makarov GD, 1982 Estimates for the distribution function of a sum of two random variables when the marginal distributions are fixed. *Theory Prob. Appl* 26, 803–806.
- Mann HB, Whitney DR, 1947 On a test of whether one of two random variables is stochastically larger than the other. *Ann. Math. Stat* 18, 50–60.
- Manski CF, 1988 *Analog Estimation Methods in Econometrics*. Chapman and Hall, London.
- Manski CF, 1996 Learning about treatment effects from experiments with random assignment of treatments. *J. Hum. Resour* 31, 709–733.
- Manski CF, 1997 Monotone treatment effect. *Econometrica* 65, 1311–1334.
- Manski CF, 2007 *Identification for Prediction and Decision*. Harvard University Press.
- Mullahy J, 2005 Individual results May vary, Barcelona (May 2005). Working Paper, Presented at the Conference on the Economics of Addiction and Health Inequality.
- Nissen SE, et al., 2016 Cardiovascular safety of celecoxib, naproxen, or ibuprofen for arthritis. *NEJM* 375, 2519–2529. [PubMed: 27959716]
- Pauly MV, et al., 1989 Positive political economy of medicare, past and future In: Pauly MV (Ed.), *Lessons from the First Twenty Years of Medicare-Research Implications for Public and Private Sector Policy*. University of Pennsylvania Press, Philadelphia.
- Ruschendorf L, 2004 Comparison of multivariate risks and positive dependence. *J. Appl. Probab* 41,391–406.
- Scharfstein DO, et al., 2004 On the construction of bounds in prospective studies with missing ordinal outcomes: application to the good behavior game trial. *Biometrics* 60, 154–164. [PubMed: 15032785]
- Stinnett AA, Mullahy J, 1998 Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis. *Med. Decis. Making* 18, S68–S80. [PubMed: 9566468]
- Tamer E, 2010 Partial identification in econometrics. *Annu. Rev. Econom* 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.
- Volpp KG, et al., 2009 A randomized, controlled trial of financial incentives for smoking cessation. *NEJM* 360, 699–709. [PubMed: 19213683]
- Willan AR, 2001 On the probability of cost-effectiveness using data from randomized clinical trials. *BMC Med. Res. Method* 1, 8.
- Williamson RC, Downs T, 1990 Probabilistic arithmetic. I. Numerical methods for calculating convolutions and dependency bounds. *Int. J. Approximate Reasoning* 4, 89–158.
- Willke RJ, et al., 2012 From concepts, theory, and evidence of heterogeneity of treatment effects to methodological approaches: a primer. *BMC Med. Res. Method* 12, 185.

Fig. 1. Survival Time Distributions: Nivolumab= $F_{niv}(t)$ versus Docetaxel= $F_{doc}(t)$ –Median Survival Times (left); Twelve-Month Survival Probabilities (right).

Fig. 2.

BFLB based on Marginals $F_j(y)$ from $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim BVN\left(\begin{bmatrix} 4.1 \\ 5.1 \end{bmatrix}, \begin{bmatrix} 1 & .5 \\ .5 & 1 \end{bmatrix}\right)$ —BFLB for Arbitrary $y = y'$ (left); Best BFLB at $y = Y_{01}$ (right).

Fig. 3. Numerical Examples: Computing Y_{jk} and $D_{jk} - F_0(y) = \mathcal{N}(0, 4)$, $F_1(y) = \mathcal{N}(.5, 1)$ (left); $F_0(y) = \text{Exp}(5)$, $F_1(y) = \text{Exp}(1)$ (right).

Fig. 4.

BFLB based on Marginals $F_j(y)$ from $\begin{bmatrix} y_0 \\ y_1 \end{bmatrix} \sim BVN\left(\begin{bmatrix} 4.1 \\ 5.1 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right) - \rho = .5$ (left); $\rho = -.5$ (right).

Fig. 5.
Illustration of Zero-Order Stochastic Dominance – $Y_{01} = [U_0, L_1]$, and $D_{01} = 1$, with.

Fig. 6. Spreading and Rectangularizing, $N(\mu, \sigma^2)$ Examples – $F_0(y) = N(0, 1)$, $F_1(y) = N(0, 4)$, $F_2(y) = N(0, 16)$ (left); $F_0(y) = N(0, 4)$, $F_1(y) = N(1, 0)$ (right).

Fig. 7.
Survival Time, Nivolumab versus Docetaxel —Sample Space for Continuous and Binary Outcomes.

Fig. 9.

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


```

. 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

```

Exhibit 1.

Using Stata's ksmirnov to Estimate D_{01} and D_{10} : $N=500$ Simulated Outcomes from $y_0 \sim N(0, 4)$ and $y_1 \sim N(.5, 1)$.

```

. by tx: sum srhs
-> tx = 0
      Variable |      Obs      Mean   Std. Dev.      Min      Max
-----+-----
      srhs |      442   3.538462   .8778532         1         5

-> tx = 1
      Variable |      Obs      Mean   Std. Dev.      Min      Max
-----+-----
      srhs |      436   3.598624   .8719896         1         5

. ksmirnov srhs, by(tx)

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

Smaller group      D      P-value
-----+-----
0:                  0.0336    0.610
1:                  0.0000    1.000
Combined K-S:      0.0336    0.966

Note: Ties exist in combined dataset;
      there are 5 unique values out of 878 observations.

. disp r(D_1)
.03357342

. disp r(D_2)
0

```

Exhibit 2.

Using Stata's ksmirnov to Estimate D_{01} and D_{10} : Volpp et al. (2009), SRHS Data.

Table 1

D_{01} and $Pr(y_1 = y_0)$ for Alternative Mean and Correlation Structures; $(y_0, y_1) \sim BVN(\mu_0, \mu_1; 1, 1, \rho)$.

		$Pr(y_1 = y_0) \text{ for } \rho =$					
$\mu_1 - \mu_0$	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 2

Binary Outcomes Bounds on $Pr(y_1 = y_0)$ — General Case (top panel); Nivolumab Example Probability Bounds (bottom panel).

		y_0		MarginalTotal
		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
Twelve-Month Survival		Docetaxel (q_{doc})		MarginalTotal
		Died	Survived	
Nivolumab	Died	π_{00}	π_{10}	.49
(q_{niv})	Survived	.12	π_{11}	.51
Marginal Total		.61	.39	1