STATISTICS IN MEDICINE
Statist. Med. 2007; **26**:20–36
Published online 10 October 2006 in Wiley InterScience
(www.interscience.wiley.com) DOI: 10.1002/sim.2739

Paper Celebrating the 25th Anniversary of *Statistics in Medicine*

The design *versus* the analysis of observational studies for causal effects: Parallels with the design of randomized trials

Donald B. Rubin*,†

Department of Statistics, Harvard University, 1 Oxford Street, 7th Floor, Cambridge, MA 02138, U.S.A.

SUMMARY

For estimating causal effects of treatments, randomized experiments are generally considered the gold standard. Nevertheless, they are often infeasible to conduct for a variety of reasons, such as ethical concerns, excessive expense, or timeliness. Consequently, much of our knowledge of causal effects must come from non-randomized observational studies. This article will advocate the position that observational studies can and should be designed to approximate randomized experiments as closely as possible. In particular, observational studies should be designed using only background information to create subgroups of similar treated and control units, where 'similar' here refers to their distributions of background variables. Of great importance, this activity should be conducted without any access to any outcome data, thereby assuring the objectivity of the design. In many situations, this objective creation of subgroups of similar treated and control units, which are balanced with respect to covariates, can be accomplished using propensity score methods. The theoretical perspective underlying this position will be presented followed by a particular application in the context of the US tobacco litigation. This application uses propensity score methods to create subgroups of treated units (male current smokers) and control units (male never smokers) who are at least as similar with respect to their distributions of observed background characteristics as if they had been randomized. The collection of these subgroups then 'approximate' a randomized block experiment with respect to the observed covariates. Copyright © 2006 John Wiley & Sons, Ltd.

KEY WORDS: assignment mechanism; causal inference; objective design; propensity scores; Rubin causal model; tobacco litigation

1. MY PERSPECTIVE ON INFERENCE FOR CAUSAL EFFECTS

Since the early 1970s, my views on the role of statistics for estimating causal effects have remained relatively constant, relatively because there is no doubt that my current views have been influenced by some wonderful colleagues and former students (e.g., see the contributions in Gelman and

*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

Received 26 July 2006
Accepted 28 September 2006

Meng [1]). But my basic perspective was formed earlier: in the early 1960s by my background in physics at Princeton; in the mid and late1960s by consulting projects on observational studies in the social sciences, and in the late 1960s by courses and conversations with Bill Cochran, who taught me classical experimental design and supervised my PhD thesis on matched sampling in observational studies. This background led me to a firm belief that all statistical studies for causal effects were seeking the same sort of answer, and that randomized experiments and observational studies were not on a dichotomy but rather on a continuum, from well suited for drawing such inferences to poorly suited. That is, for example, a randomized experiment where 90 per cent of those assigned treatment conditions do not comply with their assignments but in unknown ways and where there are many missing values is quite possibly less likely to lead to correct inferences for causal inferences than a carefully conducted observational study with many relevant covariates recorded and well-understood reasons for the assignments of treatments. In the early 1970s, this view evolved and was formalized and infused with an additional Bayesian analytic perspective stimulated by the work of Art Dempster and later George Box, who also conveyed to me the paramount importance of design.

This perspective was called the ‘Rubin causal model (RCM)’ by Paul Holland [2] for a sequence of papers I wrote in the 1970s [3–9]. The RCM can be seen as having two essential parts, together called the ‘potential outcomes with assignment mechanism’ perspective (Rubin [10, p. 476]), and a third optional part, which involves extensions to include Bayesian inference. The first part 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 articulated if causal inferences are to provide meaningful guidance for practice. The theme of this paper focuses on the second part, which concerns the assignment mechanism, and therefore the design of studies for causal effects. I believe that the careful implementation of this step is absolutely essential for drawing objective inferences for causal effects in practice, whether in randomized experiments or observational studies, yet is often basically ignored in observational studies relative to methods of analysis for causal effects. One of the reasons for this misplaced emphasis may be that the importance of design in practice is often difficult to convey in the context of most technical statistical articles.

This article is my brief attempt to refocus us on the importance of the design of observational studies, where by ‘design’ I mean all contemplating, collecting, organizing, and analysing of data that takes place prior to seeing any outcome data. Thus, 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, as well as the specification of the primary analysis plan of the outcome data, but any analysis that requires outcome data to implement is not part of design. A ‘snappy’ (to borrow one of Bill Cochran’s favourite words) title for this article would be ‘Design trumps analysis.’

A brief review of the three parts of the RCM will be given in Section 2, which introduces terminology and notation; a full-length text from this perspective is Imbens and Rubin [11]. Section 3 focuses on the assignment mechanism, the real or hypothetical rule used to assign treatments to the units, and the role of propensity scores in an observational study to try to reconstruct the hypothetical broken randomized experiment that led to the observed data. Section 3 includes a critique of the way most observational studies are conducted—essentially ignoring the reconstruction of the assignment mechanism, and advocates objectivity in observational studies using a careful design phase that never examines any outcome data. Then Section 4 illustrates the design of an observational study in the context of the recent tobacco litigation, using matched

DOI: 10.1002/sim

sampling and subclassification with diagnostics summarizing the achieved balance of covariate distributions in treated and control subsamples. This section also briefly addresses the non-causal nature of the resulting comparisons of smokers to non-smokers, and addresses the relevance of the analyses presented here nevertheless. The concluding section emphasis the sense in which the resulting observational study can be considered to be objectively designed.

2. BRIEF REVIEW OF THE RCM

2.1. Part one: Units, treatments, potential outcomes

Three basic concepts are used to define causal effects in the RCM. A unit is a physical object, for example, a patient, at a particular place and point of time, say time t. A treatment is an action or intervention that can be initiated or withheld from that unit at t (e.g., an anti-hypertensive drug, a statin); if the active treatment is withheld, we will 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, $t^*>t$: the value of some outcome measurements Y (e.g., blood pressure, cholesterol level) if the active treatment is given at t and the value of Y at the same future point in time if the control treatment is given at t. The causal effect of the treatment on that unit is defined to be the comparison of the treatment and control potential outcomes at t^* (e.g., their difference, their ratio, the ratio of their squares). 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 N 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)' [9, 10], 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. The fundamental problem facing causal inference [2; 7, Section 2.4] is that only one of the potential outcomes for each unit can ever be observed. In contrast to outcome variables, covariates are variables that take the same value for each unit no matter which treatment is applied to the units, such as quantities measured before treatments are assigned (e.g., age, pre-treatment blood pressure or cholesterol level). These values of the variables can be arranged in units by variables array consisting of: covariates, X; potential outcomes under control, $Y(0)$; and potential outcomes under 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 treatment and control potential outcomes on a common set of units; for example, the median $\log Y(1)$ *versus* the median $\log Y(0)$ for those units who are female between 31 and 35 years old as indicated by their X values, or the median $(\log Y(1) - \log Y(0))$ for those units whose $Y(0)$ and $Y(1)$ values are both positive.

This first part of the RCM is conceptual and can, and often 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' [4]. The formal use of potential outcomes to define unit-level causal effects is due to Neyman in 1923 [12] in the context of randomized experiments, and was a marvellously clarifying contribution. But evidently this notation was not formally extended to non-randomized settings until 1974 by Rubin [3], as discussed in [10, 11, 13, 14].

The intuitive idea behind the use of potential outcomes to define causal effects is very old. Nevertheless, in the context of non-randomized observational studies, prior to 1974 everyone

DOI: 10.1002/sim

appeared to use the 'observed outcome' notation when discussing causal inference. More explicitly, letting i index units and W be the column vector for the treatment assignments for the units ($W_i = 1$ if treated, $W_i = 0$ if control), the observed outcome notation replaces the potential outcomes $[Y(0), Y(1)]$ with Y_{obs}, where for the ith component of Y_{obs}

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

The observed outcome notation is inadequate in general, and can lead to serious errors—see for example, the discussion by Holland and Rubin [15] on Lord's paradox, and Rubin [14], where errors are explicated that Fisher made because of his eschewing the potential outcome notation.

2.2. Part 2: The assignment mechanism

The second part of the RCM is the formulation, or positing, 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

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

Although this general formulation, with the possible dependence of assignments on the yet to be observed potential outcomes, arose first in Rubin [4], special cases were much discussed prior to that. For example, randomized experiments (Neyman [12], Fisher [16]) are 'unconfounded' [4]

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

and they are 'probabilistic' in the sense that their propensity scores, e_i, are bounded between 0 and 1

$$0 < e_i < 1, \tag{4}$$

where

$$e_i \equiv \Pr(W_i = 1|X_i). \tag{5}$$

When the assignment mechanism is both probabilistic (4)–(5) and unconfounded (3), it generally can be written as proportional to the product of the unit level propensity scores, which emphasizes the importance of propensity scores in design

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

The term 'propensity scores' was coined by Rosenbaum and Rubin [17], where an assignment mechanism satisfying (3) and (4) is called 'strongly ignorable,' a stronger version of 'ignorable' mechanisms, coined by Rubin [5, 7], which implies possible dependence on observed values of the potential outcomes, such as in a sequential experiment

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

But until 1975 randomized experiments were not defined using equations (3)–(4), which explicitly show such experiments' freedom from any dependence on observed or missing potential outcomes.

DOI: 10.1002/sim

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. For example, in economics, Roy [18] described, without equations or notation, 'self-optimizing' behaviour where each unit chooses the treatment with the optimal outcome. And another well-known example from economics is Haavelmo's [19] formulation of supply and demand behaviour. But these and other formulations in economics and elsewhere did not use the notation of an assignment mechanism, nor did they have methods of statistical inference for causal effects based on the assignment mechanism. Instead, 'regression' models were used to predict $Y_{\text{obs},i}$ from X_i and W_i, with possible restrictions on some regression coefficients and/or error terms, where particular regression coefficients (e.g., for W_i or interactions with W_i) were interpreted as causal effects; analogous approaches were used in other social sciences, as well as in epidemiology and medical research. Such models were based on assumptions about the assignment mechanism and about the science, which were typically only vaguely explicated, and therefore could, and sometimes did, lead to mistakes.

Inferential methods based only on the assumption of a randomized assignment mechanism were proposed by Fisher [16] and Neyman [12] and further developed by others (see [10] for some references). The existence of these methods documents that the model for the assignment mechanism is more fundamental for inference for causal effects than a model for the science. These methods lead to concepts such as unbiased estimation (Neyman), p-values for null hypotheses (Fisher), significance tests (Fisher), and confidence intervals (Neyman), all defined by the distribution induced by the assignment mechanism. The collection of propensity scores defined by equation (5) is the most basic ingredient of an unconfounded assignment mechanism, and its use for objectively designing observational studies will be developed and illustrated here in Sections 3 and 4.

2.3. Part 3: Full probability model on the science

The third and final part of the RCM is optional; it is a model specification for the science, the quantities treated as fixed in the assignment-based approach and conditioned on in the assignment mechanism

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

This model on the science completes the full model specification of all observable quantities as random variables, and so is the Bayesian approach, as defined first by Rubin [4] and further developed by Rubin [7] and in other places, such as Imbens and Rubin [20]. 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 science, and thus also leads to direct posterior inference for all causal effects. Because the topic here is design in observational studies rather than the analysis of observational studies, this very brief Section 2.3 is our only digression into analysis of the observed outcome data, Y_{obs}, here to obtain the posterior predictive distribution of the missing potential outcomes, Y_{mis}, with ith component $Y_{\text{mis},i} = W_i Y_i(0) + (1 - W_i) Y_i(1)$. It is important to realize that in this formulation, the model for the assignment mechanism, which is with unconfounded designs essentially a propensity score model as in equation (6), is not a model involving what is here called 'the science,' but is conditional on the science.

DOI: 10.1002/sim

3. CONCEPTUALIZING AN OBSERVATIONAL STUDY OBJECTIVELY TO APPROXIMATE A RANDOMIZED EXPERIMENT

An observational study should be conceptualized as a broken randomized experiment. That is, Parts 1 and 2 of the RCM just described should be structured just as carefully in an observational study as in an experiment, where 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.

3.1. No outcome data in sight!

Of critical importance, in randomized experiments the design phase takes place prior to seeing any outcome data. And this critical feature of randomized experiments can be duplicated in observational studies, for example, using propensity score methods, and we should objectively approximate, or attempt to replicate, a randomized experiment when designing an observational study. 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. Of course, 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, whereas randomization eliminates bias due to all covariates, both observed and unobserved.

Blocking and matching on particular covariates are methods for eliminating extraneous variation due to those covariates, whether in the context of a randomized experiment or an observational study, and therefore these techniques are used to increase precision. The more 'conditional' way to say this is that such blocking, by creating treatment and control subgroups within which the distributions of observed covariates are more similar than would be expected if we simply assigned treatments to units completely at random, eliminates conditional (on these covariates) bias, which when averaged over in a completely randomized design becomes variance. But the critical point remains: no outcome data from the study are in sight when objectively designing either a randomized experiment or an observational study.

3.2. Common practice and an objective alternative

The prescription for design in Section 3.1 is in conflict with the typical analysis, as least as published, of observational data in epidemiology and social science. In these disciplines, rather than the outcome data, Y_{obs}, being 'not in sight,' they are used over and 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 rather than age as continuous, because the p-values for treatment effects greatly improved!' How many reported analyses that we see in journals are 'designed' *a priori* rather than are the results of repeated and unreported exploratory analyses, looking for a publishable result, where 'publishable' can imply positive or negative or simply most exciting or most likely to lead to 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 favourable experiment were used for approval of the drug for general use? Or if the drug were approved based on dozens or hundreds of re-analyses of the same data until some analysis was found that produced a significantly favourable result? Probably you would be

DOI: 10.1002/sim

skeptical of advertisements touting the great benefits of a drug approved on such analyses. Instead, typically in order to get a drug approved, US Food and Drug Administration (FDA) requires carefully specified randomized designs and carefully specified primary analyses and secondary supporting analyses, and often the data collection and first pass analyses are carried out by a agent independent from the organization trying to get approval for the drug. There is thus tremendous pressure to live with the answers that come from the pre-specified design and analyses.

Should we accept lower standards for social science and epidemiology? Objectivity can be obtained 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 truth, but I believe that it is generally a necessary ingredient if we are to find truth. The key idea is to conduct the design before ever seeing any outcome data, and to do it in such a way that any future model-based adjustments will tend to give similar point estimates. This can often be accomplished using propensity score methods to create subgroups (i.e., subclasses) of treated and control units such that within each subclass, the treated and control units have essentially the same distributions of all covariates. If this can be achieved, then model-based adjustments within each subclass will give basically the same point estimates of the subclass-specific treatment effects, although different models will yield different estimates of the precisions of those subclass-specific estimates. Therefore, ideally, the design should include the specification of the analysis that is to be carried out on Y_{obs}, especially when p-values or interval estimates will be used to make decisions.

3.3. Some recent personal examples

There are several recent examples where I have had the opportunity to implement the approach being advocated here. One concerns certain parts of the tobacco litigation, which will be discussed at greater length in Section 4 and used to illustrate some details.

The next is an on-going evaluation of 'value-added assessment (VAA)' in public schools in Pennsylvania. Part of the current administration's plan for improving the education in public schools in 'No Child Left Behind' is to hold teachers accountable for the academic performance of their students, with the hope that this accountability will lead to better teaching of the students and thus better performance by the students. This assessment is called VAA (e.g., see [21]).

A joint project with Rand set out to do this by taking school districts that had implemented VAA a year or two before the other school districts, and then subsequently comparing the children's academic performance in the treated (VAA) and control (no VAA) districts. Because there were many more non-VAA districts in the early years, at the design stage, we were able to take matched samples of treated and control districts using detailed records of past achievement, socio-economic status, percent minority, parents' education, etc. as matching variables. The variables to include were decided upon by a committee created to be composed of advocates and critics of VAA, and people knowledgeable about factors affecting both (a) administrative decisions to accept or not VAA early and (b) educational achievement. Any variable thought as important by anyone was initially included. The resulting matched samples were also passed by the committee to see if anyone objected to the balance that was created between treated and control districts with respect to these covariates. Thus, no one objected to the results of the initial effort to recover an unconfounded assignment mechanism for VAA *versus* no VAA. Of critical importance, the outcome variables were not even collected yet! They would only be collected on those districts chosen for the matched samples. The balance achieved was very good, and the objectivity of the design was enhanced by the review by the committee without any access to the implications for the

DOI: 10.1002/sim

answers. We are still waiting for the collection of outcome data; McCaffrey *et al*. [22] describes the matching effort.

Another example, which is more oriented to epidemiology, is an evaluation of the effects of peer influence on freshman smoking habits at Harvard University that was recently completed by an Economics PhD student who was visiting Harvard University from the University of Stockholm, Langenskiöld [23]. Using data that she initially collected the week before classes started and shortly thereafter, she classified a set of freshmen, who all entered as non-smokers, as either being treated, meaning rooming with a least one smoker, or control, meaning rooming with only non-smokers. The assignment of freshmen to rooms (really suites of rooms) is claimed by the housing office to be essentially random, although the office does have access to questionnaires filled out by the accepted freshmen, which we did not have access to, but whose questions we also asked during freshman week—not the identical data, but pretty close, certainly closer to having the data to justify a claim of an unconfounded assignment mechanism than in most observational studies. Later, data were collected on the outcomes of smoking behaviour near the end of the academic year. Again, critically, the outcome data were nowhere in sight when a massively time-consuming matching of treated and controls took place, involving extensive evaluations of the covariate balance between treated and control groups.

When the outcome data were finally revealed, they strongly suggested no treatment *versus* control effect whatsoever, a conclusion that was very robust to all sorts of standard model-based adjustments, as expected given the enforced balance on background covariates. This work was written in two distinct phases: first the design was finalized and implemented, and then the outcome data were examined and the preplanned analyses took place [23].

4. AN EXAMPLE FROM THE TOBACCO LITIGATION

A critical component in all the U.S. Tobacco litigation is to compare health-care costs and disease rates of smokers and 'like' never-smokers, where smokers are classified initially both as former or current, and as male or female. For ease of exposition here, I focus on the male current smokers.

4.1. The basic data and the definition of never smokers 'like' the current smokers

There exists one particular data set that both plaintiffs and defendants in this and related litigation use for many of their data analyses: the 1987 National Medical Expenditure Survey, or NMES (pronounced 'Nem-us') for short. NMES not only has copious detail on medical expenditures from ten of thousands of respondents, it also has extensive demographic and health-related covariate information on these subjects. Although we could just compare the group of male current smokers with the male never smokers in NMES, these groups are not similar to each other in background variables. There is an enormous amount of legal wrangling about what variables other than male/female are allowed to be, or in some sense, should be, used to define when a never-smoker is 'like' a smoker. Table I (from Rubin [24]) is a list of 'acceptable' variables, acceptable in the sense that the court says (at least in one case) that they may (and in some sense, should) be used to define when a never-smoker is like a smoker. Thus, the attempt will be to find a sample of male never-smokers in NMES who, as a group, look just like the male current smokers in NMES. The issue of the causal effect of the tobacco industry's alleged misconduct, and how it relates to the comparisons being sought in this section, is discussed by Rubin [25, 26] and the ethics of testifying

DOI: 10.1002/sim

Table I. Background variables in NMES.

Variables used in propensity model	Description
Seatbelt	5 levels of reported seat belt use
Arthritis	Whether reported suffering from arthritis
Census division	9 census regions
Champ insurance	
Diabetes	Doctor ever told having diabetes
Down time	6 levels of reported down time
Dump time	6 levels of reported dump time
Employment	Indicating employment status each quarter
English	English is a primary language
Retirement	Indicator for retirement status
Number of friends	7 levels measuring the number of friends
Membership in clubs	6 levels measuring memberships in clubs
Education	Completed years of education
HMO coverage	Indicating HMO coverage each quarter
High blood pressure	Doctor ever told having high blood pressure
Industry code	14 industry codes
Age	Age of the respondent
Labor union	Indicator for a member of labour union
Log height	Natural logarithm of height
Log weight	Natural logarithm of weight
Marital status	Marital status in each quarter
Medicaid	On medicaid (each quarter)
Medicare	On medicare (each quarter)
Occupation	Occupation code (13 levels)
Public assistance	Other public assistance program (each quarter)
Friends over	Frequency of having friends over (7 levels)
Physical activity	Indicator variable for physically active
Population density	3 levels
Poverty status	6 levels
Pregnant 1987	Pregnancy status in 1987 (women)
Private insurance	Other private insurance (each quarter)
Race	4 levels
Rated health	5-point self-rating of health status
Home ownership	Indicator for owning home
Rheumatism	Indicator for suffering from rheumatism
Share life	Indicator variable for having somebody to share their life
Region	4 levels of region of the country
MSA	4 levels indicating types of metropolitan statistical area
Risk	General risk taking attitude (5 levels)
Uninsured	Indicator for lack insurance (each quarter)
Veteran	Indicator for veteran status
Incapler	Survey weight in NMES database
Agesq	Age $*$ Age
Educat.sq	Education $*$ Education
Age_wt	Age $*$ logwt
Age_educt	Age $*$ Education
Age_ht	Age $*$ loght
Educat_wt	Education $*$ logwt
Educat_ht	Education $*$ loght

Table I. *Continued*.

Variables used in propensity model	Description
Loght_logwt	Loght * logwt
Loghtsq	Loght * loght
Logwtsq	Logwt * logwt
Other non-linear terms	

Table II. Original propensity score analyses in NMES: balance between male current smokers ($N=3510$) and all male never smokers ($N=4297$), and between male current smokers and the roughly comparable subset of male never smokers ($N=3510$).

Analysis of current smokers *versus*	Distributional differences in propensity scores		Percent of covariates with specified variance ratio after adjustment for propensity score		
	B	*R*	Good	Of concern	Bad
All never	1.09	1.00	57	34	9
Subset never	0.08	1.16	90	9	1

Note: $B=$ number of standard deviations between means of propensity score in current and never smokers; $R=$ ratio of current-smoker to never-smoker variance on the propensity score; also displayed is the percentage of the covariates orthogonal to the propensity score with the specified variance ratios: good = between 4/5 and 5/4; of concern = not good but between 1/2 and 2.

in such cases is discussed by Rubin [27]. But for the purposes of this article, we accept the list in Table I as giving the covariates that define 'similar' current smokers and never smokers. When creating a list such as this, some variables should not be included. For example, variables that are effectively known to have no possible connection to the outcomes, such as random numbers, or five-way interactions, or the weather half-way around the world at the time of the person's birth.

4.2. Initial distributional differences—too substantial to rely on regression adjustment

Tables II and III (modified from [24]) summarize a sequence of analyses, all conducted using only the variables in Table I, that is, with no medical expenditure data or cancer rate or other outcome data in sight. This matching/subclassification analysis was therefore in stark contrast to the numerous and repeated model-based analyses done by plaintiffs' and defendants' experts looking for results favourable to their side.

The first row of Table II shows the results of propensity score analyses of the $N=3510$ male current smokers (treated) *versus* the $N=4297$ male never smokers (control). All propensity score analyses reported in this table included all variables in Table I as well as some non-linear terms as indicated in the last row of that table. These analyses were implemented as stepwise logistic regressions predicting the treatment-control indicator variable from the background variables, as described in more detail by Rubin [24].

The 'B' value is the number of standard deviations between the means of the linear propensity score in the treated and control groups. 'Linear propensity score' means that the $\hat{\beta}X_i$ in the logistic

DOI: 10.1002/sim

Table III. Re-estimated propensity score analyses in NMES: assessing achieved balance after subclassification on re-estimated propensity scores for male current smokers ($N = 3510$) *versus* subset male never smokers ($N = 3510$).

Number of subclasses	Distributional differences in propensity scores		Percent of covariates with specified variance ratio after adjustment for propensity score		
	B	R	Good	Of concern	Bad
1	0.39	1.33	88	12	0
2	0.18	1.36	98	2	0
4	0.10	1.25	99	1	0
6	0.09	1.30	100	0	0
8	0.08	1.16	100	0	0
10	0.07	1.12	100	0	0

Note: B = number of standard deviations between means of propensity score in current and never smokers; R = ratio of current-smoker to never-smoker variance on the propensity score; also displayed is the percentage of the covariates orthogonal to the propensity score with the specified variance ratios: good = between 4/5 and 5/4; of concern = not good but between 1/2 and 2; bad = less than 1/2 or greater than 2.

regression was used instead of the corresponding estimated probability ($\hat{e}_i = \text{logit}^{-1}\hat{\beta}X_i$) because (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 [28]) assume normally distributed covariates, and (b) the linear form $\hat{\beta}X_i$ is more relevant to possible future modelling 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. The value 'R' is simply the variance ratio, S_t^2/S_c^2.

Before moving on to discuss the other columns, consider the values of B and R in the full random samples (i.e., in the first row), and suppose the linear propensity score is the only covariate. A difference of means of more than a standard deviation (i.e., $B = 1.09$) is simply too large to rely on modelling adjustments unless we are certain of the form of the model (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 as far back as Cochran [29], and specific results appear in Cochran and Rubin [28] and Rubin [30, 31]. The value of R, 1.00, is actually good news for the reliability of linear modelling adjustments, if B were not so large, and if $\hat{\beta}X_i$ were truly normally distributed. But B is large; the treatment and control groups are far apart in the direction of the propensity score. In particular, because we want to compare the current smokers to 'like' never smokers, if the never-smoker group includes individuals who look nothing like any current smokers with respect to the propensity score, or any other covariate, they should simply be discarded because they carry essentially no information about what a current smoker's health outcomes might be like if he instead were a never-smoker.

The other columns refer to the space of covariates orthogonal to (i.e., uncorrelated with) the estimated linear propensity score. More precisely, in the treatment and control groups, regress each individual covariate on the estimated linear propensity score, and then find the residuals of each

DOI: 10.1002/sim

covariate (i.e., each covariate orthogonal to the propensity scores). Calculate the mean and variance of these residuals in the treatment group and in the control group. The mean difference between the residuals in the treatment and control groups 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 II summarizes these variance ratios. Cochran and Rubin [28], Cochran [29] and Rubin [30] showed that with one covariate, even if the difference in treatment–control means is very small, linear modelling adjustments are very sensitive to non-linearities in the relationship between Y and X when the variance ratio of X approaches $\frac{1}{2}$ or 2. These other columns report the percentage of the covariates in Table I that have variance ratios in the categories shown: 'Good' implies between $\frac{4}{5}$ and $\frac{5}{4}$; 'Of concern' implies not 'Good' but between $\frac{1}{2}$ and 2; and 'Bad' implies less than $\frac{1}{2}$ or greater than 2. They reveal that about 8 per cent of the variance ratios are in the 'Bad' region, and 43 per cent 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 value of B, 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 variance differences to trust linear modelling adjustments: if a Y happened to be highly correlated with one of those orthogonal X's with a large or small variance ratio, we would be in bad shape relying on linear models to adjust. 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 sample created by discarding never smokers who are unlike any current smoker.

4.3. Initial matching incorporating the propensity score, and reassessment of balance

The second row of Table II gives the identical information as the first row except that the control group is restricted to the subset of 3510 individuals who were chosen by a 'propensity score caliper with Mahalanobis metric matching' procedure described by Rubin [24], and more generally described by Rosenbaum and Rubin [32] and Rubin and Thomas [33]. From the second row in Table II we see that, for the linear propensity scores estimated in Section 4.2, the treated and matched control samples are only 0.08 standard deviations apart, have a ratio of variances on this score equal to 1.16, and orthogonal to this score, have ratios of variances as given in the right part of that row. Discarding the $4297 - 3510 = 787$ never smokers who were least like the smokers greatly improved the balance of the covariates, except for the value of R; thus, we discarded only about 20 per cent of the original never smokers to reduce B from 1.09 to 0.08 for this originally estimated propensity score.

But what would happen if we re-estimated the propensity score 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 score would equal the re-estimated propensity score because the matching method used was affinely invariant [34]. 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. When we re-estimate the propensity scores in the matched samples, we obtain the diagnostics for balance displayed in the first row of Table III. Clearly, the re-estimated propensity score does not equal the originally estimated propensity score. For example, B is now nearly 0.4

DOI: 10.1002/sim

rather than less than 0.1 obtained with the originally estimated propensity score, and R is over 1.3 rather than less than 1.2; the matched samples are still too far apart to rely on linear regression adjustment.

4.4. Further subclassification on the propensity score, and proposed comparisons

In order to create better balance in the matched samples, we turn to subclassification as in Cochran [35] and Rosenbaum and Rubin [17, 36]. The remaining rows in Table III display the diagnostics when we subclassify into 2, 4, 6, 8 and 10 subclasses. Specifically, first consider two subclasses. To obtain the results in this row, we split the matched samples at the median propensity score in the matched samples, and then we (a) compared treated and controls who have low propensity scores, (b) compared treated and controls who have high propensity scores, and (c) weighted each comparison by the number of treated in each subclass. We weighted by the number of treated, that is, by the number of current smokers, because our objective is to estimate the outcomes for the current smokers as if they instead were never smokers. The diagnostics for two subclasses are substantially better than the diagnostics for the first row. Similarly, the diagnostics for four subclasses are even better, for six subclasses better still, for eight subclasses even better than those for six subclasses, and for ten subclasses extremely good. This result from subclassification is as expected from theoretical results by Cochran [35] and Rosenbaum and Rubin [36]. That is, if the propensity scores are well estimated, within a narrow bin of propensity score values (i.e., within any subclass), the joint distributions of all covariates that entered the propensity score estimation should be the same in the treatment and control groups in expectation.

The resulting proposed analyses would take place within each of the ten subclasses, and would then be combined by weighting the results by the number of current smokers in the subclasses. For instance, the analysis could be the simple comparison of mean total medical expenses for current smokers and never smokers. Or the subclass-specific analyses could be based on such a comparison adjusting for all X variables based on stepwise regression analyses. The claim here is that many such analyses would yield essentially the same point estimates, almost no matter what the model, because within each subclass, the distribution of X is nearly the same, at least as judged by the diagnostics in the last row of Table III. The standard errors of these estimates would vary with the specific model, with models fitting more Xs or more non-linear terms typically showing smaller standard errors and tighter interval estimates, but the point estimates should be stable across models.

Of course, there always could be and should be more refined analyses of balance conducted, for example, doing analogous propensity score analyses within each subclass, but as sample sizes become smaller, the effectiveness of propensity score analyses becomes more limited. This is not just a property of propensity score analyses but a limitation of randomization as well. Consider a randomized experiment with two units: one male and one female. How balanced is the variable male/female in the treatment and control groups each of size one? Also, our design would be more 'objective' if the exact final outcome analyses were fully specified, but this would be a very difficult task in the context of the tobacco litigation, except to say that all analyses that would be conducted would be done within the ten specified subclasses and combined as indicated.

For dramatic evidence that such an analysis can reach the same conclusion as an exactly parallel randomized experiment, see Shadish and Clark [37]. In this very carefully executed study, the

DOI: 10.1002/sim

subjects were first randomized into two arms, one of which was further randomly divided into treatment and control conditions, and the other of which was subjected to a parallel observational study with identical treatment and control exposures as in the randomized experiment. Propensity score subclassification in the observational study arm, roughly analogous to the way it was done here, and exactly following the guidelines by Rosenbaum and Rubin [36] yielded essentially the same answer as obtained in the randomized arm.

5. CONCLUSION: OBJECTIVE OBSERVATIONAL STUDY DESIGN

This brief article has argued, from what I consider a very principled perspective on statistical inference for causal effects, for the objective design of observational studies for causal effects to 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) about whose effects we wish to know; (c) the covariates (i.e., background variables) that are presumed to be unaffected by the treatments, and therefore can be used to define subgroups of units; and (d) the outcome variables that can be affected by the treatments, and all of whose observable values under all possible treatment assignments are represented by the collection of potential outcomes. We should also consider the plausibility of assumptions such as SUTVA, and try to define units and treatments to make maintained assumptions plausible.

The perspective then asserts that we should choose (if we have the option) or posit (if we are dealing with already existing data) an assignment mechanism that describes why some units will get or got assigned different treatments, formalized as a unit-exchangeable stochastic function of covariates and potential outcomes. With an unconfounded assignment mechanism, there is no dependence on the potential outcomes, and then two subgroups of units, one subgroup treated and one subgroup control, with the same distribution of the covariates involved in the assignment mechanism, may be fairly compared to estimate the effect of treatment *versus* control on outcomes because the two subgroups were essentially created by randomization. Very important for design is the fact that two subgroups of units, one treated and one control, with the same distribution of propensity scores have the same distribution of all measured covariates entering the assignment mechanism. Therefore, we should try to design an observational study such that there are subgroups of treated and control units with the same distributions of propensity scores, and this should be done without ever looking at any outcome data, and thus without looking at any answers about causal effects. Because the propensity score must be estimated, diagnostic analyses assessing the resulting balance of covariate distributions are essential.

This process of objective design was illustrated here in the context of the tobacco litigation. The result was a collection of ten subclasses, each consisting of current male smokers and male never smokers, who could be fairly compared, as defined by the court, with respect to outcomes such as medical expenditures, because each subclass has very similar distributions of background covariates for the treatment and control groups. This objective design was such that I could answer questions under oath at trial or deposition of the following type:

Question
'You claim that the regression analyses and other analyses of medical expenditures of smokers relative to similar never smokers in NMES that have been reported thus far by the plaintiffs' experts are unreliable, and that you have analyses to support your claim, is that correct?'

Answer
'Yes.'

Question
'You also claim that it is possible to do reliable analyses to compare the medical expenditures of smokers and similar never smokers in NMES, correct?'

Answer
'Yes.'

Question
'Have you ever done such analysis of the medical expenditures in NMES?'

Answer
'No.'

Question
'If you were to do them, could the estimated excess medical expenditures of smokers relative to similar never smokers be higher than the numbers calculated by plaintiffs' experts?'

Answer
'Yes.'

Question
'If you were to do these reliable analyses, could the estimated excess medical expenditures of smokers be lower than the numbers calculated by the plaintiff's experts?'

Answer
'Yes.'

Question
'So you have no idea whether the numbers being presented by these experts are too high or too low or right on target, you are simply saying that the way they were calculated was unreliable, and that it is possible to do reliable analyses on the NMES data set that compare the medical expenditures of similar smokers and never smokers?'

Answer
'Yes, that is correct.'

Question
'And you would have conducted the same analyses on background variables as presented here whether you were being retained as an expert for the plaintiffs or for the defendants?'

Answer
'Yes.'

Question
'So that is why you claim that if such analyses were to be conducted, the resulting numbers would be objective estimates of the excess medical expenditures of smokers relative to similar never smokers in NMES, and therefore the court could rely on them rather than the analyses that have been presented to the court?'

DOI: 10.1002/sim

Answer
'Yes.'

I believe that the ability of a researcher to answer such questions in this manner is an essential property of having designed an observational study objectively.

REFERENCES

1. Gelman A, Meng XL (eds). *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. Wiley: Chichester, 2004.
2. Holland PM. Statistics and causal inference. *Journal of the American Statistical Association* 1986; **81**(39):945–960.
3. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 1974; **66**(5):688–701.
4. Rubin DB. Bayesian inference for causality: the importance of randomization. *The Proceedings of the Social Statistics Section of the American Statistical Association*. American Statistical Association: Alexandria, VA, 1975; 233–239.
5. Rubin DB. Inference and missing data. *Biometrika* 1976; **63**(3):581–592 (with Discussion and Reply).
6. Rubin DB. Assignment to treatment group on the basis of a covariate. *Journal of Educational Statistics* 1977; **2**(1):1–26. Printer's correction note 3, 384.
7. Rubin DB. Bayesian inference for causal effects: the role of randomization. *Annals of Statistics* 1978; **6**(1):34–58.
8. Rubin DB. Discussion of 'conditional independence in statistical theory', by Dawid AP. *Journal of the Royal Statistical Society*, *Series B* 1979; **41**(1):27–28.
9. Rubin DB. Discussion of 'randomization analysis of experimental data in the Fisher randomization test' by Basu. *Journal of the American Statistical Association* 1980; **75**(371):591–593.
10. Rubin DB. Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science* 1990; **5**(4):472–480.
11. Imbens G, Rubin DB. *Causal Inference in Statistics*, *and in the Social and Biomedical Sciences*. Cambridge University Press: New York, 2006, in press.
12. Neyman J. On the application of probability theory to agricultural experiments: essay on principles, Section 9. *Annals of Agricultural Science* 1923; Translated in *Statistical Science* 1990; **5**(4):465–472.
13. Imbens G, Rubin DB. Rubin causal model. *The New Palgrave Dictionary of Economics* (2nd edn). Palgrave McMillan: New York, 2006, in press.
14. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. 2004 Fisher Lecture. *The Journal of the American Statistical Association* 2005; **100**(469):322–331.
15. Holland PW, Rubin DB. On Lord's paradox. *Principles of Modern Psychological Measurement*: *A Festschrift for Frederick Lord*. Erlbaum: Hillsdale, NJ, 1983; 3–25.
16. Fisher RA. *Statistical Methods for Research Workers*. Oliver and Boyd: Edinburgh, 1925.
17. Rosenbaum P, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**:41–55.
18. Roy AD. Some thoughts on the distribution of earnings. *Oxford Economic Papers* 1951; **3**:135–146.
19. Haavelmo T. The probability approach in econometrics. *Econometrica* 1944; **15**:413–419.
20. Imbens G, Rubin DB. Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics* 1997; **25**(1):305–327.
21. Rubin DB, Stuart EA, Zanutto EL. A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics* 2004; **29**(1):103–116.
22. McCaffrey DF, Stuart EA, Rubin DB, Zanutto E. Design and implementation of case–control matching to estimate the effects of value-added assessment. Unpublished Paper, Rand Corporation, 2006.
23. Langenskiöld S. Peer influence on smoking: causation or correlation? *Ph.D. Thesis*, Stockholm School of Economics, Stockholm, 2005.
24. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology* 2002; **2**:169–188.
25. Rubin DB. Statistical issues in the estimation of the causal effects of smoking due to the conduct of the tobacco industry. In *Statistical Science in the Courtroom* (Chapter 16), Gastwirth J (ed.). Springer: New York, 2000; 321–351.

26. Rubin DB. Estimating the causal effects of smoking. *Statistics in Medicine* 2001; **20**:1395–1414.
27. Rubin DB. The ethics of consulting for the tobacco industry. Special issue on 'Ethics, Statistics and Statisticians'. *Statistical Methods in Medical Research* 2002; **11**(5):373–380.
28. Cochran WG, Rubin DB. Controlling bias in observational studies: a review. *Sankhya—A* 1973; **35**(4):417–446.
29. Cochran WG. The planning of observational studies of human populations (with Discussion). *Journal of the Royal Statistical Society*, *Series A* 1965; **128**:234–266.
30. Rubin DB. Matching to remove bias in observational studies. *Biometrics* 1973; **29**(1):159–183. Printer's correction note 30, 728.
31. Rubin DB. The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics* 1973; **29**(1):184–203.
32. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling incorporating the propensity score. *The American Statistician* 1985; **39**:33–38.
33. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *Journal of the American Statistical Association* 2000; **95**(450):573–585.
34. Rubin DB, Thomas N. Affinely invariant matching methods with ellipsoidal distributions. *Annals of Statistics* 1992; **20**(2):1079–1093.
35. Cochran WG. The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* 1968; **24**:295–313.
36. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 1984; **79**:516–524.
37. Shadish WR, Clark MH. A randomized experiment comparing random to nonrandom assignment. Unpublished Paper, University of California, Merced, 2006.

DOI: 10.1002/sim