Outcome-free Design of Observational Studies: Peer Influence on Smoking

Sophie Langenskiöld and Donald B. Rubin[*,†]

SUMMARY

For estimating causal effects of treatments, randomized experiments are appropriately considered the gold standard, although they are often infeasible for a variety of reasons. Nevertheless, nonrandomized studies can and should be designed to approximate randomized experiments by using only background information to create subgroups of similar treated and control units, where "similar" here refers to their distributions of background variables. This activity should be conducted without access to any outcome data to assure the objectivity of the design. In many situations, these goals can be accomplished using propensity score methods, as illustrated here in the context of a study on whether nonsmoking Harvard freshmen are influenced by their smoking peers. In that study, propensity score methods were used to create matched groups of treated units (rooming with at least one smoker) and control units (rooming with only non-smokers) who are at least as similar with respect to their distributions of observed background characteristics as if they had been randomized, thereby approximating a randomized experiment with respect to the observed covariates.

KEY WORDS: assignment mechanism, causal inference, nonrandomized studies, objective design, observational studies, peer influence, propensity scores, randomized experiments, Rubin Causal Model, tobacco smoking behavior

[*] Correspondence to: Donald B. Rubin, Department of Statistics, Harvard University, 1 Oxford Street, 7th Floor, Cambridge, MA 02138, U.S.A.
[†] E-mail: rubin@stat.harvard.edu

1. OUR PERSPECTIVE ON INFERENCE FOR CAUSAL EFFECTS

The perspective on causal inference used in this article is called the "Rubin Causal Model (RCM)" (Holland, 1986) for a sequence of papers written in the 1970s (Rubin: 1974, 1975, 1976, 1977a, 1978a, 1979a, 1980); see also Imbens and Rubin (2007a**)** for an encyclopedia-entry length discussion of the RCM, Rubin (2007a) for a chapter length one, and Imbens and Rubin (2007b) for a full text-length treatment from this perspective. The RCM's first two parts are essential and have together been called the "potential outcomes with assignment mechanism" perspective (Rubin, 1990a, p 476). The RCM's third part is optional and involves extensions to include Bayesian inference (Rubin, 1978a).

The first part of the RCM is conceptual and defines causal effects as comparisons of potential outcomes under different treatments on a common set of units. It is critical that this first part be carefully formulated if causal inferences are to provide meaningful guidance for practice.

The focus here is on the second part of the RCM, which concerns the assignment mechanism and the design of studies for causal effects. By "design" we mean all contemplating, collecting, organizing, and analyzing of data that takes place prior to seeing any outcome data. For example, design includes analyses of covariate data used to create matched treated-control samples or to create subclasses with similar covariate distributions for the treated and control subsamples, but any analysis that requires outcome data in order to be implemented is not part of design. The careful implementation of this design step is absolutely essential for drawing objective inferences for causal effects in practice, whether in randomized experiments or nonrandomized observational studies, yet is often basically ignored with nonrandomized studies.

This perspective emphasizing the importance of design is also advocated in Rubin (2007b) using an example from the US tobacco legislation.

The third part of the RCM is briefly discussed here as well. It extends the analytic approach of the first two parts to include a joint probability model on all the potential outcomes and the covariates. This model can: (1) account for partial failures of matching or sub-classification to achieve samples with similar treated and control units with respect to their covariate distributions; (2) improve the precision of estimation by accounting for relationships between outcomes and the covariates; and (3) facilitate generalizing causal inference to contexts other than the one of the current experiment.

As will be seen in this paper, the perspective of the RCM allows the researcher first to estimate repeatedly the model for the assignment mechanism until this model fits the covariate data appropriately; this modeling never uses the outcome data. As will be seen in a subsequent paper, the RCM perspective also allows the researcher to assess different statistical models relating some noncritical outcome data to covariate data before seeing any critical outcome data. Thus, using the RCM approach helps researchers prepare for critical outcome analyses by allowing them to choose the statistical models for the final analyses without repeatedly trying different models relating critical outcome data to covariate data. In the approach traditionally followed by researchers in social and biomedical science, the assignment mechanism and the science are implicitly modeled simultaneously, for example, using ordinary least squares or logistic regression models. Typically many such models are examined sequentially, which invalidates the

standard interpretation of p-values, confidence intervals, and other formal statistical properties of procedures.

In Section 2, the RCM is briefly reviewed, thereby describing the theoretical underpinnings that make it mandatory to model separately the assignment mechanism and the science. The importance of replicating the objectivity achieved in a randomized trial through careful modeling of the assignment mechanism is described in Section 3. In Section 4, the perspective of Section 3 is illustrated using our example of peer influence on smoking. Section 5 concludes.

2. BRIEF REVIEW OF THE RCM

2.1 Part one: Units, treatments, potential outcomes

Three basic concepts define causal effects in the RCM. A unit is a physical object, for example, a person, at a particular place and point in time, say time t. An active treatment is an action or intervention that can be initiated on the unit at time t (e.g., a job training program, exposure to roommates who smoke); if the active treatment is withheld, we say that the unit has been exposed to the control treatment. Associated with that unit are two potential outcomes at a future point in time, say, time $t^* > t$: the value of some outcome measurements Y (e.g., salary, smoking behavior) if the active treatment is given at t, say $Y(1)$, and the value of Y at the same future point in time if the control treatment is given at t, say $Y(0)$. The causal effect of the active treatment versus the control treatment for the i^{th} (i=1,…,N) unit is defined to be the comparison of the treatment and control potential outcomes at t^* (e.g., their difference $Y_i(1)$- $Y_i(0)$), and this

comparison is typically called the treatment effect on the ith unit. The times t can vary from unit to unit in a population of N units, but typically the intervals, $t^* - t$, are constant across the units.

The full set of potential outcomes comprises all values of the outcome Y that could be observed in some real or hypothetical experiment comparing the active treatment to the control treatment in the population of N units. Under the "Stable Unit-Treatment Value Assumption (SUTVA)" (Rubin: 1980, 1990a), the full set of potential outcomes for two treatments and the population of N units can be represented by an array with N rows and two columns, one for $Y(1)$ and one for $Y(0)$. "The fundamental problem facing inference for causal effects" (Rubin, 1978a; Section 2.4) is that only one of the potential outcomes for each unit can ever be observed. In contrast to the collection of potential outcomes, covariates are variables X that take the same value no matter which treatment is applied to the units, such as quantities measured before treatments are assigned (e.g., pre-treatment work history, parental smoking behavior), and so can be fully observed.

The possible values of all of these variables can be arranged in units by variables array consisting of: covariates, X; potential outcomes under the control treatment, $Y(0)$; and potential outcomes under the active treatment, $Y(1)$. This array of values is the object of causal inference and is called "the science". A causal effect is a comparison of active and control treatments' potential outcomes on a common set of units; for example, the mean $Y_i(1)$ verses the mean $Y_i(0)$ for all units, or the 25^{th} percentile of log $Y_i(1)$ verses the 25^{th} percentile of log $Y_i(0)$ for those units who were unemployed before treatment assignment occurred, or male students whose parents smoked when the student was in high school.

This first part of the RCM is conceptual and typically should be conducted before seeing any data. It forces the conceptualization of causal questions in terms of real or hypothetical manipulations: "no causation without manipulation" (Rubin, 1975, p. 238). The formal use of potential outcomes to define unit-level causal effects is due to Neyman (1923) in the context of randomized experiments, and was a marvelously clarifying contribution, but evidently this notation was not formally extended to non-randomized settings until Rubin (1974), as has been discussed now in a variety of places (Rubin, 1990a; Imbens and Rubin 2007a; Imbens and Rubin 2007b; Rubin 2005, 2007a, 2007b).

For nonrandomized observational studies, prior to 1974 the standard was to use "observed outcome" notation for causal inference, where letting W be the column vector for the treatment assignments for the units ($W_i = 1$ if treated, $W_i = 0$ if control), observed outcome notation replaces the potential outcomes $[Y(0), Y(1)]$ with Y_{obs}, where the i^{th} component of Y_{obs} is defined to be:

$$(1) \qquad Y_{obs,i} = W_i Y_i(1) + (1-W_i) Y_i(0).$$

Moreover, observed outcome notation drops the clarifying subscript *obs* on $Y_{obs,i}$ and simply uses y_i in its place. This notification is inadequate in general, and can lead to serious confusion, as discussed in Holland and Rubin (1983) on Lord's paradox, and Rubin (2005) on Fisher's advice on the use of the analysis of covariance to estimate "direct" causal effects in randomized experiments.

2.2 Part 2: The assignment mechanism

The second part of the RCM is the specification of an assignment mechanism, which describes the reasons for the missing and observed values of $Y(0)$ and $Y(1)$ using a probability model for W given the science:

$$\Pr(W|X, Y(0), Y(1)). \tag{2}$$

This general formulation arose first in Rubin (1975), but special cases were much discussed prior to that. For example, randomized experiments (Neyman, 1923, 1990; Fisher, 1925) are "unconfounded" (Rubin, 1975),

$$\Pr(W|X, Y(0), Y(1)) = \Pr(W|X), \tag{3}$$

and they are "probabilistic" because the unit level probabilities of being assigned active treatment,

$$e_i \equiv \Pr(W_i = 1|X_i), \tag{4}$$

are bounded between 0 and 1:

$$0 < e_i < 1. \tag{5}$$

When the assignment mechanism is both probabilistic and unconfounded, the right-hand side of (3) can often be written as proportional to the product of the unit level probabilities, e_i, which are called propensity scores:

$$(6) \qquad \Pr(W \mid X, Y(0), Y(1)) \propto \prod_{i=1}^{N} e_1.$$

More generally (6) may hold only for those assignments W that have positive probability. For example, consider a completely randomized experiment with $N/2$ units exposed to treatment and $N/2$ exposed to control. Then, $e_i = 0.5$ and (6) holds for W such that $\Sigma W_i = N/2$, but all W such that $\Sigma W_i \neq N/2$ have probability zero.

The term "propensity scores" for the unit level probabilities in (4) was coined in Rosenbaum and Rubin (1983), where an assignment mechanism satisfying (3) - (6) is called "strongly ignorable," because (3) and (5) are stronger than the condition for an "ignorable" mechanism, coined in Rubin (1976, 1978a), which allowed possible dependence on observed values of the potential outcomes, such as in a sequential experiment:

$$(7) \qquad \Pr(W|X, Y(0), Y(1)) = \Pr(W|X, Y_{obs}) .$$

Confounded but ignorable designs arise in sequential experiments when current treatment decisions involve previously observed outcomes.

Until Rubin (1975), randomized experiments were not defined using equations (3) – (5) which explicitly show such experiments freedom from any dependence on observed or missing potential outcomes.

Other special versions of assignment mechanisms were also discussed prior to 1975, but without the benefit of explicit equations for the assignment mechanism showing possible dependence on the potential outcomes. Of note, in economics, Roy (1951) described, without equations or notation, "self-optimizing" behavior where each unit chooses the treatment with the optimal outcome (e.g., W_i=1 if $Y_i(0)<Y_i(1)$, and W_i=0 otherwise). And another well-known example from economics is Haavelmo's (1944) formulation of supply and demand behavior, but these and other formulations in economics did not use the explicit notation of an assignment mechanism, nor did they have associated methods of statistical inference for causal effects based on the assignment mechanism. Instead, "regression" models were used to predict $Y_{obs,i}$ from X_i and W_i, with possible restrictions on some regression coefficients and/or error terms, where particular regression coefficients (e.g., the coefficient of W_i or of interactions with W_i) were interpreted as causal effects.

As seen in expressions (3) – (6), the propensity scores collectively are a basic ingredient of unconfounded assignment mechanisms, and their use for objectively designing observational studies will be illustrated here in an example studying the peer influence of smoking on college freshmen at Harvard College. In this example, we work very hard to understand the actual assignment mechanism, and thereby to approximate or replicate the hypothetical randomized experiment that lead to the observed data. We think we have been successful in this example.

In some cases, in social and biomedical contexts, it can be relatively straightforward, first, to understand the variables that influence the assignment to treatment conditions (e.g., that explain the assignment or decision to live with smoking versus nonsmoking roommates), and, second, to believe that the assignment is essentially randomized within groups defined by these variables (i.e., is unconfounded as in (3)). In such cases, randomization-based modes of inference, due to Fisher (1925) and Neyman (1923; Rubin, 1990a) can be used to draw causal inferences, at least in relatively simple situations. These modes require only a model for the assignment mechanism. In other modes of inference, i.e., model-based modes of inference that will be briefly described in the next section, the science is also modeled. This critical point means that, within the RCM, the assignment mechanism, and appropriately modeling it, is fundamental for causal inference, more fundamental than formulating a model for the science.

2.3 Part 3: Full probability model on the science

The third and final part of the RCM is optional, and it creates a model specification for the science. The science is thus considered as a collection of random variables in model-based modes of inferences. In assignment-based modes of inferences, however, the science is considered to be a collection of fixed but partially unobserved values, with the treatment indicator W being the only random variable. Therefore, $Y_{obs,i}$, defined in equation (1) as a function of W_i, is a random variable in the assignment-based modes of causal inference only because W_i is a random variable, not because any part of the science is a random variable. In

contrast, in this third part of the RCM, the science itself is considered to be a matrix random variable with a joint distribution:

$$(8) \qquad \Pr(X, Y(0), Y(1)).$$

This model on the science completes the full model specification of all observable quantities, and so is Bayesian, as defined first in Rubin (1975) and further developed in Rubin (1978a), as well as in other places, such as Imbens and Rubin (1997). There are non-Bayesian versions of model-based methods as well (Rubin, 1990b), but these are fundamentally not as principled or as coherent as the Bayesian approach.

The model for the science, when combined with the model for the assignment mechanism and the observed data, leads to the posterior predictive distribution of the unknown components of the science, that is, of the missing potential outcomes, Y_{mis}, with i^{th} component $Y_{mis,i} = W_i\, Y_i(0) + (1\text{-}W_i)\, Y_i(1)$. It is important to realize that even in this Bayesian formulation, the model for the assignment mechanism, which is for unconfounded designs essentially a propensity score model as in expression (6), remains necessary to specify, just as in the assignment-based modes, and is not a model involving the science but remains conditional on the science.

3. OBSERVATIONAL STUDIES AS APPROXIMATIONS OF RANDOMIZED EXPERIMENTS

The choice and formulation of an assignment mechanism should be structured just as carefully in an observational study as in a randomized experiment. In an observational study we view the observed data as having arisen from a hypothetical complex randomized experiment with a lost rule for the propensity scores, whose values we will try to reconstruct. It is a basic characteristic of randomized experiments that the design phase takes place prior to seeing any outcome data, and this critical feature of randomized experiments nearly always should be duplicated in observational studies.

Propensity score methods are the observational study equivalent of complete (i.e., unrestricted) randomization in a randomized experiment. That is, these methods are intended to eliminate bias, but are not intended to increase precision. Propensity score methods can only perfectly eliminate bias when the assignment mechanism is truly unconfounded (given the observed covariates, X) and when the propensity scores are effectively known (or take on only few distinct values), whereas randomization eliminates bias due to all covariates, both observed and unobserved. Therefore, despite their great virtues, propensity score methods are not fully satisfactory replacements for randomization.

Blocking and matching on particular covariates are methods for eliminating extraneous variation due to measured covariates, whether in the context of a randomized experiment or an observational study. Blocking or matching creates treatment and control subgroups within which the distributions of observed covariates are more similar than would be expected if treatments were simply assigned to units completely at random, and thereby eliminates conditional bias

(i.e., conditional on these covariates), which becomes variance when averaged over the possible assignments in a completely randomized design.

This prescription for design in the absence of outcome variables is in conflict with typical analyses, at least as we (the authors) know them, of observational data in social and biomedical science. Rather than the outcome data, Y_{obs}, being "not in sight," Y_{obs} data are used over and over again to fit various models, try different transformations, look at results discarding influential outliers, etc. "Oh, I should have used five indicator variables for age in my linear regression rather than treating age only as a linear predictor, because that greatly improved the p-values for treatment effects!" How many reported analyses that we see in journals are "designed" a priori rather than are the result of repeated and unreported exploratory analyses, looking for a publishable result, where "publishable" can imply significant results of any size or in any direction, especially when they are likely to improve the prospect for promotion or tenure?

An analogy with the world of drug development is relevant. Would "you" buy a drug that was approved based on the results of hundreds of randomized experiments, where only the results of the single most favorable experiment were used for approval of the drug for general use? Or if the drug was approved based on dozens or hundreds of re-analyses of the same data until some analysis was found that produced a "significant" and favorable result? Probably you would be skeptical of advertisements touting the great benefits of a drug approved on the basis of such analyses.

Of course, there have been previous criticisms by social scientists of such statistically invalid approaches to the analysis of data. Notable among these is Leamer (1978) on specification searches. The emphasis there, however, is on models for the science rather than on objective design. Our objective design allows the careful researcher to use part of the data repeatedly, without biasing results towards an answer, even unconsciously.

Objectivity can be achieved in the design of observational studies, although it is typically not as easy as in randomized experiments. And of course, objectivity is not the same as finding the truth, but we believe that it is generally a necessary prerequisite for finding the truth. The key idea is to conduct the design before ever seeing any outcome data and to try to structure the data set in such a way that any future model-based adjustments will tend to give similar point estimates. This objective can often be accomplished using propensity score methods to create matched pairs (or subgroups, i.e., subclasses) of treated and control units such that across matched pairs (or within each subclass), the treated and control units have essentially the same distributions of all covariates. If this can be achieved, then essentially the same point estimates will be obtained irrespective of the specific model-based adjustments that are subsequently made, although the associated estimated precisions may well differ.

The next section illustrates this perspective using an evaluation of peer influence on freshman smoking habits at Harvard University carried out by Sophie Langenskiöld (2005), an Economics PhD student who was visiting Harvard University from Stockholm School of Economics. The project was conducted under the supervision of Donald B. Rubin with the support of Magnus Johannesson at Stockholm School of Economics.

4. AN EXAMPLE ON THE PEER INFLUENCE ON SMOKING

4.1 Overview of the study at Harvard University

A critical issue when studying individuals and the influences of their peers is the sheer number of covariates that may describe individuals' choice of peers and the possible connections they may have to the outcome variables. This situation makes it difficult to study the causal effects of peer smoking on smoking initiation in most contexts. However, the Harvard Housing Office (HHO) assigns freshmen into suites somewhat haphazardly using, at most, freshmen's answers to questions on the housing application form (HAF). We are thus in a situation where we know a lot about the assignment mechanism. However, we are not in as ideal a position as studies of peer influence at Dartmouth College, where assignment of freshmen to living arrangements is randomized (Sacerdote, 2001)

Moreover, we do not have access to the data on the HAFs forms for reasons of confidentiality; however, those data can be approximated or replicated by asking the same questions in a survey. Then, under the assumption that the survey data are an accurate reflection of the information in the HHO forms, the assignment of Harvard freshmen into suites becomes strongly ignorable, and propensity score methods can be applied in a valid way to design an observational study in this context. Our design matched the treated units with similar control units so that the distributions of their background covariates were similar, and this matching and evaluation of the resulting balance took place repeatedly without access to any outcome data. This plan for the design of an

observational study replicates many of the benefits of a randomized experiment, including its objectivity.

Our study of freshmen at Harvard College was conducted during the academic year 2003/2004. In the first week of classes, 589 freshmen were enrolled in the study. In the first interview (Interview 1), they were asked, first to replicate the answers they gave on the HAF, second to describe their own health habits 30 days before they enrolled in college, and third to recall their parents' health habits when they were still children. Approximately five weeks later, the students were contacted by email and were asked to complete the second interview's questionnaire online (Interview 2). A total of 462 freshmen responded to questions about their roommates' health habits during the last 30 days in this interview. After another five months, the students were again contacted for a final interview (Interview 3). A total of 411 freshmen described their health habits during the 30 days prior to Interview 3.

4.2 The Template Randomized Experiment for the Freshman Observational Study

In an ideal randomized study of peer influence on smoking behavior, the HHO would initially assign most freshmen to rooms (or more accurately, suites) except for a subgroup of nonsmoking freshmen, where we call this group of nonsmoking freshman the "target" group, and we call the remainder, who were initially assigned into suites , the "main" group. Many suites of the main group could accommodate one more suitemate, and these suites would be classified as either a smoking suite if at least one member was a smoker and as a nonsmoking suite if no member was a smoker. The smoking status of all entering freshmen could have been known from questions on

the HAF, but this information was not asked on the HAF. But for now assume, in this ideal randomized experiment that will serve as the template for our observational study, that the HAF had such questions in it.

At this point in time, the HHO would randomly assign each member of the target group to a suite, either a smoking suite – thereby becoming a treated unit, or a nonsmoking suite – thereby becoming a control unit. The use of separate suites for each member of the target group helps to justify SUTVA. The design could be either a completely randomized design or a randomized block design (e.g., individuals in the target group could be paired by their answers to questions on the HAF, and a random member of a pair would be selected for a smoking suite).

4.3 Our Observational Study as an Approximation of the Template Randomized Experiment

The study that we conducted differs from this ideal template randomized experiment, but we do think it approximates it well. We asked many freshmen during their first week of classes to replicate their answers to the questions that are on the HAF and to describe their pre-college smoking habits. We chose to study only one nonsmoking freshman from each suite who, as a group, would comprise the target group of freshman, thereby duplicating a feature of the template experiment that is designed to justify SUTVA. Also, we assume that the assignment mechanism that the HHO uses would be ignorable if we had the freshmen's answers to the questions on the HAF. Because at Interview 1 we obtain essentially the same data as obtained

on the HAF, we make the parallel assumption about ignorable treatment assignment in our observational study.

We need, however, to ask the freshmen in the target group about their roommates' smoking habits in order to know the constellation of smoking habits in the suites, that is, whether each member of the target group is a treated or control unit. Because we think the freshmen may not have this information during their first week of classes, we waited another five weeks before asking them these questions (Interview 2).

Thus, we are using variables collected after suite assignment both to define treatment status for the target freshmen and to approximate the HHO's assignment as ignorable. But it is possible that some of the questions asked on our version of the HAF are not proper covariates. If we were to control for such improper covariates, it could result in answers that underestimate the true effect of peer influence – that is, this plan could mistakenly force the treated and control units to achieve balance not only with respect to their pre-treatment (proper) covariates but also with respect to some of their observed outcome measurements correlated with the smoking behavior outcomes. We do not believe, however, either issue to be a concern with our study. We choose to study entering nonsmoking freshmen, because as nonsmokers, it is extremely unlikely that they could have affected the smoking behavior of their suitemates within a few weeks, if ever**.** In fact, freshmen who do not smoke at enrollment and who live with at least one smoker after five weeks in college, have obviously not succeeded in preventing their roommates from smoking because if they had succeeded, their roommates would not be regarded by them as smokers. The accuracy of the freshmen's replicated answers to the questions on the HAF is expected to be

good because the careful consideration that the freshmen were likely to have invested in filling out a form influencing something as important as their living conditions during their freshmen year probably made it easy for them to remember their answers. We can therefore argue that the assignment of nonsmoking freshmen into suites is strongly ignorable given the proper covariates that are derived from the questions on the HAF.

4.4 The Covariates and Missing Data in Them

Table 1 summarizes the covariates that we believe are proper and on which the similarity of treated and control freshmen should be judged. Simple comparisons assume, however, that there are no missing values on these proper covariates (see, Rosenbaum and Rubin, 1984; D'Agostino and Rubin, 2000), for discussion of alternative treatments of missing covariates for propensity score estimation). Because we have a small sample, we were reluctant to drop units with missing data. For this reason, we decided to multiply-impute the missing values of the proper covariates by drawing them from their respective posterior predictive distributions (Rubin, 1987, 2004). More explicitly, we replaced each missing value with ten imputed values**,** thereby reflecting the uncertainty with which the missing values are imputed; we used software described in Raghunathan et. al. (2001). The posterior predictive distributions in this software are defined sequentially using fully conditional model specifications because in most contexts the number of proper covariates and the complex relations among them makes it burdensome to define one joint distribution to be used for imputation. We imputed the missing values of the proper covariates ignoring the treatment indicator variable. This method implicitly assumes that, conditional on the fully observed covariates, treatment assignment was ignorable; the effect of this assumption is

assessed at the end of this section. Multiple imputation was originally proposed by Rubin (1977b, 1978b) and further developed in Rubin (1987) and subsequently by Rubin and many others, e.g., Meng et al. (1991), Little et al. (1991), Little et al. (1993), Schenker et al. (1996), Schafer (1997).

The causal estimate of peer influence is found by averaging the ten estimates of peer influence. Interval estimates for peer influence are found using the associated sampling variances. The total sampling variance is the sum of the within variance, which measures how uncertain the estimates are in each dataset on average, and the between variance, which measures how the estimates vary across the multiple datasets. These rules are presented in Rubin (1987, 2004) and improved degrees of freedom with small samples are developed in Barnard and Rubin (1999). The percent of missing information for any estimand (e.g., the mean of a covariate) is the ratio of the between variance to the total variance. For our covariates, the percents of missing information for their means are all under one percent, indicating that the multiple imputation, although allowing us to create matched samples using standard complete-data software, should have only minor effects on the results.

4.5 Dealing with Missing Treatment Indicators and Justifying SUTVA

Our target group consists of 56 treated freshmen and 285 control freshmen; 61 freshmen could either be treated or controls because even though they were not aware of having any smokers in their suites, they did not know all of their roommates' smoking habits, i.e., some of their roommates could be either smokers or non-smokers. These 61 freshmen could be regarded as

controls if we assumed that their smoking outcomes could not be influenced by smoking habits about which they are unaware. If we cannot accept this assumption, these 61 freshmen should be dropped from all analyses.

Some of the freshmen in the target group are from the same suite, which means that SUTVA would be unlikely to hold. SUTVA is far more justifiable, however, if we use only one freshman from each suite. For those suites where there were more than one treated units present, we repeatedly drew one at random; analogously for suites with more than one control unit. Millions of samples can be drawn using this method, but we drew 1,000 samples when the 61 students with undefined treatment status were regarded as control units and another 1,000 samples when they were omitted. We chose the sample that demonstrated the "best" balance between treated and control covariate distributions as described in Sections 4.6-4.9. We can perform such repeated analyses and maintain validity because no outcome data influenced our choice of the "best" sample.

4.6 Propensity Score Estimation in the Target Group

Table 2 summarizes a sequence of analyses that were performed on the covariate data in the target sample. Of essential importance, as has been emphasized repeatedly, the outcome data were not available. In effect, they were set aside and saved in a separate file during all of these analyses in order to assure that no information about the outcome data was accessed by mistake. The analyses are based on propensity scores estimated on a selection of covariates derived from

the list of proper covariates given in Table 1, that is, those covariates and their cross-products and interactions.

In a large sample, the propensity scores would have been based on all covariates in Table 1, as well as their cross-products and interactions because the biasing effects of excluding them has been shown to dominate the efficiency loss of including them (Rubin and Thomas, 2000). In a small sample such as ours of N=46 treated units versus N=204 controls[1], however, we need to be more restrictive. The covariates that are included are determined iteratively; separate iterations are run for each of the 2,000 samples' ten multiply-completed datasets. In the first iteration, we step-in the covariate that we believe is the most important to control for the probability of being treated, which is the number of desired roommates; the more roommates a student has, the more likely it is that one of them will be a smoker. This covariate remains in the subsequent analyses where we step-in the covariate that contributes the most to the R^2-value in the linear regressions, given that it is significant at the 0.10 level, and we step-out the covariates that are no longer significant at this level. A similar step-in/step-out procedure to estimate propensity scores has been used previously for larger samples (Bingenheimer et al, 2005). Once we have decided which covariates to use in each of the 2,000 samples' ten different datasets, we re-estimated the linear propensity-score using logistic regression.

4.7 Propensity Score Distributional Differences in the Target Group

[1] The samples that keep the freshmen with undefined treatment status have 235 controls whereas the samples that drop them have 204.

The "*B*" value in Table 2 is the number of standard deviations between the means of the linear propensity scores between the treated and control units in the target group. "Linear propensity score" means that the $\hat{\beta}X_i$ in the logistic regression was used instead of the corresponding estimated probability ($\hat{e}_i = \text{logit}^{-1}(\hat{\beta}X_i)$) for two principal reasons: (a) $\hat{\beta}X$ tends to be more normally distributed than the $\hat{e}_i$, and previous results on bias reduction using matching (e.g., Cochran and Rubin, 1973) assume normally distributed covariates; and (b) the linear form $\hat{\beta}X_i$ is more relevant to possible future modeling efforts using regression (i.e., covariance) adjustments, which typically model Y_i as approximately linearly related to X_i. More precisely, *B* is defined as: [the mean of $\hat{\beta}X_i$ in the treated group minus the mean of $\hat{\beta}X_i$ in the control group] divided by the within group standard deviation of $\hat{\beta}X_i$, defined as $[(S_t^2 + S_c^2)/2]^{1/2}$, where S_t^2 is the variance of $\hat{\beta}X$ in the treated group and analogously for S_c^2 in the control group. The value "*R*" is simply the variance ratio, S_t^2/S_c^2.

Before moving on to discuss the other columns in Table 2, consider the values of *B* and *R* in the target group displayed in the first row of Table 2, and suppose the linear propensity score is the only proper covariate. A difference of means of more than a standard deviation (i.e., *B* = 1.03) and a ratio of variances of a factor of nearly ½ (i.e., *R*=0.59) indicate that the active treatment and control treatment groups differ far too much on their covariate distributions to rely on models for adjustment, unless we are essentially certain of the form of the model relating outcomes to covariates (e.g., we are sure the outcomes of interest are linearly related to covariates *X*), because of the extrapolation involved when fitting straight lines to such data. This warning is quite old, and goes back as far as Cochran (1965). Older specific numerical results

appear in Cochran and Rubin (1973) and Rubin (1973b, 1979b). This lack of balance indicates that there are control units who look nothing like any treated units with respect to their propensity scores or the covariates that the propensity scores summarize. Because these "divergent" control units carry essentially no information about the treated units' missing potential outcomes (i.e., their potential outcomes under control), these divergent control units should simply be discarded.

4.8 Target Group Distributional Differences in X Orthogonal to the Propensity Scores

The other columns in Table 2 refer to the space of covariates orthogonal to (i.e., uncorrelated with) the estimated linear propensity scores. More precisely, in the treatment and control groups, regress each individual covariate on the estimated linear propensity scores, and then find the residuals of each covariate (i.e., the part of the variation in the covariate that is unaccounted for by the propensity scores, or in other words, each covariate orthogonal to the propensity scores). Calculate the mean and variance of these residuals among the treated and control units. The mean difference between the residuals among the treated and control units is essentially zero by construction; hence, there is no need to summarize those values. Next calculate the ratio of the variance of these residuals in the treated group to the variance of these residuals in the control group. The right half of Table 2 summarizes these variance ratios. Cochran and Rubin (1973) and Rubin (1973b) showed that with one covariate, even if the difference between treatment and control means is very small, linear modeling adjustments can be very sensitive to nonlinearities in the relationship between Y and X, when the variance ratio of X orthogonal to the propensity-score approaches 1/2 or 2. These other columns report the percentage of the proper covariates in

Table 1 and their cross-products and interactions (a total of 164 covariates) that have variance ratios in the categories shown: "Good" implies between 4/5 and 5/4; "Of Concern" implies not "Good" but between 1/2 and 2; and "Bad" implies less than 1/2 or greater than 2. This part of the table reveals that about 2% of the variance ratios are in the "Bad" region, and 31% are "Of Concern." An eigenvalue analysis would have been more appropriate but less intuitive to many audiences.

The conclusion from this first row is the following: (a) Based on the values of *B* and *R*, the groups are far apart, too far to trust adjustment based on linear models; (b) Based on the right side of the table, the groups have too many extreme variance ratios to trust linear modeling adjustments: if a *Y* happened to be highly correlated with one of those orthogonalized *X's* with a large or small variance ratio, we would be in bad shape relying on linear models. Consequently, we want to choose a subsample of the controls who are like the treated, and reassess the balance, that is, reassess the balance using the same calculations as in the first row, but now on a control subsample created by discarding control units who are unlike any treated unit.

4.9 Matching incorporating the propensity scores and reassessment of balance

The second row of Table 2 gives the identical information as the first row except that the control group is restricted to the subset of 46 control units[2] who were chosen by a "Mahalanobis metric matching within calipers defined by propensity scores" described in Rubin (2002b), and more generally described in Rosenbaum and Rubin (1985) and Rubin and Thomas (2000). From the

[2] There are 56 control units represented in the 10 datasets, because the treated units are matched with different control units in the respective datasets.

second row in Table 2 we see that, for the linear propensity scores estimated in Section 4.2, the treated and matched control samples are only 0.03 standard deviations apart, have a ratio of variances on this score equal to 1.04, and orthogonal to this score, have ratios of variances as given in the right part of that row. Discarding the 204 – 46 =158 "divergent" control units greatly improved the balance of the covariates.

But what would happen if we re-estimated the propensity scores in the matched samples as if they were the original random samples? Theoretically, if the X variables really were ellipsoidally symmetrically distributed with proportional covariance matrices, in expectation the originally estimated propensity scores would equal the re-estimated propensity scores because the matching method used was affinely invariant (Rubin and Thomas, 1992). Many of our matching variables, however, are not close to being ellipsoidally symmetric, but are instead, for example, dichotomous, or take on only a few values, and so by re-estimating the propensity scores, we account for this fact. When we re-estimate the propensity scores in the matched samples, we obtain the diagnostics for balance displayed in the third row of Table 2. The re-estimated propensity scores do not equal the originally estimated propensity scores. For example, B is now 0.12 rather than 0.03 as obtained with the originally estimated propensity scores, and R is now 0.85 rather than 1.04. The diagnostics in the last row of Table 2 indicate, however, that the even using the re-estimated propensity score, the matched samples are well balanced.

Other diagnostics for the estimated propensity scores are displayed in Figure 1, which reveals the distributional differences in the propensity scores between treated and control units corresponding to the three rows of Table 2. Still other diagnostics for balance can also be

conducted, such as ones based on the t-tests of the treatment and control mean differences for the proper covariates listed in Table 1, their cross-products and interactions (a total of 164 covariates). Figure 2 provides histograms for the associated 164 *p*-values for these *t*-tests in the target group and in the matched samples for the ten multiply imputed data sets. If the students in the target group had been randomly divided into treatment and control conditions, the target group histogram would be, in expectation, a uniform distribution. Clearly this is not the case, with far too many "significant" differences in covariate means. In contrast, in the matched samples, there are fewer significant *p*-values than would be expected under complete randomization – the matched samples are better balanced with respect to the observed covariates (and their non-linear terms) than we would have expected under complete randomization. In fact, in the ten multiply-imputed datasets, totaling 1,640 comparisons, we find that only four of these comparisons are significant at the 0.05 level and 14 at the 0.10 level. It is clear that the matched samples are much better balanced with respect to observed covariates than would be expected in a completely randomized experiment, and so we have achieved our design goal.

5. CONCLUSION: OBJECTIVE OBSERVATIONAL STUDY DESIGN

According to the principled perspective on causal inference that we have argued for in this article, observational studies for causal effects can often be designed objectively in ways that parallel the design of randomized experiments. This perspective asserts that we should first carefully define the science: (a) the units of study; (b) the treatments (i.e., interventions, real or hypothetical) whose effects we wish to know; (c) the covariates (i.e., background variables) that are unaffected by the treatments but may have been used to make treatment decisions, and

therefore should be used to define comparable treated and control units; and (d) the outcome variables on which we are interested in studying the effects of the treatments. Then, we should consider the plausibility of assumptions such as SUTVA and ignorability of the assignment mechanism and try to ensure that we have covariates and collections of treated and control units to make these assumptions plausible.

This process of objective design was illustrated here in the context of a study of the peer influence regarding smoking. Were the 46 nonsmoking freshmen rooming with at least one smoker more or less likely to smoke than they would have been if they instead had roomed with only non-smokers? By design, we matched them with 46 nonsmoking freshmen rooming only with non-smokers who had a similar distribution of background covariates. This matching was accomplished without any access to the critical outcome data, thereby assuring an objective design**.** Because outcome data were not used in any analysis in this paper, we could chose to use those 46 treated and control units who best matched each other as revealed by the diagnostics after repeated methods of estimating the propensity scores and of matching the units. Future work will report the results involving the outcome variables.

Finally, we would like to suggest that the design of any study should be carefully documented before any analyses of outcome data are undertaken so that the results can later be published (e.g., in an online database) irrespective of the significance or the magnitude of the results. Such a database would not only contribute to a better understanding of our economic environment, but it would also help defend the field of research against claims of publishing only studies with significant results.

Interview	Covariates and their interactions (nonlinear terms not explicitly displayed)
1	Days before the first interview
1	Number of desired roommates
1	Preference for a room that is a lively social center as opposed to a quiet/serious
1	Preference for a disordered room as opposed to a neat
1	When the days begin during the week
1	When the days end during week
1	Rank social science as academic interest
1	Rank fine arts as academic interest
1	Ranked social studies or English as academic interest
1	Ranked psychology as academic interest
1	Rank religious activities as extracurricular interest
1	Rank intramural/recreational sports or intercollegiate athletics as extracurricular interests
1	Ranked alternative, folk music, jazz, rock, and gospel as types of music they enjoyed
1	Ranked alternative, folk music, jazz, rock, gospel, and broadway tunes as types of music they enjoyed
1	Level of education completed by father[3]
1	Level of education completed by mother

[3] Questions about their father's/mother's educational level, the number of actual roommates they had, and their race were not included on the HAF. The former proxy for answers to one open-ended question on the HAF, the middle control for the increased probability of being treated in large suites, and the latter for the possible difference in peer influences across races.

1	Race
2	Number of actual roommates
2	Region where they lived before entering college

Table 1: Covariates used in propensity score models and for matching

	Propensity score diagnostics		Percent of covariates with specified variance ratio orthogonal to the propensity scores				
	B	*R*	(0, ½]	(1/2,4/5]	(4/5,5/4]	(5/4,2]	(2,∞]
Before matching	1.03	0.59	1	8	66	23	1
After matching	0.03	1.04	2	17	69	10	1
Mahalanobis metric matching	0.12	0.85	2	16	68	12	2

Table 2: Estimated propensity-scores on the logit scale for treated versus the control students

REFERENCES

Barnard J, Rubin DB. (1999), Small-sample Degrees of Freedom with Multiple Imputation. *Biometrica*, **86**(4): 948-955

Bingenheimer, JB, Robert TB, Earls JF. (2005), Firearm Violence Exposure and Serious Violent Behavior. *Science*, **308**(5726): 1323-1326.

Cochran WG, Rubin DB. (1973), Controlling bias in observational studies: a review. *Sankhya – A*, **35**(4): 417-446.

Cochran WG. (1965), The planning of observational studies of human populations (with discussion). *Journal of the Royal Statistical Society, Series A*, **128**:234-266.

D'Agostino R, Rubin DB. (2000), Estimation and Use of Propensity Scores with Incomplete Data. *Journal of the American Statistical Association*, **95**(451): 749-759.

Dorey, FJ., Little, RJA., Schenker, N. (1993), Multiple imputation for threshold-crossing data with interval censoring. *Statistics in Medicine*, **12**: 1589-1603.

Fisher RA. (1925), *Statistical methods for research workers.* Oliver and Boyd: Edinburgh.

Gelman A, Meng XL (eds.). (2004), *Applied Bayesian modeling and causal Inference from incomplete-data perspectives*. Wiley: Chichester.

Haavelmo T. (1944), The probability approach in econometrics. *Econometrica,* **15**: 413-419.

Heitjan, DF., Little, R JA. (1991), Multiple imputation for the fatal accident reporting system. *Journal of the Royal Statistical Society C*, **40**: 13-29.

Holland PW, Rubin DB. (1983), On Lord's paradox. *Principles of Modern Psychological Measurement: A Festschrift for Frederick Lord*. Erlbaum: New Jersey, pp 3-25.

Holland PM. (1986), Statistics and causal inference. *Journal of the American Statistical Association*, **81**(39): 945-960.

Imbens G, Rubin DB. (1997), Bayesian inference for causal effects in randomized experiments with noncompliance. *The Annals of Statistics,* **25(**1): 305-327.

Imbens G, Rubin DB. (2007a), Rubin causal model. Entry to appear in *The New Palgrave Dictionary of Economics*, 2nd ed. Palgrave McMillan: New York.

Imbens G, Rubin DB. (2007b), *Causal Inference in Statistics, and in the Social and Biomedical Sciences*. Cambridge University Press: New York, [to appear].

Langenskiöld S. (2005), *Peer influence on smoking: causation or correlation?* PhD Thesis. Stockholm School of Economics: Stockholm.

Leamer, E. (1978), *Specification Searches: Ad Hoc Inference with Nonexperimental Data.* John Wiley and Sons: New York.

Li, KH., Meng, XL., Raghunathan, TE., and Rubin, DB. (1991), Significance levels from repeated p-values with multiply-imputed data. *Statistica Sinica,* **1**: 65 - 92.

McCaffrey DF, Stuart EA, Rubin DB, Zanutto E. (2006), Design and implementation of case-control matching to estimate the effects of value-added assessment. Unpublished paper, Rand Corporation.

Neyman J. (1923, 1990), On the application of probability theory to agricultural experiments: essay on principles, section 9. *Annals of Agricultural Science*1923. Translated in *Statistical Science*, **5**(4): 465-472.

Raghunathan ET, Lepkowski JM, Van Hoewyk J, Solenberger P. (2001), A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models. *Survey Methodology*, **27**: 85-95.

Rosenbaum P, Rubin, DB. (1983), The central role of the propensity score in observational studies for causal effects. *Biometrika,* **70**: 41-55.

Rosenbaum P, Rubin, DB. (1984), Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association*, **79**: 516 524. (with P.R. Rosenbaum).

Rosenbaum PR, Rubin DB. (1985), Constructing a control group using multivariate matched sampling incorporating the propensity score. *The American Statistician*, **39:** 33-38.

Roy AD. (1951), Some thoughts on the distribution of earnings. *Oxford Economic Papers*, **3**: 135-146.

Rubin DB. (1973a), Matching to remove bias in observational studies. *Biometrics*, **29(**1): 159-183. Printer's correction note 30, p. 728.

Rubin DB. (1973b), The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, **29**(1): 184-203.

Rubin DB. (1974), Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**(5): 688-701.

Rubin DB. (1975), Bayesian inference for causality: the importance of randomization. *The Proceedings of the Social Statistics Section of the American Statistical Association.* American Statistical Association, Alexandria, pp. 233-239.

Rubin DB. (1976), Inference and missing data. *Biometrika*, **63**(3): 581-592. With discussion and reply.

Rubin DB. (1977a), Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics*, **2**(1): 1-26. Printer's correction note 3, p. 384.

Rubin, DB. (1977b), The Design of a General and Flexible System for Handling Non-response in Sample Surveys. *Manuscript Prepared for the U.S. Social Security Administration*, July 1.

Rubin, DB. (1978a), Multiple Imputations in Sample Surveys – Phenomenological Bayesian Approach to Nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34. Also in I*mputation and Editing of Faulty or Missing Survey Data*, U.S. Dept. of Commerce, Bureau of the Census,1-23

Rubin DB. (1978b), Bayesian inference for causal effects: the role of randomization. *The Annals of Statistics* 1978; **6**(1):34-58.

Rubin DB. (1979a), Discussion of "Conditional Independence in Statistical Theory," by A.P. Dawid. *The Journal of the Royal Statistical Society, Series* B, **41**(1): 27-28.

Rubin DB. (1979b) Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies. *The Journal of the American Statistical Association*, **74**(366): 318 328.

Rubin DB. (1980), Discussion of "Randomization Analysis of Experimental Data in the Fisher Randomization Test" by Basu. *The Journal of the American Statistical Association* **75(**371): 591-593.

Rubin, DB. (1987, 2004), *Multiple Imputation for Nonresponse in Surveys.* John Wiley & Sons.

Rubin DB. (1990a), Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, **5**(4): 472-480.

Rubin DB. (1990b), Formal Modes of Statistical Inference for Causal Effects. *Journal of Statistical Planning and Inference*, **25**: 279 292.

Rubin DB, Thomas N. (1992), Affinely invariant matching methods with ellipsoidal distributions. *The Annals of Statistics*, **20**(2): 1079-93.

Rubin DB, Thomas N. (1996), Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrika*, **52**(1): 249-264.

Rubin DB. (2000), Statistical issues in the estimation of the causal effects of smoking due to the conduct of the tobacco industry. Chapter 16 in *Statistical Science in the Courtroom*, J. Gastwirth (ed.). Springer-Verlag: New York, pp. 321-351.

Rubin DB, Thomas N. (2000), Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association*, **95**(450): 573-585.

Rubin DB. (2001), Estimating the causal effects of smoking. *Statistics in Medicine*, **20**: 1395-1414.

Rubin DB. (2002a) The ethics of consulting for the tobacco industry. Special Issue on "Ethics, Statistics and Statisticians", *Statistical Methods in Medical Research*, **11**(5): 373-380.

Rubin DB. (2002b), Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services & Outcomes Research Methodology*, **2**: 169-188.

Rubin DB, Stuart EA, Zanutto EL. (2004), A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, **29**(1): 103-116.

Rubin DB. (2005), Causal inference using potential outcomes: design, modeling, decisions. 2004 Fisher Lecture. *The Journal of the American Statistical Association*, **100**(469): 322-331.

Rubin DB. (2007a), Statistical Inference for Causal Effects, with Emphasis on Applications in Epidemiology and Medical Statistics. To appear in *Handbook of Statistics: Epidemiology and Medical Statistics*. (C.R. Rao, J.P. Miller and D.C. Rao, eds.). The Netherlands: Elsevier.

Rubin, DB. (2007b), The Design Versus the Analysis of Observational Studies for Causal Effects: Parallels With The Design of Randomized Trials. *Statistics in Medicine*, **26**(1):20-36.

Sacerdote, B (2001), Peer effects with random assignment: results for Dartmouth roommates. *The Quarterly Journal of Economics,* **116**(2): 681-704.

Schafer, JL. (1997), *Analysis of Incomplete Multivariate Data*. Chapman & Hall.

Schenker, N., Taylor, J.M.G. (1996), Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, **22**: 425-446.