THE SENSITIVITY OF LINEAR REGRESSION COEFFICIENTS' CONFIDENCE LIMITS TO THE OMISSION OF A CONFOUNDER

BY CARRIE A. HOSMAN[1], BEN B. HANSEN[2] AND PAUL W. HOLLAND

*University of Michigan, University of Michigan
and Paul Holland Consulting Corporation*

Omitted variable bias can affect treatment effect estimates obtained from observational data due to the lack of random assignment to treatment groups. Sensitivity analyses adjust these estimates to quantify the impact of potential omitted variables. This paper presents methods of sensitivity analysis to adjust interval estimates of treatment effect—both the point estimate and standard error—obtained using multiple linear regression. Central to our approach is what we term *benchmarking*, the use of data to establish reference points for speculation about omitted confounders. The method adapts to treatment effects that may differ by subgroup, to scenarios involving omission of multiple variables, and to combinations of covariance adjustment with propensity score stratification. We illustrate it using data from an influential study of health outcomes of patients admitted to critical care.

1. Introduction

1.1. *Methodological context.* In a common use of multiple linear regression, one regresses an outcome variable on a treatment variable and adjustment variables, then interprets the fitted treatment-variable coefficient as an estimate of the treatment's effect on the outcome. The interpretation relies on the assumptions of the linear model and some assumption to the effect that there either are no unmeasured confounders or at least none that demand adjustment. (Ignorability of treatment assignment [Rubin (1978); Holland (1988), Appendix], is one such assumption; there are many variants.) The linearity assumptions are often testable given the data, but the remaining assumption is not. When regression results are questioned, it's often the nonconfounding assumption that is the focus of doubt.

Because the issue arises even with the most thorough observational studies, adjusting for any number of covariates, it fuels cynicism about observational research. If the possibility of unmeasured variable bias can't be removed, then why bother with potential confounders, particularly those that are difficult to measure, or not obvious threats? It might be clear that the damage from omitting a confounder W would be reduced by adjustment for available correlates of W, yet,

because introducing these correlates would draw attention to the absence of W, not at all clear that effecting the additional adjustments would enhance the credibility of the research. Plainly, the problem here is not the methodological strategy of broadly adjusting for relevant baseline characteristics but an absence of, or lack of awareness of, suitable methods with which to quantify benefits of more comprehensive confounder controls.

Sensitivity analyses, procedures quantifying the degree of omitted variable bias needed to nullify or reverse key conclusions of a study, can help. Sensitivity analysis methods for various models and data structures are proposed in Cornfield et al. (1959), Rosenbaum and Rubin (1983), Rosenbaum (1988), Copas and Li (1997), Robins, Rotnitzky and Scharfstein (2000), Scharfstein and Irizarry (2003), Marcus (1997), Lin, Psaty and Kronmal (1998), Frank (2000) and Imbens (2003), among others, the last four bearing closest resemblance to the approach to be presented here. Invariably the methods start by in some way quantifying relationships between hypothetical omitted variables and included ones, go on to give an algorithm for converting these parameters into impacts on estimates, p-values or confidence limits, and then leave to researchers themselves the task of deciding what parameter values are plausible or relevant. Here we develop a method following the first parts of this general recipe, but then doing a bit more to help researchers calibrate intuitions about speculation parameters.

The resulting technique possesses a combination of advantages, making it both uniquely practicable and conducive to certain insights. First, it applies to inferences made with ordinary multiple regression, as we show in Section 2, as well as to inferences made with regression in combination with propensity-score stratification, a topic discussed in Section 4.3. Second, it quantifies relationships between omitted and included variables in terms intrinsic to multiple regression, permitting intuitions for the relationships to be calibrated with a few additional regression fits (Section 2). [Angrist and Imbens (2002), Section 4, suggest such calibration in a related context.] Third, it represents effects of an omission with just two quantities, one tracking confoundedness with the treatment variable and the other measuring conditional association with the outcome (Section 2). Fourth, there are important practical benefits to sensitivity analysis based on both of these quantities—"dual" sensitivity analyses, in Gastwirth, Krieger and Rosenbaum's terminology (1998)—in that analysis based only on confoundedness with the treatment variable is likely to overstate sensitivity. Our method makes this plain (Section 3), as we will demonstrate with a case study to be introduced presently, although some other methods may obscure it. Fifth, it gives closed-form characterizations of how confidence intervals, as opposed only to estimates or hypothesis tests, could be changed by inclusion of the omitted confounder (Section 3.2). Sixth, the method readily adapts to analyses in which several omissions are suspected, or where interactions with the treatment variable are used to handle possible effect heterogeneity (Section 4). Seventh, the same application brings to light certain practical advantages of the use of propensity scores which to our knowledge have not previously been noted (Section 4.3).

1.2. *A case study.*   Our application is to Connors et al.'s (1996) highly influential, and controversial, study of the critical-care procedure known alternatively as Swan–Ganz, pulmonary artery or right heart catheterization (RHC). RHC is a procedure to perform continuous measurements of blood pressure in the heart and large blood vessels of the lungs. Introduced in 1970, it became standard procedure without first being tested in clinical studies, as might be expected today, and empirical assessments that were subsequently conducted failed to uncover evidence that it improved medical outcomes [e.g., Gore et al. (1987); Zion et al. (1990)]. However, these studies were criticized for insufficient confounder controls. Using a large sample, good measures and extensive adjustments for background variables, Connors et al. (1996) echoed the disappointments of the earlier assessments and went further, finding RHC to worsen, rather than improve, both mortality and the duration of treatment in critical care. Each of these studies used nonrandomized data and is in some degree vulnerable to omitted variable bias. Although the results of subsequent randomized trials have been largely consistent with this picture [Rhodes et al. (2002); Sandhan et al. (2003); Shah and Stevenson (2004); Richard et al. (2003); Harvey et al. (2005)], the procedure remains a staple of critical care, and the surrounding debate continues. This paper examines how the omission of covariates from Connors et al.'s data might skew point and interval estimates of RHC's effect on length of stay, in the process shedding light on the degree of protection from omitted variables afforded by included ones.

1.3. *The SUPPORT data.*   Connors et al.'s (1996) data come from the Study to Understand Prognoses and Preferences for Outcomes and Risks of Treatments (SUPPORT). The study collected data on the decision-making and outcomes of seriously ill, hospitalized adult patients. Patients included in the study had to meet certain entry criteria and a predefined level of severity of illness. All 5735 SUPPORT patients who were admitted to an ICU in the first 24 hours of study were analyzed together, for reasons detailed in Connors et al. (1996). Data include initial disease categories upon admission, physiological measurements, and demographic information. Patients were coded as having had RHC if it was performed within the first 24 hours of the study.

Length-of-stay is the difference between a patient's recorded dates of entry and exit from the hospital. Here we study the omitted variable sensitivity of Connors et al.'s finding that RHC increases costs, by lengthening stays in the hospital. For cost analysis it is logical to compare lengths of stay irrespective of whether those stays ended in death, as Connors et al. do and as we do also. Since a medical procedure could, in principle, shorten stays only by increasing mortality, comparisons like this one speak directly to economic effects but not to health effects of the procedure. (A study focused on health outcomes, as opposed to resource utilization, would most naturally begin by analyzing survival and continue with analyses of patient experience, including duration of stay in the hospital, that address the issue of censoring by death [e.g., Rubin (2006)]. Such analyses require methods

other than ordinary multiple regression, however, and sensitivity analysis for them is somewhat beyond the scope of this paper.)

1.4. *First-pass regression results.* Length-of-stay is right-skewed. We log-transform it before regressing it on RHC and covariates. There are approximately 50 covariates for which regression adjustments might be considered; a backward stepwise selection procedure reduces this number to 19, and estimates the RHC effect as 0.11: taking the log transformation into account, RHC seems to increase lengths of stay by about $100 * (\exp(0.11) - 1) = 12\%$. To reflect variability added by variable selection [Faraway (1992)], we ran a nonparametric bootstrap to determine null quantiles for regression parameters' t-statistics; in this case, however, bootstrap t-quantiles were similar to quantiles of the appropriate t-distribution. (Code for these and other computations discussed in the paper appears in a supplement [Hosman, Hansen and Holland (2010)].) Either way, a 95% confidence interval for the RHC effect ranges from 0.06 to 0.16, encompassing increases of 6% up to 18%.

2. Effect estimates accounting for omitted variables

To understand how an omitted variable, W, could affect the coefficient on the treatment variable, Z, in a hypothetical regression of a given outcome on these and other variables,

$$(1) \qquad Y = \alpha \mathbf{X}^T + \beta Z + \delta W + e,$$

it is well to begin by examining how included variables affect the treatment coefficient in the regression that was actually performed,

$$(2) \qquad Y = a \mathbf{X}^T + bZ + e.$$

This process lends context to an accompanying sensitivity analysis.

2.1. *Omitted variable bias.* Perhaps the most familiar way of comparing regressors is in terms of their effects on R^2; we begin in this mode. Of 33 regressors (columns of the design matrix corresponding to the set of 19 covariates, some of which are categorical with multiple levels), the one that moves R^2 the most is "DNR status," an indicator of whether there was a do-not-resuscitate order on file for the subject on the day that he first qualified for inclusion in the study. Without it, R^2 would be reduced from 0.141 to 0.112; in contrast, removal of the next-most predictive regressor reduces R^2 only to 0.131. On this basis one might expect DNR status to have a relatively large effect on inferences about the RHC effect, and in a sense it does: omitting it from the regression equation increases the RHC coefficient estimate from 0.112 to 0.143, 1.2 standard errors. For comparison, consider a regressor which contributes more modestly to the outcome regression: omission of "bleeding in the upper GI tract," for instance, reduces R^2 only by 0.001, the 28th smallest of 33 such reductions, and removing it moves the treatment coefficient only a few percent of an SE. Blurring the distinction between parameters

and estimates just a little, we refer to this difference as bias, the bias that would have been incurred by omitting the DNR variable. It appears that R^2 does track omitted-variable bias.

There is a simple relationship between a covariate W's contribution to explaining the response, as measured by its impact on R^2, and the bias incurred by its omission. According to Proposition 2.1, this bias is the product of the standard error on Z, the treatment variable, as calculated without regression adjustment for W (another term to be explained presently) and $\rho_{y \cdot w|z\mathbf{x}}$, the partial correlation of W and the response given Z and remaining covariates, X. In turn, $\rho^2_{y \cdot w|z\mathbf{x}}$ equals $[(1 - R^2_{\text{no } W}) - (1 - R^2_{\text{with } W})]/(1 - R^2_{\text{no } W})$, the proportionate reduction in unexplained variance when W is added as a regressor [Christensen (1996), Chapter 6]. That is to say, the square of the bias due to omitting a covariate is linear in the fraction by which that covariate would reduce unexplained variance. Returning to our example, DNR status consumes 3.3% of outcome variation that would otherwise be unexplained, which is 16 interquartile ranges greater than the third quartile of $\rho^2_{y \cdot w|z\mathbf{x}}$'s associated with the 33 available regressors; the $\rho^2_{y \cdot w|z\mathbf{x}}$ for upper GI bleeding, in contrast, is 0.1%, just above the first quartile.

The degree to which an omitted variable can bias effect estimates through the value of $\rho^2_{y \cdot w|z\mathbf{x}}$ depends on the value of R^2 with which we begin. In our example, the value of R^2 is relatively small; when the R^2 value is bigger, changes in R^2 have a more pronounced effect. Imagining another regression with a higher baseline R^2 and thus less unexplained outcome variation, consider an unmeasured covariate that reduces R^2 by the same percentage as the DNR variable in our regression. As a result of having less unexplained variation to start, the proportionate reduction in unexplained variation, $\rho^2_{y \cdot w|z\mathbf{x}}$, must be larger than that of DNR in our example. Such a variable would consequently have a larger effect on the estimate of omitted variable bias.

The remaining factor in Proposition 2.1's expression for omitted variable bias expresses the degree to which the variable is confounded with treatment. This confounding turns out to figure more centrally in omitted variable bias, despite the fact that measurements of it arise less often in regression analysis than do $\rho_{y \cdot w|z\mathbf{x}}$ or R^2. A straightforward descriptive measurement falls out of the regression of Z on the remaining regressors from the outcome regression: their confoundedness with Z is reflected in their t-statistics, the ratios of their coefficient estimates to their conventional standard errors. Note that these are here considered as descriptive, not inferential, statistics; were this auxiliary regression instead being used for statistical inference, a nonlinear model may be required for binary Z, in which case these nominal standard errors may not be appropriate. Denoting by t_W the t-ratio attaching in this way to a regressor W, by b the RHC-coefficient in the absence of W and by β the same coefficient when W is included, the bias due to omission of a regressor decomposes as follows.

PROPOSITION 2.1.   *If $R^2_{y \cdot z\mathbf{x}} < 1$ and t_W is finite, then*

(3) $$\hat{b} - \hat{\beta} = \mathrm{SE}(\hat{b})t_W \rho_{y \cdot w|z\mathbf{x}}.$$

A proof is given in Section 3.2.

Besides being a strong predictor of the response, DNR status ranks highly in terms of its confounding with the treatment: if RHC is regressed on the covariates, then its t-statistic is -6.5, the fifth-largest in magnitude. (Table 1 displays t_W values, along with a few square roots of F-statistics, to be explained in Section 4.1, for multicategory variables. The magnitudes of the F-statistics are not directly comparable to those of the t-statistics.) Consistent with there being little bias associated with removing it, upper GI bleeding is as weakly related to the treatment variable as it was to the response; its t is only -0.7, the fifth-smallest

TABLE 1
Selected included covariates' relation to treatment and response variables, given remaining included variables

	Confounding with RHC ($|t_W|$) (rounded)	% decrease in unexplained variation by adding W ($100\rho^2_{y \cdot w|z\mathbf{x}}$)
Income*	6.8	0.3
Primary initial disease cat.*	48.1	3.4
Secondary initial disease cat.*	20.2	0.8
Comorbidities illness:		
Renal disease	2.1	0.2
Upper GI bleed	0.7	0.1
Day 1 measurements:		
APACHE score	5.1	0.1
White blood cell count	0.5	0.0
Heart rate	2.5	0.0
Temperature	2.3	0.1
PaO2/FIO2	15.4	0.1
Albumin	2.3	0.7
Hematocrit	3.3	0.9
Bilirubin	2.2	0.1
Sodium	3.1	0.1
PaCo2	6.8	0.2
DNR	6.5	3.3
PH	3.7	0.3
Admit diagnosis categories:		
Neurology	5.9	0.2
Hematology	3.6	0.1

*Categorical variables with multiple levels. (Strictly speaking, these rows of the table give $F_W^{1/2}$ in the middle column, not t_W; see Section 4.1.)

in magnitude. The strongest confounder among the adjustment variables is the PaO2/FIO2 ratio, a reflection of oxygenation in the lungs, with $t = -15.3$. On the other hand, its relatively small contribution toward explaining variation in the outcome, $\rho^2_{y \cdot w | z\mathbf{x}} = 0.0007$, limits the effects of its removal on the treatment coefficient; here $\hat{b} - \hat{\beta}$ is 0.011, or 40% of an SE. Although the effect on the treatment coefficient of excluding the PaO2/FIO2 variable is tempered by its small corresponding $\rho^2_{y \cdot w | z\mathbf{x}}$, (3) says that an actual omitted variable this confounded with treatment could contribute up to 15.3 standard errors' worth of bias (depending on how much more outcome variation it explains). If this possibility is not merely notional, it follows that uncertainty assessments based in the usual way on $\mathrm{SE}(\hat{b})$ are quite incomplete.

Proposition 2.1 suggests that the degree of an omitted variable's confounding with the treatment plays a potentially much larger role in determining the bias from its omission than does its conditional association with the response. As it can explain no more than 100% of otherwise unexplained variation, in general $\rho^2_{y \cdot w | z\mathbf{x}} \le 1$, ensuring in turn that $|\hat{b} - \hat{\beta}| / \mathrm{SE}(\hat{b}) \le |t_W|$. In contrast, nothing bounds the t-ratios t_W—covariates highly collinear with the treatment can introduce biases that are arbitrarily large, at least in principle. We refer to an unsigned statistic $|t_W|$ as the *treatment confounding* of W a covariate or potential covariate W.

Table 1 reports t_W and $\rho_{y \cdot w | z\mathbf{x}}$ values that actually occur in the SUPPORT data, placing observed covariates in the role of W one at a time. With these data, t_W is often of a magnitude to make large biases possible (although this is somewhat tempered by relatively small $\rho_{y \cdot w | z\mathbf{x}}$'s). The calculations permit the statistician and his audience to calibrate intuitions for t_W against actual variables and their relations to the treatment. We refer to this as *benchmarking* of treatment confounding.

2.2. *Omitted variables and the treatment's standard error.* The representation (3) of omitted-variable bias, as the product of $\mathrm{SE}(\hat{b})$ with other factors, lends hope that a meaningful combinations of errors due to variable omissions and sampling variability might be bounded by a multiple of $\mathrm{SE}(\hat{b})$. A tempting solution is simply to add to the normal 97.5 percentile 1.96, or to whatever other multiplier of the SE has been found appropriate to describe likely sampling error, a quantity large enough to bound plausible values of $|t_W \rho_{y \cdot w | z\mathbf{x}}|$. This solution will often miss the mark, however, since omitting a covariate affects standard errors as well as coefficient estimates. This is illustrated in Table 2, which shows effects on estimation of the RHC-coefficient of omitting one-by-one the regressors discussed in Section 2.1.

According to Table 2, adjusting the RHC coefficient for covariates including DNR order gives a smaller standard error, 0.0260, than does excluding DNR order and adjusting for remaining covariates only (SE = 0.0264). In contrast, inclusion of PaO2/FIO2 among the adjustment variables has the effect of increasing the standard error, from 0.0255 to 0.0260. Including or excluding upper GI tract

		Standard error	
Excluded variable	**Estimate**	**df = 5700**	**df = 50**
No exclusion	0.112	0.0260	0.278
DNR order	0.143	0.0264	0.206
GI bleed	0.112	0.0260	0.274
PaO2/FIO2	0.122	0.0255	0.115

bleeding leaves the standard error unchanged, suggesting that omitted variables associated with little bias should have little effect on the standard error. This turns out to be true, in a limited sense, as can be read from Proposition 2.2: the same statistics governing the bias due to omission of a covariate W also govern its effects on the standard error, in such a way that when both t_W and $\rho_{y \cdot w|z\mathbf{x}}$ are small, then including or excluding W has little effect on the Z-coefficient's standard error.

PROPOSITION 2.2. *If* $R^2_{y \cdot z\mathbf{x}} < 1$ *and* t_W *is finite, then*

$$(4) \qquad \mathrm{SE}(\hat{\beta}) = \mathrm{SE}(\hat{b})\sqrt{1 + \frac{1 + t_W^2}{\mathrm{df} - 1}}\sqrt{1 - \rho^2_{y \cdot w|z\mathbf{x}}}.$$

Here $\mathrm{df} = n - \mathrm{rank}(X) - 1$, *the residual degrees of freedom after Y is regressed on X and Z.*

Proposition 2.2 will be strengthened in Proposition 4.3, which is proved in the Appendix.

Whereas $|t_W|$ and $|\rho_{y \cdot w|z\mathbf{x}}|$ both associate directly with the size of the bias from omitting W, they act in opposite directions on the standard error—reflecting the fact that increasing R^2 tends to reduce coefficients' standard errors, except when it's increased at the expense of introducing colinearity among regressors. The difference explains why omitting some variables increases standard errors, whereas omitting others decreases them.

It also reaffirms that omitted variables' effects on statistical inferences are only incompletely reflected in omitted variable bias. In the context of the present example, variable omissions have only modest effects on the standard error; but this is a consequence of the sample being quite large ($n = 5735$) relative to the rank of the covariate matrix (33): barring astronomically large t_Ws, with df = 5700 the middle factor on the left side of (4) has to fall close to 1; in consequence, even a t_W of 16 inflates the standard error by at most 2%, according to (4). In moderate and small samples, standard errors are more sensitive. The third column of Table 2 presents standard errors as they would be if the sample had been much smaller, such that

df $= 50$, but with the same sample means and covariances. While the omission of a variable like GI bleed, which is only weakly related with already included variables, still leaves the treatment's standard error much the same, variables that are either moderately or strongly confounded with the treatment now cause wide shifts in the magnitude of the standard error. Adjustment for the PaO2/FIO2 measurement, for example, now increases the treatment's standard error by a whopping 59%.

2.3. *One sensitivity parameter or two?*   Should we be inclined to consider worst-case scenarios, Proposition 2.2 reaffirms Proposition 2.1's message that the omitted variable's treatment confounding, not its potential to increase R^2, most demands our attention. The $\rho_{y \cdot w|z\mathbf{x}}$-contribution to $\mathrm{SE}(\hat{\beta})$ is a factor bounded above by 1, whereas the t_W-factor can be arbitrarily large—as could the t_W-contribution to omitted-variable bias. The greater potential disturbance from confounding with the treatment than from large new contributions to the outcome regression seems to be a feature multiple regression analysis shares with other statistical techniques; indeed, sensitivity analysis methods which set out to limit inference errors in terms of a single sensitivity parameter parametrize confounding with treatment, not predictivity of the response, countenancing arbitrarily strong response-predictors in the role of the hypothetical omitted variable [e.g., Rosenbaum (1988); Rosenbaum and Silber (2009)].

What would such an analysis suggest about the sensitivity of our RHC results to the omission of a confounder not unlike included variables? PaO2/FIO2 was sufficiently confounded with treatment to indicate as many as 16 standard errors' worth of bias. Yet its inclusion in the regression adjustment moves the treatment effect by less than half a standard error. To have moved $\hat{\beta}$ so much, it would have had to consume all of the variation in Y that was not explained without it, something no one would have expected it to do. To the contrary, inspection of other variables' contributions to R^2, as enabled by Table 1, suggests that 0.1 (for instance) would be a quite generous limit on $|\rho_{y \cdot w|z\mathbf{x}}|$. This in turn would restrict omitted variable bias due to a treatment-confounder as strong as PaO2/FIO2 to 1.6 SEs—still a meaningful addition to the uncertainty estimate, if a less alarmist one. Rather than reducing the number of sensitivity parameters by permitting $\rho_{y \cdot w|z\mathbf{x}}$ to fall anywhere within its a priori limits, it is more appealing to retain $\rho_{y \cdot w|z\mathbf{x}}$, restricting it within generous bounds of plausibility.

3. Sensitivity intervals   Taken together, Propositions 2.1 and 2.2 enable a precise, closed-form description of the union of interval estimates $\{\hat{\beta} \pm q\,\mathrm{SE}(\hat{\beta}) : |t_W| \leq T, \rho_{y \cdot w|z\mathbf{x}}^2 \leq R\}$, for any nonnegative limits T, R on omitted variables' treatment confounding and contributions to reducing unexplained variation—the collection of Z-slopes falling within the confidence interval after addition of a covariate W such that $-T \leq t_W \leq T$ and $\rho_{y \cdot w|z\mathbf{x}}^2 \leq R$. Such a union of intervals is itself an interval: following Rosenbaum (2002), we call it a *sensitivity*

interval; and following Small (2007), we refer to the determining set of permissible values for $(t_W, \rho^2_{y \cdot w|z\mathbf{x}})$ as a *sensitivity zone*. The mapping of sensitivity zones to sensitivity intervals is given in Proposition 3.1.

PROPOSITION 3.1. *Let* Y, $\mathbf{X}$, Z *and* W *be as in* (1) *and* (2), *with both regressions fit either by ordinary least squares or by weighted least squares with common weights. Assume* $R^2_{y \cdot z\mathbf{x}} < 1$. *Let* $\rho_{y \cdot w|z\mathbf{x}}$, t_W *and* df *be as defined in Section* 2, *fix* $q > 0$ *and write* $C_d(t)$ *for* $[1 + (1 + t^2)/(d - 1)]^{1/2}$.

(i) *Assuming only* $t^2_W \leq T < \infty$,

$$(5) \qquad \hat{\beta} \pm q\,\mathrm{SE}(\hat{\beta}) = b \pm \left[-t_W \rho_{y \cdot w|z\mathbf{x}} + q C_{\mathrm{df}}(t_W) \sqrt{1 - \rho^2_{y \cdot w|z\mathbf{x}}} \right] \mathrm{SE}(\hat{b})$$

$$(6) \qquad \qquad \subseteq b \pm \sqrt{T^2 + q^2 C_{\mathrm{df}}(T)^2}\,\mathrm{SE}(\hat{b}).$$

(ii) *Assuming* $t^2_W \leq T < \infty$, $\rho^2_{y \cdot w|z\mathbf{x}} \leq R$ *where* $0 < R < T^2/(T^2 + q^2 C_{\mathrm{df}}(T)^2)$,

$$(7) \qquad \hat{\beta} \pm q\,\mathrm{SE}(\hat{\beta}) \subseteq b \pm [T R^{1/2} + q C_{\mathrm{df}}(T)(1 - R)^{1/2}]\,\mathrm{SE}(\hat{b}).$$

(iii) *If, on the other hand,* $T^2/(T^2 + q^2 C_{\mathrm{df}}(T)^2) < R < 1$, *then* (6) *is sharp*: *its right-hand side represents the union of* (5) *as* $(\rho_{y \cdot w|z\mathbf{x}}, t_W)$ *ranges over the sensitivity zone* $[-R^{1/2}, R^{1/2}] \times [-T, T]$.

Proposition 3.1 expresses solutions to the constrained optimization problems of determining the smallest $\hat{\beta} - q\,\mathrm{SE}(\hat{\beta})$ and largest $\hat{\beta} + q\,\mathrm{SE}(\hat{\beta})$ consistent with assumed restrictions on $\rho_{y \cdot w|z\mathbf{x}}$ and t_W: in part (i), the restrictions pertain only to t_W, while parts (ii) and (iii) impose restrictions on $\rho_{y \cdot w|z\mathbf{x}}$ also. The proof of the proposition appears in the Appendix.

REMARKS. (a) In many problems df will be large in comparison with plausible values of t^2_W, confining $C_{\mathrm{df}}(t_W)$ and $C_{\mathrm{df}}(T)$ to the immediate vicinity of 1. (b) Part (ii) says that if the magnitude of $\rho_{y \cdot w|z\mathbf{x}}$ is assumed to be small or moderate, then the extremes of the sensitivity interval correspond to speculation parameter values sitting at extremes of the sensitivity zone—the same extremes at which (signed) omitted variable bias is minimized or maximized. According to part (iii), however, if $\rho_{y \cdot w|z\mathbf{x}}$ is permitted to be large, then restricting attention to sensitivity parameter values that maximize or minimize omitted variable bias may lead the statistician to underestimate the proper extent of the sensitivity interval.

3.1. *Pegging the boundaries of the sensitivity zone.* Because our analysis began with covariate selection, there are a number of deliberately omitted variables that can be used to peg at least one boundary of the sensitivity zone. As one might

TABLE 3

95% sensitivity intervals for the treatment coefficient, with the putative unobserved variable's treatment confounding ($|t_W|$) hypothesized to be no greater than the treatment confounding of 6 deliberately omitted variables. The decrease it would bring to the variance of response residuals is hypothesized to be no greater than either of 2 index values, 1% and 10%, or is not restricted

Variable	Treatment confounding benchmark		% decrease in unexplained variation $(100\rho^2_{y \cdot w|z\mathbf{x}})$		
			1%	**10%**	**Unrestricted**
Insurance class	12.2	most	(0.03, 0.20)	(−0.04, 0.26)	(−0.21, 0.43)
Respiratory eval.	8.9	some	(0.04, 0.19)	(−0.01, 0.23)	(−0.12, 0.35)
Mean blood press.	8.6	some	(0.04, 0.19)	(−0.01, 0.23)	(−0.12, 0.34)
Cardiovascular eval.	8.5	some	(0.04, 0.19)	(−0.01, 0.23)	(−0.11, 0.34)
Weight (kg)	6.1	some	(0.04, 0.18)	(0.01, 0.21)	(−0.05, 0.28)
Immunosuppression	0.4	least	(0.06, 0.16)	(0.06, 0.16)	(0.06, 0.16)

expect, the covariates eliminated by the stepwise procedure add little to those variables that were included in terms of prediction of the response, and it is too optimistic to suppose of a genuinely unmeasured confounder that its contribution to the outcome regression would be no greater than that of measured covariates that variable selection would put aside. On the other hand, confounding with Z plays little role in common stepwise procedures like the one we used, and the deliberately omitted variables can be used to guide intuitions about plausible values of t_W. We selected six of these whose partial associations with RHC spanned the full range of such associations among stepwise-eliminated covariates and used their t_W-values to delimit the first dimension of several sensitivity zones. Table 3 delineates the t_W-part of the sensitivity zone accordingly, choosing the bound T on treatment confounding to coincide with the magnitude of confounding, conditional on stepwise-selected covariates, between the treatment and each of our six covariates.

The benchmarking method, using known variables to determine plausible values of t_W, informs targeted speculation about the potential effects of omitted variable bias. Using existing information in this way, we can speculate about the effects of omitted covariates that are of a similar nature to measured covariates—t_W benchmarks extracted from partial demographic information might reasonably predict the t_W values that would attach to additional demographic variables, were they available. To calibrate intuitions about omitted variables that are different in kind from included ones, reference values for t_W might also be obtained from external data sets.

Many analysts will have sharper intuition for potential covariates' effect on R^2, making it relatively easy to set plausible limits on $\rho^2_{y \cdot w|z\mathbf{x}}$. Table 3 considers $\rho^2_{y \cdot w|z\mathbf{x}} \leq 0.01$ or 0.10, which may be useful as general starting points. In the present case study, for instance, when included covariates are removed and put

in the role of W, $\rho^2_{y \cdot w|z\mathbf{x}} = 0.01$ corresponds approximately to the second most predictive of them, and the strongest single predictor (DNR order) gives $\rho^2_{y \cdot w|z\mathbf{x}} = 0.03$. It appears that $\rho^2_{y \cdot w|z\mathbf{x}} \leq 0.1$ is a rather conservative bound: it is difficult even to find sets of included covariates that jointly contribute so much to the outcome regression as this. Only by simultaneously placing all of the covariates into the role of W, leaving the intercept alone in the role of X, does $\rho^2_{y \cdot w|z\mathbf{x}}$ reach 0.10. Restoring these variables into the regression, and barring scenarios of still-omitted variables for which $\rho^2_{y \cdot w|z\mathbf{x}}$ exceeds 0.1, Table 3 shows the result of a positive treatment effect to be sensitive to omitted confounding on par with some of the strongest of the included confounders, but insensitive to confounding weaker than that. If benchmarking leads to accurate guesses about the values of treatment confounding and reduction in unexplained variance, and if the linear model would hold were the omitted confounder added to the regressors, then 95% sensitivity intervals will have 95% coverage, despite the variable omission.

3.2. *Basis for sensitivity formulas.* Propositions 2.1, 2.2 and 3.1 extend better-known descriptions of bias in regression coefficients' point estimates due to variable omission [e.g., Seber (1977), page 66] to interval estimates. They also have antecedents in earlier literature on numerical adjustment of multiple regression results for the addition or removal of a covariate [Cochran (1938)]. Of the three, Proposition 2.1's proof is the most illuminating. It also conveys the flavor of the others, which appear in the Appendix.

Consider $\mathbf{X}$ to be a matrix containing a column of 1's (or columns from which a column of 1's can be recovered as a linear combination) and let Y, Z and W be column vectors of common extent, equal to the number of rows of $\mathbf{X}$. An inner product is defined as $(A, B) := \sum w_i a_i b_i / \sum w_i$, where w_i is a quadratic weight for the ith observation (in the case of unweighted least squares regression, $w_i \equiv 1$). Write $\mathbf{1}$ for the column vector of 1s. For vectors A, B and C, let $\mathrm{Pj}(A|B, C)$ represent the projection of A into the subspace spanned by B and C. Variances and covariances are defined as follows: $\sigma_{ab \cdot c} := (A - \mathrm{Pj}(A|C), B - \mathrm{Pj}(B|C))$, $\sigma^2_{a \cdot c} = \sigma_{aa \cdot c}$; $\sigma_{ab} = \sigma_{ab \cdot \mathbf{1}}$, $\sigma^2_a = \sigma^2_{a \cdot \mathbf{1}}$. Partial correlations are then given as follows: $\rho_{ab} := \sigma_{ab}/(\sigma_a \sigma_b)$; $\rho_{ab \cdot c} := \sigma_{ab \cdot c}/(\sigma_{a \cdot c} \sigma_{b \cdot c})$. Denote the degrees of freedom available for estimating b as $\mathrm{df} = n - m - 1$, where $m = \mathrm{column.rank}(X)$. The nominal standard error estimates for $\hat{b}$ and $\hat{\beta}$ [cf. (2) and (1)] are then

$$(8) \qquad \mathrm{SE}(\hat{b}) = \mathrm{df}^{-1/2} \frac{\sigma_{y \cdot z\mathbf{x}}}{\sigma_{z \cdot \mathbf{x}}} \quad \text{and} \quad \mathrm{SE}(\hat{\beta}) = (\mathrm{df} - 1)^{-1/2} \frac{\sigma_{y \cdot z\mathbf{x}w}}{\sigma_{z \cdot \mathbf{x}w}}.$$

PROOF OF PROPOSITION 2.1. To show $\hat{b} - \hat{\beta} = \mathrm{SE}(\hat{b}) t_W \rho_{y \cdot w|z\mathbf{x}}$, write

$$(9) \qquad \mathrm{Pj}(W|Z, \mathbf{X}) =: B^* Z + C^* \mathbf{X}^t.$$

Using (9) to project the OLS estimate of regression (1) onto the span of $(\mathbf{X}, Z)$ and then comparing to (2) gives $\hat{b} - \hat{\beta} = B^* \delta$, a well-known result [Seber (1977), page 66].

Write $W^{\perp \mathbf{x}}$ for $W - \mathrm{Pj}(W|\mathbf{X})$, $Z^{\perp \mathbf{x}}$ for $Z - \mathrm{Pj}(Z|\mathbf{X})$, $Y^{\perp z\mathbf{x}}$ for $Y - \mathrm{Pj}(Y|Z, \mathbf{X})$, and $W^{\perp z\mathbf{x}}$ for $W - \mathrm{Pj}(W|Z, \mathbf{X})$. Then $\mathrm{Pj}(W^{\perp \mathbf{x}}|Z^{\perp \mathbf{x}}) = B^* Z^{\perp \mathbf{x}}$, and $\mathrm{Pj}(Y^{\perp z\mathbf{x}}|W^{\perp z\mathbf{x}}) = \hat{\delta} W^{\perp z\mathbf{x}}$. These formulas imply $B^* = \sigma_{wz \cdot \mathbf{x}}/\sigma^2_{z \cdot \mathbf{x}}$ and $\hat{\delta} = \sigma_{yw \cdot z\mathbf{x}}/\sigma^2_{w \cdot z\mathbf{x}} = \rho_{yw \cdot z\mathbf{x}}\sigma_{y \cdot z\mathbf{x}}/\sigma_{w \cdot z\mathbf{x}}$, so that $\hat{b} - \hat{\beta} = B^*\hat{\delta}$ can be written as the product of $\sigma_{y \cdot z\mathbf{x}}/\sigma_{z \cdot \mathbf{x}}$, $\sigma_{wz \cdot \mathbf{x}}/(\sigma_{z \cdot \mathbf{x}}\sigma_{w \cdot z\mathbf{x}})$ and $\rho_{yw \cdot z\mathbf{x}}$. Introducing mutually canceling factors of $(\mathrm{df})^{\pm 1/2}$ to the first and second of these and applying (8) turns this into the product of $\mathrm{SE}(\hat{b})$, $(\mathrm{df})^{1/2}\sigma_{wz \cdot \mathbf{x}}/(\sigma_{z \cdot \mathbf{x}}\sigma_{w \cdot z\mathbf{x}})$ and $\rho_{yw \cdot z\mathbf{x}}$. But t_W is just the ratio of $\sigma_{wz \cdot \mathbf{x}}/\sigma^2_{w \cdot \mathbf{x}}$ to $\sigma_{z \cdot w\mathbf{x}}/[(\mathrm{df})^{1/2}\sigma_{w \cdot \mathbf{x}}]$, which simplifies to the second of these terms, by way of the identity $\sigma^2_{z \cdot \mathbf{x}}\sigma^2_{w \cdot z\mathbf{x}} = \sigma^2_{w \cdot \mathbf{x}}\sigma^2_{z \cdot w\mathbf{x}}$ (an algebraic consequence of the definition of $\sigma^2_{a \cdot c}$). The result follows. $\square$

4. Extensions As it is presented in Section 2, our method explores sensitivities of covariance-adjusted estimates of a main effect to the omission of a single covariate. It may appear to be limited, then, to effect estimates made by linear covariate adjustment, without interaction terms or other allowances for heterogeneity of treatment effects, and to hidden bias scenarios involving omission of a single variable, rather than several. Such an appearance would be misleading.

4.1. *Several variables omitted at once.* Suppose now that W denotes not one but several omitted variables, or that it represents a single nominal variable with 3 or more levels, so that its encoding in terms of a design matrix would require 2 or more columns, and 2 or more degrees of freedom. Results previously presented still describe potential effects of W's omission, if t_W is reinterpreted in a natural way. (The sensitivity parameter $\rho^2_{y \cdot w|z\mathbf{x}}$ retains its original interpretation, as the proportionate decline in unexplained variance from including W as a regressor.)

When Z is regressed on X and a multivariate W, there is no one W-coefficient and corresponding t-statistic. The natural analogue of such a statistic is the ANOVA F-statistic comparing regression fits with and without W, F_W; for univariate W, $F_W = t^2_W$, as is well known. When $\mathrm{rank}(W) > 1$, define the omitted variables' treatment confounding, again denoted t_W, as the positive square root of $[(k)(\mathrm{df})/(\mathrm{df} + 1 - k)]F_W$. Proposition 2.1 then gives the following.

COROLLARY 4.1. *Suppose $R^2_{y \cdot z\mathbf{x}} < 1$, t^2_W is finite, and $\mathrm{rank}(W) = k > 1$. Then $(\hat{b} - \hat{\beta})^2 \leq \hat{V}(\hat{b})[(k)(\mathrm{df})(\mathrm{df} + 1 - k)^{-1}]F_W \rho^2_{y \cdot w|z\mathbf{x}}$ or, equivalently,*

$$|\hat{b} - \hat{\beta}| \leq \mathrm{SE}(\hat{b})t_W|\rho_{y \cdot w|z\mathbf{x}}|.$$

PROOF. Without loss of generality, W is uncorrelated with Z and X: if not, replacing W with $W - \mathrm{Pj}(W|X, Z)$ leaves Z-coefficients and their standard errors unchanged. Define $\tilde{W} = \mathrm{Pj}(Y^{\perp \mathbf{x}, z}|W)$, where $Y^{\perp \mathbf{x}, z} = Y - \mathrm{Pj}(Y|X, Z)$. Again without loss of generality, $W = (\tilde{W}, W_2, \ldots, W_k)$, where $\tilde{W} \perp (W_2, \ldots, W_k)$. Writing

$$(10) \qquad \mathrm{Pj}(Y|Z, \mathbf{X}, W) =: \hat{\alpha} + \hat{\beta}Z + \hat{\gamma}\mathbf{X}^T + \hat{\delta}_1 \tilde{W} + \hat{\delta}_2 W_2 + \cdots + \hat{\delta}_k W_k,$$

it is immediate that $\hat{\delta}_2, \ldots, \hat{\delta}_k = 0$, since $W_2, \ldots, W_k$ are orthogonal to $\mathrm{Pj}(Y^{\perp\mathbf{x},z}|W)$, and hence orthogonal to $Y^{\perp\mathbf{x},z}$. Projecting (10) onto the span of $(Z, \mathbf{X})$, and then equating the Z-coefficient in what results with the Z-coefficient in (2) yields

$$(11) \qquad\qquad \hat{\beta} + \hat{\delta}_1 B_1^* = \hat{b},$$

where B_1^* is defined by $\mathrm{Pj}(\tilde{W}|Z, X) = B_1^* Z + C^* X$. In other words, $\hat{b}$ and $\hat{\beta}$ are related just as they would have been had W been of rank 1, rather than k, consisting only of $\tilde{W}$.

We record some entailments of the definitions of $\rho_{y\cdot w|z\mathbf{x}}$, t_W and F_W in a lemma, proved in the Appendix:

LEMMA 4.2.   *Suppose* $R_{y\cdot z\mathbf{x}}^2 < 1$, t_W^2 *is finite, and* $\mathrm{rank}(W) = k$. *Then*:

$$(1) \quad \rho_{y\cdot w|z\mathbf{x}}^2 = \rho_{y\cdot\tilde{w}|z\mathbf{x}}^2; \qquad (2) \quad t_{\tilde{W}}^2 \leq k \frac{\mathrm{df}}{\mathrm{df} + 1 - k} F_W.$$

The desired result now follows from (11), Proposition 2.1 and Lemma 4.2. $\square$

When $\mathrm{rank}(W) > 1$ we have the following variant of Proposition 2.2, proved in the Appendix.

PROPOSITION 4.3.   *Suppose* $R_{y\cdot z\mathbf{x}}^2 < 1$, t_W^2 *is finite, and* $\mathrm{rank}(W) = k$, $k > 1$. *Then*

$$(12) \qquad\qquad \hat{V}(\hat{\beta}) = \hat{V}(\hat{b})\left[1 + \frac{k + t_W^2}{\mathrm{df} - k}\right](1 - \rho_{y\cdot w|z\mathbf{x}}^2).$$

Because the sensitivity intervals in Proposition 3.1 follow algebraically from the bias and standard error representations (3) and (4), they are valid for W of arbitrary rank. The proofs of Proposition 3.1 and the following are essentially the same.

PROPOSITION 4.4.   *Proposition* 3.1 *continues to hold if* $\mathrm{rank}(W) = k > 1$, *provided* $C_d(t)$ *is read as* $[1 + (k + t^2)/(d - k)]^{1/2}$ *and* t_W *is read as* $\{[k(\mathrm{df})/(\mathrm{df} + 1 - k)]F_W\}^{1/2}$.

Some of the hypothetical omissions discussed in Section 2 are of the type for which Proposition 4.4 is needed. The variable "Insurance class" appearing in Table 3, for example, is a nominal variable with 6 categories, consuming 5 degrees of freedom when added as a regressor. Its treatment confounding, t_W, was calculated as the appropriately rescaled square-root of the F-statistic comparing the linear regression of Z on X and it to the regression of Z on X alone, about 2.24.

4.2. *Treatment effects differing by subgroup.*   Recall from Section 2.1 that of the 33 X-variables selected as covariates for the regression of length of stay on RHC, DNR status most reduced R^2. Patients with do-not-resuscitate orders suffered 25% greater mortality during the study than other patients, probably making it inevitable that their outcomes on this variable should systematically differ from patients without such orders. It is natural to suspect that effects of RHC might differ for them as well. In this case our linear models require interactions between RHC and DNR status—and perhaps other interactions as well, for that matter, but it suffices for us to restrict attention to the treatment interaction with a single binary moderating variable, as all issues pertaining to sensitivity analysis arise in this simplest case. Marcus (1997) explores related problems.

Supplementing the regression of length of stay on covariates and RHC with an interaction between RHC and DNR status gave quite revealing results. The additional right-hand side variable, an indicator of RHC and DNR simultaneously, bears a coefficient of -0.43 and a t-statistic of -5: it appears that the model devoting a single term to the treatment obscured significant effect heterogeneity. Correspondingly, the main RHC coefficient, interpretable as the effect for patients without DNR orders, is larger $(+0.15)$ than it was in earlier analyses without interactions $(+0.11)$; its standard error increases slightly, from 0.026 to 0.027.

To subject these results to sensitivity analysis, we again use index values to limit the impact of the hypothesized omitted covariate on R^2: $\rho^2_{y \cdot w | z\mathbf{x}} \leq 0.10$ remains a generous limit, as simultaneously adding all 33 covariates to the regression of length of stay on RHC (now interacted with DNR status) decreases unexplained variation by only slightly more, about 11%. To set suitable limits on the omitted variable's treatment confounding, imagine for the moment that the inclusion of an interaction term had been handled somewhat differently: rather than adding ZX_1, the product of RHC and DNR indicators, add in $\tilde{X} = ZX_1 - \text{Pj}(ZX_1 | X, Z)$. The coefficient of this variable lacks any straightforward interpretation, but its addition to the right-hand side of the equation has precisely the same effect on remaining coefficients as would the addition of ZX_1 itself. To benchmark treatment confounding, we would then regress Z on X and $\tilde{X}$. By construction, however, $\tilde{X}$ is orthogonal to Z and X, so that it itself earns a t-statistic of 0 in this fitting, and its inclusion has no effect other than to remove a single residual degree of freedom. In other words, the t_W-benchmarks extracted by regressing Z on X alone, used for sensitivity analysis of the RHC effect in the absence of interactions, serve just as well here after multiplication by the factor $[(\text{df} + 1)/(\text{df})]^{1/2}$ (which in this case is effectively 1). The first 2 columns of Table 4 use benchmarks gathered in this way, finding the conclusion that RHC increases lengths of time in the hospital for patients without DNR orders to be a bit less sensitive to hidden bias than was the analogous conclusion for the analysis assuming homogeneous treatment effects, in Table 3.

When we instructed it to include the RHC–DNR interaction among the explanatory variables, our software might equally well have added an indicator of RHC

TABLE 4
Sensitivity intervals for subgroup effects and for weighted average effects

	Treatment-confounding benchmarks found with estimated					
	Effect on patients w/out DNR order		Effect on patients with DNR order		ETT-weighted average effect	
Variable	t_W	$\rho^2_{y \cdot w \mid zx} \leq 0.1$	t_W	$\rho^2_{y \cdot w \mid zx} \leq 0.1$	t_W	$\rho^2_{y \cdot w \mid zx} \leq 0.1$
Insurance class	12.2	(−0.01, 0.30)	11.9	(−0.74, 0.17)	12.2	(−0.03, 0.27)
Respiratory eval.	8.9	(0.02, 0.27)	8.1	(−0.64, 0.08)	−4.1	(0.03, 0.20)
Mean blood press.	−8.6	(0.02, 0.27)	−8.2	(−0.64, 0.08)	−5.1	(0.03, 0.21)
Cardiovascular eval.	−8.5	(0.02, 0.27)	−7.7	(−0.63, 0.07)	−7.1	(0.01, 0.23)
Weight (kg)	6.1	(0.05, 0.25)	6.4	(−0.60, 0.03)	−5.3	(0.03, 0.21)
Immunosuppression	0.4	(0.09, 0.20)	0.4	(−0.44, −0.12)	−5.9	(0.02, 0.22)

and the absence of DNR, $Z(1 - X_1)$. In this case, the main effect would be interpretable as the effect of RHC for patients *with*, rather than without, DNR orders. It follows that had that effect been the object of our interest, we could construct a sensitivity analysis for it in the same manner as just above, by persuading our regression program to expand the interaction differently. In actuality, things are still simpler than that; we really only need to take ordinary care in interpreting the regression results, and somewhat modify the benchmarking equation. The effect for patients with DNR orders is the sum of the main RHC effect and the RHC-and-DNR effects, $0.148 + (-0.430) = -0.28$, with estimated variance equal to the sum of the two estimated variances and twice their covariance, $0.0065 = (0.080)^2$. In parallel, to benchmark treatment confounding for this analysis, regress the sum of the main RHC indicator and the interaction term, the product of RHC and DNR indicators, on covariates. This gives somewhat different results than did the benchmarking for the RHC effect on patients without DNR orders, which omitted the interaction term from the left-hand side of its regression equation; compare the middle and left columns of Table 4.

The same approach yields a sensitivity analysis for any target parameter representable as a linear combination of main effect and interaction terms. Take the effect of treatment on the treated, or ETT, parameter, considered from within a model permitting treatment effects to vary within specified subgroups. If the groups are, for simplicity, patients with and without DNR orders, then since 7% of patients receiving RHC had DNR orders, the ETT can be represented as the main effect plus 0.07 times the DNR–RHC interaction effect, estimated as $0.148 + (0.07) - 0.430 = 0.118$, with corresponding standard error 0.026. For benchmarking, regress on covariates the RHC indicator plus 0.07 times the product of RHC and DNR indicators, with results as given in the rightmost two columns of Table 4.

4.3. *Propensity-adjusted estimates of treatment effects*. Regression adjusts between-group comparisons by attempting to remove adjustment variables' contributions from the outcomes before comparing them. In contrast, adjustments based on propensity scores attempt to divide the sample into strata within which treatment and control subjects have similar distributions of covariates. We estimated propensity scores using all 50 of the SUPPORT data's covariates in a logistic regression [Rosenbaum and Rubin (1984)], finding six equally-sized subclasses made treatment-control differences on the covariates jointly insignificant at level $\alpha = 0.10$ [Hansen and Bowers (2008)]. One can couple such a stratification with linear modeling to estimate treatment effects. In the simplest variant, responses are regressed on the treatment and fixed stratum effects. Fitting such a model to the SUPPORT data gives an RHC effect similar to what was estimated after ordinary covariance adjustment, as in Section 1.4, but with somewhat larger standard errors.

The main assumption for this model is that so far as the outcome, length of stay, is concerned, the only systematic difference between the RHC and non-RHC patient in a propensity stratum is RHC itself. Relax this assumption in favor of another to the effect that so far as differences between outcomes and their projections onto an omitted variable are concerned, within strata RHC and non-RHC patients do not systematically differ. Were the omitted variable to become available, we could adjust by adding it to the explanatory side of the linear model. Without it, we can do sensitivity analysis.

Benchmarking treatment-confounding levels takes a bit more effort than before: rather than simply regressing Z on covariates and recording their t- or F-statistics, we have to account in some way for the propensity stratification. We do this by removing the covariates, one at a time, from the propensity model, after each removal subclassifying the sample into sextiles, as before, but now using the modified propensity score; then regressing Z on the withheld covariate and on the propensity strata in order to associate a t- or F-statistic with that covariate. Results of this process appear in Table 5.

The results exhibit a striking pattern: adjustment based on propensity scores gives causal inferences that are far less sensitive to omitted variables than does regression-based covariate adjustment. With it, one can expect less residual confounding with the treatment than with covariate adjustment, as seen in smaller t_W-values on the propensity-score side of the table. In a sense, propensity scores focus on confounding with the treatment, whereas covariate adjustment focuses on covariates and the response. Recall from Section 2 that while both matter to omitted variable bias, confounding with the treatment is both more difficult to pin down and potentially more pernicious. It stands to reason that while propensity adjustment may pay a slight penalty up front, in terms of somewhat larger standard errors than covariance adjustment, it offers a greater return downstream, in reduced sensitivity to hidden bias.

TABLE 5

Sensitivity intervals for the treatment effect after ordinary covariance and propensity-score adjustment, illustrating that propensity adjustment better limits sensitivity to the omission of adjustment variables. For covariance adjustment, t_W is limited by the confoundedness with the treatment of 6 variables that had been eliminated by a preliminary variable-selection procedure, as in Table 3; for propensity adjustment, limits on treatment confounding are set by separately removing each of these and calculating their t_W's after propensity adjustment for remaining variables

	OLS regression			**Propensity adjusted regression**		
	$\|t_W\|$	$\rho^2_{y \cdot w\|zx} \le 0.01$	$\rho^2_{y \cdot w\|zx} \le 0.1$	$\|t_W\|$	$\rho^2_{y \cdot w\|zx} \le 0.01$	$\rho^2_{y \cdot w\|zx} \le 0.1$
Insurance class	12.2	(0.03, 0.20)	(−0.04, 0.26)	8.6	(0.02, 0.18)	(−0.03, 0.23)
Respiratory eval.	8.9	(0.04, 0.19)	(−0.01, 0.23)	3.1	(0.04, 0.17)	(0.02, 0.19)
Mean blood press.	8.6	(0.04, 0.19)	(−0.01, 0.23)	6.8	(0.03, 0.18)	(−0.01, 0.22)
Cardiovascular eval.	8.5	(0.04, 0.19)	(−0.01, 0.23)	5.4	(0.03, 0.17)	(0, 0.20)
Weight (kg)	6.1	(0.04, 0.18)	(0.01, 0.21)	5.1	(0.03, 0.18)	(0.01, 0.21)
Immunosuppression	0.4	(0.06, 0.16)	(0.06, 0.16)	0.5	(0.04, 0.16)	(0.04, 0.16)

5. Summary For effect estimates adjusted for covariates using ordinary least squares, impacts of covariate omission on point estimates and on standard errors have been represented in terms of two statistics relating the omitted variable to included ones, a measure of how adding the variable would affect R^2 and a measure of its association with the treatment variable given included variables. We refer to the latter as the omitted variable's treatment-confounding measurement. When generous limits on how the omitted variable would affect R^2 can be defended, they yield far less pessimistic assessments of sensitivity than would be possible without such a limit. Unlike the sensitivity "parameter" pertaining to R^2, plausible limits on the treatment-confounding parameter are unlikely to emerge from intuition alone; on the other hand, it is straightforward and informative to determine study-specific benchmarks for it using available data.

The changes to the treatment coefficient's point and error estimates that the addition of an omitted covariate would cause have been represented as multiples of its standard error. So these representations yield error appraisals accounting for certain hidden biases in familiar terms, as a multiple of the SE. The method adapts readily to scenarios of multivariate omission, heterogeneous treatment effect and combinations of regression with propensity scores.

APPENDIX: PROOFS

PROOF OF LEMMA 4.2. Under the conditions of the lemma, (1) and (2) can be established: (1) In a regression of Z and $\mathbf{X}$ on Y, adding $\tilde{W}$ has the same effect on the Z-coefficient and model R^2 as adding W. Thus, (1) holds.

(2) Furthermore, $\tilde{W} \in \text{span}(W)$, so $\tilde{W}$ explains no more variation in Z than does W. Thus, $R^2_{z \cdot w|x} \geq R^2_{z \cdot \tilde{w}|x}$, which implies $\frac{\rho^2_{z \cdot w|x}}{1 - \rho^2_{z \cdot w|x}} \geq \frac{\rho^2_{z \cdot \tilde{w}|x}}{1 - \rho^2_{z \cdot \tilde{w}|x}}$. By definition of the ANOVA F-statistic, $F_W = [(\sigma^2_{z \cdot \mathbf{x}} - \sigma^2_{z \cdot \mathbf{x}w})/(k)]/[\sigma^2_{z \cdot \mathbf{x}w}/(\text{df} + 1 - k)]$, or

$$(13) \qquad F_W = [(\text{df} + 1 - k)/k][\rho^2_{z \cdot w|\mathbf{x}}/(1 - \rho^2_{z \cdot w|\mathbf{x}})].$$

As $\text{rank}(\tilde{W}) = 1$, $F_{\tilde{W}} = t^2_{\tilde{W}} = \text{df}\frac{\rho^2_{z \cdot \tilde{w}|x}}{1 - \rho^2_{z \cdot \tilde{w}|x}}$. The result follows. $\square$

PROOF OF PROPOSITION 4.3. To relate $\text{SE}(\hat{b})$ and $\text{SE}(\hat{\beta})$, begin by rewriting some of the variance terms: $\sigma^2_{y \cdot z\mathbf{x}w} = \text{Var}(Y^{\perp z\mathbf{x}w}) = \text{Var}(Y^{\perp z\mathbf{x}} - \text{Pj}(Y^{\perp z\mathbf{x}}|W^{\perp z\mathbf{x}})) = \sigma^2_{y \cdot z\mathbf{x}} - \sigma^2_{y \cdot z\mathbf{x}}\rho^2_{yw \cdot z\mathbf{x}} = \sigma^2_{y \cdot z\mathbf{x}}(1 - \rho^2_{yw \cdot z\mathbf{x}})$. Similarly, $\sigma^2_{z \cdot \mathbf{x}w} = \sigma^2_{z \cdot \mathbf{x}}(1 - \rho^2_{z \cdot w|\mathbf{x}})$.

Thus, by (8), $\text{SE}(\hat{\beta}) = (\text{df} - k)^{-1/2}(\sigma_{y \cdot z\mathbf{x}}/\sigma_{z \cdot \mathbf{x}})[(1 - \rho^2_{yw \cdot z\mathbf{x}})/(1 - \rho^2_{zw \cdot \mathbf{x}})]^{1/2}$, or, invoking (8) again, $\text{SE}(\hat{b})[\text{df}/(\text{df} - k)]^{1/2}[(1 - \rho^2_{yw \cdot z\mathbf{x}})/(1 - \rho^2_{zw \cdot \mathbf{x}})]^{1/2}$.

By (13), $(1 - \rho^2_{z \cdot w|\mathbf{x}})^{-1} = 1 + k(\text{df} + 1 - k)^{-1}F_W$. Recall that t_W was defined for multivariate W in Section 4.1, as a rescaling of $F_W^{1/2}$. Applying that definition, $(1 - \rho^2_{z \cdot w|\mathbf{x}})^{-1} = (\text{df} + t_W^2)(\text{df}^{-1})$. The relationship (12) follows. $\square$

PROOF OF PROPOSITION 3.1. Equation (5) comes directly from Propositions 2.1 and 2.2. For (6), write $\hat{\beta} - q\,\text{SE}(\hat{\beta}) = \hat{b} + l_{t_W}(\arcsin \rho_{y \cdot w|z\mathbf{x}})\,\text{SE}(\hat{b})$ and $\hat{\beta} + q\,\text{SE}(\hat{\beta}) = \hat{b} + u_{t_W}(\arcsin \rho_{y \cdot w|z\mathbf{x}})\,\text{SE}(\hat{b})$, where $\arcsin \rho_{y \cdot w|z\mathbf{x}} \in (-\pi/2, \pi/2)$ and $l_t(\theta) := -t\sin\theta - qC_{\text{df}}(t)\cos\theta$ and $u_t(\theta) := -t\sin\theta + qC_{\text{df}}(t)\cos\theta$. To maximize $u_t(\arcsin\rho)$ as t ranges over $[-T, T]$ and ρ ranges over $[-1, 1]$, maximize $t \mapsto \sup\{u_t(\arcsin\rho) : -1 \leq \rho \leq 1\}$. We need only consider $t < 0$. For such a t, calculus shows that $u_t(\cdot)$ is concave unimodal on $(-\pi/2, \pi/2)$, attaining its maximum at $\arctan\{-t/[qC_{\text{df}}(t)]\}$; with some algebra and trigonometry, $\sup\{u_t(\arcsin\rho) : -1 \leq \rho \leq 1\}$ is seen to be $(t^2 + q^2C_{\text{df}}(t)^2)^{1/2}$. Consequently, $\sup\{u_t(\arcsin\rho) : |t| \leq T, |\rho| \leq 1\} = (T^2 + q^2C_{\text{df}}(T)^2)^{1/2}$. Similarly, as t and θ range over $[-T, T]$ and $(-\pi/2, \pi/2)$, $l_t(\theta)$ takes its minimum value, $-(T^2 + q^2C_{\text{df}}(T)^2)^{1/2}$, at $(T, \arctan\{T/[qC_{\text{df}}(T)]\})$. Part (i) follows.

Restating slightly an intermediate conclusion, the maximizer of $\rho \mapsto u_t(\arcsin\rho)$ over the domain $[-1, 1]$ is $-t/[t^2 + q^2C_{\text{df}}(t)^2]^{1/2}$. Under the condition of (iii), for each $t \in [-T, T]$ this falls within the narrower domain $[-R^{1/2}, R^{1/2}]$. (iii) follows.

In (ii), $\rho^2_{y \cdot w|z\mathbf{x}} \leq R$. To maximize $u_t(\arcsin\rho)$ over a domain that is symmetric in t, we again need only consider negative t. For t small enough in magnitude that $t^2/[t^2 + q^2C_{\text{df}}(t)^2] \leq R$, the maximizer of $\rho \mapsto u_t(\arcsin\rho)$ over the domain $[-1, 1]$ falls inside the narrower domain $[-R^{1/2}, R^{1/2}]$, and $\sup_{-R^{1/2} \leq \rho \leq R^{1/2}} u_t(\arcsin\rho) = \sup_{-1 \leq \rho \leq 1} u_t(\arcsin\rho) = (t^2 + q^2C_{\text{df}}(t)^2)^{1/2}$.

This function is increasing as a function of $-t$. For t such that $R \leq t^2/[t^2 + q^2 C_{df}(t)^2]$, because $u_t(\cdot)$ is concave unimodal with maximum at a point, $\arctan\{-t/[qC_{df}(t)]\}$, that falls outside of $\{\arcsin \rho : |\rho| \leq R^{1/2}\}$, $\sup\{u_t(\arcsin \rho) : -R^{1/2} \leq \rho \leq R^{1/2}\} = u_t(\arcsin R^{1/2}) = -t R^{1/2} + q C_{df}(t)(1 - R)^{1/2}$. This also is increasing as a function of $-t$. For the unique $t < 0$ such that $R = t^2/[t^2 + q^2 C_{df}(t)^2]$, $\sup\{u_t(\arcsin \rho) : -R^{1/2} \leq \rho \leq R^{1/2}\}$ is given by either of the two functions of $-t$, which shows that they coincide at that point. In consequence, $\sup\{u_t(\arcsin \rho) : -R^{1/2} \leq \rho \leq R^{1/2}\}$ is increasing as a function of $-t$ for all $t < 0$, so that the maximum of $u_t(\arcsin \rho)$ for $|t| \leq T$ and $\rho^2 \leq R$ is $\sup\{-t R^{1/2} + q C_{df}(t)(1 - R)^{1/2} : |t| \leq T\} = T R^{1/2} + q C_{df}(T)(1 - R)^{1/2}$, as required for (ii). [Similar steps yield the minimum of $l_t(\arcsin \rho)$.] □

SUPPLEMENTARY MATERIAL

Code for computations discussed in the article (DOI: 10.1214/09-AOAS315SUPP; .zip). Zip archive containing our documented R code and instructions for obtaining Connors et al.'s data, in the form of a Sweave and a corresponding PDF file.

REFERENCES

ANGRIST, J. and IMBENS, G. (2002). Comment. *Statist. Sci.* **17** 304–307. [MR1962487](MR1962487)

CHRISTENSEN, R. (1996). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer, Berlin. [MR1402692](MR1402692)

COCHRAN, W. G. (1938). The omission or addition of an independent variate in multiple linear regression. *J. Roy. Statist. Soc.* **5** (suppl.) 171–176.

CONNORS, A. J., SPEROFF, T., DAWSON, N., THOMAS, C., HARRELL, F. E. J., WAGNER, D., DESBIENS, N., GOLDMAN, L., WU, A., CALIFF, R., FULKERSON, W. J., VIDAILLET, H., BROSTE, S., BELLAMY, P., LYNN, J. and KNAUS, W. (1996). The effectiveness of right heart catheterization in the initial care of critically ill patients. Support investigators. *J. Amer. Med. Assoc.* **276** 889–897.

COPAS, J. B. and LI, H. G. (1997). Inference for non-random samples (with discussion and a reply by the authors). *J. Roy. Statist. Soc. Ser. B* **59** 55–95. [MR1436555](MR1436555)

CORNFIELD, J., HAENSZEL, W., HAMMOND, E., LILIENFELD, A., SHIMKIN, M. and WYDNER, E. (1959). Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute* **22** 173–203.

FARAWAY, J. J. (1992). On the cost of data analysis. *J. Comput. Graph. Statist.* **1** 213–229.

FRANK, K. A. (2000). Impact of a confounding variable on a regression coefficient. *Sociol. Methods Res.* **29** 147–194.

GASTWIRTH, J., KRIEGER, A. and ROSENBAUM, P. (1998). Dual and simultaneous sensitivity analysis for matched pairs. *Biometrika* **85** 907–920.

GORE, J., GOLDBERG, R., SPODICK, D., ALPERT, J. and DALEN, J. (1987). A community-wide assessment of the use of pulmonary artery catheters in patients with acute myocardial infarction. *Chest* **92** 721–727.

HANSEN, B. B. and BOWERS, J. (2008). Covariate balance in simple, stratified and clustered comparative studies. *Statist. Sci.* **23** 219–236. MR2516821

HARVEY, S., HARRISON, D. et al. (2005). Assessment of the clinical effectiveness of pulmonary artery catheters in management of patients in intensive care (pac-man): A randomised controlled trial. *Lancet* **366** 472–477.

HOLLAND, P. (1988). Causal inference, path analysis, and recursive structural equations models. *Soc. Methodol.* **18** 449–484.

HOSMAN, C., HANSEN, B. B. and HOLLAND, P. W. (2010). Supplement to "The sensitivity of linear regression coefficients confidence limits to the omission of a confounder." DOI: 10.1214/09-AOAS315SUPP.

IMBENS, G. W. (2003). Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review* **93** 126–132.

LIN, D., PSATY, B. and KRONMAL, R. (1998). Assessing the sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics* **54** 948–963.

MARCUS, S. M. (1997). Using omitted variable bias to assess uncertainty in the estimation of an AIDS education treatment effect. *J. Educ. Behav. Stat.* **22** 193–201.

RHODES, A., CUSACK, R., NEWMAN, M., GROUNDS, R. and BENNET, E. (2002). A randomized, controlled trial of the pulmonary artery catheter in critically ill patients. *Intensive Care Medicine* **348** 5–14.

RICHARD, C. et al. (2003). Early use of the pulmonary artery catheter and outcomes in patients with shock and acute respiratory distress syndrome: a randomized controlled trial. *J. Amer. Med. Assoc.* **290** 2732–2734.

ROBINS, J. M., ROTNITZKY, A. and SCHARFSTEIN, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical Models in Epidemiology, the Environment and Clinical Trials* (M. Halloran and D. Berry, eds.) 1–94. Springer, New York. MR1731681

ROSENBAUM, P. R. (1988). Sensitivity analysis for matching with multiple controls. *Biometrika* **75** 577–581. MR0967598

ROSENBAUM, P. R. (2002). *Observational Studies*. Springer, New York. MR1899138

ROSENBAUM, P. R. and RUBIN, D. B. (1983). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *J. Roy. Statist. Soc. Ser. B* **45** 212–218.

ROSENBAUM, P. R. and RUBIN, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *J. Amer. Statist. Assoc.* **79** 516–524.

ROSENBAUM, P. R. and SILBER, J. H. (2009). Amplification of sensitivity analysis in matched observational studies. *J. Amer. Statist. Assoc.* **104** 1398–1405.

RUBIN, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *Ann. Statist.* **6** 34–58. MR0472152

RUBIN, D. B. (2006). Causal inference through potential outcomes and principal stratification: Application to studies with "censoring" due to death. *Statist. Sci.* **21** 299–309. MR2339125

SANDHAN, J. D., HULL, R. D., BRANT, R. F., KNOX, L., PINEO, G. F., DOIG, C. J., LAPORTA, D. P., VINER, S., PASSERINI, L., DEVITT, H., KIRBY, A., JACKA, M. and THE CANADIAN CRITICAL CARE CLINICAL TRIALS GROUP (2003). A randomized, controlled trial of the use of pulmonary-artery catheters in high-risk surgical patients. *New England Journal of Medicine* **348** 5–14.

SCHARFSTEIN, D. O. and IRIZARRY, R. A. (2003). Generalized additive selection models for the analysis of studies with potentially nonignorable missing outcome data. *Biometrics* **59** 601–613. MR2004265

SEBER, G. (1977). *Linear Regression Analysis*. Wiley, New York. MR0436482

SHAH, M. and STEVENSON, L. (2004). Evaluation study of congestive heart failure and pulmonary artery catheterization effectiveness: The escape trial. In *American Heart Association Scientific Sessions*, New Orleans.

SMALL, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *J. Amer. Statist. Assoc.* **102** 1049–1058. MR2411664

ZION, M., BALKIN, J., ROSENMANN, D., GOLDBOURT, U., REICHER-REISS, H., KAPLINSKY, E. and BEHAR, S. (1990). Use of pulmonary artery catheters in patients with acute myocaridal infarction. *Chest* **98** 1331–1335.

C. A. HOSMAN
B. B. HANSEN
DEPARTMENT OF STATISTICS
UNIVERSITY OF MICHIGAN
ANN ARBOR, MICHIGAN 48109-1107
USA
E-MAIL: chosman@umich.edu
        ben.b.hansen@umich.edu

P. W. HOLLAND
PAUL HOLLAND CONSULTING CORPORATION
703 SAYRE DRIVE
PRINCETON, NEW JERSEY 08540
USA
E-MAIL: roberta.holland@att.net