12-2012

Near/Far Matching: A Study Design Approach to Instrumental Variables

Mike Baiocchi

Dylan S. Small
University of Pennsylvania

Lin Yang
University of Pennsylvania

Daniel Polsky
University of Pennsylvania

Peter W. Groeneveld

Recommended Citation

Baiocchi, M., Small, D. S., Yang, L., Polsky, D., & Groeneveld, P. W. (2012). Near/Far Matching: A Study Design Approach to Instrumental Variables. *Health Services and Outcomes Research Methodology, 12* (4), 237-253. http://dx.doi.org/10.1007/s10742-012-0091-0

Near/Far Matching: A Study Design Approach to Instrumental Variables

Abstract

Classic instrumental variable techniques involve the use of structural equation modeling or other forms of parameterized modeling. In this paper we use a nonparametric, matching-based instrumental variable methodology that is based on a study design approach. Similar to propensity score matching, though unlike classic instrumental variable approaches, near/far matching is capable of estimating causal effects when the outcome is not continuous. Unlike propensity score matching, though similar to instrumental variable techniques, near/far matching is also capable of estimating causal effects even when unmeasured covariates produce selection bias. We illustrate near/far matching by using Medicare data to compare the effectiveness of carotid arterial stents with cerebral protection versus carotid endarterectomy for the treatment of carotid stenosis.

Keywords

instrumental variables, matching, study design, binary outcomes, comparative effectiveness, medicare data

Disciplines

Other Statistics and Probability | Vital and Health Statistics

Near/Far Matching - A Study Design Approach to Instrumental Variables

Mike Baiocchi[1*], Dylan Small[3,4,5], Lin Yang[3], Daniel Polsky[3,4] and Peter W Groeneveld[2,3,4]

Abstract: Classic instrumental variable techniques involve the use of structural equation modeling or other forms of parameterized modeling. In this paper we use a nonparametric, matching-based instrumental variable methodology that is based on a study design approach. Similar to propensity score matching, though unlike classic instrumental variable approaches, near/far matching is capable of estimating causal effects when the outcome is not continuous. Unlike propensity score matching, though similar to instrumental variable techniques, near/far matching is also capable of estimating causal effects even when unmeasured covariates produce selection bias. We illustrate near/far matching by using Medicare data to compare the effectiveness of carotid arterial stents with cerebral protection (CAS) versus carotid endarterectomy (CEA) for the treatment of carotid stenosis.

Key terms: instrumental variables; matching; study design; binary outcomes; comparative effectiveness; Medicare data

[1]Stanford University, Department of Statistics, 390 Serra Mall, Stanford, CA 94305.
[2]Department of Veterans Affairs' Center for Health Equity Research and Promotion, Philadelphia Veterans Affairs Medical Center
[3]Department of Medicine, University of Pennsylvania School of Medicine
[4]Leonard Davis Institute of Health Economics, University of Pennsylvania
[5]Department of Statistics, The Wharton School, University of Pennsylvania

[*]Corresponding author: Mike Baiocchi. email: baiocchi@stanford.edu. telephone: 415-517-3892. Fax: 650-725-8977.

Grant support: National Heart, Lung, and Blood Institute R01HL086919, Agency for Healthcare Research and Quality R01HS018403.

1 Introduction

For comparative effectiveness research to reach its potential, there must be reliable methods to address confounding by indication using "real world" data. Without randomization, the groups receiving the treatment tend to be systematically different than those patients receiving the control and many of these differences typically go unmeasured. As a result, statistical procedures such as regression and propensity score matching are unable to properly adjust. It is common in the literature to either use a method like propensity score matching and *assume* that all important covariates are available in the data set (i.e., that strongly ignorable treatment assignment holds) or to use a classical instrumental variables approach. Near/far matching (Baiocchi et al., 2010) is a new technique that synthesizes these two approaches and thus offers opportunities to adequately address confounding by indication in observational data settings.

Near/far matching uses a study design approach to replicate the structure of a clinical trial framework within an observational setting. The following quote from Rosenbaum (2010), helps to delineate the "design" from the "analysis" of a study: "In practice, the design of an observational study consists of all activities that precede the examination of those outcome measures that will be the basis for the study's conclusions… In theory, design anticipates analysis. Analysis is ever present in design, as any goal is ever present in any organized effort, as a goal is necessary to organize effort." Most readers will be familiar with propensity score matching, which is also a study design approach. The matching phase of the procedure is study design which prepares the data for statistical analysis (e.g., using a paired t-test). In this way, near/far matching is similar to propensity score matching in that there is a matching phase to prepare the data for the evaluation of outcomes in a structure designed to mimic a clinical trial. The difference is that near/far matching harnesses the randomization of an instrument and uses this to construct an analysis which is capable of estimating treatment effects when there is selection on unobserved covariates. Additionally, near/far matching is also the correct analysis tool for many settings because it is one of just a few instrumental variable (IV) approaches which is appropriate for estimating causal effects when the outcome of interest is binary.

In this paper we demonstrate the near/far matching technique to estimate the comparative effectiveness of carotid arterial stents with cerebral protection (CAS) versus carotid endarterectomy. Section 2 introduces this motivating example. Section 3 details the data with particular attention to the instrumental variable. Section 4 offers a review of the literature with focus on methods for estimating treatment effects for binary outcomes. Section 5 is an intuitive introduction to near/far matching and places near/far matching in context with already existing techniques. We introduce the notation and mathematical framework for near/far matching in section 6. In section 7 we present the results of our

example. Section 8 of this paper discusses a few advantages of near/far matching as well pointing out a concern in designing such a study.

2 Motivating example: Comparing two interventions when there is selection bias

The motivating example for this paper comes from a comparison of carotid arterial stents (CAS) versus carotid endarterectomy (CEA) for the treatment of carotid stenosis. Carotid stenosis (i.e., narrowing of the carotid artery) is among the most common causes of stroke in the United States (Barnett et al. 1996, Dodick et al. 2004). For decades, carotid endarterectomy (CEA), a highly invasive vascular surgical technique, was the only effective interventional treatment for severe carotid stenosis. However, in late 2004 based on the result of a randomized clinical trial (Yadav et al. 2004), the FDA approved CAS for use in patients with severe carotid stenosis who were deemed "high risk" for CEA. Utilization of CAS in many U.S. hospitals grew rapidly in 2005-2006, yet uncertainty about the comparative effectiveness of the two treatment options was heightened by the publication of additional clinical trials with results that questioned the efficacy of CAS (Mas et al. 2006, Mas et al. 2008). The real-world comparative effectiveness of CAS versus CEA remains uncertain.

Use of CAS remains highly variable geographically, suggesting a lack of uniformity in which patients are being treated with CAS versus CEA nationwide. See Figure 1 for a histogram of the rates of CAS utilization by HRR. Figure 2 is a map of the HRRs and their CAS utilization. As with any new technology, there are early adopters and late adopters, resulting from a complicated process involving factors such as the number of teaching hospitals in a region, professional and institutional relationships between advocates of the new technology and those who are willing to try it, as well as logistical issues such as a hospital's existing stock of the old technology and the difficulties involved in updating to the new technology, all of which may impact the rates of use of a new technology. Many of the factors which go into determining the treatment selection occur as a process which is unrelated to patient- level covariates. We will exploit geographic variation in the design of our study.

3 Description of data

Using health care utilization and outcomes data from the Medicare program for fee-for-service beneficiaries over age 65, we compared the effectiveness (i.e., mortality rate at 180 days following the procedure) of carotid arterial stents (CAS) to carotid endarterectomy (CEA) for the treatment of carotid stenosis. The data includes patients treated from the years 2005-2008, during the period where both CAS and CEA were in use. In addition to those years, we used 2004 utilization data for CEA (i.e., pre-CAS), which we will make use of during our analysis to control for pre-existing patterns of care. The data set has information on approximately 325,000 patients treated with either CAS or CEA (approximately 13% of were CAS recipients). These data indicate each patient's

demographic information, the date and location of procedure receipt, the presence of important comorbid conditions, and subsequent major clinical outcomes such as stroke or death.

We use the geographic variation in the uptake of CAS after 2004 FDA approval as an instrumental variable. Using the geographical conventions established by the Dartmouth Atlas for Health Care, we use Hospital Referral Regions (HRRs, n=306) as our geographical unit of analysis. Our outcome of interest is mortality 90 days after intervention.

3.1 The instrument

An instrument is a random influence towards acceptance of a treatment which affects outcomes only to the extent that it affects acceptance of the treatment. Even in settings in which treatment assignment is mostly deliberate, there may nonetheless exist some essentially random influences to accept treatment, so that treatment assignment retains a random component. An instrument is weak if the random influences barely affect treatment assignment, or strong if they are decisive in influencing treatment assignment.

In this paper we use the HRR where the patient received care as the instrument. The Patients sort themselves into different geographic areas for a variety of reasons: socioeconomic, familial and cultural. These imbalances are evident in Table 1. From Table 1 we can see that the patients in our data set which are treated in high utilization HRRs tend to be more racially diverse, have higher incomes and the HRRs tend to have higher medical expenditure per patient and have more beds available *per capita* at academic institutions. We attempt to control for these socioeconomic differences by matching on these variables at the HRR level (see methods below). Regional variation will function properly as an instrument if it is uncorrelated with the patient-level confounders of concern.

The usual argument for the validity of regional variation as an instrument is: Though we note patients sort themselves based on socioeconomic differences across regions, it is unlikely that they sort themselves into different regions based on their medically relevant covariates. See **Table XX** which shows the patient-level medically relevant covariates across the instrument. Note that they are roughly similar. In fact, we will go to great lengths to control for all of the medically relevant covariates we have in our data set by pair matching at the patient-level. But we will also make use of which region the patient was from and therefore which treatment was more likely to be assigned for reasons that are extraneous to the particulars of the patient's medical history. In the example at hand we should be a bit cautious, environmental factors such as dietary habits, levels of physical fitness and exertion and other culturally influenced behaviors may have an impact on medically relevant, patient-level covariates. This would imply that HRR may be correlated with unobserved patient-level covariates. In this paper we are using this example as an

illustration of the methodology so we will not delve further into this issue; a more complete investigation of CAS vs CEA would need to engage this issue. We do point the interested reader to Section 8 and the brief discussion of sensitivity analysis for one potential statistical approach for addressing imperfect instruments.

An instrument can be thought of as an "encouragement" for the patient to take a given treatment. While a patient may be encouraged to take a treatment he/she is free to take the treatment or the control. See Holland (1988) and Angrist, Imbens and Rubin (1996) for a discussion of this framework. In this framework it is possible that the intensity of encouragement can vary. In our example, a few HRRs have rates of CAS as high as 50% whereas about a dozen have zero CAS utilization.

4 Review of literature

Propensity score matching is a common tool of choice in the health services research literature. One of the primary reasons for its wide application is that propensity score matching emulates the study design approach taken in a clinical trial. The simplicity of a clinical trial and the resulting force gained from its clarity of design are attractive. In a complex setting, where both biology and human decision making is involved, a simple statistical method which is warranted by design is often preferable to the convolutions often required by highly parametric models. But this method is inappropriate in our setting because strongly ignorable treatment assignment is not realistic. Strongly ignorable treatment assignment requires that the joint distribution of the potential outcome be independent of the treatment assignment conditional on the covariates (Rosenbaum and Rubin 1983).

Many times instrumental variables (IV) is implemented using two-stage least squares (2SLS). This is appropriate when the outcome of interest is continuous. For example, if we were considering the change in weight (measured in pounds) due to a new surgical intervention it may be appropriate to use 2SLS because weight is a continuous variable. In our motivating example we have a binary outcome – patients will either be alive or dead at 90 days after the intervention. Many research questions in health services have binary outcomes. It has been suggested that in some settings it may be appropriate to use a linear probability model in both stages of a 2SLS (Angrist 2001) as an approximation to a more correct procedure. In Bhattacharya et al. (2006) a simulation study demonstrated that bias is introduced by using linear probability models when the empirical probability of the event is up against the parameter space, that is if the event either occurs quite frequently (close to 100% of the time) or very infrequently (close to 0% of the time). In our case only 2% of the patients die, so we are in need of an approach more appropriate to binary outcomes.

In analogy to 2SLS, some researchers have used a logistic (or probit) model in the second stage of their regression when encountering a binary oucome, but this is problematic. The properties of linear models which allow 2SLS to work so nicely (e.g., orthogonality) are corrupted by the link functions in standard generalized linear models, and two stage logistic approaches can have biases even in large samples. See pages 190-192 of Angrist & Pischke (2009) and Cai, Small & Ten Have (2011) for discussion of biases for two stage logistic approaches.

In this paper we use a method we call "near/far matching," which was first described in Baiocchi et al. (2010). Near/far matching is capable of estimating treatment effects even when the outcome is binary. This method is a matching-based approach, similar to propensity score matching, but uses information about the instrument to construct the most informative matched pairs from an observational data set.

There are other IV techniques which have been developed to deal with binary outcomes. In particular, the two-stage residual inclusion model (2SRI) is a well-developed alternative see Terza, Basu & Rathouz (2008) and Cai, Small & Ten Have (2011).

5 Near/Far Matching: Overview of the method

Before explaining near/far matching we discuss a hypothetical study design approach researchers might take to investigate the relative efficacy of CAS vs CEA. We outline this hypothetical approach first in order to parallel its setup with near/far matching.

One can imagine doing an RCT to study the comparative efficacy of CAS vs CEA. A randomized, matched-pair study design would first match patients based on covariates and then randomize within the matched-pair. The pair matching in this design ensures the observed covariate distribution for the treated is similar to the control, thus reducing the extraneous variation in the null distribution due to differences in the observed covariates. The randomization supports the assumption that the unobserved covariates are also balanced between the two groups and thus strengthens the argument that the observed variation in the outcome is attributable to the difference in the level of the treatment. This is all standard thinking to most statisticians, but there are other considerations in designing a study.

Once the treatment assignments have been randomly assigned it is then necessary to ensure the patients comply with their treatment assignment. In practice, if there is minimal encouragement from the researchers then the patients may decide to become noncompliant with the randomization. This encouragement can take many forms – for example, collaboration with treating physicians, as well as free/reduced cost of care or other monetary incentives for patients who are treatment "compliers.". The objective is to have the patients stay compliant with the treatment to which they were randomly assigned.

One concern with high rates of noncompliance is that the patients would be no longer randomly assigned to treatment, and it becomes more likely that covariates, both observed and unobserved, are determining the treatment selection. Thus higher rates of compliance are desirable, hence encouragement to comply is a vital part of the study design.

In analogy to the RCT, at the outset of the analysis, it is advisable for the analyst performing a near/far matching to blind him/herself to three kinds of variables to ensure the information is not used in the matching procedure outlined below. The first variable is the outcome of interest. The second is covariate values which were recorded post-treatment (Rosenbaum 1984). The third, which is a departure from propensity score matching, is to remove the variable which records which treatment the patient actually received. The variables the analyst uses to create the matches in near/far matching are pretreatment covariates and the instrument. The assumption of strongly ignorable treatment assignment implies that, conditional on the observed covariates, the potential outcomes are independent of the treatment assignment; this assumption is why treatment assignment is used in propensity score matching and why the analyst should not use it directly to construct pairs in near/far matching.

There are two objectives in near/far matching. As in an RCT matched-pair design, one objective in near/far matching is to create matched pairs where the covariates are similar within a pair. Creating pairs with very similar covariate values (i.e., pairs which are "near" each other in covariate space) is used to improve efficiency. The other objective in near/far matching is to separate patients' instrument values within a matched pair. In our example, within a matched pair we want one patient who was highly encouraged to have CAS and the other to be highly encouraged to take CEA. This is similar to the matched pair design when there is the potential for noncompliance. If we can vary the level of encouragement then it is preferential to have two patients who are highly dissimilar ("far") in their levels of randomly assigned encouragement because it is then more likely that within the pair the one patient will comply with the encouragement and take the treatment and the other will comply with the lack of encouragement and take the control. Using an algorithm outlined in the next subsection we will construct pairs which maximize both of these objectives at the same time.

In most real-world examples there will be a trade-off between the "near" and the "far" part of the matching. The technical aspects of this trade-off, and how to construct such pairs, will be discussed in the next subsection. The intuition is that as the analyst forces separation in the instrument values between pairs of patients it becomes more difficult to find patients with quite dissimilar instrument values but very similar covariates. The Baiocchi et al (2010) paper outlines both theoretical arguments as well as practical reasons for designing studies with greater separation in the instrument.

It should be noted that we are referring to "pair" matching, but all of these arguments hold for larger block designs. Near/far matching would work with k:1 matching and other more exotic designs. The primary difference would be the optimization algorithm used to construct the sets. The nonbipartite algorithm we use in our analysis, developed in Derigs (1988) and first used in a statistical setting in Lu, et al. (2001), is useful for pair matching.

5.1 Near/far matching when the instrument is applied at a group level

In this particular application there will be two rounds of matching. First we will match hospital referral regions (HRRs) using near/far matching. Then we will use an optimal bipartite match to construct patient-level matches across HRR pairs.

The first stage of our matching uses near/far matching in order to construct the strongest instrumental variable design from what is, initially, a weak instrument. If we were to include all of the HRRs in our analysis, we would find that there are some HRRs with very different populations which would create covariate imbalance. Using all HRRs in our analysis would also mean using some with moderate use of CAS. HRRs with moderate use of CAS are not helpful for our analysis because these HRRs are not encouraging their patients very strongly in either direction, toward CAS or CEA, relative to other HRRs – thus it would be difficult to create much separation in the encouragement due to the instrument.

This step, designing our analysis to include certain HRRs and exclude others, is similar to the inclusion/exclusion criteria of a randomized controlled trial. By restricting which units of observation can enter an analysis we are gaining in precision of analysis by (1) reducing heterogeneity in the covariates of the units of observation and (2) by increasing the strength of the instrument. This must be balanced against the consideration that we are effectively limiting the generalizability of the results of our study. The issue is a bit more complicated by the fact that an instrumental variable estimate is on a subset of the population which enters our study – this is referred to as the "complier average causal effect" in Angrist, Imbens and Rubin (1996). Specific advice for the trade-off is difficult to offer as context will drive the importance of the trade-offs. Researchers should take the following three items into consideration when deciding how to design their study: the strength of the instrument (e.g., if it is weak then more separation in the instrument may benefit the study which means potentially excluding more observations), any suspected violations of the instrument (e.g., it is well known that weak instruments are particularly susceptible to bias when violations of the instrumental variable assumptions occur – see Bound, Jaeger & Baker (1995) for an excellent discussion), as well as starting sample size. Given a particular dataset, these items determine on what population the analysis can be run. By using calipers – see Rosenbaum (2010, §9.2) – the researchers may construct different matched designs (stopping short of examining the outcomes), consider the units

of observation which are included in the match, and then determine whether this sample is informative of the population the researchers are interested in investigating.

After the first round of matching we have pairs of HRRs that are quite similar in clinically relevant covariates but quite different in their usage of CAS. In the second round of matching we construct pairs of patients wherein one member of the pair – one who was treated in the HRR with higher CAS usage – is optimally matched to a patient from the paired HRR which had lower usage of CAS. Whereas the first round of matching is meant to control for HRR-level confounding covariates, the second round of matching addresses medically relevant patient-level covariates such as age, gender, race, and the presence of various comorbid conditions, thus improving the power of the inference.

To see a slightly simpler design using near/far matching see Baiocchi et al (2010). That study used proximity to treatment facility as an instrument and created pair matches of premature babies which had similar covariates (near) but were quite dissimilar on their proximity to treatment facility (far). Both the instrument and the outcome were on the patient-level. In the current example the instrument is applied on the HRR level and the outcome is on the patient level.

5.2 Near/Far Matching: Constructing the match

We will first pair-match HRRs using optimal nonbipartite matching, so they are as similar on clinically relevant HRR-level covariates as possible, while at the same time preferentially creating pairs of HRRs which are as dissimilar as possible in their percentage usage of CAS. For statistical applications of optimal nonbipartite matching, see Lu, et al. (2001), Rosenbaum and Lu (2004), Lu (2005), and Rosenbaum (2005). One of the most important covariates we will pay attention to in this first round of matching is 2004 death rates. We focus on 2004 death rates because this is pre-CAS, so all HRRs were using CEA as the only surgical treatment of carotid stenosis. If these rates are stable, then post-treatment death rates in 2004 will reflect the HRR's "base rate" of mortality before the introduction of CAS. If, pre-CAS, two HRRs have similar mortality rates but then, post-introduction of CAS, they start to diverge in both CAS utilization and subsequently mortality then we have strong evidence in support of an efficacy difference between CAS and CEA. We will also pay attention to percentage of teaching hospitals in the HRR, socioeconomic and demographics in the HRRs, beds per capita and other potentially important HRR-level variables.

In the health policy literature the most common form of matching is a matching between two distinct groups – for example when patients who received the treatment are matched to those who did not receive the treatment, as in propensity score matching. This form of matching is called bipartite matching – matching made between two distinct groups. In matching the HRRs we are attempting to create pairs where the difference between one

HRR's usage of CAS is quite different from the other HRR's. But each HRR has some level of usage of CAS so we cannot break the HRRs into two separate groups before the matching starts. We are operating under the dose matching framework as described in Lu et al. (2001). In this setting any HRR has the potential to be matched to any other HRR – this is referred to as nonbipartite matching. For implementation we used the nonbipartite algorithm developed in Derigs (1988). In this stage of the analysis we create pair matches that are as close as possible in HRR-level covariates but as dissimilar in CAS usage as possible because then the primary difference between HRRs will be their CAS usage, everything else being equal.

Let us say there are 2N HRRs. First, a discrepancy is defined between every pair of HRRs, yielding a 2N x 2N discrepancy matrix. (The term 'discrepancy' is used in place of the more common term 'distance' to avoid confusion of covariate discrepancy with the geographic distance between HRRs.) An optimal nonbipartite matching then divides the 2N HRRs into nonoverlapping pairs of two HRRs in such a way that the sum of the discrepancies within the N pairs is minimized. That is, two HRRs in the same pair are as similar as possible in their covariates while also being quite different in their utilization of CAS. In order to get the best covariate balance between the two groups, and at the same time achieve good separation in the instrument (see Baiocchi et al 2010 for a discussion of why separation in the instrument is desirable) some of the HRRs must be removed from the analysis. We do this in an optimal way by using "sinks" see Lu, et al. (2001). To remove *e* HRRs, *e* sinks are added to the data set before matching, where each sink is at zero discrepancy to each HRR and at infinite discrepancy to all other sinks. This yields a (2N + e) x (2N + e) discrepancy matrix. An optimal match will pair *e* HRRs to the *e* sinks in such a way as to minimize the total of the remaining discrepancies within N-e/2 pairs of 2N-e HRRs; that is, the best possible choice of *e* HRRs is removed.

The discrepancy matrix was built in several steps using standard devices. Because we are matching HRRs from different parts of the US, and because socioeconomic and demographic factors have been linked to health outcomes we need to control for these HRR-level covariates. Additionally, we want to make sure medically relevant covariates, such as 2004/pre-CAS mortality rates are similar within pair-matched HRRs. The discrepancy between every pair of HRRs was calculated using Mahalanobis distance. A small penalty (i.e., a positive number) was added to the discrepancy for each of the following circumstances (1) if the average HRR-level spending for inpatients during their last 6 months of life was too divergent between two HRRs (2) if the number of beds in 2005 at academic medical institutes per capita were too divergent (3) if median income of the HRRs were too dissimilar. Two independent observations drawn from the same L-variate multivariate Normal distribution have an expected Mahalanobis discrepancy of 2L, so that, speaking informally, a penalty that is typically of size 2 will double the importance of

matching on a variable. Small penalties are used to secure balance for a few recalcitrant covariates, usually those which are most systematically out of balance; see Rosenbaum (2010, §9.2) for discussion. It is typical to adjust small penalties to secure the desired balance. Finally, a substantial penalty was added to the discrepancy between any pair of HRRs whose CAS utilization differed in absolute value by at most Λ, where $\Lambda = 19\%$. Substantial (effectively infinite) penalties are used to enforce compliance with a constraint whenever compliance is possible and to minimize the extent of deviation from a constraint whenever strict compliance is not possible. This substantial penalty used a 'penalty function,' a continuous function that is zero if the constraint is respected and rises rapidly as the magnitude of the violation of the constraint increases; see Avriel (1976) for discussion of penalty functions and see Rosenbaum (2010, §8.4) for discussion of the use of penalty functions in matching.

The choice of Λ depends on the structure of the data. As Λ is increased, it is more difficult to find suitable pairs who have similar covariates. The covariance structure and distribution of the covariates in the data set, as well as the covariance of the instrument with the covariates, largely determine what values of Λ are possible. The selection of Λ occurs before we observe treatment selection or the outcomes, so the research may construct several matches using different values of Λ until a suitable match is found. In this paper we used this ad hoc approach to find a suitable match, one which maximized separation in the instrument, while keep covariates similar between the group, while also not removing an excessive amount of the observations from the analysis. These importance of these tradeoffs are driven by the specifics of the problem being analyzed. Further research is required to formalize a structure for determining optimal values for Λ.

See Table 1 to see prematch differences between HRRs and Table 2 to see postmatch differences. The tables summarize match quality by showing means and absolute standardized differences in means, that is, the absolute value of the difference in means divided by the standard deviation before matching. We started with 306 HRRs and constructed 76 pairs of HRRs. By using sinks (Lu et al. 2001) we excluded from our study some HRRs which were quite different from other HRRs. By excluding some HRRs we improved the overall quality of the matches between those HRRs with high CAS usage and those HRRs with low CAS usage. Once we have constructed pairs of HRRs which are similar in covariates, but dissimilar in CAS usage, we can now move on to patient-level matching.

After the first stage, we have a list of HRR pairs. Within a given pair, one HRR has higher usage of CAS and the other HRR within the pair has lower levels of CAS usage. For the second stage of our analysis we look at people within the HRRs. First we select a given pair of HRRs. For example, in the first stage the algorithm matched San Francisco and San Luis Obispo. San Francisco has high usage of CAS and San Luis Obispo has low usage. We now consider patients treated in San Francisco as being randomly encouraged to take the CAS

and those in San Luis Obispo as encouraged to take CEA. This is now a bipartite matching problem, matching patients treated in San Francisco to patients treated in San Luis Obispo. See tables 3 and 4 for results. We used the package optmatch in R to perform this matching (Hansen and Klopfer 2006). We summarized 27 covariates by calculating the Mahalanobis distance and using this to populate the discrepancy matrix for the function fullmatch(). Within any given HRR we allowed patients to be matched to sinks in order to improve the quality of the covariate balance between the encouraged and unencouraged groups. Out of a population of approximately 325,000 patients, our analysis was run on 85,284 patients because we were able to obtain a good study design which had (1) good covariate balance between the groups, as measured by the standardized difference column in Table 4 and (2) modest separation in the instrument between the encouraged and unencouraged groups.

6 Analyzing the near/far matching design

6.1 Notation

We follow the notation and motivation from Baiocchi et al (2010).

There are I matched pairs i = 1,…, I, with 2 subjects, j = 1, 2, one encouraged subject and one unencouraged, or 2I subjects in total. If the j^{th} subject in pair i receives the encouragement, write $Z_{ij} = 1$, whereas if this subject receives the control, write $Z_{ij} = 0$, so $1 = Z_{i1} + Z_{i2}$ for i = 1,…, I. In our study, the matched pairs consist of one patient from a high-CAS HRR, the other from a low-CAS HRR.

The matched pairs were formed by matching for an observed covariate x_{ij}, but may have failed to control an unobserved covariate u_{ij}; that is, $x_{ij} = x_{ik}$ for all i, j, k, but possibly $u_{ij} \neq u_{ik}$. This structure is in preparation for the inevitable comment or concern that the pairs in Table 1 look similar in terms of the variables in Table 1, but the table omits the specific covariate u_{ij} which might bias the comparison. Write $\mathbf{u} = (u_{11}, u_{12},\ldots,u_{I2})^T$ for the 2I-dimensional vector.

For any outcome, each subject has two potential responses, one seen under encouragement, $Z_{ij} = 1$, the other seen under unencouragement, $Z_{ij} = 0$; see Neyman (1923) and Rubin (1974). In our analysis, speaking in this way of two potential responses entails imagining that a patient ij who lived either in a low-CAS HRR (Zij = 0) or in a high-CAS HRR (Zij = 1) might instead have lived in the opposite circumstances. Here, there are two responses, (r_{Tij}, r_{Cij}) and (d_{Tij}, d_{Cij}) where r_{Tij} and d_{Tij} are observed from j^{th} subject in pair i under treatment, $Z_{ij} = 1$, while r_{Cij} and d_{Cij} are observed from this subject under control, $Z_{ij} = 0$. In our example, (r_{Tij}, r_{Cij}) indicates death in the 90 days following intervention, 1 for dead, 0 for alive, and (d_{Tij}, d_{Cij}) indicates whether the patient was treated with CAS, 1 if yes,

0 if no. For instance, if $(r_{Tij}, r_{Cij}) = (0; 1)$ with $(d_{Tij}, d_{Cij}) = (1; 0)$ then: (i) if a patient lived in a high-CAS HRR ($Z_{ij} = 1$), he/she would be treated with CAS ($d_{Tij} = 1$) and would live ($r_{Tij} = 0$), but (ii) if the patient had lived in a low-CAS HRR ($Z_{ij} = 0$), then he/she would have been treated with CEA instead of CAS ($d_{Cij} = 0$) and he/she would have died ($r_{Cij} = 0$).

A word on notation: To maintain consistency with the prior literature we use notation which the subscripts on the potential outcomes which contain a capital "C" and "T." These do not refer to control (i.e., CEA) and treatment (i.e., CAS), but rather map onto the encouragement levels from the instrument.

The effects of the treatment on a subject, $r_{Tij} - r_{Cij}$ or $d_{Tij} - d_{Cij}$, are not observed for any subject; that is, each patient received treatment in either a high or a low CAS HRR, and the outcome under the opposite circumstance is not observed. However, $R_{ij} = Z_{ij}r_{Tij} + (1 - Z_{ij})r_{Cij}$, $D_{ij} = Z_{ij}d_{Tij} + (1 - Z_{ij})\, d_{Cij}$ and Z_{ij} are observed from every subject.

Fisher's sharp null hypothesis of no treatment effect on (r_{Tij}, r_{Cij}) asserts that $H_0 : r_{Tij} = r_{Cij}$, for i = 1,…, I, j = 1, 2. In our example, this says that receiving treatment in a low-CAS HRR does not change the outcome compared to if the patient had received care in a high-CAS HRR, even if where the patient received care changes which kind of treatment the patient receives. If Fisher's null hypothesis were plausible, it would be difficult to argue that CAS and CEA produce different outcomes.

The exclusion restriction asserts that $d_{Tij} = d_{Cij}$ implies $r_{Tij} = r_{Cij}$, see Angrist, Imbens and Rubin (1996). In our example, the exclusion restriction says that patient outcomes are only affected by receiving care in a high-CAS HRR if receiving care in a high-CAS HRR changes the type of treatment the patient receives. This assumption may be dubious if we believe that there is a benefit to treating more patients; perhaps the surgeons become more skilled at performing the procedure meaning receiving CAS in a high-CAS region is different than receiving CAS in a low-CAS region. This is an important challenge to this study. If the analysis we are presenting was more than for illustrative purposes the discussion of the exclusion restriction would need to be carefully considered.

A patient with $(d_{Tij}, d_{Cij}) = (1, 0)$ is said to be a complier, in the sense that he/she would receive CAS if he lived in a high-CAS HRR ($d_{Cij} = 0$), but he/she would receive CEA if he/she lived in a low-CAS HRR ($d_{Tij} = 1$).

6.2 The Effect Ratio

The effect ratio, λ, is the parameter

$$\lambda = \frac{\sum_{i=1}^{I}\sum_{j=1}^{2}(r_{Tij} - r_{Cij})}{\sum_{i=1}^{I}\sum_{j=1}^{2}(d_{Tij} - d_{Cij})},$$

where it is implicitly assumed that the instrument does influence the treatment, $0 \neq \sum_{i=1}^{I}\sum_{j=1}^{2}(d_{Tij} - d_{Cij})$. Because (r_{Tij}, r_{Cij}) and (d_{Tij}, d_{Cij}) are not jointly observed, λ cannot be calculated from observable data so inference is required. Notice that under Fisher's sharp null of no effect, $H_0: r_{Tij} = r_{Cij}$ for all individuals *ij*, implies that $\lambda = 0$.

The effect ratio is the ratio of two average treatments effects. In a paired, randomized experiment the mean of the treated-minus-control difference provides unbiased estimates of numerator and denominator effects separately, and under mild conditions as $I \to \infty$, the ratio of these unbiased estimates is consistent for λ. The effect ratio measures the relative magnitude of two treatment effects, here the effect HRR treatment preferences on mortality compared to its effect on what treatment the patients receive. For instance, if $\lambda = 1/100$ then for every hundred patients who would have received CAS if they had sought treatment in a high-CAS region, but received CEA because they sought care in a low –CAS region, there is one additional patient death. As discussed by Angrist, Imbens and Rubin (1996), with assumptions such as the exclusion restriction and monotonicity, λ would be the average increase in mortality caused by treating with CAS among compliers, that is, patients with (d_{Tij}, d_{Cij}) = (1, 0).

6.3 Inference About an Effect Ratio

Consider the null hypothesis $H_0: \lambda = \lambda_0$. Following Baiocchi et al (2010) we can use the following test statistics for this null.

$$T(\lambda_0) = \frac{1}{I}\sum_{i=1}^{I}\left\{\sum_{j=1}^{2} Z_{ij}\left(R_{ij} - \lambda_0 D_{ij}\right) - \sum_{j=1}^{2}\left(1 - Z_{ij}\right)\left(R_{ij} - \lambda_0 D_{ij}\right)\right\} = \frac{1}{I}\sum_{i=1}^{I} V_i(\lambda_0)$$

where, because $R_{ij} - \lambda_0 D_{ij} = r_{Tij} - \lambda_0 d_{Tij}$ when $Z_{ij} = 1$ and $R_{ij} - \lambda_0 D_{ij} = r_{Cij} - \lambda_0 d_{Cij}$ if $Z_{ij} = 0$. Thus we may write

$$V_i(\lambda_0) = \sum_{j=1}^{2} Z_{ij}\left(r_{Tij} - \lambda_0 d_{Tij}\right) - \sum_{j=1}^{2}\left(1 - Z_{ij}\right)\left(r_{Cij} - \lambda_0 d_{Cij}\right).$$

Also define

$$S^2(\lambda_0) = \frac{1}{I(I-1)}\sum_{i=1}^{I}\{V_i(\lambda_0) - T(\lambda_0)\}^2.$$

From Baiocchi et al (2010) we know that if the instrument has indeed been randomly assigned then for large I the hypothesis $H_0: \lambda = \lambda_0$ can be tested by comparing $T(\lambda_0)/S(\lambda_0)$ to the standard Normal cumulative distribution.

7 Results

The quality of the matching at the individual level is summarized in Table 3. Once we have constructed pair matches the analysis is executed as outlined in section 6.3. The point estimate for death within 90 days of treatment is for an increase in the rate of death of 2.21% for the compliers in the study if they were switched from CEA to CAS. The confidence interval is (-0.37%, 4.48%). The width of the confidence interval is largely driven by the weak instrument we obtained in this example. Though we were able to force 1.23 units of standardized difference in the instrument in the HRR matching (see Table 2), the actual difference in CAS utilization once we matched on the individual level only had a separation of 0.21 units of standardized difference (see Table 3).

8 Discussion

Near/far matching is a study design approach to instrumental variables. It combines the relative simplicity of propensity score matching with the ability of an IV to address unobserved selection.

The complexity of near/far matching is in the study design portion of the procedure; that is, most time and consideration is spent on the matching. The statistical analysis is quite simple and is analogous to a paired t-test. More complex IV methods exist which are capable of estimating treatment effects for settings with binary outcomes. These methods tend to require maximization of a complex likelihood, which can be computationally taxing. Additionally, the recommendation for estimating standard errors is usually to use a bootstrapping approach, which requires additional iterations of an already complex maximization step. In contrast, near/far matching simply requires inverting the hypothesis test in the standard way in order to form a confidence interval.

One more advantage of a simple statistical procedure is researchers can construct a sensitivity analysis to help quantify the impact of violations of the assumptions of the analysis. For an introduction to sensitivity analysis see Rosenbaum (2002, §4.4-5). For discussion of alternative methods of sensitivity analysis see Imbens (2003), Robins, Rotnitzky and Scharfstein (1999) and Small (2007). Baiocchi et al. (2010) provides a sensitivity analysis for the situation where, even post near/far matching, some set of unobserved covariates is still unbalanced between the two groups.

A concern with near/far matching is that during the matching phase it is often necessary to remove observations from the analysis which are "unsuitable" – that is, their covariates are dissimilar from most other observations and/or they are not sufficiently dissimilar in their levels of encouragement from the instrument. In our example 50% of the HRRs were removed from the analysis in order to create as much separation as possible in the instrument while still maintaining good balance in the covariates. Over and above the

usual identification issue with IV estimators, which is that they are estimating a treatment effect on just the compliers, near/far matching limits the population by excluding observations from the analysis. This is a common study design problem. Consider a randomized controlled study; it is common for a study to have a list of exclusion criteria which precludes portions of the population from participating in the study and therefore narrows the population for which the estimate is valid. As in a clinical trial, the researcher must weigh the tradeoffs between having greater internal validity versus the benefits of the generalizability of the trial.

References

Angrist, J.: Estimation of Limited Dependent Variable Models with Dummy Endogenous Regressors: Simple Strategies for Empirical practice, JBES 2001; 19, 2-16

Angrist, J. D., Imbens, G. W., and Rubin, D. B: Identification of causal effects using instrumental variables (with Discussion)," Journal of the American Statistical Association, 1996; 91, 444-455.

Angrist, J., Pischke, J.:Mostly Harmless Econometrics. Princeton University Press. 2009

Avriel, M.; Nonlinear programming, New Jersey: Prentice Hall. 1976

Baiocchi, M., Small, D., Lorch, S. and Rosenbaum, P.; Building a stronger instrument in an observational study of perinatal care for premature infants. Journal of the American Statistical Association, 2010;105, 1285-1296.

Barnett HJ, Eliasziw M, Meldrum HE, Taylor DW. Do the facts and figures warrant a 10-fold increase in the performance of carotid endarterectomy on asymptomatic patients? Neurology. 1996;46(3):603-8.

Beohar N, Davidson CJ, Kip KE, Goodreau L, Vlachos HA, Meyers SN, et al. Outcomes and complications associated with off-label and untested use of drug-eluting stents. JAMA. 2007;297(18):1992-2000.

Bound, J., Jaeger, D. A., Baker, R. M.; Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak, Journal of the American Statistical Association, 1995; 90, 443-450.

Cai, B., Small, D. and Ten Have, T.: Two-stage instrumental variable methods for estimating the causal odds ratio: analysis of bias. *Statistics in Medicine*, 2011; 30, 1809-1824.

Derigs, U.; "Solving nonbipartite matching problems by shortest path techniques, Annals of Operations Research, 1988;13, 225-261.

Dodick DW, Meissner I, Meyer FB, Cloft HJ. Evaluation and management of asymptomatic carotid artery stenosis. Mayo Clin Proc. 2004;79(7):937-44.

Hansen, B.B. and Klopfer, S.O. (2006) Optimal full matching and related designs via network flows, JCGS 1988;15 609-627.

Holland, P. W.; Causal inference, path analysis, and recursive structural equations models. In Sociological Methodology, Volume 18, C. C. Clogg (ed), 449-484. Washington, D.C.: American Sociological Association

Imbens, G. W.; Sensitivity to exogeneity assumptions in program evaluation, American Economic Review 2003;93, 126-132.

Lloyd-Jones D, Adams R, Carnethon M, De Simone G, Ferguson TB, Flegal K, et al. Heart disease and stroke statistics--2009 update: a report from the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. Circulation. 2009;119(3):480-6.

Lu, B., Zanutto, E., Hornik, R. and Rosenbaum, P. R.; Matching with doses in an observational study of a media campaign against drug abuse,"Journal of the American Statistical Association, 2001;96, 1245-1253.

Mas JL, Chatellier G, Beyssen B, Branchereau A, Moulin T, Becquemin JP, et al. Endarterectomy versus stenting in patients with symptomatic severe carotid stenosis. N Engl J Med. 2006;355(16):1660-71.

Mas JL, Trinquart L, Leys D, Albucher JF, Rousseau H, Viguier A, et al. Endarterectomy Versus Angioplasty in Patients with Symptomatic Severe Carotid Stenosis (EVA-3S) trial: results up to 4 years from a randomised, multicentre trial. Lancet Neurol. 2008;7(10):885-92.

Moses JW, Leon MB, Popma JJ, Fitzgerald PJ, Holmes DR, O'Shaughnessy C, et al. Sirolimus-eluting stents versus standard stents in patients with stenosis in a native coronary artery. N Engl J Med. 2003;349(14):1315-23.

Park SJ, Shim WH, Ho DS, Raizner AE, Park SW, Hong MK, et al. A paclitaxel-eluting stent for the prevention of coronary restenosis. N Engl J Med. 2003;348(16):1537-45.

Powers, D. E. and Swinton, S. S.; Effects of self-study for coachable test item types. Journal of Educational Measurement 1984;76, 266-278

Robins, J. M., Rotnitzky, A. & Scharfstein, D.; Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference, In *Statistical Models in Epidemiology*, Ed. E. Halloran and D. Berry, pp. 1-94. NY: Springer. 1999

Rosenbaum, P. R. Observational Studies (Second Edition). New York: Springer-Verlag. 2002

Rosenbaum, P.R. Design of Observational Studies, New York: Springer. 2010

Rosenbaum, P.:The Consequences of Adjustment for a Concomitant Variable That Has Been Affected by the Treatment. *Journal of the Royal Statistical Society. Series A,* 1984; Vol. 147, No. 5, pp. 656-666

Rosenbaum, P. R., and Rubin, D. B.,; The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika 1983;70, 41–55.

Terza, J., Basu, A., and Rathouz, P.: Two-Stage Residual Inclusion Estimation: Addressing Endogeneity in Health Econometric Modeling J Health Econ. 2008 May; 27(3): 531–543.

Yadav JS, Wholey MH, Kuntz RE, Fayad P, Katzen BT, Mishkel GJ, et al. Protected carotid-artery stenting versus endarterectomy in high-risk patients. N Engl J Med. 2004;351(15):1493-501.

	1st Quartile Means	4th Quartile Means	|St-dif|
	Instrument		
% CAS utilization	1.8%	23.8%	2.32
	Covariates		
% age over 65 in HRR	14.5%	12.3%	0.66
% white in HRR	84.1%	75.0%	0.71
% urban in HRR	63.9%	75.8%	0.64
Median income in HRR	38,683	43,212	0.49
Mean education	12.8	12.8	0.27
Medicare spend	11,463	13,935	0.76
Academic beds per 1000	1.44	4.00	0.64
2004 - death within 90 day	2.46%	2.00%	0.35

Table 1: Prematch HRRs. Comparing the means of the 76 HRRs in the lowest quartile of the instrument (lowest rates of CAS utilization) to the 76 HRRs in the highest quartile. Median income in the HRR and the average Medicare spend in the last 6 months of the patients' month (both measured in dollars). Mean education in the HRR is measured as an ordinal variable.

	Unencouraged Means	Encouraged Means	|St-dif|
	Instrument		
% CAS utilization	2.9%	14.5%	1.23
	Covariates		
% age over 65 in HRR	13.6%	13.2%	0.12
% white in HRR	81.7%	79.8%	0.15
% urban in HRR	68.3%	71.6%	0.18
Median income in HRR	40,792	41,862	0.11
Mean education	12.9	12.8	0.06
Medicare spend	12,470	12,861	0.12
Academic beds per 1000	2.23	2.97	0.18
2004 - death within 90 day	2.36%	2.26%	0.08

Table 2: Postmatching HRRs. The "Encouraged Means" column summarizes the 76 HRRs within a match which had the higher rate of CAS utilization. In contrast to Table 1, the standardized differences for the covariates show the two groups have comparable means.

Matches 42,642	Type	Encouraged Mean	Unencouraged Mean	|St-dif|
CAS utilization (1/0)	Instrument	0.13	0.06	0.21
Age (years)	Covariates	74.56	74.58	0.00
Female (1/0)		0.42	0.42	0.00
CHF (1/0)		0.09	0.09	0.00
Cardiac arrhythmia (1/0)		0.18	0.18	0.00
Cardiac valvular disease (1/0)		0.08	0.08	0.01
Pulmonary circulation disease (1/0)		0.01	0.01	0.00
Peripheral vascular disease (1/0)		0.25	0.24	0.02
Paralysis (1/0)		0.02	0.03	0.01
Neurologic disorder (1/0)		0.02	0.02	0.01
Chronic pulmonary disease (1/0)		0.25	0.24	0.02
Diabetes uncomplicated (1/0)		0.29	0.29	0.00
Diabetes w/complication (1/0)		0.04	0.04	0.01
Hypothyroidism (1/0)		0.10	0.10	0.01
Renal disease (1/0)		0.08	0.08	0.01
Liver disease (1/0)		0.00	0.00	0.01
AIDS (1/0)		0.00	0.00	0.01
Lymphoma (1/0)		0.00	0.00	0.01
Metastatic cancer (1/0)		0.00	0.00	0.00
Tumor no met (1/0)		0.02	0.02	0.00
Rheumatoid arthritis (1/0)		0.02	0.02	0.01
Coagulopathy (1/0)		0.01	0.01	0.00
Obesity (1/0)		0.05	0.05	0.00
Weight loss (1/0)		0.01	0.01	0.00
Depression (1/0)		0.04	0.05	0.01
Hypertension (1/0)		0.83	0.83	0.01
Acute myocardial infarction (1/0)		0.03	0.03	0.00
Coronary artery disease, no AMI (1/0)		0.51	0.50	0.01
Death within 90 days (1/0)	Outcome	0.0202	0.0189	0.01

Table 3: Individual level matching. The "Encouraged Mean" column summarizes the means of the 42,642 individuals within a pair who were treated in a high-CAS HRR and were matched to an individual wish similar observed covariates who received treatment in a low-CAS HRR.

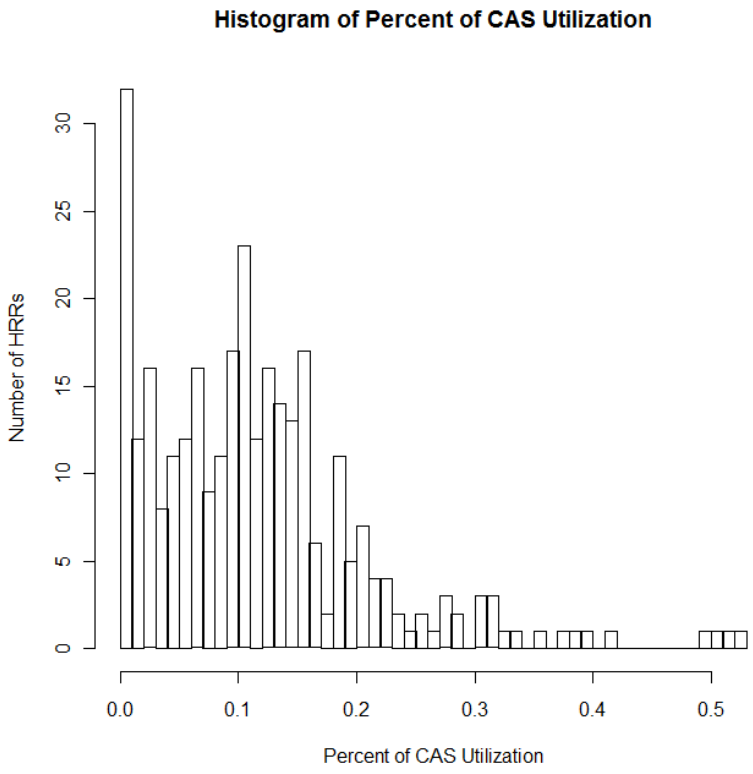

Figure 1: A histogram of the percent of CAS utilization in the 306 Hospital Referral Regions. The median value is 10.2%. The mean is 11.6%. The interquartile range goes from 4.9% up to 15.4%. Thirty-two of the HRRs (slightly more than 10% of the HRRs) had a CAS utilization of 1% or less.

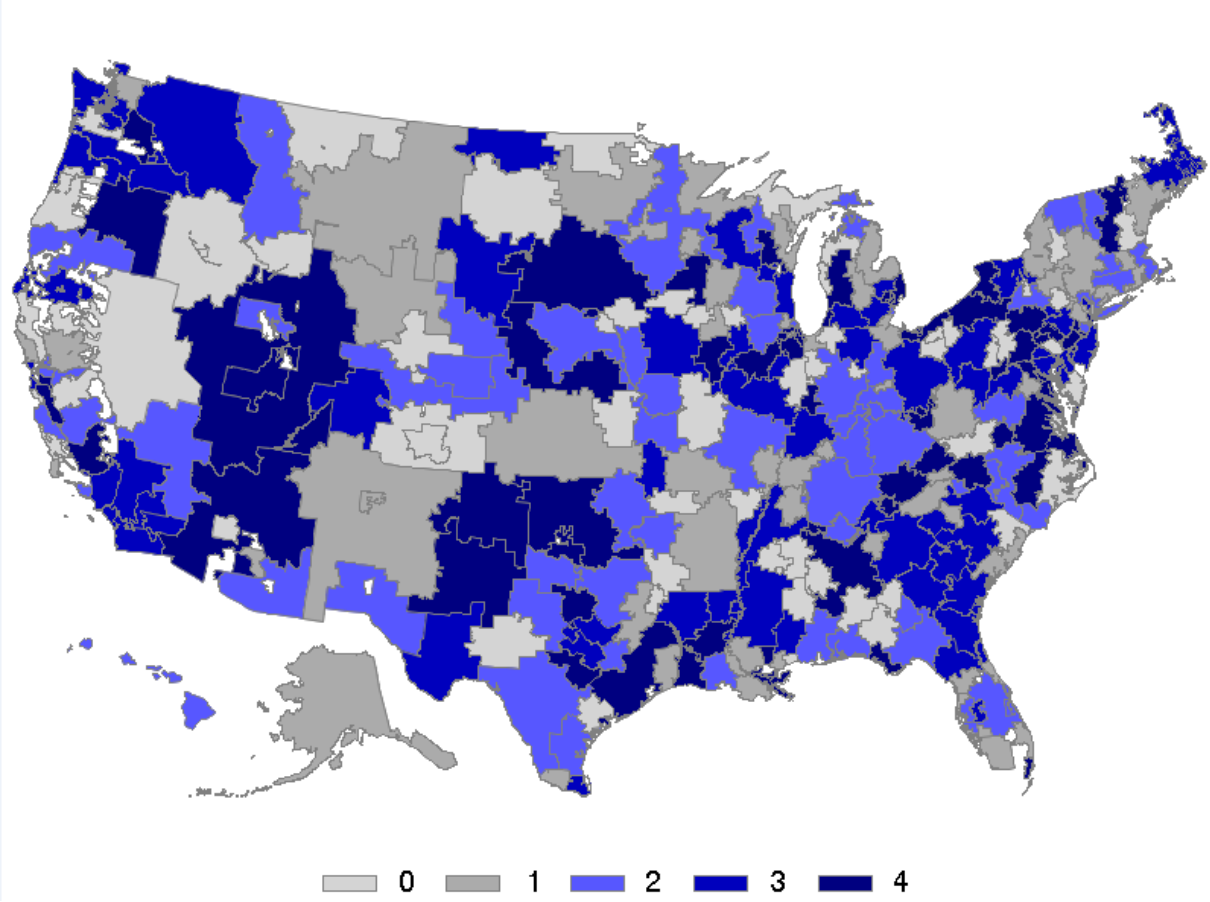

Figure 2: This is a heath map of CAS utilization in the Health Referral Regions (HRRs). To aid visualization, the HRRs have been color coded by quintiles with the HRRs with the highest rates of CAS utilization colored a deep blue and those HRRs with the lowest rates of CAS utilization colored a light gray.