ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2017)

Discovering effect modification in an observational study of surgical mortality at hospitals with superior nursing

Kwonsang Lee, Dylan S. Small, Jesse Y. Hsu, Jeffrey H. Silber and Paul R. Rosenbaum

University of Pennsylvania, Philadelphia, USA

[Received March 2016. Final revision April 2017]

Summary. There is effect modification if the magnitude or stability of a treatment effect varies systematically with the level of an observed covariate. A larger or more stable treatment effect is typically less sensitive to bias from unmeasured covariates, so it is important to recognize effect modification when it is present. We illustrate a recent proposal for conducting a sensitivity analysis that empirically discovers effect modification by exploratory methods but controls the familywise error rate in discovered groups. The example concerns a study of mortality and use of intensive care units in 23715 matched pairs of two Medicare patients, one of whom underwent surgery at a hospital that had been identified for superior nursing; the other at a conventional hospital. The pairs were matched exactly for 130 four-digit ninth international classification of diseases surgical procedure codes and balanced 172 observed covariates. The pairs were then split into five groups of pairs by the classification and regression trees method in its effort to locate effect modification. The evidence of a beneficial effect of magnet hospitals on mortality is least sensitive to unmeasured biases in a large group of patients undergoing quite serious surgical procedures, but in the absence of other life-threatening conditions, such as a comorbidity of congestive heart failure or an emergency admission leading to surgery.

Keywords: Causal inference; Classification and regression trees; Effect modification; Multivariate matching; Sensitivity analysis; Truncated product of *P*-values

1. Superior nurse staffing, surgical mortality and resource utilization in Medicare

Hospitals vary in the extent and quality of their staffing, technical capabilities and nursing work environments. Does superiority in these areas confer benefits to patients undergoing forms of general surgery that might be performed at most hospitals? To what extent and in what way do these factors affect the cost of surgical care? Are they a life-saving benefit or a pointless and unneeded expense in the case of relatively routine forms of surgery? How does a patient's choice of hospital affect that patient's outcomes and medical resource utilization?

The current paper extends a recent study by Silber *et al*. (2016) that examined health outcomes and resource utilization in general surgery at 35 'magnet' hospitals with superior nursing defined by the two attributes

(a) a nurse-to-bed ratio of 1 or more and
(b) accreditation as a 'nursing magnet hospital',

when compared with 293 control hospitals lacking both attributes. In their matched compari-

Address for correspondence: Kwonsang Lee, Department of Statistics, Wharton School, University of Pennsylvania, Philadelphia, PA 19104-6340, USA.
E-mail: kwonlee@wharton.upenn.edu

 0964–1998/17/181000

son, Silber *et al*. (2016) found significantly lower mortality at magnet hospitals than at control hospitals (4.8% *versus* 5.8%; McNemar P-value less than 0.001), substantially lower use of an expensive resource, the intensive care unit (ICU) (32.9% *versus* 42.9%), and slightly shorter length of stay, and the differences appeared to be larger for patients who were at greater risk of death. Silber *et al*. (2016) did not characterize this effect modification in terms of clinical attributes that might guide particular patients to particular hospitals. In contrast, the current paper uses a recently proposed exploratory technique to unpack effect modification, combined with a confirmatory technique that examines the sensitivity of these conclusions to unmeasured biases. Is the ostensible effect of treatment at a magnet hospital larger, more stable or more insensitive to unmeasured bias for certain categories of patients?

For discussion of the nursing magnet designation, see Aiken *et al*. (2000).

2. Review of effect modification in observational studies

2.1. Notation for causal effects, non-random treatment assignment and sensitivity analysis

In observational studies, it is known that certain patterns of treatment effects are more resistant than others to being explained away as the consequence of unmeasured biases in treatment assignment; see, for instance, Rosenbaum (2004), Zubizarreta *et al*. (2013) and Stuart and Hanna (2013).

Effect modification occurs when the size of a treatment effect or its stability varies with the level of a pretreatment covariate: the effect modifier. Effect modification affects the sensitivity of ostensible treatment effects to unmeasured biases. Other things being equal, larger or more stable treatment effects are insensitive to larger unmeasured biases; see Rosenbaum (2004, 2005). As a consequence, discovering effect modification when it is present is an important aspect of appraising the evidence that distinguishes treatment effects from potential unmeasured biases, which is a concern in every observational study. In particular, Hsu *et al*. (2013, 2015) discussed sensitivity analysis in observational studies with potential effect modification, and Section 2.2 is a concise summary. Chesher (1984), Crump *et al*. (2008), Lehrer *et al*. (2016), Lu and White (2015), Wager and Athey (2017), Athey and Imbens (2016) and Ding *et al*. (2016) have discussed effect modification from different perspectives, placing less emphasis on its role in confirmatory analyses that distinguish treatment effects from unmeasured biases in observational studies.

There are I matched pairs, $i=1,\ldots,I$, of two subjects, $j=1,2$, one treated with $Z_{ij}=1$, the other control with $Z_{ij}=0$, so $Z_{i1}+Z_{i2}=1$ for each i. Subjects are matched for an observed covariate $\mathbf{x}_{ij}$, so $\mathbf{x}_{i1}=\mathbf{x}_{i2}=\mathbf{x}_i$, say, for each i, but they may differ in terms of a covariate u_{ij} that was not measured. Each subject has two potential responses, r_{Tij} if treated, r_{Cij} if control, exhibiting response $R_{ij}=Z_{ij}r_{Tij}+(1-Z_{ij})r_{Cij}$, so the effect that is caused by the treatment, $r_{Tij}-r_{Cij}$ is not seen from any subject; see Neyman (1990) and Rubin (1974). Fisher's (1935) null hypothesis H_0 of no treatment effect asserts that $r_{Tij}=r_{Cij}$ for all i and j. Simple algebra shows that the treated-minus-control pair difference in observed responses is $Y_i=(Z_{i1}-Z_{i2})(R_{i1}-R_{i2})$ which equals $(Z_{i1}-Z_{i2})(r_{Ci1}-r_{Ci2})=\pm(r_{Ci1}-r_{Ci2})$ if Fisher's hypothesis H_0 is true. Write $\mathcal{F}=\{(r_{Tij},r_{Cij},\mathbf{x}_{ij},u_{ij}),i=1,\ldots,I,j=1,2\}$ for the potential responses and covariates, and write $\mathcal{Z}$ for the event that $Z_{i1}+Z_{i2}=1$ for each i.

In a randomized experiment, $Z_{i1}=1-Z_{i2}$ is determined by I independent flips of a fair coin, so $\pi_i=\Pr(Z_{i1}=1|\mathcal{F},\mathcal{Z})=\frac{1}{2}$ for each i, and this becomes the basis for randomization inferences, for instance for tests of Fisher's null hypothesis or for confidence intervals or point estimates formed by inverting hypothesis tests. A randomization inference derives the null distribution given $(\mathcal{F},\mathcal{Z})$ of a test statistic as its permutation distribution by using the fact that the 2^I possible

values of $\mathbf{Z}=(Z_{i1}, Z_{i2}, \ldots, Z_{I2})$ each have probability 2^{-I} in a randomized paired experiment; see Fisher (1935), Lehmann and Romano (2005), chapter 5, or Rosenbaum (2002a), chapter 2. A simple model for sensitivity analysis in observational studies says that treatment assignments in distinct pairs are independent but bias due to non-random treatment assignment may result in π_i that deviate from $\frac{1}{2}$ to the extent that $1/(1+\Gamma) \leqslant \pi_i \leqslant \Gamma/(1+\Gamma)$ for $\Gamma \geqslant 1$, and the range of possible inferences is reported for various values of Γ to display the magnitude of bias that would need to be present to alter the study's conclusion materially; see, for instance, Rosenbaum (2002a), section 4.3.2, and Rosenbaum (2002b) for the case of matched binary responses, as in the current paper. For instance, a sensitivity analysis may report the range of possible *P*-values or point estimates that are consistent with the data and a bias of at most Γ for several values of Γ.

For various approaches to sensitivity analysis in observational studies, see Cornfield *et al.* (1959), Gastwirth (1992), Gilbert *et al.* (2003), Egleston *et al.* (2009), Hosman *et al.* (2010) and Liu *et al.* (2013). For some discussion of software in R, see Rosenbaum (2015) and Rosenbaum and Small (2017).

2.2. Three strategies examining effect modification

There is effect modification if the magnitude of the effect, $r_{Tij} - r_{Cij}$, varies systematically with $\mathbf{x}_i$. We partition the space of values of $\mathbf{x}_i$ into subsets and are concerned with effects that differ in magnitude or stability between subsets. Let $\mathcal{G}$ be a subset of the values of $\mathbf{x}$, and define the null hypothesis $H_{\mathcal{G}}$ to be Fisher's null hypothesis for individual j in set i with $\mathbf{x}_{ij} \in \mathcal{G}$, so $H_{\mathcal{G}}$ asserts that $r_{Tij} = r_{Cij}$ for all ij with $\mathbf{x}_{ij} = \mathbf{x}_i \in \mathcal{G}$. Let $\mathfrak{g}=(\mathcal{G}_1, \ldots, \mathcal{G}_G)$ be a mutually exclusive and exhaustive partition of values of $\mathbf{x}_{ij} = \mathbf{x}_i$, so each pair i has an $\mathbf{x}_i$ contained in exactly one $\mathcal{G}_g$. A simple form of effect modification occurs if $H_{\mathcal{G}_g}$ is true for some g but not for other g. Write I_g for the number of pairs with $\mathbf{x}_i \in \mathcal{G}_g$, so $I = \sum_{g=1}^{G} I_g$.

There are three strategies for defining the groups $\mathfrak{g}=(\mathcal{G}_1, \ldots, \mathcal{G}_G)$, two of which are practically useful but technically straightforward, the third having interesting technical aspects that we illustrate by using the Medicare example. One useful strategy defines the groups, $\mathfrak{g}=(\mathcal{G}_1, \ldots, \mathcal{G}_G)$, *a priori*, without reference to data. For example, on the basis of clinical judgement, one might believe that certain surgical procedures are more challenging or hazardous than others and therefore divide the exactly matched procedures into a few groups based on clinical judgement alone. Alternatively, clinical judgement might separate patients with severe chronic conditions that are unrelated to the current surgery, such as congestive heart failure (CHF).

A second strategy uses an external source of data to define the groups. In particular, Silber *et al.* (2016) fitted a logit model to an external data source, predicting mortality from covariates, $\mathbf{x}_{ij}$, then formed five groups $\mathfrak{g}=(\mathcal{G}_1, \ldots, \mathcal{G}_5)$ based on this predicted risk for a given $\mathbf{x}$. This approach made no use of the mortality experience of the patients in the current study in defining the groups. A variant of the second strategy is to split one data set at random into two parts, to create the groups by using the first part, and then to analyse only the second part with these, again, externally determined groups.

In both of the first two strategies, the groups $\mathfrak{g}=(\mathcal{G}_1, \ldots, \mathcal{G}_G)$ were determined by events that were external to the outcomes' reported study. The second strategy makes explicit use of an external source of data, whereas the first strategy uses judgement that is presumably informed historically by various external sources of data. The key element in both strategies is that the groups were fixed before examining outcomes in the current study, and in that sense are unremarkable as groups, requiring no special handling because of their origin. With *a priori* groups, we could use any of a variety of methods to test the G hypotheses $H_{\mathcal{G}_g}$ in such a way as to provide strong control of the familywise error rate at α, meaning that the chance of falsely rejecting at least one true $H_{\mathcal{G}_g}$ is at most α no matter which hypotheses are true and which are false.

The third strategy that we illustrate here creates the groups $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$ by exploratory techniques using all of the current data and then goes on to perform an analysis of the same data as if the groups had been determined *a priori*. The third strategy is designed so that it controls the familywise error rate in a sensitivity analysis despite the data-dependent generation of G particular groups from among the infinitely many ways of splitting the space of values of the observed covariates $\mathbf{x}$. This strategy is discussed in detail in Hsu *et al.* (2015) and it entails certain restrictions on the way that the groups are constructed.

A simple version of the strategy regresses $|Y_i|=|(Z_{i1}-Z_{i2})(R_{i1}-R_{i2})|$ on $\mathbf{x}_i$ by using a form of regression that yields groups, such as the classification and regression trees (CART) method. For discussion of this method, see Breiman *et al.* (1984) and Zhang and Singer (2010). Note that the unsigned $|Y_i|$ not the signed Y_i are used, i.e. the regression does not know who is treated and who is control. The leaves of a CART tree become the groups $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$. The signs of the Y_i are then 'remembered', in an analysis that views the groups $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$ as fixed, so it resembles analyses that would have been appropriate with an *a priori* grouping of the type that is created by the first two strategies.

It is important to understand what is at issue in the third strategy; see Hsu *et al.* (2015) for a precise and general technical discussion. Briefly, if obscurely, the groups $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$ and hence the hypotheses $H_{\mathcal{G}_g}$ are not stable. If the observed data had been slightly different, the CART tree would have been different, and we would be testing different hypotheses. What does it mean to speak about the probability of falsely rejecting $H_{\mathcal{G}_g}$ if most data sets would not lead us to test $H_{\mathcal{G}_g}$?

Consider the simplest case: a paired randomized experiment. If Fisher's null hypothesis of no effect of any kind were true, then $Y_i=(Z_{i1}-Z_{i2})(r_{Ci1}-r_{Ci2})=\pm(r_{Ci1}-r_{Ci2})$ and, given $(\mathcal{F},\mathcal{Z})$, different random assignments Z_{ij} always yield $|Y_i|=|r_{Ci1}-r_{Ci2}|$, so all 2^I random assignments produce the same CART tree and the same $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$. In other words, under hypothesis H_0, the CART tree, and hence $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$, is a function of $(\mathcal{F},\mathcal{Z})$ and not of $\mathbf{Z}$. Therefore, under H_0, the 2^{I_g} possible treatment assignments for the I_g pairs with $\mathbf{x}_i\in\mathcal{G}_g$ each have probability 2^{-I_g}, resulting in conventional permutation tests within each of the G groups: tests that are conditionally independent given $(\mathcal{F},\mathcal{Z})$ under H_0. The problem occurs because we are interested in testing not just H_0, but also individual $H_{\mathcal{G}_g}$ when H_0 is false because some individuals are affected by the treatment. If H_0 is false, different random assignments $\mathbf{Z}$ yield different $|Y_i|$, and hence different CART trees and different hypotheses $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$. With a little care, it is possible to demonstrate two useful facts. First, if $r_{Tij}-r_{Cij}=0$ for all ij with $\mathbf{x}_i\in\mathcal{G}_g$, then the conditional distribution given $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$ and $(\mathcal{F},\mathcal{Z})$ of the corresponding Z_{ij} with $\mathbf{x}_i\in\mathcal{G}_g$ is its usual randomization distribution. In that sense, the instability of the tree over repeated randomizations has not distorted this conditional distribution of treatment assignments in groups with no treatment effect. Second, if a method is applied to test the $H_{\mathcal{G}_g}$ that would strongly control the familywise error rate at α with *a priori* fixed groups, then, conditionally given $\mathfrak{g}=(\mathcal{G}_1,\ldots,\mathcal{G}_G)$ and $(\mathcal{F},\mathcal{Z})$, the method will reject at least one null group with probability at most α. These two facts are extended to include sensitivity analyses in observational studies and are proved as propositions 1 and 2 of Hsu *et al.* (2015), who also presented some reasons to hope that subsets of $\mathbf{x}_i$ that systematically predict $|Y_i|$ may identify groups in which either the magnitude of $r_{Tij}-r_{Cij}$ or its stability varies with $\mathbf{x}_i$.

In the current paper, we present a practical application of this third strategy.

3. Discovering and using effect modification in the magnet hospital study

3.1. Forming groups of pairs for consideration as possible effect modifiers

The analyses here first broke and then re-paired the pairs in Silber *et al.* (2016) so that

(a) as in Silber *et al*. (2016), every pair was exactly matched for the 130 four-digit ninth international classification of diseases surgical procedure codes,
(b) the maximum number of pairs were exactly matched for an indicator of age greater than 75, CHF, emergency admission or not and chronic obstructive pulmonary disease (COPD). Because identically the same people were paired differently, the balancing properties of the new pairs are exactly the same as reported by Silber *et al*. (2016), Table 2, because balancing properties refer to marginal distributions of covariates and do not depend on who is paired with whom.

Using `rpart` in R, the CART tree was built by using the 22 622 pairs that were exactly matched in the sense that was described in the previous paragraph, regressing $|Y_i|$ on $\mathbf{x}_i$, where Y_i records the difference in binary indicators of mortality. So, the tree is essentially trying to locate pairs that are discordant for mortality, $|Y_i| = 1$, on the basis of exactly matched covariates. Here, a pair is discordant if exactly one patient in the pair died within 30 days. The CART algorithm was not offered all 130 exactly matched surgical procedure codes, but rather 26 mutually exclusive clusters of the 130 surgical procedures, as listed in Table 1, plus the binary covariates age>75, CHF, emergency admission and COPD. The resulting tree is depicted in Fig. 1. A few procedure clusters—e.g. liver procedures—are diverse, perhaps meriting further subdivision that we do not consider here.

We began with 25 752 matched pairs. As described above, only the 22 622 pairs that were exactly matched for five potential effect modifiers were used to build the tree in Fig. 1. Ultimately

Table 1. Grouping of procedure clusters, with and without CHF

Procedure cluster	*No CHF, proc1*	*CHF, proc3*	*No CHF, proc2*	*CHF, proc4*
Adrenal procedures	×	×		
Appendectomy	×			×
Bowel anastamoses			×	×
Bowel procedures, other			×	×
Breast procedures	×	×		
Oesophageal procedures		×	×	
Femoral hernia procedures	×	×		
Gall bladder procedures	×	×		
Incisional and abdominal hernias	×	×		
Inguinal hernia procedures	×	×		
Large bowel resection			×	×
Liver procedures	×			×
Lysis of adhesions			×	×
Ostomy procedures			×	×
Pancreatic procedures		×	×	
Parathyroidectomy	×	×		
Peritoneal dialysis access procedure			×	×
Rectal procedures	×	×		
Repair of vaginal fistulas	×	×		
Small bowel resection			×	×
Splenectomy			×	×
Stomach procedures			×	×
Thyroid procedures	×	×		
Ulcer surgery			×	×
Umbilical hernia procedures	×			×
Ventral hernia repair	×	×		

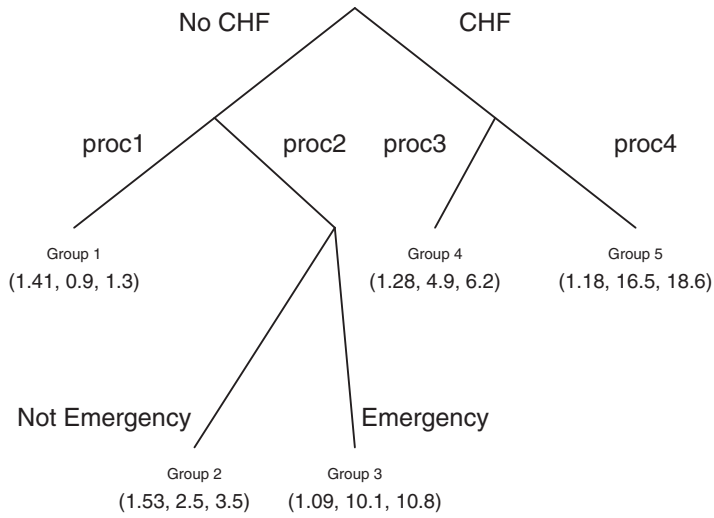

Fig. 1. Mortality in 23715 matched pairs of Medicare patients, one receiving surgery at a magnet hospital identified for superior nursing, the other undergoing the same surgical procedure at a conventional control hospital: the three values (A, B, C) at the nodes of the tree are A, McNemar odds ratio for mortality, control/magnet, B, 30-day mortality rate (%) at the magnet hospitals, and C, 30-day mortality rate (%) at the control hospitals

the CART algorithm used three of the five covariates and ignored the remaining two covariates, namely age>75 and COPD. To use the classification in Fig. 1, we need pairs that are exactly matched for three covariates, not for five covariates. Can we recover some of the pairs that we did not use because they were not matched for five covariates? To recover omitted pairs, we followed the tactic in Hsu *et al*. (2015). Specifically, we re-paired as many of the pairs that were not used to build the tree to be exact for the 130 procedures plus CHF and emergency admission, adding these additional 1093 pairs to the groups in Fig. 1, making 23715 pairs in total, or 95% of the original study. All analyses that follow refer to these 23715 pairs.

Consider the tree in Fig. 1, starting from its root at the top of Fig. 1. The tree split the population into two groups: patients without CHF and patients with CHF, which is a serious comorbid condition. It then split this divided population by grouping the 26 surgical procedure clusters. There are, of course, many ways to group 26 procedure clusters; for instance, there are $2^{26} - 1 = 67\,108\,863$ ways to split them into two groups. There are four groups of procedures: two for patients with CHF and two for patients without CHF. Table 1 displays the CART method's grouping of the 26 procedure clusters into proc1, proc2, proc3 and proc4. In Fig. 1, the CART method further divided proc2 into two subsets of patients: those admitted as emergencies and the remaining non-emergency patients. In Table 1, note that proc1 and proc3 overlap extensively, as do proc2 and proc4. To the clinical eye, with a few mild concerns, the procedures in proc2 and proc4 look riskier or more complex than those in proc1 and proc3. Groups proc1 and proc3 are very similar but not identical, and groups proc2 and proc4 are very similar but not identical. For instance, appendectomy is grouped with the less risky procedures in proc1 if the patient does not have CHF, but it was grouped with the more risky procedures in proc4 for a patient with CHF; however, it is unclear whether that switch is a profound insight or a hiccup.

The CART tree was built by predicting $|Y_i|$ from x_i. In contrast, hypothesis testing will use the signed value of Y_i.

Table 2. Mortality in 23715 matched pairs of a patient receiving surgery at a magnet hospital or a control hospital, where the pairs have been divided into five groups selected by the CART method

	Results for the following subgroups:					*Pooled*
	Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*	
CHF	No	No	No	Yes	Yes	
Procedures	proc1	proc2	proc2	proc3	proc4	
Emergency room admission	Both	No	Yes	Both	Both	
Number of pairs	10127	5636	2943	2086	2923	23715
Discordant pairs	210	293	488	217	760	1968
% discordant	2.1	5.2	16.6	10.4	26.0	8.3
Odds ratio	1.41	1.53	1.09	1.28	1.18	1.23
Mortality (%), magnet	0.9	2.5	10.1	4.9	16.5	4.7
Mortality (%), control	1.3	3.5	10.8	6.2	18.6	5.6

Table 3. Sensitivity analysis: upper bounds on *P*-values for various Γ

Γ	*P-values for the following subgroups:*					*Truncated product P*
	Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*	
1.00	0.008	0.000	0.195	0.039	0.013	0.000
1.05	0.019	0.001	0.374	0.080	0.062	0.000
1.10	0.042	0.003	0.576	0.143	0.184	0.012
1.15	0.079	0.010	0.753	0.230	0.386	0.032
1.17	0.099	0.015	0.809	0.270	0.479	0.044
1.20	0.135	0.025	0.875	0.335	0.616	0.163

3.2. Informal examination of outcomes

The first three numeric rows of Table 2 describe information that the CART method could use in building the tree, namely the number of pairs, the number of discordant pairs and the proportion of discordant pairs. In Table 2, 43% = 10 127/23 715 of pairs are in group 1, i.e. patients without CHF undergoing less risky procedures. Expressed differently, group 1 has the most pairs and the fewest discordant pairs of the five groups. As we might expect given the information that the CART algorithm was permitted to use, the proportion of discordant pairs varies markedly between the groups that it built.

The next three numeric rows of Table 2 display outcomes by treatment group, making use of Y_i and not just $|Y_i|$. The mortality rates for magnet and control groups are given, as is the odds ratio computed from discordant pairs; see Cox (1970). All the odds ratios are greater than or equal to 1, suggesting higher mortality at control hospitals. The largest odds ratio is in group 2, 1.53, whereas the largest difference in mortality rates is in group 5, $18.6\% - 16.5\% = 2.1\%$. The odds ratio that was closest to 1 is in group 3: the group that was most similar to group 2 except for admission through the emergency room.

3.3. Structured analysis of outcomes in discovered groups

The structured analysis in Hsu *et al*. (2015) starts by computing randomization tests and

upper sensitivity bounds on P-values for each of the five groups separately. In Table 3, these are based on a test of the McNemar type, essentially binomial calculations using discordant pairs; see Cox (1970) for discussion of paired binary data, and see Rosenbaum (2002), section 4.3.2, for the sensitivity analysis. In Table 3 are upper bounds on one-sided P-values testing no treatment effect in a group in the presence of a bias in treatment assignment of at most Γ. Also given in Table 2 are the odds ratios from discordant pairs associated with McNemar's test.

The final column in Table 3 gives the P-value for the truncated product of P-values as proposed by Zaykin *et al*. (2002). The truncated product generalizes Fisher's method for combining independent P-values: the test statistic is the product of those P-values that are smaller than a threshold τ, where $\tau = 0.1$ in Table 3. Zaykin *et al*. (2002) determined the null distribution of the truncated product statistic. Hsu *et al*. (2013) showed that the same null distribution may be used to combine upper bounds on P-values in a sensitivity analysis for a tree like Fig. 1, and that it often has superior power in this context compared with Fisher's product of all P-values, essentially because sensitivity analyses promise P-values that are stochastically larger than uniform for a given Γ. Truncation eliminates some very large upper bounds on P-values.

Hsu *et al*. (2015) combined the truncated product statistic with the closed testing procedure of Marcus *et al*. (1976) to provide strong control of the familywise error rate at α in a sensitivity analysis with a bias of at most Γ. Given G hypotheses $H_{\mathcal{G}_g}$, $g = 1, \ldots, G$, asserting no effect in each of G groups, closed testing begins by defining $2^G - 1$ intersection hypotheses $H_{\mathcal{L}}$, where $\mathcal{L} \subseteq \{1, \ldots, G\}$ is a non-empty set, and $H_{\mathcal{L}}$ asserts that $H_{\mathcal{G}_l}$ is true for every $l \in \mathcal{L}$. Closed testing rejects $H_{\mathcal{L}}$ at level α if and only if the P-value testing $H_{\mathcal{K}}$ is α or less for every $\mathcal{K} \supseteq \mathcal{L}$. The P-value testing $H_{\mathcal{K}}$ is based on the truncated product of P-values for $H_{\mathcal{G}_k}$ for $k \in \mathcal{K}$.

The P-value in the final column of Table 3 tests Fisher's hypothesis H_0, or $H_{\mathcal{L}}$ with $\mathcal{L} = \{1, 2, 3, 4, 5\}$. For $\Gamma = 1$, this test combines five McNemar tests using the truncated product and, in the absence of bias, the hypothesis H_0 is rejected with a one-sided P-value of 2.7×10^{-6}. To complete closed testing of subhypotheses, one performs $2^5 - 1 = 31$ tests of intersection hypotheses. Hypothesis $H_{\{3,4\}}$ has a P-value using the truncated product of 0.080, so neither $H_{\mathcal{G}_3}$ nor $H_{\mathcal{G}_4}$ is rejected at the 0.05-level by closed testing, but $H_{\mathcal{G}_1}$, $H_{\mathcal{G}_2}$ and $H_{\mathcal{G}_5}$ are rejected. In short, in the absence of bias, $\Gamma = 1$, the hypothesis of no effect is rejected in groups 1, 2 and 5.

At $\Gamma = 1.05$, Fisher's hypothesis of no effect at all is rejected at the 9.0×10^{-5} level, and closed testing rejects both $H_{\mathcal{G}_1}$ and $H_{\mathcal{G}_2}$ at the 0.05-level. At $\Gamma = 1.10$, Fisher's hypothesis H_0 of no effect is rejected at the 0.012-level, but only $H_{\mathcal{G}_2}$ is rejected at the 0.05-level. At $\Gamma = 1.17$, Fisher's hypothesis H_0 of no effect is rejected at the 0.044-level, no individual subgroup hypothesis is rejected at the 0.05-level but $H_{\{1,2\}}$ is rejected at the 0.05-level. At $\Gamma = 1.18$, no hypothesis is rejected at the 0.05-level.

A bias of $\Gamma = 1.17$ corresponds to an unobserved covariate that doubles the odds of having surgery at a control hospital and increases the odds of death by more than 60%, i.e., stated technically, $\Gamma = 1.17$ amplifies to $(\Lambda, \Delta) = (2.0, 1.61)$; see Rosenbaum and Silber (2009). McNemar's test applied to all 23 715 pairs yields a P-value bound of 0.063 at $\Gamma = 1.15$, so this overall test is slightly more sensitive to unmeasured biases and provides no information about subgroups.

What range of possible unmeasured biases, measured by Γ, should be explored? We do not know and cannot know how much bias is present in an observational study. However, in a straightforward way, we can and should determine the quantity of bias that would need to be present to alter the study's conclusions, for instance the bias that might lead to acceptance of a null hypothesis rejected in a conventional analysis that assumed no bias, $\Gamma = 1$. The degree of sensitivity to bias is a fact in the data brought to light by an appropriate analysis.

3.4. Use of the intensive care unit

In Table 2, magnet hospitals exhibited lower mortality than control hospitals for ostensibly similar patients undergoing the same surgical procedure, i.e. magnet hospitals exhibited better quality. Does better quality cost more? For resources that are allocated by a market mechanism—say, restaurants or hotels—we expect better quality to cost more, but market forces play little role in Medicare payments. In the absence of market forces, it is an open question whether better quality costs more. Silber *et al*. (2016) examined this issue in several ways, but Table 4 restricts attention to the consumption of a particularly expensive resource, namely use of the ICU. In a hospital with inadequate nursing staff, a patient may be placed in the ICU to ensure that the patient is monitored, whereas in a hospital with superior nursing this same patient might remain on a conventional hospital ward. This is one mechanism by which better quality—lower mortality rates—might cost less, not more.

Is the lower mortality in magnet hospitals associated with greater use of the ICU?: apparently not. Overall and in all five groups in Fig. 1, the use of the ICU in Table 4 is lower at magnet hospitals than at control hospitals. The odds ratio is largest in group 2, but it is not small in any group. In various other ways also, Silber *et al*. (2016) found that costs were lower at hospitals with superior nursing, despite lower mortality rates.

The closed testing procedure applied to the sensitivity analysis in Table 5 rejects the null

Table 4. Use of the ICU in 23715 matched pairs of a patient receiving surgery at a magnet hospital or a control hospital, where the pairs have been divided into five groups indicated in Fig. 1

	Results for the following subgroups:					*Pooled*
	Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*	
CHF	No	No	No	Yes	Yes	
Procedures	proc1	proc2	proc2	proc3	proc4	
Emergency room admission	Both	No	Yes	Both	Both	
Number of pairs	10127	5636	2943	2086	2923	23715
Discordant pairs	2675	2361	1282	859	970	8147
% discordant	26.4	41.9	43.6	41.2	33.2	34.4
Odds ratio	1.63	2.05	1.67	1.70	1.88	1.78
ICU (%), magnet	15.3	28.9	53.8	41.0	69.8	32.3
ICU (%), control	21.7	43.3	64.6	51.7	80.0	42.0

Table 5. Sensitivity analysis: upper bounds on *P*-values for various Γ

Γ	*P-values for the following subgroups:*					*Truncated product P*
	Group 1	*Group 2*	*Group 3*	*Group 4*	*Group 5*	
1	0.000	0.000	0.000	0.000	0.000	0.000
1.5	0.017	0.000	0.037	0.040	0.000	0.000
1.6	0.312	0.000	0.254	0.203	0.009	0.000
1.7	0.849	0.000	0.651	0.511	0.074	0.000
1.8	0.993	0.002	0.916	0.798	0.276	0.049
1.9	1.000	0.047	0.989	0.945	0.582	0.235

Table 6. Cross-tabulation of mortality and ICU use in 5636 pairs in group 2†

Control hospital	*Magnet hospital*			*Total*
	Dead	*Alive, ICU*	*Alive, no ICU*	
Dead	23	72	105	200
Alive, ICU	60	744	1493	2297
Alive, no ICU	56	726	2357	3139
Total	139	1542	3955	5636

†The table counts pairs of patients, not individual patients, with columns recording outcomes for the patient in the magnet hospital and rows recording outcomes for the matched patient in the control hospital.

hypothesis of no effect on ICU utilization in all five groups provided that the bias in treatment assignment is at most $\Gamma = 1.5$. Using the method in Rosenbaum and Silber (2009), a bias of $\Gamma = 1.5$ corresponds to an unobserved covariate that increases the odds of surgery at a control hospital by a factor of 4 and increases the odds of going to the ICU by a factor of 2. Closed testing rejects no effect only in group 2 for $1.6 \leqslant \Gamma \leqslant 1.8$ and cannot reject even Fisher's H_0 for $\Gamma = 1.9$. Detailed results for group 2 are given in Table 6.

Tables 2–6 concern the effect of going to a magnet hospital rather than a control hospital for surgery, but they do not show the specific role of nurses in this effect. It is entirely plausible that superior nurse staffing would permit more patients to stay out of the ICU, but nothing in the data indicates this directly. The main difference between the ICU and the ward of the hospital is the higher density, often higher quality, of the nurse staffing in the ICU. A hospital with a higher nurse-to-bed ratio and superior nurse staffing may be able to care for a seriously ill patient on the hospital ward, where some other hospital would be forced to send the same patient to the ICU.

3.5. Other analyses and options for analysis

The tree in Fig. 1 was built for mortality but was used also for ICU use. In an additional analysis, we applied the CART method to each leaf of Fig. 1 to predict unsigned discordance for ICU use. The two interesting aspects of this analysis were that

(a) subgroup 2 in Fig. 1 was not further divided and
(b) subgroup 5 in Fig. 1 was further divided, with more evidence of an effect on ICU use among patients in this subgroup who were not admitted through the emergency room, which is a pattern that is analogous to subgroups 2 and 3.

An interesting feature of this type of analysis is that it makes mortality the primary end point, as it would be in most surgical studies, so only mortality determines the initial tree for the mortality analysis, but it permits the secondary outcome of ICU use to affect a secondary tree.

We let the CART algorithm build the groups. Any analysis that used only $|Y_i|$ and $\mathbf{x}_i$ could be used to build the groups. In saying this, we mean that the strong control of the familywise error rate in Hsu *et al*. (2015) would not be affected by revisions to the tree that used only $|Y_i|$ and $\mathbf{x}_i$. Indeed, a surgeon who did not look at Y_i could look at Fig. 1 and Table 1 and decide to regroup some of the procedure groups. Perhaps the surgeon would view some of the CART method's

decisions as clinically unwise and would change them, or perhaps the surgeon would prefer that proc1 and proc3 be identical, and that proc2 and proc4 be identical. Indeed, the surgeon might suggest fitting the tree again, using only $|Y_i|$ and $\mathbf{x}_i$, but subdividing some procedure clusters, say liver procedures, that seem too broad to be clinically meaningful. What is critical is that the groups are formed using $|Y_i|$ and $\mathbf{x}_i$ without using the sign of Y_i.

4. Discussion: it is important to notice subgroups with larger treatment effects in observational studies

In an observational study of treatment effects, there is invariably concern that an ostensible treatment effect is not actually an effect that is caused by the treatment, but rather some unmeasured bias distinguishing treated and control groups. Larger or more stable treatment effects are more insensitive to such concerns than smaller or more erratic effects, i.e. larger biases measured by Γ would need to be present to explain a large and stable treatment effect. These considerations motivate an interest in effect modification in observational studies. Perhaps the treatment effect is larger or more stable in certain subgroups defined by observed covariates. If so, the ostensible treatment effect in such subgroups is likely to be insensitive to larger unmeasured biases, and therefore more credible, and, additionally, a larger or more stable effect is likely to be more important clinically.

The magnet hospitals had lower mortality overall, and lower or equivalent mortality in each of the five groups. However, the superior staffing of magnet hospitals was least sensitive to unmeasured bias in our group 2, consisting of patients undergoing relatively serious forms of surgery in the absence of other life-threatening conditions, such as CHF or an emergency admission leading to surgery. Moreover, not only were mortality rates lower in magnet hospitals for these patients (2.5% rather than 3.5%), but additionally the magnet hospitals cared for these patients with greatly reduced use of an expensive resource, namely the ICU (an ICU rate of 28.9% rather than 43.3%). Determining the cost of hospital care for Medicare patients is not straightforward, so Silber *et al*. (2016) contrasted several formulae to appraise the cost of magnet hospitals. In all these formulae, use of the ICU plays a substantial part, as does the length of stay in the hospital. Regardless of which formula was used, magnet hospitals appear to produce lower mortality either at no additional cost or with a cost savings.

A plausible interpretation of Fig. 1 and Tables 1–3 is that

(a) patients in groups 2, 4 and 5 should be directed to magnet hospitals, which are a limited resource,
(b) the large number of comparatively healthy patients requiring simpler surgical procedures may go to non-magnet hospitals if space in a magnet hospital is unavailable and
(c) patients in group 3 requiring emergency surgery should go to the nearest hospital.

References

Aiken, L. H., Havens, D. S. and Sloane, D. M. (2000) The magnet nursing services recognition program. *Am. J. Nursng*, **100**, 26–35.

Athey, S. and Imbens, G. (2016) Recursive partitioning for heterogeneous causal effects. *Proc. Natn. Acad. Sci. USA*, **113**, 7353–7360.

Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Chesher, A. (1984) Testing for neglected heterogeneity. *Econometrica*, **52**, 865–872.

Cornfield, J., Haenszel, W., Hammond, E., Lilienfeld, A., Shimkin, M. and Wynder, E. (1959) Smoking and lung cancer. *J. Natn. Cancer Inst.*, **22**, 173–203.

Cox, D. R. (1970) *Analysis of Binary Data*. London: Methuen.
Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2008) Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Statist.*, **90**, 389–405.
Ding, P., Feller, A. and Miratrix, L. (2016) Randomization inference for treatment effect variation. *J. R. Statist. Soc.* B, **78**, 655–671.
Egleston, B. L., Scharfstein, D. O. and MacKenzie, E. (2009) On estimation of the survivor average causal effect in observational studies when important confounders are missing due to death. *Biometrics*, **65**, 497–504.
Fisher, R. A. (1935) *The Design of Experiments*. Edinburgh: Oliver and Boyd.
Gastwirth, J. L. (1992) Methods for assessing the sensitivity of statistical comparisons used in Title VII cases to omitted variables. *Jurimetrics*, **33**, 19–34.
Gilbert, P., Bosch, R. and Hudgens, M. (2003) Sensitivity analysis for the assessment of the causal vaccine effects on viral load in HIV vaccine trials. *Biometrics*, **59**, 531–541.
Hosman, C. A., Hansen, B. B. and Holland, P. W. H. (2010) The sensitivity of linear regression coefficients' confidence limits to the omission of a confounder. *Ann. Appl. Statist.*, **4**, 849–870.
Hsu, J. Y., Small, D. S. and Rosenbaum, P. R. (2013) Effect modification and design sensitivity in observational studies. *J. Am. Statist. Ass.*, **108**, 135–148.
Hsu, J. Y., Zubizarreta, J. R., Small, D. S. and Rosenbaum, P. R. (2015) Strong control of the familywise error rate in observational studies that discover effect modification by exploratory methods. *Biometrika*, **102**, 767–782.
Lehmann, E. L. and Romano, J. (2005) *Testing Statistical Hypotheses*. New York: Springer.
Lehrer, S. F., Pohl, R. V. and Song, K. (2016) Targeting policies: multiple testing and distributional treatment effects. *Working Paper WP22950*. National Bureau of Economic Research, Cambridge.
Liu, W., Kuramoto, J. and Stuart, E. (2013) An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevn Sci.*, **14**, 570–580.
Lu, X. and White, H. (2015) Testing for treatment dependence of effects of a continuous treatment. *Econmetr. Theory*, **31**, 1016–1053.
Marcus, R., Peritz, E. and Gabriel, K. R. (1976) On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, **63**, 655–660.
Neyman, J. (1990) On the application of probability theory to agricultural experiments. *Statist. Sci.*, **5**, 463–480.
Rosenbaum, P. R. (2002a) *Observational Studies*, 2nd edn. New York: Springer.
Rosenbaum, P. R. (2002b) Attributing effects to treatment in matched observational studies. *J. Am. Statist. Ass.*, **97**, 183–192.
Rosenbaum, P. R. (2004) Design sensitivity in observational studies. *Biometrika*, **91**, 153–164.
Rosenbaum, P. R. (2005) Heterogeneity and causality: unit heterogeneity and design sensitivity in observational studies. *Am. Statistn*, **59**, 147–152.
Rosenbaum, P. R. (2015) Two R packages for sensitivity analysis in observational studies. *Observnl Stud.*, **1**, 1–17.
Rosenbaum, P. R. and Silber, J. H. (2009) Amplification of sensitivity analysis in observational studies. *J. Am. Statist. Ass.*, **104**, 1398–1405.
Rosenbaum, P. R. and Small, D. S. (2017) An adaptive Mantel–Haenszel test for sensitivity analysis in observational studies. *Biometrics*, to be published.
Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.
Silber, J. H., Rosenbaum, P. R., McHugh, M. D., Ludwig, J. M., Smith, H. L., Niknam, B. A., Even-Shoshan, O., Fleisher, L. A., Kelz, R. R. and Aiken, L. H. (2016) Comparison of the value of better nursing work environments across different levels of patient risk. *J. Am. Med. Ass. Surg.*, **151**, 527–536.
Stuart, E. A. and Hanna, D. B. (2013) Should epidemiologists be more sensitive to design sensitivity? *Epidemiology*, **24**, 88–89.
Wager, S. and Athey, S. (2017) Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Statist. Ass.*, to be published.
Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. and Weir, B. S. (2002) Truncated product method of combining *P*-values. *Genet. Epidem.*, **22**, 170–185.
Zhang, H. and Singer, B. H. (2010) *Recursive Partitioning and Applications*. New York: Springer.
Zubizarreta, J. R., Cerdá, M. and Rosenbaum, P. R. (2013) Effect of the 2010 Chilean earthquake on posttraumatic stress: reducing sensitivity to unmeasured bias through study design. *Epidemiology*, **24**, 79–87.