Trends in Successful Resuscitation after Cardiac Arrest under Trending Misclassification Error: Estimating Bounds for Partially Verified Data

November 20th, 2017

JOB MARKET PAPER

Shaun Shaikh
(shaikhsa@mcmaster.ca)

Arthur Sweetman
(arthur.sweetman@mcmaster.ca)

Abstract: Estimating trends over time, including those surrounding policy changes, typically does not address the plausible confounding issue of trends in data quality, leading to a non-classical measurement error problem. This may be a concern with either survey or administrative data, where reporting attitudes may change over time or measurement quality may improve with time. Our application is to administrative health data, which is often used in epidemiological studies to evaluate trends in binary health characteristics and treatments. We address the detection of a trend in a binary outcome – successful resuscitation following cardiac arrest – allowing for trending misclassification error. Employing a mixture model, we compute bounds on the outcome following Horowitz and Manski (1995) under contaminated and corrupt data assumptions. Identification relies on validation information from a non-random subsample of the data allowing us to place upper bounds on measurement error. We also consider how identification is improved with monotonicity assumptions (Manski and Pepper 2000), bounded variation assumptions (Manski & Pepper, 2013, 2017), and subgroup specific verification rates (Dominitz and Sherman 2004, 2006; Kreider and Pepper 2007, 2008). We show evidence of a trend in the successful resuscitation rate for the population of reported cardiac arrests in Ontario under assumptions that are weaker than those in the existing literature.

Keywords: Partial Identification – Partial Verification – Monotone Instruments - Misclassification Error – ICD

JEL codes: I10, C25

This research was funded by the Government of Ontario. The views expressed in this paper are the views of the authors and should not be taken to represent the views of the Government of Ontario.

Thanks to Charles Manski, Joel Horowitz, Ivan Canay, Adam Rosen, Matthew Webb, Andrew Jones, Paul Contoyannis and Mike Veall for helpful advice.

1 Introduction

Measurement error is ubiquitous in empirical economics, the broader social sciences and health research. It has received attention in econometrics since at least Frisch (1934), and methodological advances continue (for overviews see Bound, Brown, and Mathiowetz 2001; Hausman 2001; Chen, Hong, and Nekipelov 2011; Schennach 2016). One recent development is the partial identification literature on misclassification error (e.g., Horowitz and Manski 1995; Kreider and Pepper 2007, 2008; Molinari 2008). This approach allows us to bound potentially informative parameter sets without making the strong and sometimes arbitrary assumptions normally needed for point identification. The measurement problem in our application is misclassification of a binary outcome variable where the misclassification rates may be correlated with the regressor of interest – a time trend. This is a nonclassical measurement error problem where standard approaches will be biased. Our application is estimating trend parameters for resuscitation success rates after in-hospital cardiac arrest using administrative data from Ontario, Canada.

Our work contributes to four broad literatures. First, while Pepper (2001) looked at the related but distinct problem of trends in missing data, this is the first application of partial identification econometrics to the problem of estimating trends under trending measurement error. There are many applications in economics and the empirical social sciences involving the estimation of a trend (e.g., difference-in-differences). Furthermore, trends in measurement error are likely to be commonplace for several reasons. For instance, with survey data, stigma based misreporting may trend because of the duration dependence of stigma effects from habit formation, and/or from trends in public attitudes. Measurement error may also trend because of changes in measurement technology. This is a particular concern with administrative data held by firms, public agencies or charities. Examples of changing technologies may include information/communication system updates, internal/external audit pressures, legislation, and/or data quality monitoring/improvement programs. Major shocks to measurement technology may involve dynamic adjustment processes where measurement quality may, for example, initially drop and then steadily improve due to institutional learning until a new equilibrium is reached.

Second, this is the first application which we are aware of using an econometric approach to misclassification applied to the *international classification of diseases* (ICD) coding system. The econometrics literature has typically focused on measurement problems in survey data, sometimes treating administrative datasets as validation sources (see Bound, Brown, and Mathiowetz 2001). Administrative health datasets represent an invaluable source of rich, low cost information for the analysis of many important questions in health economics and other fields. But, the unique data complexity of ICD coding (e.g., 15, 149 diagnostic codes were used in provincial hospitals over our data period) and the variation in data quality across codes (e.g., a large validation study observed 0.0% to 100.0% recoding agreement across the codes that were examined; Juurlink et al. 2006) makes this data source unique in terms of the challenges and opportunities for work on measurement problems. This is a largely unexplored domain of applied health economics with clear implications for informing policymakers and analysts about health information quality across the large number of international health systems in which the ICD standard is implemented.

Third, this work is also a methodological contribution to an epidemiological literature where estimating trends in binary health characteristics and treatments is a routine task, but biases resulting from trending measurement error are not commonly addressed. There is also a large epidemiological literature on data quality and validation. ICD-based hospital data is often validated in re-abstraction studies where trained coders recode data from original medical charts and gauge the level of agreement between original and recoded data. When epidemiological studies on trends consider data quality, they typically use validation study results as a "permission slip" to pursue a study where coding quality passes some arbitrary threshold. Such a snapshot of data quality is not enough information to adjust for biases under trending measurement error. Our approach accounts for the time dependence of measurement error and brings validation data directly into the estimation phase. There is an abundance of useful validation information in administrative health data, however typical approaches to estimating trends in binary health variables do not take full advantage of it.

Finally, we show evidence of an increasing trend in the success rate of resuscitation for reported in-hospital cardiac arrests in Ontario using assumptions which are weaker than the existing literature. This information is of interest to a clinical cardiological audience and contributes to a small literature of large sample studies on trends in survival after cardiac arrest (e.g., Ehlenbach et al., 2009; Girotra et al., 2012; Kazaure, Roman, & Sosa, 2013; Wong et al., 2014), which has been spawned due to recent advents in medical technologies. Similar to this clinical literature, our work does not address the counterfactual problem implicit in the motivation for such work, because we do not formally account for selection over time in unobserved health. However, we take far more serious account of the measurement problem which is prerequisite to credible work on causal inference.

We are interested in measurement error in an outcome variable. Under classical measurement error, where the measurement error is independent of the latent true value, standard approaches are unbiased. This is not so when measurement error is nonclassical, which is always the case with binary variables. With binary variables, the possible values of measurement error are completely restricted by the underlying true value (i.e., false positives versus false negatives). This is known as the problem of misclassification, and it can, in and of itself, bias results significantly for standard parametric approaches to estimating binary outcome models. For example, Hausman, Abrevaya, and Scott-Morton (1998) show that misclassification rates as low as low as 2% can lead to 15-20% biases in probit estimates.

Some approaches to parametric and semiparametric identification with misclassified outcomes require independence between the misclassification rates and regressors (e.g., Card 1996; Hausman, Abrevaya, and Scott-Morton 1998). However, this is not applicable for estimating a trend when misclassification error in the outcome is also trending. One approach to semiparametric identification of misclassified outcome models which allows for misclassification rates to be correlated with regressors is Lewbel (2000). Identification here relies on an exclusion restriction in the form of a variable that is correlated with the true outcome but uncorrelated with misclassification error. However, in practice it may be difficult to find a

variable which credibly fulfils such criteria. Our identification strategy relies on bounding measurement error using validation data, which is often available in data contexts such as ours.

Our approach to identification is nonparametric, yet it allows us to estimate informative bounds on trend parameters without the need for exclusion restrictions (although assumptions of this form can be accommodated and may help identify more informative parameter regions – that is tighter bounds). We follow the seminal approach of Horowitz and Manski (1995) on the partial identification of parameters for continuous and binary variables under *contaminated* and *corrupt* data, given information on the upper bound of measurement error. Unlike the classical assumptions, the contaminated data assumption is an independence assumption that is consistent with misclassification error and can significantly tighten bounds under misclassification compared to the worst case (i.e., under corrupt data).

Horowitz and Manski's (1995) technique is further developed by Dominitz and Sherman (2004) and (2006) who show tighter bounds can be derived with additional information on the upper bound of subgroup specific error rates (or equivalently, the lower bound on *verification* rates) and monotonicity restrictions across subgroups (see Manski and Pepper 2000, 2009). Kreider and Pepper (2007, 2008) implement this *partial verification* approach with misclassified data. We most closely follow Kreider and Pepper (2008), however, we additionally consider simple bounded variation assumptions (Manski & Pepper, 2013, 2017) that impose mild restrictions on extreme forms of data corruption (which are unlikely in our applied context), while still, allowing for some degree of "bounded" data corruption.

Our application is to estimate bounds on trend parameters in successful resuscitation rates after in-hospital cardiac arrest in Ontario, Canada, allowing for time dependence in misclassification error. There are several reasons to suspect such a trend may be present. First, the coding system underwent a major revision in 2003, moving from ICD-9 to ICD-10. Following such a major shock it is conceivable that dynamic adjustment was still underway in our data period, from April 1st, 2005 to March 31st, 2010. Also, hospital administrative data of the type that we use are subject to quality monitoring and improvement.

Perversely, by improving data quality, programs of this type can induce trends in misclassification rates that confound trend estimation of a measured outcome.

Our identification strategy relies, primarily, on information in non-random validation data, which we then use to construct subgroup verification rates which bound the maximum measurement error per subgroup. We construct such verification rates based on an "internal" validation data approach employing one high quality (essentially measurement error free) measure in the data set to validate an error-prone measure for a subgroup of the data.[1] That is, we take advantage of the presence of a high-quality code - survival to discharge – which gives us information about coding error in successful resuscitation for patients who survive to discharge. This follows because successful resuscitation is an initial survival measure. Those who are not resuscitated after an arrest must have died, and must, therefore, not be alive at discharge. When we do not observe resuscitation, but do observe survival to discharge, we know a coding error has occurred in successful resuscitation.

Additionally, our data rich environment contains information on several patient characteristics and treatments for which we can make plausible monotonicity assumptions. For example, we consider restrictions such as the success rate of resuscitation being nonincreasing in age, and nondecreasing in heart stimulation (e.g., heart defibrillation). Bounds can tighten significantly when such restrictions bind in different directions.

The remainder of the paper is organized as follows. Section 2 provides background on misclassification error including standard and partial identification approaches. Background for the applied problem is in

[1] Although we are using "multiple measures" of cardiac arrest, we deliberately avoid the *repeated measurements* terminology because of its specific meaning in the econometrics measurement error nomenclature. That is, we do not have repeated measures which are both measured with error and independent of each other, allowing for recovery of a true distribution by deconvolution methods (e.g., Schennach 2004). We are using a true *validation data* approach (see Bound, Brown, and Mathiowetz 2001).

section 3, including details on calculation of verification rates. Details of bounds computations are in Section 4. Section 5 and 6 present the data and results, while section 7 concludes.

2 Background on the Misclassification Problem

2.1 Misclassification and the classical assumptions

Consider an observed scalar outcome y_i, and scalar regressor x_i, with corresponding true values y_i^* and x_i^*, and measurement errors μ_i^y and μ_i^x. A well-known result in the classical single regressor linear-in-errors model

$$y_i = \beta_0 + \beta_1 x_i + u_i,$$

where $y_i = y_i^* + \mu_i^y$, and $x_i = x_i^* + \mu_i^x$ is that measurement error in x_i results in attenuation bias of the β_1 parameter estimate, while measurement error in the response variable y_i, when it is continuous, reduces the precision of estimates without inducing bias. However, the latter unbiasedness result requires that measurement error in the outcome satisfies the classical assumptions: μ_i^y is uncorrelated with y_i^* and x_i^*, as well as μ_i^x and u_i.[2] We consider the case where the regressor is perfectly measured (i.e., $x = x^*$ and $\mu_x = 0$).

One well known case where the classical assumptions cannot hold is with misclassification of a binary response variable, in our case the outcome y_i. Measurement error takes on two values which depend on the true value of y_i^*.

[2] There is variation in the definition of "classical" measurement error within the econometrics literature. For predictors and outcomes, uncorrelated measurement error with underlying true variables, regressors, and regression error satisfies the classical assumptions according to Hausman (2001). Bound, Brown, and Mathiowetz (2001) also include, explicitly, independence of all measurement error terms from each other. Some usages of the classical measurement error terminology only explicitly require independence/uncorrelatedness of measurement error from the true outcomes (e.g., Hyslop and Imbens 2001; Chen, Hong, and Nekipelov 2011). Schennach (2016), however, distinguishes between weak and strong classical measurement error, where the former requires zero mean errors which are independent from true values, whereas the latter are mean zero conditional on true values and may allow for heteroskedasticity.

(I) a false positive, when $y^* = 0$, with $\pi_0 \equiv Prob(y = 1|y^* = 0)$, or

(II) a false negative, when $y^* = 1$, with $\pi_1 \equiv Prob(y = 0|y^* = 1)$.

This measurement error is negatively correlated with the true value. This misclassification error results in a downward bias on marginal effect estimates. Hausman, Abrevaya, and Scott-Morton (1998) show simulation results under symmetric misclassification (i.e., false positive rate = false negative rate), finding misclassification rate as low as low as 2%, can lead to probit estimates which are biased by 15-20%.

2.2 Parametric and semi-parametric identification under misclassification

Parametric and semiparametric approaches to identification in the misclassified binary outcome model have been proposed and typically estimate a modified version of the latent variable threshold model for binary outcomes of the form.

$$y_i = \pi_0 + (1 - \pi_0 - \pi_1)F(x_i\beta) \quad (1)$$

where $F()$ is a link function that bounds outcomes between 0 and 1 (e.g., a probit), y_i and x_i are, respectively, a binary outcome and a vector valued set of regressors, and π_0 and π_1 are the probabilities of false positives and negative as defined above. This collapses to the standard model when there is no misclassification error (i.e., when $\pi_0 = \pi_1 = 0$). An early application is Card (1996), who considers a transformed version of (1) with a linear link function, and identifies parameters by using external population data on π_0, π_1. Identification also requires an upper limit on π_0, π_1 and that their values do not depend on regressors x_i. Hausman, Abrevaya, and Scott-Morton (1998) show that the model is identified without population estimates of π_0, π_1 with nonlinear $F()$, including nonparametric links (for semi-parametric identification) so long as π_0 and π_1 sum to less than 1 and are independent of x_i. The latter independence condition is required under both of these approaches makes them unsuitable for the analyzing our problem

when the explanatory variable, time, is correlated with misclassification rates. When this is the case, the above models are not identified without additional information.[3]

Lewbel (2000) allows for a more general specification for (1) where $\pi_0 = \pi_0(x)$, $\pi_1 = \pi_1(x)$ are allowed to covary with x_i. He proves semiparametric identification, given a variable that determines the true outcome y_i^* which is uncorrelated with measurement error μ_y. Although this approach works in principle for our problem, given an appropriate exclusion restriction, it is difficult to find a convincing variable of this type in practice. Instead, we consider other forms of restrictions under the partial identification approaches outlined in the next subsection. Exclusion restrictions can also be accommodated in our completely nonparametric approach.

2.3 Partial identification approaches to misclassification

Instead of focusing on point identification, we follow the partial identification literature and compute sets of parameter values consistent with plausible assumptions (for overviews see Manski 2003; Tamer 2010). Parameter sets, in our case, are connected, so they are fully characterized by their bounds. We consider nonparametric estimation of bounds on a trend parameter for a binary outcome under trending misclassification error following the techniques employed in Horowitz and Manski (1995). Using a mixture model

$$y = y^*Z + \mu_y(1 - \mathrm{Z}) \tag{2}$$

where Z is a binary variable equal to 1 whenever the true outcome, y^*, is observed and equal to 0 otherwise, they derive nonparametric bounds on y^* given an upper bound on the measurement error, $\bar{\mu}$, where

$$\pi_0 + \pi_1 < \bar{\mu} \tag{3}$$

[3] Approaches similar to Card (1996) may be adapted to answer our problem if we had population estimates of misclassification rates across time, but would still require parametric assumptions.

These bounds are sharp (i.e., the tightest possible bounds using all information given by the identifying assumptions) for binary and continuous data under the following two measurement error characterizations:

(A1) Contaminated Data: Z is independent of y^*, and

(A2) Corrupt Data: Z is not independent of y^*.

Dominitz and Sherman (2004) extend Horowitz and Manski (1995) when additional information from a verification indicator is observed, for some subgroup of the data, that tells us when an observation is drawn from the true distribution. In this case, tighter bounds can be obtained compared to Horowitz and Manski (1995) when combined with monotonicity assumptions (e.g., Manski & Pepper, 2009, 2000). Dominitz and Sherman (2006) relax the verification assumptions for the case where a groups-specific verification indicator determines a lower bound on the probability that an observation is drawn from the true distribution (as opposed to assuming this probability is one). We follow this *partial verification* approach.

The partial verification approach also allows us to relax the contaminated data assumption to a weaker group specific conditional mean independence assumption on y^* given Z. The nature of the independence assumption for contaminated data differs from the classical measurement error independence assumptions. Whereas the latter are violated by misclassification error this is not so for contaminated data. Furthermore, the classical assumptions are violated when misclassification rates depend on the true value of a regressor, such as, in our case, the time-period. However, we can accommodate this using the partial verification approach by computing separate partial verification rates per time-period.

Dominitz and Sherman (2004, 2006) do not directly apply partial verification to misclassification with a binary outcome. Kreider and Pepper (2007) apply partial verification to misclassification in regressors, while Kreider and Pepper (2008) apply the approach to misclassification in the unconditional mean of a binary variable. This last derivation of bounds most closely resembles ours.

Dominitz and Sherman (2004, 2006) and Kreider and Pepper (2007, 2008) consider monotonicity restrictions to improve identification following Manski and Pepper (2000). Restrictions of this form essentially impose the expected true outcome y^* as being weakly monotone (i.e., weakly increasing or decreasing) in another variable, w, which is defined as a monotone instrumental variable. In the context of our estimation problem, such monotone instruments can improve identification depending on the responsiveness of y^* to w versus the variation in measurement error across w.

For intuition consider the following intentionally exaggerated example. Consider the plausible restriction that y^* (i.e., successful resuscitation) is weakly decreasing in age (i.e., w), where age is divided into three categories young, middle, and old. Now consider that measurement error in the middle category is so high (e.g., based on estimated partial verification rates) that bounds are completely uninformative (i.e., restricted to the bounded support [0,1] of y^*). If young and old are fully verified, we can get point estimates for the mean of y^* in each group. If the estimated mean y^* for each, young and old, is respectively 0.4 and 0.5, then monotonicity informs us that the mean for the middle group must be between 0.4 and 0.5. When the conditional bounds are tightened in this fashion then the overall bounds shrink. Note that If y^* is less responsive to age, so that the mean of y^* for the old group is 0.41, then the bounds on the middle group would tighten to [0.40, 0.41]. It should also be clear that if we increase the number of subgroups of age, the sampling variability in each subgroup would increase as the number of observations declines in each, so that some subgroup-specific bounds would restrict others simply because of randomness. This results in a finite sample bias problem. Kreider and Pepper (2007) implement a bootstrap bias correction to bound estimates based on the bootstrap bias correction for the mean (see Efron and Tibshirani 1994).

We consider several monotone instruments in our analysis which we enumerate and justify in later subsections. We also consider simple and intuitive bounded variation assumptions (Manski & Pepper, 2013, 2017) which allow us to relax the contaminated data assumptions while placing reasonable restrictions on

extreme correlations between y^* and Z. Formal derivations of our bounds will be discussed in section 4 after necessary additional details on the applied problem are introduced in the next section

Molinari (2008) makes a noteworthy contribution to the partial identification literature on misclassification. Her *direct misclassification* approach is a general bounds estimation procedure for multinomial misclassification problems. This approach can accommodate all the identifying restrictions we apply in our work, and more complex restrictions. However, this approach is unnecessary for our purposes. We only require simple frequency estimators and analytical/numerical solutions to simple linear programming problems for estimation as in Kreider and Pepper (2008).

3 Background for the Applied Measurement Problem

3.1 Cardiac arrest, resuscitation, and the measurement problem

Cardiac arrest is a major burden to the healthcare system of developed countries. Substantive interest in cardiac arrest survival trends for the medical / epidemiological audience follows from the desire to investigate whether recent advances in treatment technologies (Health Quality Ontario 2005; Morrison et al. 2015) and American Health Association guideline updates (Neumar et al., 2015) have translated into population level improvements in survival. Note that in-hospital and out-of-hospital cardiac arrest have very different prognoses and are treated separately in the literature (In the U.S., 24% of patients survive an in-hospital arrest while only 6% survive one out-of-hospital; Graham et al. 2015).

Several recent large sample American studies have attempted to determine whether an improvement in survival is observed for in-hospital cardiac arrest relying on one of: registry data (Girotra et al. 2012), random samples of hospital data (Kazaure, Roman, and Sosa 2013), and full sub-population administrative data (i.e., Medicare; limited to age 65 and above in the U.S.; Ehlenbach et al. 2009). Kazaure, Roman, and Sosa (2013), which is a 20% random sample of national hospital discharges is the largest study using a representative sample of all US adults, and finds a positive trend in adjusted survival to discharge. In parallel

work for Ontario in-hospital cardiac arrest, we found no evidence for a trend in survival to discharge from 2003 to 2010, but we find a trend in initial resuscitation successes over this time-period. As is typical in applied work, we also swiftly came to the realization that data quality for the successful resuscitation diagnoses is a concern for validity of our results and robust conclusions.

One indication of measurement error for us followed from inconsistent information across different codes. At times, this sort of inconsistency can become useful as a source of identification in measurement error problems. In particular, we found a number of observations where patients survived to discharge but, were not coded as having been successfully resuscitated. This may not be immediately puzzling to a lay audience, but it is logically inconsistent once the biology of cardiac arrest is understood.

Although frequently confused in layman's usage, cardiac arrest is not the same as a heart attack (i.e., myocardial infarction; which is heart muscle death due to blockage and oxygen deprivation). Cardiac arrest is a more severe condition in which the heart's natural rhythm and pumping action ceases. That is, the heart stops. Without resuscitation the prognosis is nearly immediate death. Crucially, for our identification strategy, successful resuscitation is both a short-term survival measure, and a necessary condition for subsequent survival (esp., survival to discharge). Thus, the biological restrictions between codes provides information about the measurement process, which is useful in restricting parameter estimates.

Since we exploit information from multiple measures, one might naturally think that our approach to identification follows the *repeated measurements* approach to identification in the econometrics measurement error. This approach, however, relies on multiple measures, which are measured with error, where the measurement error terms satisfy a mutual independence assumption (Hausman et al. 1991; Hausman, Newey, and Powell 1995; Hausman 2001; Schennach 2004, 2016). Our data environment more closely resembles the *validation data* approach to measurement error, where a validation sample is perfectly measured (or more realistically, meets some gold standard of measurement) for a subsample of the data (see Bound, Brown, and Mathiowetz 2001). In our case survival to discharge is reliably recorded with

absolutely no errors found in two large nationally representative chart studies using our data source (Canadian Institute for Health Information, 2003; Richards, Brown, & Homan, 2001).

Using the biological restrictions described above, combined with validation information indicating that survival to discharge is measured without error, we can estimate the measurement error in the (non-random) sample of those who survive to discharge. This information can then be used to estimate upper bounds on measurement error (i.e., $\bar{\mu}$ in the previous section), as well as subgroup-specific verification rates, $\bar{v}_g$, which can then be combined with either mean independence assumptions (i.e., contaminated data) or monotonicity assumptions (e.g., by treating age as a monotone instrument and computing age-specific verification rates). In the next subsection we discuss details of the computation of verification rates.

3.2 Computing partial verification rates

Panel A of figure 1 illustrates the relationship between coded and actual resuscitation status, while part B clarifies, as area 4, the subset of those with a false negative whose true status can be observed by virtue of observing them survive to discharge. However, as represented by area 5 of panel B, those with a false negative who die subsequent to resuscitation but prior to discharge cannot be identified since they are confounded with those with a true negative. Further, those with false positives cannot be directly observed. Using this information, we can estimate the error rate of false negatives in successful resuscitation coding for survivors.

In earlier work on cardiac arrest, we found evidence of a trend in the observable false negative error rate for successful resuscitation between 2003 and 2010 – that is, we observed a trend in those erroneously not recorded as being successfully resuscitated after an arrest, but who nevertheless are observed to be alive at discharge. Of course, while we observe the false negative for those who survive to discharge after failing to have been coded as surviving the initial arrest, we have no direct measure of the total extent of measurement error. In particular, for those who do not survive to discharge, we know neither the false negative nor false positive rate of coded successful resuscitation. However, we can make aggregate

predictions of unobserved measurement error based on assumed relationships between the observed and unobserved error rates. For example, in panel A of figure 2, the lower curve shows the observed false negative error rate for those who survive to discharge. The upper curve is a prediction (weighted at the hospital level) of the false negative measurement error, assuming (perhaps unrealistically – as addressed below) that this rate is the same in each hospital-year group for those who actually survive to discharge (our observed error) and those who do not (which is unobserved). Since those who do not survive to discharge are much more numerous than those alive at discharge, their contribution to the aggregate error is much larger assuming the same error rate. This prediction does not include errors from the false positives.

In panel B of figure 2, we show what would happen to the trend in successful resuscitation if we incorporate information about observed and predicted false negative errors as illustrated in panel A (assuming no false positive errors). Under these unrefined assumptions, we find that successful resuscitation no longer exhibits a systematic increasing trend and measurement error can fully account for the observable trend in successful resuscitation. Although these are (overly) strong assumptions, this illustration demonstrates the very serious potential confounding problem which may occur if trending measurement error is ignored. The partial identification approach we undertake provides a powerful and convenient framework to understand how alternative assumptions will translate into bound estimates on the successful resuscitation rate over time.

It is easier to see how group-specific verification rates, $\bar{v}_g$, are computed if we first consider the trivial case when the entire dataset is a single group (i.e., $\bar{v}_g = \bar{v}$). To get the total verification rate, we must consider separately verification rates for survivors and non-survivors to discharge, $\bar{v}_1$ and $\bar{v}_0$, and corresponding upper bounds on each associated error, $\bar{\mu}_1$ and $\bar{\mu}_0$. We use survivor to discharge data to estimate the error rate in successful resuscitation for survivors, which we can then translate into information on verification rates on non-survivors under some assumption about the relationship between the two.

For survivors to discharge, we know that false positives are zero and we can compute an estimate of false negatives for survivors to discharge as $P(R = 0|S = 1)$, where R = 1 when a resuscitation success is recorded in the ICD coding, and R = 0 otherwise; and S = 1 when survival to discharge is true, and S = 0 otherwise. Then, given (3), we can get an estimate for an upper bound, $\bar{\mu}_1$, of errors for survivors to discharge based on

$$\bar{\mu}_1 = P(R = 0|S = 1) \tag{4}$$

We relate the error rate upper bound for survivors, $\bar{\mu}_1$, to its counterpart for non-survivors, $\bar{\mu}_0$, by η, a non-survivor error bound magnification factor as follows:

$$\bar{\mu}_0 = \eta\bar{\mu}_1 \tag{5}$$

We can then compute how bounds vary under alternative assumptions about η. When $\eta = 1$ then the upper bound on the error rate for survivors is taken as equal to the upper bound that for non-survivors. *A priori*, we could articulate arguments supporting measurement error for non-survivors as being either greater than or less than its counterpart for survivors to discharge. For instance, we may surmise that a coder may feel that data quality is not important if someone has died (other than the code for death itself) and thus may exhibit less attention to detail, which would lead to more errors for those who die. On the other hand, there may be institutional pressure to make sure all of the proverbial i's are dotted and t's are crossed when someone dies, leading to greater attention to detail and a lower error rate when a patient dies. This specification allows us a convenient way to consider sensitivity to assumptions on this dimension.

The partial verification rate for non-survivors follows from (4) and (5) as

$$\bar{v}_0 = 1 - \eta P(R = 0|S = 1) \tag{6}$$

For survivors to discharge, we know all patients are fully verified as successful resuscitations, meaning $\bar{v}_1$= 1. Combined with (6), this gives us the verification rate for the full dataset as

$$\bar{v} = [1 - \eta P(R = 0|S = 1)]P(S = 0) + P(S = 1) \quad (7)$$

It is easy to see how subgroup and time-period specific verification rates can be similarly computed as

$$\bar{v}_{g,t} = [1 - \eta P(R = 0|S = 1, g, t)]P(S = 0|\, g, t) + P(S = 1|g, t) \quad (8)$$

4 Empirical Methods

4.1 Horowitz and Manski (1995) bounds under data corruption

When deriving bounds, it is useful to return to the mixture formulation (2). Then, it is easy to see that $\pi_1 = P(y = 0|Z = 0)$ and $\pi_0 = P(y = 1|Z = 0)$. Horowitz and Manski (1995; henceforth HM) bounds are derived from the basic relationship between the observed and true outcome. We compute time-period specific bounds using the relationship:

$$P(y^* = 1|t) = P(y = 1|t) + P(y = 0|Z = 0, t) - P(y = 1|Z = 0, t).$$

That is, the true successful resuscitation rate is equal to the coded (or perhaps miscoded) rate plus the false negative rate minus the false positive rate. If we know the probability of having correct data is greater than some known minimum (i.e., $P(Z = 1|t) \geq \bar{v}_t$), then HM bounds are computed as the minimum and maximum values of $P(y^* = 1|t)$, under constraint (3) given the corrupt data assumption (A2), with a bounded support condition of $y^* \in [0,1]$. It is easy to see that these bounds are given by:

$$[LB_{HM}(t), UB_{HM}(t)] \equiv [\max\{P(y = 1|t) - (1 - \bar{v}_t),\, 0\}, \min\{P(y = 1|t) + (1 - \bar{v}_t), 1\}].$$

For small enough $\bar{v}_t$ the bounds will not be informative. These bounds represent the worst-case scenario resulting from the correlation between Z and y^* (i.e., from data corruption).

Given that, in practice the worst case of either: (1) all false positives; or (2) all false negatives is not commonly observed in chart validation studies (e.g. Juurlink et al. 2006), we consider reasonable bounds on the worst case while still allowing for some degreee of data corruption. We do this by placing simple

restrictions on the minimum proportion of false negatives / false positives. That is we consider the bounded variation assumptions

(A3) $\pi_0 \equiv P(y = 1|Z = 0) \geq \lambda_0$, and

(A4) $\pi_1 \equiv P(y = 0|Z = 0) \geq \lambda_1$.

When λ_0 or λ_1 is strictly greater than zero, we impose that there is at least some proportion of errors that are false positives or false negatives. This allows for placing simple restrictions on the worst case of data corruption without needing independence as in (A1). We refer to this formulation as bounded corruption (BC). The resulting bounds are given by

$LB_{BC}(t) \equiv \max\{P(y = 1|t) - (1 - \bar{v}_t)(1 - \lambda_1) + (1 - \bar{v}_t)\lambda_1, 0\}$, and

$UB_{BC}(t) \equiv \min\{P(y = 1|t) - (1 - \bar{v}_t)\lambda_0 + (1 - \bar{v}_t)(1 - \lambda_0), 1\}$.

4.2 Partial verification bounds under contaminated data

We implement bounds under partial verification (henceforth PV bounds) following Kreider and Pepper (2008) using partial verification rates for each subgroup and time-period. We treat the hospital as the cross-sectional subgroup when calculating PV bounds.

Group and time specific bounds, $UB(g,t)$ and $LB(g,t)$, can be, respectively, derived by maximizing and minimizing the objective function $P(y^* = 1|g,t)$ over $v_{g,t} \equiv P(Z = 1|g,t)$ subject to the constraint that $0.5 < \bar{v}_{g,t} \leq v_{g,t} \leq 1$. The lower bound of 0.5 implies that measurement is at least better that chance (see Card 1996; Hausman, Abrevaya, and Scott-Morton 1998; Kreider and Pepper 2008).

Since y^* is not observable, the problem must be recast in terms of observables. Noting that $P(y = 1|g,t)$ is equal to $P(y^* = 0|Z = 0, g, t)$ when $Z = 0$, we can state the observable outcome probability as:

$$P(y=1|g,t) = P(y^*=1|Z=1,g,t)P(Z=1|g,t) + P(y^*=0|Z=0,g,t)P(Z=0|g,t).$$

From contaminated data we have

(A5) $P(y^*=1|g,t) = P(y^*=1|Z,g,t)$.

Substituting (A5) and rearranging, we get,

$$P(y^*=1|Z=1,G) = \frac{V_{g,t} - P(y=0|g,t)}{2V_{g,t}-1} \equiv f(v_{g,t};y).$$

Thus, we can get conditional PV bounds by solving:

$UB_{PV}(g,t) \equiv Max_{v_{g,t}} f(v_{g,t};y)$ s.t. $\bar{v}_{g,t} \leq v_{g,t} \leq 1$, and

$LB_{PV}(g,t) \equiv Min_{v_{g,t}} f(v_{g,t};y)$ s.t. $\bar{v}_{g,t} \leq v_{g,t} \leq 1$.

Time-specific bounds are computed using the law of total probability:

$$UB_{PV}(t) = \sum_{g=1}^{N_g} UB_{PV}(g,t)P(g|t),$$

$$LB_{PV}(t) = \sum_{g=1}^{N_g} LB_{PV}(g,t)P(g|t).$$

The derivation of these bounds required only a weaker conditional mean independence assumption in the form of (A5) as compared to the contaminated data assumption (A1).

4.3 Bounds under monotonicity

Manski and Pepper (2000) formalize the notion of a monotone instrumental variable (MIV), wherein we know an expected outcome, y^*, is weakly monotone in some covariate, w. Without loss of generality consider expected y^* nondecreasing in a single monotone instrument w in each time-period, t, so that

$$E[y^*|t,w=i] \geq E[y^*|t,w=j], \text{ for } \forall\ i > j \qquad (9)$$

Then from (9) it follows that

$$\sup_{i \leq j} \; LB(i,t) \leq P(y^* = 1 | w = j, t) \leq \inf_{k \geq j} \; UB(k,t),$$

where $LB(\cdot,t)$, $UB(\cdot,t)$ are bounds for each $w = \cdot$, that can be derived according to any of the HM, BC, or PV formulations defined above. Time-specific bounds are given by the law of total probability:

$$UB_{MIV}(t) = \sum_j \inf_{k \geq j} \; UB(k,t) P(w = k|t),$$

$$LB_{MIV}(t) = \sum_j \sup_{i \leq j} \; LB(i,t)\, P(w = i|t).$$

More generally, consider a vector of monotone instruments $\boldsymbol{w} = [w_1, \dots, w_M]$ which can be partitioned into w_m and $w_{(-m)}$ for $m = 1, \dots, M$. For ease of exposition we transform any nonincreasing MIV so that y^* is nondecreasing in all w_m. Then we have M separate MIV conditions.

(A6) $E\left[y^* \middle| t, w_{(-m)}, w_m = i\right] \geq E\left[y^* \middle| t, w_{(-m)}, w_m = j\right]$, for $\forall\, i > j$, and $m = 1, \dots, M$

Conditional bounds can be computed by imposing each MIV condition in (A6) sequentially, and aggregating using the law of total probability until time-period specific bounds are produced. Finding the greatest lower bound and least upper bound requires imposing all M conditions sequentially for all $M!$ possible sequences.

To correct a well-known small sample bias problem when computing bounds with suprema and infima conditions, Kreider and Pepper (2007) implement a bootstrap bias correction to the MIV bounds estimates shown above. Given T_n, a consistent estimator of some a lower/upper bound parameter of interest, we can compute the small sample bias of T_n as $E^*(T_n) - T_n$, where $E^*()$ is the bootstrap expectation operator. The unbiased estimates are given by $T_n^u = 2T_n - E^*(T_n)$ (see Efron and Tibshirani 1994). We implement this bias correction for MIV bounds, resampling at the hospital cluster level.

Since we have population data, we do not consider a sampling process for any of the previous bound estimates. But for MIV bounds, it may be the case that the joint MIV assumptions only hold in a super population. In this case the bias correction procedure, which presumes a sampling process, is considered. We do not formally consider inference for these bound estimates, but in the partial identification literature inference can be performed on both identified sets or a true parameter which must be contained in these sets. When the bias corrected MIV bounds for different time-periods do not overlap, this implies rejection of a null hypothesis of equality of the true successful resuscitation rate in those two time-periods at any significance level. Thus, we take non-intersecting bounds across time-periods as clear evidence of a trend.

More generally, given any time-period specific bound computation procedure discussed above we can bound the difference in time-specific successful resuscitation rates according to the following formula:

$$\Delta ij \equiv E[y^*|t=i] - E[y^*|t=j] \in [LB(i) - UB(j), UB(i) - LB(j)\,] \quad (10)$$

If Δij does not contain zero, then bounds in each time period do not overlap.

5 Data

The data for this study are provided by the Ontario Ministry of Health and Long Term Care, and taken from the Discharge Abstract Database (DAD) for Ontario, Canada, which contains the full population of hospital admission-related discharges. We consider admissions to all acute care hospitals over the fiscal years (FY) Apr. 1st, 2005 until Mar. 31st, 2011 (i.e., FY 2005 to 2010). This excludes emergency department visits and day procedures.

The DAD contains a rich set of admission-related clinical and demographic information, including up to 25 diagnosis-related codes per hospital admission. Diagnostic data are encoded under the Canadian implementation of the 10th revision of ICD (ICD-10-CA). The coding details include a field that identifies comorbidities that occur after admission. Our interest is restricted to in-hospital cardiac arrests, which are typically treated separately from out-of-hospital cardiac arrests in the medical literature as both prognosis

and response differ. Accordingly, we restrict the analysis to patients with at least one clearly demarcated post-admission cardiac arrest. Of the patients who have at least one post-admission arrest, we also exclude patients who also had a clearly demarcated pre-admission arrest (i.e., when the first post-admission arrest is a subsequent arrest, re-arrest after a primary pre-admission arrest event).

We also exclude subsequent hospitalizations with cardiac arrest for the same patient. Thus, our analysis is restricted to a single arrest related admission per patient. It is, however, possible that a patient has more than one post-admission arrest during a single admission, but, because the number of arrests during admission is not likely to be reliably measured, we do not take this into consideration in sample selection. Furthermore, from this point on, unless made explicit otherwise, we use the shorthand "arrest" (e.g. when referring to the number of arrests per year, hospital, etc.) as referring to an arrest-related admission, given all the aforementioned caveats.

Cardiac arrest is identified in the data by ICD-10-CA code I.46 where a subsequent fourth digit code determines if the patient was successfully resuscitated (I.46.0), had a *sudden cardiac death* (1.46.1), or was attempted to be resuscitated without success (I.46.9). For sudden cardiac deaths, it is unclear from the coding guidelines whether a resuscitation attempt was made. For this reason, we exclude patients where all post-admission arrests were sudden cardiac deaths (although if a patient had one I.46.0 or I.46.9 as well as an I.46.1 they are retained in the analysis). Finally, we exclude patients under age 18 and patients without a valid Ontario health number, and also, we drop the bottom quintile of observations associated with small hospitals defined by the average number of arrests per year (N = 14,605; 37 hospitals retained).

6 Results

Hospital-specific partial verification rates computed according to (8) – without a time series element - are shown in figure 3. We can see significant variation across hospitals, with the rates ranging from 0.618 to 0.980.

Table 1 shows both HM and PV bounds without any bounded variation or monotonicity assumptions. These initial tables are computed under the assumption that non-survivor measurement error is at most as high as survivor measurement error (i.e., $\eta = 1$). We consider departures from this assumption in later results.

HM bounds are shown in the first two columns of results. Verification rates for HM bounds are computed for three consecutive two-year time-periods of the dataset. Panel A of the table reports bound estimates on successful resuscitation rates for each of these time-periods. Panel B reports bound estimates for three trend parameters, $\Delta 31, \Delta 32, \Delta 21$, based on differences across the three two-year time-periods. We can see that HM bounds computed under the (A2) data corruption assumption do not yield informative estimates of any trend parameter, but are informative in terms of bounding successful resuscitation rates to values strictly within their bounded support.

The next two results columns of table 1 show PV bounds under the conditional mean independence assumption (A5). Under this formulation, we see evidence of a trend since $\Delta 31 \in [0.041, 0.109]$ and $\Delta 21 \in [0.010, 0.102]$. That is, we find that that the successful resuscitation rate is higher in both the second and third time-period compared to first. We cannot make such a claim for comparison between the second and third time-periods. Figure 4 shows how PV bounds for $\Delta 31$ are affected by relaxation of the equality assumption on the upper limit of non-survivor error to survivor error (i.e., $\eta = 1$). We see that the trend parameter, $\Delta 31$, is strictly positive for η as high as 1.7, so that even with significantly higher error rates among non-survivors, evidence of a trend is present under assumption (A5).

Recall that Assumption (A5) is a weak version of the contaminated data assumption (A1) because independence between the true outcome, y^*, and the true outcome observation indicator (i.e., Z) is not required to hold in the whole dataset but only within each hospital and time-period grouping. It is useful to recall that both the contaminated data assumption (A1) and its weaker conditional mean independence counterpart (A5) are consistent with misclassification error being correlated with the time trend. Thus, they both constitute assumptions that have greater validity and robustness than standard approaches to binary outcomes which do not account for misclassification (i.e., probit, logit, etc.) as well as those approaches

which account for misclassification but do not allow dependence of misclassification rates with regressors (e.g., Card, 1996; Hausman, Abrevaya, & Scott-Morton, 1998). We see that significant gains in identification power are made with this simple assumption while being logically consistent with the structure of our model. Nevertheless, it also clarifies that data corruption is a concern that needs to be considered.

In table 2, we compute HM and PV bounds for the lowest and highest measurement error quartiles of hospitals (weighted by number of patients). That is, we compute the error rate for survivors at the level of hospitals, rank hospitals in order of this average measurement error metric, and then compute quartiles for patients based on this data quality measure of their admitted hospital. This allows us to show two important results in table 2.

First, in the first two results columns in Panel A of table 2 we see, that for the lowest measurement error quartile, we can detect a trend between the first and third time-period. This is fairly robust evidence that for a large subset of high data quality hospitals, there is likely to be a real positive trend in successful resuscitation for reported cardiac arrests (assuming $\eta = 1$ is valid). Of course, for high measurement error hospitals, HM bounds reveal no useful information about the trend, absent additional assumptions.

Second, in Panel B of table 2, we show PV bounds on successful resuscitation rates per time-period for low and high measurement error quartiles. The point of this illustration is not to report any information about the trend parameters, but rather consider evidence around the issue of data contamination vs. corruption. For each time-period, a comparison of the low measurement error lower bounds to the high measurement error upper bounds yields the important result that low measurement error hospitals tend to have higher successful resuscitation rates. This provides evidence that there is some degree of data corruption in the measurement problem we are analyzing. If so, this would immediately invalidate assumption (A1), and results from HM bounds under data contamination would be invalid. But, since we are computing subgroup specific partial verification rates, which require only the conditional mean independence assumption (A5), it is not immediately clear that the PV bounds are invalid. This is because

the PV bounds formulation we use allows for y^* and Z to be correlated between hospital and time-period groupings, but not correlated within a time-period for each hospital. Nevertheless, evidence of between hospital data corruption raises question about within hospital data corruption as well, and it is useful to consider departures from assumption (A5).

To address the possibility of within hospital - time-period data corruption, we consider HM bounds with additional restrictions in the form of monotonicity and bounded variation conditions as discussed in section 4. We consider three values (0, 0.1, and 0.25) for the bounded variation parameters on the minimum proportion of false positives and false negatives, with $\lambda_0 = \lambda_1$.

Four monotonicity conditions are imposed. Three of them are motivated by the biology and medical practice around cardiac arrest. First, we impose that successful resuscitation rates are nonincreasing in age. This is intuitive and noncontroversial. Second, we impose that successful resuscitation is nondecreasing in heart stimulation interventions (ICD intervention code 1.HZ.09; e.g., defibrillation). This is because it is well known in the cardiac arrest medical literature that cardiac arrests with "shockable" rhythms have a higher survival probability (Girotra et al., 2012). Finally, we also consider successful resuscitation as nondecreasing in days spent in ICU. It is well known that an earlier response time to resuscitation is an important factor associated with higher survival (Girotra et al., 2012), and that this is more likely to occur in an ICU which typically has higher levels of monitoring and nurse to patient ratios than wards. Furthermore, for patients who are already in the ICU at the time of arrest, those that are successfully resuscitated are likely to remain in ICU for more time, compared to those were not resuscitated, who would die and leave the ICU immediately.

Panel A of Table 3 shows the correlation of observed successful resuscitation and the aforementioned three monotone instruments. The directions of correlation are as expected by our theoretical arguments. Panel B of Table 3 shows that these individual characteristics are also correlated with measurement error for survivors. We may consider this additional evidence of possible data corruption. With respect to

identification power, the bounds should tighten when measurement error varies significantly in the monotone instruments.

A fourth monotonicity condition is motivated based on the pattern of correlation we observed between y^* and low versus high measurement error hospitals in table 2. That is, we find that lower measurement error hospitals seem to have higher successful resuscitation rates. Thus, we consider imposing this restriction and seeing how the bounds respond.

Figure 5 shows a histogram of group specific partial verification rates for all cells of our MIV variables. Hospital measurement error is treated as four quartiles. Age is treated as three groups (18-74 years, 75-84 years, and 85+ years). ICU days is treated as 5 groups (0 days; (0,1] days; (1,3] days; (3,10] days; and >10 days). Heart stimulation is a binary category. From figure 5 we see that verification rates across all cells of these groups for all time-periods vary between 0 and 1. With many fully verified subgroups, many monotonicity restrictions may bind in different directions and be quite informative about conditional bounds in other cells.

But with many subgroups, smaller cell sizes lead to a concern of sampling variability. Although we effectively have population level data, it is not necessarily true the all monotonicity assumptions simultaneously hold in this definition of the population. In fact, we find that a small number of conditional bounds are empty sets, which means that, as articulated, the monotonicity assumptions can hold only a superpopulation. Proper inferential procedures would require careful specification of a sampling process appropriate to this problem. For the purposes of the present application, we find it useful to report bias corrected bound estimates using clustered resampling. As mentioned earlier, if bounds in different time-periods are non-overlapping (i.e., trend parameter lower bounds are greater than zero) then inference on trend parameters being zero should reject at any level. So, we can still identify trends, but are not able to benefit from the full power of formal inference techniques which are appropriate to this problem.

Results of bias corrected MIV bounds with BV conditions (i.e., under assumptions A3, A4, and A6) are shown in table 4. Although we considered three values for the minimum proportion of false positives/negatives ($0, 0.1, 0.25;\ \lambda_0 = \lambda_1$), we report only the lowest value (i.e., weakest restriction) for which a trend can be detected. The first two columns of table 4 show we can detect a trend between time-periods 1 and 3, with monotonicity assumptions and modest bounds (i.e., $\lambda_0 = \lambda_1 = 0.1$) on data corruption. But this result requires that non-survivor measurement error is not higher than survivor measurement error (i.e., $\eta = 1$). We relax this latter assumption for results in the next two columns of table 4, allowing non-survivor measurement error to be as much as twice the survivor measurement error bound (i.e., $\eta = 2$). In this case we can still detect a trend between time-period 1 and 3, but require a slightly stronger restriction on the extent of data corruption (i.e., $\lambda_0 = \lambda_1 = 0.25$).

7 Conclusion

Using partial identification techniques and information on partial verification rates of subgroups, as well as monotonicity and bounded variation assumptions, we demonstrate that informative bounds can be identified on a trend parameter in the success rate of resuscitation after in-hospital cardiac arrest in Ontario.

Although seemingly a trivial problem, the estimation of trends with misclassification in binary dependent variables with confounding time-dependent misclassification rates is a difficult problem to solve that is not taken seriously in most applied work involving trend estimation in health and applied economics.

We show that a weaker conditional mean independence formulation of the contaminated data assumption can be quite powerful in terms of tightening the identification region to an informative range. In our application, we found modest evidence of some data corruption. Possibly, this is due to institutional features of high data quality hospitals which translate into systematic differences in treatment technologies for cardiac arrest. We also show how minor restrictions on worst case bounds combined with monotonicity assumptions can identify a trend, while still allowing for some plausible range of data corruption. In general,

more sophisticated use of such bounded corruption assumptions may involve the use of information from chart validation studies.

We make a case that a large set of applied work involving similar ICD-based administrative health data and a similar methodology can be undertaken using either information from multiple measures within the dataset, or by taking advantage of a large cache external validation data from chart studies that is available for such administrative dataset.

The methods employed are relatively simple and have low computational cost. Furthermore, it is easy to extend this approach to multiple regressors, with correlated measurement error using standard density estimation techniques.

8 References

Bound, J., Brown, C., & Mathiowetz, N. (2001). Measurement Error in Survey Data. In E. E. Leamer & J. J. Heckman (Eds.), *Handbook of Econometrics* (Vol. 5, pp. 3705–3843). Amsterdam: North Holland/Elsevier.

Canadian Institute for Health Information. (2003). *Discharge Abstract Database (DAD)/CMG/Plx Data Quality Re-abstraction Study*. Ottawa, Canada.

Card, D. (1996). The Effect of Unions on the Structure of Wages: A Longitudinal Analysis. *Econometrica*, *64*(4), 957–979.

Chen, X., Hong, H., & Nekipelov, D. (2011). Nonlinear Models of Measurement Errors. *Journal of Economic Literature*, *49*(4), 901–937. https://doi.org/10.1257/jel.49.4.901

Dominitz, J., & Sherman, R. P. (2004). Sharp bounds under contaminated or corrupted sampling with verification, with an application to environmental pollutant data. *Journal of Agricultural Biological and Environmental Statistics*, *9*(3), 319–338. https://doi.org/10.1198/108571104x3389

Dominitz, J., & Sherman, R. P. (2006). Identification and Estimation of Bounds on School Performance Measures: A Nonparametric Analysis of a Mixture Model with Verification. *Journal of Applied Econometrics*, *21*, 1295–1326. https://doi.org/10.1002/jae

Efron, B., & Tibshirani, R. (1994). *An Introduction to the Bootstrap*. CRC Press.

Ehlenbach, W. J., Barnato, A. E., Curtis, J. R., Kreuter, W., Koepsell, T. D., Deyo, R., & Stapleton, R. D. (2009). Epidemiologic study of in-hospital cardiopulmonary resuscitation in the elderly. *New England Journal of Medicine*, *361*(1), 22–31. https://doi.org/10.1056/NEJMoa0810245

Frisch, R. (1934). *Statistical Confluence Analysis byMeans of Complete Regression Systems*. Universitetets Økonomiske Instituut.

Girotra, S., Nallamothu, B. K., Spertus, J. A., Li, Y., Krumholz, H. M., & Chan, P. S. (2012). Trends in Survival After In-Hospital Cardiac Arrest. *New England Journal of Medicine*, *367*(20), 1912–1920. https://doi.org/10.1016/j.biotechadv.2011.08.021.Secreted

Graham, R., McCoy, M. A., Schultz, A. M., & Editors. (2015). *Strategies to improve cardiac arrest survival: a time to act*. National Academies Press.

Hausman, J. (2001). Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left. *Journal of Economic Perspectives*, *15*(4), 57–67. https://doi.org/10.1257/jep.15.4.57

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, *87*(2), 239 269.

Hausman, J. A., Newey, W., Ichimura, H., & Powell, L. (1991). Identification and estimation of polynomial errors-in-variables models. *Journal of Econometrics*, *50*(3), 273–295.

Hausman, J. A., Newey, W. K., & Powell, J. L. (1995). Nonlinear errors in variables Estimation of some Engel curves. *Journal of Econometrics*, *65*(1), 205–233.

Health Quality Ontario. (2005). Use of automated external defibrillators in cardiac arrest: An evidence-based analysis. *Ontario Health Technology Assessment Series*, *5*(19), 29.

Horowitz, J. L., & Manski, C. F. (1995). Identification and Robustness with Contaminated and Corrupted Data. *Econometrica*, *63*(2), 281–302.

Hyslop, D. R., & Imbens, G. W. (2001). Bias from classical and other forms of measurement error. *Journal of Business & Economic Statistics*, *19*(4), 475–481.

Juurlink, D., Preyra, C., Croxford, R., Chong, A., Austin, P., Tu, J., & Laupacis, A. (2006). Canadian Institute for Health Information Discharge Abstract Database: A Validation Study. *Institute for Clinical Evaluative Sciences*, (June).

Kazaure, H. S., Roman, S. A., & Sosa, J. A. (2013). Epidemiology and outcomes of in-hospital cardiopulmonary resuscitation in the United States, 2000-2009. *Resuscitation*, *84*(9), 1255–1260. https://doi.org/10.1016/j.resuscitation.2013.02.021

Kreider, B., & Pepper, J. V. (2008). Inferring Disability Status from Corrupt Data. *Journal of Applied Econometrics*, *23*(3), 329–349. https://doi.org/10.1002/jae

Kreider, B., & Pepper, J. V. (2007). Disability and Employment. *Journal of the American Statistical Association*, *102*(478), 432–441. https://doi.org/10.1198/016214506000000997

Lewbel, A. (2000). Identification of the Binary Choice Model with Misclassification. *Econometric Theory*, *16*(4), 603–609.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F., & Pepper, J. V. (2000). Monotone Instrumental Variables: With an Application to the Returns to Schooling. *Econometrica*, *68*(4), 997–1010.

Manski, C. F., & Pepper, J. V. (2009). More on monotone instrumental variables. *The Econometric Journal*, *12*(1), S200–S216. https://doi.org/10.1111/j.l

Manski, C. F., & Pepper, J. V. (2013). Deterrence and the Death Penalty: Partial Identification Analysis Using Repeated Cross Sections. *Journal of Quantitative Criminology*, *29*(1), 123–141. https://doi.org/10.1007/s10940-012-9172-z

Manski, C. F., & Pepper, J. V. (2017). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *Review of Economic Dynamics Economics and Statistics*, forthcoming.

Molinari, F. (2008). Partial identification of probability distributions with misclassified data. *Journal of Econometrics*, *144*(1), 81–117. https://doi.org/10.1016/j.jeconom.2007.12.003

Morrison, L. J., Brooks, S. C., Dainty, K. N., Dorian, P., Needham, D. M., Ferguson, N. D., … Zhan, C. (2015). Improving use of targeted temperature management after out-of-hospital cardiac arrest: A stepped wedge cluster randomized controlled trial. *Critical Care Medicine*, *43*(5), 954–964.

Neumar, R. W., Eigel, B., Callaway, C. W., Estes, N. A. M., Jollis, J. G., Kleinman, M. E., … Sendelbach, S. (2015). American heart association response to the 2015 institute of medicine report on strategies to improve cardiac arrest survival. *Circulation*, *132*(11), 1049–1070.

Pepper, J. V. (2001). *How Do Response Problems Affect Survey Measurement of Trends in Drug Use?* in Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us. (C. F. Manski, J. V. Pepper, & C. V. Petrie, Eds.). Washington, DC: The National Academies Press.

Richards, J., Brown, A., & Homan, C. (2001). The data quality study of the Canadian discharge abstract database. *Proceedings of Statistics Canada Symposium 2001*, 1–10.

Schennach, S. M. (2004). Estimation of Nonlinear Models with Measurement. *Econometrica*, *72*(1), 33–75.

Schennach, S. M. (2016). Recent Advances in the Measurement Error Literature. *Annual Reviews of Economics*, *8*, 341–379. https://doi.org/10.1146/annurev-economics-080315-015058

Tamer, E. (2010). Partial Identification in Econometrics. *Annual Review of Economics*, *2*(1), 167–195. https://doi.org/10.1146/annurev.economics.050708.143401

Wong, M. K. Y., Morrison, L. J., Qiu, F., Austin, P. C., Cheskes, S., Dorian, P., … Ko, D. T. (2014). Trends in short- and long-term survival among out-of-hospital cardiac arrest patients alive at hospital arrival. *Circulation*, *130*(21), 1883–1890. https://doi.org/10.1161/CIRCULATIONAHA.114.010633

9 Figures

Figure 1 – Coding Relationships

Panel A

		Truth	
		Resuscitation	**No Resuscitation**
Coded	**Resuscitation (I.46.0=1)**	True Positive	False Positive
	No Resuscitation (I.46.0=0)	False Negative	True Negative

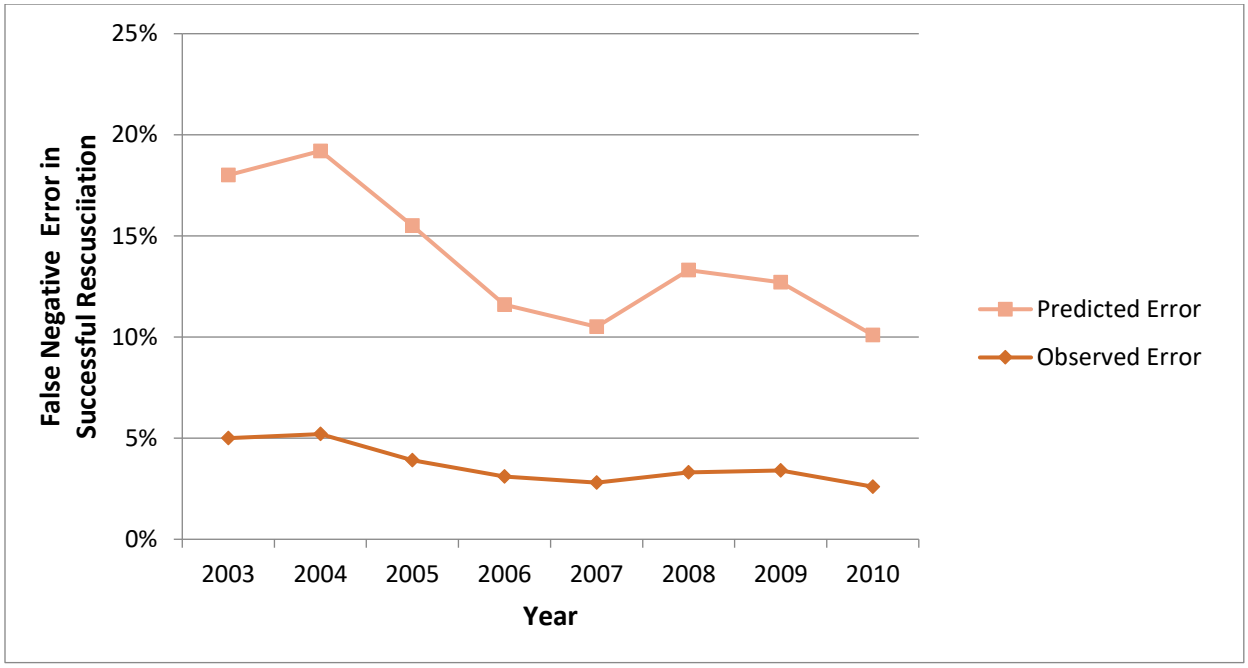
Panel B

		Truth		
		Resuscitation		**No Resuscitation**
Coded	**Resuscitation (I.46.0=1)**	1. Survive to Discharge	2. Died by Discharge	3. Died by Discharge
	No Resuscitation (I.46.0=0)	***4. Survive to Discharge***	5. Died by Discharge	6. Died by Discharge
Observe at Discharge		*Survive*	*Died by Discharge*	

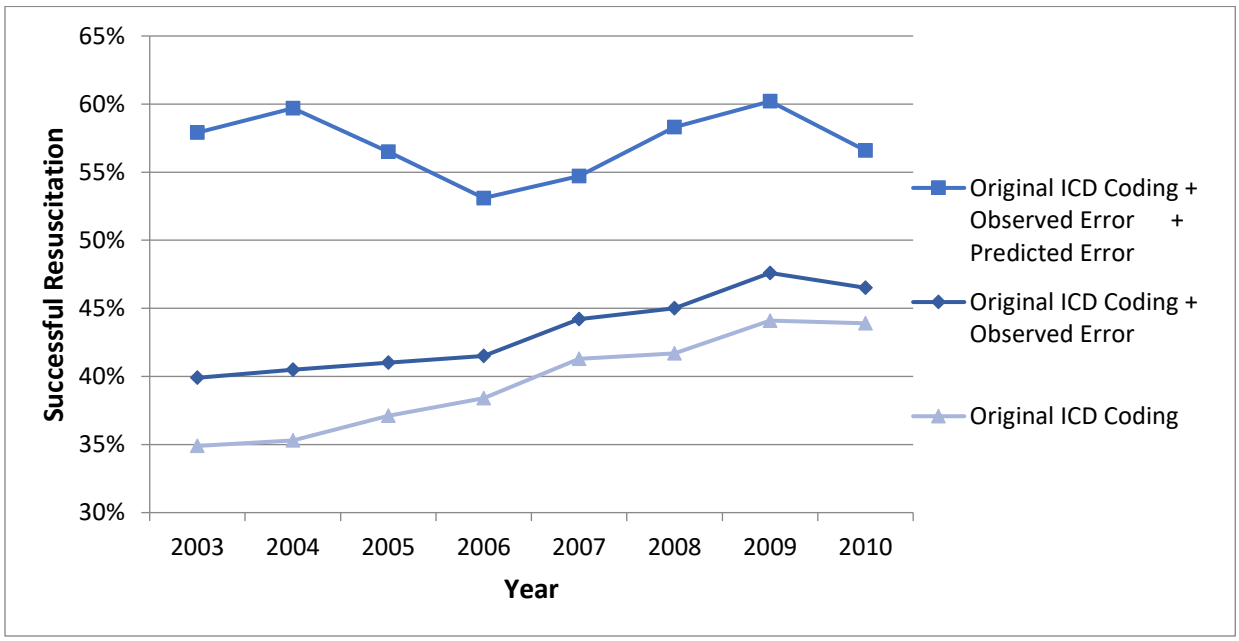
Notes: In Panel B, section 4 is miscoded, whereas 1 is correctly coded. We have no information about 2, 3, 5 or 6.

Figure 2 – Trends in Successful Resuscitation and Measurement Error

Panel A

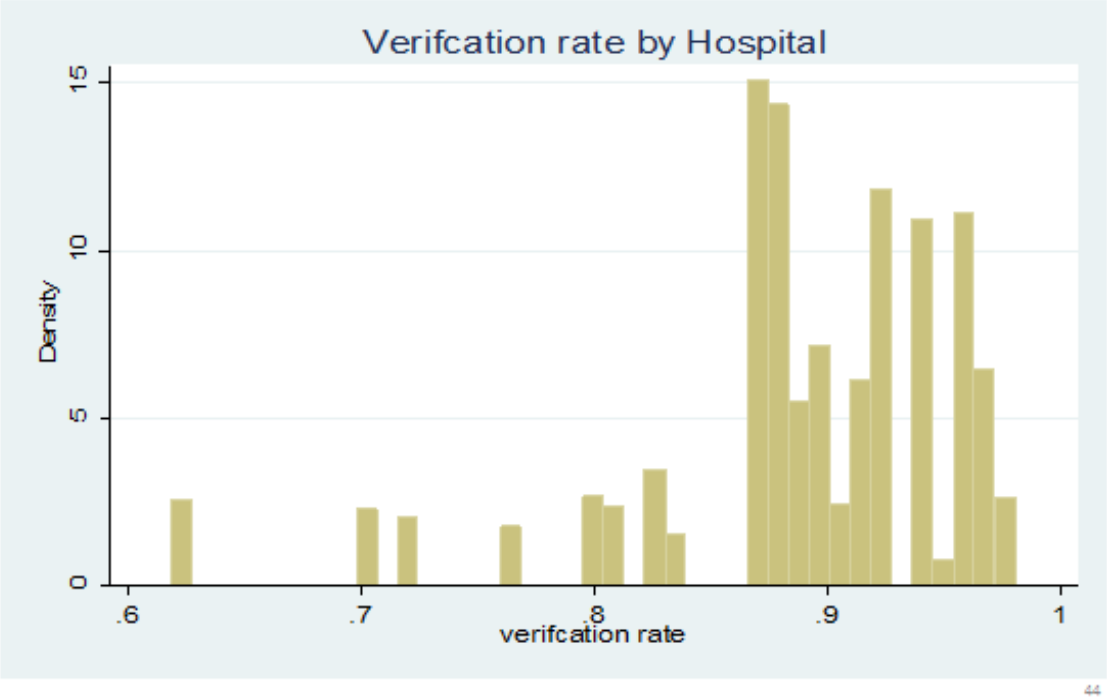


Panel B



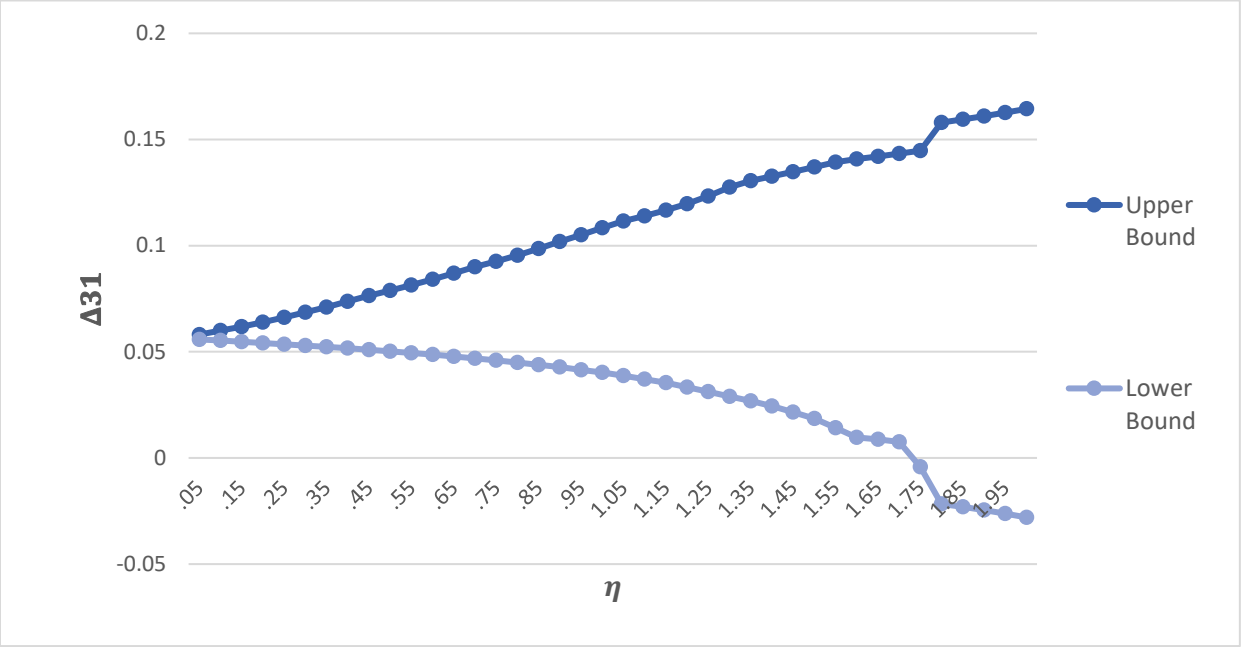
Notes: In both Panels, Observed Error recodes miscoded false negatives in successful resuscitation for survivors to discharge. Predicted Error takes hospital-year average observed false negative errors for survivors to discharge and forces equal rates of false negatives for non-survivors. The above scenarios assume no false positives.

Figure 3 – Histogram of Hospital Partial Verification Rates



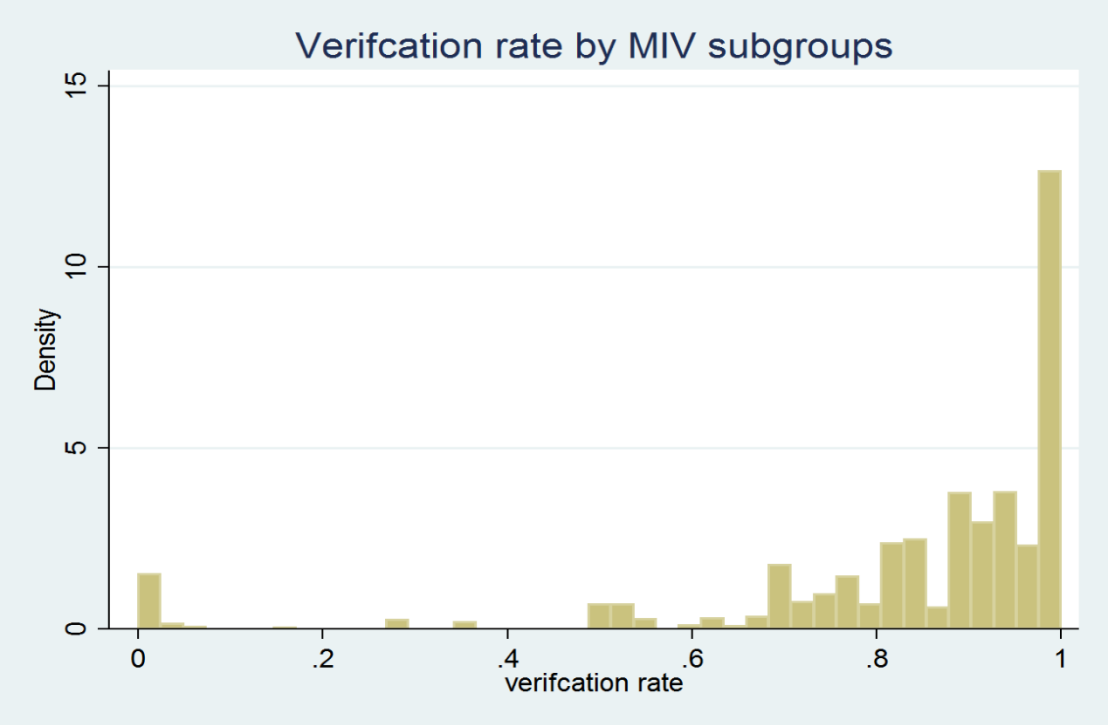
Notes: Verification rates are calculated as: $\bar{v}_g = [1 - \eta P(R = 0|S = 1, g)]P(S = 0|\, g) + P(S = 1|g)$ where g is a hospital, R is successful resuscitation, S is survival to discharge, and $\eta = 1$ is the magnification factor of the resuscitation error bound for non-survivors compared to its counterpart for survivor to discharge.

Figure 4 – PV Bounds of Trend Parameter, Δ31, by Non-survivor Error Bound Factor, η



PV: Partial verification bounds computed under assumption A5: $P(y^* = 1|g,t) = P(y^* = 1|Z,g,t)$, where g is the hospital.
η: The magnification factor of the resuscitation error bound for non-survivors compared to its counterpart for survivors to discharge. When $\eta > 1$, Non-survivor measurement error is allowed to exceed survivor error.
$\Delta 31 = E[y^*|t = 3] - E[y^*|t = 1]$

Figure 5 – Histogram of Partial Verification Rates by MIV subgroups



Notes: Verification rates are calculated as: $\bar{v}_{g,t} = [1 - \eta P(R = 0|S = 1, g, t)]P(S = 0|\, g, t) + P(S = 1|g, t\,)$ where g is each cell defined by age (18-74 years, 75-84 years, and 85+ years), ICU days (0 days; (0,1] days; (1,3] days; (3,10] days; and >10 days), heart stimulation, and hospital measurement error quartiles.

10 Tables

Table 1 – HM and PV Bounds ($\eta = 1$)

	HM bounds		PV bounds	
A. Mean resuscitation per time-period				
t	LB	UB	LB	UB
1	0.329	0.565	0.400	0.452
2	0.361	0.631	0.463	0.502
3	0.409	0.598	0.493	0.509
B. Differences in mean resuscitation between time-periods				
Δij	LB	UB	LB	UB
$\Delta 31$	-0.156	0.269	0.041	0.109
$\Delta 32$	-0.222	0.237	-0.009	0.046
$\Delta 21$	-0.204	0.222	0.010	0.102

HM: Horowitz and Manski (1995) bounds computed under assumption A2: Z is not independent of y^*

PV: Partial verification bounds computed under assumption A5: $P(y^* = 1|g,t) = P(y^* = 1|Z,g,t)$, where g is the hospital.

η: The magnification factor of the resuscitation error bound for non-survivors compared to its counterpart for survivors to discharge. In all above results, $\eta = 1$, so that non-survivor measurement error has the same bound as survivor error.

Table 2 – HM and PV Bounds by Hospital Measurement Error Quartile ($\eta = 1$)

	Lowest measurement error hospitals		Highest measurement error hospitals	
A. HM bounds - Differences in mean resuscitation between time-periods				
	LB	UB	LB	UB
$\Delta 31$	0.0005	0.1566	-0.2989	0.3744
B. PV bounds - Mean resuscitation per time-period				
t	LB	UB	LB	UB
1	0.475	0.486	0.343	0.438
2	0.526	0.533	0.339	0.412
3	0.560	0.566	0.434	0.471

HM: Horowitz and Manski (1995) bounds computed under assumption A2: Z is not independent of y^*

PV: Partial verification bounds computed under assumption A5: $P(y^* = 1|g,t) = P(y^* = 1|Z,g,t)$, where g is the hospital.

η: The magnification factor of the resuscitation error bound for non-survivors compared to its counterpart for survivors to discharge. In all above results, $\eta = 1$, so that non-survivor measurement error has the same bound as survivor error.

Table 3 – Correlation of Monotone Instruments with Resuscitation / Resuscitation Error

A. Correlation table - Full Sample				
	Successful Resuscitation	Age	Days in ICU	Heart Stimulation
Successful Resuscitation	1			
Age	-0.1146*	1		
Days in ICU	0.1920*	-0.0916*	1	
Heart Stimulation	0.0963*	-0.0574*	-0.0160*	1
B. Correlation table - Survivors to discharge				
	Resuscitation Coding Error - Survivors to Discharge	Age	Days in ICU	Heart Stimulation
Resuscitation Coding Error - Survivors	1			
Age	-0.0153*	1		
Days in ICU	-0.0058*	-0.0870*	1	
Heart Stimulation	-0.1024*	-0.0198*	-0.0646*	1

(*) Significant at 1% level

Table 4 – MIV Bounds with Bounded Data Corruption

	$\eta = 1, \lambda_0 = \lambda_1 = 0.10$		$\eta = 2, \lambda_0 = \lambda_1 = 0.25$	
A. Mean resuscitation per time-period				
	LB	UB	LB	UB
1	0.389	0.458	0.377	0.462
2	0.432	0.517	0.422	0.524
3	0.48	0.523	0.473	0.522
B. Differences in mean resuscitation between time-periods				
Δij	LB	UB	LB	UB
$\Delta 31$	0.022	0.134	0.012	0.145
$\Delta 32$	-0.037	0.092	-0.051	0.100
$\Delta 21$	-0.026	0.037	-0.040	0.051

MIV: Monotone instrumental variable bounds computed under assumption A6:
$E[y^*|t, w_{(-m)}, w_m = i] \geq E[y^*|t, w_{(-m)}, w_m = j]$, for $\forall\, i > j$, and $m = 1, \ldots, 4$.
The four MIVs, w_m, are age (18-74 years, 75-84 years, and 85+ years), ICU days (0 days; (0,1] days; (1,3] days; (3,10] days; and >10 days), heart stimulation, and hospital measurement error quartiles.
λ_0, λ_1: Lower bounds for bounded corruption assumptions A3: $\pi_0 \equiv P(y = 1|Z = 0) \geq \lambda_0$ & A4; $\pi_1 \equiv P(y = 0|Z = 0) \geq \lambda_1$.
η: The magnification factor of the resuscitation error bound for non-survivors compared to its counterpart for survivors to discharge. When $\eta > 1$, Non-survivor measurement error is allowed to exceed survivor error.
$\Delta ij \equiv E[y^*|t = i] - E[y^*|t = j] \in [LB(i) - UB(j), UB(i) - LB(j)]$, for $i, j = 1, \ldots, 3$.