Biased Encouragements and Heterogeneous Effects in an Instrumental Variable Study of Emergency General Surgical Outcomes

Colin B. Fogarty\textsuperscript{a}, Kwonsang Lee\textsuperscript{b}, Rachel R. Kelz\textsuperscript{c}, and Luke J. Keele\textsuperscript{d}

\textsuperscript{a}Operations Research and Statistics Group, MIT Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA; \textsuperscript{b}Department of Statistics, Sungkyunkwan University, Seoul, Republic of Korea; \textsuperscript{c}Center for Surgery and Health Economics, Department of Surgery, Hospital of the University of Pennsylvania, Philadelphia, PA; \textsuperscript{d}Department of Surgery, University of Pennsylvania, Philadelphia, PA

\begin{abstract}
We investigate the efficacy of surgical versus nonsurgical management for two gastrointestinal conditions, colitis and diverticulitis, using observational data. We deploy an instrumental variable design with surgeons’ tendencies to operate as an instrument. Assuming instrument validity, we find that nonsurgical alternatives can reduce both hospital length of stay and the risk of complications, with estimated effects larger for septic patients than for nonseptic patients. The validity of our instrument is plausible but not ironclad, necessitating a sensitivity analysis. Existing sensitivity analyses for IV designs assume effect homogeneity, unlikely to hold here because of patient-specific physiology. We develop a new sensitivity analysis that accommodates arbitrary effect heterogeneity and exploits components explainable by observed features. We find that the results for nonseptic patients prove more robust to hidden bias despite having smaller estimated effects. For nonseptic patients, two individuals with identical observed characteristics would have to differ in their odds of assignment to a high tendency to operate surgeon by a factor of 2.34 to overturn our finding of a benefit for nonsurgical management in reducing length of stay. For septic patients, this value is only 1.64. Simulations illustrate that this phenomenon may be explained by differences in within-group heterogeneity. Supplementary materials for this article, including a standardized description of the materials available for reproducing the work, are available as an online supplement.
\end{abstract}

\section{1. Introduction}
\subsection{1.1. Emergency General Surgery Conditions and Operative Care}
Emergency general surgery (EGS) refers to medical emergencies where the injury is most often endogenous (a burst appendix) while trauma care refers to injuries that are exogenous (a gunshot wound). There are a set of medical conditions known as EGS conditions that are thus distinct from trauma injuries (Shafi et al. 2013; Gale et al. 2014). While operative management is often the primary course of treatment for EGS conditions, it carries additional risk of complications and adverse reactions to anesthesia. For many EGS conditions, nonoperative alternatives including observation, minimally invasive procedures and supportive care exist. Therefore, a critical clinical question faced by the acute care community is determining the overall efficacy of surgical interventions for EGS conditions relative to alternative courses of treatment.

We focus on the effectiveness of operative care for two common EGS conditions: ulcerative colitis and diverticulitis. Diverticula are marble-sized pouches that can form in the lining of the lower part of the large intestine. Diverticulitis occurs when one or more diverticula tear and become inflamed or infected, causing severe abdominal pain, fever, and nausea. Ulcerative colitis is a form of inflammatory bowel disease that causes long-lasting inflammation and ulcers in the lining of the large intestine. The condition most often begins gradually but can worsen, leading to life-threatening complications. Both conditions may be managed through diet and medication, but may require surgery to remove sections of the intestine.

One challenge in studying the effectiveness of surgery for EGS conditions is that randomized comparisons between surgical and nonsurgical care are generally judged to violate the principle of clinical equipoise. Evidence must instead be collected from observational studies. Unsurprisingly, patients who undergo emergency operative treatment are often the sickest patients (Shafi et al. 2013). While statistical adjustment for measures of patient frailty may render treated and control groups comparable on observables, such adjustments are often insufficient due to the potential presence of hidden biases (Keele, Harris, and Grieve 2019).

\subsection{1.2. Instrumental Variables and Physician Preference as an Instrument}
Instrumental variable (IV) methods are a set of statistical techniques that can facilitate the identification of causal effects in the presence of hidden bias. An IV is a variable that
is associated with the treatment of interest but affects outcomes only indirectly through its impact on treatment assignment, providing a haphazard nudge toward taking a treatment (Angrist, Imbens, and Rubin 1996). For a variable to be an instrument, (1) the IV must be associated with the exposure; (2) the IV must be randomly or as-if randomly assigned; and (3) the IV cannot have a direct effect on the outcome (Angrist, Imbens, and Rubin 1996). Under (1)–(3), an IV provides a consistent estimate of a causal effect even in the presence of unobserved confounding between the exposure and the outcome. See Baiocchi, Cheng, and Small (2014) and Imbens (2014) for reviews.

One frequently proposed instrument in comparative effectiveness research is a physician's preference for a specific course of treatment. Here, it is assumed that for a patient of a given level of disease severity, physicians may differ in their preferred course of treatment for primarily idiosyncratic reasons. When a patient receives care from a physician, it is then assumed they are being exposed to an as-if randomized encouragement toward that physician's preferred treatment. Brookhart et al. (2006) first used this type of IV in the context of drug prescriptions. See Brookhart and Schneeweiss (2007) for an overview.

We use a surgeon's preference for operative management as an instrument for whether a patient receives surgery after admission to the emergency department. Following Keele et al. (2018), we construct a measure of a surgeon's tendency to operate (TTO) by calculating the percentage of times a surgeon operates when presented with an EGDS condition. For the surgeons in the data, TTO varied from less than 5% to 95%. Consider two patients in the emergency department with similar baseline characteristics. Being assigned to a patient with a higher or lower TTO may then be viewed as a nudge toward or away from surgery (Keele et al. 2018). Assignment of surgeons with varying TTO to patients is a plausibly haphazard process, with clear impact on whether or not a patient receives operative care.

Our dataset merges the American Medical Association (AMA) Physician Masterfile with all-payer hospital discharge claims from New York, Florida, and Pennsylvania in 2012–2013. The study population consists of all patients admitted for emergency or urgent inpatient care. The data include an identifier for each patient's surgeon, patient sociodemographic and clinical characteristics including indicators for frailty, severe sepsis or septic shock, and 31 comorbidities based on Elixhauser indices (Elixhauser et al. 1998). The data also include measures of the surgeon's age and years of experience. Surgeons were excluded if they did not perform at least 5 operations for one of the 51 specific EGS conditions per year within the two-year study time-frame. This resulted in \( N = 44,082 \) patients presenting with either colitis or diverticulitis. Our primary outcome was the presence of any in-hospital complication (infection, bleeding) as measured by a binary variable. Secondary outcomes are hospital length-of-stay; and in-hospital mortality within 30 days after admission. We also investigated whether the presence of sepsis (an inflammatory response to infection that can result in organ malfunction) acts as a modifier for the effect of surgery.

1.3. Probing the Robustness of an Instrumental Variable Analysis and the Roles of Effect Heterogeneity

For surgeons' tendency to operate to be a valid instrument one must assume that there are no unobserved variables which simultaneously influence the TTO for the surgeon assigned to a patient and the patient's health outcomes. Imagine that surgeons with a higher TTO are more skilled surgeons. If the sickest patients are treated by the most skilled surgeons, then the resulting estimate of surgery's efficacy may be biased. While the data include measures of surgeon age and experience, these are imperfect measures of surgical skill, such that a critic of the study may be unconvinced that bias has been removed.

Suspicious of residual bias motivate sensitivity analyses for IV studies. A sensitivity analysis asks how strong the impact of a hidden bias would have to be on the IV to alter the substantive conclusions of a study. Existing methods for sensitivity analysis share a common limitation: they are derived under a prespecified model of effects for each individual. The most prevalent model for effects is that of proportional doses—that the difference in potential outcomes under encouragement to treatment versus encouragement to control is proportional to the difference in treatment actually received under encouragement to treatment versus encouragement to control (Imbens and Rosenbaum 2005). Under this model, all patients who would undergo surgery if and only if assigned to a surgeon with a high TTO would have the same treatment effect.

Treatment effect heterogeneity refers to variation in the effects of an intervention across individuals. While some of this variation may be ascribed to intrinsically random variation, a portion of it may be predictable on the basis of observed covariates known as effect modifiers. Our work synthesizes and builds upon a growing recent literature on effect heterogeneity in matched designs. Baiocchi et al. (2010) presented a model of inference for matched IV studies with heterogeneous effects, but only does so under an assumption of no unmeasured confounding. Fogarty (2018) presented a general recipe for variance estimation for finely stratified designs but does not address variance estimation within a sensitivity analysis. Fogarty (2020) develops a sensitivity analysis valid under effect heterogeneity, but employs a standard error estimator that can be unduly conservative and does not address the validity of sensitivity analyses using McNemar's test with binary outcomes when effects vary.

We first present a sensitivity analysis for matched IV studies that is asymptotically correct for a weaker null hypothesis on a parameter known as the effect ratio (Baiocchi et al. 2010; Kang et al. 2016). This method obviates concerns that the nonidentified components of effect heterogeneity may conspire with hidden bias to render attempts at sensitivity analysis insufficient. In the particular case of binary outcomes, we prove that sensitivity analysis in Baiocchi et al. (2010), justified therein only under the sharp null, is actually a special case of our method and hence also provides an asymptotically correct sensitivity analysis even with heterogeneous effects. We next highlight how the identified aspects of effect heterogeneity can improve the performance of a sensitivity analysis. We develop improved standard error estimates for use within sensitivity analyses with heterogeneous effects. These standard errors are valid regardless of the truth of an underlying model for effect modification as a function of
observed covariates, but provide more substantial gains should the posited model of effects be correct. By further comparing results between the septic and nonseptic subgroups, we provide an illustration of the role that latent heterogeneity can play in the resulting power of a sensitivity analysis as originally discussed in Rosenbaum (2005).

2. A Matched Instrumental Variable Study and Its Experimental Ideal

2.1. A Near-Far Match

Stronger instruments provide both more informative effect estimates and estimates that are more resistant to hidden bias, providing motivation to actively strengthen an instrument as part of the analysis (Small and Rosenbaum 2008). One approach for strengthening an instrument is a near-far matched design (Baiocchi et al. 2010). Near-far matching creates matched pairs that are similar in terms of observed covariates but highly dissimilar on the values of the instrument. We first categorize surgeons as high versus low tendency to operate depending on whether or not their TTO value is above or below the median TTO (0.604). We then use a form of near-far matching that combines refined covariate balance with near-far matching (Pimentel et al. 2015; Keele et al. 2020). This method allows us to balance a large number of nominal covariates while maintaining a large distance within pairs on the numerical value for TTO.

Each matched pair is exactly matched on both hospital and an indicator of sepsis, such that across-hospital differences in quality of care and fundamental differences in physiology between septic and nonseptic patient cannot bias the analysis. For the remaining covariates, we minimized the total of the within-pair distances on covariates as measured by the Mahalanobis distance and applied a caliper to the propensity score. We further applied near-fine balance constraints for three indicators for surgical volume at the hospital at the time of admission and a dummy variable for whether a patient had a specific disability. Fine balance constrains two groups to be balanced on a particular variable without restricting matching on the variable within individual pairs (Rosenbaum, Ross, and Silber 2007). A near-fine balance constraint returns a finely balanced match when one is feasible, and otherwise minimizes the deviation from fine balance (Yang et al. 2012). As a result, the marginal distribution for these three indicators will be exactly or nearly exactly the same across levels of the IV. Finally, we applied optimal subsetting to discard matched pairs with high levels of imbalance on covariates (Rosenbaum 2012). After matching, we calculated the standardized difference for each covariate, which is the mean difference between matched patients divided by the pooled standard deviation before matching. We attempted to produce absolute standardized differences of less than 0.10, a commonly recommended threshold (Rosenbaum 2010).

Table 1 contains balance statistics before and after matching. Matching moved most absolute standardized differences below 0.1, suggesting successful attenuation of potential biases due to discrepancies in observed surgeon characteristics and available patient physiology metrics. Based upon observed covariates, Table 1 suggests that surgeons with high versus low TTO appear to have been assigned to patients in an asymmetric manner.

2.2. Notation for Randomized Encouragement Designs and Instrumental Variable Studies

We now introduce notation for the experimental design that near-far matching emulates. To assist the reader, essential notation and definitions introduced here and elsewhere are summarized in Section A of the web-based supplementary materials. There are 2n individuals partitioned into n matched pairs. Each pair contains one individual encouraged to take the treatment, denoted $Z_{ij} = 1$, and one individual not encouraged to take the treatment ($Z_{ij} = 0$). In our study, the patient assigned to the surgeon with the higher value of the IV TTO in each pair is the patient for whom $Z_{ij} = 1$. Individual $ij$ has two potential responses: one under encouragement to treatment, $y_{ij}(1)$, and the other under no encouragement, $y_{ij}(0)$. Further, individual $ij$ has two values for the level of treatment actually received: that under encouragement, $d_{ij}(1)$, and under no encouragement, $d_{ij}(0)$. In this work we will consider binary values for $d_{ij}(z)$ for $z = 0, 1$, such that $d_{ij}(z)$ represents whether or not an individual would actually take the treatment when assigned encouragement level $z$. That said, the methods we develop readily extend to the case of continuous $d_{ij}(z)$. The observed outcome and exposure are $Y_{ij} = y_{ij}(Z_{ij})$ and $D_{ij} = d_{ij}(Z_{ij})$, respectively. Let $\mathcal{F} = \{y_{ij}(1), y_{ij}(0), d_{ij}(1), d_{ij}(0), x_{ij}, u_{ij} : i = 1, \ldots, n; j = 1, 2\}$, where $x_{ij}$ represent the observed covariates for each individual and $u_{ij}$ is an unobserved covariate with domain $0 \leq u_{ij} \leq 1$. While a univariate hidden variable constrained to the unit interval may appear restrictive, in fact there always exists a hidden variable between 0 and 1 such that if the practitioner had access to it, adjustment for it would be sufficient for identifying a causal effect (Rosenbaum 2017, sec. 9, footnote 15). We consider finite-population inference, which will be reflected by conditioning upon $\mathcal{F}$ in the probabilistic statements that follow. Inferential statements will pertain to parameters defined in the observed study population, without reliance upon the existence of a hypothetical superpopulation. Let $Z = (Z_{i1}, Z_{i2}, \ldots, Z_{in})$, and let boldface similarly denote other vector quantities such as $u = (u_{i1}, u_{i2}, \ldots, u_{in})$.

Patients are paired using observed covariates $x_{ij}$ such that $x_{i1} \approx x_{i2}$. Despite this, paired subjects may not be equally likely to be exposed to high versus low values of the IV due to discrepancies on an unobserved covariate, that is, $u_{ij} \neq u_{i2}$. Let $\Omega = \{z : z_{i1} + z_{i2} = 1\}$ be the set of the $2^n$ possible values of $z$ of $Z$, and let $\mathcal{Z}$ denote the event $\{Z \in \Omega\}$. In finite-population causal inference the only source of randomness is the assignment of the encouragement $Z_{ij}$. Where $Z$ chosen uniformly as in a randomized experiment, we would have $\Pr(Z = z | \mathcal{F}, \mathcal{Z}) = 1/|\Omega|$ for each $z \in \mathcal{Z}$, where $|A|$ is the number of elements in $A$.

2.3. Conventional IV Assumptions

Here, we outline common assumptions made within IV designs. First, our notation implicitly assumes the stable unit treatment value assumption (Rubin 1980): both the potential treatments received $d_{ij}(z)$ and potential outcomes $y_{ij}(z), z = 0, 1$, depend solely on the value of the instrument for individual $ij$ and are not affected by the value of $Z_{ij}$ for $i'j \neq ij$. Next, we assume that the instrument has a nonzero effect on the treatment received, $n^{-1} \sum_{i=1}^{n} (d_{ij}(1) - d_{ij}(0)) \neq 0$, an assumption which may be
verify with empirical tests known as weak instrument tests. IV designs with stronger instruments are more robust (Small and Rosenbaum 2008; Baiocchi et al. 2010; Keele and Morgan 2016), which motivated our use of near-far matching to increase instrument strength.

For the IV to yield identification of a causal effect, one must assume that instrument assignment is unconfounded once we condition on baseline covariates: \( \pi_{ij} = \Pr(Z_{ij} = 1 \mid x_{ij}) \). For instruments that arise from noncompliance in randomized trials this assumption is satisfied by design, but for naturally occurring IVs this assumption is frequently subject to debate. Since we are focused on emergency admissions it is unlikely that patients themselves are selecting their surgeon, such that assignment to a surgeon with a specific TTO is a plausibly haphazard process. While we employ matching to remove possible bias from haphazard assignment, we cannot ensure that the study results are insensitive to the presence of an unobserved confounding between the instrument and the outcome. This motivates our development in Section 3.3 of a sensitivity analysis to assess robustness to departures from this assumption.

In defining an IV, Angrist, Imbens, and Rubin (1996) further required that the variable satisfies the exclusion restriction: the instrument can only affect the outcome by influencing the treatment received, stated formally as \( d_{ij}(z) = d_{ij}(z') \Rightarrow y_{ij}(z) = y_{ij}(z') \). Informally, under the exclusion restriction, any effect of a surgeon’s TTO on the outcome must only be a consequence of the medical effects of operative management. We bolster this assumption by comparing patients who receive care in the same hospital, which should hold fixed other systemic factors of care that might affect outcomes. Typically, IV studies further invoke a monotonicity assumption of the following form: for all individuals, \( d_{ij}(z') \geq d_{ij}(z) \) for \( z' \geq z \). Together the exclusion restriction and monotonicity imply that the causal estimand being estimated by an IV design is the average treatment effect among individuals who comply with their encouragement to receive or not receive the treatment, referred to as the local average treatment effect. For preference-based instruments the assumption of stochastic monotonicity is more plausible, which yields a different causal estimand where individuals more strongly influenced by the instrument receive larger weight (Small et al. 2017). See Keele et al. (2018) for a detailed discussion of the monotonicity assumption in the context of TTO. As will be described in Section 3.1, inference may still proceed using a variable which violates monotonicity (be it deterministic or stochastic) and/or the exclusion restriction; however, the estimand will no longer be the local average treatment effect.

### Table 1. Balance before and after matching.

| | Unmatched data \( N = 44,082 \) (11,344 septic) | Matched data \( N = 6068 \) (2336 septic) |
|---|---|---|
| | High TTO | Low TTO | St. dif. | High TTO | Low TTO | St. dif. |
| TTO | 0.83 | 0.33 | 2.82 | 0.84 | 0.27 | 4.18 |
| Age | 63.48 | 65.24 | −0.11 | 68.21 | 68.13 | 0.08 |
| No. comorbidities | 3.09 | 3.11 | −0.02 | 3.99 | 3.96 | 0.03 |
| Surgeon age | 54.16 | 52.95 | 0.12 | 53.02 | 51.63 | 0.15 |
| Surgeon experience (years) | 17.07 | 15.45 | 0.16 | 15.59 | 14.70 | 0.09 |
| Female | 0.52 | 0.55 | −0.07 | 0.56 | 0.56 | 0.00 |
| Hispanic | 0.11 | 0.12 | −0.01 | 0.08 | 0.09 | −0.03 |
| White | 0.80 | 0.73 | −0.13 | 0.76 | 0.76 | 0.00 |
| African-American | 0.10 | 0.15 | −0.05 | 0.13 | 0.13 | −0.00 |
| Other racial cat. | 0.09 | 0.12 | −0.03 | 0.11 | 0.11 | 0.00 |
| Sepsis | 0.29 | 0.24 | 0.11 | 0.38 | 0.38 | 0.00 |
| Disability | 0.07 | 0.08 | −0.01 | 0.10 | 0.10 | 0.00 |

Select comorbidities

| Comorbidity | Unmatched data | Matched data |
|---|---|---|
| Congestive heart failure | 0.15 | 0.15 | 0.00 |
| Cardiac arrhythmias | 0.26 | 0.23 | 0.03 |
| Valvular disease | 0.07 | 0.07 | 0.00 |
| Pulm. circulation disorders | 0.05 | 0.04 | 0.01 |
| Peripheral vascular disorders | 0.08 | 0.07 | 0.01 |
| Hypertension, uncomplicated | 0.43 | 0.45 | 0.02 |
| Paralysis | 0.02 | 0.01 | 0.01 |
| Other neurological disorders | 0.08 | 0.08 | 0.00 |
| Chronic pulmonary disease | 0.22 | 0.22 | 0.00 |
| Diabetes, uncomplicated | 0.17 | 0.21 | −0.04 |
| Diabetes, complicated | 0.05 | 0.04 | 0.01 |
| Hypothyroidism | 0.13 | 0.13 | −0.01 |
| Renal failure | 0.15 | 0.17 | −0.02 |
| Liver disease | 0.04 | 0.06 | −0.02 |
| Peptic ulcer, excl. bleeding | 0.01 | 0.02 | −0.01 |
| Hypertension, complicated | 0.14 | 0.16 | −0.02 |

NOTE: Instrument split into high versus low categories at the median value for TTO (0.604). Before and after matching, average values for each variable are given for the high and low groups along with the standardized difference.
proposes a logit form for $\pi_{ij}$ dependent upon a matched-set specific parameter $\kappa_i$ along with the unobserved covariate $u_{ij}$,

$$\log \left( \frac{\pi_{ij}}{1 - \pi_{ij}} \right) = \kappa_i + \log(\Gamma)u_{ij}, \quad (1)$$

for some $\Gamma \geq 1$. This is equivalent to assuming that for two individuals $j$ and $k$ in the same matched set $i$, $\pi_{ij}$ and $\pi_{ik}$ may differ in the odds ratio scale by at most a factor $\Gamma$, that is, $\Gamma^{-1} \leq \pi_{ij}(1 - \pi_{ik})/(\pi_{ik}(1 - \pi_{ij})) \leq \Gamma$; see Rosenbaum (2002, sec. 4, Proposition 12) for a proof. Returning attention to the paired structure by conditioning upon $Z$, (1) implies $1/(1 + \Gamma) \leq \Pr(Z_{ij} = 1 | F, Z) \leq \Gamma/(1 + \Gamma)$ for $i = 1, \ldots, n; j = 1, 2$. In our study, (1) suggests that patients in the same matched set could differ in their odds ratio of being assigned to a surgeon with a high TTO by at most $\Gamma$. $\Gamma = 1$ corresponds to a randomized encouragement design where $u_{ij}$ does not impact the assignment probabilities (Holland 1988), while pushing $\Gamma$ beyond one allows unmeasured confounding to increasingly bias the encouragement to treatment. In terms of the conventional IV assumptions, $\Gamma > 1$ allows for a violation of instrument unconfoundedness conditional upon baseline covariates.

### 3. The Effect Ratio and Methods for Sensitivity Analysis

#### 3.1. The Effect Ratio and Its Interpretation

In an IV design, one estimand of interest is the effect ratio. The effect ratio is

$$\lambda = \frac{\sum_{i=1}^{n} \sum_{j=1}^{2} (y_{ij}(1) - y_{ij}(0))}{\sum_{i=1}^{n} \sum_{j=1}^{2} (d_{ij}(1) - d_{ij}(0))}, \quad (2)$$

where it is assumed that $\sum_{i=1}^{n} \sum_{j=1}^{2} (d_{ij}(1) - d_{ij}(0)) \neq 0$. It is the ratio of two average treatment effects. In our application, the treatment effect in the numerator is the effect of being assigned to a surgeon with high versus low TTO on any one of the health outcomes of interest, while the effect of surgeon TTO on whether or not a patient has surgery is in the denominator. Under the assumption of monotonicity and the exclusion restriction described in Section 2.3, $\lambda$ may then be interpreted as the sample average treatment effect among compliers, that is, the individuals who would undergo surgery if and only if assigned to a surgeon with a higher preference for surgery. Assuming unconfoundedness, Baiocchi et al. (2010) provided a large-sample method for constructing confidence intervals and performing inference for the effect ratio under effect heterogeneity.

Table 2 shows estimated effect ratios and confidence intervals using their method for septic and nonseptic patients on our three surgical outcomes. The estimated length of stay among compliers is much longer—4 days for nonseptic patients and nearly a week for septic patients. Additionally, we find that compliers that underwent surgery were much more likely to experience a post-operative complication. Among septic patients, the risk of a complication was 25% higher, and the lower bound on the 95% confidence interval indicates the risk of a complication was 19% or higher. Finally, the effect of surgery on mortality among compliers is positive but the confidence intervals are wide and include zero.

### 3.2. Sensitivity Analysis for the Effect Ratio With Effects Proportional to Dose

The estimates in Table 2 are valid with heterogeneous effects under the assumption of no hidden bias. Rosenbaum (1996) and Rosenbaum (2002, sec. 5.4) developed methods for exact sensitivity analyses in paired IV studies under the proportional dose model, which states

$$H^{(\lambda_0)}_{\text{prop}} : y_{ij}(1) - y_{ij}(0) = \lambda_0(d_{ij}(1) - d_{ij}(0)), i = 1, \ldots, n; j = 1, 2.$$ 

This extends the model of constant effects common in randomized experiments to randomized encouragement designs. If an individual complies with their encouraged treatment then $y_{ij}(1) - y_{ij}(0) = \lambda_0$, while if an individual defies their encouragement then $y_{ij}(1) - y_{ij}(0) = -\lambda_0$ as $d_{ij}(1) - d_{ij}(0) = -1$. If the encouragement does not influence the treatment received, $y_{ij}(1) - y_{ij}(0)$ is set to zero and hence the exclusion restriction holds. A sensitivity analysis proceeds by finding the worst-case inference for a particular $\Gamma$ in (1). One then iteratively increases the value of $\Gamma$ until the null hypothesis can no longer be rejected. Baiocchi et al. (2010) described how McNemar's test can be used to test $H^{(\lambda_0)}_{\text{prop}}$ with binary outcomes.

The proportional dose model implies that all individuals who would comply with their assigned encouragement have an identical treatment effect $\lambda$, which precludes essential heterogeneity among other forms of effect heterogeneity. Furthermore, with binary outcomes, the only plausible value for this effect would be zero: $y_{ij}(1) - y_{ij}(0)$ can only take the values $-1, 0, 1$, with $\lambda = -1$ or 1 reflecting an extremely strong effect. See Rosenbaum (2002, sec. 5.1) for a related discussion on the restrictiveness of the constant effect model with binary outcomes outside of encouragement designs. We instead develop a sensitivity analysis for the weaker null hypothesis

$$H^{(\lambda_0)}_{\text{weak}} : \sum_{i=1}^{n} \sum_{j=1}^{2} (y_{ij}(1) - y_{ij}(0)) = \lambda_0$$

while leaving individual effects unspecified. The proportional dose hypothesis $H^{(\lambda_0)}_{\text{prop}}$ is simply one element of $H^{(\lambda_0)}_{\text{weak}}$ and need not yield the supremum $p$-value over the composite null.

### 3.3. A Valid Approach With Heterogeneous Effects

Observe that under the null hypothesis, $(2n)^{-1} \sum_{i=1}^{n} \sum_{j=1}^{2} [y_{ij}(1) - y_{ij}(0) - \lambda_0(d_{ij}(1) - d_{ij}(0))] = 0$. Define $\xi_i^{(\lambda_0)} = (Z_{i1} - Z_{i2})(Y_{i1} - Y_{i2} - \lambda_0(D_{i1} - D_{i2}))$ as

| Complication | Septic patients | Nonseptic patients |
|--------------|----------------|-------------------|
| Length of stay | 6.80 [4.4, 9.2] | 4.10 [3.4, 4.8] |
| Mortality | 1.96 [−1.9, 5.0] | 0.59 [−0.5, 1.6] |

NOTE: Point estimates for binary outcomes are risk differences expressed as percentages. Brackets are 95% confidence intervals from inverting the test of Baiocchi et al. (2010).
the encouraged-minus-non encouraged differences in the terms $Y_{ij} - \lambda_0 D_{ij}$, which may be thought of as the observed outcome adjusted for the treatment level received for each individual $ij$. Observe that at $\Gamma = 1$, $E(\zeta(\lambda_0) \mid F, Z) = (1/2) \sum_{y=1}^{Y_{ij}} (y - y_{ij}(0)) - \lambda_0 (d_{ij}(1) - d_{ij}(0))$, and that $\sum_{y=1}^{Y_{ij}} \zeta(\lambda_0)$ hence forms an unbiased estimating equation for $\lambda_0$ in a randomized encouragement design. Further observe that under $H_{prop}^{(\lambda_0)}$, the absolute value $|\zeta(\lambda_0)|$ is fixed across randomizations, with only its sign varying. This observation underpins the use of permutation-based methods for sensitivity inference described in Rosenbaum (2002); however, for other elements of $H_{weak}^{(\lambda_0)}$, the value of $|\zeta(\lambda_0)|$ may itself be random over $z \in \Omega$.

Suppose one wants to conduct a sensitivity analysis for whether the effect ratio equals $\lambda_0$ at level of unmeasured confounding $\Gamma$ without assuming proportional doses. Define $L_{\Gamma; i}$ as

$$L_{\Gamma; i} = \zeta(\lambda_0) - \left(\frac{\Gamma - 1}{1 + \Gamma}\right) |\zeta(\lambda_0)|.$$

The term $\{\Gamma - 1\}/(1 + \Gamma)$ is the worst-case expectation for $\zeta(\lambda_0)$ at $\Gamma$ if the proportional dose model actually held. As proven in Fogarty (2020), this provides an upper bound even if effects are heterogeneous. Based upon the $n$ random variables $L_{\Gamma; i}$, define $SE(L_{\Gamma})$ to be the conventional standard error for a paired design, $SE(L_{\Gamma})^2 = \{n(n-1)\}^{-1} \sum_{i=1}^{n} (L_{\Gamma; i} - L_{\Gamma})^2$, and consider using as a test statistic with a greater-than alternative $T(Z, Y - \lambda_0 D) = L_{\Gamma} / SE(L_{\Gamma})$.

We now construct a reference distribution for the sensitivity analysis. When testing with a greater-than alternative, the sensitivity analysis bounds the upper-tail probability for the employed test statistic. Let $V_{\Gamma; i} (i = 1, \ldots, n)$ be conditionally independent given $F, Z$ and take values $\pm 1$ with $p(r(V_{\Gamma; i} = 1 | F, Z) = \Gamma/(1 + \Gamma)$, and define the random variable $B_{\Gamma; i} = V_{\Gamma; i} |\zeta(\lambda_0) - (\Gamma - 1)/(1 + \Gamma)| |\zeta(\lambda_0)|$. The construction of $B_{\Gamma; i}$ is motivated by the proportional dose model: under $H_{prop}^{(\lambda_0)}$, $B_{\Gamma; i}$ stochastically dominates $L_{\Gamma; i}$ if (1) holds at $\Gamma$. Consider using as a reference distribution the distribution of $B_{\Gamma}/SE(B_{\Gamma})$ given $|\zeta(\lambda_0)|$; call its distribution function $G_{\Gamma}(\cdot)$. The decision whether or not to reject the null hypothesis in a sensitivity analysis at $\Gamma$ using the studentized statistic $L_{\Gamma}/SE(L_{\Gamma})$ is

$$\psi(\alpha, \Gamma) = \{L_{\Gamma}/SE(L_{\Gamma}) \geq G_{\Gamma}^{-1}(1 - \alpha)\},$$

where $G_{\Gamma}^{-1}(q) = \inf\{k : G_{\Gamma}(k) \geq q\}$ is the $q$th quantile of the distribution of $B_{\Gamma}/SE(B_{\Gamma})$ given $|\zeta(\lambda_0)|$. See Algorithm 2 of Fogarty (2020) for more on constructing $G_{\Gamma}(\cdot)$.

**Proposition 1.** Suppose that $H_{weak}^{(\lambda_0)}$ is true and that (1) holds at $\Gamma$. Then, under mild regularity conditions and for $\alpha \leq 0.5$, $\lim_{n \to \infty} E(\psi(\alpha, \Gamma) \mid F, Z) \leq \alpha$.

The proof of Proposition 1 mirrors that of Theorem 2 in Fogarty (2020) and is sketched in the web-based supplement. For testing the effect ratio without assuming proportional doses, regularity conditions are required to ensure that $n \times SE(L_{\Gamma})^2$ converges in probability to a limiting value, and that a central limit theorem applies to $\sqrt{n}(L_{\Gamma} - E(L_{\Gamma} \mid F, Z))$.

**Remark 1.** At $\Gamma = 1$, our method is asymptotically equivalent to that of Baiocchi et al. (2010). The test statistic proposed in that work for inference at $\Gamma = 1$ is precisely $L_{\Gamma}/SE(L_{\Gamma})$, and the methods differ solely in the reference distribution employed. Baiocchi et al. (2010) used the Normal distribution, while our approach uses a reference distribution generated by biased permutations of the observed data under the assumption of proportional doses to maintain exactness under that model for effects.

Under suitable regularity conditions the reference distribution $G_{\Gamma}(k)$ converges in probability to $\Phi(k)$, the CDF of a standard Normal, pointwise at all points $k$. Their method can be viewed as a large-sample approximation of our method at $\Gamma = 1$, and replacing our reference distribution $G_{\Gamma}(k)$ at $\Gamma > 1$ with the standard Normal $\Phi(k)$ would provide the natural extension of their large-sample approach for inference on the effect ratio to a sensitivity analysis.

### 3.4. Binary Responses: Equivalence With McNemar’s Test and Testing Nonzero Nulls

Baiocchi et al. (2010) suggested conducting a sensitivity analysis using McNemar’s test while restricting attention to the narrower null hypothesis $H_{prop}^{(\lambda_0)}$, amounting to a test of Fisher’s sharp null of no effect. One may be concerned that this sensitivity analysis would be misleading if instead $H_{weak}^{(\lambda_0)}$, held, such that the effect ratio equalled zero but Fisher’s sharp null did not hold. As we now demonstrate, McNemar’s test is actually equivalent to the test $\psi(\alpha, \Gamma)$ presented in the previous section with $\lambda_0 = 0$, and hence also provides a sensitivity analysis that is both finite-sample exact for $H_{prop}^{(\lambda_0)}$ and asymptotically correct for $H_{weak}^{(\lambda_0)}$.

**Proposition 2.** Suppose that outcomes are binary and one employs a sensitivity analysis based on McNemar’s test statistic, $T_{M}(Z, Y) = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{ij} Y_{ij}$, using its worst-case distribution under the assumption of $H_{prop}^{(\lambda_0)}$ as described in Rosenbaum (1987) and Baiocchi et al. (2010). Denote the resulting candidate level-$\alpha$ sensitivity analysis by $\psi_{M}(\alpha, \Gamma)$. Then, for any observed vector $Z \in \Omega$ and any vector of binary responses $Y$, $\psi_{M}(\alpha, \Gamma) = \psi(0, \alpha, \Gamma)$. That is, the two sensitivity analyses are equivalent, furnishing identical $p$-values for all $\Gamma$.

The proof of Proposition 2 is presented in the web-based supplement, and requires showing that over elements of $\Omega$, $L_{\Gamma}/SE(L_{\Gamma})$ is a strictly increasing function of McNemar’s statistics $T_{M}(Z, Y) = \sum_{i=1}^{n} d_{ij} Y_{ij}$, the number of encouraged individuals for whom an event occurred. Because $\psi(0, \alpha, \Gamma)$ is valid under the weak null of no effect ratio, the sensitivity analysis for binary outcomes presented in Baiocchi et al. (2010), motivated under Fisher’s sharp null, is also an asymptotically correct sensitivity analysis for the effect ratio equaling zero in the presence of effect heterogeneity. Sensitivity analyses using McNemar’s tests are ubiquitous in paired observational studies with binary outcomes, and Proposition 2 provides a useful fortification for those analyses should a critic be concerned about effect heterogeneity.

The equivalence with McNemar’s test pertains to the test that the effect ratio equals zero. The procedure $\psi(\lambda_0, \alpha, \Gamma)$ remains asymptotically valid for testing $H_{weak}^{(\lambda_0)}$ for $\lambda_0 \neq 0$ even with binary outcomes. Hence, the test $\psi(\lambda_0, \alpha, \Gamma)$ can be inverted to
provide asymptotically valid sensitivity intervals for the effect ratio with binary outcomes.

### 3.5. Application to the Study of Surgical Outcomes for EGS Patients

The estimates in Table 2 assumes that patients are assigned to surgeons in an as-if random fashion after accounting for observed covariates. Since our IV is not the result of a randomized encouragement, we cannot rule out the presence of hidden bias from IV—outcome confounders. In Table 3, we summarize the sensitivity analysis using the changepoint value—the value of \( \gamma \) at which the estimate is no longer statistically significant at \( \alpha = 0.05 \) (Zhao 2019).

Despite the large point estimates for the complications outcome, we find that a somewhat modest amount of confounding could explain this result. For example, for nonseptic patients an unobserved covariate would have to increase the odds of treatment by a surgeon with a high TTO by a factor of 1.59 within matched pairs to overturn our finding of a positive effect ratio at \( \alpha = 0.05 \). Table 3 illustrates that when effect modification is present, hypothesis tests for different subgroups may vary not only in their degree of statistical significance assuming \( \Gamma = 1 \), but also in their sensitivity to bias should the test at \( \Gamma = 1 \) reject. For related results, see also Lee et al. (2018). While the point estimates are larger for septic patients as displayed in Table 2, the \( \gamma \) changepoint values are larger in the nonseptic group for both complication and hospital length of stay. For septic patients an unobserved covariate would have to increase the odds of treatment by a surgeon with a high TTO by a factor of 1.64 within matched pairs to overturn our finding of a positive effect ratio, whereas a factor of 2.36 would be required for nonseptic patients.

Sepsis is but one of many patient-level health characteristics which might influence whether or not surgery will be more effective than nonsurgical alternatives. We next explore how accounting for additional heterogeneity that is predictable on the basis of effect modifiers may further decrease reported sensitivity to hidden bias. After doing so, we present theoretical results to help explain the pattern in Table 3 of smaller treatment effects being considerably less sensitive to hidden bias.

### 4. Improved Standard Errors by Exploiting Effect Modification

#### 4.1. Conservativeness of Finite Population Sensitivity Analysis Under Effect Heterogeneity

Proposition 1 tells us that rejecting through the test \( \psi^{(*)}(\alpha, \Gamma) \) provides an asymptotically valid sensitivity analysis, in the sense that if (1) holds at \( \Gamma \) then the test will commit a Type I error with probability at most \( \alpha \) in the limit. In practice, it could be that \( \psi^{(*)}(\alpha, \Gamma) < \alpha \) under \( H_{\text{weak}}^{(*)} \), resulting in a conservative test. Imagine that we had access to \( E(\tilde{L}_1 | F, Z) \) and \( SD(\tilde{L}_1 | F, Z) \), the true expectation and standard deviation of the conditional distribution \( \tilde{L}_1 | F, Z \). In this case, rejecting the null hypothesis when the test statistic \( \{\tilde{L}_1 - E(\tilde{L}_1 | F, Z)/SD(\tilde{L}_1 | F, Z)\} \) exceeds \( \Phi^{-1}(1 - \alpha) \) would, under mild conditions ensuring that a central limit theorem holds, result in a test statistic with asymptotic level exactly equal to \( \alpha \). Unfortunately, this test statistic cannot be deployed in practice. For \( \Gamma > 1 \) the expectation \( E(\tilde{L}_1 | F, Z) \) is unknown to its dependence on both the vector of unmeasured confounders \( \mathbf{u} \) and the missing potential outcomes, which are not imputed under \( H_{\text{weak}}^{(*)} \). Further, for any value of \( \Gamma \), \( SD(\tilde{L}_1 | F, Z) \) is also unknown when effects are heterogeneous. By using the test statistic \( \tilde{L}_1/SE(\tilde{L}_1) \), our procedure overcomes these difficulties by means of the bounds \( E(\tilde{L}_1 | F, Z) \leq 0 \) and \( var(\tilde{L}_1 | F, Z) \leq E(SE(\tilde{L}_1)^2 | F, Z) \) (Fogarty 2020, Lemmas 2 and 3). While we cannot generally hope to improve upon the bound \( E(\tilde{L}_1 | F, Z) \leq 0 \) for \( \Gamma > 1 \) without risking an anti-conservative procedure, we illustrate that improvements on \( SE(\tilde{L}_1) \) can be attained while preserving the asymptotic level.

### 4.2. A General Construction of Valid Standard Errors for Sensitivity Analysis

Extending Lemma 3 of Fogarty (2020) to encouragement designs, we have

\[
E(SE(\tilde{L}_1)^2 | F, Z) = var(\tilde{L}_1 | F, Z) + \frac{1}{n(n-1)} \sum_{i=1}^{n} (\mu_{\Gamma i} - \tilde{\mu}_{\Gamma i})^2,
\]

where \( \mu_{\Gamma i} = E(L_1 | F, Z) \) is the true, but unknowable, expectation of \( L_1 \) and \( \tilde{\mu}_{\Gamma i} = n^{-1} \sum_{j=1}^{n} H_{\Gamma i} \). The bias in \( SE(\tilde{L}_1) \) thus depends upon the degree of heterogeneity in the expectations \( \mu_{\Gamma i} \). We can interpret the magnitude of the bias as the mean squared error estimate from a regression of \( \mu_{\Gamma i} \) on an intercept column \( \mathbf{1}_n \), a vector containing \( n \) ones. Imagine now that we were able to account for some of the variation in \( \mu_{\Gamma i} \) through covariance adjustment. This would reduce the mean squared error, and hence the degree of bias.

Let \( \mathbf{Q} \) be an \( n \times p\) matrix with \( p < n \) that is constant across all \( z \in \Omega \), and let \( \mathbf{H} = \mathbf{Q}(\mathbf{Q}^T\mathbf{Q})^{-1}\mathbf{Q}^T \) be the orthogonal projection of \( \mathbb{R}^n \) onto the column space of \( \mathbf{Q} \). Let \( h_{ik} \) be the \( (i,k) \) element of \( \mathbf{H} \) and let \( \tilde{L}_{1 i} = L_{1 i}/\sqrt{1-h_{ii}} \). Define \( SE(\tilde{L}_1; \mathbf{Q})^2 \) as

\[
SE(\tilde{L}_1; \mathbf{Q})^2 = \frac{1}{nh_{ii}^2}(\mathbf{I} - \mathbf{H})\tilde{L}_{1 i},
\]

where \( \mathbf{I} \) is the \( n \times n \) identity matrix and \( \tilde{L}_1 = (\tilde{L}_{1 i}, \ldots, \tilde{L}_{1 n})^T \).

**Proposition 3.** For any value of \( \Gamma \), \( E(SE(\tilde{L}_1; \mathbf{Q})^2 | F, Z) = var(\tilde{L}_{1 i} | F, Z) = n^{-2} \tilde{\mu}_{1 i} (\mathbf{I} - \mathbf{H})\tilde{\mu}_{1 i} \geq 0 \), where \( \tilde{\mu}_{1 i} = (\tilde{\mu}_{1 i}, \ldots, \tilde{\mu}_{1 n})^T \), and \( \tilde{\mu}_{1 i} = \mu_{1 i} / \sqrt{1-h_{ii}} \).

The proof of Proposition 3 follows that of Proposition 1 in Fogarty (2018). Regardless of the true value of \( \Gamma \) for which (1)
holds and regardless of the form for the matrix $Q$, $\text{SE}(\bar{L}_r; Q)^2$ will be conservative in expectation for $\text{var}(L_r \mid F, Z)$ so long as $Q$ does not vary across $\Omega$.

### 4.3. A Regression-Based Standard Error

Setting $Q = 1_n$ recovers the usual standard error estimator. Guided by the form of the bias, we see that $Q$ should be chosen to predict the individual-level expectations $\bar{\mu}_r \approx \mu_r$. One choice would be to let $Q = (1, X)$ be a $n \times (k + 1)$ matrix where the $qth$ column of $X$ is $(\bar{x}_{1q}, \ldots, \bar{x}_{nq})^T$, and where $\bar{x}_{i.q} = (x_{i1q} + x_{i2q})/2$, the average of the $qth$ covariate values in the $i$th pair. This reflects a hope that the heterogeneous expectations $\mu_r$ may be linear in the within-pair covariate averages. Under mild regularity conditions, it can be shown that the modified variance estimator may be linear in the within-pair covariate averages. Under mild regularity conditions, it can be shown that the modified variance estimator $\text{SE}(\bar{L}_r; Q_{reg})$ with $Q_{reg} = (1_n, \bar{X})$ is asymptotically never worse than $\text{SE}(L_r)$, regardless of whether or not $\mu_r$ is truly linear in the within-pair covariate averages $\bar{X}$.

**Proposition 4.** Under regularity conditions and defining $0/0 = 1$,

$$\frac{\text{SE}(\bar{L}_r; Q_{reg})^2 - \text{var}(L_r \mid F, Z)}{\text{SE}(L_r; 1_n)^2 - \text{var}(L_r \mid F, Z)} \to 1 - R^2,$$

where $R^2$ is the coefficient of determination from a regression of $\mu_r$ on $(1_n, \bar{X})$.

Sufficient conditions for Proposition 4 are presented along with its proof in the supplementary materials. The conclusion of Proposition 4 would hold replacing $L_r$ with $L_r$, and with $1/n^2$ replaced by $1/(n(n - k - 1))$ in (4). In that case, the resulting standard error estimator would simply be the RMSE from a regression of $L_r$ on $(1_n, \bar{X})$ divided by $\sqrt{n}$.

### 4.4. A Nonparametric Approach: Pairing the Pairs Through Nonbipartite Matching

While the regression-based standard error does not require a properly specified linear model for its validity, the gains are dependent on the extent to which $\mu_r$ is linear in $Q$. While averages of polynomial terms can be included within $X$, alternative nonparametric approaches may be preferred. We now illustrate that a variant of the approach considered in Abadie and Imbens (2008) can also be employed in a finite population sensitivity analysis.

Suppose $n$ is even and consider "pairing the pairs" through the use of nonbipartite matching based on Mahalanobis distances between the average values of the within-pair covariates $\bar{x}_i$. This results in $n/2$ "pairs of pairs," where pairs have been matched to other pairs with similar average values for $\bar{x}_i$. For each pair, let $\mathcal{J}(i)$ be the index of the pair that was matched with the $i$th pair such that $\mathcal{J}(\mathcal{J}(i)) = i$. Consider the variance estimator

$$\text{SE}(\bar{L}_r; Q_{PoP})^2 = \frac{1}{2n^2} \sum_{i=1}^{n} (L_r - L_r, \mathcal{J}(i))^2,$$

which simply squares the differences in the terms $L_r \mid F, Z$ within each pair of pairs.

**Proposition 5.**

$$E(\text{SE}(\bar{L}_r; Q_{PoP})^2 \mid F, Z) - \text{var}(\bar{L}_r \mid F, Z) \geq 1 - \frac{1}{2n^2} \sum_{i=1}^{n} (\mu_r - \mu_r, \mathcal{J}(i))^2.$$

While a proof unique to this estimator follows by expanding the square, as we show in the web-based supplement the result can also be obtained through Proposition 3 by letting $Q_{PoP}$ be the $n \times (n/2)$ matrix with the $qth$ column containing membership indicators for the $qth$ of $n/2$ pairs of pairs. See the supplementary material for how to use an odd number of pairs and for a comparison of the regression and pairs of pairs approaches.

### 5. Improved Sensitivity Analysis

#### 5.1. An Improved Sensitivity Analysis for Sample Average Effects

The standard errors in Section 4 can be used to improve the power of the sensitivity analysis while maintaining the asymptotic level of the test. Consider conducting a sensitivity analysis using $\bar{L}_r/\text{SE}(\bar{L}_r; Q)$ as a test statistic for some matrix $Q$. Let $G_r(\cdot; Q)$ be the distribution of $B_r/\text{SE}(B_r; Q)$ given $|\mathcal{J}|^{(\alpha)}$, where $B_r$ is defined as in Section 3.3. The modified test is $\varphi_Q^{(\alpha)}(\alpha, \Gamma) = \mathbb{I}[\bar{L}_r/\text{SE}(\bar{L}_r; Q) \geq G_r^{-1}(1 - \alpha; Q)]$.

**Proposition 6.** Suppose (1) holds at $\Gamma$ and that $L_r^{(\alpha)}$ holds. Under regularity conditions $\lim_{n \to \infty} E(\varphi_Q^{(\alpha)}(\alpha, \Gamma)) \leq \alpha$, such that the procedure yields a valid level-$\alpha$ sensitivity analysis for the effect ratio with heterogeneous effects. Furthermore, for $Q = Q_{reg}$ or $Q = Q_{PoP}$ and regardless of whether or not the null holds, then under regularity conditions $\lim_{n \to \infty} E(\varphi_Q^{(\alpha)}(\alpha, \Gamma)) = \lim_{n \to \infty} E(\varphi_Q^{(\alpha)}(\alpha, \Gamma))$, such that the new procedure is both less conservative under the null and more powerful under the alternative than the procedure employing the conventional standard error estimator.

The magnitude of the improvement will depend upon the extent to which $\text{SE}(\bar{L}_r; Q)$ improves upon the conventional standard error estimator, $\text{SE}(\bar{L}_r)$. In the web-based supplementary materials, we present a detailed simulation study illustrating the potential benefits conferred by these improved standard error estimators when heterogeneous effects are present. The simulations include scenarios both with no hidden bias and where hidden bias is present. The web-based supplement also includes a sketch of the proof of Proposition 6, a discussion of sufficient conditions, and an algorithm illustrating the procedure $\varphi_Q^{(\alpha)}(\alpha, \Gamma)$.

#### 5.2. Application to the EGS Study

Table 4 contains a comparison of 95% confidence intervals calculated at $\Gamma = 1$ and sensitivity values within the septic and nonseptic subgroups for the hospital length of stay outcome using three standard errors: the conventional standard error for a paired design, the regression-based standard error, and one
formed by pairing together pairs on the basis of observed covariates. In our application, the different standard error estimators did not materially alter the confidence interval lengths. We found that the $\Gamma$ changepoints were also similar across standard errors. Results for the complication and mortality outcomes indicated a similar trend and are omitted here.

This lack of improvement suggests that there is little evidence for substantive effect modification within the sepsis and nonsepsis subgroups on the basis of other patient characteristics. In the web-based supplement, we formally derive an exact omnibus test for effect modification which leverages our improved standard errors. Using this test, for all outcome variables and within both the septic and nonseptic subgroup we fail to reject the null hypothesis that the proportional dose model in Section 3.2 holds, finding little evidence that the IV estimates vary as a function of over 90 additional observed covariates.

6. Heterogeneity and the Power of a Sensitivity Analysis in IV Studies

6.1. Larger But Britter Effect Estimates

While there was no evidence for effect modification within the septic and nonseptic subgroups on the basis of observed covariates, the effect ratio estimates within the septic group were larger than those in the nonseptic subgroup for all three outcomes as presented in Table 2. For the complication outcome, we were able to reject the null of no effect ratio until $\Gamma = 1.56$ and $\Gamma = 1.59$ for the septic and nonseptic groups, respectively. For the length of stay variable, the discrepancy is even more substantial: for two individuals in the same matched set, a hidden variable would need to produce discrepancies in the odds that the individuals were assigned to a high TTO surgeon by a factor of 1.64 in the septic group, while in the septic-group the odds would need to differ by a factor of 2.36. Despite the septic subgroup having the larger estimated effects, the estimates for the nonseptic group proved to be more robust to hidden bias.

We now demonstrate how this phenomenon is reflective of the importance of reduced within-pair homogeneity for reducing sensitivity to hidden bias.

6.2. A Favorable Reality Unknown to the Practitioner

Imagine that $Z_{ij}$ were actually a valid instrument and that the true value for the effect ratio exceeded its hypothesized value. Given that our instrument is not randomly assigned, a critic could always counter that a large effect estimate is merely due to bias from lurking variables invalidating the proposed instrument. Even in this favorable situation of a valid instrument and a positive treatment effect, we would hope that our inferences would prove robust to moderate degrees of hidden bias to protect ourselves against such criticism. Calculations within this section proceed under this favorable setting: there is no hidden bias, there is truly an effect, but the practitioner, blind to this reality, hopes that her inferences perform well under the stress of a sensitivity analysis. We thus assess our method’s ability to discriminate between (1) no treatment effect and hidden bias; and (2) a treatment effect without hidden bias.

Until this point inference has been performed conditional upon $F$, and a generative model for $F$ has been neither required nor assumed. For the calculations in this section, it is convenient to assume a superpopulation model. Because $F$ will be viewed as random, the procedures exploiting effect modification presented in Section 5.1 no longer provide valid tests for the effect ratio (Fogarty 2018, sec. 5). We focus on the sensitivity analysis given in (3) using the conventional standard error estimate, $\psi^{(\lambda_0)}$, which remains valid under a superpopulation model. Following Small and Rosenbaum (2008), we imagine that $\chi_i^{(\lambda_0)}$ is generated as

$$\chi_i^{(\lambda_0)} = \epsilon_i + S_i(\lambda - \lambda_0),$$

where $\epsilon_i = (Z_{i1} - Z_{i2})((Y_{i1} - Y_{i2}) - \lambda(D_{i1} - D_{i2}))$ are the adjusted encouraged-minus-non encouraged differences in responses, and $S_i = (Z_{i1} - Z_{i2})(D_{i1} - D_{i2})$ are the encouraged-minus-non encouraged differences in the treatment received, reflecting the strength of the instrument. Note that this generative model does not imply the proportional dose model, such that the individual-level effects are allowed to be heterogeneous.

We assume that $\epsilon_i$ are iid from a symmetric distribution with mean zero and finite variance $\sigma^2$. The treatments received are assumed binary. We assume that there are no defiers, that the exclusion restriction holds, and that individuals $ij$ are assigned status as compliers, never-takers, and always-takers independently with probability $p_C, p_N$, and $p_A$, respectively. This results in $\Pr(S_i = 1) = p_C + p_A p_N$, $\Pr(S_i = -1) = p_A p_N$ and $\Pr(S_i = 0) = 1 - p_C - 2 p_A p_N$. The true treatment effect among compliers is $\lambda$, while $\lambda_0$ is its value under the null.

6.3. Making Sense of Our Sensitivity Analysis

Through simulation, we now compare the role that the effect size relative to individual-level variability plays in an observational study. We use the length of stay outcome to motivate parameter values, assuming that the estimated effect ratios are actually the true values of $\lambda$ and that the standard deviations of $\chi_i^{(\lambda_0)}$ and $\chi_i^{(\lambda_0)}$ in our sample reflect the true standard deviations for these distribution. We assume there are no defiers, and set the probability of compliance to $p_C = 0.58$ (the estimated compliance rate in our dataset), and set $p_A = p_N = 0.21$. We simulate from the generative model (6) using different sample sizes for the septic and nonseptic groups. The ratio of the variances of $\epsilon_i$ in the septic and nonseptic groups is $\sigma^2 C / \sigma^2 N \approx 8$, and we set $\sigma^2 S / \sigma^2 N$, the ratio of the sample sizes for the septic and nonseptic groups, to also equal 8. We do this so that the variances of the sample averages of the $\epsilon_i$ terms in both groups, $\sigma^2 C / \sigma^2 S$ and $\sigma^2 N / \sigma^2 N$, are approximately equal. The equality of these variances for the sample mean may appear to put inference within the nonseptic subgroup at a disadvantage: as $\lambda_S > \lambda_N$.
The four plots show the power of the test against the null $\lambda = 0$ in the absence of unmeasured confounding as a function of $\Gamma$ for $I = 200, 1000, 2000,$ and $10,000$. The septic subgroup has $n_S = I$ pairs with a larger effect ratio and larger degree of heterogeneity, while the nonseptic has $n_{NS} = I/8$ pairs, a smaller effect ratio and a smaller degree of heterogeneity. The probability of compliance is $0.58$ in all simulations.

Inference at $\Gamma = 1$ for the null of no effect ratio would be more powerful in the septic group. Does this disadvantage carry over to a sensitivity analysis at $\Gamma > 1$?

In each simulation, $n_S$ and $n_{NS}$ observations are drawn from the distribution (6), with parameters $(\lambda_S, \sigma^2_S)$ and $(\lambda_{NS}, \sigma^2_{NS})$ and with $\epsilon_i$ normally distributed. After drawing the samples, we conduct a sensitivity analysis at $\Gamma$, and record whether or not we correctly reject the null of zero effect ratio for each subgroup. We proceed with $n_S = 200, 1000, 2000, 10,000$ and $n_{NS} = n_S/8$ over a range of $\Gamma$ values. We run 10,000 simulations for each setting. Through this simulation, we highlight the different roles that sample size plays for inference assuming no hidden bias ($\Gamma = 1$) and in a sensitivity analysis. Figure 1 presents the results of the simulation study. Each plot shows the performance of the sensitivity analyses in the septic and nonseptic subgroups, varying $n_S$ while maintaining $n_S/n_{NS} = 8$. For instance, the upper-left panel shows the performance in a sensitivity analysis as a function of $\Gamma$ with $n_S = 200$, and $n_{NS} = 25$. Assuming no hidden bias ($\Gamma = 1$), the test in the septic subgroup has higher power as $\lambda_S > \lambda_{NS}$ and $\sigma^2_{NS}/n_{NS} \approx \sigma^2_S/n_S$. As $\Gamma$ increases, we see that despite the smaller effect size the test in the nonseptic subgroup begins to outperform that in the septic group, owing to the fact that $\sigma^2_{NS} < \sigma^2_S$. As $n_S$ increases with $n_S/n_{NS}$ fixed, we see that the larger sample size in the septic-group provides a smaller and smaller benefit. In sufficiently large samples, at $\Gamma = 1$ the tests in both subgroups reject the null of no effect ratio with near certainty. As $\Gamma$ increases in this larger sample regime, the power functions of the tests in both subgroups converge to step functions, with changepoints at $\Gamma = 1.91$ for the nonseptic group and $\Gamma = 1.47$ for the septic group. These values are the design sensitivities (Rosenbaum 2004) for these two generative models: asymptotically a test of no effect will reject with probability zero when $\Gamma$ is above the design sensitivity and one below it; see the web-based supplementary materials for a formula for the design sensitivities along with a numerical evaluation of the impact of compliance on design sensitivity in IV designs. Overall, the comparative performance between the septic and nonseptic subgroups highlights the importance of reduced heterogeneity of responses within pairs for increasing robustness (Rosenbaum 2005).

7. Summary

Assuming the validity of surgeon’s tendency to operate as an instrument, our effect estimates indicate that surgery has an adverse, statistically significant effect on both hospital length of stay and presence of a complication within both the septic and nonseptic subgroups, with a larger estimated effect within the septic group. While the effect estimates for 30-day mortality indicate that surgery may also have an adverse effect, these results were not statistically significant at $\alpha = 0.05$ even assuming no hidden bias ($\Gamma = 1$). The complication outcome
was robust to a moderate degree of hidden bias: for both septic and nonseptic patients, two individuals with the same observed covariates would have to differ in their odds of being assigned to a high versus low TTO surgeon by a factor of roughly 1.6 to overturn the findings of the study. For septic patients a similar odds ratio discrepancy would be required to overturn the finding of an adverse effect on hospital length of stay, while for nonseptic patients the odds would have to differ by nearly 2.4. This helps to frame the debate about what arguments against the validity of TTO as an instrument would actually matter: if a pattern of hidden bias could not influence the odds of assignment of surgeons to patients to this extent, it could not explain away the finding of an effect.

These conclusions required innovations in sensitivity analyses for IV designs. Given surgery’s suspected variation in efficacy due to baseline patient characteristics we developed a sensitivity analysis that does not require effect homogeneity. Instead, we show how a sensitivity analysis for the effect ratio may proceed without specifying a pattern of effect heterogeneity. We then developed variance estimators that can result in improved sensitivity analyses when effects vary with observed covariates. Motivated by the finding that treatment effect estimates in the nonseptic group were more robust to hidden bias despite being of a smaller estimated magnitude, we showcased the role of strong instruments and reduced within-pair effect variation for design sensitivity and the power of a sensitivity analysis.

Supplementary Materials

Appendix  The web-based appendix contains proofs for Propositions 1–6, additional insights and simulations for the improved standard errors, and a discussion of design sensitivity for IV designs. (.pdf file)

R Code for Example and Simulations  An R script illustrates our methods with a simulated dataset. Code for the simulations producing Figure 1 is also included, as is code to both conduct the simulations and reproduce the tables in the supplement. (.R file)

Funding

The dataset used for this study was purchased with a grant from the Society of American Gastrointestinal and Endoscopic Surgeons. Although the AMA Physician Masterfile data is the source of the raw physician data, the tables and tabulations were prepared by the authors and do not reflect the work of the AMA. The Pennsylvania Health Cost Containment Council (PHC4) is an independent state agency responsible for addressing the problems of escalating health costs, ensuring the quality of health care, and increasing access to health care for all citizens. While PHC4 has provided data for this study, PHC4 specifically disclaims responsibility for any analyses, interpretations or conclusions. Some of the data used to produce this publication was purchased from or provided by the New York State Department of Health (NYSDOH) Statewide Planning and Research Cooperative System (SPARCS). However, the conclusions derived, and views expressed herein are those of the author(s) and do not reflect the conclusions or views of NYSDOH. NYSDOH, its employees, officers, and agents make no representation, warranty or guarantee as to the accuracy, completeness, currency, or suitability of the information provided here. The authors declare no conflicts.

References

Abadie, A., and Imbens, G. W. (2008), "Estimation of the Conditional Variance in Paired Experiments," *Annales d’Economie et de Statistique*, 91–92, 175–187. [1632]  
Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [1626,1628]  
Baiochi, M., Cheng, J., and Small, D. S. (2014), "Instrumental Variable Methods for Causal Inference," *Statistics in Medicine*, 33, 2297–2340. [1626]  
Baiochi, M., Small, D. S., Lorch, S., and Rosenbaum, P. R. (2010), "Building a Stronger Instrument in an Observational Study of Perinatal Care for Premature Infants," *Journal of the American Statistical Association*, 105, 1285–1296. [1626,1627,1628,1629,1630]  
Brookhart, M. A., and Schneeweiss, S. (2007), "Preference-Based Instrumental Variable Methods for the Estimation of Treatment Effects: Assessing Validity and Interpreting Results," *The International Journal of Biostatistics*, 3, 14. [1626]  
Brookhart, M. A., Wang, P., Solomon, D. H., and Schneeweiss, S. (2006), "Evaluating Short-Term Drug Effects Using a Physician-Specific Prescribing Preference as an Instrumental Variable," *Epidemiology*, 17, 268. [1626]  
Elzhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998), "Comorbidity Measures for Use With Administrative Data," *Medical Care*, 36, 8–27. [1626]  
Fonteyn, C. B. (2018), "On Mitigating the Analytical Limitations of Finely Stratified Experiments," *Journal of the Royal Statistical Society*, Series B, 80, 1035–1056. [1626,1631,1633]  
Fogarty, C. B. (2018), "On Mitigating the Analytical Limitations of Finely Stratified Experiments," *Journal of the Royal Statistical Society*, Series B, 80, 1035–1056. [1626,1631,1633]  
Holland, P. W. (1988), "Causal Inference, Path Analysis, and Recursive Structural Equation Models," *Sociological Methodology*, 18, 449–484. [1629]  
Imbens, G. (2014), "Instrumental Variables: An Econometrician's Perspective," *Statistical Science*, 29, 323–358. [1626]  
Imbens, G. W., and Rosenbaum, P. (2005), "Robust, Accurate Confidence Intervals With a Weak Instrument: Quarter of Birth and Education," *Journal of the Royal Statistical Society*, Series A, 168, 109–126. [1626]  
Kang, H., Krueger, B., May, J., and Small, D. S. (2016), "Full Matching Approach to Instrumental Variables Estimation With Application to the Effect of Malaria on Stunting," *The Annals of Applied Statistics*, 10, 335–364. [1626]  
Keele, L., Harris, S., and Grieve, R. (2019), "Does Transfer to Intensive Care Units Reduce Mortality? A Comparison of an Instrumental Variables Design to Risk Adjustment," *Medical Care*, 57, e73–e79. [1625]  
Keele, L. J., Harris, S., Pimentel, S., and Grieve, R. (2020), "Stronger Instruments and Refined Covariate Balance in an Observational Study of the Effectiveness of Prompt Admission to the ICU," *Journal of the Royal Statistical Society*, Series A, 183, 1501–1521, DOI: 10.1111/rssa.12437. [1627]  
Keele, L. J., and Morgan, J. (2016), "How Strong Is Strong Enough? Strengthening Instruments Through Matching and Weak Instrument Tests," *Annals of Applied Statistics*, 10, 1086–1106. [1628]  
Keele, L. J., Sharocky, C. E., Sellers, M. M., Wirtalla, C. J., and Kelz, R. R. (2018), "An Instrumental Variables Design for the Effect of Emergency General Surgery," *Epidemiologic Methods*, 7, 20170012. [1626,1628]  
Lee, K., Small, D. S., Hsu, J. Y., Silber, J. H., and Rosenbaum, P. R. (2018), "Discovering Effect Modification in an Observational Study of Surgical Mortality at Hospitals With Superior Nursing," *Journal of the Royal Statistical Society*, Series A, 181, 535–546. [1631]  
Pimentel, S. D., Kelz, R. R., Silber, J. H., and Rosenbaum, P. R. (2015), "Large, Sparse Optimal Matching With Refined Covariate Balance in an Observational Study of the Health Outcomes Produced by New Surgeons," *Journal of the American Statistical Association*, 110, 515–527. [1627]  
Rosenbaum, P. R. (1987), "Sensitivity Analysis for Certain Permutation Inferences in Matched Observational Studies," *Biometrika*, 74, 13–26. [1630]
Rubin, D. B. (1980), "Comment on 'Randomization Analysis of Experimental Data: The Fisher Randomization Test,' Journal of the American Statistical Association, 75, 591–593. [1627]
Shafii, S., Aboutanos, M. B., Agarwal, S., Jr., Brown, C. V., Crandall, M., Feliciano, D. V., Guillamondegui, O., Haider, A., Inaba, K., Osler, T. M., and Ross, S. (2013), "Emergency General Surgery: Definition and Estimated Burden of Disease," Journal of Trauma and Acute Care Surgery, 74, 1092–1097. [1625]
Small, D., and Rosenbaum, P. R. (2008), "War and Wages: The Strength of Instrumental Variables and Their Sensitivity to Unobserved Biases," Journal of the American Statistical Association, 103, 924–933. [1627,1628,1633]
Small, D. S., Tan, Z., Ramsahai, R. R., Lorch, S. A., and Brookhart, M. A. (2017), "Instrumental Variable Estimation With a Stochastic Monotonicity Assumption," Statistical Science, 32, 561–579. [1628]
Yang, D., Small, D. S., Silber, J. H., and Rosenbaum, P. R. (2012), "Optimal Matching With Minimal Deviation From Fine Balance in a Study of Obesity and Surgical Outcomes," Biometrics, 68, 628–636. [1627]
Zhao, Q. (2019), "On Sensitivity Value of Pair-Matched Observational Studies," Journal of the American Statistical Association, 114, 713–722. [1631]