Who Should Get Blood? Personalizing Medicine with Heterogeneous Treatment Effects

Jason Abaluck, Leila Agha and David Chan

PRELIMINARY AND INCOMPLETE *

March 6, 2019

Abstract

Randomized clinical trials typically estimate average treatment effects within selected populations. With modern medical records and quasi-experimental research designs, it is now possible to estimate heterogeneous treatment effects using vastly more information. We present evidence that applying such estimates to inform clinical decisions could lead to large health benefits, outperforming both status quo physician decisions and strict applications of current medical guidelines. We study blood transfusion decisions for 1.6 million patients with anemia receiving inpatient care at Veteran Health Administration hospitals from 2000-2015. We first show that observed treatment decisions are largely invariant to a wide array of observable patient characteristics, with the exception of blood hemoglobin levels. Treatment effects estimated by naively assuming unconfoundedness vary substantially with patient characteristics. Using instruments based on quasi-random assignment of patients to physicians, we find that much of the measured heterogeneity in the naive "observational" treatment effects reflects heterogeneity in underlying causal effects rather than selection. In counterfactual simulations, we find that better targeting the existing number of transfusions would reduce the total 30-day mortality rate in the study population by 1.1 percentage points, from a base of 9% .

*Abaluck: Yale University and NBER, jason.abaluck@yale.edu. Agha: Dartmouth College and NBER, leila.agha@dartmouth.edu. Chan: Stanford University and NBER, david.chan@stanford.edu. Thanks to Brian Abaluck, Joe Altonji, Joshua Aronson, Pinar Karaca Mandic, Danielle Li, Erzo Luttmer, Nicole Maestes, Adam Sacarny, Jon Skinner, Doug Staiger, Chris Walters, and Kevin Williams as well as seminar participants at Dartmouth, Wharton, the National Bureau of Economic Research, the Caribbean Health Economics Symposium, the National Tax Association annual meeting, and Yale. Funding for this work was provided by NIA Grant Number T32-AG0000186 to the NBER.

1 Introduction

Recent advances in machine learning and genetics, as well as the widespread adoption of electronic medical records, make possible more personalized assessments of the benefits of alternative treatments (Collins and Varmus 2015). Efforts to personalize medicine face a fundamental challenge: Existing randomized experiments are not powered to uncover heterogeneity in treatment effects, yet attempts to do so using observational data are confounded by selection into treatment based on unobservable determinants of outcomes.

Further, bringing evidence-based clinical guidelines into practice can introduce an important tradeoff. Physicians often possess more knowledge of patients' clinical conditions and potential benefits from treatment along dimensions not fully observable to analysts who construct guideline recommendations. As a result, strict adherence even to sophisticated, evidence-based rules may perform badly if physicians would otherwise use privately observed information to target treatment to patients who will benefit most.

In this paper, we develop and apply methods to estimate heterogeneous treatment effects in settings where physicians select patients into treatment based on both observable and unobservable patient characteristics. We use the model to evaluate the 433,517 blood transfusions administered to patients admitted to VA hospitals from 2000 to 2015. We compare mortality under the status quo transfusion decisions and in counterfactual scenarios where total transfusions are held constant, but transfusions are reallocated to patients who would benefit most.

While health economics has had a long-standing interest in overuse of care, there has been a new and growing literature on whether care is being allocated efficiently (Chandra and Staiger 2007; Chandra and Staiger 2017; Chandra et al. 2016; Currie et al. 2016). Recent work by Abaluck et al. (2016) found that, in the case of a common diagnostic imaging test, the welfare costs of failing to allocate the test to the highest yield patients is several times larger than the welfare costs of overuse. In this paper, we consider the more general challenge of optimal treatment.[1] We measure the extent of misallocation and further ask whether this misallocation could be mitigated by stringent application of current guidelines or by hypothetical guidelines that incorporate more clinical detail. Additionally, our framework allows us to identify whether physician discretion leads to better treatment decisions because physicians use private information in deciding who

[1] Analytically, the most important difference is that the yield of a diagnostic test–whether it is positive or negative–is observed for each tested patient. This would be analogous to a case where the effect of treatment on the treated was known and we wanted to recover the relationship between patient characteristics and average treatment effects. In our setting here, the ETT must likewise be estimated.

to treat or worse treatment decisions because physicians fail to optimally weight factors observable to both physicians and analysts.

Our model builds on several recent innovations in applied econometrics. First, we apply an important insight from the recent literature on school "value-added' that researchers may combine observational and quasi-experimental estimates to construct mean-squared error minimizing value-added estimates (Chetty et al. 2014; Angrist et al. 2017; Hull 2018). Rather than estimate the benefits of a particular school (as in the school value-added literature), we seek to estimate the benefits for every possible combination of observable patient characteristics. To do so, we further draw on the recent literature on estimating heterogeneous treatment effects using machine learning techniques (Belloni et al. 2014; Athey and Imbens 2016; Wager and Athey 2017; Asher et al. 2016). Finally, in the medical context, it is likely that physicians select patients into treatment based on characteristics that at once are unobservable to the econometrician but also relate to treatment effects, so we distinguish between average and marginal treatment effects for a given set of observable characteristics explicitly into our model, following Heckman and Vytlacil (2005).

Our analysis proceeds as follows. We begin by comparing physician treatment decisions with treatment effects measured in OLS or machine learning models that do not explicitly account for selection on unobservable characteristics. Our first finding is that physicians consider hemoglobin level in deciding which patients to transfuse, but their decisions otherwise do not seem to use a large number of other observed patient attributes. Although these patient attributes are unused in decision-making, comparing 30-day mortality for treated and untreated patients suggests observational treatment effects that vary substantially depending on these unused characteristics. For example, assuming that observational treatment effects are causal, patients with fewer recent emergency department or hospital admissions derive a larger mortality reduction from transfusions; nevertheless, patients in the 1st quartile of recent inpatient visits benefit nearly four times as much as patients in the 4th quartile but the two groups are almost equally likely to be transfused.

Next, we develop a structural model which permits treatment effects to vary with observable characteristics as well as unobservable selection into treatment. Our model relies on quasi-experimental variation in the assignment of patients to physicians, who may have different propensities to transfuse blood, within a hospital and service. This experimental design is similar to a growing "judges-design" literature that exploits random assignment to a decision-maker to estimate a treatment effect on a population of compliers who are induced by the decision-maker into treatment (Aizer and Doyle 2013;

Kling 2006). Consistent with other papers in this literature, we examine the validity of our approach with respect to two key assumptions. First, we assess balance in the characteristics of patients assigned to physicians with different treatment propensities. Our checks suggest that patients appear to be quasi-experimentally assigned to physicians with different treatment propensities. Second, we assess a monotonicity assumption that requires no defiers when patients are assigned to higher-propensity physicians, by showing that subgroups of patients with different observable characteristics show the same first-stage relationship between physician treatment propensity and patient treatment (Bhuller et al. 2016; Dobbie et al. 2018).

In order to assess the relationship between observational treatment effects and treatment effects implied by our quasi-experimental design, we run regressions interacting the observational treatment effect with our physician treatment-propensity instrument. Intuitively, a coefficient of unity on this interaction suggests that the measured heterogeneity reflects heterogeneous causal effects while a coefficient of 0 suggests that it has no bearing on actual treatment effects, due to either selection bias or measurement error in the original observational approach to measuring treatment effects. We find that roughly half of the heterogeneity in naively estimated observational treatment effects reflects true differences in quasi-experimental treatment effects. In counterfactual simulations, we find that better targeting the existing number of transfusions would reduce total mortality in the study population by 1.1 percentage points, from a base of 9 percentage points.

Our model can also be applied to a wide variety of settings outside of medical care, in which the goal is to estimate heterogeneous treatment effects using a large number of case attributes, some quasi-experimental variation, and a larger amount of observational data. For example, problems of a firm deciding which workers would benefit the most from further training, or a policymaker trying to determine which community would benefit most from a new schooling investment share the underlying features that we model. One would like to estimate heterogeneous treatment effects by comparing treated and untreated beneficiaries, but to do so one must allow for potential selection. Selection impacts both the way in which we interpret the observational data as well as the balance between explicit policies and policies that make use of discretion. The benefit of strict guidelines will depend, in part, on how well current experts are making these allocation decisions and how important unobservable characteristics are for predicting heterogeneity in returns.

2 Data and Observational Evidence

2.1 Data

Blood transfusions are among the most common medical interventions, with approximately 11 million units transfused annually in the United States (Carson et al. 2017). The single most important factor determining transfusion decisions is the patient's blood hemoglobin level, which reflect his red blood cell count. There are a wide range of underlying causes of anemia, defined as low hemoglobin levels, many of which are frequently seen among hospitalized patients. Some conditions, such as traumatic injury or gastrointestinal bleeding, may need prompt transfusion to save a patient's life. Other conditions, such as iron deficiency, intravenous fluid administration, or repeated laboratory testing, are either chronic or do not reflect any urgent need for more blood.

Furthermore, clinicians often recognize that patient characteristics unrelated to the root cause of anemia should influence the benefit or cost of transfusing blood. For example, for patients with coronary artery disease, inadequate blood levels can lead to myocardial ischemia or infarction, and blood transfusions may improve mortality for these patients to a greater degree, regardless of the cause of anemia. On the other hand, patients with congestive heart failure may be sensitive to volume overload from unnecessary transfusions. Transfusing these patients could actually worsen mortality. Clinical guidelines incorporate some of this reasoning but usually do not make hard recommendations for transfusion thresholds. Transfusion thresholds are almost always stated as hemoglobin levels, and there has been a wide range of conflicting thresholds recommended by different recent guidelines. While many randomized trials of transfusion strategies have been performed, even a recent meta-analysis of 31 trials had a large confidence interval around the potential benefits of more "liberal" transfusion policies, running from a 20% reduction to a 16% increase in 30-day mortality associated with more liberal transfusion strategies (Carson et al. 2017). The estimated benefits of the marginal transfusion are even more uncertain.

This study relies on electronic health records from the Veteran's Health Administration (VHA) to construct a detailed database of hospitalized patients who may receive a blood transfusion. We collected data on each patient's clinical characteristics relevant for transfusion, whether the patient was transfused during the hospital stay, and patient outcomes. These data are collected from the VA Corporate Data Warehouse, which includes inpatient visits, bed locations and ward assignments, physician orders, laboratory tests, diagnoses, and demographics. The data span the years 2000-2014.

We select patients who are admitted in the hospital in an acute-care bed section

(e.g., general medicine, surgery, intensive care unit). Out of this cohort of 7.9 million admissions, we select 2.7 million admissions for patients who had a minimum hemoglobin level that would have placed them in a range, between 6 mg/dl and 11 mg/dl, with any reasonable clinical uncertainty about the need for transfusion. We assign each admission to the first attending physician of record.[2] In order to exploit variation across physicians in transfusion propensities, we restrict attention to physicians who had at least 100 admissions within this hemoglobin range practicing in bed-sections that treated at least 3,000 patients. This restriction results in our final analytic sample of 1.6 million admissions and 4,778 physicians.

To capture whether or not patients are transfused during an admission, we rely on physician order entry data (available from year 2009 onward) and procedure codes (from the years 2000 to 2008), similar to those available in billing data. We observe rich patient characteristics including 31 comorbid conditions. We also construct spline variables describing past VA health care use over the past year: number of emergency department, inpatient, and primary care visits. We capture demographic information (race, ethnicity and spline in age, race). We construct measures of average, maximum, minimum, and missing indicator for vital signs (pulse rate, temperature, respiration rate, and blood pressure) over the hospital stay. Further, we consider laboratory tests related to the patient's blood level (hemoglobin), the risk of bleeding (platelet count, coagulation studies, liver function tests), and potential myocardial ischemia (troponin, creatine kinase-muscle/brain). For each laboratory test, we calculating average, maximum, and minimum values, as well as a missing indicator.

Our primary outcome is 30-day patient mortality, which the VHA tracks using its own records and data from Medicare and the Social Security Administration. The baseline mortality rate is 9.4%. Our key laboratory test of interest for blood transfusions is a patient's hemoglobin level; we focus primarily on the minimum hemoglobin level observed during the hospital stay. Our primary treatment variable is an indicator for whether the patient was transfused during his stay, which we obtain from physician order entry records. As we will describe below, our quasi-experimental design relies on the identity of the physician identity who is assigned the patient. We also observe this identity from internal patient assignment records. Additional summary statistics are reported in Table 1.

[2]In our full sample, 16.7% of patients lack any reported hemoglobin test during their stay; we exclude these patients from our analysis. We are left with 5.1 million tested patients. While patients with hemoglobin levels greater than 11 comprise 73% of all tested patients, they represent only 10% of blood transfusions.

2.2 Observational Evidence

In this section, we consider how patient characteristics appear correlated with "observational" estimates of the treatment effect of blood transfusion and the extent to which physicians tailor transfusion decisions using these observable patient characteristics. The relationship between treatment effects and the patient selection for transfusion underlies our central question of the efficiency of physician treatment decisions relative to those implied by a statistical algorithm.

Denote $D_i \in \{0, 1\}$ as an indicator for whether a patient in admission i was transfused during the admission. We are interested in counterfactual mortality for this patient with or without transfusion, which we denote as

$$Y_i\left(D_i\right) = \mu_i D_i + a_i. \tag{1}$$

$Y_i(1)$ is the potential mortality that patient i would have under transfusion, while $Y_i(0)$ is the potential mortality that the same patient would have under no transfusion. a_i can be thought of as the patient's mortality outcome absent transfusion (i.e., $Y_i(0)$), and μ_i can be thought of as the patient-specific transfusion effect on mortality (i.e., $Y_i(1) - Y_i(1)$). We only observe one realized outcome, corresponding to whether the patient is transfused or not: $Y_i = \left(1 - D_i\right) Y_i\left(0\right) + D_i Y_i\left(1\right)$.

To mitigate selection bias, we can write

$$a_i = \mathbf{X}_i'\gamma + \mathbf{T}_i'\eta + \zeta_{\ell(i)} + \varepsilon_i, \tag{2}$$

as a regression of a_i on controls $\mathbf{X}_i$, time dummies $\mathbf{T}_i$, and clinical location dummies $\zeta_{\ell(i)}$ for each location ℓ. For $\mathbf{X}_i$, we use a rich vector of patient characteristics, including cubic splines of minimum hemoglobin level during the admission, patient demographics, prior medical conditions, and cubic splines of prior visit utilization in outpatient, emergency department, and inpatient settings. By definition, ε_i is uncorrelated with $\mathbf{X}_i$, $\mathbf{T}_i$, and ℓ.

We make an *unconfoundedness* (i.e., "selection-on-observables") assumption that potential outcomes are independent with D_i, conditional on observed patient characteristics, time dummies, and the location of treatment:

$$\left\{Y_i\left(0\right), Y_i\left(1\right)\right\} \perp\!\!\!\perp D_i \left|\mathbf{X}_i, \mathbf{T}_i, \ell\left(i\right)\right., \tag{3}$$

allows us to estimate the treatment effect μ_i in the causal model of Equation (1) by OLS. Finally, in order to measure heterogeneity in observational treatment effects as a function of patient characteristics, we approximate μ_i as a linear function of $\mathbf{X}_i$, or

$\mu_i = f(\mathbf{X}_i) = \beta_0 + \mathbf{X}_i'\beta_1$, which yields

$$Y_i = D_i\left(\beta_0 + \mathbf{X}_i'\beta_1\right) + \mathbf{X}_i'\gamma + \mathbf{T}_i'\eta + \zeta_{\ell(i)} + \varepsilon_i. \tag{4}$$

Under the assumption in Equation (3), $\hat{\mu}_i = \hat{\beta}_0 + \mathbf{X}_i'\hat{\beta}_1$ is an unbiased estimate of the treatment effect conditional on $\mathbf{X}_i$. We label this estimate as the *observational treatment effect*, conditional on $\mathbf{X}_i$.

The plausibility of the assumption in Equation (3) depends on the richness of patient observable characteristics in $\mathbf{X}_i$ and the extent to which unobserved characteristics are captured by time and clinical location. In later sections of this paper, we will devote substantial attention to the issue of selection on unobservable characteristics.

Here, we simply ask whether physicians respond to *observable* characteristics that appear to predict μ_i. Thus, we examine how the probability of treatment is related to estimated observational treatment effects $\hat{\mu}_i = \beta_0 + \mathbf{X}_i'\beta_1$. Because we are interested in this relationship within time and clinical location, we regress the transfusion indicator variable, D_i, on $\mathbf{T}_i$ and dummies for $\ell(i)$, and focus on the residuals $\tilde{D}_i$. Panel A of Figure 1 shows this relationship, conditional on time and clinical locations, taking into account patient characteristics of demographics, comorbidities, and prior utilization that can be easily observed in claims data. The observational treatment effects of transfusion on mortality is negative for about 80% of admissions in our sample and ranges from -4.9 percentage points for patients at the 5th percentile to 2.0 percentage points to patients at the 95th percentile. The difference in mean residualized observational treatment effects between the highest and lowest ventiles is 8.7 percentage points. The corresponding difference in mean transfusion probabilities between these groups of patients is -57.1 percentage points, which suggests that physicians are quite responsive to OLS treatment effects, increasing the probability of treatment as transfusion is more likely to save lives.

However, in Panel B of Figure 1, we investigate the relationship between observational treatment effects and transfusion that is orthogonal to hemoglobin. Hemoglobin is particularly salient to physicians, and the vast majority of guideline recommendations are based on hemoglobin level. Specifically, we residualize transfusion and observational treatment effects by cubic splines of minimum hemoglobin, in addition to the time and clinical location dummies we considered for Panel A. We find that once we account for hemoglobin, physicians are much less responsive to observational treatment effects, despite the substantial residual variation in treatment effects. The difference in mean residualized observational treatment effects between the highest and lowest ventiles is 6.9 percentage points, but the corresponding difference in mean residualized transfusion probabilities is only -10.9 percentage points.

Table 2 presents results more systematically by type of patient characteristics $\mathbf{X}_i$ that we include in Equation (4). Corresponding to claims-observable characteristics used in Figure 1, we consider "base" characteristics of demographics, comorbidities, and prior utilization. We also consider characteristics that are observable in electronic health records data, such as those in the VHA. These notably include laboratory test values and vital signs. We consider only five laboratory tests related to cardiac markers and blood counts (i.e., creatine kinase-muscle/brain or CK-MB, troponin I, troponin T, white blood count, and platelets). We consider vital signs of heart rate, systolic and diastolic blood pressure, oxygen saturation, and respiratory rate. For each of these categories of characteristics, or combinations of them, we estimate observational treatment effects specified in Equation (4). We also order patients by their treatment effects and calculate average treatment effects in the first and fifth quintiles, as well as treatment probabilities corresponding for these patients.

As shown in Table 2, we find that overall average observational treatment effects become more negative as we include more patient characteristics in Equation (4). This suggests that physicians select patients with higher baseline mortality for treatment, so that controlling for patient characteristics reveals a higher mortality benefit from transfusion. Consistent with Figure 1, we find large variation in observational treatment effects between the first and fifth quintiles, regardless of the patient characteristics that we use. However, laboratory test values and vital signs are particularly informative, such that either of these categories allows for similar or more discrimination among treatment effects than all of the base characteristics observable in claims data. This suggests that access to real-time electronic health data yields significant gains in identifying heterogeneous treatment effects. While there are large differences in treatment probabilities between patients grouped in the first and last quintile of treatment effects (Panel A), these differences are mostly eliminated when controlling for hemoglobin (Panel B). When considering patient characteristics by category, the differences in treatment probabilities are even smaller when controlling for other patient characteristics, but the heterogeneity in treatment effects notably remains substantial (Panel C).

3 Selection and Heterogeneous Treatment Effects

So far, we have seen that doctors principally consider hemoglobin in their decision of which patients to transfuse despite the fact that many other variables predict variation in observational treatment effects. If we interpret the heterogeneity in these treatment effects as causal, we could substantially lower mortality by reallocating treatments to

the patients who would benefit most.

But to what extent does variation in observational treatment effects reflect true differences in underlying causal effects? Variation in observational treatment effects may not be useful to policymakers for two reasons: First, even with a rich set of observational characteristics, the unconfoundedness assumption in Equation (3) may be violated. In this case, differences in realized outcomes, holding fixed observed patient characteristics, may also reflect selection bias. Second, even with a large number of observations and unconfoundedness, variation in observational treatment effects will include random noise. In this section, we develop and estimate a model of selection with heterogeneous treatment effects to measure the degree to which heterogeneity in observational treatment effects reflects policy-relevant treatment effect heterogeneity. The model shows how we can distinguish between these hypotheses using quasi-experimental variation in the assignment of patients to physicians. The model leads to a parsimonious test of the extent to which observational treatment effects heterogeneity reflects true treatment effect heterogeneity. The model also characterizes selection on gains: are treated patients likely to benefit more, holding fixed observable patient characteristics?

Given parametric assumptions, we can recover all of the marginal treatment effects necessary to simulate the mortality impact of counterfactual treatment rules. We consider both the optimal rule given the information in observational treatment effects as well as an optimal function of observable characteristics in a "moment forest" model.

3.1 Model of Selection

In the inpatient setting, assignment of patients i to physicians j is plausibly quasi-random within groups of similar physicians on the same service $g(j)$ and within a period of time t (e.g., a month or quarter). Our model exploits this quasi-experimental variation that in part drives selection. Following Vytlacil (2002), we consider a threshold-crossing representation of selection:

$$D_i = 1\left\{\psi\left(Z_i, X_i, g(j), t\right) \geq V_i\right\}, \tag{5}$$

where Z_i is a patient-specific instrument, X_i are patient observable characteristics, and V_i is a latent variable distributed i.i.d., conditional on $X_i = x$, according to some cumulative distribution function $F_{V|x} : \mathbb{R} \to [0, 1]$. Heckman and Vytlacil (2005) show that this model can be equivalently stated according to propensity scores:

$$D_i = 1\left\{P\left(X_i, Z_i, g(j), t\right) \geq U_i\right\}, \tag{6}$$

where $P(Z_i, X_i, g(j), t) \equiv F_{V|x}(\psi(X_i, Z_i, g(j), t))$ is the propensity score and $U_i \equiv F_{V|x}(V_i)$ follows a uniform distribution. Patients with a given set of observable characteristics are ordered by their tendency to be treated and U_i denotes quantiles of this ordering. For example, if $U_i = 0.4$, patient i will be treated if 41% of patients with that set of observable characteristics are treated but not if 39% of such patients are treated.

Denoting $D_i(z)$ as treatment status of patient i when $Z_i = z$, we impose the following two conditions for the validity of our instrument under heterogeneous treatment effects. First, we assume that, conditional on $g(j)$ and t, the instrument is independent of potential outcomes and the unobserved latent variable, and that the instrument does not otherwise affect potential outcomes except through its relationship with treatment status (*independence* and *exclusion*):

$$(Y_{i1}, Y_{i0}, U_i) \perp\!\!\!\perp Z_i \,|\, g(j), t \,, \tag{7}$$

where Y_{i1} and Y_{i0} are potential outcomes under transfusion and no transfusion, respectively. Intuitively, our instrument Z_i uses variation in treatment that is driven by the quasi-random assignment of patients to physicians, which we detail below in 3.4. Second, we assume a *monotonicity* (or *uniformity*) condition (Imbens and Angrist 1994; Heckman and Vytlacil 2007):

$$D_i(z) \geq D_i(z') \text{ for all } i, \text{ or } D_i(z) \leq D_i(z') \text{ for all } i, \text{ for any } z \text{ and } z'. \tag{8}$$

This condition is embedded in the notation in Equation (6) and states that there can be no "defiers" when patients are assigned to one instrument over another.

For potential outcomes conditional on X_i, we assume the following semiparametric form:

$$E[Y_{id} \,|X_i, Z_i, U_i] = \alpha_d(X_i) + \gamma_d(X_i)\, J(U_i) + \xi_{g(j),t}, \tag{9}$$

for some strictly increasing continuous function $J(\cdot) : [0,1] \to \mathbb{R}$, such that $E[J(U_i)] = 0$, and where $\xi_{g(j),t}$ is a fixed effect for physician group $g(j)$ in time period t. Taking iterated expectations, we can derive the following expectation of outcomes conditional on treatment status from Equation (9):

$$E[Y_{id} \,|X_i, Z_i, D_i = d] = \alpha_d(X_i) + \gamma_d(X_i)\, \lambda_d(P_{ij}) + \xi_{g(j),t}, \tag{10}$$

where $P_{ij} \equiv P(X_i, Z_i, g(j), t)$, and $\lambda_d(P_{ij}) \equiv E[J(U_i) \,|P_{ij}, D_i = d]$.

This is the standard model in the marginal treatment effects literature, e.g. Kline

and Walters (2016), with one importance difference: The functions $\gamma_d(X_i)$ are permitted to vary with patient observable characteristics rather than being constants. This allows estimated heterogeneity in observational treatment effects to potentially be explained by heterogeneous selection bias.

3.2 Treatment Effects

The marginal treatment effect, or MTE, for patients with $X_i = x$ and $U_i = u$ is then

$$\begin{aligned} \text{MTE}(x,u) &= E\left[Y_{i1} - Y_{i0} \,|X_i = x, U_i = u\right] \\ &= \Delta\alpha(x) + \Delta\gamma(x)J(u). \end{aligned} \tag{11}$$

The term $\Delta\alpha(x) \equiv \alpha_1(x) - \alpha_0(x)$ represents average treatment effects conditional on characteristics $X_i = x$, or ATE(x). The second term, $\Delta\gamma(x)J(u) \equiv (\gamma_1(x) - \gamma_0(x))J(u)$, represents treatment effect heterogeneity related to selection. Since $J(\cdot)$ is an increasing function, $\Delta\gamma(x) < 0$ would imply decreasing treatment effects with the latent index U_i (i.e., positive selection on gains), while $\Delta\gamma(x) > 0$ would imply the opposite. Given that $E\left[J(U_i)\right] = 0$, ATE(x) = E[MTE(x, U_i)].

The observational treatment effects, or the treatment effect estimated by OLS under the unconfoundedness assumption in Equation (3), for patients with $X_i = x$ is:

$$\begin{aligned} \text{OLS}(x) &= E\left[Y_{i1} \,|X_i = x, D_i = 1\right] - E\left[Y_{i0} \,|X_i = x, D_i = 0\right] \\ &= \Delta\alpha(x) + \gamma_1(x)E_{Z|x}\left[\lambda_1(P_{ij})\right] - \gamma_0(x)E_{Z|x}\left[\lambda_1(P_{ij})\right] \\ &= \Delta\alpha(x) + \Delta\gamma(x)E_{Z|x}\left[\lambda_1(P_{ij})\right] + \gamma_0(x)E_{Z|x}\left[\lambda_1(P_{ij}) - \lambda_0(P_{ij})\right], \end{aligned} \tag{12}$$

where $E_{Z|x}$ denotes the expectation given variation in Z_i, conditional on $X_i = x$. The first term in Equation (12) is the average treatment effect, or ATE(x). The second term represents selection on gains, or the difference between treatment on the treated and ATE(x). The third term represents selection on levels implied by $\gamma_0(x)J(U_i)$. The OLS assumption in Equation (3) implies that $\gamma_0(x) = \gamma_1(x) = 0$ for all x, so that OLS(x) = ATE(x).

Note that setting γ_0 and γ_1 equal to constants as is standard in the literature would imply that all variation in OLS coefficients with x conditional on treatment propensities reflected heterogeneity in ATE(x), which would imply large benefits of reallocating treatments based on OLS coefficients. We will allow for the possibility that variation in OLS(x) reflects differential selection, in which case it would not imply large benefits of

reallocating treatments.

3.3 Policy-Relevant Estimation

Equation (10) fully specifies potential outcomes conditional on any set of arbitrary patient characteristics X_i. In principle, with infinite data and sufficient variation in P_{ij} within each set of observable characteristics X_i, one could directly estimate Equation (10) separately for $D_i = 1$ and $D_i = 0$ to obtain $(\alpha_0(x), \alpha_1(x), \gamma_0(x), \gamma_1(x))$ for any x. In practice, however, as the set of characteristics becomes increasingly rich, the number of observations i for which $X_i = x$ becomes increasingly limited. Additionally, feasible policies based on patient characteristics will generally rely on a simplified space of patient characteristics. In ongoing work, we are estimating a "moment forest" machine learning model to determine optimal treatment guidelines using all available variables. For now, we consider the optimal rule conditioning on a smaller set of variables.

Specifically, we reduce the dimensionality of X_i by focusing on a coarser projected space, $q(X_i)$. As in the recent value-added literature (e.g., Chetty et al. 2014; Angrist et al. 2017), a particularly well-motivated and readily available projection of X_i is the observational treatment effect of X_i, estimated in Section 2.2 by OLS, $q(X_i) = \text{OLS}(X_i)$. We consider a few other simple rules based on existing guidelines as well. As we will discuss below, the key parameter for policies that recommend either treatment or not—i.e., $D(q(X_i)) \in \{0, 1\}$—is the average treatment effect conditional on $q(X_i)$, or $\Delta\alpha(q) \equiv E_X[\Delta\alpha(X_i)|q(X_i) = q]$, with some abuse of notation.

To ease estimation, we make the additional simplifying assumption that $\Delta\gamma(x) = \Delta\gamma$ for all x, or that selection on gains is constant across patient characteristics. This assumption implies that the difference in marginal treatment effects between patients last to be treated and those first to be treated, or $\text{MTE}(x, 1) - \text{MTE}(x, 0) = \Delta\gamma(J(1) - J(0))$, is the same regardless of x.

As discussed in Heckman and Vytlacil (2007), we can estimate relevant structural parameters using two approaches: a control-function approach, as in Equation (10), or an equivalent instrumental-variable (IV) approach. We will describe the latter approach here, as this focuses attention on the relevant parameter of $\Delta\alpha(q)$, although a version of the control-function approach would also retrieve the same parameter. We start by restating Equation (10) as an instrumental-variables reduced form relationship:

$$
\begin{aligned}
E[Y_{id}|X_i, Z_i] &= \alpha_0(X_i) + \Delta\alpha(X_i)P_{ij} + \\
&\quad \gamma_0(X_i)\lambda_0(P_{ij})(1 - P_{ij}) + \gamma_1(X_i)\lambda_1(P_{ij})P_{ij} + \xi_{g(j),t} \\
&= \alpha_0(X_i) + \Delta\alpha(X_i)P_{ij} + \Delta\gamma\lambda_1(P_{ij})P_{ij} + \xi_{g(j),t}, \qquad (13)
\end{aligned}
$$

where the second equality makes use of the fact that $\lambda_0(P_{ij})(1-P_{ij})+\lambda_1(P_{ij})P_{ij} = 0$ for any pair of control functions $\lambda_0(\cdot)$ and $\lambda_1(\cdot)$, and substitutes $\Delta\gamma$ for $\Delta\gamma(X_i)$. Following Olsen (1980), we assume a linear function for $J(\cdot)$, or $J(U_i) = U_i$, which implies that $\lambda_1(P_{ij}) = E[J(U_i)|U_i < P_{ij}] - E[J(U_i)] = (P_{ij} - 1)/2$. Identification of the average treatment effect, $\Delta\alpha(X_i)$, is that $\partial E[Y_{id}|X_i = x, Z_i]/\partial P_{ij}$ approaches $\Delta\alpha(x)$ for any x as P_{ij} approaches 1, because $\lambda_1(P_{ij})$ approaches 0 as P_{ij} approaches 1, regardless of the functional form of $J(\cdot)$.

For estimation purposes, we focus on relationships conditional on $q(X_i) = q$ and P_{ij}:

$$
E[Y_{id}|q(X_i) = q, Z_i] = \alpha_0(X_i) + \Delta\alpha(q)P_{ij} + \Delta\gamma\lambda_1(P_{ij})P_{ij} + \xi_{g(j),t}. \qquad (14)
$$

If the exclusion restriction in Equation (7) holds, then we could also estimate Equation (14) with $E_X[\alpha_0(X_i)|q(X_i) = q]$ in place of $\alpha_0(X_i)$, although in our baseline specification, we use $\alpha_0(X_i)$ to increase precision. Including the full set of controls in $\alpha_0(X_i)$ also yields consistent estimation if Equation (7) only holds conditional on X_i.

3.4 Our Quasi-Experiment

In the ideal experiment, patients with a given set of observable characteristics would be sorted according to how a representative physician would treat them, and random fractions of these patients, starting from the patient in first sorted position to the last, would receive treatment. Since this experiment is not feasible, we employ a quasi-experimental analogue of this in a judges-design framework, in order to estimate heterogeneous marginal treatment effects laid out in Sections 3.1 to 3.3. In our framework, we consider patients as plausibly quasi-experimentally assigned to physicians with different transfusion propensities within clinical groups. Physicians with higher propensities choose a greater fraction of patients for transfusion, and through their transfusion decisions affect mortality.

In the framework in Sections 3.1 to 3.3, identification of average treatment effects and selection on gains follows familiar arguments laid out in the active marginal treatment effects literature. We can identify average treatment effects by comparing outcomes

for patients assigned to doctor A, who treats 0% of the time, to patients assigned to doctor B, who treats 100% of the time. We can then ask whether outcomes for doctor C, who treats 40% of the time, are better than we would expect given average treatment effects. If so, this suggests that doctors are selecting on gains and allocating the patients who benefit most to treatment. In practice, we do not observe doctors treating 0% and 100% of the time, but we can semiparametric assumptions of the form in (9) to extrapolate from mortality relationships we observe when comparing doctors who transfuse 30% vs. 40% of patients, and doctors who transfuse 60% vs. 70% of patients. For this strategy, we construct a continuous, single-dimensional instrument to capture a physician's empirical propensity to transfuse. Our approach is similar to other leave-one-out "jackknife" instruments (e.g., Aizer and Doyle 2013; Dobbie, Goldin, and Yang 2018), except that we increase its power by accounting for the overall number of patients that we observe with a physician. The instrument propensity for physicians with fewer patients is shrunken towards a mean, using an empirical Bayes procedure described in Appendix A (Chetty, Friedman, and Rockoff 2014).

The independence and exclusion assumption in Equation (7) corresponds to the idea that, conditional on time and location cells, physicians are as good as randomly assigned patients, and that physicians who transfuse more do not do anything else differently which impacts outcomes. The institutional setting of the VHA supports this conditional random assignment assumption: Patients are assigned to physicians according to inpatient rotations and it is rare that care is transferred across doctors by patient request or due to differential expertise within an inpatient service. The assignment of a given patient to a physician thus depends principally on which doctors have openings when that patient arrives and this should be unrelated to patient attributes. We also investigate the conditional random assignment assumption empirically through a series of balance tests.

For these balance tests, we can exclude a subset of observable characteristics from our conditioning set, use it to construct predicted mortality, and then ask whether physicians who are more likely to transfuse also have patients who look systematically different based on observable characteristics. To implement this test, we continue to condition on variables that define relatively uniform sets of patients over which there is plausible quasi-random assignment of patients to physicians. Specifically, we control for a patient's bed section by hemoglobin bin fixed effects, for five categories of hemoglobin levels capturing different patient acuity levels. We further control for month-year fixed effects and day of week fixed effects, and basic patient demographic variables (age, gender, race).

These regressions are reported in column 1 of Table 4. In row 1, we regression predicted mortality given comorbid conditions and lab values on the controls noted above as well as transfusions instrumented with the empirical Bayes jackknife treatment propensity (this is the reported coefficient). We find no evidence that healthier patients (with low predicted mortality based on their comorbid conditions and lab values) are assigned to doctors with high transfusion propensities. The point estimate suggests any selection countervails our estimated treatment effects by assigning sicker patients to doctors with high transfusion propensities, but it is both small in magnitude and not statistically significant.

We also find no evidence that there are differences in treatment assignment to patients with different observational treatment effects. In Figure 2, we plot the difference in predicted mortality for high treatment propensity doctors and low treatment propensity doctors across a range of predicted observational treatment effects. We find no evidence that the difference in predicted mortality risk between high transfusion and low transfusion doctors varies with the predicted observational treatment effects. This same finding is corroborated by column 1 of Table 4, in the panel with two endogenous variables. This specification includes both transfusions and transfusions interacted with observational treatment effects as regressors, instrumented with both the jackknife treatment propensity and the jackknife treatment propensity interacted with observational treatment effects (this regression also controls flexibly for splines of the observational treatment effects). In each case, we find no significant effect on predicted mortality, consistent with our assumption of conditional random assignment.

The exclusion assumption for our instrumental variable also requires that physicians who transfuse more do nothing else differently that impacts patient outcomes. While transfusion decisions are a salient dimension of physician decision-making for patients with low hemoglobin levels, there are a number of other treatment decisions made during the same hospital stay that could also affect patient care outcomes. We measure other dimensions of treatment directly and test whether accounting for them changes our central findings. These results are reported alongside our main regression results.

Finally, we explore the validity of our monotonicity assumption in Equation (8) by testing the sign of the "first stage" using the jackknife instrumental variable to predict transfusion propensity separately within key patient subgroups (Bhuller et al. 2016; Dobbie et al. 2018). These results are reported in Table 3. We find a strong first stage relationship between our jack-knife instrument of the doctor's transfusion propensity and whether or not the index patient was transfused. The finding is of similarly large magnitude and highly statistically significant ($p < .01$) in each tested

subsample, including patients with high predicted mortality, low predicted mortality, high hemoglobin, and low hemoglobin.

4 Treatment Effect Estimates and Simulation Results

4.1 Heterogeneous Treatment Effects

We first estimate a simplified version of equation 14, which excludes the nonlinear $\lambda_1(P_{ij})P_{ij}$ term from the estimation. This equation will identify local average treatment effects and, assuming selection on levels but not gains, it will identify average treatment effects; in other words, we will recover ATEs in this baseline specification if treated patients are sicker or healthier than untreated patients but do not have systematically different treatment effects from untreated patients. Next, we will enrich the model by allowing for selection on gains and estimating the complete version of equation 14.

We begin with a basic instrumental variable estimate of local average treatment effects, using the empirical bayes jackknifed treatment propensity as an instrumental variable for transfusion. This specification is reported in the first row of Table 4. All regressions reported in this table controls for hospital section by hemoglobin bin and timing fixed effects, as well as a basic set patient characteristics that may influence the sorting of patients to physicians. In Panel A column 2, we find that transfusions reduce mortality by an estimated 1.6 percentage points ($p < 0.05$). In column 3, we control for a richer set of patient covariates beyond those included in the basic balance regression and find similar results. The estimated benefit of transfusion actually becomes slightly larger, implying a 2 percentage point mortality reduction, once we account for these additional control variables.

To operationalize equation 14 (with $\Delta\gamma = 0$), we assume that $\Delta\alpha(q) = \delta_0 + \delta_1 \cdot OLS(X_i)$. We then estimate δ_0 and δ_1 by including terms for T_{ij} and $OLS(X_i) \cdot T_{ij}$ instrumented respectively by the empircial Bayes jackknife treatment propensity Z_{ij} and $OLS(X_i) \cdot Z_{ij}$. We also flexibly control for the direct effect of $OLS(X_i)$ using 30-knot splines. These results are shown in Panel B. For this initial table, the observational treatment effects heterogeneity is estimated using a limited set of patient characteristics including hemoglobin, vital signs, utilization history and demographics. In later results, we expand the set of variables used to predict treatment effect heterogeneity.

In column 2 with limited controls for patient characteristics, we estimate a coefficient of 0.972 on the interaction between transfusion and the observational treatment effects prediction ($p < 0.01$). In our preferred column 3 specification with additional

patient controls, this attenuates slightly to 0.901 ($p < 0.05$). This finding suggests that a 1 percentage point increase in observational treatment effects correlates with a 0.9 percentage point increase in average treatment effects. The linear transfusion dummy variable has a coefficient of -0.005, which is not statistically distinguishable from 0.

Our posterior estimates of average treatment effects are a linear transformation of the observational treatment effects where 0.9 is the scale parameter and -0.005 provides a location parameter. For 85% patients in our sample, this posterior estimate suggests that transfusions reduce mortality, and the average treatment effect across all patients in our sample would be 1.6 percentage points. In this specific case, the IV posteriors are very close to the OLS estimates, although that was not ex ante obvious. Below we discuss alternative specifications that do not share this feature.

To see graphical evidence of this relationship, we consider the reduced form version of our IV equation. Specifically, we estimate equation 13 but now include only control variables in the regression model. We exclude the instrument and the interaction between the instrument and the observational treatment effects prediction from the regression. We then break the sample into two groups based on whether the doctor's transfusion propensity is above or below the mean transfusion propensity. For each decile of the observational treatment effects distribution, we calculate the average residual mortality for both the high IV and low IV groups. The graph plots the difference between these mortality residuals (y-axis) against the observational treatment effects (x-axis).

We expect that patients with large OLS-predicted benefit of transfusion will have lower mortality when they are assigned to high transfusion propensity doctors compared to similar patients who are treated by low transfusion propensity doctors. This gap in mortality rates should close as the OLS-predicted benefit of transfusion approaches 0. This pattern is apparent in Figure 2, where we find a strong upward sloping relationship between the OLS predicted treatment effect and the difference in residual mortality for the high IV and low IV groups.

In the next set of results, reported in Table 4 column 4, we augment our baseline specification to account for possible correlations between transfusion propensity and the propensity to provide other types of medical treatments. Specifically, we construct jackknife counts of the number of physician orders per hospital stay for pharmacy medications, imaging, nursing, and diet, and use them as instrumental variables for the number of each type of orders that the patient receives.

Accounting for these additional treatments leads to some attenuation of our IV estimates. In Panel A, the overall effect of transfusion attenuates slightly to a to 1.7 percentage point decline in mortality, from a 2 percentage point decline in the model

that did not account for these additional treatments. In the Panel B regression that interacts the OLS predicted treatment effect with transfusion, the coefficient on that interaction attenuates from 0.901 to 0.753, but remains statistically significant ($p < .05$) and clinically important.

Finally, in Table 4 column 5, we augment our regression specification with $\lambda_1(P_{ij})P_{ij}$, a quadratic term in the predicted treatment propensity (i.e. fitted values from the first stage regression). This specification allows us to investigate whether doctors use unobservable patient characteristics to select on gains. Note that when the first stage estimation of the predicted treatment propensity (as a function of the instrumental variable and other controls) is estimated with error, this regression will suffer from "forbidden regression" bias.

In the Panel A specification with transfusion as the only endogenous variable, the coefficient on transfusions will now estimate the average treatment effect across all patients (i.e. when the treatment propensity equals 1 so that $\lambda_1(P_{ij})P_{ij} = 0$). This estimate of the treatment effect is larger, at 3.6 percentage points. The coefficient on the quadratic term implies that there is reverse roy selection–in other words, physicians are prioritizing transfusions for the patients who benefit least from treatment. The point estimate would imply that the marginal treatment effect for a doctor who transfuses 25% of her patients has a marginal treatment effect of 1.9 percentage points while a doctor who treats 75% of her patients has a marginal treatment effect of 5.3 percentage points. We find similar results on discretion in Panel B, when including the interaction of transfusion with the observational treatment effects.

While these results are provocative—suggesting large failures on the part of physicians to select the most appropriate patients for treatment–we should be cautious in our interpretation. As noted in 3, estimating marginal treatment effects requires strict monotonicity conditions which may not hold in this setting if doctors disagree about the ranking of patients' suitability for transfusion. In addition, the estimates may be biased due to the "forbidden" regression problem described above. Incorporating these findings into our simulations will only bolster the case for restricting physician discretion.

We now turn to Table 5. In this table, we replace the OLS predicted treatment effects used previously in 4 with a set of predicted treatment effects estimated to allow heterogeneity along a richer array of patient covariates. In this table, the OLS treatment effect is predicted using the full set of observed patient covariates, adding 31 comorbidities and results from five lab tests, to the baseline variables included previously (hemoglobin, demographics, vital signs, and past utilization). Allowing these additional sources of treatment effect heterogeneity substantially increases the standard

deviation of predicted observational treatment effects across individuals from 0.017 with the limited variable set to 0.027 with the full variable set (as reported in Table 1).

In column 1 of Table 5, we estimate that each 1 pp change in this new OLS variable predicts a 0.475 pp change in treatment effects. A major factor behind the observed attenuation in the coefficient relative to the estimates reported in Table 4 may be that the full observational treatment effects, which is now predicted by over 100 covariates, is more noisily estimated due to over-fitting.

In Table 5 column 2, we estimate a diminished, but still significant, pattern of the reverse roy selection documented previously. Now that we have accounted for additional sources of heterogeneity in average treatment effects, the scope for selection on gains may be more limited.

4.2 Treatment on the Treated Under Current Guidelines

Next, we investigate how our estimated treatment effects correlate to popular transfusion guidelines. For this exercise, we use the IV estimated treatment effects from Table 4 column 1; this makes use of the full heterogeneity in predicted treatment effects using all the possible covariates.

Figure 3 breaks up the patients in our sample into 20 ventiles. The blue triangle series groups patients according to their hemoglobin levels, with the lowest hemoglobin values in the lowest numbered ventile. Within each ventile of hemoglobin level, we calculate the average IV estimated treatment effect, using estimates from Table 5 column 1. The red dot series groups patients according to their estimated treatment effect, and calculates the average IV estimate treatment effect in each group.

The comparison of these two series allows us to visualize how a transfusion guideline based solely on hemoglobin might perform relative to a transfusion guideline that incorporated more dimensions of treatment effect heterogeneity. More negative values along the y-axis correspond to larger benefits of transfusion. The figure illustrates assigning transfusions based on Hemoglobin captures only a fraction of possible benefits: treatment effects range from -1.5 pp to 0 as Hemoglobin varies, but they range from -4 pp to +2 pp as we vary observational treatment effects estimated using all available variables. We explicitly simulate the benefits to reassigning transfusions based on the more complete index in the next section.

4.3 Counterfactual Transfusion Policies

In this section, we apply our estimates of heterogeneous treatment effects from the selection model to consider a number of counterfactuals. Specifically, we want to compare mortality rates that would be associated with status quo treatment decisions and contrast them with outcomes under strict adherence to existing guidelines and with treatment decisions that strictly follow the optimal guideline based on our estimated treatment effects.

We apply cross validation techniques to avoid overfitting and thus avoid overstating the benefits of adherence to the optimal strict guidelines. Specifically we identify the treatment rule given results estimated on half of our sample, which has been randomly assigned as the "training" data set. We then evaluate the benefits of guideline adherence by predicting stroke and bleed outcomes using treatment effects estimating on the other "test" half of our sample. We perform 50 bootstrap repetitions of this cross-validation procedure for the estimates described below.

When considering strict adherence to existing guidelines, we consider the counterfactual whereby physicians transfuse all patients within a hospital section, starting with the lowest hemoglobin level and stopping when the expected transfusion rate equals the current observed transfusion rate. This reflects the fact that while current guidelines vary in the precise hemoglobin threshold recommended for transfusion, they focus on hemoglobin as the single key determinant of transfusion recommendations.

To construct the optimal strict guideline, we consider minimizing 30-day mortality rate subject to a constraint that holds the total number of transfusions constant within each hospital section at the rate currently observed in our sample. By restricting reallocation to within a hospital section, we illustrate the benefits to transfusion reallocation that does not require changing the geographic distribution of blood bank supply.

For each of these simulations, we compare actual mortality rates to the mortality rates predicted from applying our treatment effects estimated in the training data to predict outcomes in the test data.

Results of these simulations are reported in Table 6. In column 1, we naively use the observational treatment effects to evaluate the benefits of reassigning transfusions to patients with the largest predicted benefits, on the basis of their observational treatment effect ordering. We estimate a 1.9 percentage point reduction in mortality with this rule. Results in column 2 show that this estimate would overstate the benefit of reallocation by not adjusting the estimate to account for the fact that only 48% of the variation in observational treatment effects is predictive of true treatment effect heterogeneity. When using our IV model to simulate the benefits of reassignment, we find that reassignment

of transfusions could reduce mortality rates by 1.1 percentage points, from a base of 9.4%.

In Table 6 column 3, we estimate the benefits of reassigning transfusions to patients with the lowest hemoglobin levels within each hospital section. This assignment rule would outperform status quo physician decisions, reducing mortality by 0.3 percentage points. However, the gains are much smaller than the proposed alternative guideline that incorporates more clinical detail.

These estimates suggest that stricter adherence to existing guidelines could modestly reduce patient mortality. However, there are potentially large gains to guideline improvements which use information beyond hemoglobin levels to better tailor treatment decisions.

5 Conclusion

This paper develops a new methodology for estimating heterogeneous returns to treatment within a model that accounts for selection on unobservable characteristics. This methodology can be applied for development of new guidelines that perform substantially better than current guidelines or current observed treatment decisions in counterfactual simulations.

Current efforts at applying machine learning to medical applications frequently fail to account for selection into treatment on the basis of unobserved factors, limiting the plausibility of the resulting estimates. We estimate our instrumental variables model to understand the tradeoffs between allowing physician discretion and requiring strict guideline adherence. We observe that strict adherence to existing guidelines would lead to modest improvements in mortality, but more nuanced guidelines could reduce mortality in the targeted population by 1.1 percentage point or approximately 12%.

Our selection model is identified by a jackknife instrumental variable approach which relies on monotonicity and exclusion assumptions for identification. Our estimation also makes use of functional form restrictions for tractability of estimation, and these could also contribute to identification. While in principle, nonparametric identification is possible, even in our large sample of patients we do not have sufficient power for a completely nonparametric approach.

Our findings suggest that current approaches to medical decision-making lead to significant misallocation of common treatments. Applying quasi-experimental methods to analyze medical records databases provides new opportunities to estimate heterogeneous returns to treatment and develop guidelines that effectively tailor treatment plans. We

demonstrate that physicians' current discretionary decisions underperform relative to our proposed assignment rule, as do current medical guidelines.

References

Abaluck, J., L. Agha, C. Kabrhel, A. Raja, and A. Venkatesh (2016, December). The determinants of productivity in medical testing: Intensity and allocation of care. *American Economic Review 106*(12), 3730–64.

Aizer, A. and J. J. Doyle (2013). Juvenile incarceration, human capital and future crime: Evidence from randomly-assigned judges. *The Quarterly Journal of Economics 130*(2), 759–803.

Angrist, J. D., P. D. Hull, P. A. Pathak, and C. R. Walters (2017). Leveraging lotteries for school value-added: Testing and estimation. *The Quarterly Journal of Economics 132*(2), 871–919.

Asher, S., D. Nekipelov, P. Novosad, and S. P. Ryan (2016). Classification trees for heterogeneous moment-based models. Technical report, National Bureau of Economic Research.

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Belloni, A., V. Chernozhukov, and C. Hansen (2014, May). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives 28*(2), 29–50.

Bhuller, M., G. B. Dahl, K. V. Loken, and M. Mogstad (2016, September). Incarceration, Recidivism and Employment. Technical Report w22648, National Bureau of Economic Research, Cambridge, MA.

Carson, J. L., D. J. Triulzi, and P. M. Ness (2017). Indications for and adverse effects of red-cell transfusion. *New England Journal of Medicine 377*(13), 1261–1272.

Chandra, A., A. Finkelstein, A. Sacarny, and C. Syverson (2016, August). Health care exceptionalism? performance and allocation in the us health care sector. *American Economic Review 106*(8), 2110–44.

Chandra, A. and D. O. Staiger (2007). Productivity spillovers in health care: evidence from the treatment of heart attacks. *Journal of Political Economy 115*(1), 103–140.

Chandra, A. and D. O. Staiger (2017, November). Identifying sources of inefficiency in health care. Working Paper 24035, National Bureau of Economic Research.

Chetty, R., J. N. Friedman, and J. E. Rockoff (2014, September). Measuring the impacts of teachers i: Evaluating bias in teacher value-added estimates. *American Economic Review 104*(9), 2593–2632.

Collins, F. S. and H. Varmus (2015). A new initiative on precision medicine. *New England Journal of Medicine 372*(9), 793–795.

Currie, J., W. B. MacLeod, and J. Van Parys (2016). Provider practice style and patient health outcomes: the case of heart attacks. *Journal of health economics 47*, 64–80.

Dobbie, W., J. Goldin, and C. S. Yang (2018, February). The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review 108*(2), 201–240.

Heckman, J. J. and E. Vytlacil (2005). Structural equations, treatment effects, and econometric policy evaluation1. *Econometrica 73*(3), 669–738.

Heckman, J. J. and E. J. Vytlacil (2007). Econometric evaluation of social programs, part i: Causal models, structural models and econometric policy evaluation. *Handbook of econometrics 6*, 4779–4874.

Hull, P. (2018). Estimating hospital quality with quasi-experimental data.

Imbens, G. W. and J. D. Angrist (1994, March). Identification and Estimation of Local Average Treatment Effects. *Econometrica 62*(2), 467–475.

Kline, P. and C. R. Walters (2016). Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics 131*(4), 1795–1848.

Kling, J. R. (2006). Incarceration length, employment, and earnings. *The American Economic Review 96*(3), pp. 863–876.

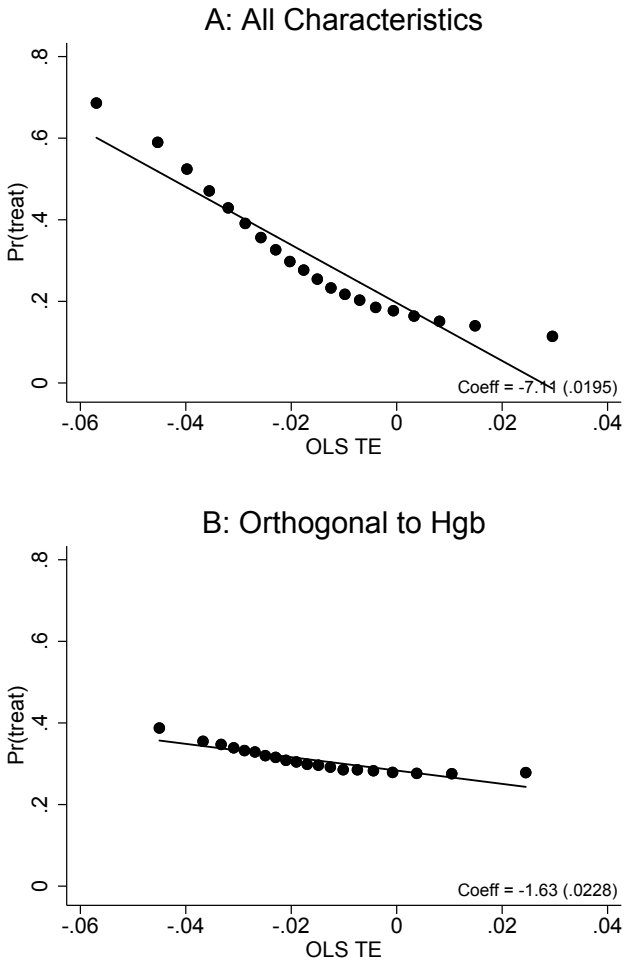
Olsen, R. J. (1980). A least squares correction for selectivity bias. *Econometrica: Journal of the Econometric Society*, 1815–1820.

Vytlacil, E. (2002). Independence, Monotonicity, and Latent Index Models: An Equivalence Result. *Econometrica 70*(1), 331–341.

Wager, S. and S. Athey (2017). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 0*(ja), 0–0.

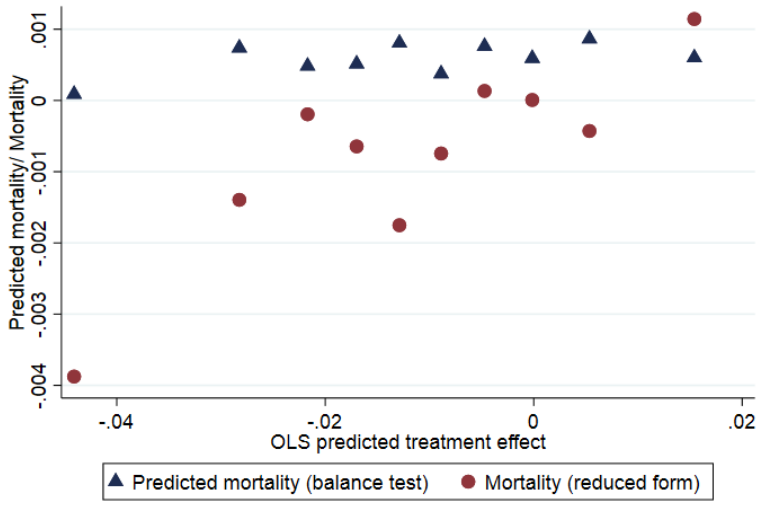
6 Tables and Figures

Figure 1: Treatment Probability Given Observational Treatment Effects



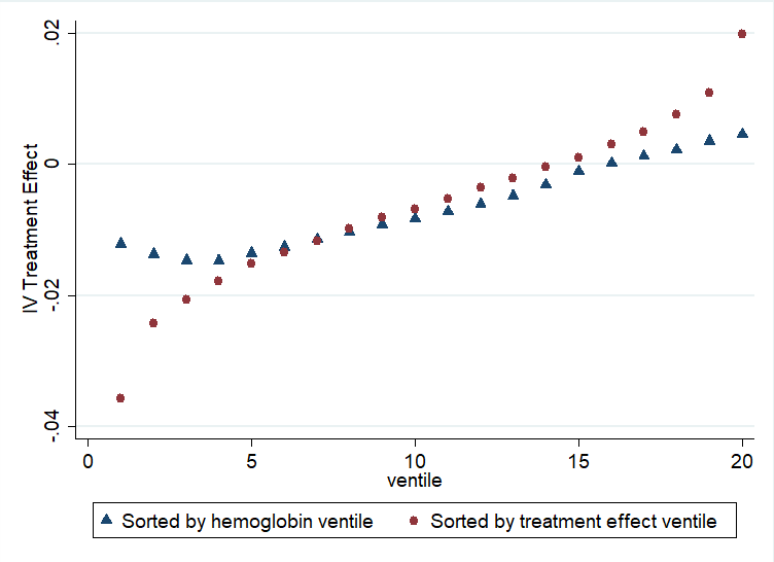
Note: This figure shows the probability of treatment as a function of observational treatment effects. OLS treatment effects are estimated as a function of patient characteristics, $\mathbf{X}_i$, given by $\beta_0 + \mathbf{X}_i'\beta_1$ in Equation (4). Panel A is a binned scatterplot in which each dot represents 5% of the data ordered by observational treatment effects, residualized by time and clinical location dummies. Each dot shows the average residualized observational treatment effects for the bin on the x-axis and the average residualized probability of transfusion on the y-axis. For interpretation, the mean treatment effect for all the data is added back to the residuals on the x-axis, and the overall probability of transfusion is added back to the residuals on the y-axis. Panel B is a similar binned scatterplot, in which each dot is residualized not only by time and clinical location dummies, but also by cubic splines of minimum hemoglobin.

Figure 2: Balance Test and Reduced-Form Effect of Transfusion



Notes: This graph reports non-parametric version of the balance test and main IV results. First, we calculate residuals of mortality and predicted mortality variables, controlling for patient characteristics, time, and location controls. Then, in each decile of the observational treatment effects, we calculate the difference between average residual mortality (or predicted mortality) for patients with above average values of the treatment propensity instrumental variable compared to patients with below average values of the treatment propensity instrumental variable.

Figure 3: Comparison of Transfusion Assignment Policies



Notes: This graph breaks up the patients in our sample into 20 ventiles. The blue triangle series groups patients according to their hemoglobin levels, with the lowest hemoglobin values in the lowest numbered ventile. Within each ventile of hemoglobin level, we calculate the average IV estimated treatment effect, using estimates from Table 5 column 1. The red dot series groups patients according to their estimated treatment effect, and calculates the average IV estimate treatment effect in each group.

Table 1: Summary statistics

	Doctor transfusion propensity is below mean		Doctor transfusion propensity is above mean	
	Mean	Std. Dev.	Mean	Std. Dev.
	(1)	(2)	(3)	(4)
Doctor transfusion propensity	-0.044	0.037	0.052	0.051
Transfusion	0.181	0.385	0.248	0.432
Age	67.861	11.910	67.964	11.838
Minimum hemoglobin value	9.146	1.213	9.166	1.193
Mortality	0.093	0.290	0.097	0.290
Predicted mortality	0.093	0.100	0.097	0.102
Limited variable observational treatment effects	-0.012	0.017	-0.012	0.017
Full variable observational treatment effects	-0.012	0.027	-0.012	0.027
Sample size	841,676		729,623	

Note: This table reports mean and standard deviation of key variables used in our analysis. The sample is split into two groups, one for patients whose physician has below mean transfusion rate, and the other for patients whose physician has above mean transfusion rate. Note that this does not correspond to a balance test, because we have not controlled for timing and hospital section.

Table 2: Observational Treatment Effects and Treatment Probabilities by Patient Characteristics

	Observational Treatment Effects (p.p. Mortality)			Pr (Treat)	
	Overall	1st Quintile	5th Quintile	1st Quintile	5th Quintile
Panel A: Unresidualized					
Base	−1.590	−4.441	1.395	56.77	14.24
All	−2.656	−8.602	4.370	46.69	26.57
Demographics	−2.059	−3.813	−0.295	59.70	9.06
Comorbidities	−1.539	−4.368	1.430	58.22	14.74
Prior utilization	−1.748	−4.022	0.490	58.04	10.37
Laboratory test values	−3.045	−8.650	4.206	45.91	28.83
Vital signs	−2.210	−4.507	−0.115	55.97	9.28
Panel B: Residualized by Hemoglobin					
Base	−1.590	−3.649	0.949	35.73	27.73
All	−2.656	−8.152	4.327	33.96	29.77
Demographics	−2.059	−3.085	−1.130	32.55	26.97
Comorbidities	−1.539	−3.434	0.988	35.32	28.22
Prior utilization	−1.748	−3.176	−0.215	33.98	28.63
Laboratory test values	−3.045	−8.198	4.199	31.36	31.06
Vital signs	−2.210	−3.766	−0.974	32.33	29.17
Panel C: Residualized by Other Characteristics					
Demographics	−2.059	−2.911	−1.260	31.70	29.03
Comorbidities	−2.656	−4.023	−1.278	33.71	28.44
Prior utilization	−1.748	−2.820	−0.56	31.81	29.86
Laboratory test values	−3.045	−8.382	3.922	30.92	31.29
Vital signs	−2.210	−3.744	-0.977	32.39	29.24

Note: This table shows OLS average treatment effects and treatment probabilities when using various patient observed characteristics. Average treatment effects are calculate for all patients, for patients with treatment effects in the first quintile (the most negative), and for patients with treatment effects in the fifth quintile (the least negative or most positive). Treatment probabilities are also given for patients with treatment effects in the first and fifth quintiles. Panel A presents treatment effects and probabilities that are "unresidualized" by any other patient characteristics, i.e., only residualized by clinic and time dummies, which corresponds to Panel A in Figure 1. Panel B presents treatment effects and probabilities that are also residualized by hemoglobin, corresponding to Panel B in Figure 1. Panel C presents treatment effects and probabilities that are residualized by hemoglobin and any other patient characteristics. Base characteristics include demographics, comorbidities, and prior utilization.

Table 3: First Stage and Monotonicity Results

	Dependent variable: Transfusion				
	Pooled sample	High mortality	Low mortality	Low hemoglobin	High hemoglobin
Model with two endogenous variables					
Doctor transfusion propensity	0.543	0.6817	0.7534	1.0927	0.6048
	(0.005)	(0.0079)	(0.0056)	(0.0129)	(0.0041)
Doctor transfusion propensity	0.77	0.8783	0.8436	1.0741	0.5924
$\times$ Observational treatment effect	(0.006)	(0.0092)	(0.0066)	(0.0137)	(0.0045)
Model with one endogenous variable					
Doctor transfusion propensity	0.615	0.8328	0.8209	1.1256	0.6065
	(0.005)	(0.0067)	(0.0054)	(0.0084)	(0.0041)
Sample size	1,571,299	785,649	785,650	640,338	930,961

Note: This table reports estimates of the first stage regression associated with our instrumental variable strategy, and tests for monotonicity of the instrument in various subgroups of the patient population. Each column reports a single coefficient from each of three separate regressions. In the bottom panel, the dependent variable is a dummy variable for whether the patient is transfused and the instrumental variable of interest is the doctor's transfusion propensity. In the top panel, the first result reports a similar regression with the transfusion dependent variable, but the regression also controls for interactions between the predicted observational treatment effects and the instrumental variable. The second result reports the coefficient on the instrumental variable interacted with the patient's observational treatment effects prediction, controlling for the main effect of the IV. All regressions control for a rich set of patient characteristics, a 40-knot spline in the observational treatment effects, timing fixed effects (day of week, and month-by-year), and a fixed effect for each hospital section and 1 point hemoglobin category. Patient characteristics include hemoglobin levels, lab values, comorbidities, vital signs, demographic factors and 1-year utilization history. Standard errors clustered at the hospital section level are reported in parentheses.

Table 4: Balance and IV results using limited variable set OLS predictions of treatment effects

	Balance test	Instrumental variables results			
		Dependent variable:			
	Predicted 30-day mortality	30-day mortality			
	(1)	(2)	(3)	(4)	(5)
A. Model with one endogenous variable					
Transfusion	0.003	-0.016**	-0.020**	-0.017**	-0.036***
	(0.003)	(0.008)	(0.008)	(0.008)	(0.005)
Selection on gains					-0.069***
					(0.015)
B. Model with two endogenous variables					
Transfusion	0.045	0.000	-0.005	-0.004	-0.027***
	(0.109)	(0.010)	(0.011)	(0.011)	(0.009)
Transfusion × Observational TE	0.004	0.972***	0.901**	0.753**	0.315
	(0.004)	(0.367)	(0.377)	(0.354)	(0.323)
Selection on gains					-0.062***
					(0.016)
Control variables included					
Hemoglobin, vital signs, utilization history, demographics	Yes	Yes	Yes	Yes	Yes
Comorbidities, lab values	No	No	Yes	Yes	Yes
IV for imaging, pharmacy, nursing, diet orders	No	No	No	Yes	No

Note: This table reports estimates of our balance test and instrumental variable regressions. The first column reports a balance test, where the outcome variable is predicted 30-day mortality, where the prediction is formed based on a linear regression that includes a rich vector of patient characteristics, including hemoglobin, vital signs, utilization history, demographics, comorbidities and lab values. The balance test then controls for hospital bed section, timing of admission, and a limited set of patient characteristics including demographics, hemoglobin levels, vital signs, and past utilization. The remaining columns report results from our instrumental variable regressions. The top panel reports a simple IV specification that uses physician's jackknife transfusion propensity as an instrumental variable for transfusions. The bottom panel includes a main effect of transfusion, as well as an additional endogenous variable for the interaction between receiving a transfusion and the observational treatment effects. In this table, we use a limited set of variables to construct the observational treatment effects estimates, including hemoglobin, vital signs, demographics, and past utilization. Column 5 further includes a selection on gains estimate, identified by a quadratic function of transfusion propensity, as described in the main text. In addition to the control variables described at the bottom of the table, all regressions control for hospital section by hemoglobin group, timing variables, and a flexible function of the predicted observational treatment effects. Standard errors (in parentheses) are clustered at the hospital section level.

Table 5: IV results using full variable set OLS predictions of treatment effects

	Dependent variable: 30-day mortality	
	(1)	(2)
Transfusion	-0.014	-0.022***
	(0.009)	(0.007)
Transfusion × Observational TE	0.475**	0.368***
	(0.229)	(0.179)
Selection on gains		-0.033**
		(0.015)
Control variables included		
Hemoglobin, vital signs, utilization history, demographics	Yes	Yes
Comorbidities, lab values	Yes	Yes
IV for imaging, pharmacy, nursing, diet orders	No	No

Note: This table reports estimates of additional instrumental variable regressions, similar to the specifications reported in Panel B of 4. The difference between these results and the Table 4 comes from the richer version of the OLS predicted treatment effect. In this table, the observational treatment effects is predicted using the full set of observed patient covariates, adding 31 comorbidities, and results from 5 additional lab tests, in addition to the baseline covariates included in Table 4 (hemoglobin, demographics, vital signs, and past utilization). This allows for more predicted heterogeneity in observational treatment effects. For more details on these regressions, including control variables and standard errors, see notes to Table 4.

Table 6: Simulated mortality benefits of re-assigning blood transfusions

	Reassign to largest OLS treatment effect first	Reassign to largest OLS treatment effect first	Reassign to lowest hemoglobin level first
	(1)	(2)	(3)
Using OLS to evaluate gains:	-0.0185***		
	(0.0002)		
Using IV to evaluate gains:		-0.0111***	-0.0029***
		(0.0007)	(0.0002)

Note: This table reports mortality reductions from simulations that reallocate blood transfusions, relative to current observed transfusion assignments. These estimates apply cross-validation, so that we estimate the benefits of treatment using a training sample, and then simulate the benefits to the reallocation in a disjoint testing sample. We repeat 50 bootstraps of the cross validation procedure. Column 1 uses the naive OLS estimates of treatment effects to evaluate the benefits of different assignment rules. Columns 2 and 3 use IV estimates of treatment effects to evaluate the benefits of reallocation. Standard errors from the bootstrapping procedure are reported in parentheses.

A Empirical Bayes Jackknife Instrument

NEED TO REORGANIZE AND FIX NOTATION

We start by constructing an empirical Bayes jackknifed propensity to treat. Relative to the raw jackknife propensity, the EB estimator increases power by reweighting based on the number of patients in each period for a given doctor and accounting for drift over time, similar to Chetty et al. (2014). This variable is constructed as a weighted average of physician residual treatment propensities in each period where the weights depend on the number of observations observed in each period. Our first step is to estimate the probability of treatment equation:

$$D_i = h^*(x_i) + \tau_{pt} + \theta_{g(p)t} + e_{ip} \tag{15}$$

and then compute $\tilde{D}_i = D_i - h^*(x_i) = \tau_{pt} + \theta_{g(p)t} + e_{ip}$ (we do not residualize by $\theta_{g(p)t}$ in order to properly account for imprecision in $\theta_{g(d)t}$ in our weighting procedure) [NEED TO DO!]. We split our entire sample length into periods of 90 days, and for each doctor in each 90 day period we compute:

$$\hat{\tau}_{pt} = \frac{1}{N_{pt}} \sum_{j \in p(t)} (\tilde{D}_{jp} - \tilde{D}_{g(p)t}) \tag{16}$$

where $\tilde{D}_{g(p)t} = \frac{1}{N_{g(p)t}} \sum_{j \in g,t} \tilde{D}_{jp}$. For the period in which patient i's treatment decision is made, we compute $\hat{\tau}_{(-i)pt} = \frac{1}{N_{pt}-1} \sum_{j \neq i \in p(t)} (\tilde{D}_{jp} - \tilde{D}_{(-i),g(p)t})$ where $\tilde{D}_{(-i),g(p)t} = \frac{1}{N_{g(p)t}-1} \sum_{j \in g,t} \tilde{D}_{jp}$. Given the $\hat{\tau}_{dt}$, we compute our empirical Bayes estimator as:

$$Z_i = \sum_s \psi_s(N_{p1}, ..., N_{pT})[(1 - \mathbb{1}(s = t))\hat{\tau}_{ps} + \mathbb{1}(s = t)\hat{\tau}_{(-i)pt}] \tag{17}$$

where the weights $\psi_s(N_1, ..., N_T)$ so that Z_{idt} is the MSE minimizing predictor of τ_{dt} given the number of observations in each period and given the assumption that $Cov(e_{id}, e_{jd})$ depends only on the distance in time between the treatment decisions for i and j. More details are given in Appendix QQQ (NEED TO UPDATE – CURRENTLY WRITTEN FOR WARFARIN). Note that Z_{idt} defined in this way can be written as:

$$Z_i = \tau_{pt} + \sum_{j \neq i \in p} w_j e_j \tag{18}$$

for some set of weights w_j.

Suppose $Y_{id} \sim_{i.i.d} Bernoulli(x_{id}\beta + \delta_c + \theta\mu_d \cdot x_{id} + p_{dt})$. Note that we can also rewrite

this assumption as $Y_{id} = x_{id}\beta + \delta_c + \theta\mu_d \cdot x_{id} + p_{dt} + e_{idt}$ where $E(e_{idt}|x_{dt}, \delta_c, p_{dt}) = 0$ and $e_{idt} = 1 - (x_{id}\beta + \delta_c + p_{dt})$ with probability $x_{id}\beta + \delta_c + p_{dt}$ and $-(x_{id}\beta + \delta_c + p_{dt})$ with probability $1-(x_{id}\beta+\delta_c+p_{dt})$. Note that μ_d are doctor-level observable characteristics. When we regress outcomes on Warfarin use, we include as instruments our estimates of p_{dt} and $\theta\mu_d \cdot x_{id}$ and we include as controls clinic fixed effects as well as all the x_{id}. We assume further that e_{idt} are uncorrelated across time or across different physicians (in other words, conditional on the stroke probability at a given doctor time, and conditional on covariates, the fact that a given patient happens to have a stroke at time s has no further bearing on whether some other patient has a stroke at time t).

Assume that $Cov_d(p_{d(t+s)}, p_{dt}) = \sigma_{ps}$ where the subscript d denotes the empirical covariance of these parameters across doctors. We can consistently estimate β in a fixed effects regression of Y_{id} on x_{id} and fixed effects for p_{dt}. I assume below that β is precisely estimated so we can ignore any noise in $\hat{\beta}$. I will however allow for the possibility of imprecisely estimated clinic fixed effects. Define: $\tilde{Y}_{idt} \equiv Y_{id} - x_{id}\beta - \theta\mu_d \cdot x_{id} = \delta_c + p_{dt} + e_{idt}$. I assume without loss of generality that the mean of p_{dt} within each clinic is 0.

If we estimate this regression by OLS, then our estimates of p_{dt} will be given by: $\hat{p}_{dt} = \frac{1}{N_{dt}} \sum_{i \in d(t)} (\tilde{Y}_{idt} - \bar{Y}_c)$ where $d(t)$ denotes the set of doctor d's patients at time t and $\bar{Y}_c = \frac{1}{N_c} \sum_{i \in c} \tilde{Y}_{idt}$. Note that we can write this as:

$$\begin{aligned} \hat{p}_{dt} &= \frac{1}{N_{dt}} \sum_{i \in d(t)} (\tilde{Y}_{idt} - \bar{Y}_c) \\ &= \frac{1}{N_{dt}} \sum_{i \in d(t)} (\delta_c + p_{dt} + e_{idt} - (\delta_c + \frac{1}{N_c} \sum_{i \in c} e_{idt})) \\ &= p_{dt} + \frac{1}{N_{dt}} \sum_{i \in d(t)} (e_{idt} - \bar{e}_c) \end{aligned} \tag{19}$$

where $\bar{e}_c = \frac{1}{N_c} \sum_{i \in c} e_{idt}$.

In this model, we want to construct the best linear unbiased estimator of p_{dt} given $\hat{p}_{d(-t)}$ where $\hat{p}_{d(-t)} = (\hat{p}_{d1}, ..., \hat{p}_{d(t-1)})$. We write this estimator as:

$$\hat{p}_{dt} = \sum_{s=1}^{t-1} \psi_s(N_1, ..., N_{t-1}) \hat{p}_{ds} \tag{20}$$

The weights ψ_s in each period are allowed to vary flexibly with the number of observations observed in all periods (you might give less weight to a prior period because you know that other periods are estimated with greater precision).

We will find ψ_s as the solution to the problem of minimizing:

$$\psi = \arg \min_{\{\psi_1,...,\psi_{(t-1)}\}} \sum_d \left(p_{dt} - \psi_0(N_{d1},...,N_{d(t-1)}) - \sum_{s=1}^{t-1} \psi_s(N_{d1},...,N_{d(t-1)})\hat{p}_{ds} \right)^2 \quad (21)$$

As in Chetty et al. (2014), the resulting coefficients ψ are equivalent to those obtained from an OLS regression of $\hat{p}_{dt}$ on $\hat{p}_{d(-t)}$ identified using across doctor variation. If we observed a large number of doctors for every possible set of $N_d = \{N_{d1},...,N_{d(t-1)}\}$, then we could run this regression to recover the relevant coefficients. In lieu of that, we can make use of the underlying structural assumptions in the model to compute the regression coefficients. These are given by $\psi = \Sigma_p^{-1}\gamma$ where $\gamma = (Cov_d(p_{dt}, \hat{p}_{d1}|N_d), ..., Cov_d(p_{dt}, \hat{p}_{d(t-1)}|N_d))'$ and Σ_p is the across-doctor covariance matrix of $\hat{p}_{d(-t)}$ (again conditional on N_d). The subscript d is used to indicate that this covariance is empirical covariance across doctors as opposed to the covariance of the random variables. This distinction will be important below for the variance terms (for example, the across doctor variance of p_{dt} is positive, whereas the variance of the "random variable" p_{dt} is 0 since p_{dt} is a constant).

Firstly, note that for $t \neq s$, $Cov_d(p_{dt}, \hat{p}_{ds}|N_{dt}, t) = E_d(p_{dt}\hat{p}_{ds}|N_{dt}, t) - E_d(p_{dt}|N_{dt}, t)E_d(\hat{p}_{ds}|N_{dt}, t) = \sigma_{p(t-s)}$. This quantity does not depend on N_{dt}, and it only depends on $t - s$, not t or s individually. We estimate this using the covariance of all $\tilde{Y}_{idt}$ and $\hat{p}_{ds}$ for all t and s that are the appropriate number of years apart and for which doctor d has at least 2 observations in year t. I denote this covariance by $Cov_i(\tilde{Y}_{idt}, \hat{p}_{ds})$. Let N_{t-s} denote the number of patients whose doctors have at least 2 observations in the current year and at least 1 observation $t - s$ years earlier. Then we have:

$$\begin{aligned}
Cov_i(\tilde{Y}_{idt}, \hat{p}_{ds}) &= E_i(\tilde{Y}_{idt}\hat{p}_{ds}) - E_i(\tilde{Y}_{idt})E_i(\hat{p}_{ds}) \\
&= E_i((\delta_c + p_{dt} + e_{idt})(p_{ds} - \bar{e}_c + \frac{1}{N_{ds}} \sum_{j \in d(s)} e_{jds})) - E_i(\tilde{Y}_{idt})E_i(\hat{p}_{ds}) \\
&= E_i((\delta_c + p_{dt} + e_{idt})(p_{ds} - \bar{e}_c + \frac{1}{N_{ds}} \sum_{j \in d(s)} e_{jds})) - E_i(\tilde{Y}_{idt})E_i(\hat{p}_{ds}) \\
&= Cov_i(p_{dt}, p_{ds}) - E_i(e_{idt}, \bar{e}_c) \\
&= Cov_i(p_{dt}, p_{ds}) - \frac{1}{N_{t-s}} \sum_i \frac{1}{N_{c(i)}} Var(e_{jdt}) \qquad (22)
\end{aligned}$$

where the fourth line follows since e_{idt} and e_{ids} are independent by assumption and so $E_i(e_{idt}e_{ids}) = 0$. Note further that $Cov_i(p_{dt}, p_{ds}) \to_p \sigma_{p(t-s)}$. Let $\hat{e}_{idt}$ denote the residuals from a regression of $\tilde{Y}_{idt}$ on doctor-time fixed effects. We can estimate the

second term using:

$$\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\frac{N_{dt(j)}}{N_{dt(j)}-1}\hat{e}^2_{jdt} \rightarrow_p$$
$$\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\frac{N_{dt(j)}}{N_{dt(j)}-1}Var(\hat{e}_{jdt}) =$$
$$\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\frac{N_{dt(j)}}{N_{dt(j)}-1}\left(\left(\frac{N_{dt(j)}-2}{N_{dt(j)}}\right)Var(e_{jdt}) + \frac{1}{N^2_{dt(j)}}\sum_{k\in d(t)} Var(e_{kdt})\right) =$$
$$\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\left(\left(\frac{N_{dt(j)}-2}{N_{dt(j)}-1}\right)Var(e_{jdt}) + \frac{1}{N_{dt(j)}(N_{dt(j)}-1)}\sum_{k\in d(t)} Var(e_{kdt})\right) =$$
$$\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}\left(\left(\frac{N_{dt(j)}-2}{N_{dt(j)}-1}\right)Var(e_{jdt}) + \frac{1}{N_{dt(j)}-1}Var(e_{jdt})\right) =$$
$$\frac{1}{N_{t-s}}\sum_i \frac{1}{N_{c(i)}}Var(e_{jdt}) \tag{23}$$

The case of the current year requires special consideration. We want to know $Cov(p_{dt}, \hat{p}^{JK}_{idt})$ where $\hat{p}^{JK}_{idt}$ is the jackknife estimate which we compute either using only prior observations or as a leave one out estimate. Let $t(i)$ denote the specific date on which individual i was treated and let N_{idt} denote the number of patients tested prior to patient i by doctor d in period t. In that case:

$$\begin{aligned} Cov_d(p_{dt}, \hat{p}^{JK}_{idt}) &= Cov_d(p_{dt}, \frac{1}{N_{idt}}\sum_{t(j)<t(i)} \tilde{Y}_{jdt}) \\ &= Cov_d(p_{dt}, \frac{1}{N_{idt}}\sum_{t(j)<t(i)} p_{dt} + \delta_c + e_{jdt}) \\ &= Var_d(p_{dt}) \end{aligned} \tag{24}$$

We estimate this term below.

The above derivation tells us the off-diagonal terms of Σ_p as well as the vector γ. The diagonal terms are given by $Var_d(\hat{p}_{dt}|N_d) = Var_d(\hat{p}_{dt}|N_{dt})$ since the variance of the estimated $\hat{p}_{dt}$ depends only on the number of observations in that period. If we observed enough doctors in every period with each possible number of observations, we could compute this variance empirically. Since we do not, we again rely on structure to relate the variance for each N_d to a single underlying across patient variance. Specifically, we impose that the mean and variance variance of true p_{dt} across doctors is the same for all N- $Var_{d,N}(p_{dt}) = Var_d(p_{dt})$ and $E_{d,N}(p_{dt}) = E_d(p_{dt})$.

For each value of $N_{dt} = N$, let D_{Nt} denote the set of patients belonging to doctors with exactly N patients at time t and let $|D_{Nt}|$ denote the number of such patients. Further, let $N_{c(i),D_{Nt}}$ denote the number of patients in clinic c belonging to D_{Nt}. Then we have:

$$
\begin{aligned}
& Var_i(\hat{p}_{dt}|N_{dt},t) =_p \\
& Var_d(p_{dt}|N_{dt},t) + Var_i\left(-\bar{e}_c + \frac{1}{N_{dt}}\sum_{i\in d(t)} e_{idt}|N_{dt},t\right) \\
= \quad & Var_d(p_{dt}|N_{dt},t) + \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(-\bar{e}_{c(i)} + \frac{1}{N_{dt(i)}}\sum_{j\in dt(i)} e_{jdt}\right)^2 \\
\to_p \quad & Var_d(p_{dt}|N_{dt},t) \\
& + \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{1}{N_{c(i)}^2}\sum_{j\in c(i)} e_{jtd}^2 + \frac{1}{N_{dt(i)}^2}\sum_{j\in dt(i)} e_{jdt}^2 - \frac{2}{N_{c(i)}N_{dt(i)}}\sum_{j\in dt(i)} e_{jdt}^2\right) \\
= \quad & Var_d(p_{dt}|N_{dt},t) + \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{N_{c(i),D_{Nt}}}{N_{c(i)}^2}e_{itd}^2 + \frac{1}{N_{dt(i)}}e_{idt}^2 - \frac{2}{N_{c(i)}}e_{idt}^2\right) \\
= \quad & Var_d(p_{dt}|N_{dt},t) \\
& + \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{N_{c(i),D_{Nt}}N_{dt(i)} + N_{c(i)}^2 - 2N_{c(i)}N_{dt(i)}}{N_{c(i)}^2 N_{dt(i)}}e_{idt}^2\right) \qquad (25)
\end{aligned}
$$

Analogous reasoning to equation 23 shows that we can compute the parenthetical term by substituting: $\frac{N_{dt(i)}}{N_{dtij)}-1}\hat{e}_{idt}^2$ for e_{idt}^2 when $N_{dt} > 1$.

With sufficient data, we could estimate $Var_d(\hat{p}_{dt}|N_{dt},t)$ for every (N_{dt},t) pair and use this to determine $Var_d(p_{dt}|N_{dt},t)$. Note however that this requires we observe enough physicians at that pair that the law of large numbers applies in the above derivation. Because this is not true for many (N_{dt},t) pairs, we instead assume that $Var_d(p_{dt}|N_{dt},t) = Var_d(p_{dt})$, a constant. We additionally assume that $Var_d(\hat{p}_{dt}|N_{dt},t) = Var_d(\hat{p}_{dt}|N_{dt})$ which follows if, for example, we assume that the distribution of e_{idt} is the same in each period.

We then proceed as follows. First, we estimate $Var_d(\hat{p}_{dt}|N_{dt},t)$ for all (N_{dt},t) pairs where we observe at least 20 doctors with 2 observations. Second, we estimate $Var_d(p_{dt})$ as the physician weighted average of these estimates. For each such pair, our estimate

is given by:

$$\begin{aligned} Var_d(p_{dt}) &= Var_d(\hat{p}_{dt}|N_{dt},t) - \\ &\quad \frac{1}{|D_{Nt}|}\sum_{i\in D_{Nt}}\left(\frac{N_{c(i),D_{Nt}}N_{dt(i)}+N_{c(i)}^2-2N_{c(i)}N_{dt(i)}}{N_{c(i)}^2(N_{dt(i)}-1)}\hat{e}_{idt}^2\right) \end{aligned} \tag{26}$$

Finally, we compute $Var_d(\hat{p}_{dt}|N_{dt})$ using the empirical distribution - we bin cases of $N_{dt} > 5$ (3% of all doctors - 2/3 of which have 6 or 7 observations) into a single bin - for the very small number of doctor-times with more than 5 doctors, this will tend to overstate the noise in the variance and thus understate the weight they should receive.

Finally, consider the constant term $\psi_0(N_{d1},...,N_{d(t-1)})$. This term is given by:

$$\begin{aligned} \psi_0(N_{d1},...,N_{d(t-1)}) &= E_d(p_{dt}-\sum_{s=1}^{t-1}\psi_s(N_{d1},...,N_{d(t-1)})\hat{p}_{ds}) \\ &= \bar{p}_t \end{aligned} \tag{27}$$

since $E_d(\hat{p}_{ds}) = 0$ (by construction, since we partial out time fixed effects when we define $\tilde{Y}_{idt}$). To deal with this term we can just include time fixed effects.