HCEO WORKING PAPER SERIES

Severity of Illness and the Duration of Intensive Care

Anand Acharya
Lynda Khalaf
Marcel Voia
Myra Yazbeck
David Wensley

Working Paper

2021-003

01/2021

HUMAN CAPITAL AND
ECONOMIC OPPORTUNITY
GLOBAL WORKING GROUP

The University of Chicago
1126 E. 59th Street Box 107
Chicago IL 60637

www.hceconomics.org

Severity of Illness and the Duration of Intensive Care[1]

January 19, 2021

Abstract

Severity of illness scores may introduce or perpetuate bias when used to ration or prioritize intensive care. Using an economic framework that accounts for both demand and supply–side pathways, we find direct physiology to be the relevant driver of intensive care utilization. A deeper implication and key take–away is that (i) including treatments and diagnosis in severity scores provides a channel to perpetuate bias in the triage process and (ii) evidence of this bias is drawn from unobserved patient–level factors working from both demand and supply–side directions. *JEL code: I10*

Anand Acharya, Carleton University (anand.acharya@carleton.ca)
Lynda Khalaf, Carleton University (lynda.khalaf@carleton.ca)
Marcel Voia, Université d'Orléans (marcel.voia@univ-orleans.fr)
Myra Yazbeck, University of Ottawa, (myra.yazbeck@uottawa.ca)
David Wensley, University of British Columbia (dwensley@cw.bc.ca)

[1]Data for this study was collected under the guidance of the Canadian Critical Care Trials Group. Data collection was funded by British Columbia Children's Foundation, Vancouver Foundation, and British Columbia Services Medical Grant. IRB approval was obtained from all participating centers.

I. Introduction

The availability of intensive care unit services providing timely interventions such as mechanical ventilation is a vital component of societal well–being. Intensive care units (ICUs) are an essential, but limited resource which are often highly congested due to the fixed supply of ICU beds and the unpredictable duration of intensive care stay (Sprung et al., 2020; OECD, 2020). ICUs are therefore at risk of being overwhelmed during periods of high demand leading to difficult policy decisions around prioritizing patients and rationing care (Emanuel et al., 2020). One method for rationing and prioritizing care is to triage patients using severity of illness scores which are the *de–facto* clinical markers of the need for intensive care (Christian et al., 2014; Fischkoff et al., 2020; Killien et al., 2020). However, the use of severity of illness scores is complicated by potential biases, presumably driven by unobserved factors – all of which may lead to broad inequities in the allocation of treatment resources (Soto, Martin and Gong, 2013).

In this paper we draw pragmatic insights on this problem by contrasting two intensive care severity scores intended to capture the same underlying health status – the severity of critical illness. Although these two competing scores are in practice used interchangeably, our study reveals differing implications, attributable to the components, methods, and premises by which they are constructed – differences that are similar to a wide range of general severity measures of potential use in triage algorithms. These differences may provide insights on how biases enter the triage process. For example, some severity scores include mechanical ventilation as a weighted component in the score construction. Mechanical ventilation is a treatment choice that in principle is driven by scientifically grounded patient physiology (Tobin, 2019) – hence the rationale for its inclusion in assessing how sick a patient is. However, in practice, this treatment decision appears to be influenced by wider, often unobserved, patient, physician, and systems factors (Bauer et al., 2017). These factors lead to variability in use of mechanical ventilation across geography, income, and ethnicity – a variability that leads to inequities of care (Schmidt, 2020). Thus, severity scores incorporating wider therapeutic markers such as mechanical ventilation (rather than direct patient markers such as oxygen levels) perpetuate these broader biases, and furthermore, potentially entrench circular and

self-confirming treatment regimes (Tobin, 2019; Miller and Toltzis, 2020).

Triage algorithms use severity scores to identify (i) patients most likely to benefit from care (Christian et al., 2014; Fischkoff et al., 2020), and (ii) those likely to consume the most resources (Killien et al., 2020; Gall et al., 2016). An important goal of triage processes is fair and ethical rationing or prioritizing of resources so as to minimize inequities amongst disadvantaged groups (Emanuel et al., 2020). However, implementing these principles in practice is difficult (Galiatsatos et al., 2020). To address this, formulating a new application of market design, Pathak et al. (2020) propose a *priority reserve system* in which a specific quantity of resources is set aside for disadvantaged groups and the remaining resources allocated using severity score–based algorithms. In practice, the size and scope of the reserve depends on the potential inequities introduced by the triage algorithm – the greater the biases embedded in the triage process, the greater the reserve. Therefore, our motivation is that the reserve system is pragmatically optimized when using severity scores that have been themselves evaluated for potential biases[2].

To contrast the two severity scores of focus in this paper, we propose an economic framework in which unobserved patient factors may drive both demand and supply side channels of ICU utilization. Based on this framework, we advance an econometric methodology that accounts for multiple sources of unobserved biases within a model for the duration of intensive care, a central ICU utilization measure in triage (Killien et al., 2020; Miller and Toltzis, 2020; Toltzis et al., 2015). Pivotal to our economic structure and empirical framework is a recognition of latent patient–level health related factors influencing both demand and supply–side pathways driving the duration of intensive care, often persisting despite extensively controlling for observable patient characteristics (Skinner, 2011).

An innovative component of our study is to develop econometric duration methods to disentangle multiple channels (i.e. both demand and supply–side) of endogeneity – at the

[2]For example Pathak et al. (2020) use the SOFA score (Vincent et al., 1996) as an objective physiology–based severity measure. However, SOFA includes mechanical ventilation as a variable in its construction, which may introduce and perpetuate bias as discussed above. In contrast, a score such as APACHE (Zimmerman et al., 2006) does not include mechanical ventilation.

source of severity score related biases. We specify a semi–parametric accelerated failure time duration model, to which we propose an exact, distribution–free rank instrumental variables (IV) inference strategy to account for multiple endogenous covariates with a single instrument (Keiding, Andersen and Klein, 1997; Kalbfleisch and Prentice, 2003; Andrews and Marmer, 2008). Our identification strategy is based on intensive care admissions resulting from accidental events, using pediatric trauma as an instrumental variable. Our simulation based exact–rank inference is straightforward to implement, robust to weak instruments, and accounts for possibly complex convolutions of presumed frailty and baseline duration distributions.

We derive our empirical results by applying our econometric framework to an original multi–center administrative data set of 10,044 Canadian pediatric ICU admissions. This setting is important to our research question for the following reasons. First and perhaps most importantly, treating critically ill children is highly consequential. The returns to timely and effective care have an enormous impact on a child's life trajectory. Economists have established that effective early interventions avoid costly remediation in adulthood, a feature that echoes in this clinical context (e.g. Currie (2020); Hendren and Sprung-Keyser (2020); Heckman (2007); Case, Lubotsky and Paxson (2002)). Second, universal health coverage in Canada implies that patients are not selected or treated based on their insurance coverage or ability to pay, making our results generalizable and not a function of a specific group's access to care. Furthermore, single payer universal coverage controls for the wide variation in costs and utilization that have been shown to be attributable to insurance status or participation (Cooper et al., 2019). Third, since each of the six hospitals in our sample is the only referral ICU in their respective health delivery regions, we do not require strategies to control for hospital selection based on demand preference factors (e.g. Doyle, Graves and Gruber (2019)). However, each hospital and delivery region maintain autonomy over institutional policies, funding, and physician remuneration. This feature retains an important element of our study design, as it allows for regional variation in practice patterns and institutional policies, an important potential supply–side channel that drives health utilization (Chandra and Staiger, 2020; Cutler et al., 2019; Finkelstein, Gentzkow and

Williams, 2016). Finally, the general severity of illness scores in pediatric intensive care are well developed and validated, with one score constructed on physiologic stability (Pollack, Patel and Ruttimann, 1996), and a second based on an outcome propensity of mortality (Slater, Shann and Pearson, 2003). Both scores are in principle considered to be physiology–based and are typical candidates for triage algorithms (Fischkoff et al., 2020; Killien et al., 2020).

The main implications of our study are: (i) direct physiology is the relevant driver of ICU utilization, (ii) including treatments and diagnoses in severity scores provides a channel to perpetuate bias and (iii) evidence of this bias is drawn from unobserved patient–level factors working from both demand and supply–side directions. These findings contribute to the economic literature on the use of clinical practice guidelines versus alternative data synthesis in medical decision making (Currie and MacLeod, 2017; Manski, 2018*b*,*a*; Chandra and Staiger, 2020; Currie and MacLeod, 2020) and also contribute to the extensive literature on risk adjustment in health and wider settings (Ellis, 2011; Einav et al., 2016; Finkelstein et al., 2017). Perhaps more urgently, our results have immediate implications for the timely debate on the appropriateness and nature of severity adjustment in the context of prioritizing and rationing critical care resources (Pathak et al., 2020; Emanuel et al., 2020; Maves et al., 2020; Fischkoff et al., 2020; Killien et al., 2020).

II. Data Measures

Our economic framework examines the relationship between severity of illness and duration of intensive care to study the potential for severity scores to introduce or perpetuate biases when evaluating ICU utilization. We use severity scores as a marker for the exposure; severity of illness, and duration of intensive care (the extensive margin of ICU utilization) as a central and relevant ICU utilization measure in triage (Killien et al., 2020; Miller and Toltzis, 2020; Toltzis et al., 2015).

We analyse observations on 10,044 patients from six academic centers in Canada which provide intensive care services to all children in their respective geographic referral areas. Patient level data includes precise elapsed time (measured in hours) from ICU admission to

ICU discharge (duration of intensive care) along with the Pediatric Risk of Mortality Score (PRISM III) (Pollack, Patel and Ruttimann, 1996) and the Pediatric Index of Mortality (PIM II) (Slater, Shann and Pearson, 2003). Also observed are, age (neonates $<$ 1 month old, infants 1-12 months old, children 12-144 months old, and adolescents $>$ 144 months old), chronic conditions (0,1,2, and $\geq$3), previous ICU admissions (0,1), congenital cardiac conditions (0,1) and whether the admission is the result of a trauma (0,1). There are no missing observations, including all admissions and listed data, during the study period.

Severity scores - The pediatric ICU population is characterized by heterogeneous admission diagnoses ranging from motor vehicle accidents to severe viral infections. However, admission to ICU is not determined by the underlying diagnosis but rather by the severity of the resulting illness. Therefore, there is a need for a consistent and reliable method of classifying illness severity which has resulted in internationally derived and validated severity scoring tools.

Two widely accepted and utilized pediatric illness severity measures are PRISM III (Pollack, Patel and Ruttimann, 1996) and PIM II (Slater, Shann and Pearson, 2003). The PRISM score was developed using over 10,000 admissions from 32 pediatric ICUs in the United States and PIM was developed using over 5000 admissions from 60 pediatric ICUs in the United Kingdom, Australia and New Zealand. While these scores are both intended to measure illness severity[3], the underlying premises behind the two scores reveal some important differences.

Pediatric Risk of Mortality Score – PRISM score is a composite measure that quantifies a child's health status using 16 physiologic variables derived directly from the patient at the time of admission to ICU. The composite score is comprised of (i) cardiovascular (heart rate, blood pressure and temperature), (ii) neurologic (pupillary reactivity and level of consciousness), (iii) respiratory and metabolic (acid-base status, blood oxygen and carbon dioxide levels), (iv) chemical (glucose, potassium, measures of kidney function) and (v) hematologic (white blood cell count, platelet count and blood clotting factor levels) compo-

[3]While it is common, in practice to equate mortality risk and illness severity, the distinction is a topic of debate in the clinical literature (Pollack, 2016).

nents. The PRISM score, although a standalone measurement, is then further combined with a proprietary algorithm to provide a probability of mortality estimate. The standalone score component has a numerical range of 0-71 and is based on multivariate regression analyses restricted to the listed physiologic variables.

The variables used to construct the PRISM score are patient-derived physiologic indications for specific treatments provided in the pediatric ICU. These variables are not subject to observer error or variations in diagnostic intensity as they are obtained directly from patient monitors and laboratory assays. Furthermore, measurement of these variables is not unique to specific populations and therefore less prone to cultural and socioeconomic biases. The PRISM score, however, is based on the magnitude of the deviations of these variables from a physiologically established baseline. Therefore, the definition of these deviations or measures of instability, may nevertheless introduce a subjective component into the PRISM score[4].

Pediatric Index of Mortality – PIM is another score used as a marker of illness severity derived from wider variables determined at the time of admission to ICU. These variables are based on the patient's physiology (blood pressure, acid-base status, oxygen level, pupillary response), treatment (mechanical ventilation), admission category (elective, post-operative, post-operative cardiac) and diagnostic classification (high or low risk). PIM, does not provide a numerical illness severity score but instead provides a probability of mortality derived using a logistic weighting of the listed variables and is expressed as a log-odds.

The variables used to construct PIM include patient-derived physiologic indications for ICU treatments and in addition, therapeutic and diagnostic categories. While a patient's oxygen level is an objective physiologic measurement, the decision to mechanically ventilate a patient is based on numerous, more subjective, factors. These include variation in physician practices (intra and inter-regional) and institutional constraints (e.g. availability of ventilators and trained support staff) leading to varying rates of ventilation for otherwise similar indications. The inclusion of diagnostic categories raises the possibility of variable diagnostic intensity (e.g. those in areas with greater diagnostic intensity may have more health problems diagnosed making them appear sicker relative to those with undiagnosed

[4]This may partly explain the re-calibration of the score over time.

conditions) (Song et al., 2010; Finkelstein et al., 2017). Furthermore, the presence of certain diagnoses are highly dependent on the availability of primary medical care and have been linked to socioeconomic status (Thakur et al., 2013). For example, asthma is a low-risk diagnosis category in the PIM score which serves to decrease the score. A patient with undiagnosed asthma would appear more likely to die (according to PIM) than had they, in the counter-factual, been diagnosed. If the score is used to prioritize care to those less likely to die (Christian et al., 2014; Fischkoff et al., 2020), then an undiagnosed case would be given lower priority. If undiagnosed conditions were more prevalent amongst certain groups, these groups may potentially suffer inequities in care.

Both scores are intended to capture the same underlying health status and are constructed and evaluated on the basis of discriminatory power for (i) mortality prediction and (ii) ICU utilization. However, the scores are not evaluated for confounding (arising from either omitted factors or simultaneity), an issue we address with an economic framework for ICU utilization.

III. Economic Framework

We draw on Finkelstein, Gentzkow and Williams (2016) to specify an economic framework to form the basis for our main estimating equation, where our primary empirical focus is on the relationship between the duration of ICU and the severity of illness. In particular, we specify a framework in which a patient's unobserved health factors may influence both patient demand and physician supply-side channels of the equilibrium ICU utilization.

A country with $g = 1...G$ geographic regions, has a population of $i = 1..N$ children, who differ on latent health ability, $\tilde{\nu}_{ig} \in \mathbb{R}$. All children, conditional on $\tilde{\nu}_{ig}$ have a probability of experiencing a severe health event $\mathbb{P}(a_{ig} = 1|\tilde{\nu}_{ig})$, resulting in intensive care, where the indicator $a_{ig} \in \{0, 1\}$ represents admission to ICU in region g. On admission to ICU, a standardized medical assessment provides a severity of illness marker, $s_{ig}(\tilde{\nu}_{ig}) \in \mathbb{R}$. The marker reflects the nature of the event, severity of exposure, and the child's physiologic response[5] – partly driven by unobserved factors $\tilde{\nu}_{ig}$.

[5]The severity marker is monotone in its arguments, $s'_{ig}(\tilde{\nu}_{ig}) > 0$ and furthermore, $s''_{ig}(\tilde{\nu}_{ig}) \leq 0$ (Hurley, 2000).

The key identifying feature in our analysis is that some health events are entirely accidental, in the sense that the health event occurs irrespective of their health ability, $\mathbb{P}(a_{ig} = 1|\tilde{\nu}_{ig})=\mathbb{P}(a_{ig} = 1)$. Furthermore, within such health events, children with varying degrees of ability are at equal risk. For example, children involved as passengers in motor vehicle accidents appear to be randomly selected in terms of observable characteristics (e.g. age, gender, etc.) and in principal, in terms of unobservable characteristics (e.g. health ability).

Patient Demand – On admission to ICU ($a_{ig} = 1$), the goal is to recover the child's health, which establishes the primary normative objective in the model (Culyer and Wagstaff, 1993; Hurley, 2000). We define a child's prospective quantity of intensive care $y_{ig} \equiv \Lambda(t_{ig}) \in \mathbb{R}$ as a transformation[6] of time, $t_{ig} \in \mathbb{R}^+$ in ICU. A patient's duration of ICU is a composite treatment choice, since it aggregates multiple treatments, invasive interventions, diagnostics, therapeutic, or monitoring decisions.

A child will benefit from a quantity of intensive care that is first and foremost appropriately matched to their severity of exposure – proxied by $s_{ig}(\tilde{\nu}_{ig})$. Second, a child in intensive care will benefit from a quantity of care that is proportional to their underlying health ability $\tilde{\nu}_{ig}$, and observable patient factors $x_{ig} \in \mathbb{R}$. Accordingly, the child's prospective benefit from critical care is,

$$\mathcal{U}(y_{ig}|s_{ig}, x_{ig}, \tilde{\nu}_{ig}) = -\frac{1}{2}(y_{ig} - s_{ig}(\tilde{\nu}_{ig}))^2 + (\tilde{\nu}_{ig} + x_{ig})y_{ig}. \tag{1}$$

Equation (1) provides a parsimonious representation of two distinct features of intensive care. First, the quadratic term captures an intuitively appealing notion of appropriately matching the level of intervention with the need for care, such that $y_{ig} = s_{ig}(\tilde{\nu}_{ig})$ (Wong et al., 2015). Too much or too little care leads to a decreased benefit. Many therapies delivered in intensive care are equally sensitive to over or under use. For example, a critically ill child presenting with a severe infection would require a course of antibiotics. Early withdrawal could result in a rapid escalation to septic shock. Equivalently detrimental, an overly prolonged course could result in antibiotic resistance or organ damage.

The second term in equation (1) captures the complementarity between health care and

[6]The general transformation function, $\Lambda(\cdot)$ is monotone (Ridder, 1990).

health ability. In ICU, this term recognizes that some patients may require a seemingly disproportionate level of care. The optimal quantity of care that maximizes the child's benefit is,

$$y_{ig}^{*} = s_{ig}(\tilde{\nu}_{ig}) + x_{ig} + \tilde{\nu}_{ig}. \tag{2}$$

Therefore, a child's need-based quantity of care is determined by (i) the severity of exposure, (ii) observable characteristics, and (iii) unobserved ability to recover. The unobserved individual level differences impact the demand-side via two pathways. First, a child with frail health ability (i.e. a higher value of $\tilde{\nu}_{ig}$) scores higher on the severity measure and second, would also require a greater amount of residual care for the given level of severity.

Physician Supply – The supply side of the model decomposes the care delivered into (i) the clinical care team's treatment choices and (ii) the institutional constraints faced by the care team. First, the medical care team provides a quantity of care to maximize their perceived benefit to the child,

$$\tilde{\mathcal{U}}(y_{ig}|s_{ig}, x_{ig}, \tilde{\nu}_{ig}) = \mathcal{U}(y_{ig}|s_{ig}, x_{ig}, \tilde{\nu}_{ig}) + \lambda_g(\tilde{\nu}_{ig})y_{ig}, \tag{3}$$

where $\lambda_g(\tilde{\nu}_{ig}) \in \mathbb{R}$, one component of a hospital effect, captures the regional differences in practice patterns. The variability and uncertainty around the efficacy of treatment are driven by unobserved patient characteristics, an important source of group practice differences. Indeed the literature has noted significant regional variation in physician beliefs towards the efficacy of therapy, that "are not highly correlated with demographics, background, and practice characteristics, and are often not consistent with professional guidelines for appropriate care" (Cutler et al., 2019, p. 195).

Institutional Constraints – The clinical care team, mindful of the scarcity of available health resources, chooses a quantity of care,

$$y_{ig}^{*} = \arg\max_{y} \tilde{\mathcal{U}}(y_{ig}|s_{ig}, x_{ig}, \tilde{\nu}_{ig}) - C_g(\tilde{\nu}_{ig})y_{ig}, \tag{4}$$

where $C_g(\tilde{\nu}_{ig}) \in \mathbb{R}^{+}$ reflects their regional institutional setting and constraints, providing a second component of a hospital effect. This represents financial and resource constraints,

which are again possibly influenced by unobserved patient characteristics. The optimal quantity of care provided by the care team is,

$$y^*_{ig} = s_{ig}(\tilde{\nu}_{ig}) + x_{ig} + \tilde{\nu}_{ig} + \lambda_g(\tilde{\nu}_{ig}) - C_g(\tilde{\nu}_{ig}), \tag{5}$$

which forms the theoretical basis for our main estimating equation, where our primary focus is on the relationship between the utilization y^*_{ig} and the exposure $s_{ig}(\tilde{\nu}_{ig})$. The key feature and complication is that $\tilde{\nu}_{ig}$ is unobserved leading to confounding through multiple channels.

IIIa. Econometric Duration

Specification – Model (5) translates to the accelerated failure time (AFT) model (Ridder, 1990; Kalbfleisch and Prentice, 2003),

$$\Lambda(t_{ig}) = \beta s_{ig} + \delta X_{ig} + \epsilon_{ig}, \tag{6}$$

where, $\Lambda(t_{ig}) = ln(t_{ig}) \equiv y_{ig}$ (see also Manning and Mullahy (2001)), s_{ig} is an observed severity score, the k-dimension row vector X_{ig} is the net hospital effect $(\lambda_g(\tilde{\nu}_{ig}) - C_g(\tilde{\nu}_{ig}))$ and the pre-admission controls x_{ig}, and $\beta \in \mathbb{R}, \delta \in \mathbb{R}^k$ are finite dimensional parameters. The unobserved error ϵ_{ig} is the convolution of (η_{ig}, ν_{ig}), where η_{ig} represents the presumed baseline duration with unspecified distribution (Tsiatis, 1990). The unobserved ν_{ig} assumes the role of omitted variable $\tilde{\nu}_{ig}$ from model (5), with unspecified distribution (Keiding, Andersen and Klein, 1997). The fundamental statistical assumption in model (6) is $\{(\epsilon_{ig}, X_{ig}) \; \forall \; i, g\}$ are *exchangeable*[7].

Identification – To address the confounding in model (5), we extend the accelerated life model (6) with an *identification-robust* instrumental variables approach, using a single instrument, $z_{ig} \in \{0, 1\}$, the *trauma* status of a patient. Trauma is an ICU admission category that is the result of moderate to severe insults associated with a range of physiologic instability. Children suffering a trauma do so accidentally and presumably irrespective of their underlying health ability. We assume trauma is related to the duration of ICU solely through severity of exposure. These assumptions are summarized by,

$$\{((z_{ig} = 1), \nu_{ig}, \eta_{ig}) : \mathbb{P}(a_{ig} = 1 | \nu_{ig}, \eta_{ig}) = \mathbb{P}(a_{ig} = 1)\}. \tag{7}$$

[7]This is a weaker assumption to i.i.d. (Young, 2019*a*; Randles and Wolfe, 1979).

Therefore, although trauma patients are not randomized in a strict experimental sense, their admission to ICU occurs in an 'as though' random manner (Rosenbaum, 2010).

Rank Inference – For ease of notation we use the matrix formulation to define the inference function,

$$\mathcal{G}(\beta_o) = c(\beta_o)' \, (p_z) \, c(\beta_o), \tag{8}$$

where the $(n \times n)$ matrix $p_z = z(z'z)^{-1}z'$ and the n-column vector, $c : [0, 1) \to \mathbb{R}$ is a rank preserving non stochastic score. Inverting statistic (8) over β forms a (i) Hodges-Lehman-Sen estimator and (ii) size-controlled confidence set for the coefficient associated with severity in model (6).

Statistic (8) is a rank test of the sharp null $H_o : \beta = \beta_0$ in model (6), which implies testing $H_o : \gamma = 0$ in the transformed regression,

$$W = Z\gamma + \mu, \tag{9}$$

where, $W(\beta_o) = \Lambda(t) - s\beta_o - X\hat{\delta}(\beta_o)$, is the n-column vector of aligned residuals about the hypothesized β, and $\hat{\delta} = (X'X)^{-1}X'(\Lambda(t) - s\beta_o)$ is the null restricted least squares estimator of δ. Andrews and Marmer (2008) developed this statistic as the rank analogue of Anderson and Rubin (1949), which we extend to the duration framework[8] via scores which account for possibly complicated convolution distributions of ϵ_{ig}.

Two asymptotically equivalent scores for (8) are the quantile F_o scores and the expected value F_o scores (Randles and Wolfe, 1979),

$$c^{(i)} = F_o^{-1}\left(\frac{(i)}{(n+1)}\right), \quad c^{*(i)} = E_{F_o}[V^{(i)}],$$

where $V^{(i)}$ is the ith order statistic, $O = [w_{(1)} \leq \cdot \leq w_{(n)}]$, (i) is the ith rank label $r = [(1), \cdot, (n)]$, and F_o is a presumed distribution function. For a well know historic example, if (6) is presumed log-normal, the quantile F_o and expected value F_o scores follow van der Waerden (1953) and Fisher and Yates (1938), respectively. Irrespective of the choice of scores, Andrews and Marmer (2008) show that inference based on statistic (8) is both (i)

[8] Although we adapt this inference strategy to censored data (Prentice, 1978), we do not require this in our analysis, as all durations are fully observed.

exact and (ii) distribution free. The choice of quantile scores offers a tractable method for dealing with complex convolutions ϵ_{ig}. Under mild regularity conditions (Hyndman and Fan, 1996), an empirical quantile function (drawn from a simulated convolution distribution or a split sample) forms the basis of a score for an exact and distribution–free inference using statistic (8).

A valid confidence set $\mathcal{C}_\beta(\alpha) = [\beta_o : \mathcal{G}(\beta_o) < g_{calc}(\alpha)]$, is constructed as follows,

1. Calculate the critical value $g_{calc}(\alpha)$ drawing the n-vector u_l from the uniform distribution and replacing the random variate values with the rank labels. Apply the desired score, $c^{(i)}$ to the rank labels and construct the statistic (8), saving the resulting value as a single element of an m-vector. Repeat this for m times and order the resulting m-vector. Select $g_{calc}(\alpha)$ as the element corresponding to the desired significance level from the ordered m-vector.

2. Calculate value of statistic $\mathcal{G}(\beta_o)$ for a chosen β_o value by replacing the realized variate values of W with the associated rank labels. Apply the desired score, $c^{(i)}$ to the rank labels and construct the statistic (8) saving the resulting value as a single element in the grid search over β_o.

3. Retain all β_o not rejected, that is, satisfying $\mathcal{G}(\beta_o) < g_{calc}(\alpha)$. The set of all hypothesized values not rejected form the confidence set $\mathcal{C}_\beta(\alpha)$.

4. The Hodges-Lehman-Sen estimator is the value (or set of values) of β_o that minimizes $\mathcal{G}(\beta_o)$.

Exchangeability of the aligned residuals $W(\beta_o)$ (which follows from model (6)), combined with the assumptions on the instrument admit (i) arbitrary forms of conditional heteroskedasticity, (ii) correlation between X_{ig}, ϵ_{ig}, and (iii) convolutions of unspecified baseline survival and heterogeneity distributions. Furthermore, $\mathcal{G}(\beta_o)$ does not require variance estimation. These notable properties permit key insights in our analysis, that overcome otherwise non-trivial inferential challenges.

IV. Results and Discussion

We begin with a brief graphical exploratory analysis and then contrast the two severity markers with two sets of analyses, (i) ignoring and (ii) accounting for endogeneity, that are otherwise similarly specified as detailed in model (6) for each score,

$$\ln(t_{ig}) = \delta_o(Hospital_g) + \beta(PRISM_{ig}) + \delta X_{ig} + \epsilon_{ig}, \tag{10}$$

$$\ln(t_{ig}) = \delta_o(Hospital_g) + \beta(PIM_{ig}) + \delta X_{ig} + \epsilon_{ig}, \tag{11}$$

where our primary interest is on the coefficient β which captures the exposure effect-size. The vector X_{ig} are the *pre-admission* covariates, age (neonates $<$ 1 month old, infants 1-12 months old, children 12-144 months old, and adolescents $>$ 144 months old), chronic conditions (0,1,2, and $\geq$3), previous ICU admission (0,1), and congenital cardiac conditions (0,1). These factors are observable controls for unobserved health ability.

Regression (10) does not directly include treatments or diagnosis received after admission to ICU. Furthermore, since these factors are not indirectly included in the PRISM score, they are captured by the hospital effect. Regression (11) also does not directly include treatments or diagnosis received after admission to ICU. However, in contrast to (10), the PIM score, as previously described, includes various treatments and diagnoses, removing a portion of this effect from the hospital effect. Both regressions include, although differing portions of, physiologic based indications for treatment via the severity scores.

Exploratory analysis of duration of ICU and PRISM – Since the primary role of critical care is the monitoring and maintenance of physiology stability, severity of illness is a plausible driver of duration of intensive care. Figure 1 suggests an association between increasing severity of illness (PRISM scores) and increasing duration of intensive care. To further investigate this relationship, the Kaplan-Meier function reveals near uniform increases in duration of intensive care per severity category at all points of the distribution (Figure 2). Interestingly, although not a specification test *per se*, proportional shifts on the time scale suggest general agreement with a time ratio motivating an accelerated time metric (Kalbfleisch and Prentice, 2003).

Background analyses ignoring endogeneity – The first set of analyses begins with a standard likelihood inference for the Weibull model (ignoring ν_{ig}) that serves as a reference, comparable to those in the literature (Pollack et al., 2018; Straney et al., 2010). Next, we add a Gamma frailty specification, again for comparability to the literature and as a first approach to the unobserved factors ν_{ig}, in the absence of endogeneity. Using the data described in Section I, the point estimate and 95% confidence intervals for the PRISM coefficient presented in row 1, column 2 of Table 1 closely match results for a similar regression in the literature (Pollack et al., 2018). A standard Gamma frailty modeling does not materially change these results (Row 2, Column 2). In contrast to the descriptive graphical analysis of Figure 1, in the regression context, although statistically significant, PRISM has a small exposure effect-size. In contrast, the point estimate and 95% confidence intervals for the PIM coefficients presented in row 1, column 3 of Table 1 suggest a moderate role in the severity-duration relationship with a notable increase in exposure effect-size when compared to the PRISM results. This may reflect the wider scope of the PIM score, particularly the possibly endogenous explanatory component of certain admission criteria or treatment choices for duration of ICU. We find evidence of mild unobserved heterogeneity with the Gamma frailty modeling indicating an approximate 10% reduced effect (Row 2, Column 3). These results serve as reference for the remainder of our central analysis.

Central analyses accounting for endogeneity – Informed by an economic model and utilizing robust rank IV duration methods, our central analysis relaxes a key distributional assumption, both baseline and frailty, relative to the above reference results. Further, and perhaps more importantly, we also relax the exogeneity assumption on the marker for illness severity. Following our economic model, we also specify a possibly confounded hospital effect and other pre–exposure controls; age, chronic conditions, previous ICU admission, and cardiac surgery. As noted, all of these controls are permitted to be correlated with unmeasured and unobserved individual factors. Our instrumental variable is the trauma status of each patient, $z_{ig} \in \{0, 1\}$. Figure 4 is a descriptive graphical summary broadly revealing that the binary instrumental variable Trauma has comparable support over both left (duration) and right-side (severity score) variables. Furthermore, the histograms do not show any ev-

idence of differences in clusters, outliers or shape, despite the small sample size of Trauma – suggesting that our results are likely not driven by problematic issues of leverage (Young, 2019*a*,*b*).

The size-controlled confidence sets and Hodges-Lehman-Sen estimates for the PRISM coefficient, reported in row 4, column 2 of Table 1 are different in multiple dimensions from the reference results.

First, accounting for confounding leads to an *increased* exposure effect. A plausible mechanism explaining this result is as follows. Suppose the unobserved health status of the patient is masking an underlying vulnerability that if observed would be associated with an increased length of stay. If the PRISM score did not capture this vulnerability then we would expect to see the above result if PRISM were negatively correlated with the unobserved factor. Using an example of mechanical ventilation demonstrates a pathway for this to occur. Consider two patients presenting with similar respiratory indications at the threshold that would indicate mechanical ventilation, but differing in underlying vulnerability. It is more likely that the patient with greater vulnerability would be mechanically ventilated. Therefore, for all ventilated patients, it is more likely that those with greater underlying vulnerability would have lower PRISM scores as compared to otherwise similar patients with less underlying vulnerability. If this were the case, PRISM would be negatively correlated with the unobserved patient factors.

Second, there is a three–fold change in the magnitude of the exposure effect size, with no overlap in the confidence sets. A change of this magnitude would likely have a material impact on most forms of severity-adjustment. Furthermore, since the results are identification-robust size controlled confidence sets, there is additional information from the form of the confidence sets. In this case, the closed sets indicate an informative instrument and reflect the compatibility of the data with model specification. The increased length of the confidence sets, reflect the (i) inherent sampling variability in the data (common to all analysis), and (ii) the informative content of the instrument, here likely reflecting the small sample size of those instrumented.

Row 4, column 3 of Table 1 presents the PIM rank IV results for an otherwise identically

specified analysis. Again the size controlled confidence sets reflect the informative content of the instrument along with the sampling variability. The 95% confidence sets for the PIM coefficient have some overlap with the reference sets in row 1 and 2. However, there is now an approximately two-fold *decrease* in exposure effect size. This correction is in the opposite direction from the PRISM results. It is instructive to consider the mechanisms that could result in a decrease in the attributable effect of PIM in the severity-duration relationship.

The PIM score includes a wider number of admission factors in its construction. Many of these factors are included on the basis of improving explanatory or predictive criteria in a propensity outcome model for mortality. Contrasting the reference analysis between PRISM and PIM, when utilized as a surrogate marker in the related but different analysis of the severity–duration relationship, these additional factors seem to offer improved explanatory power using PIM. However, a closer look at the PIM variables suggests possible sources of endogeneity. For example, the inclusion of mechanical ventilation increases the PIM score, implying increased severity. However, increased severity implies an increased chance of receiving mechanical ventilation as a treatment. This possibility of a reversal of pathways would suggest the potential for endogeneity. The reported confidence sets would be in agreement with this type of explanation.

Moreover, following up on the mechanical ventilation example from above, a patient with a greater underlying vulnerability was more likely to be ventilated as compared to an otherwise similar patient. If this were the case, the unobserved patient factor would be positively correlated with the PIM score, because mechanical ventilation is a direct variable in the score.

Analyses partially accounting for endogeneity – The generalized Anderson–Rubin method[9] provides an intermediate least–squares analysis to investigate the magnitude of exposure effect-size differences between the reference analysis the rank inference. We consider this an intermediate analysis in the sense that although this least–squares inference shares desirable identification properties with the rank inference, a crucial difference in our context is that

[9]See the on–line appendix for derivations and proofs that generalize the Anderson–Rubin method to the parametric duration case.

it does not account for endogenous hospital effects. This is an important component to our economic model, providing a second channel for endogeneity, for which instruments are lacking. Row three of Table 1 present the 95% size controlled confidence sets for PRISM and PIM in the Weibull parametric accelerated failure time specification.

The intermediate results for PRISM are notably different from the rank IV results, but mildly different for PIM. Including wider admission factors, the PIM score possibly absorbs the severity related treatment and diagnostic heterogeneity that would otherwise be captured by the hospital effect. If these factors are confounded, as our economic model implies, then the associated endogeneity would be accounted for in a correction for PIM alone. The PIM results would reflect this logic, whereas, the intermediate PRISM results suggest that severity related treatment and diagnostic heterogeneity are possibly captured in an endogenous hospital effect. Using the logic of the mechanical ventilation example, this would explain the difference between the intermediate and central results.

Taken together, the *prima facia* implications of our results is that direct physiology is the relevant driver of ICU utilization. The deeper implications and key take-away of our paper is that (i) including treatments and diagnosis in severity scores perpetuate bias and (ii) evidence of this bias is drawn from unobserved factors working through both demand and supply side channels.

V. Conclusion

In this paper we have investigated the role of severity scores introducing or perpetuating bias in evaluating ICU outcomes. We describe an economic framework in which uncertainty around a patient's unobserved health ability may drive both demand and supply side channels of ICU utilization, an outcome relevant to the rationing context. Under this framework, unobserved health ability leads to possible confounding of observed severity markers.

By contrasting two severity scores, intended to capture the same underlying health status, we find evidence of confounding rooted in treatment choices and diagnosis. Because many treatments and diagnosis have been shown to vary with socio-economic status, geography, race, and ethnicity, their inclusion perpetuates the upstream choices that may embed bias in

downstream rationing of ICU resources.

These results have a number of implications. First, not all severity scores are created equal. If the goal of fair and ethical rationing or prioritizing is to minimize inequities amongst certain groups then severity scores should also reflect this. We have demonstrated that inclusion of treatments or diagnosis in score construction, may violate these principals. Remaining bias reflected in objective physiology may then be addressed with the reserve system (Pathak et al., 2020). Finally, our results, derived and viewed through the lens of economists is highly consistent and supportive of following evidence based practice where sound physiology is the relevant driver of valuable ICU resources.

'*Never before in 45 years of active practice have I witnessed physicians coping with inadequate medical resources - specifically a shortage of ventilators. Given this situation, it is pivotal that caregivers have the requisite knowledge to interpret arterial oxygenation scientifically, know when to institute mechanical ventilation, and equally know how to remove the ventilator expeditiously to make it available for the next patient.*' (Tobin, 2020, p. 1320)

References

Anderson, Theodore W, and Herman Rubin. 1949. "Estimation of the parameters of a single equation in a complete system of stochastic equations." *The Annals of Mathematical Statistics*, 20(1): 46–63.

Andrews, Donald WK, and Vadim Marmer. 2008. "Exactly distribution-free inference in instrumental variables regression with possibly weak instruments." *Journal of Econometrics*, 142(1): 183–200.

Bauer, Philippe R, Ashok Kumbamu, Michael E Wilson, Jasleen K Pannu, Jason S Egginton, Rahul Kashyap, and Ognjen Gajic. 2017. "Timing of intubation in acute respiratory failure associated with sepsis: a mixed methods study." *Mayo Clinic Proceedings*, 92(10): 1502–1510.

Case, Anne, Darren Lubotsky, and Christina Paxson. 2002. "Economic status and health in childhood: The origins of the gradient." *American Economic Review*, 92(5): 1308–1334.

Chandra, Amitabh, and Douglas O Staiger. 2020. "Identifying Sources of Inefficiency in Healthcare." *The Quarterly Journal of Economics*, 135(2): 785–843.

Christian, Michael D, Charles L Sprung, Mary A King, Jeffrey R Dichter, Niranjan Kissoon, Asha V Devereaux, and Charles D Gomersall. 2014. "Triage: care of the critically ill and injured during pandemics and disasters: CHEST consensus statement." *Chest*, 146(4): e61S–e74S.

Cooper, Zack, Stuart V Craig, Martin Gaynor, and John Van Reenen. 2019. "The price aint right? Hospital prices and health spending on the privately insured." *The Quarterly Journal of Economics*, 134(1): 51–107.

Culyer, Anthony J, and Adam Wagstaff. 1993. "Equity and equality in health and health care." *Journal of Health Economics*, 12(4): 431–457.

Currie, Janet. 2020. "Child health as human capital." *Health Economics*, 29(4): 452–463.

Currie, Janet M, and W Bentley MacLeod. 2017. "Diagnosing expertise: Human capital, decision making, and performance among physicians." *Journal of labor economics*, 35(1): 1–43.

Currie, Janet M, and W Bentley MacLeod. 2020. "Understanding Doctor Decision Making: The Case of Depression Treatment." *Econometrica*, 88(3): 847–878.

Cutler, David, Jonathan S Skinner, Ariel Dora Stern, and David Wennberg. 2019. "Physician beliefs and patient preferences: a new look at regional variation in health care spending." *American Economic Journal: Economic Policy*, 11(1): 192–221.

Doyle, Joseph, John Graves, and Jonathan Gruber. 2019. "Evaluating measures of hospital quality: Evidence from ambulance referral patterns." *Review of Economics and Statistics*, 101(5): 841–852.

Dufour, Jean-Marie. 1997. "Some impossibility theorems in econometrics with applications to structural and dynamic models." *Econometrica*, 1365–1387.

Dufour, Jean-Marie, and Mohamed Taamouti. 2005. "Projection-based statistical inference in linear structural models with possibly weak instruments." *Econometrica*, 73(4): 1351–1365.

Einav, Liran, Amy Finkelstein, Raymond Kluender, and Paul Schrimpf. 2016. "Beyond statistics: the economic content of risk scores." *American Economic Journal: Applied Economics*, 8(2): 195–224.

Ellis, Randall P. 2011. "Risk adjustment." *The New Palgrave Dictionary of Economics*.

Emanuel, EJ, G Persad, R Upshur, B Thome, M Parker, A Glickman, C Zhang, C Boyle, M Smith, and JP Phillips. 2020. "Fair Allocation of Scarce Medical Resources in the Time of Covid-19." *The New England Journal of Medicine*, 382(21): 2049–2066.

Finkelstein, Amy, Matthew Gentzkow, and Heidi Williams. 2016. "Sources of geographic variation in health care: Evidence from patient migration." *The Quarterly Journal of Economics*, 131(4): 1681–1726.

Finkelstein, Amy, Matthew Gentzkow, Peter Hull, and Heidi Williams. 2017. "Adjusting risk adjustment–accounting for variation in diagnostic intensity." *The New England Journal of Medicine*, 376(7): 608.

Fischkoff, Katherine, Mary Faith Marshall, Regina Okhuysen-Cawley, Preeti John, Sabrina Derrington, Denise Dudzinski, Maria Susan Gaeta, Robert J. Walter, Chinyere OConnor, and Jessica Turnbull. 2020. "Society of Critical Care Medicine crisis standard of care recommendations for triaging critical resources during the COVID-19 pandemic." *Society of Critical Care Medicine*.

Fisher, Ronald A, and Frank Yates. 1938. *Statistical tables: For biological, agricultural and medical research.* Oliver and Boyd.

Galiatsatos, Panagis, Allen Kachalia, Harolyn ME Belcher, Mark T Hughes, Jeffrey Kahn, Cynda H Rushton, Jose I Suarez, Lee Daugherty Biddison, and Sherita H Golden. 2020. "Health equity and distributive justice considerations in critical care resource allocation." *The Lancet Respiratory Medicine*, 8(8): 758–760.

Gall, Christine, Randall Wetzel, Alexander Kolker, Robert K Kanter, and Philip Toltzis. 2016. "Pediatric triage in a severe pandemic: maximizing survival by establishing triage thresholds." *Critical care medicine*, 44(9): 1762–1768.

Heckman, James J. 2007. "The economics, technology, and neuroscience of human capability formation." *Proceedings of the national Academy of Sciences*, 104(33): 13250–13255.

Hendren, Nathaniel, and Ben Sprung-Keyser. 2020. "A unified welfare analysis of government policies." *The Quarterly Journal of Economics*, 135(3): 1209–1318.

Hurley, Jeremiah. 2000. "An overview of the normative economics of the health sector." In *Handbook of Health Economics*. Vol. 1, 55–118. Elsevier.

Hyndman, Rob J, and Yanan Fan. 1996. "Sample quantiles in statistical packages." *The American Statistician*, 50(4): 361–365.

Kalbfleisch, John D, and Ross L Prentice. 2003. *The statistical analysis of failure time data.* John Wiley & Sons.

Keiding, Niels, Per Kragh Andersen, and John P Klein. 1997. "The role of frailty models and accelerated failure time models in describing heterogeneity due to omitted covariates." *Statistics in Medicine*, 16(2): 215–224.

Killien, Elizabeth Y, Brianna Mills, Nicole A Errett, Vicki Sakata, Monica S Vavilala, Frederick P Rivara, Niranjan Kissoon, and Mary A King. 2020. "Prediction of Pediatric Critical Care Resource Utilization for Disaster Triage." *Pediatric Critical Care Medicine*, e491–e501.

Manning, Willard G, and John Mullahy. 2001. "Estimating log models: to transform or not to transform?" *Journal of Health Economics*, 20(4): 461–494.

Manski, Charles F. 2018*a*. "Credible ecological inference for medical decisions with personalized risk assessment." *Quantitative Economics*, 9(2): 541–569.

Manski, Charles F. 2018*b*. "Reasonable patient care under uncertainty." *Health Economics*, 27(10): 1397–1421.

Maves, Ryan C, James Downar, Jeffrey R. Dichter, John L. Hick, Asha Devereaux, James A. Geiling, Niranjan Kissoon, Nathaniel Hupert, Alexander S. Niven, Mary A. King, Lewis L. Rubinson, Dan Hanfling, James G. Hodge Jr, Mary Faith Marshall, Katherine Fischkoff, Laura E. Evans, Mark R. Tonelli,

Randy S. Wax, Gilbert Seda, John S. Parrish, Robert D. Truog, Charles L. Sprung, and Michael D. Christian. 2020. "Triage of scarce critical care resources in COVID-19: an implementation guide for regional allocation." *Chest*, 158(1): 212–225.

Miller, Kathryn E, and Philip Toltzis. 2020. "Finding the Right Ethical Framework for PICU Resource Allocation During a Pandemic." *Pediatric Critical Care Medicine*.

OECD. 2020. "Beyond containment: Health systems responses to COVID 19 in the OECD."

Pathak, Parag A, Tayfun Sönmez, M Utku Ünver, and M Bumin Yenmez. 2020. "Fair allocation of vaccines, ventilators and antiviral treatments: leaving no ethical value behind in health care rationing." *arXiv preprint arXiv:2008.00374*.

Pollack, Murray M. 2016. "Severity of illness confusion." *Pediatric critical care medicine*, 17(6): 583.

Pollack, Murray M, Kantilal M Patel, and Urs E Ruttimann. 1996. "PRISM III: an updated Pediatric Risk of Mortality score." *Critical Care Medicine*, 24(5): 743–752.

Pollack, Murray M, Richard Holubkov, Ron Reeder, J Michael Dean, Kathleen L Meert, Robert A Berg, Christopher JL Newth, John T Berger, Rick E Harrison, Joseph Carcillo, et al. 2018. "Pediatric Intensive Care Unit (PICU) Length of Stay: Factors Associated with Bed Utilization and Development of a Benchmarking Model." *Pediatric Critical Care Medicine*, 19(3): 196.

Prentice, Ross L. 1978. "Linear rank tests with right censored data." *Biometrika*, 65(1): 167–179.

Randles, Ronald H, and Douglas A Wolfe. 1979. *Introduction to the theory of nonparametric statistics.*

Ridder, Geert. 1990. "The non-parametric identification of generalized accelerated failure-time models." *The Review of Economic Studies*, 57(2): 167–181.

Rosenbaum, Paul R. 2010. *Design of observational studies.* Vol. 10, Springer.

Schmidt, H. 2020. "The way we ration ventilators is biased." *New York Times*, 15.

Skinner, Jonathan. 2011. "Causes and consequences of regional variations in health care." In *Handbook of Health Economics*. Vol. 2, 45–93. Elsevier.

Slater, Anthony, Frank Shann, and Gale Pearson. 2003. "PIM2: a revised version of the Paediatric Index of Mortality." *Intensive Care Medicine*, 29(2): 278–285.

Song, Yunjie, Jonathan Skinner, Julie Bynum, Jason Sutherland, John E Wennberg, and Elliott S Fisher. 2010. "Regional variations in diagnostic practices." *New England Journal of Medicine*, 363(1): 45–53.

Soto, GJ, GS Martin, and MN Gong. 2013. "Healthcare disparities in critical illness." *Critical care medicine*, 41(12): 2784.

Sprung, Charles L, Gavin M Joynt, Michael D Christian, Robert D Truog, Jordi Rello, and Joseph L Nates. 2020. "Adult ICU Triage During the Coronavirus Disease 2019 Pandemic: Who Will Live and Who Will Die? Recommendations to Improve Survival." *Critical care medicine*, 48(8): 1196–1202.

Straney, Lahn, Archie Clements, Jan Alexander, and Anthony Slater. 2010. "Quantifying variation of paediatric length of stay among intensive care units in Australia and New Zealand." *Quality and Safety in Health Care*, 19(6): e5–e5.

Thakur, Neeta, Sam S Oh, Elizabeth A Nguyen, Melissa Martin, Lindsey A Roth, Joshua Galanter, Christopher R Gignoux, Celeste Eng, Adam Davis, and Kelley Meade. 2013. "Socioeconomic status and childhood asthma in urban minority youths. The GALA II and SAGE II studies." *American Journal of Respiratory and Critical Care Medicine*, 188(10): 1202–1209.

Tobin, Martin J. 2019. "Why physiology is critical to the practice of medicine: a 40-year personal perspective." *Clinics in Chest Medicine*, 40(2): 243–257.

Tobin, Martin J. 2020. "Basing respiratory management of COVID-19 on physiological principles." *American Journal of Respiratory and Critical Care Medicine*, 201(11): 1319–1320.

Toltzis, Philip, Gerardo Soto-Campos, Evelyn M Kuhn, Ryan Hahn, Robert K Kanter, and Randall C Wetzel. 2015. "Evidence-based pediatric outcome predictors to guide the allocation of critical care resources in a mass casualty event." *Pediatric Critical Care Medicine— Society of Critical Care Medicine*, 16(7): e207–e216.

Tsiatis, Anastasios A. 1990. "Estimating regression parameters using linear rank tests for censored data." *The Annals of Statistics*, 18(1): 354–372.

van der Waerden, B. L. 1953. "Ein neuer Test für das Problem der zwei Stichproben." *Math. Annalen*, 126: 93–107.

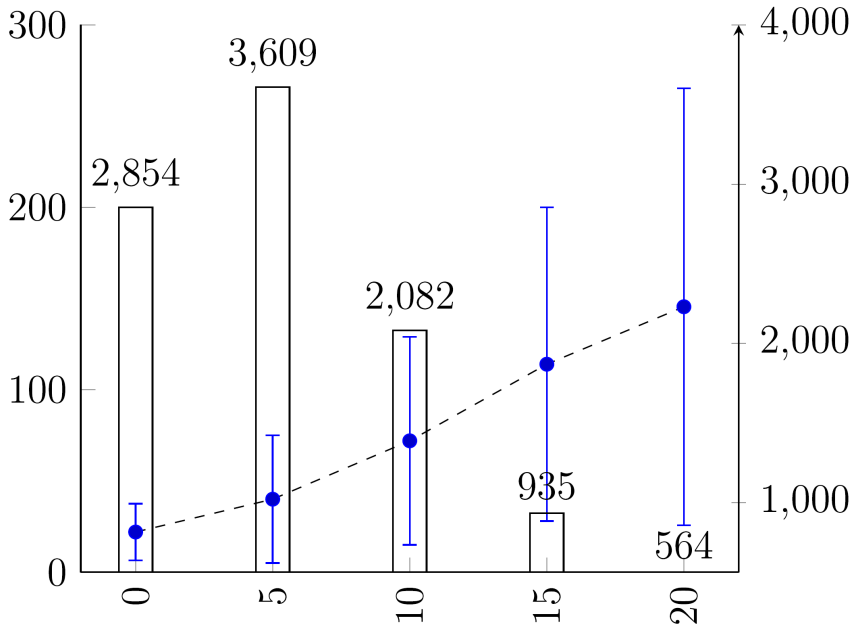
Vincent, J-L, Rui Moreno, Jukka Takala, Sheila Willatts, Arnaldo De Mendonça, Hajo Bruining, CK Reinhart, PeterM Suter, and Lambertius G Thijs. 1996. "The SOFA (Sepsis-related Organ Failure Assessment) score to describe organ dysfunction/failure." *Intensive Care Medicine*, 22: 707–710.

Wong, Hector R, Natalie Z Cvijanovich, Nick Anas, Geoffrey L Allen, Neal J Thomas, Michael T Bigham, Scott L Weiss, Julie Fitzgerald, Paul A Checchia, and Keith Meyer. 2015. "Developing a clinically feasible personalized medicine approach to pediatric septic shock." *American Journal of Respiratory and Critical Care Medicine*, 191(3): 309–315.

Young, Alwyn. 2019*a*. "Channeling fisher: Randomization tests and the statistical insignificance of seemingly significant experimental results." *The Quarterly Journal of Economics*, 134(2): 557–598.

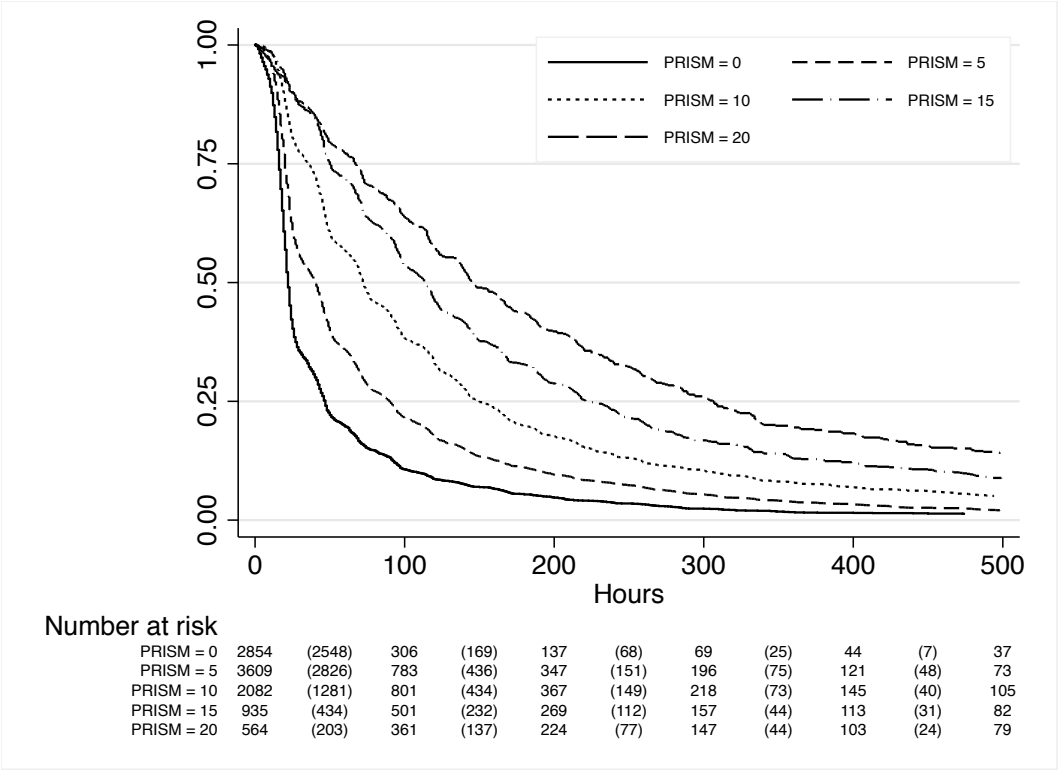
Young, Alwyn. 2019*b*. "Consistency without inference: Instrumental variables in practical application."

Zimmerman, Jack E, Andrew A Kramer, Douglas S McNair, and Fern M Malila. 2006. "Acute Physiology and Chronic Health Evaluation (APACHE) IV: hospital mortality assessment for todays critically ill patients." *Critical care medicine*, 34(5): 1297–1310.



Duration and PRISM

Figure I: This figure is an exploratory and descriptive analysis for the severity-duration relationship. The PRISM score is categorized into five groups (0=0, 1–5=5, 6–10=10, 11–15=15, and >16=20), with the distribution of patients per group represented by the bar values and right vertical axis. The left vertical axis is length of stay in hours with the median and interquartile range displayed using the blue dot and whiskers. The graph suggests an association with increased median length of stay and PRISM categories. Also noticeable are, (i) the increasing inter-quartile ranges with increasing severity categories and (ii) the large proportion of patients with lower PRISM scores.



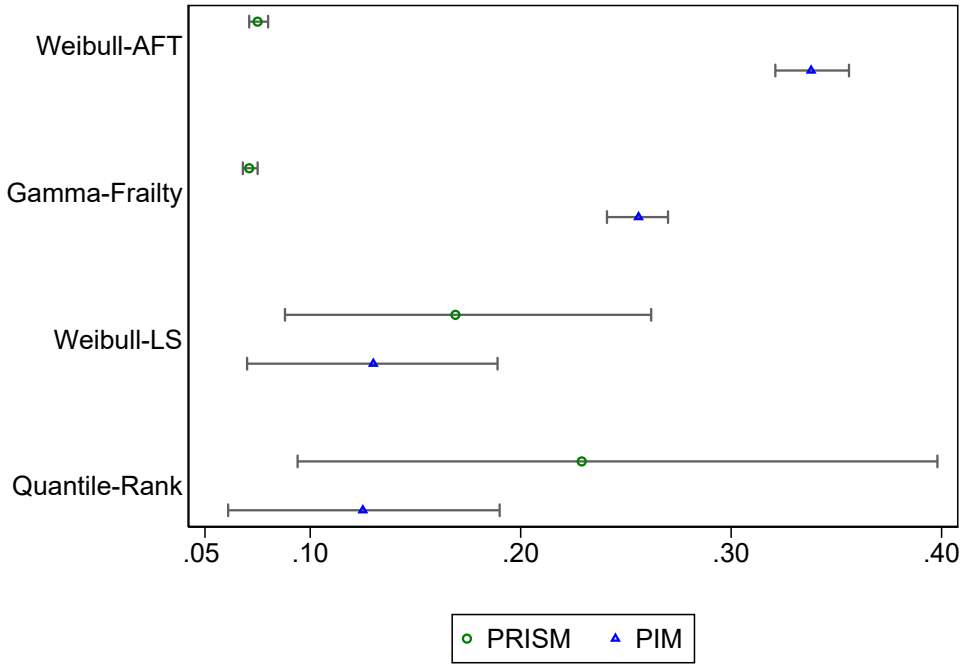
Stratified Kaplan-Meier function

Figure II: This figure is a graph of the empirical Kaplan-Meier function stratified by five groupings of the PRISM score corresponding to FIGURE 1. The Kaplan-Meier functions monotonically shift outward from the origin for increasing PRISM groupings. Proportional shifts on the time (horizontal) scale is consistent with an accelerated failure time metric. Numbers below the figure detail the respective risk sets for each PRISM grouping (0=0, 1–5=5, 6–10=10, 11–15=15, and >16=20).

Duration analysis of severity scores PRISM and PIM

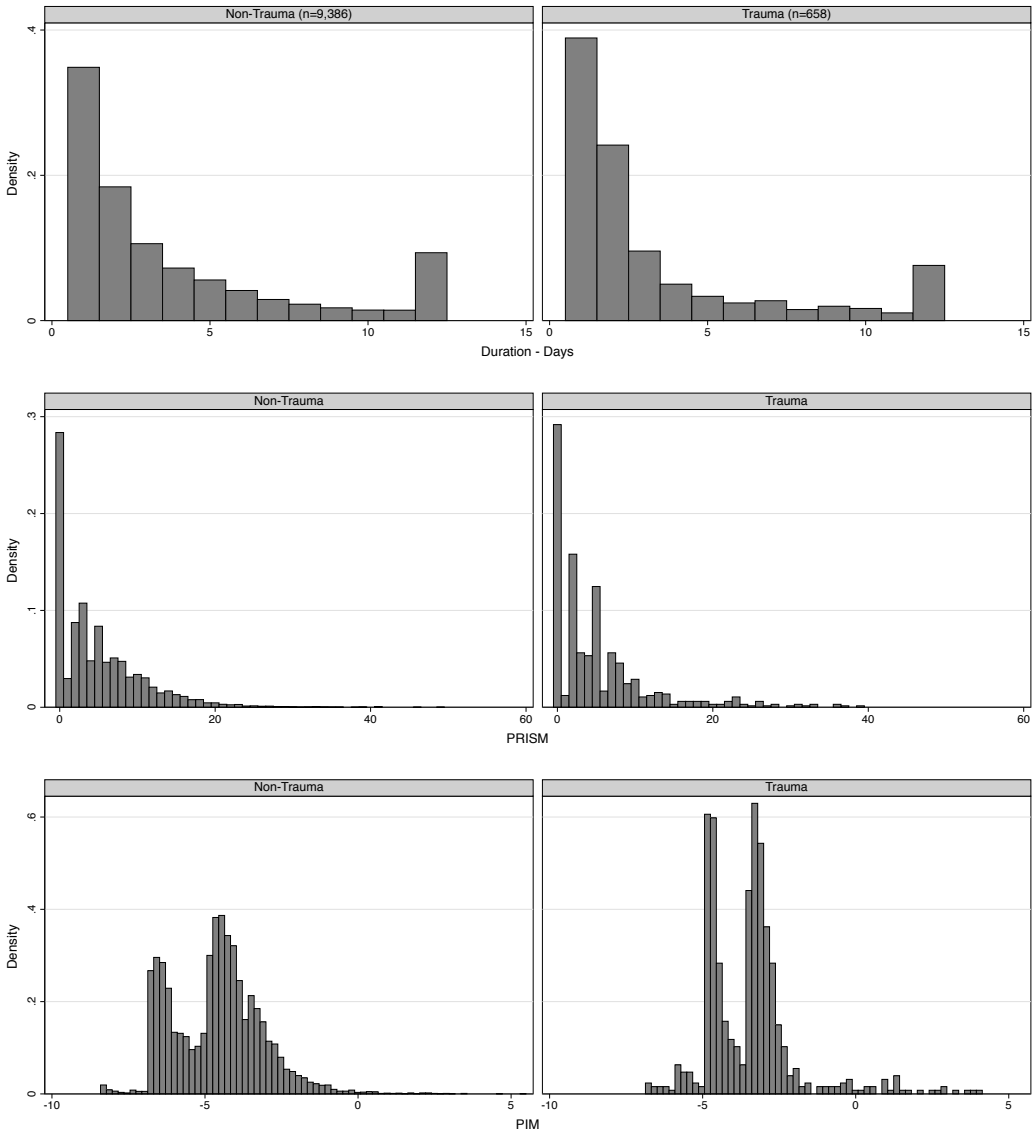
Analysis	PRISM	PIM
Reference analysis		
Weibull – Likelihood	(0.071, **0.075**, 0.080)	(0.321, **0.338**, 0.356)
Gamma – frailty	(0.068, **0.071**, 0.075)	(0.241, **0.256**, 0.270)
Identification Robust IV		
Weibull – Least Squares	(0.088, **0.169**, 0.262)	(0.070, **0.130**, 0.189)
Empirical Quantile – Rank	(0.094, **0.229**, 0.398)	(0.061, **0.125**, 0.190)

Table I: This table reports the estimates of β for four accelerated failure time ICU length of stay analysis of equation (10) – PRISM column two, and equation (11) – PIM column three. Outer values are 95% confidence sets and the center value in bold is either the point estimate or Hodges-Lehman-Sen estimator. Other covariates – not reported, are a hospital effect (six hospitals) and patient pre-exposure controls; age, chronic conditions, previous ICU admission, and congenital cardiac condition. The top two analysis serve as a reference for ignoring the endogeneity of the severity score. The first analysis (Weibull – Likelihood) ignores endogeneity and frailty, while the second analysis (Gamma – frailty) accounts for frailty with a presumed gamma frailty distribution but ignores endogeneity. The third and fourth analysis account for endogeneity using the trauma status of each patient as an instrumental variable. These exact, simulation based confidence sets are size controlled and robust to weak instruments. The third analysis provides intermediate estimates that serves as a comparator to the central analysis of the paper, the rank IV (Empirical Quantile – Rank) with empirical quantile scores of the baseline–frailty convolution distribution. In this analysis, all covariates are permitted to be jointly correlated with unmeasured and unobserved individual factors, which accounts for the concurrent endogeneity of the hospital effect. The fundamental assumption is the exchangeability of W, the ordered aligned residuals. Sample size 10,044 of which 658 are trauma patients.



95% CONFIDENCE SETS FOR PRISM AND PIM

Figure III: This figure is a graphical representation corresponding to TABLE 1. The whiskers represent the 95% confidence sets. For the first two rows, the green dot is the point estimate for PRISM and the blue triangle is the point estimate for PIM. For the third and fourth row the blue dot is the Hodges-Lehman-Sen estimator for PRISM and the blue triangle is the Hodges-Lehman-Sen estimator for PIM. The confidence sets for the third and fourth row are weak-instrument robust – where the closed sets in all cases reveal an informative instrument and model specification compatible with the data. Furthermore, the reference analysis (row one and two) for both PRISM and PIM exhibit confidence sets that do not overlap with the IV identification-robust confidence sets of row three and four. The correction for PRISM and PIM (i.e. movement from rows one and two to rows three and four) is in opposite directions, reflecting the central role of direct physiology as a driver of ICU utilization. The intermediate analysis result (row three) for PIM is in close agreement with the central results (row four) explained by the endogenous treatment and diagnoses being absorbed directly in the PIM score. Whereas, the intermediate and central results are notably different for PRISM, reflecting the endogenous treatment and diagnosis choices absorbed by the hospital effect – which is uniquely accounted for with the quantile rank inference. The wider confidence sets for PRISM capture both the inherent sampling uncertainty, small sample of instrumented, and the instrument quality which varies with the choice of severity marker.



Non-Trauma vs Trauma

Figure IV: This figure is an exploratory graphical summary of the distribution of the left-hand side variable duration (categorized to days – with the last category top coded to 12 days or greater), and the right-hand side variables, respectively PRISM, and PIM for the non-Trauma and Trauma groups of patients. The histograms broadly reveal that the binary instrumental variable Trauma has comparable support over both left and right-side variables. Although the Trauma patients have a mildly lower median duration, the histogram of duration does not reveal differences in clusters or shape between Trauma and Non-Trauma. Similarly the distribution of severity scores are similar but, with higher median values for Trauma.