NBER WORKING PAPER SERIES

DISCOVERING TREATMENT EFFECTIVENESS VIA MEDIAN TREATMENT EFFECTS— APPLICATIONS TO COVID-19 CLINICAL TRIALS

John Mullahy

Working Paper 27895
http://www.nber.org/papers/w27895

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2020

Thanks are owed to Chris Adams, Marguerite Burns, Domenico Depalo, Chuck Manski, Dan Millimet, Ciaran O'Neill, and Jon Skinner for helpful comments on an earlier draft. NICHD grant P2CHD047873 to the Center for Demography and Ecology and NIA grant P30AG017266 to the Center for Demography of Health and Aging, both at UW-Madison, provided logistical support. 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 They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Discovering Treatment Effectiveness via Median Treatment Effects—Applications to COVID-19 Clinical Trials
John Mullahy
NBER Working Paper No. 27895
October 2020
JEL No. C18,I10,I18

ABSTRACT

Comparing median outcomes to gauge treatment effectiveness is widespread practice in clinical and other investigations. While common, such difference-in-median characterizations of effectiveness are but one way to summarize how outcome distributions compare. This paper explores properties of median treatment effects as indicators of treatment effectiveness. The paper's main focus is on decisionmaking based on median treatment effects and it proceeds by considering two paths a decisionmaker might follow. Along one, decisions are based on point-identified differences in medians alongside partially identified median differences; along the other decisions are based on point-identified differences in medians in conjunction with other point-identified parameters. On both paths familiar difference-in-median measures play some role yet in both the traditional standards are augmented with information that will often be relevant in assessing treatments' effectiveness. Implementing both approaches is shown to be straightforward. In addition to its analytical results the paper considers several policy contexts in which such considerations arise. While the paper is framed by recently reported findings on treatments for COVID-19 and uses several such studies to explore empirically some properties of median-treatment-effect measures of effectiveness, its results should be broadly applicable.

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

In a study published online on March 18, 2020, comparing lopinavir–ritonavir treatment for COVID-19 with standard care, Cao et al., 2020, report:

> Patients assigned to lopinavir–ritonavir did not have a time to clinical improvement different from that of patients assigned to standard care alone in the intention-to-treat population (median, 16 days vs. 16 days; hazard ratio for clinical improvement, 1.31; 95% confidence interval [CI], 0.95 to 1.85; P=0.09)…

On April 29, 2020, the U.S. National Institute on Allergy and Infectious Diseases (NIAID, 2020a) announced in a news release summarizing a separate study:

> Hospitalized patients with advanced COVID-19 and lung involvement who received remdesivir recovered faster than similar patients who received placebo…. Specifically, the median time to recovery was 11 days for patients treated with remdesivir compared with 15 days for those who received placebo.[1]

On the basis of that study the U.S. Food and Drug Administration (FDA) two days later issued an Emergency Use Authorization for remdesivir (U.S. FDA, 2020a), noting:

> While there is limited information known about the safety and effectiveness of using remdesivir to treat people in the hospital with COVID-19, the investigational drug was shown in a clinical trial to shorten the time to recovery in some patients.

This study, the Adaptive Covid-19 Treatment Trial or ACTT Study (Beigel et al., 2020), was subsequently published online on May 22, 2020.

In a third study, published online May 8, 2020, Hung et al., 2020, report on a trial comparing combination therapy for COVID-19 with lopinavir–ritonavir alone. The authors note:

> The combination group had a significantly shorter median time from start of study treatment to negative nasopharyngeal swab (7 days [IQR 5–11]) than the control group (12 days [8–15]; hazard ratio 4·37 [95% CI 1·86–10·24], p=0·0010).

Apart from their focus on treating COVID-19, a common feature of these three studies[2] is that they summarize their respective primary endpoints as medians of each arm's time-to-

[1] Mass media wasted no time reporting these findings. An April 29 headline in the *Washington Post* read: "Gilead's remdesivir improves recovery time of coronavirus patients in NIH trial."

[2] The Cao et al., 2020, Beigel et al., 2020, and Hung et al., 2020, studies will be referenced henceforth simply as Cao, Beigel, and Hung. They will be revisited later in the paper.

event (TTE) distribution, and then judge the treatments' efficacy or effectiveness by the difference between those medians. Such median-based comparisons are common in the clinical literature, so common perhaps that they and the information they convey to decisionmakers may be taken for granted.

Much is at stake on the manner in which the effectiveness of candidate COVID-19 treatments and vaccines is assessed, using clinical trial data or otherwise. Loading this burden onto two parameters—median outcomes under treatment and comparator—is a high-stakes gamble. Consider for instance figure 1 whose top panel depicts by the two indicated points the results reported in the NIAID news release (NIAID, 2020a); until the publication of the Beigel study over three weeks later this was the only information about the study's effectiveness known by the public. (The bottom panel of figure 1 reproduces the results reported by Beigel.) Is this information alone sufficient for a decisionmaker to conclude that remdesivir is effective since it will "shorten the time to recovery in some patients"? Analogously does the finding that median outcomes do not differ in the Cao study mean that decisionmakers should not prefer one treatment over the other (see figure 2)? Or might in both instances regulators, clinicians, patients, and others wish to know more about the distributions of outcomes before arriving at such conclusions?

[figures 1 and 2 about here]

The perhaps-obvious points are that there are many ways subject-level outcome data can be summarized or aggregated across a sample[3] and that these alternative approaches may not yield the same conclusion about treatments' effectiveness, nor should they. Understanding the nature, merits, and shortcomings of the methods chosen as well as how those methods stack up against alternatives should thus be instructive for COVID-19 decisionmaking and beyond.

Numerous comparisons of candidate COVID-19 vaccines and treatments will be contemplated, studied, and reported. It is thus timely to review and assess the properties and decisionmaking value of treatment effect (TE) measures defined by distributions' medians—like those featuring in the studies described above—as summaries of treatments' effectiveness: In a nutshell, what is learned about treatments' effectiveness from consideration and comparison of two distributions' medians?

The paper argues that a reasonable answer to this question is: "Probably something but often little." To learn more about treatments' effectiveness the paper suggests two paths decisionmakers might follow. Along path 1 decisions are based on *point-identified* differences in median outcomes—as is customary—in conjunction with medians of the distribution of differences in outcomes, i.e. the *treatment effect distribution* (section 2). The latter medians are

[3] See Zarin et al., 2011, who discuss data aggregation in the context of clinical-trial reporting.

typically *partially identified* since the distribution of differences may not be observable.[4] Along path 2 decisions are again based on point-identified differences in medians in conjunction with additional point-identified parameters whose nature will be discussed in section 5. On either path it is proposed that the basis of decisions is enhanced by considering features of distributions in addition to differences in median outcomes. In both cases conventional differences in medians play some role so that decisionmakers accustomed to relying on such criteria should find nothing particularly foreign in the proposed strategies.

Given the urgency of discovering effective treatments and vaccines for COVID-19 it is timely to consider measures of effectiveness that align with what is important to decisionmakers and those affected by their decisions. What is proposed here may be helpful to this end, particularly since the paper strives to offer intuitive and easily implemented strategies.

Section 2 offers definitions and other preliminaries. Section 3 discusses basic elements of decisions based on median treatment effects. Sections 4 and 5 present the key issues and results for paths 1 and 2 as described above. Section 6 illustrates the applicability of these ideas using three COVID-19 clinical studies. Section 7 summarizes. While the discussion throughout is motivated and framed by attempts to discover the effectiveness of COVID-19 treatments, the issues addressed should be broadly applicable in contexts where considerations of median treatment effects arise.[5]

2. Definitions and Other Preliminaries

Consider two subject-specific outcomes, y_0 and y_1, which might be health status under treatment 0 (T_0) and treatment 1 (T_1) and which may or may not—depending on particulars—be potential outcomes where only one of y_0 or y_1 is observed. Suppose throughout that each y_j is non-negative and has either continuous or discrete (integer) measurement. Whether or not both y_j are observed for each subject, define the subject-level treatment effect as $\Delta = \mathrm{y}_1 - \mathrm{y}_0$.[6] Unless noted otherwise *smaller* values of y will correspond to better health outcomes.

Let $\mathrm{Pr}\left(\ldots\right)$ denote a probability mass or density and $\mathrm{F}\left(\ldots\right)$ its corresponding cumulative.

[4] Partial identification strategies have recently been used to understand aspects of COVID-19 prevalence and treatment decisions (see Manski, 2020b, and Manski and Molinari, 2020).

[5] To streamline the paper many technical details and in-depth discussions appear in appendixes and footnotes.

[6] In this paper the term "treatment effect" when unqualified has the precise meaning indicated here. The expressions "median treatment effect," "quantile treatment effect," "treatment effectiveness," and others are generic; their meaning depends on particular context.

Define $F_j(y) = \Pr(y_j \leq y)$ for $j \in \{0,1\}$.[7] $F_j(y)$ will sometimes be abbreviated as F_j. Each y_j has support Y_j, which may be continuous or discrete. $Y = Y_0 \cup Y_1$ is the common support of y_0 and y_1; often $Y_0 = Y_1$. Define the α–quantile of $F_j(y)$ as[8]

$$q_j(\alpha) = \inf_{y \in Y}\left\{y \middle| F_j(y) \geq \alpha\right\} \text{ for } \alpha \in (0,1). \tag{1}$$

m_j is shorthand used henceforth for the marginal medians $q_j(.5)$.[9] Define the interquartile range (IQR) of each F_j as

$$IQR_j = \left\{q_j(.25),\ q_j(.75)\right\}. \tag{2}$$

Note that IQR_j is defined as a two-element set not as an interval as perhaps more conventional.

It is assumed at this point that the F_j are point identified over relevant subsets of Y; requiring identification to be over all $y \in Y$ or only over subsets will be considered in context.[10] Whether identification results from a randomized trial or some other method is unimportant; indeed each F_j can be learned from a different data source.

Define the difference-in-y-probabilities treatment effect as

[7] This paper will focus only on issues of identification leaving considerations of inference involving sampling variation for future study. (Goldman and Kaplan, 2018, offer innovative perspectives on hypothesis testing in contexts like those considered here.) Distinguishing population from sample parameters as might be typical is thus of little consequence. The N-observation sample data can be treated as if they are a finite population of size N. This is noted here so that notation can be streamlined, e.g. one needn't distinguish population distributions $F_j(y)$ from sample or empirical distributions $F_{j,N}(y)$. Thus interpreted a randomized trial splits the population of size 2N into subpopulations of size M and 2N–M and administers accordingly T_0 and T_1; the resulting F_j characterize the counterfactual full-population distribution of outcomes under each treatment.

[8] "min" suffices for most cases covered here but "inf" is technically appropriate (see Hansen, 2020, section 11.13).

[9] See appendix A for discussion of medians' computation and measurement.

[10] In some cases (e.g. right censoring) point identification need hold only up to some value.

$$\Delta F(y) = F_1(y) - F_0(y), \tag{3}$$

a familiar endpoint in TTE and other studies (e.g. difference in 12-month survival probabilities). Define the difference-in-α-quantiles treatment effect as

$$\Delta q(\alpha) = q_1(\alpha) - q_0(\alpha). \tag{4}$$

A special case of (4) is the difference-in-medians treatment effect

$$\Delta m \ = \ m_1 - m_0 \ = \ \Delta q(.5). \tag{5}$$

Δm is a prominent measure used to compare treatments' effectiveness in clinical studies, particularly albeit not exclusively for TTE outcomes.[11] Δm is point identified under the assumption that both $F_j(y)$ are point identified at least for $\left\{y \middle| F_j(y) \le .5\right\}$.

The treatment effect distribution is $F(\Delta)$, where the distribution of Δ derives from the joint distribution $\Pr(y_0, y_1)$. In potential-outcomes and other contexts $\Pr(y_0, y_1)$ is not point identified so that $F(\Delta)$ is consequently not point identified.[12] When $F(\Delta)$ is not point identified, that it can generally be informatively partially identified is an important consideration in what follows.

Quantiles of the treatment effect distribution are

$$q(\alpha)\Delta = \min\left\{\Delta \middle| F(\Delta) \ge \alpha\right\}. \tag{6}$$

with the median-difference treatment effect defined as

[11] Other characterizations of median TEs based on ratios or percentage changes instead of differences have been considered; see Lee and Kobayashi, 2001, and Rogawski et al., 2017.

[12] In some instances $\Pr(y_0, y_1)$ is point identified even though the corresponding $F(\Delta)$ would not readily admit interpretation as a distribution of counterfactual outcome differences. Examples include pre-post and crossover designs. See Fan et al., 2003, for an example in ophthalmology research where individuals' left eyes and right eyes are randomized to serve as treatment and control "subjects."

$$\mathrm{m}\Delta \;=\; \mathrm{q}\left(.5\right)\Delta \;=\; \mathrm{med}\left(\mathrm{F}\left(\Delta\right)\right) \;=\; \min\left\{\Delta \middle| \mathrm{F}\left(\Delta\right) \geq .5\right\}. \tag{7}$$

With a few exceptions this paper's discussion focuses on $\Delta\mathrm{m}$ and $\mathrm{m}\Delta$ rather than general quantile treatment effects, this owing to the prominence[13] of medians in empirical clinical and health research; most results generalize readily to other quantiles. Manski (2007, chapter 7) notes the inequality

$$\mathrm{m}\Delta \neq \Delta\mathrm{m}\,, \tag{8}$$

and proceeds to suggest that "a researcher reporting a median treatment effect must be careful to state whether the object of interest is the left- or right-hand side [of (8)]."[14]

As will be seen in section 4 understanding and computing bounds on $\mathrm{m}\Delta$ is facilitated by reference to the inequality probabilities (IPs) $\Pr\left(y_1 > y_0\right) = \Pr\left(\Delta > 0\right)$. Like $\mathrm{m}\Delta$ IPs are generally not point identified but are partially identified using information on the F_j. Define

$$\mathrm{D}_{jk} = \max_{y \in Y}\left\{\mathrm{F}_j\left(y\right) - \mathrm{F}_k\left(y\right), 0\right\}, \text{ for } j,k \in \left\{0,1\right\}. \tag{9}$$

The D_{jk} are the largest vertical differences between $\mathrm{F}_j\left(y\right)$ and $\mathrm{F}_k\left(y\right)$ over the common support Y.[15] These measures, known commonly as Kolmogorov's D-statistics, are depicted in figure 3. Note that at most one of the D_{jk} can exceed .5, a result that will prove useful later.

[figure 3 about here]

[13] A pubmed.gov search conducted on September 8, 2020 yielded:

— 90,192 hits using the search string: "median time to" OR "median survival" OR "median progression-free" OR "median length" OR "median duration"

— 185,333 hits using the search string: "median difference" OR "median change" OR "median percentage change" OR "median percent change" OR "median relative change" OR "difference in median" OR "difference between median"

— 253,125 hits using the union of these search strings

[14] $\Delta\mathrm{m}$ and $\mathrm{m}\Delta$ are what Manski, 1997, refers to generically and respectively as "ΔD" and "DΔ" treatment effects; thus the $\Delta\mathrm{m}$ and $\mathrm{m}\Delta$ notation used here.

[15] The corresponding y-ordinates that define the D_{jk} may not be unique.

3. How $\Delta\mathrm{m}$ May Inform Decisions

There are many standards by which two or more outcome distributions might be compared: moments, quantiles, dominance, etc. On what principles might $\Delta\mathrm{m}$ be advocated to inform decisions? While $\Delta\mathrm{m}$ may have reasonable statistical properties this is a different matter from it possessing well grounded conceptual properties. For instance comparison of the medians (or other quantiles) of observed marginal distributions are uninformative about the distribution of gains and losses arising from an imagined change from T_0 to T_1, i.e. $\mathrm{F}\left(\Delta\right)$.

Imbens and Wooldridge, 2009, raise two issues to support their claim that $\Delta\mathrm{m}$ may often be of greater interest than $\mathrm{m}\Delta$ or other features of $\mathrm{F}\left(\Delta\right)$. First, they assert that it is natural for decisionmakers to compare policies via differences between outcome distributions, which differences "can often be summarized by differences in the quantiles." Second, they note that $\mathrm{F}\left(\Delta\right)$ and associated parameters like $\mathrm{m}\Delta$ cannot generally be point identified. Each claim has some merit.

Yet two counterclaims might be advanced to support consideration of $\mathrm{m}\Delta$. The first is statistical: While $\mathrm{m}\Delta$ cannot generally be point identified it can generally be partially identified informatively, and in a subset of such cases can be sign identified (section 4 and appendix C). The second is conceptual: While decisionmakers may assess treatments' effectiveness by comparison of resulting marginal outcome distributions, they may also care about policies' distributional consequences, e.g. the fractions of the population that benefit or suffer from policy change *and* how much they benefit or suffer. The latter concerns are not informed by examination of the marginal distributions but require consideration of the joint distribution $\Pr\left(\mathrm{y}_0,\mathrm{y}_1\right)$ or, specifically, $\mathrm{F}\left(\Delta\right)$. Heterogeneous response to policy in direction and magnitude may be important considerations: As articulated nicely by Koenker and Bilias, 2000, "treatment may make otherwise weak subjects especially robust, and turn the strong to jello."

Decisionmakers may or may not feel that the basis of their decisions is enhanced by information on $\mathrm{F}\left(\Delta\right)$ or $\mathrm{m}\Delta$. Either way a revealed preference argument suggests that $\Delta\mathrm{m}$ measures of treatments' effectiveness provide them with *at least some* useful information about their choices (see footnote 13). Ease of computation, applicability with right-censored data, and parsimony in summarizing data are three plausible reasons for such popularity. The following two sections discuss decisionmaking that does (path 1; section 4) and does not (path 2; section 5) admit roles for partially identified parameters like $\mathrm{m}\Delta$ working in conjunction with $\Delta\mathrm{m}$. In both instances it is suggested that the basis of decisions may be enhanced by considering parameters beyond $\Delta\mathrm{m}$. To implement the strategies proposed along either path analysts need appeal only to information contained in F_0 and F_1; how that information gets used is quite different along the two paths, however.[16]

Finally, while the analytical results reported in sections 4 and 5 may be of interest per se, considering how they may be applicable in real-world decision and policy settings is appropriate. Rather than disrupting the flow of the paper here with a lengthy discussion, appendix E describes several policy or regulatory contexts wherein considerations of median or more general quantile TEs are or at least arguably should be prominent.[17] To anticipate that discussion it is worth noting briefly one example of regulatory language that naturally motivates such considerations. The regulatory language governing FDA's determination of effectiveness of biological products is:

> Effectiveness means a reasonable expectation that, in a *significant proportion of the target population*, the pharmacological or other effect of the biological product, when used under adequate directions, for use and warnings against unsafe use, will serve a *clinically significant function* in the diagnosis, cure, mitigation, treatment, or prevention of disease in man. [21 CFR 601.25(d)(2)] (emphasis added)

Note that this standard entails considerations of both population quantiles ("significant proportion") and treatment effect magnitudes ("clinically significant function"). Appendix E suggests statistical formalization of this language and suggests further how the discussion in sections 4 and 5 may usefully address such regulatory questions.

4. Path 1: Assessing Treatments' Effectiveness via $\Delta\mathrm{m}$ and Partially Identified $\mathrm{m}\Delta$

The essence of decision problems in this context is captured by a simple picture. Figure 4 depicts some population's discrete $\Pr\left(y_0,y_1\right)$ that puts probability mass 1/3 on three $\left(y_0,y_1\right)$

[16] Regulatory language suggests that parameters like medians on their own may inadequately describe outcomes. In its guidance for product labeling, the FDA (FDA, 2006) notes:

> When time-to-event endpoints (e.g., mortality) are used, median or mean survival alone is not usually an adequate descriptor. Survival curves (or event-free survival curves) and hazard ratios are often effective ways to display such data. Data can also be summarized at specific times (e.g., prevalence at 3, 6, 9, 12 months) or at specific event frequency (e.g., time to 25 percent, 50 percent, and 75 percent prevalence of events).

[17] While it is straightforward to imagine how understanding IPs may be valuable in informing certain decisions (see Mullahy, 2018a) it is maybe less obvious how knowledge of, or knowledge of bounds on, $\mathrm{m}\Delta$ is informative. While applicability to practical questions may be more nuanced there are nonetheless real-world settings where knowledge of $\mathrm{m}\Delta$ —or more generally of $\mathrm{q}\left(\alpha\right)\Delta$ —should be decision-relevant as will be suggested in appendix E.

values, $\{(3,8),\ (15,11),\ (17,21)\}$. Imagine $\Pr(y_0, y_1)$ is known. There is nothing particularly peculiar about the depicted probability structure (Pearson correlation .77, rank correlation 1.0). Yet $\Delta m = -4 < 0$ and $m\Delta = 4 > 0$. A decisionmaker knowing nothing more than the pictured information must select T_0 or T_1. Which should be chosen? (Which would *you* choose?)

[figure 4 about here]

In general one would appeal to the decisionmaker's utility or loss function to answer this question. But this is not the main issue here: Important here is whether the assumed knowledge of $m\Delta$ might influence even to some degree a decisionmaker's attitudes about the relative merits of T_0 and T_1 given that Δm is presumed known. Would such knowledge be even partially influential then the obstacle to decisionmaking based on $m\Delta$ in settings where $\Pr(y_0, y_1)$ is not known is $m\Delta$'s point identification not, as Imbens and Wooldridge hint, irrelevance of $m\Delta$ *per se*. In these cases partial identification may support such decisionmaking.

Partial Identification of and Bounds on mΔ

Since $m\Delta$ is generally partially but not point identified, a decisionmaker whose choice would depend to at least some degree on knowing it must accept that knowing a range of possible values is the best to be hoped for. Whether the knowable range of its possible values suffices for decisionmaking is context dependent.

A parameter θ is *partially identified* when it is known to be in the closed, half-open, or open bounds interval $b(\theta)$. Define the bounds set $B(\theta) = \{L(\theta), U(\theta)\}$ where $L(\theta) = \inf(b(\theta))$ and $U(\theta) = \sup(b(\theta))$.[18] $L(\theta)$ and $U(\theta)$ are valid bounds since $\theta \in b(\theta)$ but may not be the tightest bounds. The bounds $B(\theta)$ are considered sharp when $L(\theta)$ and $U(\theta)$ are the largest and smallest values learned from the data that are consistent with knowledge of θ. These tightest bounds and corresponding bounds interval are denoted $B^*(\theta) = \{L^*(\theta), U^*(\theta)\}$ and $b^*(\theta)$. A partially identified parameter θ is *sign identified* when $b^*(\theta)$ excludes zero.[19]

[18] Using inf/sup instead of min/max covers cases where $b(\theta)$ may be half or fully open.

[19] The paper does not consider monotone treatment response (Manski, 1997) or related considerations (Lee, 2000). See appendix B for a brief discussion.

Strategies for computing $B^*(m\Delta)$ are discussed in appendix C.[20]

Bounds with Partially Observed Outcome Data

Because they are relevant in the case studies appearing in section 6, a few considerations involved in computing bounds on $m\Delta$ under two forms of partially observed outcomes—right-censored data and IQR data—are sketched here and discussed more fully in appendix C.

Right-censoring of outcome data is common in TTE studies. Given noninformative right-censoring at y_c (outcomes unobserved if they exceed y_c) three scenarios can be considered:

a. both $F_0(y_c)$ and $F_1(y_c)$ are greater than .5 (both m_0 and m_1 are point identified)

b. one of $F_0(y_c)$ and $F_1(y_c)$ is greater than .5, the other is not (one of m_0 or m_1 is point identified, the other is not)

c. both $F_0(y_c)$ and $F_1(y_c)$ are less than than .5 (neither m_0 nor m_1 is point identified)

These cases, depicted in the top panel of figure 5, have different implications for the identification of Δm and $m\Delta$:

a. Δm is point identified; $m\Delta$ is partially identified with finite bounds

b. both Δm and $m\Delta$ are partially identified with one finite bound

c. neither Δm nor $m\Delta$ is informatively bounded (no finite bounds)

None of these results is surprising. Yet even under scenario (a) it is noteworthy that while the particular right-censoring threshold y_c is of no consequence for point identifying Δm—i.e. m_0, m_1, and Δm are invariant with respect to any y_c that exceeds both m_0 and m_1—this is not so for computation of $m\Delta$ bounds. The extent of right-censoring affects the amount of

[20] A Stata program, `medte`, computes $B^*(m\Delta)$ with uncensored or right-censored outcomes. A zip file containing the do-file defining the `medte` program is available here. `medte` reports Δm, $B^*(m\Delta)$, the IP bounds defined in appendix C ((C.1) and (C.2)), and the approximate central aperture measures defined in section 5. The zip file also contains a `readme` file and the two datasets used for the analyses reported in those tables. (It may be of interest to note that $B^*(m\Delta) = \{-6, 6\}$ for the data depicted in figure 4, with this $B^*(m\Delta)$ computed using the permutation approach described in appendix C.)

information used to compute the $m\Delta$ bounds. Reducing right-censoring—i.e. increasing y_c—can never widen $m\Delta$ bounds but may narrow them. See appendix C for discussion.[21]

[figure 5 about here]

The second form of partially observed data of interest here is interquartile-range data. IQRs are often reported alongside medians, sometimes in settings where data are also right censored. For instance IQRs are reported for the primary endpoints in the Cao study that reported entire data (their figure 2 and table 3) as well as in the Hung study that reported the two arms' quartiles only (their table 2; figure 6 here). Indeed it is not unusual to find the only outcome information reported to be that on each outcome's three sample quartiles.

[figure 6 about here]

Recall the definition of IQR_j in (2). It turns out that $m\Delta$ can be bounded informatively using only the IQR_j. When outcomes are continuously distributed and the $F_j(y)$ are everywhere-increasing in y the IQR-based bounds are

$$B_{\mathrm{IQR}}(m\Delta) = \left\{ \left(q_1(.25) - q_0(.75)\right),\ \left(q_1(.75) - q_0(.25)\right) \right\}, \qquad (10)$$

where the $q_j(.25)$ and $q_j(.75)$ are the observed data. When outcomes are measured as integers the corresponding IQR-based bounds are

$$B_{\mathrm{IQR}}(m\Delta) = \left\{ \left(q_1(.25) - q_0(.75) - 1\right),\ \left(q_1(.75) - q_0(.25) + 1\right) \right\}. \qquad (11)$$

Appendix C provides details.

Relationships between Δm and $B^*(m\Delta)$

While Δm and $m\Delta$ describe different quantities, understanding relationships between

[21] For ethical and statistical reasons stopping clinical trials "early for benefit" is controversial (Pocock, 1992). Ethical considerations aside, if the only endpoints of interest are the median TTE outcomes m_0 and m_1 then there is no further statistical benefit to be realized by waiting to amass more data once the events for half the subjects in each treatment arm have occurred.

them may enhance understanding of treatment effectiveness when knowledge of Δm alone inadequately informs choice. If so, whether knowing $B^*(m\Delta)$ instead of $m\Delta$ itself will satisfy a decisionmaker will depend on context.

This subsection presents results on relationships between Δm and $B^*(m\Delta)$. Knowledge of Δm turns out to be partially informative about $B^*(m\Delta)$ and vice versa; since both Δm and $m\Delta$ derive from $\Pr(y_0, y_1)$ this is not surprising. The results are explained in appendix D.

Result 1 (R1): Suppose $\Delta m < 0$. Then:

a. $0 \le D_{01} \le .5$

b. $0 < D_{10} \le 1$

c. $L^*(m\Delta) \le 0$

R1(c) means that knowing $\text{sign}(\Delta m)$ suffices to sign $L^*(m\Delta)$ when $\Delta m < 0$. An interpretation is that the data cannot reject $\text{sign}(m\Delta) = \text{sign}(\Delta m)$. Alternatively, if $m\Delta$ is sign identified—with $\Delta m < 0$ this means $U^*(m\Delta) < 0$—then its sign must be the same as Δm's, i.e. $m\Delta$ and Δm cannot be sign identified in opposite directions. This result is generalized with Result 2:

Result 2 (R2): $L^*(m\Delta) \le \Delta m \le U^*(m\Delta)$

Result 3 (R3): Suppose $D_{10} > .5$. Then:

a. $\Pr(y_1 \le y_0) > .5$

b. $\Delta m < 0$

c. $m\Delta < 0$

Finding $D_{jk} > .5$ is powerful, sufficing to sign identify $m\Delta$ and offer a strong statement about $\Pr(y_j < y_k)$.[22] Indeed, $m\Delta$ is sign identified if and only if one of the D_{jk} exceeds .5.

[22] With smaller values of y corresponding to better health it might be reasonable in such cases to consider T_j to be a *breakthrough* treatment. See appendix E.

Figure 7 integrates these ideas. Depicted are two different joint probability structures. $\Pr_{(a)}\left(y_0,y_1\right)$, where y_0 and y_1 are proximate, has $\Delta m = 2$, $m\Delta = 2$, $B^*\left(m\Delta\right) = \left\{-1,4\right\}$, $D_{01} = .4$, and $B^*\left(\Pr\left(y_1 > y_0\right)\right) = \left\{.4,1\right\}$. Conversely $\Pr_{(b)}\left(y_0,y_1\right)$, where y_0 and y_1 are distant, has $\Delta m = 6$, $m\Delta = 6$, $B^*\left(m\Delta\right) = \left\{3,8\right\}$, $D_{01} = 1$ (zero-order dominance; see Castagnoli, 1984), and $B^*\left(\Pr\left(y_1 > y_0\right)\right) = \left\{1,1\right\}$. Roughly speaking, joint distributions placing probability mass further northwest of the $y_0 = y_1$ locus yield stronger identifying information along with larger median treatment effects; in such instances the marginals F_0 and F_1 are "farther apart" in an important sense explored further in section 5.

[figure 7 about here]

Informing Decisions with Δm and Partially Identified $m\Delta$

Imagine a decisionmaker for whom both Δm and $m\Delta$ are decision-relevant parameters. Consider a stylized policy or decision function

$$p_\beta = \beta\Delta m + \left(1-\beta\right)m\Delta\,, \tag{12}$$

where $\beta \in \left[0,1\right]$ is presumed known and reflects the relative importance to the decisionmaker of Δm and $m\Delta$. With smaller y corresponding to better health, positive and negative values of p_β favor T_0 or T_1, respectively. In general p_β is only partially identified, with

$$B^*\left(p_\beta\right) = \left\{\beta\Delta m + \left(1-\beta\right)L^*\left(m\Delta\right),\ \beta\Delta m + \left(1-\beta\right)U^*\left(m\Delta\right)\right\}. \tag{13}$$

However sign identification of p_β would be valuable for decisionmaking. Consider three cases:

a. Suppose $\Delta m = 0$. Then for any $\beta \in \left[0,1\right]$ p_β would be sign-identified only if $m\Delta$ is sign-identified. But from R2 $m\Delta$ is not sign identified when $\Delta m = 0$.

b. Suppose $\Delta m \neq 0$ and $m\Delta$ is sign-identified. Then p_β is sign-identified for any $\beta \in \left[0,1\right]$ since from R2 the signs of Δm and $m\Delta$ coincide when $m\Delta$ is sign-identified.

c. Suppose $\Delta m \neq 0$ and $m\Delta$ is not sign-identified. Then p_β is sign-identified for some values of β. E.g. suppose $\Delta m < 0$ and $U^*\left(m\Delta\right) > 0$. Then p_β is sign identified for $\beta \in \left(U^*\left(m\Delta\right) \middle/ \left(U^*\left(m\Delta\right) - \Delta m\right),\ 1\right]$. Knowing this range may be instructive.

Decisionmakers may be positioned to make coherent choices even when they know only (13) and not (12). If p_β is sign identified then the decision at hand is well supported and there can be no regret. If not then it is appropriate to admit reservations about whatever decision is made since it might be regretted and since certitude about its merits is not credible (Manski, 2020a). Either way to the extent that $m\Delta$ is at least minimally important to a decisionmaker its partial identification needn't hinder information on its bounds supporting or cautioning choices that must be made.[23,24]

5. Path 2: Assessing Treatments' Effectiveness via Δm and Other Point-Identified Parameters

Recall from section 3 that Imbens and Wooldridge, 2009, have argued that typical policy choices will appropriately be based on consideration of quantiles of F_0 and F_1, writing: "choice should be governed by preferences of the policymaker over [F_0 and F_1] (which can often be summarized by differences in the quantiles)." One might thus imagine a decision function in which various quantiles hold different importance for a decisionmaker:

$$p_\omega\left(J\right) = \sum\nolimits_{\alpha \in J} \omega_\alpha \times \Delta q\left(\alpha\right), \tag{14}$$

where J is the set of quantiles of interest and ω_α are importance weights. Should all the $\Delta q\left(\alpha\right)$ in (14) be point identified then Imbens and Wooldridge's claim suggests that the sign of p_ω will determine the policymaker's choice; its magnitude might also be of interest in some contexts. Basing choices on Δm alone is a particular version of (14) with $J = \left\{.5\right\}$ and $\omega_{.5} = 1$.

[23] See Manski, 2018a,b, for important insights on decisionmaking in clinical and related contexts.

[24] When β is unknown, a more modest yet quite pragmatic suggestion—one that still acknowledges both the decision-relevance of $m\Delta$ as well as its partial identification—is simply to routinely report $B^*\left(m\Delta\right)$ alongside point estimates of Δm when a study's results are tabulated. For instance, one might report $\boxed{\texttt{-4 \{-15,10\}}}$ in a tabular summary of the results reported in table 1 (being careful to note that this is not a conventional confidence interval).

Still in the spirit of Imbens and Wooldridge's suggestion of comparing marginal distributions, policy choices might alternatively depend on a set of probability treatment effects,

$$p_\pi\left(K\right)=\sum_{k\in K}\pi_k\times\Delta F\left(y_k\right), \tag{15}$$

where K is the set of y-values of interest (e.g. 6-, 12-, and 24-month survival) and π_k are importance weights. Again the sign and perhaps the magnitude of p_π may be decisive for treatment choice. Familiar single-outcome criteria, e.g. differences in 12-month survival probabilities, are specific versions of (15).

Of course (14) and (15) might both be of interest in which case a general choice function

$$p_{both}=v\left(p_\omega\left(J\right),p_\pi\left(K\right)\right) \tag{16}$$

may underpin decisions.

Aperture and Weighted-Δm Measures of Treatment Effectiveness

Point-identified parameters beyond Δm may provide decisionmakers more-nuanced perspectives on treatment effectiveness than those offered by Δm alone. The search is for broadly applicable parameters that provide such perspectives while also being straightforward to implement in empirical studies. The perspective here is that of a decisionmaker who is not prepared to entirely abandon Δm but is willing to temper decisions by additional criteria. One approach, seemingly novel, is suggested here.[25]

As used in the preceding paragraph "treatment effectiveness" is intended to describe a broad sense of the extent to which one treatment makes people better off than another. In the following discussion greater treatment effectiveness means some amalgam of larger α–quantile treatment effects $\Delta q\left(\alpha\right)$ over some $\alpha\in\left(0,1\right)$ *and* larger y–probability treatment effects $\Delta F\left(y\right)$ over some $y\in Y$. In essence, greater treatment effectiveness means F_0 and F_1 are "farther apart" in some directions. The practical challenge is how to summarize this notion with a single parameter, i.e. how to define decision-informative yet simple and parsimonious representations of "farther apart" and "some direction."

Define the area between F_0 and F_1 over some interval $y\in\left(r,s\right)$ as the *local aperture* of

[25] The approach suggested here is admittedly speculative; its merits—should it appear to have any—would need to be vetted more thoroughly than the scope of this paper permits.

F_0 and F_1 and denote this area $A(r,s)$.[26] "Aperture" signifies that $A(r,s)$ measures the "opening" between F_0 and F_1 over the interval (r,s). Intuitively, the larger is $A(r,s)$ the more effective is T_j relative to T_k in the vicinity of (r,s) as F_0 and F_1 are locally "farther apart."[27]

For any point (y,α) in the region bounded by F_0 and F_1 over $y \in (r,s)$ define two approximations to $A(r,s)$ as

$$\begin{aligned} a_1(y,\alpha) &= \left(F_0(y) - F_1(y)\right) \times \left(F_1^{-1}(\alpha) - F_0^{-1}(\alpha)\right) \\ &= -\Delta F(y) \times \Delta q(\alpha) \end{aligned}, \tag{17}$$

and

$$\begin{aligned} a_2(\alpha) &= .5 \times \left(F_0\left(F_1^{-1}(\alpha)\right) - F_1\left(F_0^{-1}(\alpha)\right)\right) \times \left(F_1^{-1}(\alpha) - F_0^{-1}(\alpha)\right) \\ &= .5 \times \left(F_0\left(F_1^{-1}(\alpha)\right) - F_1\left(F_0^{-1}(\alpha)\right)\right) \times \Delta q(\alpha). \end{aligned} \tag{18}$$

Using the idea of aperture and its approximations to devise broadly applicable and point-identifiable measures of treatment effectiveness, it is natural to consider aperture at a "central" location in the data, or *central aperture*. Letting $(r,s) = (m_j, m_k)$ (i.e. r is the smaller of m_0 or m_1), define central aperture as $A(m_j, m_k)$. From (17) and (18) it follows that two easily computed approximations to central aperture are:

$$\begin{aligned} a_1(m^*,.5) &= -\Delta F(m^*) \times \Delta m \\ &= \delta_1 \times \Delta m \end{aligned} \tag{19}$$

and

$$\begin{aligned} a_2(.5) &= .5 \times \left(F_0\left(F_1^{-1}(.5)\right) - F_1\left(F_0^{-1}(.5)\right)\right) \times \Delta m \\ &= .5 \times \delta_2 \times \Delta m \end{aligned} \tag{20}$$

[26] With discrete data "area" should be interpreted as a scaled sum over a set $y \in Q = \{r, \ldots, s\}$, e.g. $A(r,s) = \frac{R-1}{R} \sum_{y \in Q} \left(F_0(y) - F_1(y)\right)$ if Q is a set of consecutive integers (with $R = \#Q$).

[27] Imagine at this point that F_0 and F_1 do not cross on the interval (r,s). Crossovers are seen below to be irrelevant for the particular approaches proposed here.

where $m^* = .5(m_0 + m_1)$.[28,29]

$a_1(m^*,.5)$ and $a_2(.5)$ are necessarily non-negative since for any y between m_0 and m_1 Δm has the opposite sign of the terms multiplying it in (19) and (20).[30] For practical purposes it is useful to re-define $a_1(m^*,.5)$ and $a_2(.5)$ so their signs indicate the Δm -direction of treatment effectiveness, i.e.

$$a_1(m^*,.5) = \text{sign}(\Delta m) \times (\delta_1 \times \Delta m) \tag{21}$$

and

$$a_2(.5) = \text{sign}(\Delta m) \times (.5 \times \delta_2 \times \Delta m), \tag{22}$$

with $A(m_j, m_k)$ analogously signed. These revised definitions are used henceforth.

Figure 8 depicts $a_1(m^*,.5)$ and $a_2(.5)$ when y_0 and y_1 are imagined to be gamma-distributed with $m_0 = 4$ and $m_1 = 2$. In the top panel $-a_1(m^*,.5) = .70$ is the area of the shaded rectangle while in the bottom panel $-a_2(.5) = .64$ is the area of the shaded trapezoid.[31]

[figure 8 about here]

[28] Note that m^* may or may not be in the common support Y (e.g. if Y is the set of non-negative integers).

[29] The vertical distances between F_0 and F_1 described by the δ_j not only define probability treatment effects but also correspond roughly to the degree of informativeness (or tightness) of bounds on parameters like $m\Delta$ and IP (see section 4 and appendix C). Moreover while the roles played by the δ_j in (21) and (22) may be intuitive per se, it might also be noted that they are respectively (if positive) Boole-Fréchet lower bounds on $\Pr(y_j < m^*, y_k > m^*)$ and $\Pr(y_j < m_k, y_k > m_j)$ (if the $\delta_j < 0$ then reverse the inequality directions in these probability statements). These translate roughly as "y_j is small and y_k is large so $y_k - y_j$ is large."

[30] Note that F_0 and F_1 cannot cross over their local domains in the definitions of $a_1(m^*,.5)$ and $a_2(.5)$. E.g. suppose $m_1 < m_0$; then $F_1(y) \geq .5$ and $F_0(y) < .5$ for all $y \in (m_1, m_0)$.

[31] The .5 multiplier in the expression for $a_2(.5)$ in (22) arises since the area of the trapezoid PQRS is the combined area of the two right triangles PSQ and RQS.

Noteworthy from (21) and (22) is that $a_1\left(m^*,.5\right)$ and $a_2\left(.5\right)$ respect but augment Δm. Both are intuitive, easily computed indicators of treatment effectiveness,[32] measures that provide broader characterizations of effectiveness than does Δm on its own. That is $a_1\left(m^*,.5\right)$ and $a_2\left(.5\right)$ describe more comprehensively than does Δm the divergence of F_0 and F_1 in the "middle" of the data. In an important sense $a_1\left(m^*,.5\right)$ and $a_2\left(.5\right)$ combine the motivations underlying (14) and (15); for instance

$$a_1\left(m^*,.5\right) = \operatorname{sign}\left(\Delta m\right) \times \left(-p_\omega\left(\left\{.5\right\}\right) \times p_\pi\left(\left\{m\,*\right\}\right)\right). \tag{23}$$

The sense in which central aperture and its approximations are indicators of treatment effectiveness can be appreciated with reference to figure 9 where $m_0 = 4$ and $m_1 = 2$ in both panels wherein F_0 is the same but F_1 differs (each F_j is gamma distributed). In the top panel $\delta_1 = -.08$ and $a_1\left(m^*,.5\right) = -.16$ while in the bottom panel $\delta_1 = -.18$ and $a_1\left(m^*,.5\right) = -.35$. This difference in approximate central aperture conforms to a notion that the effectiveness of T_1 relative to T_0 is greater in the scenario depicted in the figure's bottom panel even though $\Delta m = -2$ in both cases. If a decisionmaker must choose among T_0, $T_{1(\text{top})}$, or $T_{1(\text{bottom})}$ which do they select? Might knowledge of the respective $a_1\left(m^*,.5\right)$ influence or support their choice?[33]

[figure 9 about here]

[32] $a_1\left(m^*,.5\right)$ and $a_2\left(.5\right)$ can be estimated even with right censoring; they are point identified so long as both $F_j\left(y\right)$ are point identified for all $y \le \max\left\{m_0, m_1\right\}$.

[33] For positive y the *global aperture* of F_0 and F_1—the area between F_0 and F_1 over the entirety of Y—is the difference in means $E_0\left[y\right] - E_1\left[y\right] = -\Delta E\left[y\right]$ (necessarily finite in a sample but perhaps not so in a population where Y is unbounded). If a decisionmaker uses $\Delta E\left[y\right]$ to inform choice they are appealing to the global aperture of F_0 and F_1 even if they do not appreciate this explicitly. While $\Delta E\left[y\right]$ could be used to gauge treatment effectiveness it will often be impractical in clinical studies to do so due to right censoring. Also note that F_0 and F_1 can cross over Y, making any corresponding notion of aperture as "opening" somewhat fuzzy.

One might also consider *tail aperture* or *quartile aperture* to assess treatment effectiveness, akin to how one might consider comparisons of IQRs, e.g. $\mathrm{A}\left(\mathrm{q}_{\mathrm{j}}(\alpha),\mathrm{q}_{\mathrm{k}}(\alpha)\right)$ for $\alpha=.25$, $\alpha=.75$, etc., implemented using $\mathrm{a}_{\mathrm{j}}(\ldots)$ approximations. A vector of such measures might complement the Imbens and Wooldridge, 2009, approach described in (14) and (15), e.g. $\mathrm{p}_{\xi}(\mathrm{J})=\sum_{\alpha\in \mathrm{J}}\xi_{\alpha}\times \mathrm{A}\left(\mathrm{q}_{\mathrm{j}}(\alpha),\mathrm{q}_{\mathrm{k}}(\alpha)\right)$. Such indicators may be instructive regarding treatment effectiveness even when the information conveyed by $\mathrm{a}_1(\mathrm{m}^*,.5)$ or $\mathrm{a}_2(.5)$ is tenuous, as with the Cao et al., 2020, data where $\mathrm{A}\left(\mathrm{m}_{\mathrm{j}},\mathrm{m}_{\mathrm{k}}\right)=\mathrm{a}_1(\mathrm{m}^*,.5)=\mathrm{a}_2(.5)=\Delta\mathrm{m}=0$ (see figure 2).

Inspection of (21) and (22) suggests that $\mathrm{a}_1(\mathrm{m}^*,.5)$ and $\mathrm{a}_2(.5)$ can be interpreted as *weighted* Δm. One might thus consider more generally a class of weighted $\Delta\mathrm{m}$, $\mathrm{w}\times\Delta\mathrm{m}$, where $\mathrm{w}\in[0,1]$ are point-identifiable weights describing some decision-relevant aspects of the divergence of F_0 and F_1. Conventional analysis effectively sets $\mathrm{w}=1$ but there is no reason to believe $\mathrm{w}=1$ best informs a decisionmaker's choice. Such weighted $\Delta\mathrm{m}$ parameters build on $\Delta\mathrm{m}$ but downweight it the smaller is w. Reasonable candidates for w might include δ_1 and $.5\times\delta_2$ as in (21) and (22), as well as other parameters that characterize vertical distances between F_0 and F_1, e.g. the D_{jk}.[34]

The preceding discussion is not advocacy for using a particular aperture measure or weighted $\Delta\mathrm{m}$ as a decision standard. Instead it is an appeal for future research to assess the merits of easily implementable and point-identified criteria like $\mathrm{a}_1(\mathrm{m}^*,.5)$ and $\mathrm{a}_2(.5)$ for decisionmaking using point-identified parameters, the premise being that decisions should be informed more comprehensively than by appealing to $\Delta\mathrm{m}$ alone. Consideration of central aperture might serve as a useful starting point for such exploration but is unlikely to be its destination. Moving beyond the intuition that parameters akin to $\mathrm{a}_1(\mathrm{m}^*,.5)$ and $\mathrm{a}_2(.5)$ describe treatment effectiveness to a formal assessment of welfare differences they may describe might also be a valuable step.

Determining what values of $\mathrm{a}_1(\mathrm{m}^*,.5)$ or $\mathrm{a}_2(.5)$—or indeed any other parameters

[34] Referring to $\mathrm{w}\times\Delta\mathrm{m}$ to gauge treatment effectiveness recalls the Harberger, 1971 (eq. 2), first-order approximation to welfare change due to a "treatment": the product of marginal value and quantity change. Quantity change alone—here $\Delta\mathrm{m}$—tells an incomplete story; a fuller story unfolds by considering as well the worth of quantity along the margin of quantity change.

advanced in such discussions—correspond to "clinically significant" differences between F_0 and F_1 will also require novel consideration. (For instance if Δ_{CS} is considered a clinically significant Δm in a particular treatment context then might a magnitude of $a_1\left(m^*,.5\right)$ or $a_2\left(.5\right)$ exceeding $.25 \times \Delta_{CS}$ suggest clinical significance? Or $.1 \times \Delta_{CS}$? Or … ?) While ultimately essential, such pragmatic considerations ought not postpone exploring the merits of parameters that better inform decisionmakers about treatment effectiveness than do those conventionally used, particularly when such novel parameters are straightforward to implement in practice.

6. COVID-19: Three Case Studies

Case Study 1: Remdesivir for COVID-19 Treatment (Beigel et al., 2020)

The Beigel study reports results of a randomized, double-blind, placebo-controlled trial of intravenous remdesivir in adults hospitalized with COVID-19. Its primary outcome is time to recovery (TTR) measured in days, with recovery determined by a patient meeting a pre-specified threshold on an ordinal scale. The intention-to-treat sample analyzed here[35] consists of 538 and 521 observations on treated and control patients, respectively, of which 147 and 181 observations are right-censored after 28 days. The bottom panel of figure 1 depicts these data.

Table 1 summarizes the results. As noted earlier $\Delta m = -4$, favoring treatment over control, while $B^*\left(m\Delta\right) = \left\{-15,10\right\}$. The bounds interval is fairly wide as can be expected from the wide IP bounds. While the Δm result might lead a decisionmaker to favor treatment over control, wide bounds on $m\Delta$ might encourage that same decisionmaker to be conservative in advancing such a recommendation should considerations beyond Δm be relevant. Using earlier arguments, however, note that p_β is sign identified (negative) for $\beta \in \left[5/7,1\right]$, so that a decisionmaker who appeals to (12) and who weighs Δm no more than $2/7$ is positioned to recommend treatment over control.

Finally, for the Beigel data the central aperture measures are $A\left(m_1, m_0\right) = -.394$, $a_1\left(m^*,.5\right) = -.386$, and $a_2\left(.5\right) = -.375$.

Case Study 2: Lopinavir-Ritonavir for COVID-19 Treatment (Cao et al., 2020)

The Cao study reports the results of an RCT involving hospitalized patients with

[35] The data used in this section's analyses of the first two case studies as well as in producing figure 1 (bottom panel) and figure 2 were coded from "eyeball analysis" of Beigel's figure 2A and Cao's figure 2 then calibrated as necessary to match the reported sample medians.

confirmed SARS-CoV-2 infection. Treatment consisted of receipt of lopinavir-ritonavir plus standard care, while control consisted exclusively of standard care. The trial's primary endpoint is time to clinical improvement (TTI), measured in days. The trial was open label; moreover the authors report that placebo treatment in the control group was not possible due to the trial's emergency nature. The intention-to-treat sample consists of 199 patients with 99 and 100 randomized to treatment and control, respectively. By the administrative censoring time of 28 weeks clinical improvement was observed for 77 of the 99 treatment subjects and 70 of the 100 control subjects; thus 22 and 30 subjects' data, respectively, are treated as right-censored. Figure 2 reproduces the data depicted in Cao's figure 2.

Median TTI is 16 days in both arms of the trial so $\Delta m = 0$. A decisionmaker concerned only with marginal median outcomes would favor neither treatment. But $B^*\left(m\Delta\right) = \left\{-9,7\right\}$, so the magnitude of $m\Delta$ could be quite substantial in either direction. Regarding central aperture $A\left(m_j, m_k\right) = a_1\left(m^*,.5\right) = a_2\left(.5\right) = 0$ in the Cao sample as $\Delta m = 0$.[36,37]

Case Study 3: Combination Therapy for COVID-19 Treatment (Hung et al., 2020)

The Hung study reports the results of an open-label randomized phase-2 trial comparing combination treatment for COVID-19 (interferon beta-1b plus lopinavir–ritonavir plus ribavirin) with lopinavir–ritonavir alone in a sample of hospitalized patients. The study's primary outcome is time to providing a nasopharyngeal swab negative for severe acute respiratory syndrome coronavirus 2 RT-PCR. 86 and 41 patients were randomly assigned to treatment and control. The time to negative test (TTNT) outcomes are reported as integers (days) and analyzed in an intention-to-treat context.[38]

[36] As seen in figure 2 F_0 first-order dominates F_1 (at least up to the censoring threshold), so that a decisionmaker might have other reasons to favor T_1.

[37] While this paper has sidestepped issues of sampling variation and inference it might be noted that $a_1\left(m^*,.5\right)$ and $a_2\left(.5\right)$ are themselves subject to sampling variation. Even in a case like the Cao data where the point estimates of both measures are zero (since $\Delta m = 0$) there will generally arise non-degenerate sampling distributions. A simple bootstrap with 1,000 replications shows for those data 95% central sampling intervals of $\left[-.85,.09\right]$ for $a_1\left(m^*,.5\right)$ and $\left[-.80,.07\right]$ for $a_2\left(.5\right)$ with respective sampling distribution medians of $-.04$ and $-.04$. (For the Beigel data the corresponding results are $\left[-1.15,-.05\right]$ for $a_1\left(m^*,.5\right)$ and $\left[-1.11,-.04\right]$ for $a_2\left(.5\right)$ with respective sampling distribution medians of $-.42$ and $-.40$.)

[38] The Hung study reports no information on right-censoring. The analysis proceeds as if the data are uncensored, i.e. that the reported quantiles are those corresponding to the full sample.

Figure 6 depicts the medians and IQRs reported in Hung's table 2; this is the only information the study provides on the outcomes' marginal distributions. As such, the reported TTNT data are right-censored and interval-measured (e.g. $q_1(.25)=5$ implies $F_1(5)\in[.25,.5)$, as depicted by the closed and open symbols in figure 6). For these data $\Delta m=-5$. Applying (11) to these data yields $B_{IQR}(m\Delta)=\{-11,4\}$.[39]

7. Summary

Ease of computation, broad applicability, and parsimony in summarizing outcome data are three plausible *statistical* reasons for the prominence of Δm in clinical research. By a revealed preference argument, Δm must routinely provide decisionmakers with useful information about the choices they confront. Yet this popularity could arise either because parameters other than Δm are not of interest[40] *or* because parameters beyond Δm are of interest but the cost of using them to inform choices is perceived to be too great.

This paper has suggested that it is straightforward to augment the signals about treatment effectiveness sent by Δm with information about other features of outcomes' distributions in such ways as should more comprehensively inform decisionmakers who must choose on the basis of the data at hand, *whether or not* those other features are point identified. In essence the strategies presented in this paper may serve to reduce the perceived costs of appealing to parameters beyond just Δm for those decisionmakers who do find them of interest.

While one might imagine a portfolio of alternative approaches, the particular ideas advanced here are designed to be broadly roadworthy since they are easily implemented and require—whether a decisionmaker follows path 1 or path 2—assumptions hardly more stringent than those needed for identification of Δm itself. That the measures of treatment effectiveness proposed here are all anchored to Δm should also be of comfort to decisionmakers who have traditionally and broadly relied on Δm in making choices. How these approaches perform in practice, whether they provide useful information to decisionmakers, and how corresponding

[39] A curiosity is that IQRs and medians are sometimes reported when right-censoring probabilities exceed .25 or .5. For example, the Cao study reports IQR_0 as $\{15,18\}$ even though the right-censoring fraction in the control group is .30. Other clinical studies report analogous results. Whether what is reported in such instances derives from model-based predictions (e.g. from proportional hazard models), from other sources, or from erroneous calculations is not always evident. In no event, however, are right-censoring probabilities exceeding .5 or .25 logically consistent with point-identified medians or .75-quantiles.

[40] For example, $\beta=1$ in (12) or $J=\{.5\}$, $\omega_{.5}=1$, and $v(\ldots)$ is invariant w.r.t. p_π in (16).

standards for clinical significance should be defined are among the issues still to be resolved.

Finally the paper has focused on issues involving identification, ignoring considerations of inference used to understand implications of sampling variation. Such considerations may be of interest in some decisionmaking contexts and, if so, would be useful to tackle in future study.

Acknowledgments

Thanks are owed to Chris Adams, Marguerite Burns, Domenico Depalo, Chuck Manski, Dan Millimet, Ciaran O'Neill, and Jon Skinner for helpful comments on an earlier draft.

References

Beigel, J.H. et al. 2020. "Remdesivir for the Treatment of Covid-19—Preliminary Report." *NEJM*. DOI: 10.1056/NEJMoa2007764

Cao, B. et al. 2020. "A Trial of Lopinavir–Ritonavir in Adults Hospitalized with Severe Covid-19." *NEJM* 382: 1787-1799.

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

Cravens, S.M.R. 2002. "The Usage and Meaning of 'Clinical Significance' in Drug-Related Litigation." *Washington and Lee Law Review* 59: 553-597.

European Medicines Agency (EMA). 2020. *ICH E9 (R1) Addendum on Estimands and Sensitivity Analysis in Clinical Trials to the Guideline on Statistical Principles for Clinical Trials—Step 5*. Document EMA/CHMP/ICH/436221/2017. Amsterdam: EMA.

Fan, D.S.P. et al. 2003. "Ocular-Hypertensive and Anti-Inflammatory Response to Rimexolone Therapy in Children." *Archives of Ophthalmology* 121: 1716-1721.

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

Ganz, R.A. et al. 2013. "Esophageal Sphincter Device for Gastroesophageal Reflux Disease." *NEJM* 368: 719-727.

Goldman, M. and D.M. Kaplan. 2018. "Comparing Distributions by Multiple Testing across Quantiles or CDF Values." *Journal of Econometrics* 206: 143-166.

Hansen, B. 2020. *Introduction to Econometrics* (Vol. 1). Web textbook. University of Wisconsin-Madison, Department of Economics. (https://www.ssc.wisc.edu/~bhansen/probability/, accessed May 21, 2020) (May 2020 version).

Harberger, A.C. 1971. "Three Basic Postulates for Applied Welfare Economics: An Interpretive Essay." *Journal of Economic Literature* 9: 785-797.

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

Hung, I. F.-N. et al. 2020. "Triple Combination of Interferon beta-1b, Lopinavir–Ritonavir, and Ribavirin in the Treatment of Patients Admitted to Hospital with COVID-19: An Open-label, Randomised, Phase 2 Trial." *The Lancet* 395: 1695-1704.

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

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

Lee, M.-J. 2000. "Median Treatment Effect 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 Demands." *Health Economics* 10: 411-428.

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

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

Manski, C.F. 2018a. "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment." *Quantitative Economics* 9: 541–569.

Manski, C.F. 2018b. "Reasonable Patient Care under Uncertainty." *Health Economics* 27: 1397-1421.

Manski, C.F. 2020a. "The Lure of Incredible Certitude." *Economics and Philosophy* 36: 216-245.

Manski, C.F. 2020b. "Bounding the Predictive Values of COVID-19 Antibody Tests." Working Paper (May 14, 2020, version). Department of Economics and Institute for Policy Research, Northwestern University.

Manski, C.F. and F. Molinari. 2020. "Estimating the COVID-19 Infection Rate: Anatomy of an Inference Problem." *Journal of Econometrics* (in press) https://doi.org/10.1016/j.jeconom.2020.04.041.

Manski, C.F. and A. Tetenov. 2020. "Statistical Decision Properties of Imprecise Trials Assessing COVID-19 Drugs" NBER Working Paper 27293.

Mullahy, J. 2018a. "Individual Results May Vary: Inequality-Probability Bounds for Some Health-Outcome Treatment Effects." *Journal of Health Economics* 61: 151-162.

Mullahy, J. 2018b. "Health Status Measurement." Chapter in *Oxford Research Encyclopedia of Economics and Finance*, A. Jones, Ed. Oxford: Oxford University Press.

Pocock, S.J. 1992. "When to Stop a Clinical Trial." *BMJ* 305: 235-240.

Rogawski, E.T. et al. 2017. "Estimating Differences and Ratios in Median Times to Event." *Epidemiology* 27: 848-851.

U.S. Dept. of Health and Human Services. 2016. *42 CFR Part 11, Clinical Trials Registration and Results Information Submission; Final Rule.* Federal Register 81(183), Part II (September 21, 2016).

U.S. Food and Drug Administration. 2006. *Clinical Studies Section of Labeling for Human Prescription Drug and Biological Products—Content and Format: Guidance for Industry.* Rockville, MD: U.S. FDA.

U.S. Food and Drug Administration. 2009. *Guidance for Industry: Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims.* Rockville, MD: U.S. FDA.

U.S. Food and Drug Administration, Center for Devices and Radiological Health. 2012. Transcript of the January 11, 2012, Meeting of the Gastroenterology and Urology Devices Advisory Panel. https://wayback.archive-it.org/7993/20170113191551/

http://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDevicesAdvisoryCommittee/Gastroenterology-UrologyDevicesPanel/UCM291391.pdf (accessed May 28, 2020).

U.S. Food and Drug Administration. 2020a. *Coronavirus (COVID-19) Update: FDA Issues Emergency Use Authorization for Potential COVID-19 Treatment.* https://www.fda.gov/news-events/press-announcements/coronavirus-covid-19-update-fda-issues-emergency-use-authorization-potential-covid-19-treatment (May 1, 2020; accessed May 6, 2020).

U.S. Food and Drug Administration. 2020b. *FDA Guidance on Conduct of Clinical Trials of Medical Products during COVID-19 Public Health Emergency.* Rockville, MD: U.S. FDA (version May 14, 2020).

U.S. National Institute on Allergy and Infectious Diseases (NIAID). 2020a. *NIH Clinical Trial Shows Remdesivir Accelerates Recovery from Advanced COVID-19.* https://www.niaid.nih.gov/news-events/nih-clinical-trial-shows-remdesivir-accelerates-recovery-advanced-covid-19 (April 29, 2020; accessed May 6, 2020).

U.S. National Institute on Allergy and Infectious Diseases (NIAID). 2020b. *Adaptive COVID-19 Treatment Trial (ACTT). ClinicalTrials.gov* Identifier: NCT04280705. https://clinicaltrials.gov/ct2/show/record/NCT04280705 (accessed May 6, 2020).

Washington Post. 2020. "Gilead’s remdesivir improves recovery time of coronavirus patients in NIH trial." https://www.washingtonpost.com/business/2020/04/29/gilead-says-positive-results-coronavirus-drug-remdesivir-will-be-released-by-nih/ (reported April 29, 2020; accessed May 6, 2020).

Zarin, D.A. et al. 2011. "The ClinicalTrials.gov Results Database—Update and Key Issues." *NEJM* 364: 852-860.

Zarin, D.A. et al. 2016. "Trial Reporting in ClinicalTrials.gov—The Final Rule." *NEJM* 375: 1998-2004.

Appendix A: Computing Medians

The definition of $\mathrm{m_j}$ in section 2 may not coincide with how statistical packages compute medians for samples where N is even. In these cases the definition (1) selects the .5N low-to-high order statistic, $\mathrm{y_{(.5N)}}$, while some packages (e.g. Stata commands, including `centile`, `summarize`, and `tabstat`) compute the median as the average of $\mathrm{y_{(.5N)}}$ and $\mathrm{y_{(.5N+1)}}$. While this discrepancy will typically be inconsequential when analyzing large samples this is not necessarily so for smaller samples (see table A1).

As such, understanding what formulae are used in published research and whether these align with the parameters that are conceptually of interest may be worthwhile. Also noteworthy is that Stata has several commands that compute percentiles, and for quantiles other than the median these may not yield the same results (though the different formulae are documented; e.g. compare `centile` and `summarize`). In small-to-medium-sized samples such differences may be nontrivial in estimating some parameters, e.g. IQRs.

Additional Measurement Considerations

Measuring TTEs as integers (e.g. days) may introduce a form of measurement error. Suppose the outcome of interest is TTR. As in the Beigel study, a patient is assessed once daily at an unreported time with respect to recovery status. Suppose an assessment performed at one second after midnight on day two indicates the patient as not recovered but that recovery occurs at some point later on day two. Assuming recovery is an absorbing state, then the assessment taken sometime on day three would indicate recovery. That day-three assessment could in theory take place as late as one second before midnight. Thus observing the patient's TTR as day=3 means that the true recovery time could be anywhere in the open interval $\left(2,4\right)$. Therefore the true difference in median outcomes is only partially identified, known only to be in the interval $\left(\Delta\mathrm{m}-2,\Delta\mathrm{m}+2\right)$ where $\Delta\mathrm{m}$ is the observed integer difference in medians. While such complications could be relevant in practice—especially when TTEs are "small"—they are ignored in this paper.

Appendix B: Monotone Treatment Response and Distributional Assumptions

Monotone Treatment Response

While this paper does not make use of assumptions about monotone treatment response, its prominence in related work merits brief mention. In a foundational paper Manski, 1997, shows the potential identifying power of monotone response to treatment in partial identification settings. While typically not testable, assumptions about monotone response—outcomes are increasing or at least non-decreasing in treatment—will generally provide tighter bounds on partially identified parameters.

Several researchers have considered special cases in the spirit of monotone response that may yield tighter bounds on $m\Delta$ and related parameters. Lee, 2000, describes five cases ("Special Models," or SM):

(SM1) $y_1 = y_0 + q$, $q \geq 0$

(SM2) $y_1 = y_0 + \varepsilon$, $\varepsilon \geq 0$

(SM3) $y_1 = y_0 + \varepsilon$, ε unrestricted

(SM4) $y_1 = y_0 + \varepsilon$, ε symmetric

(SM5) $y_1 = \xi(y_0)$, $\xi(y_0) \geq y_0$ and $\xi(y_0)$ increasing in y_0

Relating $sign(m\Delta)$ *to* $sign(\Delta m)$

Lee draws on this framework to derive necessary and sufficient conditions under which $\Delta m = 0 \rightarrow m\Delta = 0$ and $\Delta m > 0 \rightarrow m\Delta > 0$. Specifically $\Delta m = 0 \rightarrow m\Delta = 0$ if and only if

$$\Pr(y_0 - m_0 < y_1 - m_1 < 0) \;=\; \Pr(0 < y_1 - m_1 < y_0 - m_0), \tag{B.1}$$

what Lee calls "equal probability of octants." Moreover, $\Delta m > 0 \rightarrow m\Delta > 0$ if and only if

$$\Pr(m_1 < y_1 < y_0) \;<\; \Pr(y_0 < y_1 < m_1). \tag{B.2}$$

These conditions are not verifiable when the joint distribution $F(y_0, y_1)$ is unknown.

Stochastic Dominance

Determining that $F_0(y)$ and $F_1(y)$ obey a stochastic dominance relationship is of limited value for understanding magnitude or sign relationships between Δm and $m\Delta$. For example, a first-order dominance relationship between $F_0\left(y\right)$ and $F_1\left(y\right)$ implies knowledge of $\text{sign}\left(\Delta m\right)$ as well as the signs of all marginal quantiles' differences, but implies nothing correspondingly about $\text{sign}\left(m\Delta\right)$. Higher-order dominance is no more informative.

Appendix C: Details on Computation of Bounds on $\mathbf{m\Delta}$

Computing Bounds using Permutation Distributions

A intuitive strategy for computing tightest bounds on $\mathrm{m}\Delta$ imagines a sample of $\mathrm{N}_0 = \mathrm{N}_1 = \mathrm{N}$ observations on y_0 and y_1, each gathered into N-vectors $\mathbf{y}_\mathrm{j}$. Define $\mathbf{\Pi}\left(\mathbf{y}_1\right)$ as the $\mathrm{N} \times \mathrm{N}!$ matrix whose columns are the N! permutations of $\mathbf{y}_1$. Then the distributions of the elements of each column of the $\mathrm{N} \times \mathrm{N}!$ matrix $\mathbf{D}$ whose p-th column is $\mathbf{\Pi}\left(\mathbf{y}_1\right)_\mathrm{p} - \mathbf{y}_0$, $\mathrm{p} \in \left\{1,\ldots,\mathrm{N}!\right\}$, are all the possible treatment-effect distributions consistent with the data.

Let $\mathbf{m} = \left[\mathrm{m}_1,\ldots,\mathrm{m}_{\mathrm{N}!}\right]$ be the $1 \times \mathrm{N}!$ vector of column medians of $\mathbf{D}$. Given the data the tightest possible bounds $\mathrm{L}^*\left(\mathrm{m}\Delta\right)$ and $\mathrm{U}^*\left(\mathrm{m}\Delta\right)$ are the smallest and largest elements of $\mathbf{m}$; this holds since any claimed lower bound larger than $\min\left(\mathbf{m}\right)$ or upper bound smaller than $\max\left(\mathbf{m}\right)$ is inconsistent with the data. When N is odd these permutation bounds are unique; when N is even one must decide how to define medians, as discussed in appendix A.

While intuitively straightforward, this approach is of limited value in applications. When $\mathrm{N}_0 \neq \mathrm{N}_1$ $\mathbf{D}$ cannot be defined as above.[41] Even if $\mathrm{N}_0 = \mathrm{N}_1$ relying on permutations becomes less practical as N increases.[42] Moreover, with right-censored data it is not evident how the permutation approach can be used even if $\mathrm{N}_0 = \mathrm{N}_1$.

Computing Bounds using D_{jk}

In lieu of permutation-based solutions one can appeal to a generally applicable approach that relies on D_{01} and D_{10} and the inequality probabilities defined in (9). Tightest bounds on the IP (Fan and Park, 2010; Mullahy, 2018) are given by

$$\mathrm{D}_{01} \ \leq \ \Pr\left(\mathrm{y}_1 > \mathrm{y}_0\right) \ = \ \Pr\left(\Delta > 0\right) \ \leq \ 1 - \mathrm{D}_{10} \qquad \text{(C.1)}$$

and

[41] If $\mathrm{N}_0 \neq \mathrm{N}_1$ then imagine the imbalance arises from different sampling probabilities. Let C be the least common multiple of N_0 and N_1. Then replicating each y_j $\mathrm{C}/\mathrm{N}_\mathrm{j}$ times yields a balanced sample of C observations from which the permutation distribution can be computed.

[42] Heckman et al., 1997, suggest approximating such data by defining the y_j as selected quantiles of each F_j , so that $\mathrm{N}_0 = \mathrm{N}_1$, and undertaking permutation analysis of the quantiles.

$$D_{10} \ \leq \ \Pr\left(y_0 > y_1\right) \ = \ \Pr\left(\Delta < 0\right) \ \leq \ 1 - D_{01}. \qquad \text{(C.2)}$$

For an intuition on determining the tightest bounds on $m\Delta$ use (C.2) and consider a case where it is observed that $D_{10} > .5$. Then it follows from (C.2) and R3 that $\Pr\left(\Delta < 0\right) > .5$ so that $m\Delta$ must be negative. To determine the value of $U^*\left(m\Delta\right)$ consider an algorithm where the elements of y_1 are all displaced by some positive amount d (i.e. $y_1 + d$) and then compute the corresponding $D_{10}\left(d\right) = \max_y \left\{F_1\left(y+d\right) - F_0\left(y\right)\right\}$. If $D_{10}\left(d\right) \geq .5$ then $U^*\left(m\Delta\right)$ is no greater than -d; if $D_{10}\left(d\right) < .5$ then $U^*\left(m\Delta\right)$ is between -d and zero. Repeating this search over all possible d identifies the largest value of d for which $D_{10}\left(d\right) \geq .5$, thus defining $U^*\left(m\Delta\right) = -d$.

If initially $D_{10} < .5$ then $U^*\left(m\Delta\right)$ is positive. To determine this bound reverse the previous procedure: displace the elements of y_1 as $y_1 - d$ until is determined the smallest d for which $D_{10}\left(d\right) \geq .5$ which value of d will define $U^*\left(m\Delta\right) = d$. Determining $L^*\left(m\Delta\right)$ is analogous, using D_{01} and corresponding $D_{01}\left(d\right)$. With integer-valued outcomes like the TTE data examined in this paper the determination of $B^*\left(m\Delta\right)$ is straightforward since the d-search is over a regular (integer) grid; with continuous outcomes the search will rely on some reasonably selected step sizes.[43]

This discussion establishes a link between the D_{jk} and sign identification of $m\Delta$. Specifically $m\Delta$ is not sign identified unless one of the D_{jk} exceeds .5. That is sign identification of $m\Delta$ corresponds to a lower bound greater than .5 on one of the D_{jk} . $\Pr\left(y_1 > y_0\right) > .5$ if and only if $L^*\left(m\Delta\right) > 0$.[44]

Figure C1 demonstrates the intuition of the search algorithm using the Cao sample where the outcomes are measured as integers. For these data $D_{10} < .5$. The top panel of Figure

[43] This search algorithm is equivalent to, but the inverse of, the algorithm proposed in lemma 2.3 of Fan and Park, 2010, that searches over u to determine the largest difference $F_1^{-1}\left(u\right) - F_0^{-1}\left(u+.5\right)$ over the domain $u \in \left[0,.5\right]$.

[44] The distinction between strict and weak inequalities in the preceding discussion is largely irrelevant if the focus is on continuously distributed outcomes. With discrete/integer outcomes, however, such distinctions may be important.

C1 indicates that a displacement of $d=-8$ results in $D_{10}\left(-8\right)=.48$ so that the displacement of -8 is not sufficiently large to determine $U^{*}\left(m\Delta\right)$. However, with a displacement of $d=-9$ (bottom panel of Figure C1), $D_{10}\left(-9\right)=.52$. As such $U*\left(m\Delta\right)=9$.

[figure C1 about here]

Using simulated integer data and small $N_0=N_1=N$ (N≤12), the `medte` program mentioned in section 4, which uses the $D_{jk}-\text{based}$ algorithm described here, is seen to have the following properties when compared with the exact permutation algorithm: when N is even `medte` returns bounds between those returned by the permutation algorithm based on the .5N and $.5N+1$ order-statistic median definitions (e.g. $B^{*}=\left\{-1,6\right\}$ versus $\left\{-2,5\right\}$ or $\left\{0,7\right\}$); when N is odd `medte` returns bounds one unit more conservative than those returned by the permutation algorithm (e.g. $B^{*}=\left\{-2,9\right\}$ versus $\left\{-1,8\right\}$).

Computing Bounds with Right-Censored Data

Computation of $m\Delta$ bounds with right-censored outcomes is no different than with fully observed outcomes so long as the censoring is properly accommodated in the various probabilities involved in computation.

While such bounds will generally not be tightest in the sense of the bounds that would arise were the data not censored they will still generally be informative. Moreover as noted in section 4 reducing the amount of right-censoring—i.e. increasing y_c—can never widen $m\Delta$ bounds but may narrow them. To see this consider an extreme example depicted in the bottom panel of figure 5 where $y_c=11$ (i.e. all values of y greater than 11 are unobserved). In this example the marginal medians and Δm are identified, with $m_0=m_1=10$ and $\Delta m=0$. Based on the observable (uncensored) data the $m\Delta$ bounds are $B^{*}\left(m\Delta\right)=\left\{-10,10\right\}$. Were the data fully uncensored the resulting bounds would be $B^{*}\left(m\Delta\right)=\left\{-10,2\right\}$ whereas with a less stringent censoring threshold, say $y_c=15$, the bounds would be $B^{*}\left(m\Delta\right)=\left\{-10,7\right\}$.

Computing Bounds with Quartile Data

Because it does not use information from the entirety of the F_j marginals $B_{IQR}\left(m\Delta\right)$

defined in section 4 is not generally sharp though in some specific cases (e.g. $y_j \sim N\left(\mu_j, \sigma^2\right)$) it is. In any event bounding $m\Delta$ using information only on quartiles requires that the IQR_j be point identified, which may be problematic when right-censoring is prevalent. Just as the m_j are not point identified when right-censoring fractions exceed .5, IQR_j is not identified when $F_j\left(y\right)$ is right censored at $y < F_j^{-1}\left(.75\right)$.[45]

The basic idea in computing $B_{IQR}\left(m\Delta\right)$ can be illustrated concretely for integer-measured outcomes using the Hung data.[46] The estimated bounds from (11) are $B_{IQR}\left(m\Delta\right) = \left\{-11, 4\right\}$ (note $\Delta m = -5$). Figure C2 depicts the computation of the lower bound. The distribution $F_1\left(y\right)$ is displaced rightward by one-unit increments until the maximum (over y) vertical distance between $F_0\left(y\right)$ and the displaced $F_1\left(y+d\right)$ can be determined to be at least .5. Imagine a rightward displacement of $F_1\left(y\right)$ of $d = -10$. At y=15 $F_1\left(y-10\right) \in \left[.25, .5\right)$ and $F_0\left(y\right) - F_1\left(y-10\right) \in \left(.25, .75\right]$. Consequently the distribution has not been displaced sufficiently far rightward to determine $L_{IQR}\left(m\Delta\right)$. Consider now a rightward displacement of $F_1\left(y\right)$ of $d = -11$. At y=15 $F_1\left(y-11\right) \in \left[0, .25\right)$ and $F_0\left(y\right) - F_1\left(y-11\right) \in \left(.5, 1\right]$. Thus

$$L_{IQR}\left(m\Delta\right) = q_{.25,1} - q_{.75,0} - 1 = 5 - 15 - 1 = -11 \qquad (C.3)$$

since $d = -11$ is the smallest possible shift guaranteeing $F_0\left(y\right) - F_1\left(y+d\right) \geq .5$.

[45] Some simulations suggest that $B_{IQR}\left(m\Delta\right)$ is often close to the sharp bounds obtained from fully observed data. E.g. with $F_j = \text{uniform}\left(0, a_j\right)$ and $a_0 < a_1$ $B^*\left(m\Delta\right) = \left\{.5a_1 - a_0,\ .5a_1\right\}$ and $B_{IQR}\left(m\Delta\right) = \left\{.25a_1 - .75a_0,\ .75a_1 - .25a_0\right\}$; for $a_0 = .8$ and $a_1 = 1$ $B^*\left(m\Delta\right) = \left\{-.3,\ .5\right\}$ while $B_{IQR}\left(m\Delta\right) = \left\{-.35,\ .55\right\}$.

[46] One can also consider outcomes having continuous distributions. In samples drawn therefrom, however, F_0 and F_1 will not be smooth but instead will have a finite number of jump points at irregularly spaced non-integer y-values. Determining the bounds in such instances is conceptually no different but practically would need to accommodate irregularly spaced jump points in the search algorithm.

[figure C2 about here]

Computing Bounds with Transformed Outcomes

The signs but not the magnitudes of both Δm and $m\Delta$ are invariant to monotone-increasing transformations $t\left(y_j\right)$. Define $\Delta_t = t\left(y_1\right) - t\left(y_0\right)$. Then[47]

$$\operatorname{sign}\left(\operatorname{med}\left(F_1\left(t\left(y\right)\right)\right) - \operatorname{med}\left(F_0\left(t\left(y\right)\right)\right)\right) \; = \; \operatorname{sign}\left(\Delta m\right) \qquad \text{(C.4)}$$

and

$$\operatorname{sign}\left(\operatorname{med}\left(F\left(\Delta_t\right)\right)\right) \; = \; \operatorname{sign}\left(m\Delta\right). \qquad \text{(C.5)}$$

To determine bounds on $\operatorname{med}\left(F\left(\Delta_t\right)\right)$ two considerations arise: the first is whether the monotone positive transformation $t\left(y_j\right)$ is affine; the second is whether both the original and transformed outcomes are integers.

The general result is that the signs of the tightest lower and upper bounds on $\operatorname{med}\left(F\left(\Delta_t\right)\right)$ are the same as those on $m\Delta$. While the monotone positive transformations $t\left(y_j\right)$ change the shapes of the distribution functions $F\left(y_j\right)$ they do not change D_{01} or D_{10}, only their corresponding y-ordinates. As shown above it is the D_{jk}—in particular whether or not they exceed .5—that determine the bounds' signs. When furthermore the outcomes and transformed outcomes are continuous and $t\left(y_j\right)$ is affine (i.e. $t\left(y_j\right) = a + by_j$ with $b>0$) then $L^*\left(\operatorname{med}\left(F\left(\Delta_t\right)\right)\right) = a + bL^*\left(m\Delta\right)$ and $U^*\left(\operatorname{med}\left(F\left(\Delta_t\right)\right)\right) = a + bU^*\left(m\Delta\right)$.

When the y_j and $t\left(y_j\right)$ are all integers—e.g. converting outcomes from weeks to days by multiplying by seven—and $t\left(y_j\right)$ is affine then

$$L^*\left(\operatorname{med}\left(F\left(\Delta_t\right)\right)\right) = a + b\left(L^*\left(m\Delta\right) + 1\right) - 1 \qquad \text{(C.6)}$$

[47] Such sign-invariance does not hold for average treatment effects (ATEs) where $\operatorname{sign}\left(E\left[y_1 - y_0\right]\right)$ has no necessary implications for $\operatorname{sign}\left(E\left[t\left(y_1\right) - t\left(y_0\right)\right]\right)$.

and

$$U^*\left(\text{med}\left(F\left(\Delta_t\right)\right)\right)=a+b\left(U^*\left(m\Delta\right)-1\right)+1. \tag{C.7}$$

As such the sharp bounds on $\text{med}\left(F\left(\Delta_t\right)\right)$ have the same sign as those on $m\Delta$ except when $L^*\left(m\Delta\right)=0$ or $U^*\left(m\Delta\right)=0$ in which cases $L^*\left(\text{med}\left(F\left(\Delta_t\right)\right)\right)$ takes the sign of $U^*\left(m\Delta\right)$ if $L^*\left(m\Delta\right)=0$ and $U^*\left(\text{med}\left(F\left(\Delta_t\right)\right)\right)$ takes the sign of $L^*\left(m\Delta\right)$ if $U^*\left(m\Delta\right)=0$. While perhaps curious on first impression this result arises because the bounds in such cases are defined as integers and respect the mechanical definition medians in (1).

Appendix D: Explanation of Results 1-3

Result 1 (R1): Suppose $\Delta\mathrm{m} < 0$. Then:

a. $0 \leq \mathrm{D}_{01} \leq .5$

b. $0 < \mathrm{D}_{10} \leq 1$

c. $\mathrm{L}^{*}\left(\mathrm{m}\Delta\right) \leq 0$

Explanation

a. $\mathrm{F}_0\left(\mathrm{y}\right) < .5$ for all $\mathrm{y} < \mathrm{m}_0$ so $0 \leq \mathrm{D}_{01} = \max\left\{\mathrm{F}_0\left(\mathrm{y}\right) - \mathrm{F}_1\left(\mathrm{y}\right), 0\right\} < .5$ for all $\mathrm{y} < \mathrm{m}_0$. Since $\mathrm{m}_1 < \mathrm{m}_0$ then $\mathrm{F}_1\left(\mathrm{y}\right) \geq .5$ for all $\mathrm{y} \geq \mathrm{m}_0$, and since $\mathrm{F}_0\left(\mathrm{y}\right) \leq 1$ then $0 \leq \mathrm{D}_{01} \leq .5$ for all $\mathrm{y} \geq \mathrm{m}_0$. Thus $0 \leq \mathrm{D}_{01} \leq .5$.

b. Since $\mathrm{m}_1 < \mathrm{m}_0$ there is at least one value of $\mathrm{y} \leq \mathrm{m}_1$ where $\mathrm{F}_1\left(\mathrm{y}\right) > \mathrm{F}_0\left(\mathrm{y}\right)$ so that $\mathrm{D}_{10} > 0$. Since $0 \leq \mathrm{F}_0\left(\mathrm{y}\right) < .5$ for all $\mathrm{y} < \mathrm{m}_0$ and since $.5 \leq \mathrm{F}_1\left(\mathrm{y}\right) \leq 1$ for all $\mathrm{y} \geq \mathrm{m}_1$ then $0 \leq \mathrm{D}_{10} \leq 1$ for all $\mathrm{y} \in \left[\mathrm{m}_1, \mathrm{m}_0\right]$. Thus $0 < \mathrm{D}_{10} \leq 1$.

c. Follows directly from (a).

Result 2 (R2): $\mathrm{L}^{*}\left(\mathrm{m}\Delta\right) \ \leq \ \Delta\mathrm{m} \ \leq \ \mathrm{U}^{*}\left(\mathrm{m}\Delta\right)$

Explanation:

Consider the permutations $\boldsymbol{\Pi}\left(\mathbf{y}_1\right)$ defined in appendix C. One column of $\boldsymbol{\Pi}\left(\mathbf{y}_1\right)$ (say the p-th) must be perfectly negatively rank correlated with the elements of $\mathbf{y}_0$.[48] The median of that permutation's TE distribution $\left[\min\left(\mathbf{y}_1\right)_{\mathrm{p}} - \max\left(\mathbf{y}_0\right), \ldots, \max\left(\mathbf{y}_1\right)_{\mathrm{p}} - \min\left(\mathbf{y}_0\right)\right]$ is $\mathrm{m}_1 - \mathrm{m}_0 = \Delta\mathrm{m}$. Thus $\Delta\mathrm{m}$ is one element of the vector of permutation-distribution medians, **m**. Result 2 holds since $\mathrm{L}^{*}\left(\mathrm{m}\Delta\right)$ and $\mathrm{U}^{*}\left(\mathrm{m}\Delta\right)$ are the smallest and largest elements of **m**.

Result 3 (R3): Suppose $\mathrm{D}_{10} > .5$. Then:

a. $\Pr\left(\Delta < 0\right) > .5$

b. $\Delta\mathrm{m} < 0$

[48] With ties in the data more than one column may have this property.

c. $\mathrm{m}\Delta < 0$

Explanation:

a. Follows from (C.2).

b. Since $\mathrm{D}_{10} > .5$ then there is at least one value of y where $\mathrm{F}_1\left(\mathrm{y}\right) > .5 > \mathrm{F}_0\left(\mathrm{y}\right)$. Such values of y must satisfy $\mathrm{m}_1 \leq \mathrm{y} < \mathrm{m}_0$. Thus $\mathrm{m}_0 > \mathrm{m}_1$ or $\Delta\mathrm{m} < 0$.

c. Follows from (C.2) with $\mathrm{U}^*\left(\mathrm{m}\Delta\right) < 0$ implying $\mathrm{m}\Delta < 0$.

Appendix E: Median and Related Treatment Effects in Policymaking Contexts

This appendix reviews several policy contexts within which considerations of median and other quantile measures and treatment effects do (or should) figure prominently.

FDA Effectiveness Criteria for Biological Products

FDA's standards for determining effectiveness of biological products (21 CFR 601.25(d)(2)) were quoted in section 3. Similar regulatory language defining effectiveness governs medical devices (21 CFR 860.7(e)(1)) and over-the-counter drugs (21 CFR 330.10(a)(4)(ii)). These standards rely on two key parameters: a *significant proportion* of the target population being affected, and effect magnitudes that are *clinically significant*.[49]

One perhaps fair interpretation of "clinically significant function" is that the outcome under T_1 is better than the outcome under T_0 by at least the some prespecified amount (say Δ_{CS}). That is, for a given subject $y_1 - y_0 = \Delta \leq \Delta_{CS}$ when smaller outcomes correspond to better health.[50]

Consider the top panel of figure E1. Assume that larger values of y correspond to worse health ($\Delta < 0$ is an improvement in health). Let $\Delta_{CS} < 0$ be the threshold for the *clinically significant* improvement in outcome; as such $\Delta \leq \Delta_{CS}$ is required to demonstrate effectiveness. (Note that smaller—more negative—values of Δ_{CS} correspond to more stringent standards.) Let $\alpha_{SP} \in (0,1)$ be the threshold for the *significant proportion* of the population experiencing the clinically significant improvement; as such $\alpha \geq \alpha_{SP}$ is required to demonstrate effectiveness, where α is the fraction of the population for which $\Delta \leq \Delta_{CS}$. (Note that larger values of α_{SP} correspond to more stringent standards.) Then the two shaded regions in figure E1's top panel indicate the combinations $(\Delta_{CS}, \alpha_{SP})$ corresponding to T_1 being effective or ineffective relative to T_0 given the treatment effect distribution $F(\Delta)$ indicated in the figure.

[49] Cravens, 2002, offers a detailed assessment of the role of "clinical significance" in legal and regulatory contexts, noting: "At its worst, 'clinical significance' is assigned no explicit meaning at all, but simply appears in passing.…it makes sense either that the phrase should be endowed with some specific meaning or that it should not be used at all."

[50] Often policy does not work from a position of equipoise between T_0 and T_1 but rather privileges or gives deference to a status quo (e.g. non-inferiority trials). (See Manski and Tetenov, 2020.)

[figure E1 about here]

In general some $\alpha -$quantile of $F(\Delta)$ determines whether the CFR standard is met. If $\alpha_{SP} = .5$ then the magnitude of $m\Delta$ is determinative. While $\alpha_{SP} = .5$ is not privileged in this regulatory context is nonetheless has some intuitive appeal (though see the case study reported in the final subsection of this appendix).

When $m\Delta$ cannot be point identified appealing to its bounds is partially informative about whether these regulatory standards are met. The bottom panel of figure E1 depicts the regions where it is possible to determine effectiveness or ineffectiveness when only bounds (sharp or otherwise) on the quantiles of the treatment effect distribution $q(\alpha)\Delta$ are available.

Related considerations arise in the determination of *breakthrough therapies*. Suppose smaller values of y represent better health. It might be reasonable in cases like those covered in R3 (section 4) to consider T_1 a breakthrough therapy.[51] Consider FDA policy for expedited approval of drugs that treat serious or life-threating illnesses:

> The Secretary shall, at the request of the sponsor of a drug, expedite the development and review of such drug if the drug is intended, alone or in combination with 1 or more other drugs, to treat a serious or life-threatening disease or condition and preliminary clinical evidence indicates that *the drug may demonstrate substantial improvement over existing therapies on 1 or more clinically significant endpoints*, such as substantial treatment effects observed early in clinical development. (In this section, *such a drug is referred to as a "breakthrough therapy"*.) [21 USC 356(a)(1)] (emphasis added)

As with the determination of effectiveness of biological products, what constitutes "substantial improvement" presumably affects determination of breakthrough therapy status.

Outcome Measurement in ClinicalTrials.gov Registration and Reporting

At the stage when an "applicable" clinical trial is registered, as required, in the *ClinicalTrials.gov* registry the registrant is not required to specify the particular "measure type" they will be estimating as the parameter that summarizes the trial's primary or secondary outcome, e.g. a median, mean, or probability. It is instructive to compare 42 CFR 11.28 ("What constitutes clinical trial registration information?") with 42 CFR 11.48 ("What constitutes clinical trial results information?").

[51] FDA approvals of breakthrough therapies are catalogued at https://www.fda.gov/drugs/nda-and-bla-approvals/breakthrough-therapy-approvals

While recording information on the measure type is not mandated when a trial is registered in *ClinicalTrials.gov*—although it can be provided optionally (Zarin et al., 2016)—this information is required when a trial's results are reported. Whether it is appropriate to report the "measure type" at the initial registration stage or not until results are reported has been debated,[52] but in principle allowing measure types (parameters) to be specified after a trial's data have been realized raises possibilities of cherry picking (Zarin et al., 2011).

What is variously called the "method of aggregation," (Zarin et al., 2011, 2016), the "population-level summary" (EMA, 2020), or the "measure type" (42 CFR 11.48(a)(3)(iii)(E); 42 CFR 11.48(a)(3)(v)(B)(2)(ii)) is what would be known familiarly as a parameter.[53]

> When specifying the Measure Type, the responsible party is required to select one option from the following limited list of options: "Count of participants," "count of units," "number," "mean," "median," "least squares mean," "geometric mean," and "geometric least squares mean." When specifying the associated Measure of Dispersion, the responsible party is required to select one option from the following limited list of options: "Standard deviation," "inter-quartile range," "full range," and "not applicable" (which would be permitted only if the specified measure type is "count of participants," "count of units," or "number"). [U.S. DHHS, 2016, p. 25084]

"median time to response" is noted explicitly as a measure type that may be suitable for time-to-event data (U.S. DHHS, 2016, p. 25087). Though perhaps of little practical importance it is noteworthy that the median is the only quantile recognized in the regulatory language:

> No "other" option is available for either Measure Type or Measure of Dispersion, but responsible parties have the option of voluntarily providing additional information about the baseline measures as part of a freetext description of the baseline measure. [U.S. DHHS, 2016, p. 25084]

A Case Study—Torax Medical, Inc. and the LINX Reflux Management System

In 2010, Torax Medical, Inc., submitted to the FDA a premarket approval application for its LINX Reflux Management System, an "implantable device for the treatment of pathologic Gastroesophageal Reflux Disease (GERD) in patients who continue to have chronic GERD

[52] Zarin et al., 2011: "Some argue that the method of aggregation…is part of the statistical analysis plan and may properly be specified later—after data accrual but before unblinding."

[53] Confusion may arise with the regulatory terminology: "outcome measure type" refers to primary vs. secondary outcomes; "measure type" refers to the parameter; compare CFR 11.48(a)(3)(iii)(D) with CFR 11.48(a)(3)(iii)(E). Moreover the meaning of an "estimand" may also not align with standard usage of that term in econometrics; see EMA, 2020.

symptoms despite antireflux therapy" (U.S. FDA, 2012). The key results supporting the sponsor's PMA application are from a pre-post clinical study reported in Ganz et al., 2013. The primary outcome is the proportion of patients experiencing a reduction of at least 50% in the fraction of a 24-hour period with a pH less than four. The sponsor specified *ex ante* that "effectiveness" would be determined by whether at least 60% of subjects met or bettered the 50% criterion (but see transcript below).

The following are excerpts from the official transcript of the January 11, 2012, meeting of the FDA's Medical Devices Advisory Committee, Gastroenterology and Urology Devices Panel.[54]

> DR. VENKATARAMAN-RAO [FDA Presenter]: ...As shown by this table, 64 out of 100 subjects met the primary effectiveness endpoint. Of these, 56 had pH normalization and 8 had an at least 50% reduction in total time that their pH was less than 4. As the lower limit of the confidence interval was found to be 53.8%, the results for this endpoint do not support the claim that the success probability is greater than 60%.
>
> . . .
>
> DR. SCHROEDER [FDA Presenter]: ...From the statement of the hypothesis given on this slide, it can be seen that the Sponsor was attempting to show that the true success probability associated with this endpoint is greater than 60%. Of the 100 implanted subjects, 64 could be classified as treatment success based on the esophageal pH monitoring endpoint. The observed success rate was 64%, and the 95% confidence interval for this success rate ranged from 54% to 73%. Note that 4 subjects with missing responses, including 3 explants, were treated as failures in this analysis. Since the lower bound of the confidence interval is less than 60%, there is not enough evidence to conclude that the true success rate is greater than 60%, so the primary endpoint for the study was not met.
>
> . . .
>
> DR. AFIFI [Panel Member]: ... My question is about the choice of the 60% figure for the target improvement. How was that arrived at? And was it really sort of a compromise between sample size available or anticipated sample size? And because there could've been other choices. If you could answer that question.
>
> MR. PULLING [Sponsor Advisor]: ...The 60% performance goal was, admittedly, largely arbitrary in choice. We did a careful assessment of the, you know, current feasibility data we had available at the time. We

[54] While the FDA is not required to follow the recommendations of its Advisory Committees, it does so in the great majority of cases.

considered other performance goals and really between 50% and 60%. As you know, had we gone down to 50, sample size requirements would've been much, much less; settled on 60 with a high degree of confidence that even though the bar was set a little higher than we probably needed to, we felt it was something we would've met.

DR. SHAHEEN [Panel Member]: I'd like to first say that, unlike some that may speak later, I'm not particularly -- I don't feel very off-put by the fact that you didn't make the primary outcome variable. It's actually laudable that you used a physiologic outcome variable as opposed to just symptoms, which we know, in almost all of these previous studies, will respond to these kind of therapies. So I think that, especially given that there's nothing magical about the lower 95% confidence interval on 60%, I think that those numbers are fine.

. . .

MS. OLVEY [FDA Advisor]: ...The primary effectiveness endpoint is based on esophageal pH testing at baseline and 12 months. An individual subject was defined as a success if either of the following criteria were met:

- normalization of pH, with normalization defined as a pH < 4 for no more than 4.5% of the monitoring time; or

- reduction of at least 50% in total time that pH < 4 relative to baseline.

For the primary endpoint, the LINX pivotal study required that the success rate be at least 60% as indicated by the lower bound of a 97.5% confidence interval. Sixty-four of the 100 implanted patients achieved success, resulting in an observed success rate of 64% with a 97.5% lower confidence bound of 54%. Since this lower bound falls below 60%, the primary endpoint was not met.

Please discuss whether these data support the effectiveness of the LINX device.

DR. WOODS [Panel Chair]: Okay, so I'm going to move around the table and ask everyone for their answers. If it appears that we're quickly coming to consensus, then we may not need everyone to answer. But we will start with Dr. Shaheen.

DR. SHAHEEN [Panel Member]: Well, clearly by the a priori outcome, the device did not make the primary endpoint. However, it's been discussed here already, the primary endpoint in this situation is fairly arbitrary and could just as easily have been picked to be 50%. I think that the important thing is to look at the totality of the data and to understand that within the field, which is that most previous devices that have attempted to do something similar to this device haven't even used the physiologic primary outcome and have relied primarily on symptom

> control. Given that situation, I think that there is substantial evidence that the device does have efficacy in the setting, although I think we do need to acknowledge it did not meet this primary endpoint.
>
> DR. WOODS [Panel Chair]: Okay. Dr. Inge.
>
> DR. INGE [Panel Member]: I don't disagree. In fact, I think that he's made the exact points that need to be made.
>
> . . .
>
> DR. WOODS [Panel Chair]: Okay. So, Dr. Lerner, the Panel is unanimous in its belief that the data do support the effectiveness of the LINX device. Is that sufficient for you?
>
> DR. LERNER [FDA Representative]: Yes, thank you.
>
> MS. WATERHOUSE: ...We will proceed to Question 2. Question 2 reads as follows: Is there reasonable assurance that the LINX Reflux Management System is effective for use in patients who meet the criteria specified in the proposed indication? Please lock in your votes.
>
> (Panel vote.)
>
> MS. WATERHOUSE [Panel Member, Designated Federal Officer]: The poll is now closed.
>
> ...On Question 2, the Panel members voted unanimously yes, so they voted 9 to 0 that there is reasonable assurance that the LINX Reflux Management System is effective for use in patients who meet the criteria specified in the proposed indications for use.

Note foremost the seemingly arbitrary choice by the sponsor of α_{SP} , the target population fraction parameter.

The FDA followed the recommendations of the Advisory Committee and approved the Linx System for marketing on March 22, 2012. How the data were summarized mattered, and could have mattered much more had the Advisory Committee been convinced by the FDA staff's assessments. (See Mullahy, 2018b, for extended discussion.)

Figure 1 — Time-to-Recovery Outcomes in the Beigel et al., 2020, ACTT Remdesivir Study: Outcomes reported in NIAID, 2020 (top); Reproduction of Beigel's figure 2A (bottom)

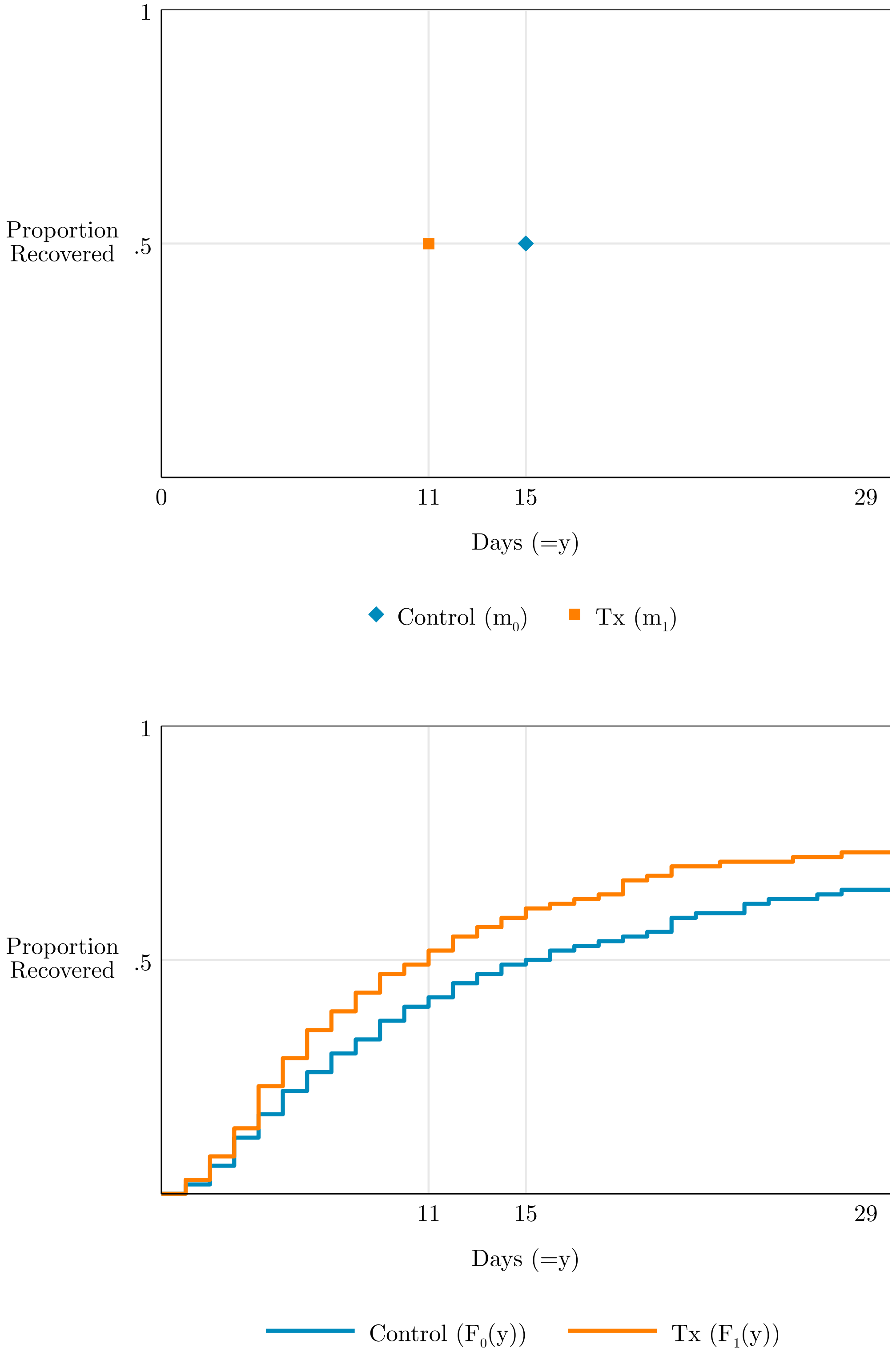


Figure 2 — Time-to-Improvement Outcomes in the Cao et al., 2020, Lopinavir–Ritonavir Study
(Reproduction of Cao's figure 2)

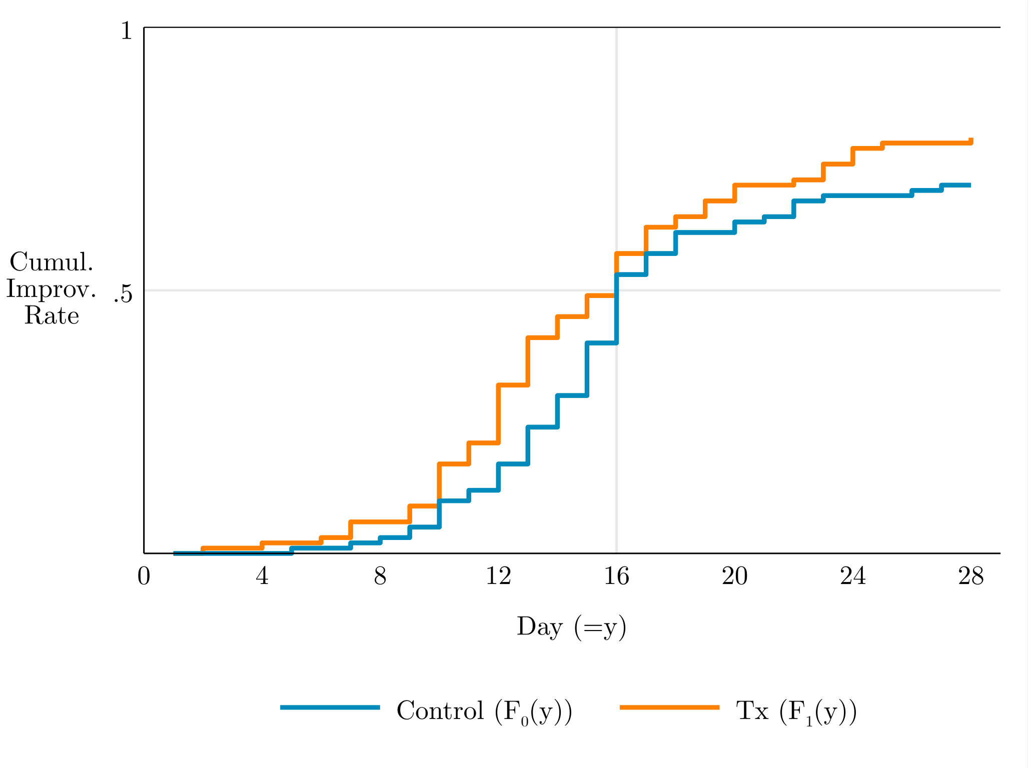


Figure 3 — Defining D_{01} and D_{10}

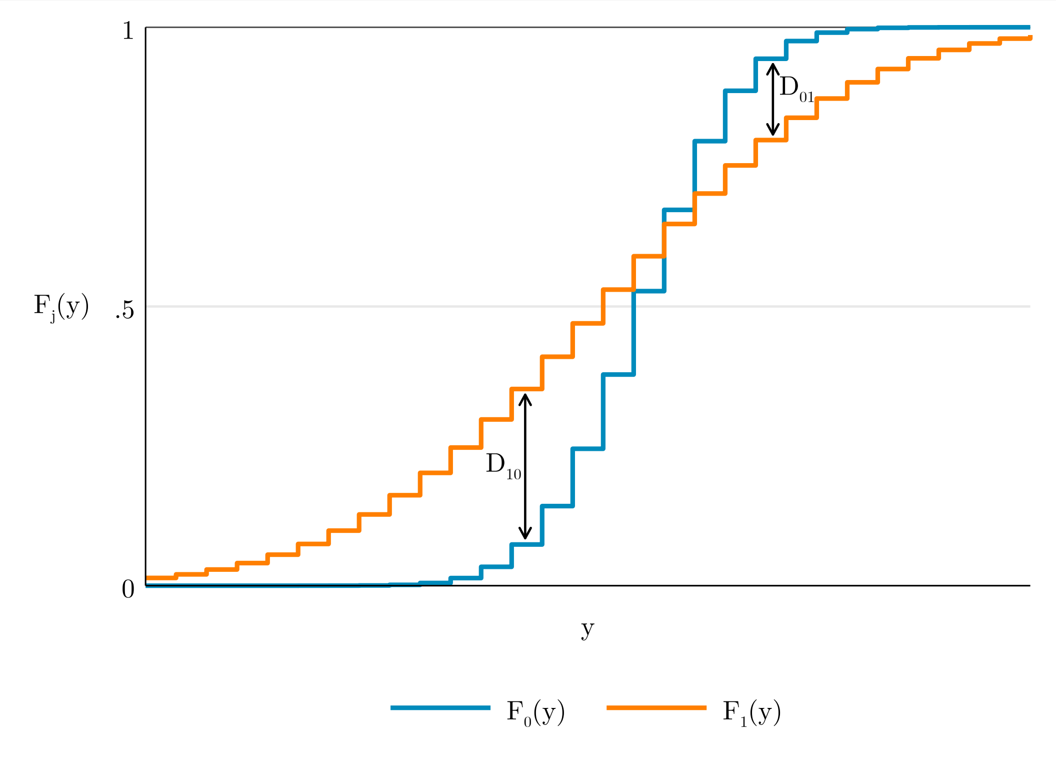


Figure 4 — Defining Δm and $m\Delta$ with Known $\Pr\left(y_0, y_1\right)$

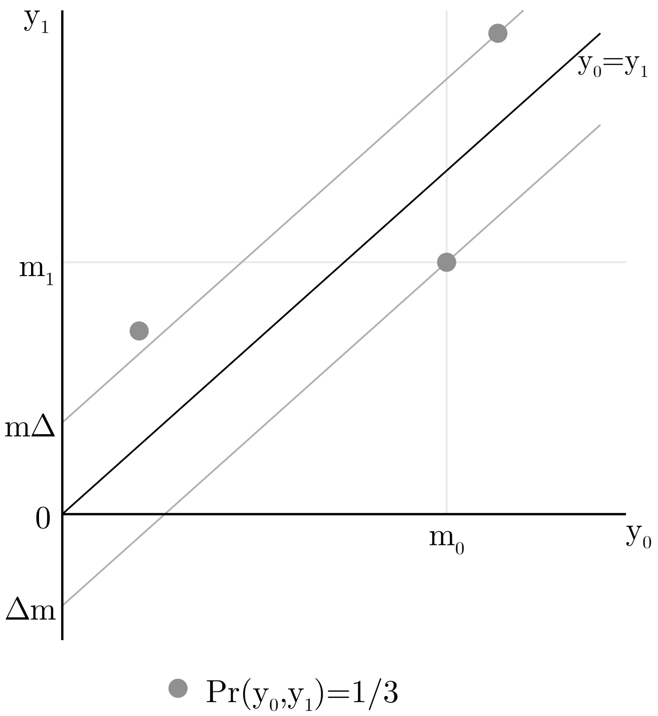


Figure 5 —Identifying Δm and $m\Delta$ with Right-Censored Outcomes:
Three right-censoring scenarios (top); Degree of right censoring affects $m\Delta$ bounds (bottom)

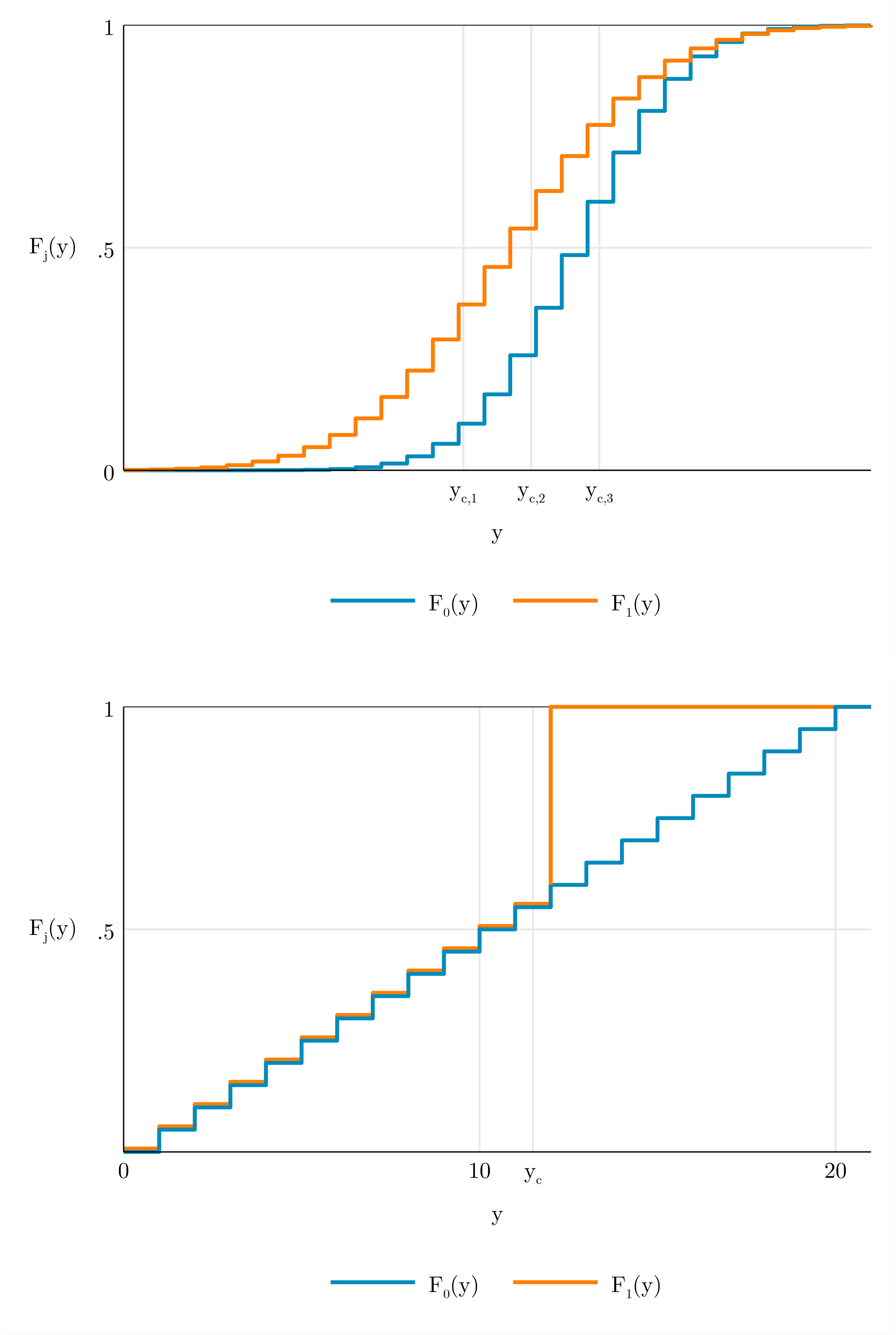


Figure 6 — Time-to-Negative-Test Outcomes in Hung et al., 2020, Combination-Treatment Study (Quartiles and IQR reported in Hung's table 2)

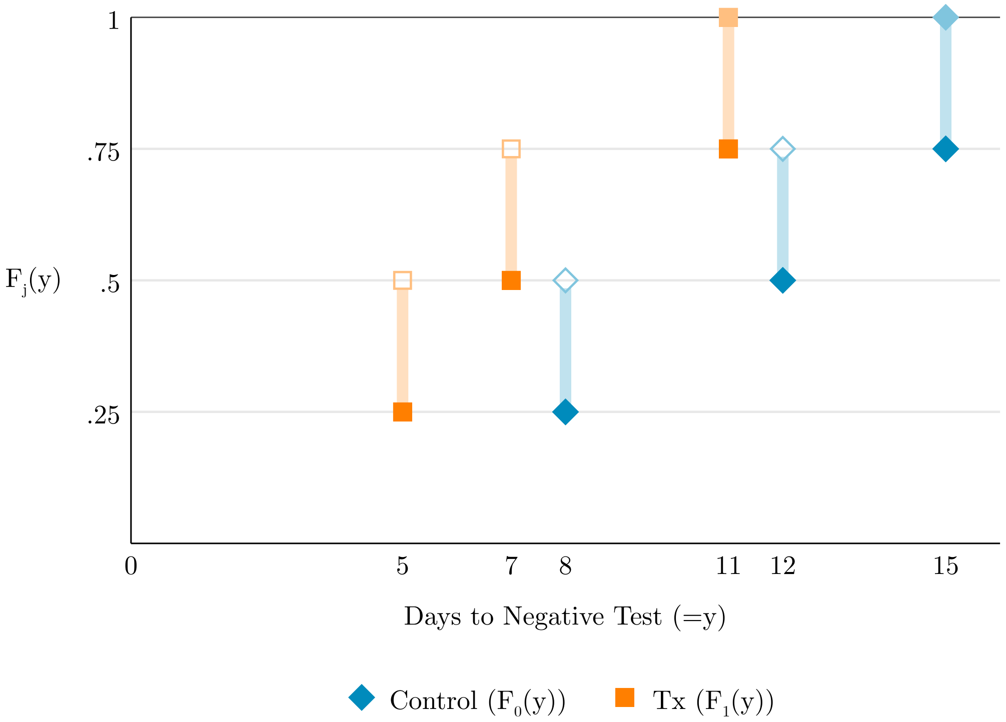


Figure 7 — Comparing Implications of Two Different Joint Distributions

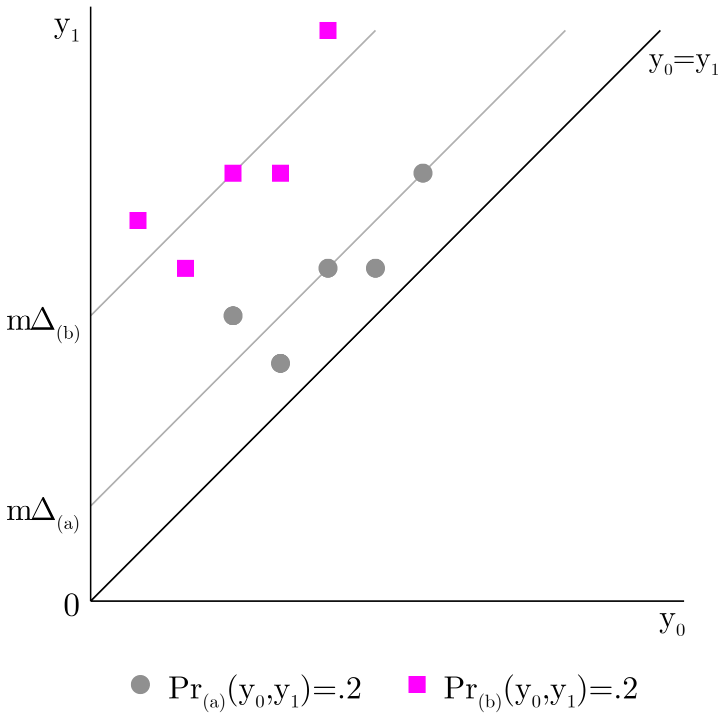


Figure 8 — Approximate Central Aperture, $m_0 = 4$, $m_1 = 2$, $y_j \sim \text{gamma}\left(\gamma_j,\delta_j\right)$ with $\gamma_0 = 4$ and $\gamma_1 = 1$: $a_1\left(m^*,.5\right) = -.70$ (top); $a_2\left(.5\right) = -.64$ (bottom) (Note: $\left(m_j,\gamma_j\right)$ fix δ_j)

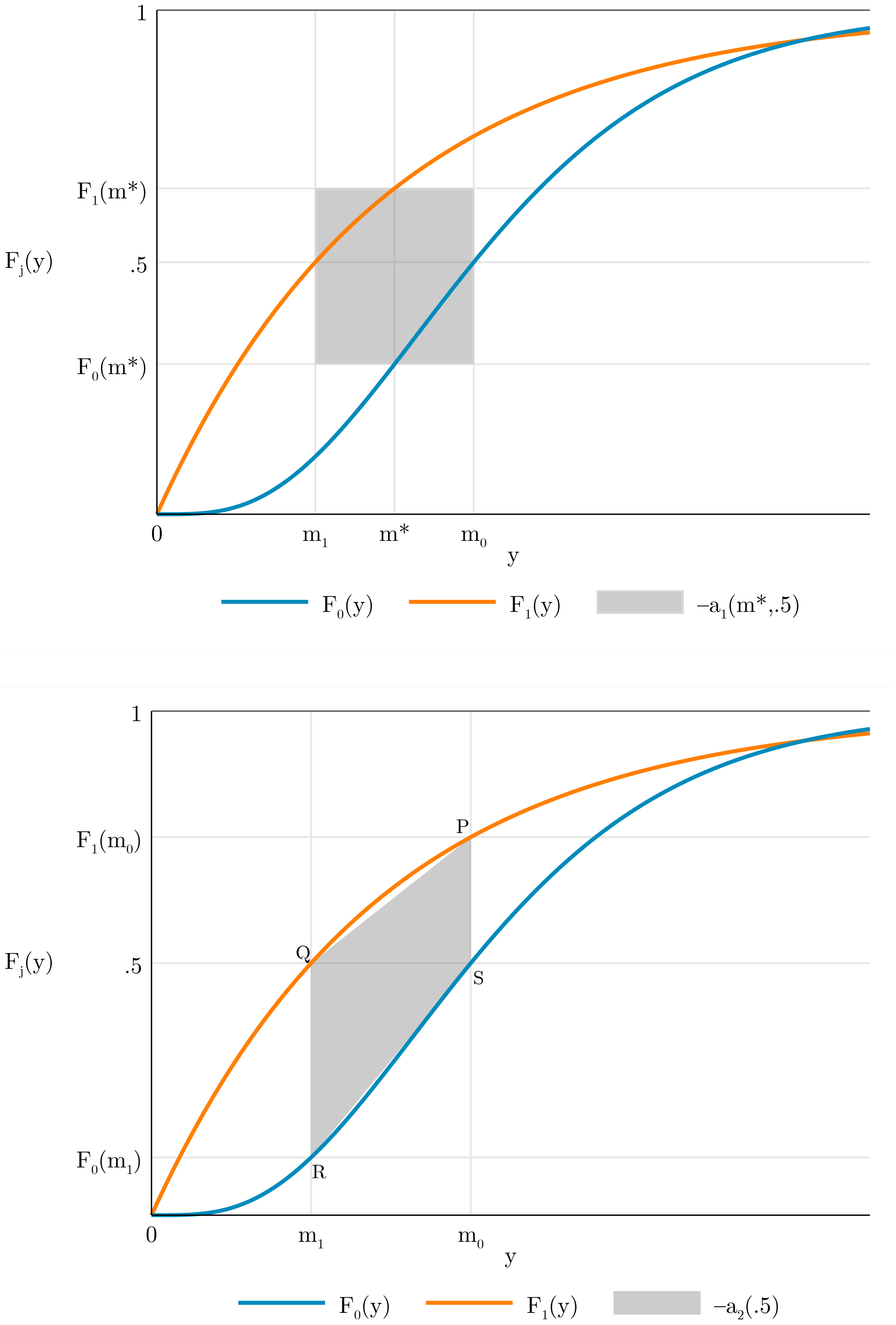


Figure 9 — $a_1\left(m^*,.5\right)$, with $m_0 = 4$, $m_1 = 2$, $y_j \sim \text{gamma}\left(\gamma_j,\delta_j\right)$ (Note: $\left(m_j,\gamma_j\right)$ fix δ_j): $\gamma_0 = .3$, $\gamma_1 = .2$, $a_1\left(m^*,.5\right) = -.16$ (top); $\gamma_0 = .3$, $\gamma_1 = .9$, $a_1\left(m^*,.5\right) = -.35$) (bottom)

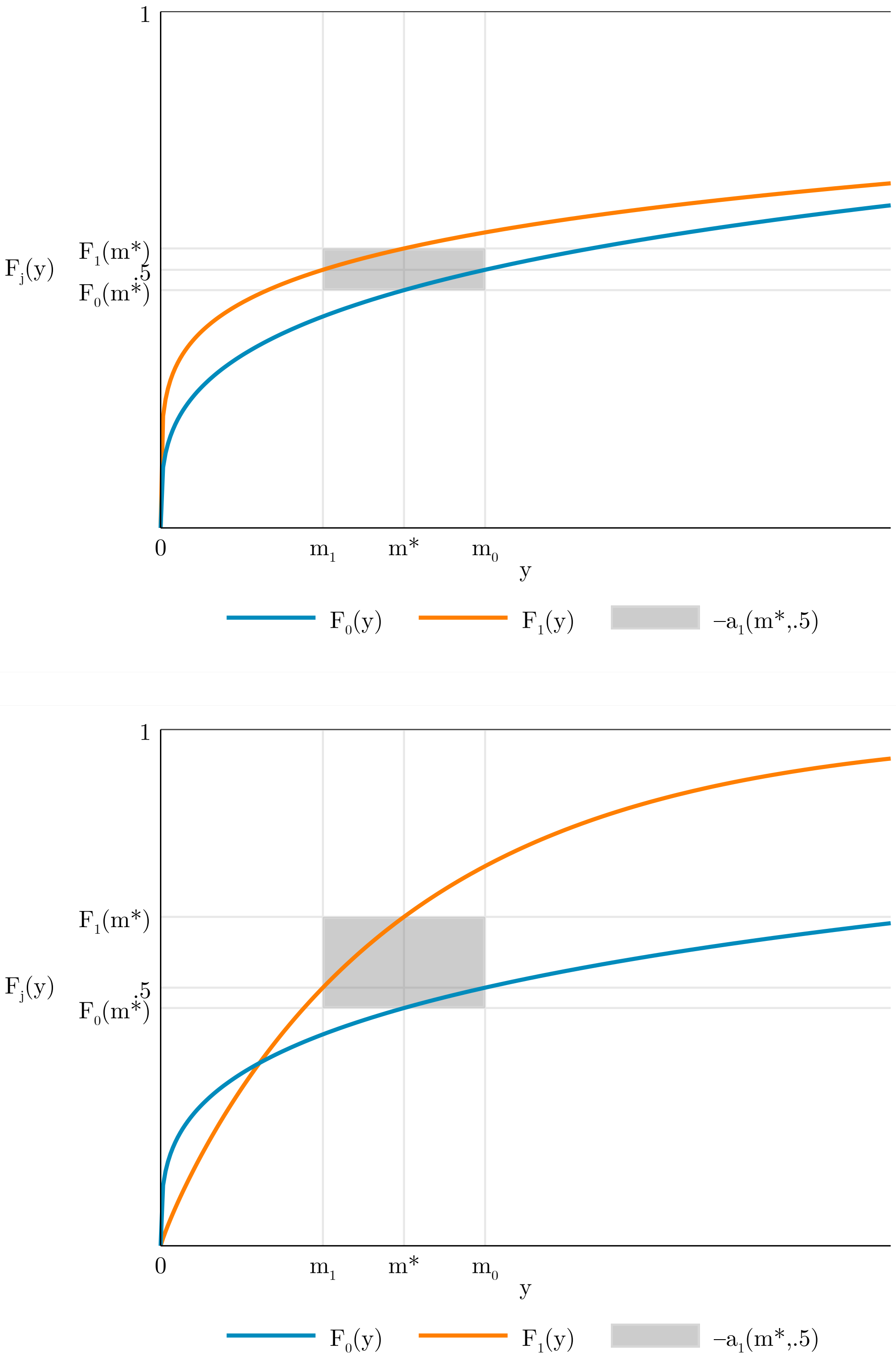


Figure C1 — Computation of $\mathrm{L}\left(\mathrm{MTE}\right)$, Cao et al., 2020, Data:

$\mathrm{F}_1\left(\mathrm{y}-\mathrm{d}\right)$ with $\mathrm{d}=-8$ (top); $\mathrm{F}_1\left(\mathrm{y}-\mathrm{d}\right)$ with $\mathrm{d}=-9$ (bottom)

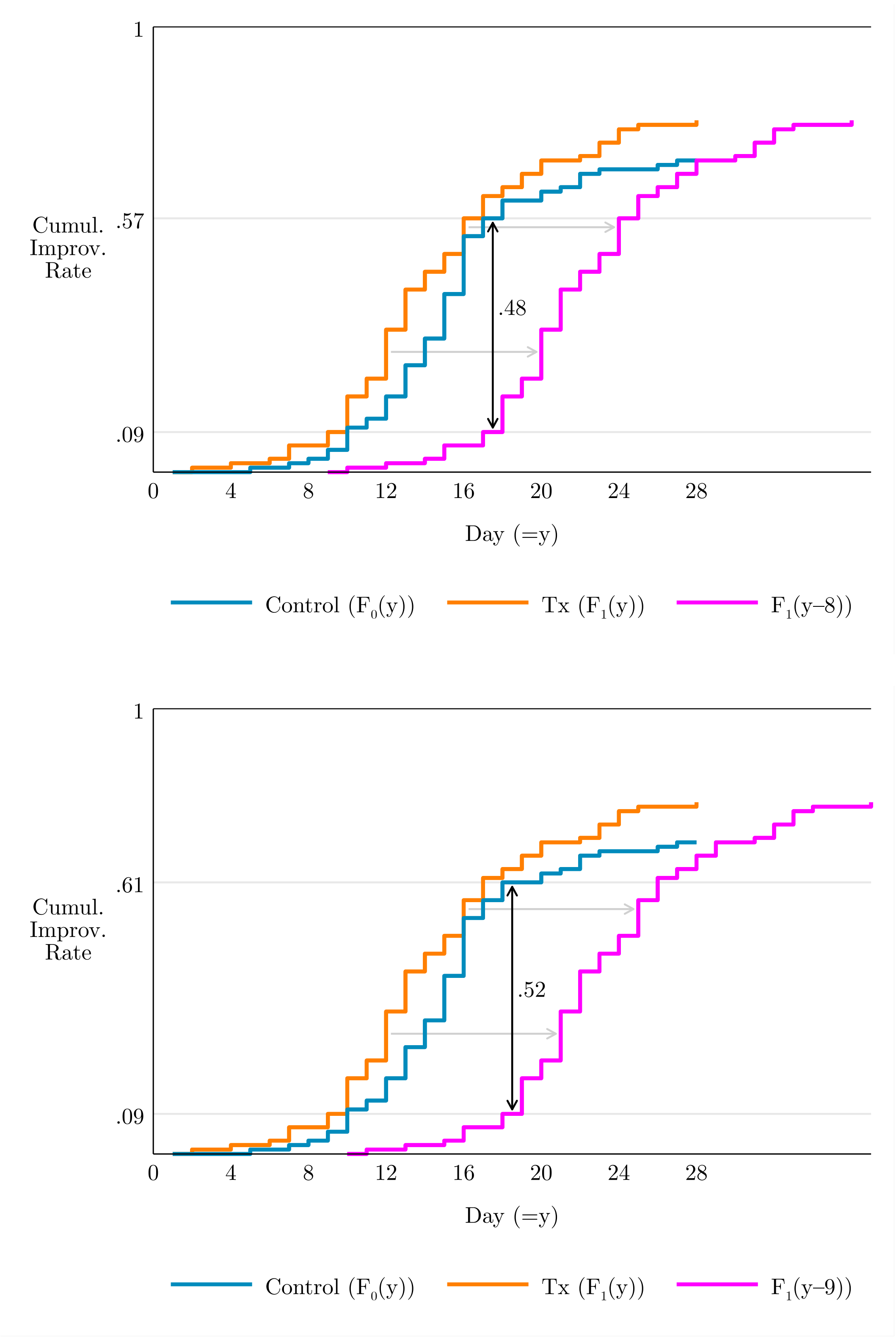


Figure C2 — IQR-Based Computation of $L_{IQR}\left(m\Delta\right)$, Hung et al., 2020, Data

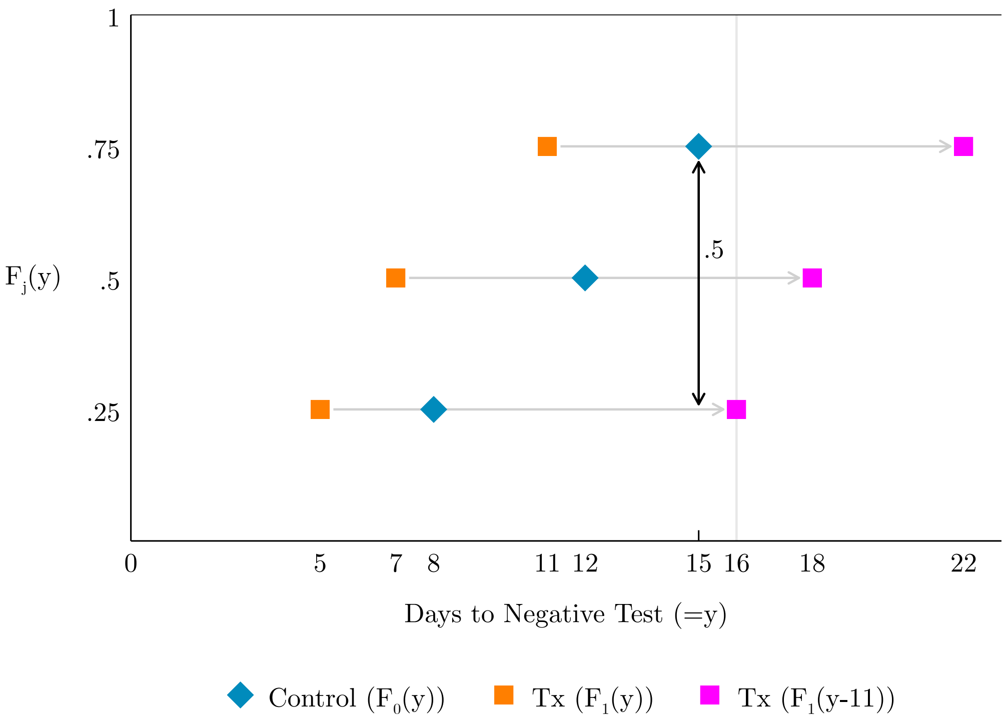
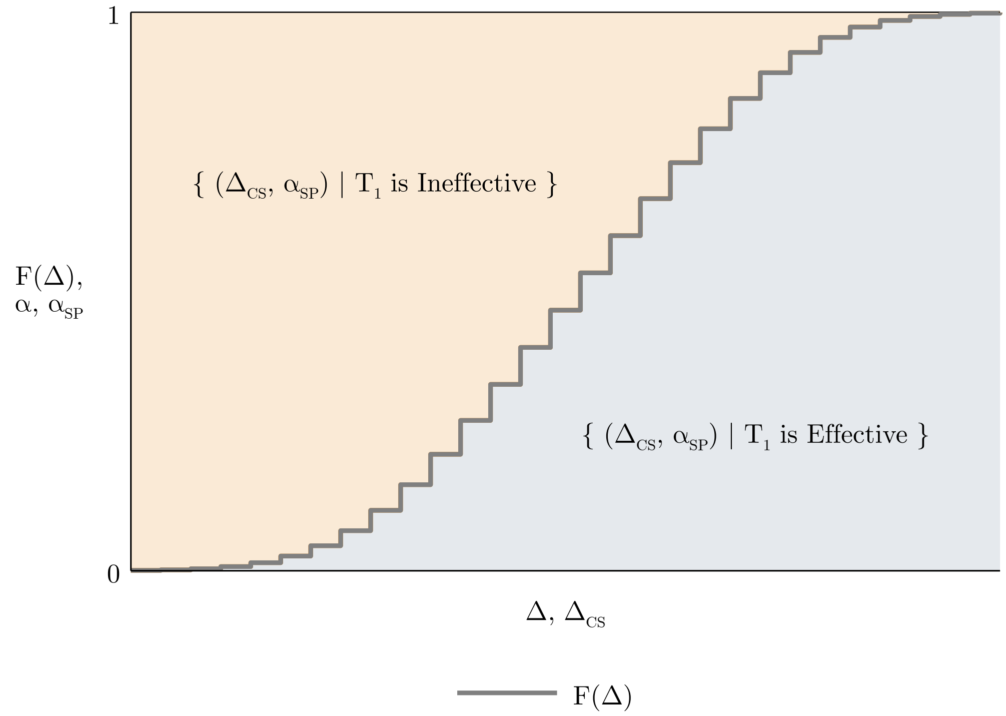


Figure E1 — Visualizing FDA Effectiveness Criteria for Biological Products:
$F\left(\Delta\right)$ and $q\left(\alpha\right)\Delta$ point identified (top); $F\left(\Delta\right)$ and $q\left(\alpha\right)\Delta$ partially identified (bottom)



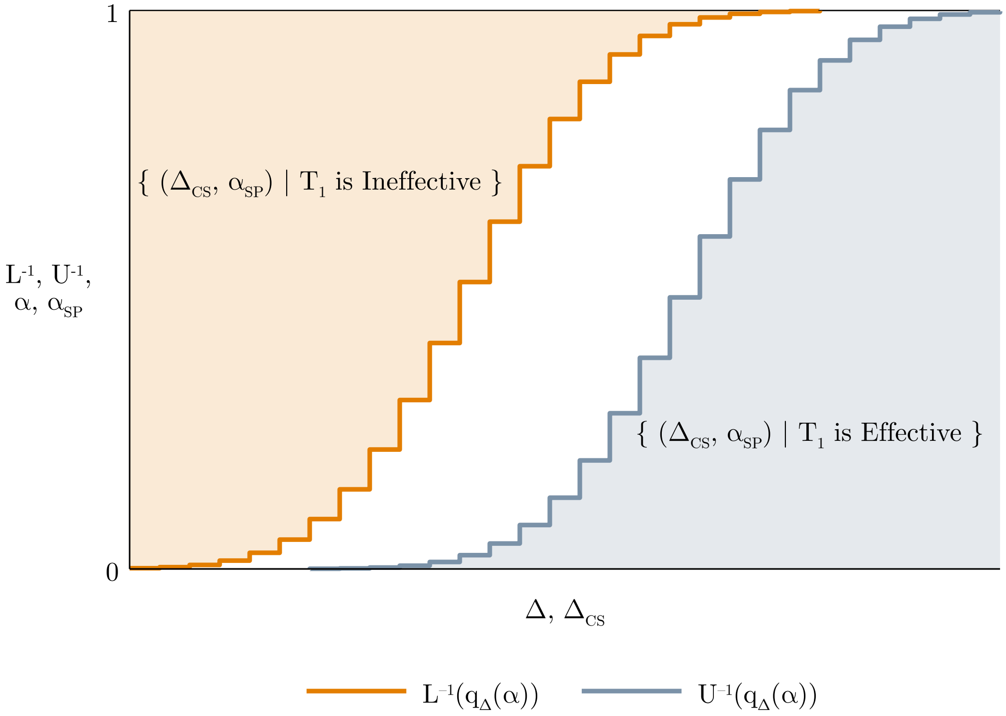


Table 1 — Empirical Results, Beigel et al., 2020, Remdesivir Trial
(Output from medte Stata Program)

```
Median Treatment Effects and Related Parameters

Outcome variable:       ttr
(integer-valued)

Group variable (g):     tx

  Uncensored Obs. for g=0:         340
  Uncensored Obs. for g=1:         391

  Right-censored Obs. for g=0:     181
  Right-censored Obs. for g=1:     147

Median Treatment Effects

  Sample median for g=0 (m0):       15
  Sample median for g=1 (m1):       11

  Diff. in medians m1-m0:           -4

  Lower bound on med(F(y1-y0)):    -15
  Upper bound on med(F(y1-y0)):     10

Central Aperture Measures

  a1(m*,.5):                       -0.3862
  a2(.5):                          -0.3747

Bounds on Inequality Probabilities

  Lower bound on Prob(y1>y0):       0.0000
  Upper bound on Prob(y1>y0):       0.8856
```

Table 2 — Empirical Results, Cao et al., 2020, Lopinavir–Ritonavir Trial
(Output from medte Stata Program)

```
Median Treatment Effects and Related Parameters

Outcome variable:       tti
(integer-valued)

Group variable (g):     tx

  Uncensored Obs. for g=0:          70
  Uncensored Obs. for g=1:          77

  Right-censored Obs. for g=0:      30
  Right-censored Obs. for g=1:      22

Median Treatment Effects

  Sample median for g=0 (m0):       16
  Sample median for g=1 (m1):       16

  Diff. in medians m1-m0:            0

  Lower bound on med(F(y1-y0)):     -9
  Upper bound on med(F(y1-y0)):      7

Central Aperture Measures

  a1(m*,.5):                        0.0000
  a2(.5):                           0.0000

Bounds on Inequality Probabilities

  Lower bound on Prob(y1>y0):       0.0000
  Upper bound on Prob(y1>y0):       0.8259
```

Table A1 —Simulated Relative Frequencies of Differences between Two Median Computation Methods, $.5\times\left(y_{(.5N)}+y_{(.5N+1)}\right)-y_{(.5N)}$ ($y\sim$ Discrete Uniform $\in\left\{0,1,\ldots,100\right\}$; 10,000 replications for each N)

	Sample Size = N			
Difference	20	50	100	1,000
0	.09	.21	.37	.90
.5	.17	.31	.41	.10
1	.14	.19	.15	0
1.5	.11	.12	.05	0
2	.09	.07	.02	0
2.5	.07	.04	.01	0
3	.06	.02	<.005	0
3.5	.05	.01	<.005	0
4	.04	.01	<.005	0
4.5	.03	.01	0	0
≥5	.14	.01	0	0