Statistics Surveys
Vol. 9 (2015) 106–208
ISSN: 1935-7516
DOI: 10.1214/15-SS110

Some models and methods for the analysis of observational data

José A. Ferreira

Department of Statistics, Informatics and Modelling
National Institute for Public Health and the Environment (RIVM)
Antonie van Leeuwenhoeklaan 9
3721 MA Bilthoven, The Netherlands
e-mail: jose.ferreira@rivm.nl

Abstract: This article provides a concise and essentially self-contained exposition of some of the most important models and non-parametric methods for the analysis of observational data, and a substantial number of illustrations of their application. Although for the most part our presentation follows P. Rosenbaum's book, "Observational Studies", and naturally draws on related literature, it contains original elements and simplifies and generalizes some basic results. The illustrations, based on simulated data, show the methods at work in some detail, highlighting pitfalls and emphasizing certain subjective aspects of the statistical analyses.

MSC 2010 subject classifications: Primary 62G10, 62G05, 62G15, 62G99; secondary 62P10, 62P25.
Keywords and phrases: Observational data, confounding, testing, estimation, treatment effect, stratification, propensity scores.

Received June 2015.

Contents

1. Introduction

For ethical reasons, or because of practical difficulties or impossibilities, many questions concerning the effect of a 'treatment' (a medical treatment, a state intervention, etc.) on a population cannot be studied experimentally—i.e. by assigning different forms of the treatment to virtually identical individuals, or by assigning different forms of the treatment randomly to different individuals, and then determining the treatment effect by comparing individuals who have undergone different forms of treatment with respect to some 'response' (e.g. survival time following a medical treatment). Instead, such questions have to be studied by means of observational data, namely of observations made on individuals who by their own volition or through the vicissitudes of life happen to get certain forms of the treatment. Since the characteristics or attributes of these individuals (their age, sex, economic status, health, etc.) usually contribute to determine their responses to treatment as well as the treatments assigned to them, individuals undergoing different forms of treatment generally fail to be comparable enough for it to be possible to extricate the differences in their responses that are due to treatment from those that are due to differences in their characteristics. This confounding of effects due to treatment with effects due to individual characteristics, which is a result of the inability to control the treatment assignment and of the impossibility of manipulating the characteristics of individuals, is the main obstacle to the drawing of valid conclusions about the existence and magnitude of a treatment effect from observational data. While, for example, in an experimental study involving a medicine and a placebo the effect of the medicine is easily estimated by subtracting the average response in a placebo group from the average response in a treated group—because, just before the trial, the two groups are statistically indistinguishable—in an

observational study this procedure will normally lead to biased estimates and conclusions.

Awareness of the problem of confounding and the use of analytic methods to circumvent it go back at least to the 19th century; the books by Freedman [9, 10] and Rothman [30, 31], for instance, sketch some of the historical developments connected with observational studies and describe classical examples of analyses—most of which, interestingly, required very little statistical theory—based on observational data. In addition to the classical method of stratification, there exist by now a number of more or less well-established statistical methods and underlying models designed to 'control for confounding' and which are potentially useful for the analysis of observational data. Expositions and outlines of these models and methods, from varying standpoints and at various levels of detail, can be found in a number of books and articles, among which we may single out those of Rosenbaum [26, 27], Rubin [33], Pearl [20, 21], Imbens and Rubin [15], Imbens [14], and Stuart [38]. Despite this, however, when we decided to learn the newer methods, in particular those based on the propensity score, we have found no single reference that was simultaneously concise, self-contained and uniformly clear. Moreover, we realized that our sources contained a couple of misconceptions, or at least ambiguous statements,[1] that few of them managed to conform to mathematical usage throughout, usually to the detriment of clarity,[2] and that some of the assumptions under which they presented the methods could be weakened and clarified.[3] Finally, we saw that there was a lack of consensus, which sometimes led to disputes, concerning the correctness or superiority of this or that approach.[4] This is somewhat surprising because, provided one avoids stepping too far into the treacherous ground of 'causality'—which one can afford to do when describing the methods, though not when applying them—, there is relatively little 'theory' to talk about (just a few central ideas needed to understand a few methods) and the mathematics is of undergraduate level.

The present work represents an attempt to provide a concise and essentially self-contained exposition of certain models and methods which we think are potentially useful for the analysis of observational data, and to illustrate some aspects of their application—an attempt that, hopefully, will not contribute a substantial share of obscurities, misconceptions, errors or biases. Although for

[1] For example, about the responses having to be assumed constant, about their being constant under the hypothesis of no treatment effect, or about 'how easy' it is to get accurate estimates of the propensity score.

[2] For example, confusing numerical variables with random variables, or using the same symbol to denote simultaneously a set function, a function of one variable, and a function of several variables.

[3] For example, the assumption that the treatment assignments are independent in the derivation of conditional tests is unnecessary, and the basic assumptions that make it possible to 'correct for confounding' as well as the definition of treatment effect are best formulated in terms of random variables rather than in terms of distributions, as Pearl [20, 21] has done (and has been arguing for), though in our opinion not to the same extent that we do here.

[4] See, for example, the exchange initiated by Shrier's letter [35] and pursued in Pearl [22] and elsewhere, and the claims or implications about estimators weighted by propensity scores (e.g. [19]).

the most part we have been guided by Rosenbaum [26], our presentation of the basic model for observational data in section 2 and of methods of testing and estimating a treatment effect in sections 3 and 4 contains original elements and it simplifies, generalizes and clarifies some results. A substantial portion of the article—sections 5 and 6 and the many appendices with figures—are devoted to the illustration of the methods of analysis with simulated data and to the description of the model used for simulation. We thought it very important to show the methods at work in some detail, to point out pitfalls and emphasize certain necessarily subjective aspects of the statistical analyses, for which purposes it seems best to use data simulated from a model that is neither completely unrealistic nor overly complicated. On the other hand, in our illustrations we have not attempted to use, nor purport to use, 'optimal' procedures of any kind; that would have been futile in an area where the least of all problems is lack of optimality. In particular, we do not consider methods based on parametric models, which may be optimal when the correct model for the observational data is known but are seldom justifiable in 'causal inference' problems, and we also do not consider estimation methods such as 'propensity score weighting', which may be more efficient in certain cases but whose virtues are far from being pacific—but we do offer occasional comments on these methods.

There is one aspect in which we certainly fare no better than other authors—that of realism. Indeed, we do not mention nor provide 'real-life' examples. In an article that aims to explain well-known methods and their workings as clearly as possible there is plenty of justification for that. Moreover, as demonstrated by Freedman [9, 10], and despite the optimism shared by many authors (see, for example, the conclusions of Imbens [14] and the explanations of Pearl [21] for the yet unfulfilled potential of his methods), 'the number of successful applications [of the methods described here and in the books and articles cited above] is at best quite limited'. This is not hard to understand: While the 'theory' is more or less well-established, its application can be extremely difficult, if not altogether impossible. In fact, the theory is readily applied—even if only to show that no definite conclusion can be drawn from a given data set—provided one knows which confounders must be 'corrected for' and provided the truly indispensable confounders are represented in the data set. But these provisions, in turn, presuppose the existence of some scheme—a 'causal diagram' or 'structural model' like those studied by Pearl [20, 21] and co-workers—to describe with a sufficient degree of realism the main elements and their interrelations appertaining to the question being investigated. Unfortunately, in most investigations we have met or read about—especially, and perhaps tellingly, those from areas that seem particularly anxious to use statistical methods—knowledge of the subject matter is often insufficient for the construction of a plausible causal scheme. The reader who looks for applications will find a good presentation of some realistic ones in [26, 27], for example; for our part, we hope that by helping to simplify and clarify the presentation of the methods we may contribute a little to their realistic application.

2. Basic Model

We regard a set of observational data as a sample (not necessarily a random sample) of N random vectors $(\mathbf{X}_1, T_1, R_1), (\mathbf{X}_2, T_2, R_2), \ldots, (\mathbf{X}_N, T_N, R_N)$, the i-th random vector $(\mathbf{X}_i, T_i, R_i)$ being composed of a vector $\mathbf{X}_i$ of *covariates* taking values in $\mathbb{R}^d$, a *treatment assignment* T_i with a distribution on $\mathbb{N}_0$, and a *response* R_i taking values in $\mathbb{R}$. Each such random vector is thought of as pertaining to a 'unit' or 'individual' whose treatment assignment is partly determined by (is associated/correlated with) its covariates and whose response may be determined in part by the treatment assignment. For example, in a study to investigate whether or not the regular intake of vitamin supplements increases life expectancy the vector $\mathbf{X}_i$ would stand for a number of personal characteristics or attributes of an individual labelled i, such as age, educational level, income, type of household, geographical location, sporting activities, weight, height, etc., T_i for the amount of vitamins taken by the individual, and R_i for the individual's age at death, or perhaps for quality of life at old age measured according to certain criteria. The crucial aspect of an observational study—a study based on observational data—is that the treatment is not assigned randomly to the individuals, as is the case in an experimental study; rather, it is influenced by a number of characteristics, some of which are embodied by the vector of covariates. Typically, those individuals who, due to their personal characteristics, are more likely to be treated or to undergo a more intensive form of treatment are also those who tend to have better responses. Thus, individuals who take vitamins on a regular basis may typically be more conscious about their health and may typically be better off than those who do not, and therefore will tend to live longer. Consequently, in an observational study the potential effect of the treatment on the response may be *confounded* by the characteristics of the units: even if vitamins have a positive effect on life expectancy, a greater average life expectancy observed in individuals who take vitamins may be explained, at least in part, by the individuals who choose to be treated tending to live longer than the rest of the population.

In order to be able to draw correct inferences about the effect of treatment on the response based on observational data one must somehow 'correct for' the *confounders* (the components of the $\mathbf{X}_i$s that confound the effect of the treatment on the response). Unless one possesses a sufficiently accurate formula describing how the response is affected by the treatment and by the characteristics of the units, the only general way of correcting for confounders is to compare the responses obtained under different treatments within groups of units whose values of the confounders are essentially the same, or, more specifically, within groups of units which at the outset were indistinguishable regarding their chance of being assigned a given treatment. The creation of such groups may be achieved by *stratification* or by *matching*. In stratification the data are partitioned into a number of disjoint subsets—the *strata*—each of which corresponds to a group of units with (approximately) the same values of the covariates and, ideally, a variety of levels of treatment. Matching can be seen as a more directed way of

forming homogeneous groups: as in stratification, units are grouped together—in so-called *matched sets*—if their covariate vectors are equal or similar in some sense, but the grouping may involve constraints on the characteristics of the units or on the numbers of units that undergo each level of treatment. For instance, each group may be required to contain at least two units with differing treatments, or each group may be required to contain exactly one unit with a particular treatment. Because of these constraints, matching is often carried out in parallel with the selection of units from a database or even in parallel with the sampling of data. For example, if a particular form of treatment is rare because it is not the ordinary one and can only be ethically justified under special circumstances, then the process of matching an individual undergoing the ordinary treatment to an individual undergoing the rare treatment is subordinate to the appearance as well as to the characteristics of the latter individual. Although the strata obtained by stratification may happen to coincide with the matched sets obtained by matching, stratification and matching generally yield different groups. Stratification typically discards a substantial proportion of units at the outset, namely those in strata consisting of units with a single treatment level; matching can make use of somewhat more data, though in reality that is partly at the cost of greater dissimilarity between matched units. Sometimes, stratification and matching on *propensity scores*—certain functions of the covariates defined later in this section—make use of most of the data.

Remarks. (i) In most situations of interest the treatment takes a finite number of values—the *levels* of treatment—and often only the values 0 and 1; we take the range of the T_is as $\mathbb{N}_0$ to simplify the presentation.

(ii) Examples of observational studies are given in chapter 1 of [26]. Freedman [9] provides a lucid and concise account of the difficulties posed by observational data and of the typical flaws of observational studies. The distinction between 'covariate' (or 'concomitant') and 'response' is usually clear from the problem at hand, but subsection 3.1.3 of [26] provides a discussion of possible confusions and overlaps between the two concepts. □

The exact form of the approach to correct for confounders outlined above depends on the model assumed for the observational data, namely on the assumptions about the distribution of the vector $(\mathbf{X}_i, T_i, R_i)$ pertaining to unit i in the sample and about the joint distribution of vectors pertaining to different units. Perhaps the most general *basis* for such a model is obtained by assuming that there are functions τ_i, ρ_i and uniform random variables U_i, V_i $(i = 1, \dots, N)$ such that

$$T_i = \tau_i(U_i, \mathbf{X}_i), \quad R_i = \rho_i(V_i, \mathbf{X}_i, T_i) \equiv \rho_i(V_i, \mathbf{X}_i, \tau_i(U_i, \mathbf{X}_i)), \tag{2.1}$$

$\tau_i(u, \mathbf{x})$ is not constant in $\mathbf{x}$ for fixed u nor constant in u for fixed $\mathbf{x}$, and U_i and V_i are independent conditionally on $\mathbf{X}_i$ (but neither the U_is nor the V_is need to be independent). This *basic model* says that the treatment depends on the covariates but is not merely a function of the covariates (for otherwise the response would be fully determined by the covariates and by another variable

and then the treatment would vanish from the picture), and that the response depends partly on the covariates and possibly on the treatment. To say that there is a *treatment effect* is then to say that T_i and R_i are not independent conditionally on $\mathbf{X}_i$ or, what is essentially the same, that $\rho_i(v, \mathbf{x}, t)$ is not constant in t for fixed $(v, \mathbf{x})$. Finally, the $\mathbf{X}_i$s confound the effect of the treatment on the response to the extent that the variation of $\rho_i(v, \mathbf{x}, t)$ with t is constrained by $\mathbf{x}$ through the equation $t = \tau_i(u, \mathbf{x})$.

We will expand upon the basic model below in subsection 2.1; in particular, we will explain in detail in what sense $\mathbf{X}_i$ is a confounder of the treatment, why conditioning (stratifying, matching) on it removes confounding, and why the independence of U_i and V_i conditionally on $\mathbf{X}_i$ is indispensable. Before embarking on explanations, however, let us make some observations and spell out a couple of immediate but crucial implications of the conditions around (2.1).[5] First, the basic model constitutes our most fundamental assumption—an assumption under which one can hope to draw 'unconfounded inferences' about the existence and nature of a treatment effect—and for that reason it will be *assumed to hold in everything that follows*.

Secondly, the model is consistent with—but more explicit than—the so-called Neyman-Rubin model (e.g. [26, 28]) and with the assumptions normally associated with it: If $t \neq t'$, $\rho_i(V_i, \mathbf{X}_i, t)$ and $\rho_i(V_i, \mathbf{X}_i, t')$ are two *potential responses*, and if the datum observed is $\rho_i(V_i, \mathbf{X}_i, t)$ then $\rho_i(V_i, \mathbf{X}_i, t')$ is termed a *counterfactual* of it; by the independence of U_i and V_i conditionally on $\mathbf{X}_i$, the vector $(\rho_i(V_i, \mathbf{X}_i, t), \rho_i(V_i, \mathbf{X}_i, t'))$ is independent of T_i given $\mathbf{X}_i$; and the conditions on τ_i ensure that, for each $\mathbf{x}$, $0 < P(T_i = t|\mathbf{X}_i = \mathbf{x}) < 1$ for some t. To emphasize the idea of 'causality' in equations (2.1) one may regard $\mathbf{X}_i$ as being generated first, followed by T_i and then by R_i.

Thirdly, since a uniform random variable can be 'unfolded' into a sequence of independent uniform random variables, and since the τ_is can always be thought of as functionals acting jointly on that sequence and on the $\mathbf{X}_i$s, and similarly for the ρ_is, the basic model is completely general as a model for the effect of a treatment on a response, both of which are functions of a set of other variables.

Finally, it should be kept in mind that the purpose of the basic model is not at all theoretical: in any given investigation, the 'applied researcher' is expected to stare hard at it and provide convincing arguments to the effect that everything besides the covariates (the $\mathbf{X}_i$) enters into the treatment and into the response via separate, independent ways (via the U_i and V_i). Of course, the question of whether one is entitled to separate the data into components U_i, V_i and $\mathbf{X}_i$ in such clear-cut fashion is one of metaphysics—probably, in most studies the most one can hope for is to be able to say that the dependence between U_i and V_i conditionally on $\mathbf{X}_i$ is negligible. Still, the degree of validity of an empirical investigation based on the methods described in this work depends on the degree to which the basic model may be expected to hold, so for such an investigation

[5] The reader who feels the need for a concrete example of equations (2.1) may wish to read section 5 already at this point.

to have a modicum of credibility it must be founded on careful consideration and justification of its assumptions. Unfortunately, providing justification for a given set of confounders is perhaps the most vexing task faced by observational studies that set out to establish the existence of a 'causal effect'. Although methods exist—the ones presented by Pearl [20, 21], mentioned later on—that can help in selecting confounders and in providing a rationale for a given selection, much of the effort required for presenting a convincing argument has to be based on extra-mathematical, extra-statistical knowledge appertaining to the subject of the investigation. We suspect that, when required to show that the basic model holds at least approximately, applied researchers will often have to recognize that it cannot possibly hold or that it cannot hold unless additional covariates are measured; but this realization alone should have a beneficial effect on many studies, and may even lead to real progress.

Turning to the implications of the basic model, note first that according to (2.1) varying T_i in $\rho_i(V_i, \mathbf{X}_i, T_i)$ may (but need not) cause variation in R_i; if it does then there is a treatment effect in the tautological sense that varying T_i does have the effect of varying R_i. This is a *proper* treatment effect in the sense that it is distinct from that of the covariates, i.e. is not fully determined by $\mathbf{X}_i$: thanks to the proviso about τ_i (that this vary in its first argument), two potential 'draws' or 'realizations' ω and ω' from the underlying probability space may satisfy $T_i(\omega) \neq T_i(\omega')$ even when $\mathbf{X}_i(\omega) = \mathbf{X}_i(\omega')$. In principle, this allows us to *choose* or *manipulate* the value of T_i for a fixed value of $\mathbf{X}_i$ and opens up the *possibility* of investigating whether and how the third argument of ρ_i is capable of making the response change—i.e. it opens up the possibility of investigating the existence and nature of a treatment effect.

This very concrete definition of treatment effect—expressed in terms of realizations ω from the underlying probability space and of how and to what extent the third argument of ρ_i, in combination with those realizations, makes the response vary—has an automatic translation into a statistical hypothesis. Indeed, by the independence of U_i and V_i conditionally on the value of $\mathbf{X}_i$, it follows directly from (2.1) that *there is a treatment effect if and only if* T_i *and* R_i *are dependent conditionally on the event* $\{\mathbf{X}_i = \mathbf{x}\}$ for all $\mathbf{x} \in \mathbb{R}^d$. In particular, to say that there is no treatment effect is to say that T_i *and* R_i *are independent conditionally on* $\{\mathbf{X}_i = \mathbf{x}\}$ for all $\mathbf{x} \in \mathbb{R}^d$—the *null hypothesis of no treatment effect*. Although this hypothesis refers to a generic unit i, it readily applies to an arbitrary set I of units (pertaining to a sample—random or not, stratified or not—from some population): the null hypothesis of no treatment effect relative to I holds if and only if it holds for each $i \in I$.

It follows that under our basic model a study about the existence and extent of a treatment effect amounts to a study of the joint distribution of T_i and R_i conditionally on $\mathbf{X}_i$. Because this joint distribution is determined by the probabilities

$$\begin{aligned} P(T_i = t, R_i \leq r | \mathbf{X}_i = \mathbf{x}) &= P(\tau_i(U_i, \mathbf{x}) = t, \rho_i(V_i, \mathbf{x}, t) \leq r | \mathbf{X}_i = \mathbf{x}) \\ &= P(\rho_i(V_i, \mathbf{x}, t) \leq r | \mathbf{X}_i = \mathbf{x}) f_i^{(\mathbf{x})}(t), \end{aligned}$$

where we write

$$f_i^{(\mathbf{x})}(t) = P(T_i = t|\mathbf{X}_i = \mathbf{x}),$$

we see that the null hypothesis of no treatment effect holds for unit i if and only if $P(\rho_i(V_i, \mathbf{x}, t) \le r|\mathbf{X}_i = \mathbf{x})$ is a function of r and $\mathbf{x}$ alone (is independent of t).

Probabilities such as $f_i^{(\mathbf{x})}(t)$ ($t \in \mathbb{N}_0$) will be referred to as *propensity scores*; they constitute a probability distribution indexed by $\mathbf{x}$ and i and as such will sometimes be referred to as *the* propensity score (of unit i). There will be no risk of confusion if we also refer to the random variables $f_i^{(\mathbf{X}_i)}(t)$, which as we shall see are sometimes covariates in their own right, as propensity scores (of unit i). Finally, as a vector or scalar function of $\mathbf{x}$ the value(s) of $f_i^{(\mathbf{x})}(t)$ for t in a subset of $\mathbb{N}_0$ will sometimes be called the propensity score *function*. Propensity scores play an important role in much of what follows because of the following further consequence of the basic model: Assuming for simplicity that $\mathbf{X}_i$ is discrete, writing f for a generic probability function on $\mathbb{N}_0$, abbreviating the statement $f_i^{(\mathbf{x})}(t) = f(t)$ for all t such that $f_i^{(\mathbf{x})}(t) > 0$ as $f_i^{(\mathbf{x})} = f$, and using $\sum_f$ to indicate summation over $\mathbf{x}$ such that $f_i^{(\mathbf{x})} = f$, we have

$$P\left(T_i = t, R_i \le r \middle| f_i^{(\mathbf{X}_i)} = f\right) = \frac{P\left(T_i = t, R_i \le r, f_i^{(\mathbf{X}_i)} = f\right)}{P\left(f_i^{(\mathbf{X}_i)} = f\right)} =$$

$$\frac{\sum_f P(\tau_i(U_i, \mathbf{x}) = t, \rho_i(V_i, \mathbf{x}, t) \le r, \mathbf{X}_i = \mathbf{x})}{P\left(f_i^{(\mathbf{X}_i)} = f\right)} =$$

$$\frac{\sum_f P(\rho_i(V_i, \mathbf{x}, t) \le r|\mathbf{X}_i = \mathbf{x}) P(T_i = t|\mathbf{X}_i = \mathbf{x}) P(\mathbf{X}_i = \mathbf{x})}{P\left(f_i^{(\mathbf{X}_i)} = f\right)} =$$

$$\frac{\sum_f P(\rho_i(V_i, \mathbf{x}, t) \le r|\mathbf{X}_i = \mathbf{x}) f_i^{(\mathbf{x})}(t) P(\mathbf{X}_i = \mathbf{x})}{P\left(f_i^{(\mathbf{X}_i)} = f\right)},$$

whence, using the fact that $f_i^{(\mathbf{x})}(t) = f(t)$ for the $\mathbf{x}$ under the summation sign,

$$P\left(T_i = t, R_i \le r \middle| f_i^{(\mathbf{X}_i)} = f\right) = f(t) \frac{\sum_f P(\rho_i(V_i, \mathbf{x}, t) \le r|\mathbf{X}_i = \mathbf{x}) P(\mathbf{X}_i = \mathbf{x})}{P\left(f_i^{(\mathbf{X}_i)} = f\right)}. \quad (2.2)$$

Letting $r \to \infty$ and using $\sum_f P(\mathbf{X}_i = \mathbf{x}) = P\left(f_i^{(\mathbf{X})} = f\right)$ this yields in particular

$$P\left(T_i = t \middle| f_i^{(\mathbf{X}_i)} = f\right) = f(t), \quad (2.3)$$

whence (take $f = f_i^{(\mathbf{x})}$) the distribution of the treatment is the same conditionally on $\mathbf{X}_i = \mathbf{x}$ and conditionally on $f_i^{(\mathbf{X}_i)} = f_i^{(\mathbf{x})}$—that is, $P(T_i = t|\mathbf{X}_i = \mathbf{x}) = P(T_i = t|f_i^{(\mathbf{X})} = f_i^{(\mathbf{x})})$. Moreover, as just seen, the probabilities $P(\rho_i(V_i, \mathbf{x}', t) \le$

$r|\mathbf{X}_i = \mathbf{x})$ appearing on the right side of (2.2) under the summation sign depend on t if and only if there is a treatment effect; so it follows from (2.2)–(2.3) that T_i *and* R_i *are independent conditionally on the propensity score if and only if there is no treatment effect*. What this says is that under the basic model inferences on the treatment effect can be drawn not only by conditioning (stratifying, matching) on the covariates but also by conditioning (stratifying, matching) on the associated propensity score—although, as will be seen in sections 3 and 4, statistical procedures based on the propensity score require stronger assumptions than do statistical procedures based on the covariates.

This observation is a variant of a result of Rosenbaum and Rubin [28], who first advocated the use of the propensity score in observational studies. Since $\mathbf{X}_i = \tilde{\mathbf{X}}_i$ implies $f_i^{(\mathbf{X}_i)}(t) = f_i^{(\tilde{\mathbf{X}}_i)}(t)\ \forall t$ but the converse is not true, a stratification based on the propensity score will be coarser, and hence yield bigger and more useable strata (strata containing units with different levels of treatment), than a stratification based on $\mathbf{X}_i$ whenever the latter is high-dimensional and the former low-dimensional; and for the same reason it will usually be easier to find matches based on propensity scores than based on the covariates.[6] The logic of the Rosenbaum-Rubin approach to observational studies is thus to replace the typically high-dimensional vector of covariates by the typically low-dimensional propensity score in order to increase the number and the size of useable strata. This approach is particularly attractive when the T_is are binary, say with 1 indicating that the unit is treated and 0 that it is a 'control', for in that case stratification is based on a single variable, namely $f_i^{(\mathbf{X}_i)}(1)$ (the probability that unit i receives treatment conditionally on it having covariate value $\mathbf{X}_i$), and the resulting strata often contain practically all the data. However, one does not normally get something for nothing, and there is a price to pay for this reduction in dimensionality: the estimation of the propensity score, which is unknown in virtually all applications. Subsection 2.2 below outlines the problem of estimating propensity scores; the method of stratification on estimated propensity scores is illustrated in subsections 6.3 and 6.4.

In the rest of the paper we shall consider methods based on conditioning (stratifying, matching) on the d-dimensional vector $\mathbf{X}_i$ involved in the basic model; let it be said once and for all that these methods *apply without change with the propensity score of unit* i *in place of* $\mathbf{X}_i$ at the cost of somewhat stronger assumptions, to be formulated in each case. Evidently, when conditioning on *estimated* propensity scores, or when stratifying continuous data into $\mathbb{R}^d$ cells (rectangles), the methods will apply at most in an approximate sense.

To anticipate a little, let us mention that the methods in question—the methods of testing and estimation presented in sections 3 and 4—require assumptions about the joint distribution of the responses and/or treatments *from different units*; so far, we have found no need for assumptions of this sort, and the basic model is really a collection of analogous but generally different statements, each statement concerning one unit (in particular, the treatment effect may exist only for some of the units).

[6] Practical aspects of stratification and matching are discussed in subsections 3.3 and 3.4.

Remarks. (i) The conditional independence of U_i and V_i in the basic model can be regarded as an explicit version of a condition proposed by [2] (cf. 'Definition 1', also in [3]), which of course is itself related to the conditions of the Neyman-Rubin model.

(ii) Some of the preceding statements (e.g. the validity of the Rosenbaum-Rubin result when $\mathbf{X}_i$ has a continuous distribution) and certain statements involving conditional probabilities or expectations that appear below can be made more precise (e.g. by mentioning continuity conditions on $\mathbf{x} \to f_i^{(\mathbf{x})}$); however, in a work like this it would be somewhat pedantic to do so.

(iii) Observe that in (2.1) R_i need not vary with V_i, so the response could be constant conditionally on $(\mathbf{X}_i, T_i)$. This is as it should be because, while units are often drawn randomly from a population, in some applications it is more appropriate to think of them as specific entities that may respond differently to different levels of treatment. Of course, everything that has been said above holds true if ρ_i is constant in the first argument. □

2.1. Explication of the basic model

In order to describe the sense in which $\mathbf{X}_i$ is a confounder of the treatment let ω and ω' represent two draws from the underlying probability space—two potential 'runs of reality'—satisfying $T_i(\omega) = 1$ and $T_i(\omega') = 0$ (say) and which may be called counterfactuals of each other. For concreteness (but without loss of generality) assume that $\mathbf{X}_i$ is one-dimensional and that $\mathbf{x} \to \tau_i(u, \mathbf{x})$ has inverse $\tau_{i,u}^{-1}$, so that we can write the first equation in (2.1) as $\mathbf{X}_i = \tau_{i,U_i}^{-1}(T_i)$. If we cannot see $\mathbf{X}_i(\omega)$ nor $\mathbf{X}_i(\omega')$ (hence do not condition on their values being equal or, in other words, do not stratify/match on the random variable $\mathbf{X}_i$) then we cannot see whether and to what extent a possible difference between the two *potential outcomes*

$$\begin{aligned} R_i(\omega) &= \rho_i(V_i(\omega), \mathbf{X}_i(\omega), T_i(\omega)) = \rho_i(V_i(\omega), \tau_{i,U_i(\omega)}^{-1}(1), 1), \\ R_i(\omega') &= \rho_i(V_i(\omega'), \mathbf{X}_i(\omega'), T_i(\omega')) = \rho_i(V_i(\omega'), \tau_{i,U_i(\omega')}^{-1}(0), 0), \end{aligned} \tag{2.4}$$

is due to the difference in treatment (the difference in the third argument of ρ_i) or to a possible difference between $\tau_{i,U_i(\omega)}^{-1}(1)$ and $\tau_{i,U_i(\omega')}^{-1}(0)$ (which occur here in the second argument of ρ_i). Indeed, while the unobservable difference between $V_i(\omega)$ and $V_i(\omega')$ (which occur in the first argument of ρ_i) can, in principle, be averaged out by drawing 'runs of reality' (ω, ω'), the difference between $\tau_{i,U_i(\omega)}^{-1}(1)$ and $\tau_{i,U_i(\omega')}^{-1}(0)$ cannot, by the simple reason that $u \to \tau_{i,u}^{-1}(1)$ and $u \to \tau_{i,u}^{-1}(0)$ are different functions (in contrast, the difference between $\tau_{i,U_i(\omega)}^{-1}(1)$ and $\tau_{i,U_i(\omega')}^{-1}(1)$, for example, can in principle be averaged out), and it follows that the difference between the two potential outcomes cannot be averaged out in order to exhibit the difference due to having a 1 instead of a 0 in the treatment variable. In contrast, if we are able to see $\mathbf{X}_i(\omega)$ and $\mathbf{X}_i(\omega')$ then we can focus on draws (ω, ω') for which $T_i(\omega) \neq T_i(\omega')$ but $\mathbf{X}_i(\omega) = \mathbf{X}_i(\omega') = \mathbf{x}$ (say), in

which case a possible difference between the two potential outcomes,

$$\begin{aligned} R_i(\omega) &= \rho_i(V_i(\omega), \mathbf{X}_i(\omega), T_i(\omega)) = \rho_i(V_i(\omega), \mathbf{x}, 1), \\ R_i(\omega') &= \rho_i(V_i(\omega'), \mathbf{X}_i(\omega'), T_i(\omega')) = \rho_i(V_i(\omega'), \mathbf{x}, 0), \end{aligned} \tag{2.5}$$

can, *in principle*, be averaged out in (ω, ω') to bring out the difference due to having a 1 instead of a 0 in the third argument of the response function.[7]

We have emphasized the cautionary 'in principle' because the averaging out over (ω, ω') is possible only if the conditioning on the events

$$\begin{aligned} \mathbf{X}_i(\omega) = \mathbf{X}_i(\omega') = \mathbf{x}, \\ 1 = T_i(\omega) \equiv \tau_i(U_i(\omega), \mathbf{x}) \neq \tau_i(U_i(\omega'), \mathbf{x}) \equiv T_i(\omega') = 0 \end{aligned} \tag{2.6}$$

does not create a systematic difference between the $V_i(\omega)$ and $V_i(\omega')$ involved in (2.5) above. If the values of $U_i(\omega)$ and $U_i(\omega')$ cannot tell us anything about the values of $V_i(\omega)$ and $V_i(\omega')$ then it is also the case that the second constraint in (2.6) cannot tell us anything about the values of $V_i(\omega)$ and $V_i(\omega')$, and then we can average out the difference between the two responses in (2.5) that is due to V_i and identify the difference that is due to having a 1 instead of a 0 in the third argument of the response function. If, on the contrary, U_i and V_i are entangled (in the sense that knowing the value of one tells us something about the probable range of the other), say that $U_i = \varphi(\tilde{U}_i, Y_i)$ and $V_i = \varphi(\tilde{V}_i, Y_i)$ with $\tilde{U}_i, \tilde{V}_i$ and Y_i independent standard uniforms and φ some $\mathbb{R}^2$-valued function, then, assuming for concreteness (but without loss of generality) that all the necessary inverses exist (hence writing $Y_i(\omega) = \varphi^{-1}_{\tilde{U}_i(\omega)}(U_i(\omega))$, and so on), the second constraint in (2.6) can be solved as

$$U_i(\omega) = \tau^{-1}_{i,\mathbf{x}}(1), \quad U_i(\omega') = \tau^{-1}_{i,\mathbf{x}}(0),$$

and used in (2.5) to yield

$$\begin{aligned} R_i(\omega) = \rho_i(V_i(\omega), \mathbf{x}, 1) &= \rho_i\left(\varphi(\tilde{V}_i(\omega), Y_i(\omega)), \mathbf{x}, 1\right) \\ &= \rho_i\left(\varphi(\tilde{V}_i(\omega), \varphi^{-1}_{\tilde{U}_i(\omega)}(U_i(\omega))), \mathbf{x}, 1\right) \\ &= \rho_i\left(\varphi(\tilde{V}_i(\omega), \varphi^{-1}_{\tilde{U}_i(\omega)}(\tau^{-1}_{i,\mathbf{x}}(1))), \mathbf{x}, 1\right) \end{aligned}$$

and

$$\begin{aligned} R_i(\omega') = \rho_i(V_i(\omega'), \mathbf{x}, 0) &= \rho_i\left(\varphi(\tilde{V}_i(\omega'), Y_i(\omega')), \mathbf{x}, 0\right) \\ &= \rho_i\left(\varphi(\tilde{V}_i(\omega'), \varphi^{-1}_{\tilde{U}_i(\omega')}(U_i(\omega'))), \mathbf{x}, 0\right) \\ &= \rho_i\left(\varphi(\tilde{V}_i(\omega'), \varphi^{-1}_{\tilde{U}_i(\omega')}(\tau^{-1}_{i,\mathbf{x}}(0))), \mathbf{x}, 0\right). \end{aligned}$$

[7] Note that we cannot see what τ^{-1}_{i,U_i} is doing with the treatments 1 and 0 in the second argument of the responses in (2.4); in contrast, both responses in (2.5) are being forced a given $\mathbf{x}$ in the second argument.

But then, just as in the case of (2.4) considered earlier, a possible difference between these two potential outcomes is due to a difference in treatment and to a difference between $\varphi^{-1}_{\tilde{U}_i(\omega)}(\tau^{-1}_{i,\mathbf{x}}(1))$ and $\varphi^{-1}_{\tilde{U}_i(\omega')}(\tau^{-1}_{i,\mathbf{x}}(0))$ (occurring in the second argument of the response function), and the latter difference cannot be averaged out. To remedy the situation one can add Y_i to $\mathbf{X}_i$ (i.e. to 'correct for the confounder' $(\mathbf{X}_i, Y_i)$) and thus focus on draws (ω, ω') for which $T_i(\omega) \neq T_i(\omega')$ but $\mathbf{X}_i(\omega) = \mathbf{X}_i(\omega') = \mathbf{x}$ and $Y_i(\omega) = Y_i(\omega') = y$ (say), for then the two potential outcomes are

$$\begin{aligned} R_i(\omega) &= \rho_i(V_i(\omega), \mathbf{x}, 1) = \rho_i(\varphi(\tilde{V}_i(\omega), y), \mathbf{x}, 1) =: \tilde{\rho}_i(\tilde{V}_i(\omega), (\mathbf{x}, y), 1), \\ R_i(\omega') &= \rho_i(V_i(\omega'), \mathbf{x}, 0) = \rho_i(\varphi(\tilde{V}_i(\omega'), y), \mathbf{x}, 0) =: \tilde{\rho}_i(\tilde{V}_i(\omega'), (\mathbf{x}, y), 0), \end{aligned} \tag{2.7}$$

say, and the difference between them can be averaged out in (ω, ω') to bring out the difference due to having a 1 instead of a 0 in the third argument of the response—because, $\tilde{U}_i$ and $\tilde{V}_i$ being independent, the events

$$\mathbf{X}_i(\omega) = \mathbf{X}_i(\omega') = \mathbf{x}, \quad Y_i(\omega) = Y_i(\omega') = y,$$

$$\tilde{\tau}_i(\tilde{U}_i(\omega), (\mathbf{x}, y)) := \tau_i(\varphi(\tilde{U}_i(\omega), y), \mathbf{x}) = 1$$

and

$$\tilde{\tau}_i(\tilde{U}_i(\omega'), (\mathbf{x}, y)) := \tau_i(\varphi(\tilde{U}_i(\omega'), y), \mathbf{x}) = 0$$

inform nothing about the $\tilde{V}_i(\omega)$ and $\tilde{V}_i(\omega')$ figuring in (2.7).

This argumentation, based on 'realizations' (draws (ω, ω') from the underlying probability space), has a parallel but more efficient expression in terms of probability laws: First, the independence of U_i and V_i conditional on $\mathbf{X}_i$ ensures that the probability law of R_i conditional on $\mathbf{X}_i = \mathbf{x}$ and $T_i = t$ is equal to the probability law of $\rho_i(V_i, \mathbf{x}, t)$ conditional on $\mathbf{X}_i = \mathbf{x}$, that is

$$\begin{aligned} \mathcal{L}(R_i | \mathbf{X}_i = \mathbf{x}, T_i = t) &= \mathcal{L}(\rho_i(V_i, \mathbf{x}, t) | \mathbf{X}_i = \mathbf{x}, \tau_i(U_i, \mathbf{x}) = t) \\ &= \mathcal{L}(\rho_i(V_i, \mathbf{x}, t) | \mathbf{X}_i = \mathbf{x}), \end{aligned}$$

and it opens up the possibility of studying the effect of the treatment on the response by 'fixing' $\mathbf{x}$ and 'varying' t. Secondly, the violation of the conditions of the basic model and its rectification through the inclusion of additional confounders can be concisely illustrated by the following elaboration of an example mentioned by [35] (see also [22]).

Example 2.1. Let U, V, W, Y be independent standard uniform variables and set $U' = \varphi(U, Y)$, $V' = \varphi(V, Y)$, where, for example, $\varphi(u, y) = \log(y)/\log(u\,y)$; then U' and V' are *dependent* standard uniform variables. Suppose that the data $\mathbf{X}$, T, R on an arbitrary unit satisfy

$$T = \tau(U', \mathbf{X}), \quad R = \rho(V', \mathbf{X}, T), \quad \mathbf{X} = \chi(W, Y),$$

where τ is defined in terms of a continuous function $F : \mathbb{R}^d \to]0, 1[$ by $\tau(u, \mathbf{x}) = 1$ if $F(\mathbf{x}) \geq u$ and $\tau(u, \mathbf{x}) = 0$ otherwise, and ρ and χ are certain functions. The probability law of R conditional on $\mathbf{X} = \mathbf{x}$ and $T = 1$ satisfies

$$\begin{aligned}\mathcal{L}(R|\mathbf{X}=\mathbf{x},T=1) &= \mathcal{L}(\rho(V',\mathbf{x},1)|\mathbf{X}=\mathbf{x},T=1)\\ &= \mathcal{L}(\rho(V',\mathbf{x},1)|\chi(W,Y)=\mathbf{x},\tau(U',\mathbf{x})=1)\\ &= \mathcal{L}(\rho(V',\mathbf{x},1)|\chi(W,Y)=\mathbf{x},F(\mathbf{x})\geq\varphi(U,Y))\\ &= \mathcal{L}\left(\rho(\varphi(V,Y),\mathbf{x},1)\middle|\chi(W,Y)=\mathbf{x},Y\leq\exp\left\{\frac{F(\mathbf{x})\log U}{1-F(\mathbf{x})}\right\}\right),\end{aligned}$$

and similarly

$$\mathcal{L}(R|\mathbf{X}=\mathbf{x},T=0)=\mathcal{L}\left(\rho(\varphi(V,Y),\mathbf{x},0)\middle|\chi(W,Y)=\mathbf{x},Y>\exp\left\{\frac{F(\mathbf{x})\log U}{1-F(\mathbf{x})}\right\}\right).$$

Thus, an investigation of a treatment effect *based on* $(\mathbf{X},T,R)$, which would necessarily consist of comparing $\mathcal{L}(R|\mathbf{X}=\mathbf{x},T=1)$ with $\mathcal{L}(R|\mathbf{X}=\mathbf{x},T=0)$, would compare the probability laws on the right of the two identities above. But these laws differ not only with respect to the third argument of ρ, where treatment potentially exerts its influence, but also with respect to the first argument, since the law of $\varphi(V,Y)$ conditioned on the event that $\chi(W,Y)=\mathbf{x}$ and $Y\leq\exp\{\log(U)F(\mathbf{x})/(1-F(\mathbf{x}))\}$ generally differs from the law of $\varphi(V,Y)$ conditioned on the event that $\chi(W,Y)=\mathbf{x}$ and $Y>\exp\{\log(U)F(\mathbf{x})/(1-F(\mathbf{x}))\}$: even if $\rho(v,\mathbf{x},t)$ were constant in t the two laws would generally differ and their difference would point, erroneously, to the existence of a treatment effect.

This unpleasant result is a reflection of the fact that in general U' and V' are dependent conditionally on $\mathbf{X}$. However, if we replace $\mathbf{X}$ by $\tilde{\mathbf{X}}:=(W,Y)$ in our definitions and write

$$T=\tilde{\tau}(U,\tilde{\mathbf{X}}):=\tau(\varphi(U,Y),\mathbf{X}),\quad R=\tilde{\rho}(V,\tilde{\mathbf{X}},T):=\rho(\varphi(V,Y),\mathbf{X},T),$$

then the basic model holds (U and V being independent), and comparing

$$\begin{aligned}\mathcal{L}(R|\tilde{\mathbf{X}}=(w,y),T=1)&=\mathcal{L}\left(R\middle|(W,Y)=(w,y),y>\exp\left\{\frac{F(\chi(w,y))\log U}{1-F(\chi(w,y))}\right\}\right)\\ &=\mathcal{L}(\rho(\varphi(V,y),\chi(w,y),1))\end{aligned}$$

with

$$\mathcal{L}(R|\tilde{\mathbf{X}}=(w,y),T=0)=\mathcal{L}(\rho(\varphi(V,y),\chi(w,y),0))$$

should lead to unbiased conclusions.

To interpret the difference between conditioning on $\mathbf{X}$ and conditioning on $\tilde{\mathbf{X}}$ we may say that although $\mathbf{X}$ properly *entangles* T and R, there is a residual element of entanglement, Y, that underlies $\mathbf{X}$, is not fully accounted by it, and acts separately from $\mathbf{X}$ on the treatment (through $\varphi(U,Y)$) and on the response (through $\varphi(V,Y)$). By completing $\mathbf{X}$ with Y—equivalently: replacing it by $\tilde{\mathbf{X}}:=(W,Y)$—no residual entanglement remains between treatment and response: what remains is U and V, which act independently on T and R.

Interestingly, the determination of the correct set of confounders may appear to be less crucial when the response is constant conditionally on the confounders

and on treatment (when ρ is constant in its first argument; cf. remark (iii) before the beginning of the present subsection), for no biases of the sort illustrated here obtain. But, of course, the assumption of constancy itself requires a judicious, if not exhausting, consideration of the relevant factors. Thus, for example, the idea that a plot of land is 'predetermined' in the sense of its yield under each of several different treatments being (approximately) predictable presupposes—in particular—the knowledge of very many of its soil characteristics. □

One moral that can be drawn from this example (which, like all morals, need not offer consolation at all times) is that one should strive to take *all* the confounders into account, for bias can result not only from neglecting confounders but also from taking account of surrogates of neglected ones: $\mathbf{X} = \chi(W, Y)$ is a surrogate of the confounders W and Y, which lie deeper and must be conditioned upon if one is to draw unbiased inferences about the treatment effect.

When we talk of taking all the confounders into account, however, what we really mean is including as many as possible of the potentially confounding covariates in the basic model around (2.1) or in a causal diagram (cf. the introduction) underlying it, not necessarily using them in the subsequent statistical analysis. For it may happen that several different sets of covariates—among which there may be a smaller or more convenient set—satisfy the assumptions of the basic model and can therefore be used to investigate the treatment effect. Pearl [20, 21] has developed criteria and algorithms to determine such subsets of covariates on the basis of a causal diagram. Alternatively, at least if the causal diagram is not too complex, one can also identify the relevant subsets of covariates by inspection of the law of R conditional on $\mathbf{X} = \mathbf{x}$ and $T = t$; an illustration of this procedure is given in appendix A. Finally, we must emphasize that taking *more* potential confounders is not necessarily better, for conditioning on certain variables may actually create or increase bias: in example 2.1, conditioning on W and Y takes care of confounding, but conditioning on an arbitrary number of covariates $f_1(\mathbf{X}), f_2(\mathbf{X}), \ldots$ defined by functions $f_1, f_2, \ldots$ generally does not; see also [22] and [35].

2.2. Stratification on the estimated propensity score

We have seen that under the basic model the propensity score—a covariate in its own right—can, in principle, replace the covariate vector to great advantage in drawing inferences about the treatment effect. Thus, if the propensity score can be accurately estimated then stratification or matching on the *estimated* propensity score should yield practically unbiased inferences. Evidently, the estimation of the propensity score requires assumptions about the joint distribution of the data from different units—assumptions that are similar to those needed for testing and estimating the treatment effect, introduced in sections 3 and 4, respectively.

Probably the least that needs to be assumed is that conditionally on the $\mathbf{X}_i$s the T_is have the same distribution, so that the propensity score $f^{(\mathbf{x})} \equiv f_i^{(\mathbf{x})}$ is the same for all units sharing the same value $\mathbf{x}$ of their covariate. Once this

holds, and even if the T_is are dependent,[8] one can hope to estimate $f^{(\mathbf{x})}$ from the N pairs $(\mathbf{X}_i, T_i)$, for instance by a non-parametric regression estimator (a Nadaraya-Watson estimator, a nearest neighbour estimator, a random forest estimator, or any other estimator that is consistent or at least has some chance of being consistent) when N is sufficiently large, or by a sufficiently flexible parametric or semi-parametric model.

There seems to be a widespread belief that propensity score estimates are always warranted (e.g. in the case of binary treatments, by taking a logistic regression model for $f^{(\mathbf{x})}(1)$ and estimating its parameters from the data) or at least that the biased or inaccurate estimation of the propensity score does not represent a serious problem for the testing of treatment effects (see, for example, p. 7 of [38]). However, we have found no convincing evidence in the literature that this is so, and the careful study of [11] indicates that in realistic settings estimated propensity scores can lead to high bias in estimates of treatment effects. Moreover, we see no reasons to expect that inferences based on estimated propensity scores dispense with the consistency of propensity score estimators.

3. Testing for a treatment effect

The actual investigation of a treatment effect—in particular the *testing* for a treatment effect—requires further assumptions, in addition to those of the basic model.[9] The first additional assumption is really a more stringent version of the independence of U_i and V_i conditionally on $\mathbf{X}_i$ contained in the basic model; in particular, it is designed to ensure that the conditioning on several units having covariate vectors equal to the same $\mathbf{x} \in \mathbb{R}^d$—units which fall into the same stratum or matched set—does not destroy the independence between the U_is and V_is of those units, and it *will be assumed throughout this section*:

A0 For every $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$, conditionally on the event that $\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_n}$ are all equal to a given arbitrary $\mathbf{x} \in \mathbb{R}^d$, the random vectors $(U_{i_1}, U_{i_2}, \ldots, U_{i_n})$ and $(V_{i_1}, V_{i_2}, \ldots, V_{i_n})$ are independent. Moreover, if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K \in \mathbb{R}^d$ are distinct and $I_1, I_2, \ldots, I_K$ form a partition of $I = \{i_1, i_2, \ldots, i_n\}$ then for all $(u_i)_{i \in I}, (v_i)_{i \in I} \in \mathbb{R}^n$,

$$P(U_i \leq u_i, V_i \leq v_i, i \in I | \mathbf{X}_i = \mathbf{x}_j, i \in I_j, j = 1, 2, \ldots, K) =$$

$$\prod_{j=1}^{K} P(U_i \leq u_i, V_i \leq v_i, i \in I_j | \mathbf{X}_i = \mathbf{x}_j, i \in I_j).$$

The second part of this assumption is intended to make inferences from different strata/matched sets independent of each other; in particular, the conditioning event of one stratum implies nothing about the treatment assignments and the responses of other strata. The role of the first part will become clear

[8] Though, as will be seen in section 3, statistical procedures based on the propensity score probably require independence.

[9] Alternative assumptions will be introduced in section 4 for purposes of estimating a treatment effect.

in a moment, but let us note that it is not implausible that the information that several units (as opposed to one unit in isolation) have covariate vectors equal to a given $\mathbf{x}$ will 'entangle' U_i and V_i: for example, the fact that *at a given moment* there are so many patients who possess the same set of personal characteristics and who therefore are equally eligible for treatment might lead a doctor responsible for the treatment assignment to pick the patients to be treated on the basis of a new covariate (a 'tie-breaker'), which could act both on U_i and on V_i.

Write

$$f_I^{(\mathbf{x})}((t_i)_{i\in I}) = P(T_i = t_i, i \in I | \mathbf{X}_i = \mathbf{x}, i \in I), \tag{3.1}$$

$(t_i)_{i\in I} = (t_{i_1}, t_{i_2}, \ldots, t_{i_n}) \in \mathbb{N}_0^n$, for the probability function of the treatments assigned to an arbitrary set $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$ of n units, conditional on the event that their covariate vectors all take the value $\mathbf{x}$; when $I = \{i\}$ this reduces to the propensity score of section 2. Recall that the null hypothesis of no treatment effect relative to the units $1, 2, \ldots, N$ holds if and only if T_i and R_i are independent conditionally on $\mathbf{X}_i$ for all i, and that this is equivalent to all $\rho_i(v, \mathbf{x}, t)$ being constant in t for fixed $(v, \mathbf{x})$. By **A0** we have, for $(r_i)_{i\in I} \in \mathbb{R}^n$,

$$P(T_i = t_i, R_i \le r_i, i \in I | \mathbf{X}_i = \mathbf{x}, i \in I) =$$

$$f_I^{(\mathbf{x})}((t_i)_{i\in I}) P(\rho_i(V_i, \mathbf{x}, t_i) \le r_i, i \in I | \mathbf{X}_i = \mathbf{x}, i \in I).$$

Since the null hypothesis holds relative to $I = \{i_1, i_2, \ldots, i_n\}$ if and only if $\rho_i(V_i, \mathbf{x}, t)$ is constant in t for all $i \in I$, it follows from this factorization that $(T_i)_{i\in I} \equiv (T_{i_1}, T_{i_2}, \ldots, T_{i_n})$ and $(R_i)_{i\in I} \equiv (R_{i_1}, R_{i_2}, \ldots, R_{i_n})$ are independent conditionally on $\{\mathbf{X}_i = \mathbf{x}, i \in I\}$ for all $\mathbf{x}$ if and only if the null hypothesis holds relative to I (if and only if the last probability above is a function of the r_is and $\mathbf{x}$ alone). Consequently, the null hypothesis holds relative to $\{1, 2, \ldots, N\}$ if and only if the vectors $(T_i)_{i\in I}$ and $(R_i)_{i\in I}$ are independent conditionally on $\{\mathbf{X}_i = \mathbf{x}, i \in I\}$ for all $I \subset \{1, 2, \ldots, N\}$ and all $\mathbf{x}$.

The role of the first part of **A0** is thus to allow one to *formulate* the null hypothesis relative to a given set of units in terms of the conditional independence of vectors of treatments and vectors of responses, and hence in principle to *test it* on the basis of such vectors. Of course, the null hypothesis relative to a set of units is either true or false irrespectively of whether **A0** holds or not (even irrespectively of whether U_i and V_i are independent conditionally on the value of $\mathbf{X}_i$ for each i, as the basic model demands); but **A0** provides us with a form of sampling or replication (the basis of most statistical methods), namely with two or more pairs of treatments and responses per stratum—something that the conditional independence of the basic model cannot do because it concerns a single unit at a time. The second part of **A0** extends the range of this sampling/replication, varying the 'conditions' $\mathbf{x}$ under which hypotheses about treatment effects can be considered and tested.

The other assumptions that need to be introduced should provide us with a *null distribution* for testing the conditional independence of a vector of treat-

ments and a vector of responses per stratum/matched set.[10] There are at least two ways of proceeding, each of which covers a class of observational studies commonly encountered. The first is to assume that

A1 For every $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$, conditionally on the event that $\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_n}$ are all equal to a given arbitrary $\mathbf{x} \in \mathbb{R}^d$, the random variables $T_{i_1}, T_{i_2}, \ldots, T_{i_n}$ are exchangeable in the sense that for $t_{i_1}, t_{i_2}, \ldots, t_{i_n} \in \mathbb{N}_0$

$$P(T_i = t_i, i \in I | \mathbf{X}_i = \mathbf{x}, i \in I) = P(T_i = t'_i, i \in I | \mathbf{X}_i = \mathbf{x}, i \in I)$$

whenever $(t'_{i_1}, t'_{i_2}, \ldots, t'_{i_n})$ results from a permutation of the coordinates of $(t_{i_1}, t_{i_2}, \ldots, t_{i_n})$.

This assumption implies that the treatment assignment among units characterized by equal covariate vectors is equally likely to be $(1, 0, 0, 2, \ldots, 0, 1, 0, 1)$, say, as $(2, 0, 0, 1, \ldots, 1, 0, 1, 0)$, or any other vector obtained by permuting the coordinates of these vectors. It corresponds to the assumption that the observational study has *overt bias* or is *free of hidden bias* (cf. section 3.2 of [26]), and it certainly holds if the τ_is in (2.1) are all equal and the U_is are exchangeable; however, it is easy to see that **A1** may hold even when the basic model does not, so it alone does not guarantee the possibility of drawing 'unconfounded inferences'. As will be seen in subsection 3.1, **A1** (together with **A0**) implies that if units are stratified or matched on their covariates then the assignment of treatments within a stratum or matched set is equivalent to the assignment of treatments in an experimental study, whence the observational data may be treated by the same methods as those used to treat experimental data.

Assumption **A1** means that the $f_I^{(\mathbf{x})}$ of (3.1) is a symmetric function: permutation of its arguments does not change its value; the dependence on the set I could, for example, indicate a group effect, or a group *size* effect, on the distribution of the treatments. Clearly, for each I, $P(T_j = t | \mathbf{X}_i = \mathbf{x}, i \in I) = P(T_k = t | \mathbf{X}_i = \mathbf{x}, i \in I)$ for all t and all $j, k \in I$; that is, all units with the same values of the covariate vectors have the same marginal conditional distribution of treatment.[11]

[10] To see the need for further assumptions suppose that τ_i varies systematically with i; for example, say that i is a time index related to the moment when a unit arrives for (possible) treatment, that τ_i tends to become smaller on average with increasing i, and that this has nothing to do with the characteristics $\mathbf{X}_i$ of the successive units but rather with some 'exogenous' trend in treatment policy. Then the basic model still offers the theoretical possibility of drawing inferences about the treatment effect, but the possibility cannot be realized unless one possesses the correct model for τ_i. Similarly, if $\tau_i = \tau$ for all i but the U_i, though uniform, are dependent and not exchangeable, then the drawing of inferences requires the knowledge of the joint distribution of $U_1, \ldots, U_N$.

[11] As examples of $f_I^{(\mathbf{x})}((t_i)_{i \in I})$ for $I = \{1, 2, \ldots, n\}$ we cite $p(\mathbf{x})^{\sum_{i=1}^n t_i} (1 - p(\mathbf{x}))^{\sum_{i=1}^n (1 - t_i)}$ for $t_i \in \{0, 1\}$ and $0 < p(\mathbf{x}) < 1$ (binary independent treatments), $\mathrm{r}(\mathbf{x})!\, n^{-\mathrm{r}(\mathbf{x})} / (t_1! t_2! \cdots t_n!)$ for $t_i \in \{0, 1, \ldots, \mathrm{r}(\mathbf{x})\}$ and $\sum_{i=1}^n t_i = \mathrm{r}(\mathbf{x}) \in \mathbb{N}$ (non-independent treatments with number of treatment levels $\mathrm{r}(\mathbf{x})$ depending on $\mathbf{x}$), and $e^{-n\lambda(\mathbf{x})} \lambda(\mathbf{x})^{\sum_{i=1}^n t_i} / (t_1! t_2! \cdots t_n!)$ for $t_i \in \mathbb{N}_0$ and $\lambda(\mathbf{x}) > 0$ (independent Poisson treatments with covariate-dependent parameter). However, the particular form of $f_I^{(\mathbf{x})}$ is irrelevant for our purposes: it is its symmetric character that matters and must be justified in applications.

The following examples exhibit responses and treatment assignments with various probabilistic structures, all of which are compatible with **A1** (though, as just pointed out, they need not comply with the basic model).

Example 3.1. Suppose that the T_is are binary, $T_i = 1$ indicating that the i-th unit has been 'treated', $T_i = 0$ that it has been kept as a 'control'. Let the covariate vectors have a finite number of possible values, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K \in \mathbb{R}^d$, and regard the $(\mathbf{X}_i, T_i, R_i)$ pertaining to units with the same value of $\mathbf{X}_i$ as data on the same patient obtained at different times. Thus if, for example, $\mathbf{X}_1 = \mathbf{X}_2 = \cdots = \mathbf{X}_8 = \mathbf{x}$ and $\mathbf{X}_i \neq \mathbf{x}$ for $i > 8$ then $R_1, R_2, \ldots, R_8$ represent all the responses of the same patient measured at eight different times and $T_1, T_2, \ldots, T_8$ the treatments assigned to that patient at those times. Clearly, neither $R_1, R_2, \ldots, R_8$ nor $T_1, T_2, \ldots, T_8$ need to be independent, since the responses are measured on the same patient and the patient's treatment assignment may for instance be constrained to include treatment exactly four times. If **A1** holds then $T_1, T_2, \ldots, T_8$ are exchangeable—so the assignment $(1, 0, 0, 1, 0, 1, 0, 1)$ is as likely as $(1, 1, 1, 0, 0, 0, 1, 0)$, $(1, 0, 0, 0, 0, 1, 0, 0)$ is as likely as $(0, 0, 1, 1, 0, 0, 0, 0)$, and so on—and the treatment assignments for different individuals are made independently.

This example may be specialized to that of Fisher's 'lady tasting tea' experiment (section 2.2 of [26]), possibly involving several ladies participating in independent tasting experiments; see its continuation in examples 3.7 and 3.9. □

Example 3.2. Suppose that the T_is are binary and that the vectors $(\mathbf{X}_i, T_i, R_i)$ pertaining to units with the same value of $\mathbf{X}_i$ represent data on different patients sampled from medical records and sharing (for instance) the same age, sex, and education and income levels. Let there be K possible values, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K \in \mathbb{R}^d$, for the covariate vectors. Assumption **A1** says that the treatment assignments among patients with covariate vectors equal to $\mathbf{x}_k$, say, are exchangeable. Suppose that such patients form a simple random sample from a subpopulation characterized by the same $\mathbf{x}_k$; then the treatment assignments are exchangeable indeed; and if the subpopulation is *very large* then the treatment assignments are also *approximately* independent. □

Example 3.3. Take the situation in example 3.2 with the difference that the patients are sampled in such a way that, of the n_k patients with covariate vectors equal to $\mathbf{x}_k$, m_k are treated and $n_k - m_k$ are kept as controls $(k = 1, 2, \ldots, K)$. Then the treatment assignments are dependent. □

Example 3.4. In the situation of example 3.3 assume that the responses of the n_k patients with covariate vectors equal to $\mathbf{x}_k$ consist of measurements (e.g. blood pressure readings) made on those patients. Then it is probably all right to regard the responses as independent random variables. □

Example 3.5. In the situation of example 3.3 assume that the responses of the n_k patients with covariate vectors equal to $\mathbf{x}_k$ do not represent measurements made on those patients but rather the *ranks* of such measurements. Then the responses are dependent random variables. □

The model used later in our illustrations follows (2.1) with $\tau_i = \tau$, $\varrho_i = \varrho$ for all i and certain τ and ϱ and the U_is and the V_is independent, so it satisfies **A0** and **A1**. Certain observational studies, however, such as *case-referent studies* (example 3.6 below) require a slight variant of **A1**:

A1′ For every $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$, conditionally on the event that $\mathbf{X}_{i_1}, \mathbf{X}_{i_2}, \ldots, \mathbf{X}_{i_n}$ are all equal to a given arbitrary $\mathbf{x} \in \mathbb{R}^d$ and $R_{i_1} = R_{i_2} = \cdots = R_{i_n} = r \in \mathbb{R}$, the random variables $T_{i_1}, T_{i_2}, \ldots, T_{i_n}$ are exchangeable in the sense that for $t_{i_1}, t_{i_2}, \ldots, t_{i_n} \in \mathbb{N}_0$

$$P(T_i = t_i, i \in I | \mathbf{X}_i = \mathbf{x}, R_i = r, i \in I)$$

is invariant to permutations of $(t_{i_1}, t_{i_2}, \ldots, t_{i_n})$.

Example 3.6. In a case-referent (or case-control) study a sample of *cases*—individuals with a certain, typically rare, disease—is matched with a sample of individuals drawn from a more general 'population' containing *referents* or non-cases, and possibly cases as well, and the two samples are compared with respect to the frequency of a treatment—typically exposure or non-exposure to a toxic substance (see, for example, pp. 7, 83–86 of [26], or pp. 73–93 of [31]). The purpose of the study is to investigate whether the exposure contributes to the disease, and the matching of the two samples, which consists of arranging cases and referents in groups sharing similar personal characteristics (age, sex, occupation, etc.), is intended to make cases and referents comparable with respect to everything except disease status. A particularity of a case-referent study is that the sampling of cases is drawn first and the sampling of referents is, to some extent, subordinate to the sample of cases that was drawn; this 'sampling plan' is often the only practical one if, due to the rarity of the disease, a simple random sample of the general population would have to be very large in order to guarantee the drawing of a substantial number of cases.

To specify a model for a case-referent study assume for simplicity that all variables are discrete and let $p_{\mathbf{X},T|R=r}$, $p_{T,R|\mathbf{X}=\mathbf{x}}$, etc., denote conditional probability functions of the vector $(\mathbf{X}, T, R)$ pertaining to a unit drawn randomly from the conceptually infinite general population, and let the event $R = 1$ mean that the unit is a case and $R = 0$ that it is a non-case. Let the first unit be a case: $(\mathbf{X}_1, T_1, R_1) \equiv (\mathbf{X}_1, T_1, 1)$, where $(\mathbf{X}_1, T_1)$ is drawn according to $p_{\mathbf{X},T|R=1}$. To find a match $(\mathbf{X}_2, T_2, R_2) \equiv (\mathbf{X}_1, T_2, R_2)$ for the first unit we draw R_2 according to $p_{R|\mathbf{X}=\mathbf{X}_1}$ and then T_2 according to $p_{T|\mathbf{X}=\mathbf{X}_1, R=R_2}$, each draw being independent of the preceding draws. Since the event $R_1 = 1$ is certain, we have

$$\begin{aligned} f^{(\mathbf{x})}_{\{1,2\}}(t_1, t_2) &\equiv P(T_1 = t_1, T_2 = t_2 | \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}) \\ &= P(T_1 = t_1, T_2 = t_2 | \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}, R_1 = R_2 = 1) p_{R|\mathbf{X}=\mathbf{x}}(1) + \\ &= P(T_1 = t_1, T_2 = t_2 | \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}, R_1 = 1, R_2 = 0) p_{R|\mathbf{X}=\mathbf{x}}(0). \end{aligned}$$

Conditionally on $\{\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}, R_1 = R_2 = 1\}$, the variables T_1 and T_2 have the same distribution and are independent, hence $P(T_1 = t_1, T_2 = t_2 | \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}, R_1 = R_2 = 1)$ is symmetric in t_1 and t_2; in contrast, t_1 and t_2 are not

interchangeable in $P(T_1 = t_1, T_2 = t_2 | \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}, R_1 = 1, R_2 = 0)$ if there is a treatment effect, since in that case the distribution of T_2 conditional on $R_2 = 0$ differs from that of T_1 conditional on $R_1 = 1$. Thus, **A1** does not hold in general because $f^{(\mathbf{x})}_{\{1,2\}}(t_1, t_2)$ need not be symmetric in its arguments. On the other hand, **A1′** holds with $I = \{1, 2\}$ since as noted above

$$P(T_1 = t_1, T_2 = t_2 | \mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}, R_1 = R_2 = r)$$

is symmetric in t_1, t_2 for $r = 1$, and for $r = 0$ this conditional probability can be defined arbitrarily.

To formulate the sampling of the first two units explicitly in terms of the equations in (2.1) write the conditional distribution functions of $(\mathbf{X}, T, R)$ as $P_{\mathbf{X},T|R=r}$, $P_{T,R|\mathbf{X}=\mathbf{x}}$, etc., and let U_1, U_2, V_2 and W_1 be independent standard uniform random variables. First, $(\mathbf{X}_1, T_1, R_1)$ can be generated by $R_1 = 1$ and $\mathbf{X}_1 = P^{-1}_{\mathbf{X}|R=1}(W_1)$ followed by $T_1 = P^{-1}_{T|\mathbf{X}=\mathbf{X}_1,R=1}(U_1)$ (as usual, the superscript '-1' indicates the inverse of a non-decreasing function). Conditionally on this first draw, $(\mathbf{X}_2, T_2, R_2) \equiv (\mathbf{X}_1, T_2, R_2)$ can be generated by $T_2 = P^{-1}_{T|\mathbf{X}=\mathbf{X}_1}(U_2)$ followed by $R_2 = P^{-1}_{R|\mathbf{X}=\mathbf{X}_1,T=T_2}(V_2)$. Clearly, $P^{-1}_{T|\mathbf{X}=\mathbf{X}_1,R=1}$ and $P^{-1}_{T|\mathbf{X}=\mathbf{X}_1}$ need not be equal if T and R are not independent, so T_1 and T_2 need not be exchangeable conditionally on $\mathbf{X}_1 = \mathbf{X}_2 = \mathbf{x}$.

The drawing of $(\mathbf{X}_i, T_i, R_i)$ for $i > 2$ is an iteration of the procedure for generating $(\mathbf{X}_1, T_1, R_1)$ and $(\mathbf{X}_2, T_2, R_2)$, in any of its two versions now described. Thus, if the first case is to be matched to k referents one draws $(\mathbf{X}_3, T_3, R_3), \ldots,$ $(\mathbf{X}_k, T_k, R_k)$ by repeating the procedure used to generate $(\mathbf{X}_2, T_2, R_2)$; then the second case, $(\mathbf{X}_{k+1}, T_{k+1}, R_{k+1}) \equiv (\mathbf{X}_{k+1}, T_{k+1}, 1)$, is generated by repeating the procedure used to generate $(\mathbf{X}_1, T_1, R_1)$; and so on.

An essential point about the method of sampling in a case-referent study is that the referents, even if they do not include cases, must be sampled as if they *could* include cases; otherwise, 'selection bias' (pp. 85–86 of [26], pp. 96–101 of [31]) may occur and alter the form of probabilities such as $f^{(\mathbf{x})}_{\{1,2\}}$. □

Under the null hypothesis of no treatment effect (and under **A0**, as always in the present section), **A1** and **A1′** are equivalent since the probabilistic identity in the latter reduces to that in the former. Moreover, under both **A1** and **A1′** to say that there is no treatment effect is to say that conditioning *further* on the values of the responses does not change the distribution of the treatments:

$$P(T_i = t_i, i \in I | \mathbf{X}_i = \mathbf{x}, R_i = r_i, i \in I) = f^{(\mathbf{x})}_I((t_i)_{i \in I}) \tag{3.2}$$

for $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$, $(t_i)_{i \in I} \in \mathbb{N}_0^n$, and $(r_i)_{i \in I} \in \mathbb{R}^n$. Conversely, to say that there is a treatment effect is to say that the probabilities on the left here depend on the r_is for some choice of the t_is. In other words, the treatment has an effect if and only if the knowledge of the responses (in addition to the knowledge that all covariate vectors are identical to some $\mathbf{x}$) provides information on the treatments.

Because of the equivalence of **A1** and **A1′** under the null hypothesis, the methods of testing to be described in subsections 3.1 and 3.2 are the same under

the two assumptions—they are based solely on (3.2) and on the symmetrical character of $f_I^{(\mathbf{x})}$.

Before going on to present the testing procedures, we must show that there is a strengthening of **A0** under which **A1** holds with the propensity scores in place of the $\mathbf{X}_i$s and the methods of subsections 3.1 and 3.2 are applicable with the propensity score as a covariate (hence should be approximately valid with the *estimated* propensity score as a covariate):

A0′ For every $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$, conditionally on the event $\{\mathbf{X}_i = \mathbf{x}_i, i \in I\}$, where $\mathbf{x}_{i_1}, \mathbf{x}_{i_2}, \ldots, \mathbf{x}_{i_n} \in \mathbb{R}^d$ are arbitrary, the random vectors $(U_{i_1}, U_{i_2}, \ldots, U_{i_n})$ and $(V_{i_1}, V_{i_2}, \ldots, V_{i_n})$ are independent. Moreover, if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K \in \mathbb{R}^d$ are distinct and $I_1, I_2, \ldots, I_K$ form a partition of $I = \{i_1, i_2, \ldots, i_n\}$ then for all $(u_i)_{i\in I}, (v_i)_{i\in I} \in \mathbb{R}^n$,

$$P(U_i \le u_i, V_i \le v_i, i \in I | \mathbf{X}_i = \mathbf{x}_j, i \in I_j, j = 1, 2, \ldots, K) =$$

$$\prod_{j=1}^{K} P(U_i \le u_i, V_i \le v_i, i \in I_j | \mathbf{X}_i = \mathbf{x}_j, i \in I_j).$$

Finally, $(\mathbf{X}_{i_1}, T_{i_1}), (\mathbf{X}_{i_2}, T_{i_2}), \ldots, (\mathbf{X}_{i_n}, T_{i_n})$ are independent and identically distributed.

To show that this assumption (which differs from **A0** only in its third and last proviso) yields the desired results we have to show that if the $\mathbf{X}_i$s are replaced by the propensity scores then (i) **A1** holds, (ii) the null hypothesis to be tested is the same when conditioning on the propensity scores as when conditioning on the $\mathbf{X}_i$s, and (iii) the second part of **A0** holds (so that inferences from different strata/matched sets continue to be independent if based on propensity scores). For simplicity we consider the case where $\mathbf{X}_i$ is discrete, when the propensity score is also discrete.

First, for arbitrary $I = \{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$ the set of random vectors $\{(\mathbf{X}_i, T_i) : i \in I\}$ is exchangeable (its elements being independent and identically distributed), and in particular the propensity score $f^{(\mathbf{x})}(t) \equiv f_i^{(\mathbf{x})}(t) = P(T_i = t | \mathbf{X}_i = \mathbf{x})$ is independent of i. Then, given a permutation π on I, we have, on writing f for a generic probability function on $\mathbb{N}_0$ and using $f^{(\mathbf{X}_i)} = f$ as an abbreviation of the event that $f^{(\mathbf{X}_i)}(t) = f(t)$ for all t such that $f^{(\mathbf{X}_i)}(t) > 0$,

$$P\left(T_i = t_{\pi(i)}, i \in I \middle| f^{(\mathbf{X}_i)} = f, i \in I\right) = \frac{P\left(T_i = t_{\pi(i)}, f^{(\mathbf{X}_i)} = f, i \in I\right)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\frac{P\left(T_{\pi^{-1}(j)} = t_j, f^{(\mathbf{X}_{\pi^{-1}(j)})} = f, j \in I\right)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} = \frac{P\left(T_j = t_j, f^{(\mathbf{X}_j)} = f, j \in I\right)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$P\left(T_i = t_i, i \in I \middle| f^{(\mathbf{X}_i)} = f, i \in I\right)$$

for $(t_i)_{i\in I} \in \mathbb{N}_0^n$. Thus $P(T_i = t_i, i \in I | f^{(\mathbf{X}_i)} = f, i \in I)$ is a symmetric function of the t_is and **A1** is satisfied with the propensity score $f^{(\mathbf{X}_i)}$ in place of $\mathbf{X}_i$.

Secondly, abbreviating summation over $\{\mathbf{x}_i, i \in I\}$ such that $f^{(\mathbf{x}_i)} = f$ for $i \in I$ as $\sum_f$, using the conditional independence of the U_i and V_i (first part of **A0′**) and the identity

$$P(T_i = t_i, i \in I | \mathbf{X}_i = \mathbf{x}_i, i \in I) = \prod_{i \in I} P(T_i = t_i | \mathbf{X}_i = \mathbf{x}_i) = \prod_{i \in I} f^{(\mathbf{x}_i)}(t_i)$$

(which follows from the last part of **A0′**), and finally noting that under the summation sign we have $f^{(\mathbf{x}_i)}(t_i) = f(t_i)$, we see that for $(r_i)_{i \in I} \in \mathbb{R}^n$

$$P\Big(T_i = t_i, R_i \le r_i, i \in I \Big| f^{(\mathbf{X}_i)} = f, i \in I\Big) = \frac{P\big(T_i = t_i, R_i \le r_i, f^{(\mathbf{X}_i)} = f, i \in I\big)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\frac{\sum_f P(\tau_i(U_i, \mathbf{x}_i) = t_i, \varrho_i(V_i, \mathbf{x}_i, t_i) \le r_i, \mathbf{X}_i = \mathbf{x}_i, i \in I)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\frac{\sum_f P(\tau_i(U_i, \mathbf{x}_i) = t_i, \varrho_i(V_i, \mathbf{x}_i, t_i) \le r_i, i \in I | \mathbf{X}_i = \mathbf{x}_i, i \in I) P(\mathbf{X}_i = \mathbf{x}_i, i \in I)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\frac{\sum_f P(\varrho_i(V_i, \mathbf{x}_i, t_i) \le r_i, i \in I | \mathbf{X}_i = \mathbf{x}_i, i \in I) \prod_{j \in I} f^{(\mathbf{x}_j)}(t_j) P(\mathbf{X}_i = \mathbf{x}_i, i \in I)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\prod_{j \in I} f(t_j) \frac{\sum_f P(\varrho_i(V_i, \mathbf{x}_i, t_i) \le r_i, i \in I | \mathbf{X}_i = \mathbf{x}_i, i \in I) P(\mathbf{X}_i = \mathbf{x}_i, i \in I)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\prod_{j \in I} f(t_j) \frac{\sum_f P(\varrho_i(V_i, \mathbf{X}_i, t_i) \le r_i, \mathbf{X}_i = \mathbf{x}_i, i \in I)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\prod_{j \in I} f(t_j) \frac{P\left(\varrho_i(V_i, \mathbf{X}_i, t_i) \le r_i, f^{(\mathbf{X}_i)} = f, i \in I\right)}{P\left(f^{(\mathbf{X}_i)} = f, i \in I\right)} =$$

$$\prod_{j \in I} f(t_j) P\left(\varrho_i(V_i, \mathbf{X}_i, t_i) \le r_i, i \in I \Big| f^{(\mathbf{X}_i)} = f, i \in I\right).$$

Thus (let $r_i \to \infty$ in the first and last terms)

$$P\left(T_i = t_i, R_i \le r_i, i \in I \Big| f^{(\mathbf{X}_i)} = f, i \in I\right) =$$

$$P\Big(T_i = t_i, i \in I \Big| f^{(\mathbf{X}_i)} = f, i \in I\Big)\, P\Big(\varrho_i(V_i, \mathbf{X}_i, t_i) \le r_i, i \in I \Big| f^{(\mathbf{X}_i)} = f, i \in I\Big),$$

and it follows (recall the argument after (3.1)) that the null hypothesis of no treatment effect as formulated at the beginning of this section (after the introduction of **A0**) holds if and only if it holds with $\left\{f^{(\mathbf{X}_i)} = f, i \in I\right\}$ in place of $\{\mathbf{X}_i = \mathbf{x}_i, i \in I\}$.

Finally, let I_1, I_2 be disjoint subsets of indices and put $I = I_1 \cup I_2$; in obvious notation, we have, on using the second and third parts of **A0′**,

$$P\Big(U_i \le u_i, V_i \le v_i, f^{(\mathbf{X}_j)} = f_1, f^{(\mathbf{X}_k)} = f_2, i \in I, j \in I_1, k \in I_2\Big) =$$

$$\sum_{\mathbf{x}_1 : f^{(\mathbf{x}_1)} = f_1} \sum_{\mathbf{x}_2 : f^{(\mathbf{x}_2)} = f_2} P(U_i \le u_i, V_i \le v_i, \mathbf{X}_j = \mathbf{x}_1, \mathbf{X}_k = \mathbf{x}_2, i \in I, j \in I_1, k \in I_2) =$$

$$\sum_{f_1, f_2} P\big(U_i \le u_i, V_i \le v_i, i \in I \big| \mathbf{X}_i = \mathbf{x}_j, i \in I_j, j = 1, 2\big) P(\mathbf{X}_i = \mathbf{x}_j, i \in I_j, j = 1, 2) =$$

$$\sum_{f_1, f_2} \prod_{j=1}^{2} P\big(U_i \le u_i, V_i \le v_i, i \in I_j \big| \mathbf{X}_i = \mathbf{x}_j, i \in I_j\big) P(\mathbf{X}_i = \mathbf{x}_i, i \in I_j) =$$

$$\sum_{\mathbf{x}_1 : f^{(\mathbf{x}_1)} = f_1} \sum_{\mathbf{x}_2 : f^{(\mathbf{x}_2)} = f_2} \prod_{j=1}^{2} P(U_i \le u_i, V_i \le v_i, \mathbf{X}_i = \mathbf{x}_j, i \in I_j) =$$

$$\prod_{j=1}^{2} \left\{ \sum_{\mathbf{x}_j : f^{(\mathbf{x}_j)} = f_j} P(U_i \le u_i, V_i \le v_i, \mathbf{X}_i = \mathbf{x}_j, i \in I_j) \right\} =$$

$$\prod_{j=1}^{2} P\Big(U_i \le u_i, V_i \le v_i, f^{(\mathbf{X}_i)} = f_j, i \in I_j\Big).$$

Dividing by $P\big(f^{(\mathbf{X}_j)} = f_1, f^{(\mathbf{X}_k)} = f_2, j \in I_1, k \in I_2\big)$ and noting that (by the last part of **A0′**) this probability equals $P\big(f^{(\mathbf{X}_i)} = f_1, i \in I_1\big) P\big(f^{(\mathbf{X}_i)} = f_2, i \in I_2\big)$, we conclude (the proof for K sets $I_1, I_2, \ldots, I_K$ being completely analogous) that the second part of **A0** also holds with the propensity scores as covariates, and hence inferences from different strata continue to be independent when based on propensity scores.

Remarks. (i) Regarding the last part of **A0′**, it can be seen that if $\tau_i = \tau$ for all i in the basic model then the exchangeability of the $(\mathbf{X}_i, U_i)$s implies that of the $(\mathbf{X}_i, T_i)$s and similarly the independence of the $(\mathbf{X}_i, U_i)$s implies that of the $(\mathbf{X}_i, T_i)$s.

(ii) The first part of **A0** (**A0′**) could perhaps be described as 'independence within strata', and the second part as 'independence between strata'; but in the first part it is the U_i and V_i of each unit i in the same stratum/matched set that are independent and in the second it is the (U_i, V_i)s from different strata/matched sets. The assumption could be more concisely formulated in somewhat more abstract terms: If $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K \in \mathbb{R}^d$ and $I_1, I_2, \ldots, I_K$ form a partition of $I = \{i_1, i_2, \ldots, i_n\}$ then for all $(u_i)_{i \in I}, (v_i)_{i \in I} \in \mathbb{R}^n$,

$$P(U_i \le u_i, V_i \le v_i, i \in I | \mathbf{X}_i = \mathbf{x}_j, i \in I_j, j = 1, 2, \ldots, K) =$$

$$\prod_{j=1}^{K} P(U_i \le u_i, i \in I_j | \mathbf{X}_i = \mathbf{x}_j, i \in I_j) P(V_i \le v_i, i \in I_j | \mathbf{X}_i = \mathbf{x}_j, i \in I_j).$$

If $K = 1$ this reduces to the first part of **A0**.

(iii) The method of assessing strata and matched sets described in subsection 3.4 and illustrated in section 6—which is essentially a means of assessing whether **A1** and **A1′** are at least plausible and which is an integral part of any analysis of

observational data based on the type of methods presented here—can sometimes help detecting violations of **A1** and **A1′**. But since, as already mentioned, these assumptions may hold even if the basic model does not, their plausibility in the face of a data set need not provide assurance that all the important confounders are being taken into account: the justification of the basic model and of **A0** (**A0′**) will nearly always require extra-statistical justification. Chapters 6 to 8 of [26] describe some general strategies for detecting violations of **A1** and **A1′**—and in some cases violations of the basic model and of **A0** (**A0′**)—on the basis of several observational data sets. □

3.1. Testing per stratum

A1 and **A1′** have both one consequence that can be used to test the null hypothesis of no treatment effect in a way that does not involve the probabilities (3.1), which in applications are unknown. For fixed $I = \{i_1, i_2, \ldots, i_n\}$ let $C_I^{(t)}$ count the number of T_is with $i \in I$ assuming the value $t \in \mathbb{N}_0$:

$$C_I^{(t)} = \sum_{i \in I} \mathbb{1}_{\{T_i = t\}}, \tag{3.3}$$

where $\mathbb{1}_A$ is the indicator function of the event A ($\mathbb{1}_A = 1$ if A occurs, $\mathbb{1}_A = 0$ otherwise). For $c_0, c_1, c_2, \ldots \in \mathbb{N}_0$ such that $n = \sum_{t \in \mathbb{N}_0} c_t$ let $\mathcal{T} \equiv \mathcal{T}(c_0, c_1, c_2, \ldots)$ stand for the set of vectors $(t_i)_{i \in I} = (t_{i_1}, t_{i_2}, \ldots, t_{i_n})$ of treatment assignments such that

$$c_0 = \sum_{i \in I} \delta_{t_i,0}, \;\; c_1 = \sum_{i \in I} \delta_{t_i,1}, \;\; c_2 = \sum_{i \in I} \delta_{t_i,2}, \ldots$$

(as usual, $\delta_{x,y}$ denotes Kronecker's delta: $\delta_{x,x} = 1$ and $\delta_{x,y} = 0$ if $x \neq y$). Clearly, if $(t_i^*)_{i \in I} = (t_{i_1}^*, t_{i_2}^*, \ldots, t_{i_n}^*)$ is one particular element of $\mathcal{T}$ then any other element of $\mathcal{T}$ can be obtained by permuting the coordinates of $(t_{i_1}^*, t_{i_2}^*, \ldots, t_{i_n}^*)$. Thus, since by assumption the probabilities $f_I^{(\mathbf{x})}((t_i)_{i \in I})$ are invariant under permutations of their arguments, we have, on abbreviating summation over the indices in $(t_i)_{i \in I} \in \mathcal{T}$ by $\Sigma_{\mathcal{T}}$,

$$P\left(C_I^{(t)} = c_t, t \in \mathbb{N}_0 \middle| \mathbf{X}_i = \mathbf{x}, i \in I\right) = \sum_{\mathcal{T}} f_I^{(\mathbf{x})}((t_i)_{i \in I}) = \#\mathcal{T} \times f_I^{(\mathbf{x})}((t_i^*)_{i \in I}),$$

where $(t_i^*)_{i \in I}$ is one particular element of $\mathcal{T}$ and as usual $\#\mathcal{T}$ stands for the number of elements of $\mathcal{T}$. Now consider the ratio of

$$P\left(T_i = t_i^*, i \in I, C_I^{(t)} = c_t, t \in \mathbb{N}_0 \middle| \mathbf{X}_i = \mathbf{x}, i \in I\right) = f_I^{(\mathbf{x})}((t_i^*)_{i \in I})$$

to the preceding probability:

$$\frac{P\left(T_i = t_i^*, i \in I, C_I^{(t)} = c_t, t \in \mathbb{N}_0 \middle| \mathbf{X}_i = \mathbf{x}, i \in I\right)}{P\left(C_I^{(t)} = c_t, t \in \mathbb{N}_0 \middle| \mathbf{X}_i = \mathbf{x}, i \in I\right)} = \frac{1}{\#T}.$$

Since—as is easy to see if the covariates are discrete—the term on the left here is the probability that $T_i = t_i^* \ \forall\, i$ conditionally on the event that $\mathbf{X}_i = \mathbf{x} \ \forall\, i$ and $C_I^{(t)} = c_t \ \forall\, t$, we conclude that

$$P\left(T_i = t_i^*, i \in I \middle| \mathbf{X}_i = \mathbf{x}, i \in I, C_I^{(t)} = c_t, t \in \mathbb{N}_0\right) = \frac{1}{\#T}$$

for each $(t_1^*, t_2^*, \ldots, t_n^*) \in \mathcal{T}(c_0, c_1, c_2, \ldots)$. Since by definition of $\mathcal{T}$ the number of elements in it corresponds to the number of distinct *strings* of length $n = \sum_{t\in\mathbb{N}_0} c_t$ made up of c_0 elements labelled 0, c_1 elements labelled 1, c_2 elements labelled 2, etc., we also see that

$$\#\mathcal{T} = \frac{n!}{\prod_{t\in\mathbb{N}_0} c_t!}.$$

The same argument (one need only replace the event $\{\mathbf{X}_i = \mathbf{x}, i \in I\}$ by $\{\mathbf{X}_i = \mathbf{x}, R_i = r_i, i \in I\}$ in the above) shows that *under the null hypothesis* we have, for all $(t_1^*, t_2^*, \ldots, t_n^*) \in \mathcal{T}(c_0, c_1, c_2, \ldots)$ and $(r_i)_{i\in I} = (r_{i_1}, r_{i_2}, \ldots, r_{i_n}) \in \mathbb{R}^n$,

$$P\left(T_i = t_i^*, i \in I \middle| \mathbf{X}_i = \mathbf{x}, R_i = r_i, i \in I, C_I^{(t)} = c_t, t \in \mathbb{N}_0\right) = \frac{\prod_{t\in\mathbb{N}_0} c_t!}{n!}. \quad (3.4)$$

It is this result that allows us to develop tests of the null hypothesis by conditioning on the covariates and on the number of times that each level of treatment occurs and then confronting the actual treatment assignments with the *conditional null distribution*, or *null probabilities*, (3.4). Such tests are *conditional* in the sense that it is by conditioning *further* (on the $C_I^{(t)}$s of (3.3), not just on the covariates) that the null distribution of treatment assignments becomes fully specified (in contrast to $f_I^{(\mathbf{x})}$, which is an unknown function of $\mathbf{x}$ and I). Many tests are possible, of course, since there are many ways of looking for discrepancies between the observed treatment assignments and the treatment assignments expected under the null. One possibility is to consider a test based on the sample covariance between treatments and responses. This test is appealing because the covariance is a measure of association and the relationship between a treatment and a response often manifests itself in a straightforward manner, higher levels of treatment being generally accompanied by higher responses or else by lower responses. We shall now elaborate on this test.

Assume that the treatment can only manifest itself by a positive association with the response, increasing levels of treatment tending to cause increasing responses. Suppose that $\{1, 2, \ldots, N\}$, or a subset of it, can be partitioned into K disjoint subsets $I_1, I_2, \ldots, I_K$ of sizes $n_1, n_2, \ldots, n_K$ such that all units with labels in I_k have covariate vectors equal to the same $\mathbf{x}_k \in \mathbb{R}^d$. For simplicity, in what follows we shall refer to the K sets of indices, and to the corresponding sets of units, as *strata*, keeping in mind that they may also denote matched sets (groups obtained by matching). The sample covariance associated with the k-th stratum is

$$\mathcal{K}_k \equiv \mathcal{K}_k((T_i, R_i)_{i\in I_k}) := \frac{1}{n_k} \sum_{i\in I_k} R_i T_i - \bar{R}_k \bar{T}_k,$$

where $\bar{R}_k = n_k^{-1} \sum_{i \in I_k} R_i$ and $\bar{T}_k = n_k^{-1} \sum_{i \in I_k} T_i$. For fixed k, condition on the event that the covariates $(\mathbf{X}_i)_{i \in I_k}$ are all equal to $\mathbf{x}_k \in \mathbb{R}^d$, the responses $(R_i)_{i \in I_k}$ are equal to $(r_i)_{i \in I_k} = (r_{i_1}, r_{i_2}, \ldots, r_{i_{n_k}}) \in \mathbb{R}^{n_k}$, and the numbers of times that the different levels of treatment occur satisfy

$$c_t^{(k)} = C_{I_k}^{(t)} := \sum_{i \in I_k} \mathbb{1}_{\{T_i = t\}} \quad (t \in \mathbb{N}_0) \tag{3.5}$$

for some $c_t^{(k)}$s in $\mathbb{N}_0$ such that $\sum_{t \in \mathbb{N}_0} c_t^{(k)} = n_k$; as shown above, under the null hypothesis the (conditional) distribution of the vector of treatment assignments $(T_i)_{i \in I_k}$ is uniform on the set $\mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots)$ that consists of the $n_k! / \prod_{t \in \mathbb{N}_0} c_t^{(k)}!$ points $(t_i)_{i \in I_k} \in \mathbb{N}_0^{n_k}$ satisfying $c_t^{(k)} = \sum_{i \in I_k} \delta_{t_i, t}$ $\forall t$. In principle, these points $(t_i)_{i \in I_k}$ may be enumerated and the corresponding values of $\mathcal{K}_k((t_i, r_i)_{i \in I_k})$ computed, yielding the conditional null distribution of the sample covariance. Then, with $q_{1-\alpha}$ denoting the quantile of probability $1-\alpha$ of this distribution, one rejects the null hypothesis—more precisely the null hypothesis relative to the k-th stratum—at the significance level of α if and only if the sample covariance actually observed, $\mathcal{K}_k((T_i, R_i)_{i \in I_k})$, exceeds $q_{1-\alpha}$.

To show that the size of this test is indeed α, let us define $q_{1-\alpha}$ as the smallest number q satisfying

$$\frac{\#\{(t_i)_{i \in I_k} \in \mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots) : \mathcal{K}_k((t_i, r_i)_{i \in I_k}) \geq q\}}{\#\mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots)} \leq \alpha.$$

Under the null hypothesis and with $q = q_{1-\alpha}$, the left-hand side here equals

$$P\Big(\mathcal{K}_k((T_i, R_i)_{i \in I_k}) \geq q_{1-\alpha} \Big| \mathbf{X}_i = \mathbf{x}_k, R_i = r_i, i \in I_k, (T_i)_{i \in I_k} \in \mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots)\Big);$$

thus, under the null, this probability is $\leq \alpha$, and integrating out the r_is and the $c_t^{(k)}$s yields

$$P\left(\mathcal{K}_k((T_i, R_i)_{i \in I_k}) \geq q_{1-\alpha} \Big| \mathbf{X}_i = \mathbf{x}_k, i \in I_k\right) \leq \alpha.$$

The conditional test can be carried out in another, equivalent manner, namely by computing the p-value—call it $p_k((T_i, R_i)_{i \in I_k})$—of the observed sample covariance:

$$\frac{\#\{(t_i)_{i \in I_k} \in \mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots) : \mathcal{K}_k((t_i, r_i)_{i \in I_k}) \geq \mathcal{K}_k((T_i, R_i)_{i \in I_k})\}}{\#T(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots)};$$

if this is $\leq \alpha$ then the null hypothesis is rejected at the level of α.

For large n_k the complete enumeration of the elements of $\mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots)$ required for the computation of the quantile or of the p-value can be impracticable. Fortunately, simulation can always be used to compute the two quantities approximately, namely by drawing a treatment assignment $(t_i)_{i \in I_k}$ pseudo-randomly from $\mathcal{T}(c_0^{(k)}, c_1^{(k)}, c_2^{(k)}, \ldots)$ and computing the corresponding value of

$\mathcal{K}_k((t_i, r_i)_{i\in I_k})$ a large number of times to get an estimate of the null distribution of the test statistic $\mathcal{K}_k((T_i, R_i)_{i\in I_k})$, from which an estimate of $q_{1-\alpha}$ and an estimate of $p_k((T_i, R_i)_{i\in I_k})$ are readily calculated (as the sample quantile of the values of $\mathcal{K}_k((t_i, r_i)_{i\in I_k})$ thus generated and as the proportion of times that those values exceed the observed test statistic $\mathcal{K}_k((T_i, R_i)_{i\in I_k})$, respectively). In addition, there exist excellent approximations based on central limit theorems for rank statistics which make it possible to compute the p-values of 'randomization tests' like the present one easily and accurately (see proposition 2, p. 35, of [26], and more generally chapter 12 of [8]).

Finally, let us observe that the conditional test based on the sample covariance $\mathcal{K}_k((T_i, R_i)_{i\in I_k})$ is equivalent to the conditional test based on the inner product

$$\mathbf{R}^{(k)} \cdot \mathbf{T}^{(k)} := \sum_{i\in I_k} R_k T_k$$

between the vector $\mathbf{R}^{(k)} := (R_i)_{i\in I_k}$ of responses and the vector $\mathbf{T}^{(k)} := (T_i)_{i\in I_k}$ of treatments, which Rosenbaum ([26], p. 35) calls a *sum statistic*. This follows from the fact that conditionally on the R_is and on the number of times that each level of treatment occurs (viz. (3.5)) the term $\bar{R}_k\bar{T}_k$ in $\mathcal{K}_k((T_i, R_i)_{i\in I_k})$ is a constant, so the conditional distribution of $\mathcal{K}_k((T_i, R_i)_{i\in I_k})$ differs from that of $\mathbf{R}^{(k)} \cdot \mathbf{T}^{(k)}$ by a change of scale and location.

Remarks. (i) The method used here to derive conditional tests is a special case of a standard method (see, for example, pp. 145–7 of [36]) for deriving tests by conditioning on sufficient statistics (which in our case are the $C^{(t)}_{I_k}$s of (3.5)).

(ii) Lehmann ([17], sections 5.10–5.13) establishes the unbiasedness and optimality of this test under general assumptions in the case of treatments with two levels and independent R_is. Rosenbaum ([26], pp. 44, 54) proves the unbiasedness of tests based on sum statistics in the case of binary treatments and responses of the form $R_i = r_iT_i + r'_i(1 - T_i)$ and under assumptions about the *constants* r_i and r'_i. □

Example 3.7. Consider the situation of example 3.1 with $K = 1$, so that there is a single stratum characterized by covariate vectors equal to a given $\mathbf{x} \in \mathbb{R}^d$. Let the stratum have size $n = 8$ and let there be exactly $m = 4$ treated units. The set $\mathcal{T}$ of treatment assignments satisfying these restrictions consists of the

$$\#\mathcal{T} = \frac{8!}{4!4!} = 70$$

vectors $(t_1, t_2, \ldots, t_8)$ with binary coordinates such that $\sum_{i=1}^8 \delta_{t_i,0} = 4$ and $\sum_{i=1}^8 \delta_{t_i,1} = 4$, or equivalently such that $\sum_{i=1}^8 t_i = 4$. The conditional distribution of the treatment assignments under the null hypothesis is therefore

$$P\left(T_i = t_i, i = 1, \ldots, 8 \middle| \mathbf{X}_i = \mathbf{x}, R_i = r_i, i = 1, \ldots, 8, \sum\nolimits_{j=1}^8 t_j = 4\right) = \frac{1}{70}$$

for $(t_1, t_2, \ldots, t_8) \in \mathcal{T}$. The elements of $\mathcal{T}$ are listed in table 1, sorted by increasing values of a test statistic S to be introduced below.

In order to illustrate the workings of the test and its conditional character let us first consider two situations where the responses may take arbitrary integer values: In the first, the vector of responses that turns up is $\mathbf{r} = (1, 2, 3, 4, 5, 6, 7, 8)$, and in the second it is $\mathbf{r}' = (1, 1, 1, 2, 2, 3, 6, 7)$; in both situations we take the observed treatment assignment as $\mathbf{t} := (0, 0, 1, 0, 1, 0, 1, 1)$. The observed values of the inner product statistics in the two situations are therefore

$$s := \mathbf{r} \cdot \mathbf{t} = 3 + 5 + 7 + 8 = 23 \quad \text{and} \quad s' := \mathbf{r}' \cdot \mathbf{t} = 1 + 2 + 6 + 7 = 16.$$

The inner product statistics themselves are the random variables $S := \mathbf{r} \cdot \mathbf{T}$ and $S' := \mathbf{r}' \cdot \mathbf{T}$, which depend on the random vector $\mathbf{T} := (T_1, T_2, \ldots, T_8)$ of treatment assignments. The possible values of these random variables are also given in table 1 next to the 70 possible values of $\mathbf{T}$ (the elements of $\mathcal{T}$). The conditional probability functions of S and S' under the null hypothesis,

$$g_S(s) := P\left(S = s \middle| \mathbf{X}_i = \mathbf{x}, i = 1, \ldots, 8, \sum\nolimits_{j=1}^{8} T_j = 4\right), \quad s = 10, 11, \ldots, 26,$$

$$g_{S'}(s) := P\left(S' = s \middle| \mathbf{X}_i = \mathbf{x}, i = 1, \ldots, 8, \sum\nolimits_{j=1}^{8} T_j = 4\right), \quad s = 5, 6, \ldots, 18,$$

are represented in figure 1; they are readily computed from table 1 by aggregating and adding up the probabilities corresponding to the possible values of S and S'. It is seen that, although the size of the test is always α, the form of the conditional distribution can vary quite much with the responses.

From the bottom of the last two columns in the table we see that seven treatment assignments yield a value of S at least as large as $s = 23$ and 12 assignments yield a value of S' at least as large as $s' = 16$. Thus, under the null hypothesis the probability that S would yield a value at least as large as the one observed is $\frac{7}{70} = 0.1$ and the probability that S' would yield a value at least as large as the one observed is $\frac{12}{70} = 0.17$. These are the p-values of the conditional tests based on the inner product statistic in the two situations; for instance, in none of them is the null rejected at the level of 0.05.

As pointed out by Rosenbaum ([26], pp. 34–35), as n increases the size of $\mathcal{T}$ increases at such a high rate—for instance, if we increased n from 8 to 10 and m from 4 to 5 in this illustration then the size of table 1 would quadruple—that it becomes a problem to compute p-values by complete enumeration of its elements. The method of simulation that we have described as an alternative to enumeration is in this case equivalent to the pseudo-random generation of rows from table 1 and the selection of the corresponding values of s or s', which serve as random samples from the distribution of S or S'.

Let us now consider a situation where the random vector $\mathbf{R}$ of responses is binary and its coordinates are constrained by the condition $\sum_{i=1}^{8} R_i = 4$, so that the set of possible outcomes for $\mathbf{R}$ coincides with $\mathcal{T}$. This situation corresponds to Fisher's tea tasting experiment, alluded to at the end of example 3.1, in which the lady tries to distinguish the four cups of tea prepared by pouring milk

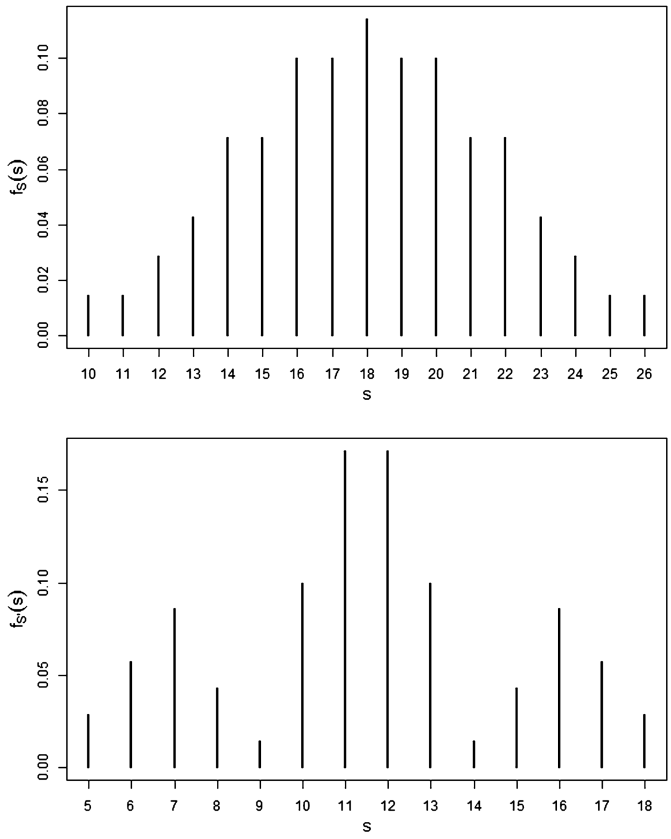


FIG 1. *Null conditional probability functions of the test statistics S and S'.*

first from the four cups prepared by pouring milk last; identifying the binary treatments with the two ways of serving the tea and the binary responses with the determination by the lady of the cups prepared according to her favourite recipe, the null hypothesis corresponds to the assumption that the lady has no discriminating powers whatsoever. Here, the mechanics of the conditional test based on the statistic $\mathbf{R}\cdot\mathbf{T}$ are exactly the same as before. What is worth noting in this case is that the test reduces to a Fisher exact test, namely to Fisher's test for a 2×2 contingency table with all marginal totals fixed at 4. This follows by arranging the pairs (R_i, T_i) into a contingency table and observing that the number of pairs $(1,1)$ in the table equals $\mathbf{R}\cdot\mathbf{T}$. Moreover, the conditional distribution of this statistic is multinomial, viz.

$$\binom{4}{s}\binom{4}{4-s}\Big/\binom{8}{4}$$

for $s = 0, 1, 2, 3, 4$. Indeed, the *randomization* picks 4 out of 8 cups to treat and under the null hypothesis the lady picks 4 of the 8 at random; the ex-

TABLE 1
Enumeration of the elements of $\mathcal{T}$ and of the corresponding values of the inner product statistics S and S'

t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	S	S'	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8	S	S'
1	1	1	1	0	0	0	0	10	5	1	0	1	0	0	1	0	1	18	12
1	1	1	0	1	0	0	0	11	5	0	1	1	0	1	0	0	1	18	11
1	1	1	0	0	1	0	0	12	6	0	1	1	0	0	1	1	0	18	11
1	1	0	1	1	0	0	0	12	6	1	0	0	1	1	0	0	1	18	12
1	1	0	1	0	1	0	0	13	7	1	0	0	0	1	1	1	0	19	12
1	1	1	0	0	0	1	0	13	9	0	0	1	1	1	0	1	0	19	11
1	0	1	1	1	0	0	0	13	6	1	0	1	0	0	0	1	1	19	15
0	1	1	1	1	0	0	0	14	6	0	1	1	0	0	1	0	1	19	12
1	1	1	0	0	0	0	1	14	10	1	0	0	1	0	1	0	1	19	13
1	1	0	1	0	0	1	0	14	10	0	1	0	1	1	0	0	1	19	12
1	0	1	1	0	1	0	0	14	7	0	1	0	1	0	1	1	0	19	12
1	1	0	0	1	1	0	0	14	7	0	0	1	1	0	1	1	0	20	12
1	1	0	0	1	0	1	0	15	10	0	0	1	1	1	0	0	1	20	12
1	0	1	0	1	1	0	0	15	7	0	1	0	0	1	1	1	0	20	12
0	1	1	1	0	1	0	0	15	7	1	0	0	0	1	1	0	1	20	13
1	0	1	1	0	0	1	0	15	10	0	1	1	0	0	0	1	1	20	15
1	1	0	1	0	0	0	1	15	11	1	0	0	1	0	0	1	1	20	**16**
1	1	0	0	1	0	0	1	16	11	0	1	0	1	0	1	0	1	20	13
1	0	1	0	1	0	1	0	16	10	0	1	0	1	0	0	1	1	21	**16**
1	1	0	0	0	1	1	0	16	11	0	1	0	0	1	1	0	1	21	13
0	1	1	0	1	1	0	0	16	7	0	0	1	1	0	1	0	1	21	13
0	1	1	1	0	0	1	0	16	10	0	0	1	0	1	1	1	0	21	12
1	0	0	1	1	1	0	0	16	8	1	0	0	0	1	0	1	1	21	**16**
1	0	1	1	0	0	0	1	16	11	0	1	0	0	1	0	1	1	22	**16**
0	1	0	1	1	1	0	0	17	8	0	0	1	0	1	1	0	1	22	13
1	1	0	0	0	1	0	1	17	12	0	0	0	1	1	1	1	0	22	13
1	0	1	0	0	1	1	0	17	11	1	0	0	0	0	1	1	1	22	**17**
0	1	1	1	0	0	0	1	17	11	0	0	1	1	0	0	1	1	22	**16**
1	0	1	0	1	0	0	1	17	11	0	0	1	0	1	0	1	1	**23**	**16**
0	1	1	0	1	0	1	0	17	10	0	1	0	0	0	1	1	1	**23**	**17**
1	0	0	1	1	0	1	0	17	11	0	0	0	1	1	1	0	1	**23**	14
1	1	0	0	0	0	1	1	18	15	0	0	1	0	0	1	1	1	**24**	**17**
0	1	0	1	1	0	1	0	18	11	0	0	0	1	1	0	1	1	**24**	**17**
1	0	0	1	0	1	1	0	18	12	0	0	0	1	0	1	1	1	**25**	**18**
0	0	1	1	1	1	0	0	18	8	0	0	0	0	1	1	1	1	**26**	**18**

pression above gives the probability that the lady picks exactly s cups of the 4 treated. □

Example 3.8. In any of the situations of examples 3.3–3.5 with $n_k = 8$ and $m_k = 4$ the workings of the conditional test on the k-th stratum are exactly as in example 3.7. However, for general n_k and m_k examples 3.4 and 3.5 exhibit interesting connections with the z-test and the Wilcoxon-Mann-Whitney test, respectively, provided an additional assumption is introduced, namely that the responses in the treated group are not only independent but also identically distributed and that the same is true of the responses in the control group. It is evident that under this assumption the null hypothesis of no treatment effect is equivalent to the equality of the distributions in the two groups.

Consider example 3.4: let $\mathbf{R}^{(k)}$ be the vector of responses and $\mathbf{T}^{(k)}$ that of the treatment assignments in stratum k, and set $S_1^{(k)} = \mathbf{R}^{(k)} \cdot \mathbf{T}^{(k)}$. Then $S_1^{(k)}$

is the sum of the responses among the treated patients (those with treatment equal to 1). Conditionally on the responses and on the value of the covariates, $S_1^{(k)}$ is equivalent—in the sense that it yields an equivalent conditional test—to

$$S_0^{(k)} := \mathbf{R}^{(k)} \cdot \mathbf{1} - \mathbf{R}^{(k)} \cdot \mathbf{T}^{(k)} = \mathbf{R}^{(k)} \cdot \left(\mathbf{1} - \mathbf{T}^{(k)}\right),$$

where $\mathbf{1}$ is the n_k-vector with coordinates equal to 1; this $S_0^{(k)}$ is of course the sum of the responses among the control patients (those with treatment equal to 0). Thus the test statistic

$$S^{(k)} := \frac{S_1^{(k)}}{m_k} - \frac{S_0^{(k)}}{n_k - m_k},$$

the difference between the average responses in the treated and control groups, yields a conditional test that is equivalent to the one provided both by $S_1^{(k)}$ and $S_0^{(k)}$. On the other hand, conditionally on the T_is and as a function of the R_is, $S^{(k)}$ is a difference between independent sample averages of independent and identically distributed variables (thanks to the additional assumption), and so, provided m_k and $n_k - m_k$ are not too small, it is approximately normally distributed and therefore serves as the basis for the usual z-test. In fact, it turns out (see pp. 174–175 of [39]) that the *exact* conditional test based on the inner product statistic is asymptotically equivalent to the z-test comparing the mean responses in the treated and control groups. [By 'exact' we mean that its type I error is smaller than or equal to the size of the test (e.g. 0.05 or 0.01); the type I error of an asymptotic test is *approximately* equal to the size of the test (hence could exceed it somewhat).]

In example 3.5 the response vectors $\mathbf{R}^{(k)}$ stand for the ranks of the n_k measurements made on the patients in stratum k. Consequently, $S_1^{(k)} = \mathbf{R}^{(k)} \cdot \mathbf{T}^{(k)}$ is the sum of the ranks in the treated group and, as is well known (e.g. pp. 405–8 of [24]), forms the basis of the Wilcoxon-Mann-Whitney test. Moreover, $S_1^{(k)}$ is approximately normally distributed, so the exact conditional test based on it is also asymptotically equivalent to a well-known 'unconditional' test. □

3.2. Testing for an overall effect

So far, our discussion has been centred on tests for a treatment effect within a fixed stratum—tests involving units that share the same value of the covariates. *Testing per stratum* is of interest when the particular value of the covariates determines interesting subpopulations (e.g. women in a certain age group) and the strata are not too small. However, in many applications the strata reflect a range of representative conditions, subpopulations, types of patients, etc., and it is of interest to test for an overall treatment effect, i.e. an effect averaged across that range. In other applications each stratum consists of units that depend on each other in some sense (e.g. they are siblings) and are sampled randomly from a population (e.g. of families with two or more children) and the interest lies

in the existence of a treatment effect in that population. In this subsection we consider the problem of *combining evidence* for a treatment effect across strata. A general approach to this problem is to develop a test based on a statistic that is a sum or weighted sum of statistics computed per stratum and whose distribution under the null hypothesis that no treatment effect exists in any of the strata is determined from the joint distribution of *all* treatment assignments. By the second part of **A0** (by **A0′** when stratifying on propensity scores) and **A1** (**A1′** in case-referent studies), the latter distribution is given by the product of probability functions of the form (3.4), one probability function per stratum.

Example 3.9. Consider the situation of example 3.1 in the guise of several tea tasting experiments like the one described at the end of example 3.7. Let the K different values of the covariate vectors represent K ladies and write $S^{(k)}$ for the inner product statistic computed with the data from the k-th lady; $S^{(k)}$ can be used to test whether the k-th lady possesses any discriminating power. Suppose first that each lady has come forward with the claim that she can discriminate between the two ways of preparing tea. A legitimate point of view would be to test the null hypothesis that *none of the ladies has discriminating powers*. In this case the rejection of the null would indicate that at least one of the ladies has discriminating powers, and if rejection occurred then we would probably want to identify the ladies more likely to be the discriminating ones—which could be done by performing the K conditional tests separately and picking the ladies with smaller p-values (taking $s = 4$ in the probabilities at the end of example 3.7 shows that the smaller p-value attainable in each test is $1/70$). To test the null, one could use the *Mantel-Haenszel statistic* (p. 31 of [26])

$$S := \sum_{k=1}^{K} S^{(k)}$$

(the inner product of the vector of treatments obtained by concatenating the vectors of treatments of the K ladies with the vector of responses obtained by concatenating their vectors of responses). Under the null hypothesis, the $S^{(k)}$s have the same distribution—the multinomial distribution given at the end of example 3.7—and are independent, so the null distribution of S can, in principle, be tabulated and the p-value of the test computed from it; alternatively the p-value can be estimated by simulation (generating independent sets of K observations from the multinomial distribution and adding them to get observations from the null distribution of S).

Whatever the result of this test (rejection or non-rejection of the null), any inference derived from it would have to refer to the particular group of ladies who made the claim—not, for instance, to an arbitrary person—and perhaps even more specifically to the group of ladies *at the time* that the test was carried out (for even a person's tasting buds can change with time).

Now suppose that the ladies have been sampled *randomly* from a certain population. A legitimate null hypothesis in this case would reflect the sampling mechanism, and would therefore state that a lady randomly sampled from the

population has no discriminating powers. The population being finite, this would be equivalent to the hypothesis that *no lady in the population has discriminating powers*. The test based on S, exactly with the same mechanics, would serve here as well. The inferences from the test, however, even if based on a small fraction of the population, would concern the whole population (not just the ladies who turned up in the sample). □

Example 3.10. Consider the two situations of example 3.8: In the first, $S^{(k)}$, the difference between the average responses in the treated group and the average response in the control group, is used to test the null hypothesis within stratum k; in the second, $S_1^{(k)}$, the sum of the ranks in the treated group, is used for the same purpose. In both cases, the data giving rise to the responses in stratum k consist of two random samples, one drawn from a subpopulation of treated patients and the other from a subpopulation of controls (in the first case the responses are the measurements and in the second they are the ranks of those measurements).

Suppose that the K strata represent disjoint and exhaustive classes of individuals of similar age, of the same sex, belonging to the same ethnic group, etc. Then a null hypothesis pertaining to all strata is that in every class of individuals the distribution of the treated individuals is equal to the distribution of the controls. To test this hypothesis we may consider

$$S := \sum_{k=1}^{K} S^{(k)} \quad \text{and} \quad S' := \sum_{k=1}^{K} S_1^{(k)},$$

respectively in the first and second situations. As in example 3.9, if the result of the test is found to be significant then the p-values of the tests per stratum may be examined in order to identify the classes of individuals for which the treatment effect is likely to be stronger.

The mechanics of the tests based on S and S' are completely analogous to those based on the Mantel-Haenszel statistic of example 3.9; thus, one can tabulate the distribution of each statistic and use it to compute the p-value, or else estimate the p-value by simulation. Under the additional assumption introduced in example 3.8 one can even compute the approximate null distributions of S and S', which are normal (at least if the m_k and $n_k - m_k$ are not too small).

Evidently, S and S' are not the only possible statistics that can be used to test the null hypothesis just defined. For example, S', often called the *stratified rank sum statistic*, may be replaced by a statistic proposed by Hodges and Lehmann which sometimes (namely when K is large compared to N; cf. p. 33 of [26]) is more powerful (better at detecting departures from the null) than S'. The Hodges-Lehmann statistic has exactly the same inner product form as S', but the responses on which it is based are defined differently: first, the original measurements within each stratum have their sample means subtracted from them, and then they are pooled into a single set and ranked; finally, these ranks are used as the responses. For the rest, the mechanics of the resulting test are again completely analogous to those of the Mantel-Haenszel statistic. □

Example 3.11. In the situation of example 3.3 let $m_k = n_k - m_k = 1$ for all k, let the responses be binary—1 standing for a positive or successful outcome—and interpret the different values of the covariate vectors, $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$, as labels of *pairs* of patients—siblings, identical twins, individuals matched on personal characteristics, etc. Thus, each stratum contains a pair of patients, only one of which is treated, and there are $N = 2K$ patients in total. By **A1**, within each stratum both treatment assignments have probability 1/2; under the null (no treatment effect in stratum k), conditionally on the responses of the two patients the two treatment assignments still have probability 1/2.

Evidently, it is not possible to test for a treatment effect per stratum. To test for a treatment effect on the set of pairs of patients, or on a population they may represent, we may consider the number of positive outcomes among treated patients,

$$S := \sum_{k:T_k=1} R_k = \sum_{k=1}^{K} T_k R_k,$$

or the difference between the proportion of positive outcomes among treated patients and the proportion of positive outcomes among controls,

$$D := \frac{1}{K} \sum_{k:T_k=1} R_k - \frac{1}{K} \sum_{k:T_k=0} R_k = \frac{1}{K} \sum_{k=1}^{K} T_k R_k - \frac{1}{K} \sum_{k=1}^{K} (1 - T_k) R_k.$$

Since $K \cdot D = 2 \sum_{k=1}^{K} T_k R_k - \sum_{k=1}^{K} R_k = 2 \cdot S - \sum_{k=1}^{K} R_k$, conditionally on the responses the distributions of S and D differ only by a change of scale and location and hence yield equivalent tests (cf. the argument on p. 133, before the remarks). In fact, both tests are equivalent to *McNemar's test*. To see this, recall that in the latter test the K pairs of responses corresponding to the paired patients are arranged in the table

		Response in control		
		0	1	Totals
Response in	0	$n_{1,1}$	$n_{1,2}$	$n_{1\cdot}$
treated patient	1	$n_{2,1}$	$n_{2,2}$	$n_{2\cdot}$
	Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	K

where $n_{1,1}$ stands for the number of pairs whose treated and control patients have both a negative outcome, $n_{1,2}$ for the number of pairs whose treated patient has a negative outcome and whose control has a positive outcome, etc., so that $n_{2\cdot}$ is the number of treated patients with a positive outcome and $n_{\cdot 2}$ the number of control patients with a positive outcome. McNemar's statistic (pp. 492–3 of [24]) is equivalent to $n_{1,2} - n_{2,1}$, and clearly $K \cdot D = n_{2\cdot} - n_{\cdot 2} = n_{2,1} - n_{1,2}$.

The test based on D can be carried out by tabulating the null distribution of the statistic or by estimating the p-value by simulation. Asymptotically (or for large $n_{i,j}$s), and provided *the pairs* of patients are randomly sampled from a large population, the test is also equivalent to the McNemar test as this is usually applied, appealing to a normal/chi-square approximation. □

Example 3.12. In a case-referent study (example 3.6) a matched set consists of individuals sharing the same covariate vector, at least one of whom has the disease. The data from such a study are sometimes arranged in a contingency table, even though this disregards the structure of the matched sets:

		Exposure		
		1	0	Totals
Disease	1	$n_{1,1}$	$n_{1,2}$	$n_{1\cdot}$
	0	$n_{2,1}$	$n_{2,2}$	$n_{2\cdot}$
	Totals	$n_{\cdot 1}$	$n_{\cdot 2}$	N

A natural test statistic for the null hypothesis of no treatment effect (again equivalent to $\sum_i T_i R_i$) is the difference between the proportions of individuals exposed in the diseased and non-diseased individuals,

$$D := \frac{\sum_{i=1}^N T_i R_i}{\sum_{i=1}^N R_i} - \frac{\sum_{i=1}^N T_i (1-R_i)}{\sum_{i=1}^N (1-R_i)} = \frac{n_{1,1}}{n_{1\cdot}} - \frac{n_{2,1}}{n_{2\cdot}}.$$

Its null distribution is obtained by computing the formula for D for all permutations of the T_is within each matched set, and an approximation to it is obtained by permuting the T_is pseudo-randomly within matched sets (independently in different matched sets) to get a pseudo random sample from the null distribution of D. The usual approximate test for comparing two probabilities cannot serve to approximate nor to replace the test based on D, since the T_is are permuted within matched sets and, due to the sampling plan of the study, the summands in $n_{2,1}/n_{2\cdot}$ need not be identically distributed. □

3.3. The problem of few, sparse strata

Following stratification, a stratum can be included in the testing procedure only if it contains at least two units with different values of treatment (otherwise the conditional null distribution of treatment assignments, given by (3.4), is degenerate at a single value of treatment). However, stratification usually yields several strata with the same level of treatment and which therefore have to be discarded. This is particularly likely to happen if the majority of units share the same treatment or if the strata are too small. Unfortunately, strata are bound to be small if there are too many covariates relative to the total number of units. For example, even in the most favourable situation where the d coordinates of the $\mathbf{X}_i$s are binary there are 2^d distinct possible values for $\mathbf{X}_i$ and as many potential strata; and if we are to get an average of two observations per stratum when the coordinates of $\mathbf{X}_i$ are independent then a sample size of $N = 2^{d+1}$ is required. Sometimes, due to only a few and small strata being available for testing, no test procedure can provide substantial evidence for a treatment effect because none of the probabilities (3.4) is small enough. To illustrate these difficulties it is enough to mention a quick example from subsection 3.4.2 of [26]: Fourteen patients were divided into nine strata on the basis of three covariates; six of the strata contained only patient, one contained only two treated patients, another

contained two treated patients and two control patients, and a last one contained one treated patient and one control. Clearly, only the last two strata can be used for testing and there are only $\binom{4}{2}\binom{2}{1} = 6\times 2 = 12$ possible treatment assignments, so the smallest probability attained by the distribution of treatment assignments (and hence the smallest p-value attainable) is $\frac{1}{12} \approx 0.083$.

By definition, matched sets contain at least two units with different values of treatment. However, matching suffers from essentially the same difficulties as stratification does: unless the dimension of the $\mathbf{X}_i$s is low, it is difficult to find good matches for all units.

When it works—when propensity scores can be accurately estimated—the Rosenbaum–Rubin approach mitigates these problems, and in some cases makes use of most of the data. One may anticipate that the Rosenbaum–Rubin approach will be particularly useful in situations where the treatment takes only a few of values, the sample size is large and the $\mathbf{X}_i$s have more than just a couple of components; if the sample size or the dimension of $\mathbf{X}_i$ are small then stratifying/matching on the $\mathbf{X}_i$s, if possible at all, may still be the better approach.

3.4. Checking 'balance': assessment of strata and matched sets

Because $\mathbf{X}_i$ typically includes some continuous random variables, stratification is often achieved by splitting the range of each covariate into intervals or finite subsets and taking their Cartesian product as the strata. Thus, the conditional probabilities often involve $\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}}$ for some neighbourhood $\mathcal{N}_{\mathbf{x}}$ of $\mathbf{x}$ rather than the $\mathbf{X}_i = \mathbf{x}$ appearing in the probabilities $f_I^{(\mathbf{x})}$. Fine partitions—partitions that result from splitting the range of each covariate into many intervals or finite subsets—conform better to assumptions **A1** and **A1′** and therefore remove more of the bias due to confounding, but they lead to fewer useable strata—strata containing at least two different levels of treatment—and smaller sample sizes per stratum, and hence to lower power in testing for a treatment effect and lower efficiency in estimating the treatment effect. Conversely, coarser partitions yield more and larger strata, and hence lead to tests with greater power of rejecting the null hypothesis and to estimates with smaller variance; however, the tests have a greater type I error and the estimates a greater bias than desired. Like in many statistical procedures there is thus a trade-off between bias and variance in the choice of the stratification, and this calls for means of assessing the quality of a stratification. Under **A1**, an intuitively obvious way of checking whether a given stratification is fine enough (and which evidently applies to matched sets as well) consists of comparing the joint distribution of the covariates across the different levels of treatment per stratum: if the levels of treatment appear to be *balanced*—i.e. if the joint distribution does not appear to vary with the treatment—then the stratification should be appropriate. Ideally, one should choose the coarser stratification among all sufficiently fine stratifications, but in practice the choice is not always straightforward.

In order to provide a justification for this method, suppose that **A1** holds with $f_I^{(\mathbf{x})}$ continuous in $\mathbf{x}$ (more precisely, in those coordinates of $\mathbf{x}$ that correspond

to continuous covariates in $\mathbf{X}_i$). Fix $\mathbf{x} \in \mathbb{R}^d$ and a neighbourhood $\mathcal{N}_\mathbf{x}$ of it, and write as usual $I = \{i_1, i_2, \ldots, i_n\}$ for a set of unit labels and $(t_i)_{i\in I} \in \mathbb{N}_0^n$. For each $\mathbf{x}' \in \mathcal{N}_\mathbf{x}$ and each neighbourhood $N_{\mathbf{x}'}$ of $\mathbf{x}'$ contained in $\mathcal{N}_\mathbf{x}$ we have

$$P(T_i = t_i, i \in I | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, i \in I) P(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, T_i = t_i, i \in I) =$$

$$\frac{P(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, T_i = t_i, i \in I)}{P(\mathbf{X}_i \in \mathcal{N}_\mathbf{x}, i \in I)} = \frac{P(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I)}{P(\mathbf{X}_i \in \mathcal{N}_\mathbf{x}, i \in I)} P(T_i = t_i, i \in I | \mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I) =$$

$$P(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, i \in I) P(T_i = t_i, i \in I | \mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I).$$

If $\mathcal{N}_\mathbf{x}$ is small enough,

$$\begin{aligned} P(T_i = t_i, i \in I | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, i \in I) &\approx f_I^{(\mathbf{x})}((t_i)_{i\in I}) \\ &\approx f_I^{(\mathbf{x}')}((t_i)_{i\in I}) \approx P(T_i = t_i, i \in I | \mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I), \end{aligned}$$

so dropping the left- and rightmost probabilities here from the left- and rightmost members of the identities above yields the approximate identity

$$P(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, T_i = t_i, i \in I) \approx P(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, i \in I)$$

for $\mathbf{x}' \in \mathcal{N}_\mathbf{x}$. In words, conditionally on the covariate vectors of the units in I being contained in a neighbourhood $\mathcal{N}_\mathbf{x}$ their joint distribution is about the same irrespectively of which units received which treatment.

To make this property operational one needs to assume that the pairs $(\mathbf{X}_i, T_i)$ with $\mathbf{X}_i \in \mathcal{N}_\mathbf{x}$ constitute a random sample, or at least that they are identically distributed, so that empirical distributions or sample means of the covariates within subgroups of units receiving different treatments have a meaning and hence can be compared. In terms of expectations, such an assumption implies that if $\mathcal{N}_\mathbf{x}$ is small then

$$E(\mathbf{X}_i | \mathbf{X}_i \in \mathcal{N}_\mathbf{x}, T_i = t) \tag{3.6}$$

is practically independent of t and of i. One way of checking the quality of a stratification is thus to compare sample versions of the expectation in (3.6) for different values of t. Another, more thorough way, which is especially convenient when the strata are large, is to test the equality of the distributions corresponding to different values of t. Illustrations of this method will be provided in section 6, although there, in order to save space, we shall only check the equality of *marginal* distributions.

This method applies in particular when the covariates consist of propensity scores, assuming **A0′** and in particular the same propensity score $f^{(\mathbf{x})} \equiv f_i^{(\mathbf{x})}$ for all units; however, there is another, perhaps more detailed way of assessing the quality of a given stratification/matching based on the propensity score and which follows from the following result of Rosenbaum and Rubin [28].

Writing f for a probability function on $\mathbb{N}_0$ and $\mathcal{S}_f = \{\mathbf{x} : f^{(\mathbf{x})} = f\}$, we have

$$P\left(T_i = t \middle| \mathbf{X}_i = \mathbf{x}, f^{(\mathbf{X}_i)} = f\right) = P(T_i = t | \mathbf{X}_i = \mathbf{x}) = f(t), \quad \mathbf{x} \in \mathcal{S}_f,$$

and, as seen earlier (see (2.3)), $P\left(T_i = t \middle| f^{(\mathbf{X}_i)} = f\right) = f(t)$. Assuming that the $\mathbf{X}_i$s are discrete and using these identities in

$$P\Big(\mathbf{X}_i = \mathbf{x}\Big|T_i = t, f^{(\mathbf{X}_i)} = f\Big) = P\Big(\mathbf{X}_i = \mathbf{x}\Big|f^{(\mathbf{X}_i)} = f\Big)\frac{P\big(T_i = t\big|\mathbf{X}_i = \mathbf{x}, f^{(\mathbf{X}_i)} = f\big)}{P\big(T_i = t\big|f^{(\mathbf{X}_i)} = f\big)}$$

gives

$$P\left(\mathbf{X}_i = \mathbf{x}\middle|T_i = t, f^{(\mathbf{X}_i)} = f\right) = P\left(\mathbf{X}_i = \mathbf{x}\middle|f^{(\mathbf{X}_i)} = f\right), \quad \mathbf{x} \in \mathcal{S}_f.$$

In other words, conditionally on the event $\left\{f^{(\mathbf{X}_i)} = f\right\}$ the joint distribution of the covariates of unit i is independent of that unit's treatment assignment. An analogous result is valid if the $\mathbf{X}_i$s are continuous and have densities.

The result is especially useful when there are only a couple of different treatments. In particular, if the T_is are binary, 1 indicating that the unit is treated and 0 that it is a control, it reads

$$\begin{aligned} P\Big(\mathbf{X}_i = \mathbf{x}\Big|T_i = 0, f^{(\mathbf{X}_i)}(1) = p\Big) &= P\Big(\mathbf{X}_i = \mathbf{x}\Big|f^{(\mathbf{X}_i)}(1) = p\Big) \\ &= P\Big(\mathbf{X}_i = \mathbf{x}\Big|T_i = 1, f^{(\mathbf{X}_i)}(1) = p\Big) \end{aligned}$$

for $\mathbf{x}$ such that $p = f^{(\mathbf{x})}(1) \in]0, 1[$, and is called the *balancing property* of the propensity score (pp. 297–9 of [26]). Thus, in this case, in order to assess the quality of a stratification or matching one checks whether the subgroup of treated units and the subgroup of control units within each stratum or matched set have approximately the same distribution of covariates.

Under **A1′**, and if the responses are discrete, the assessment of a stratification on the covariates is based on the approximate identity

$$\begin{aligned} P\left(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I\middle|\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}}, R_i = r, T_i = t_i, i \in I\right) \approx \\ P\left(\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}'}, i \in I|\mathbf{X}_i \in \mathcal{N}_{\mathbf{x}}, R_i = r, i \in I\right), \end{aligned}$$

according to which the distribution of the $\mathbf{X}_i$s conditionally on their falling in a neighbourhood $\mathcal{N}_{\mathbf{x}}$ and on the responses being equal to a given r is about the same irrespectively of which units received which treatment. If the responses are continuous one replaces $R_i = r$ by $R_i \in \mathcal{N}_r$, say, in the approximate identity.

4. Estimation of a treatment effect

It is often of interest to estimate the magnitude of a treatment effect rather than just testing for the existence of one. Naturally, this more ambitious task requires different, stronger assumptions than those used for testing. In this section we consider two assumptions that appear to be valid in many situations and in *some* situations—when a stratification (typically a stratification on the propensity score) succeeds in using practically the whole sample—lead to useful estimation

procedures. In subsection 4.1, under the same assumptions, we digress a little from stratification methods to consider an entirely different, perhaps more direct method of estimating the magnitude of a treatment effect, based on estimating 'counterfactuals' by a non-parametric estimator of the response conditionally on the covariates. Case-referent studies (examples 3.6 and 3.12) require a somewhat different approach to estimation and are treated in subsection 4.2.

In the first place, we need to be able to *quantify* a treatment effect. Treatment effect was defined at the beginning of section 2 in a very general way, namely as the dependence of the responses on the treatments given the covariates (dependence which could even vary with the units considered). Here we continue to assume the basic model and hence to adopt the same definition of treatment effect, but we want to quantify it. This can be done in many ways, depending on the type of response and on what is considered practically relevant; however, treatment effects are most conveniently quantified in terms of *parameters* of the joint distribution of a unit's response and treatment conditionally on the values of its covariates, and on averaged versions of those parameters. This leads us to assume, in addition to the basic model, that

A2 For any $\mathbf{x} \in \mathbb{R}^d$, conditionally on $\mathbf{X}_i = \mathbf{x}$ the vector of the response and treatment of an arbitrary unit i, (R_i, T_i), has a distribution function $G^{(\mathbf{x})}$ that may depend on $\mathbf{x}$ but not on i.

An immediate consequence of **A2** is that

$$f^{(\mathbf{x})}(t) \equiv f_i^{(\mathbf{x})}(t) = P(T_i = t | \mathbf{X}_i = \mathbf{x}) \tag{4.1}$$

is independent of i. Also, the distribution function of R_i conditional on the event $\{\mathbf{X}_i = \mathbf{x}, T_i = t\}$, denoted by $H^{(\mathbf{x},t)}$, satisfies

$$P(T_i = t | \mathbf{X}_i = \mathbf{x}) H^{(\mathbf{x},t)}(r) = G^{(\mathbf{x})}(r, t), \tag{4.2}$$

and therefore is also independent of i. (In contrast, whether the (R_i, T_i)s are identically distributed depends on whether the $\mathbf{X}_i$s are identically distributed.)

Under **A2** we can in principle quantify the treatment effect conditionally on the covariates: If for example $E(R_i)$ exists and the T_is are finite then the covariance of R_i and T_i conditional on $\mathbf{X}_i = \mathbf{x}$,

$$\kappa(\mathbf{x}) := E[R_i T_i | \mathbf{X}_i = \mathbf{x}] - E[R_i | \mathbf{X}_i = \mathbf{x}] E[T_i | \mathbf{X}_i = \mathbf{x}],$$

exists and may be used as a measure of the effect of treatment on the response conditionally on the value of $\mathbf{X}_i$. Under the null hypothesis of no treatment effect we have $\kappa(\mathbf{x}) = 0$, though of course $\kappa(\mathbf{x}) = 0$ does not imply that there is no treatment effect.

Apart from its relative simplicity—if we compare it to parameters expressing the variation of $H^{(\mathbf{x},t)}$ or of its mean with t for fixed $\mathbf{x}$—there are at least three reasons for recommending $\kappa(\mathbf{x})$: First, in most applications, if a treatment has an effect on a response then that effect usually consists of an average increase or decrease in the response, and hence can be detected by estimating $\kappa(\mathbf{x})$ or

averages of it for varying $\mathbf{x}$. Secondly, $\kappa(\mathbf{x})$ is the theoretical analogue of the sample covariance introduced in subsection 3.1 for the testing of a treatment effect, which was seen to be equivalent to the inner product or sum statistic. Finally, and paralleling the property described in example 3.8, if the treatment assumes only two values—say it is binary—then the covariance is essentially equivalent to the difference between the mean responses under one and the other treatment,

$$\delta(\mathbf{x}) := E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 1] - E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 0].$$

Indeed, it is easy to check that if T_i is binary then

$$\kappa(\mathbf{x}) = \delta(\mathbf{x}) \cdot f^{(\mathbf{x})}(0) \cdot f^{(\mathbf{x})}(1).$$

For simplicity, we refer to $\kappa(\mathbf{x})$ and $\delta(\mathbf{x})$ as *the* treatment effect at $\mathbf{x}$ (or conditionally on $\mathbf{X}_i = \mathbf{x}$); and when referring to the treatment effect at $\mathbf{x}$ without further specification we shall denote it by $\mathrm{e}(\mathbf{x})$.

Of course, there are other definitions of $\mathrm{e}(\mathbf{x})$ which are more convenient in other situations. For example, if the responses are binary, so that the expectations involved in the $\delta(\mathbf{x})$ above are conditional probabilities, a ratio or a logarithm of ratios rather than a difference may be preferable.

Now suppose that $E[\mathrm{e}(\mathbf{X}_i)]$ exists for each i; we define the *overall treatment effect* as

$$\boldsymbol{\epsilon} := E[\mathrm{e}(\mathbf{X})],$$

where the distribution of $\mathbf{X}$ is *an average* (to be specified) of the distributions of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$.

This is just one of the possible ways of averaging $\mathrm{e}(\mathbf{x})$ with respect to $\mathbf{x}$. It may be interpreted as the treatment effect on a unit drawn randomly (according to a uniform distribution, for example) from the N units represented in the observational data set, and it accommodates a few standard situations. For example, if the $\mathbf{X}_i$s have the same distribution function F then the expectation $E[\mathrm{e}(\mathbf{X})]$ may be taken with respect to F. If instead the $\mathbf{X}_i$s are 'stratified', in the sense that there exists a partition $I_1, I_2, \ldots, I_K$ of $\{1, 2, \ldots, N\}$ such that $\mathbf{X}_i$ has distribution function F_k whenever $i \in I_k$ $(k = 1, 2, \ldots, K)$, then the expectation may be taken with respect to $F := \sum_{k=1}^{K} F_k w_k$, where $w_k = \#I_k/N$. And if the w_ks in the latter average do not reflect the proportions of elements of each stratum in a certain population of interest then the expectation may be taken with respect to $F := \sum_{k=1}^{K} F_k w'_k$, where the w'_ks are the correct proportions.

Remark. The overall treatment effect has to be distinguished from any parameter that quantifies differences between mean responses at different levels of treatment; in fact, this distinction lies at the root of the problem of confounding and is what prompts the need for 'correcting for confounding' by methods based on stratification and matching. To elaborate a little on this let us consider the situation of a binary treatment.

If $T_i = 1$ indicates that unit i receives treatment and $T_i = 0$ that it is kept as a control, then the mean difference between the responses in the treated and

control groups,

$$\Delta := E[R_i|T_i = 1] - E[R_i|T_i = 0],$$

quantifies the difference between mean responses at the two levels of treatment. Since $E[R_i|T_i = t]$ is obtained by integrating $E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = t]$ with respect to the distribution of $\mathbf{X}_i$ conditional on $T_i = t$, this Δ is obtained by integrating $E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 1]$ and $E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 0]$ with respect to potentially *different* distributions and subtracting the results. In contrast, $\boldsymbol{\epsilon}$ is obtained by integrating $E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 1] - E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 0]$ with respect to *the same* distribution.

Despite the differences in the computation of Δ and $\boldsymbol{\epsilon}$ and the obvious correctness of the latter, one could be tempted to treat the observational study as an experimental study and to test for a treatment effect by estimating Δ (which indeed would be the correct thing to do in the latter type of study). This 'naive approach' would ignore the fact that $\mathbf{X}_i$ and T_i are dependent, for which reason $\Delta \neq \boldsymbol{\epsilon}$ in general. For example, it can be seen that if the vectors $(\mathbf{X}_i, T_i, R_i)$ are identically distributed, the distribution of $\mathbf{X}_i$ conditional on $T_i = t$ has a positive density $f_{\mathbf{X}_i|T_i=t}$ and the distribution of $(\mathbf{X}_i, R_i)$ conditional on $T_i = t$ has a density $f_{\mathbf{X}_i,R_i|T_i=t}$ then the bias of the naive approach is

$$\boldsymbol{\epsilon} - \Delta = \int\int r \left\{ \left(\frac{f_{\mathbf{X}_i}(\mathbf{x})}{f_{\mathbf{X}_i|T_i=1}(\mathbf{x})} - 1 \right) f_{\mathbf{X}_i,R|T_i=1}(\mathbf{x}, r) - \left(\frac{f_{\mathbf{X}_i}(\mathbf{x})}{f_{\mathbf{X}_i|T_i=0}(\mathbf{x})} - 1 \right) f_{\mathbf{X}_i,R|T_i=0}(\mathbf{x}, r) \right\} \mathrm{d}\mathbf{x}\, \mathrm{d}r,$$

where $f_{\mathbf{X}_i}$ is a density of $\mathbf{X}_i$; if $\mathbf{X}_i$ and T_i are independent then $f_{\mathbf{X}_i} = f_{\mathbf{X}_i|T_i=0} = f_{\mathbf{X}_i|T_i=1}$ and both terms inside the integral vanish, so $\boldsymbol{\epsilon} - \Delta = 0$; otherwise, only a miraculous cancellation will bring the bias to 0. □

Having defined parameters that quantify treatment effects, we need an assumption that allows us to estimate and find confidence intervals for them:

A3 For any $\{i_1, i_2, \ldots, i_n\} \subset \{1, 2, \ldots, N\}$, conditionally on

$$\mathbf{X}_{i_1} = \mathbf{x}_{i_1}, \mathbf{X}_{i_2} = \mathbf{x}_{i_2}, \ldots, \mathbf{X}_{i_n} = \mathbf{x}_{i_n}, \quad T_{i_1} = t_{i_1}, T_{i_2} = t_{i_2}, \ldots, T_{i_n} = t_{i_n},$$

the responses $R_{i_1}, R_{i_2}, \ldots, R_{i_n}$ are independent.

Except for the conditioning on the covariates, this corresponds to the assumption of 'no interference between units' (pp. 41–42 of [26]). Thanks to it, conditionally on their covariates and treatments the responses of a subset of units form a random sample, so that standard statistical methods based on the central limit theorem can in principle be used for estimating and testing hypotheses on e($\mathbf{x}$) and, by suitable modifications, on $\boldsymbol{\epsilon}$. As pointed out on p. 42 of [26], there are situations where **A3** may not hold: for example, if treatment consists of vaccination, non-vaccinated units in contact with vaccinated ones may have better responses than 'genuine' controls.

We now consider the problem of estimating and finding confidence intervals for $\mathrm{e}(\mathbf{x})$ and $\boldsymbol{\epsilon}$. The estimation of $\mathrm{e}(\mathbf{x})$ is straightforward: Estimate $\mathrm{e}(\mathbf{x})$ by the empirical covariance of the T_is and R_is,

$$\hat{\mathrm{e}}(\mathbf{x}) := \frac{\sum_{i:\mathbf{X}_i=\mathbf{x}} R_i T_i}{\#\{j : \mathbf{X}_j = \mathbf{x}\}} - \bar{R}_{\mathbf{x}} \bar{T}_{\mathbf{x}},$$

where

$$\bar{T}_{\mathbf{x}} := \frac{1}{\#\{j : \mathbf{X}_j = \mathbf{x}\}} \sum_{i:\mathbf{X}_i=\mathbf{x}} T_i \quad \text{and} \quad \bar{R}_{\mathbf{x}} := \frac{1}{\#\{j : \mathbf{X}_j = \mathbf{x}\}} \sum_{i:\mathbf{X}_i=\mathbf{x}} R_i,$$

or by its unbiased version (with $\#\{j : \mathbf{X}_j = \mathbf{x}\} - 1$ in place of $\#\{j : \mathbf{X}_j = \mathbf{x}\}$). If the treatments are binary one should instead use

$$\hat{\delta}(\mathbf{x}) := \frac{\sum_{i:\mathbf{X}_i=\mathbf{x}} R_i T_i}{\#\{j : \mathbf{X}_j = \mathbf{x}, T_j = 1\}} - \frac{\sum_{i:\mathbf{X}_i=\mathbf{x}} R_i (1 - T_i)}{\#\{j : \mathbf{X}_j = \mathbf{x}, T_j = 0\}}$$

to estimate $\delta(\mathbf{x})$, since this parameter has a more straightforward interpretation and is easier to estimate than the covariance.

To estimate $\boldsymbol{\epsilon}$ one may take

$$\hat{\boldsymbol{\epsilon}} := \frac{1}{N} \sum_{i=1}^{N} \hat{\mathrm{e}}(\mathbf{X}_i).$$

This should be (approximately) unbiased for $\boldsymbol{\epsilon}$ when this parameter is computed by averaging the distributions of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$ with equal weights; if other weights are used to compute $\boldsymbol{\epsilon}$ then those weights must be used to compute $\hat{\boldsymbol{\epsilon}}$.

The simplest general method for obtaining confidence intervals for $\mathrm{e}(\mathbf{x})$ and $\boldsymbol{\epsilon}$ (within the scope of our assumptions) is based on the normal approximation to the distributions of $\hat{\mathrm{e}}(\mathbf{x})$ and $\hat{\boldsymbol{\epsilon}}$. For binary treatments this approximation amounts to treating the T_is as fixed and using the central limit theorem: Writing $\mathrm{SE}(\hat{\delta}(\mathbf{x}))$ for the square root of the unbiased estimate of the variance of $\hat{\delta}(\mathbf{x})$ (namely the sum of the sample variance of the R_is in the treated group divided by $\#\{j : \mathbf{X}_j = \mathbf{x}, T_j = 1\} - 1$ and the sample variance of the R_is in the control group divided by $\#\{j : \mathbf{X}_j = \mathbf{x}, T_j = 0\} - 1$) and

$$\mathrm{SE}\left(\frac{1}{N}\sum\nolimits_{i=1}^{N} \hat{\delta}(\mathbf{X}_i)\right)^2 = \frac{1}{N^2}\sum\nolimits_{i=1}^{N} \mathrm{SE}(\hat{\delta}(\mathbf{X}_i))^2,$$

we have, provided neither $\#\{j : \mathbf{X}_j = \mathbf{x}, T_j = 1\}$ nor $\#\{j : \mathbf{X}_j = \mathbf{x}, T_j = 0\}$ is 'too small', that

$$Z(\mathbf{x}) := \frac{\hat{\delta}(\mathbf{x}) - \delta(\mathbf{x})}{\mathrm{SE}(\hat{\delta}(\mathbf{x}))} \quad \text{and} \quad \mathbf{Z} := \frac{\frac{1}{N}\sum_{i=1}^{N} \hat{\delta}(\mathbf{X}_i) - \boldsymbol{\epsilon}}{\mathrm{SE}\left(\frac{1}{N}\sum_{i=1}^{N} \hat{\delta}(\mathbf{X}_i)\right)}$$

are approximately standard normally distributed, and hence may be used to compute approximate confidence intervals for the unknown parameters.

Clearly, these statements implicitly assume that the $\mathbf{X}_i$s are discrete, while in reality they may have continuous coordinates and in general events such as $\{\mathbf{X}_i = \mathbf{x}\}$ must be seen as abbreviations of $\{\mathbf{X}_i \in \mathcal{N}_\mathbf{x}\}$ for some neighbourhood $\mathcal{N}_\mathbf{x}$ determined by the stratification/matching. So the estimation method proposed is approximate not just because of the normal approximation but also because the assumptions it relies on may *in actuality* hold only approximately (even if *formally* they hold exactly). It is also clear from the definitions of $\boldsymbol{\epsilon}$ and $\hat{\boldsymbol{\epsilon}}$ that the method is practicable only in situations where almost all the units fall into useable and relatively large strata. Unfortunately, when stratifying the data on many covariates a substantial portion of them will have to be left out, even if the sample is large. The most favourable situation is perhaps the one in which the one-dimensional propensity score can be accurately estimated (this situation will be considered below); even then, because our normal approximation ignores the variability in the estimate of the propensity score and in the subsequent stratification on it, the formula for the approximate variance of $\hat{\boldsymbol{\epsilon}}$ and the associated confidence intervals for $\boldsymbol{\epsilon}$ will tend to be somewhat optimistic. These observations will be illustrated by the examples of section 6.

A similar method can be used for more general types of treatment, but it is somewhat more involved and the correctness of the corresponding approximation is even more uncertain. First, in order to be able to regard the T_is as fixed and to use the normal approximation on the R_is it appears necessary to rewrite $\kappa(\mathbf{x})$ in terms of expectations conditional on $T_i = t$ as well as on $\mathbf{X}_i = \mathbf{x}$. For example, suppose the T_is only take the values 0, 1 and 2. Then it is seen that

$$\begin{aligned}\kappa(\mathbf{x}) = f^{(\mathbf{x})}(0)f^{(\mathbf{x})}(1)\,\{E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 1] - E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 0]\} + \\ f^{(\mathbf{x})}(1)f^{(\mathbf{x})}(2)\,\{E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 2] - E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 1]\} + \\ 2\,f^{(\mathbf{x})}(0)f^{(\mathbf{x})}(2)\,\{E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 2] - E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 0]\}.\end{aligned}$$

An estimate $\hat{\kappa}(\mathbf{x})$ of $\kappa(\mathbf{x})$ is obtained by replacing $E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = t]$ and the propensity scores $f^{(\mathbf{x})}(t)$ in this formula by their empirical counterparts. Neglecting the variability of the T_is and of the estimates of $f^{(\mathbf{x})}(t)$, a formula for the variance of $\hat{\kappa}(\mathbf{x})$ can be computed. This variance can be estimated in terms of the sample variance of the R_is, yielding a standard error $\mathrm{SE}(\hat{\kappa}(\mathbf{x}))$, and the usual standardized version of $\hat{\kappa}(\mathbf{x})$, formed by subtracting $\kappa(\mathbf{x})$ and dividing by $\mathrm{SE}(\hat{\kappa}(\mathbf{x}))$, can be used to compute approximate intervals for $\kappa(\mathbf{x})$. Intervals for the overall treatment effect are obtained by averaging the $\hat{\kappa}(\mathbf{x})$ estimates as indicated for the case of binary treatments.

Note that nowhere has it been assumed that the $\mathbf{X}_i$s form a random sample—they could, for example, form a stratified sample from a population—nor that the T_is are independent conditionally on the $\mathbf{X}_i$s—they are merely assumed to have the same probability distribution given by (4.1). [For example, if the number of treatments is limited then the treatment assignments to different units may be dependent (but still exchangeable) conditionally on the covariates.] It is only conditionally on the values of the covariates *and* treatment assignments that the responses are assumed to be independent.

However, if we consider estimating ϵ by the method just described but with the propensity scores $f^{(\mathbf{X}_i)}$ in place of the $\mathbf{X}_i$s then we probably need to assume more, namely a combination of **A0′** and **A3**, which together imply that the $(\mathbf{X}_i, T_i, R_i)$s form a random sample:

A3′ $(\mathbf{X}_1, T_1, R_1), (\mathbf{X}_2, T_2, R_2), \ldots, (\mathbf{X}_N, T_N, R_N)$ are independent and identically distributed.

To see that this condition yields **A3** with the $f^{(\mathbf{X}_i)}$s in place of the $\mathbf{X}_i$s—and hence provides some justification for the estimation method when stratifying on the propensity score—note that by **A3′** we have, in the usual notation,

$$P\Big(R_i \le r_i, i \in I \Big| f^{(\mathbf{X}_i)} = f_i, T_i = t_i, i \in I\Big) = \prod_{i \in I} P\Big(R_i \le r_i \Big| f^{(\mathbf{X}_i)} = f_i, T_i = t_i\Big).$$

Finally, let us note that although our definition of overall treatment effect ϵ involved conditioning on $\mathbf{X}_i$ an equivalent definition is obtained by conditioning on $f^{(\mathbf{X}_i)}$. In order to show this, it is enough to verify that integrating

$$P(R_i \le r | \mathbf{X}_i = \mathbf{x}, T_i = t)$$

with respect to the distribution of $\mathbf{X}_i$ gives the same result as integrating

$$P\left(R_i \le r \Big| f^{(\mathbf{X}_i)} = f, T_i = t\right)$$

with respect to the distribution of $f^{(\mathbf{X}_i)}$. For simplicity we do this in the case where $\mathbf{X}_i$ is discrete, when $f^{(\mathbf{X}_i)}$ also is discrete: Using $P\left(T_i = t \middle| f^{(\mathbf{X}_i)} = f\right) = f(t)$ (see (2.3)), we have

$$\sum_f P\Big(R_i \le r \Big| f^{(\mathbf{X}_i)} = f, T_i = t\Big) P\Big(f^{(\mathbf{X}_i)} = f\Big) = \sum_f \frac{P\big(R_i \le r, f^{(\mathbf{X}_i)} = f, T_i = t\big)}{P\left(T_i = t \middle| f^{(\mathbf{X}_i)} = f\right)} =$$

$$\sum_f \sum_{\mathbf{x}: f^{(\mathbf{x})} = f} \frac{P(R_i \le r, \mathbf{X}_i = \mathbf{x}, T_i = t)}{f(t)} = \sum_f \sum_{\mathbf{x}: f^{(\mathbf{x})} = f} \frac{P(R_i \le r, \mathbf{X}_i = \mathbf{x}, T_i = t)}{f^{(\mathbf{x})}(t)} =$$

$$\sum_f \sum_{\mathbf{x}: f^{(\mathbf{x})} = f} \frac{P(R_i \le r, \mathbf{X}_i = \mathbf{x}, T_i = t)}{P(\mathbf{X}_i = \mathbf{x}, T_i = t)} \frac{P(\mathbf{X}_i = \mathbf{x}, T_i = t)}{P(T_i = t | \mathbf{X}_i = \mathbf{x})} =$$

$$\sum_f \sum_{\mathbf{x}: f^{(\mathbf{x})} = f} P(R_i \le r | \mathbf{X}_i = \mathbf{x}, T_i = t) P(\mathbf{X}_i = \mathbf{x}) =$$

$$\sum\nolimits_{\mathbf{x}} P(R_i \le r | \mathbf{X}_i = \mathbf{x}, T_i = t) P(\mathbf{X}_i = \mathbf{x}),$$

as required.

Remark. The method of estimation described here is essentially the one presented in section 2.3 of Lunceford and Davidian [19], who review methods of estimation based on stratification on the propensity score. A class of estimators

that we do not consider consists of weighted averages of the responses weighted by functions of the propensity scores. These seem to be very popular but we have found no evidence that they are superior to the natural, simple and transparent estimator considered here. In fact, the arguments on pp. 279–294 of [10] and the results of [11] indicate that the contrary is probably true in general. □

4.1. Digression: an approach based on predicting 'counterfactuals'

In general, the applicability of the methods considered in this work depends on the availability of large samples. Since this is especially true of the method of stratification on the propensity score function when the latter is estimated by a non-parametric predictor, it is natural to ask whether, in situations where a large sample is available, treatment effects can be estimated directly by a non-parametric predictor of the response conditionally on the covariates and on the treatment. To see that this is possible in principle, recall first from the beginning of this section that integrating

$$\mathrm{e}(\mathbf{x}) = E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 1] - E[R_i|\mathbf{X}_i = \mathbf{x}, T_i = 0] =: \varphi(\mathbf{x}, 1) - \varphi(\mathbf{x}, 0)$$

with respect to the distribution of the $\mathbf{X}_i$s yields the treatment effect

$$\boldsymbol{\epsilon} = E[\mathrm{e}(\mathbf{X})] = E[\varphi(\mathbf{X}, 1)] - E[\varphi(\mathbf{X}, 0)].$$

If N is large, this is close to

$$\boldsymbol{\epsilon}_N := \frac{1}{N}\sum\nolimits_{i=1}^{N} \varphi(\mathbf{X}_i, 1) - \varphi(\mathbf{X}_i, 0).$$

Consequently, if N is large and if for each $\mathbf{x} \in \mathbb{R}^d$ and $t = 0, 1$ we can find an estimator $\hat{\varphi}_N(\mathbf{x}, t)$ of $\varphi(\mathbf{x}, t) = E[R|\mathbf{X} = \mathbf{x}, T = t]$ then we can estimate both $\boldsymbol{\epsilon}_N$ and $\boldsymbol{\epsilon}$ by

$$\hat{\epsilon}_N := \frac{1}{N}\sum\nolimits_{i=1}^{N} \hat{\varphi}_N(\mathbf{X}_i, 1) - \hat{\varphi}_N(\mathbf{X}_i, 0). \tag{4.3}$$

Now let $\Pi(\mathbf{x}, t)$ be a consistent estimator of $\varphi(\mathbf{x}, t)$ constructed from the data; for example, $\Pi(\mathbf{x}, t)$ may be a random forest predictor or a non-parametric regression (e.g. Nadaraya-Watson) predictor. Then an estimate of $\varphi(\mathbf{X}_i, 1)$ is

$$\hat{\varphi}_N(\mathbf{X}_i, 1) := T_i R_i + (1 - T_i)\Pi(\mathbf{X}_i, 1),$$

and an estimate of $\varphi(\mathbf{X}_i, 0)$ is

$$\hat{\varphi}_N(\mathbf{X}_i, 0) := (1 - T_i)R_i + T_i\Pi(\mathbf{X}_i, 0).$$

Indeed, if $T_i = 1$ then R_i is an unbiased estimate of $\varphi(\mathbf{X}_i, 1)$; and if $T_i = 0$ then $\Pi(\mathbf{X}_i, 1)$ is an approximately unbiased estimate of $\varphi(\mathbf{X}_i, 1)$, which in that case can be regarded as the expected value of the response R_i of the (unobserved) *counterfactual observation* $(\mathbf{X}_i, 1, R_i)$ associated with the observation $(\mathbf{X}_i, 0, R_i)$. Similarly, if $T_i = 0$ then R_i is an unbiased estimate of $\varphi(\mathbf{X}_i, 0)$; and

if $T_i = 1$ then $\Pi(\mathbf{X}_i, 0)$ is an estimate of $\varphi(\mathbf{X}_i, 0)$, the expected value of the response of the counterfactual $(\mathbf{X}_i, 0, R_i)$ associated with $(\mathbf{X}_i, 1, R_i)$.

With these estimates of $\varphi_N(\mathbf{X}_i, 1)$ and $\varphi_N(\mathbf{X}_i, 0)$, the estimator (4.3) has similarities with the 'matching estimators' studied in [1], with the parametric estimator of Snowden, Rose and Mortimer [37], who attribute it to Robins [25], and with the non-parametric estimator studied in [4]. We expect it to be *approximately* unbiased for the treatment effect provided $\Pi(\mathbf{x}, t)$ is a consistent predictor. Because its bias is difficult to gauge—if only because it is intended to estimate ϵ_N rather than ϵ—and its variance difficult to estimate (some simulations of ours and the work of [1] concerning matching estimators suggest that the bootstrap cannot be used to obtain variance estimates), such an estimator may be of limited value. However, it can be useful as a check on the estimates obtained by other means (e.g. by stratification on the propensity score). With a very large sample one may even consider splitting the data into a number of subsets of equal size, compute an estimate of the treatment effect with each subset, and then compute a variance estimate from the sample of treatment effect estimates; this variance estimate is typically an upper bound for the variance of the treatment effect estimate based on the whole data set and hence may be used to produce a conservative confidence interval for the treatment effect. The method will be illustrated in subsection 6.6.

4.2. Case-referent studies

Case-referent studies serve to illustrate other ways of quantifying and estimating a treatment effect. In principle, any measure of discrepancy $\mathrm{e}(\mathbf{x})$ between

$$P(R = 1|T = 1, \mathbf{X} = \mathbf{x}) \quad \text{and} \quad P(R = 1|T = 0, \mathbf{X} = \mathbf{x}) \tag{4.4}$$

can be used to quantify the treatment effect in a matched set of units characterized by the same value $\mathbf{x}$ of the covariate. Once such a measure has been chosen we can integrate $\mathbf{x}$ out of it—more precisely: integrate it with respect to $P(\mathbf{X} \leq \mathbf{x}|R = 0)$—to get the corresponding overall treatment effect

$$\boldsymbol{\epsilon} := E[\mathrm{e}(\mathbf{X})|R = 0],$$

where the conditioning on $R = 0$ reflects the fact that, by virtue of the sampling scheme described in example 3.6, all the values of the covariates that show up in the sample result from sampling conditionally on that event.

The problem is that the probabilities in (4.4) cannot be estimated from the data in a case-control study, because the number of diseased individuals is practically fixed by the sampling scheme; and it is not obvious that there is *a function* of those probabilities that can be estimated. Perhaps surprisingly, there is one such function—one that is difficult to estimate with small samples (at least if one does not possess a credible parametric model for the data) but still worth considering. The function in question is the *odds ratio*, or, more precisely, the odds ratio conditional on the event $\mathbf{X} = \mathbf{x}$, defined by

$$\mathrm{e}(\mathbf{x}) = \frac{\mathcal{O}(R = 1|T = 1, \mathbf{X} = \mathbf{x})}{\mathcal{O}(R = 1|T = 0, \mathbf{X} = \mathbf{x})}, \tag{4.5}$$

where

$$\mathcal{O}(R=1|T=t,\mathbf{X}=\mathbf{x})=\frac{P(R=1|T=t,\mathbf{X}=\mathbf{x})}{1-P(R=1|T=t,\mathbf{X}=\mathbf{x})}$$

is the *odds* of an individual with characteristics $\mathbf{x}$ being diseased given that he was exposed to treatment t, an increasing function of $P(R=1|T=t,\mathbf{X}=\mathbf{x})$. Under the null hypothesis of no treatment effect, $\mathcal{O}(R=1|T=t,\mathbf{X}=\mathbf{x})$ is independent of t and hence $\mathrm{e}(\mathbf{x})=1$, while if the exposure is associated with the disease then $P(R=1|T=1,\mathbf{X}=\mathbf{x})>P(R=1|T=0,\mathbf{X}=\mathbf{x})$ and therefore $\mathrm{e}(\mathbf{x})>1$. The significance of the odds ratio in connection with case-control studies is due to the property (first used by Cornfield [6] and easily verified) that

$$\mathrm{e}(\mathbf{x})=\frac{\mathcal{O}(T=1|R=1,\mathbf{X}=\mathbf{x})}{\mathcal{O}(T=1|R=0,\mathbf{X}=\mathbf{x})}=\frac{p_{\mathbf{x}}(1-q_{\mathbf{x}})}{q_{\mathbf{x}}(1-p_{\mathbf{x}})}, \tag{4.6}$$

where we write

$$p_{\mathbf{x}}:=P(T=1|R=1,\mathbf{X}=\mathbf{x}) \quad \text{and} \quad q_{\mathbf{x}}:=P(T=1|R=0,\mathbf{X}=\mathbf{x});$$

that is, $\mathrm{e}(\mathbf{x})$ also compares the odds of an individual having been exposed given that he is diseased with the odds of an individual having being exposed given that he is not diseased. Although it is (4.5) that is regarded as meaningful—because (assuming that disease is associated with exposure) it is usually the exposure that causes the disease—the identity (4.6) shows that the odds ratio is not just a function of the inestimable probabilities (4.4) but also a function of the probabilities $p_{\mathbf{x}}$ and $q_{\mathbf{x}}$, which *can* be estimated, by

$$\hat{p}_{\mathbf{x}}:=\frac{\sum_{\{i:R_i=1,\mathbf{x}_i=\mathbf{x}\}}T_i}{\#\{j:R_j=1,\mathbf{X}_j=\mathbf{x}\}}=:\frac{1}{m_{\mathbf{x}}}\sum\nolimits_{i:R_i=1,\mathbf{X}_i=\mathbf{x}}T_i$$

and

$$\hat{q}_{\mathbf{x}}:=\frac{\sum_{\{i:R_i=0,\mathbf{x}_i=\mathbf{x}\}}T_i}{\#\{j:R_j=0,\mathbf{X}_j=\mathbf{x}\}}=:\frac{1}{n_{\mathbf{x}}-m_{\mathbf{x}}}\sum\nolimits_{i:R_i=0,\mathbf{X}_i=\mathbf{x}}T_i.$$

Since, besides being unbiased, $\hat{p}_{\mathbf{x}}$ and $\hat{q}_{\mathbf{x}}$ are approximately normal for large $m_{\mathbf{x}}$ and $n_{\mathbf{x}}-m_{\mathbf{x}}$,

$$\hat{\mathrm{e}}(\mathbf{x}):=\frac{\hat{p}_{\mathbf{x}}(1-\hat{q}_{\mathbf{x}})}{\hat{q}_{\mathbf{x}}(1-\hat{p}_{\mathbf{x}})}$$

may be taken as a slightly biased but consistent estimator of the 'conditional' odds ratio $\mathrm{e}(\mathbf{x})$, and consequently

$$\hat{\boldsymbol{\epsilon}}:=\sum\nolimits_{\mathbf{x}}\hat{\mathrm{e}}(\mathbf{x})\frac{n_{\mathbf{x}}}{n},$$

with $n=\sum_{\mathbf{x}}n_{\mathbf{x}}$, as a consistent and approximately normally distributed estimator of the overall treatment effect, or 'average odds ratio', $\boldsymbol{\epsilon}$. For large $m_{\mathbf{x}}$ and $n_{\mathbf{x}}-m_{\mathbf{x}}$ one can even compute estimates of $\mathrm{Var}[\hat{\mathrm{e}}(\mathbf{x})]$, and from them a rough estimate of $\mathrm{Var}(\hat{\boldsymbol{\epsilon}})$ and a corresponding confidence interval for $\boldsymbol{\epsilon}$.

Unfortunately, as already hinted, this program is seldom practicable because case-control studies are typically characterized by small values of $m_{\mathbf{x}}$. In the absence of a reliable estimate of the average odds ratio, one may try to quantify the treatment effect *indirectly*, as a function of $p_{\mathbf{x}}$ and $q_{\mathbf{x}}$. For example,

$$\mathrm{e}'(\mathbf{x}) := p_{\mathbf{x}} - q_{\mathbf{x}} \equiv P(T=1|R=1,\mathbf{X}=\mathbf{x}) - P(T=1|R=0,\mathbf{X}=\mathbf{x})$$

and $\boldsymbol{\epsilon}' := E[\mathrm{e}'(\mathbf{X})|R=0]$ could sometimes be useful measures of *unconfounded* association between the treatment and the response, and they are comparatively easy to estimate.

Remark. Note that **A2** does not hold in a case-referent study, since the responses of different units in a matched set are not identically distributed, nor does **A3** (the number of positive responses in a matched set being fixed). What we have instead is that conditionally on the covariates being equal to a given $\mathbf{x}$ *and* on the responses being equal to a given $r \in \{0,1\}$ the treatments are independent and identically distributed. This can be regarded as a version of **A2** (conditioning on the responses as well) combined with a version of **A3** with the responses and treatments interchanged. □

5. A model for simulating observational data

Our purpose in this section is to define a simple model for simulating observational data. The model will be used in section 6 to illustrate several aspects of the methods presented in sections 3 and 4. Evidently, the whole point of using simulated data to study the workings of a statistical method is that one knows exactly the answers one should get—for example: which conclusion concerning the existence or non-existence of a treatment effect—and hence can ascertain the correctness and accuracy of the method at least in some specific situations.

We shall define the model by specifying the distribution of a random vector $(\mathbf{X}, T, R)$; once this is done we show how $\mathbf{X}$, T and R arise from a special case of the basic model of section 2.

Let a p-dimensional random vector $\mathbf{U}$ represent a set of covariates of an individual randomly sampled from some population and let T indicate the individual's treatment assignment, $T=1$ indicating that the individual is treated and $T=0$ that the individual is not treated—i.e. that the individual is a control. The distribution of $\mathbf{U}$ conditionally on $T=t$ is normal with mean vector $\boldsymbol{\mu}_t$ and full-rank covariance matrix Σ_t $(t=0,1)$. Thus the conditional density of $\mathbf{U}$ given $T=t$ is

$$f_{\mathbf{U}|T=t}(\mathbf{u}) = (2\pi)^{-p/2}|\Sigma_t|^{-1/2}e^{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_t)^{\mathrm{T}}\Sigma_t^{-1}(\mathbf{u}-\boldsymbol{\mu}_t)} \quad (\mathbf{u}\in\mathbb{R}^p, t=0,1).$$

Writing $p_t = P(T=t)$ and $f_{T|\mathbf{U}=\mathbf{u}}(t) = P(T=t|\mathbf{U}=\mathbf{u})$ we have by Bayes rule

$$f_{T|\mathbf{U}=\mathbf{u}}(t) = \frac{p_t|\Sigma_t|^{-1/2}e^{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_t)^{\mathrm{T}}\Sigma_t^{-1}(\mathbf{u}-\boldsymbol{\mu}_t)}}{p_0|\Sigma_0|^{-1/2}e^{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_0)^{\mathrm{T}}\Sigma_0^{-1}(\mathbf{u}-\boldsymbol{\mu}_0)} + p_1|\Sigma_1|^{-1/2}e^{-\frac{1}{2}(\mathbf{u}-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{u}-\boldsymbol{\mu}_1)}}$$

($t = 0, 1, \mathbf{u} \in \mathbb{R}^p$). This probability function describes the effect of $\mathbf{U}$ on the treatment assignment T.

In order to specify the distribution of an individual's response R let us first introduce a q-dimensional random vector $\mathbf{V}$ to represent another set of covariates and set $\mathbf{X} = (\mathbf{U}, \mathbf{V})$. We wish that $\mathbf{X}$ influence R, that $\mathbf{U}$ and $\mathbf{V}$ be dependent, but that T be influenced by $\mathbf{X}$ only through $\mathbf{U}$ (as already specified). The latter requirement is that

$$f_{T|\mathbf{X}=\mathbf{x}}(t) := P(T = t|\mathbf{X} = \mathbf{x}) = P(T = t|\mathbf{U} = \mathbf{u}) = f_{T|\mathbf{U}=\mathbf{u}}(t),$$

for $t = 0, 1$, $\mathbf{x} = (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{p+q}$, which completes the specification of the distribution of T conditionally on $(\mathbf{U}, \mathbf{V})$ and, provided $\mathbf{U}$ and $\mathbf{V}$ are given a joint density, implies

$$f_{\mathbf{V}|\mathbf{U}=\mathbf{u},T=t}(\mathbf{v}) = f_{\mathbf{V}|\mathbf{U}=\mathbf{u}}(\mathbf{v}) \quad (t = 0, 1, \mathbf{u} \in \mathbb{R}^p, \mathbf{v} \in \mathbb{R}^q),$$

where both the left and right hand sides denote conditional densities. To specify the distribution of $\mathbf{U}$ and $\mathbf{V}$ we require that the distribution of $\mathbf{V}$ conditionally on $\mathbf{U} = \mathbf{u}$ be normal with mean vector $\boldsymbol{\nu} + \mathbf{B}(\mathbf{u} - \boldsymbol{\mu})$, where $\boldsymbol{\nu} \in \mathbb{R}^q$, $\boldsymbol{\mu} := p_0\boldsymbol{\mu}_0 + p_1\boldsymbol{\mu}_1$ and $\mathbf{B}$ is a $q \times p$ matrix, and full-rank covariance matrix $\mathbf{M}$:

$$E[\mathbf{V}|\mathbf{U} = \mathbf{u}] = \boldsymbol{\nu} + \mathbf{B}(\mathbf{u} - \boldsymbol{\mu}), \quad \mathrm{Var}[\mathbf{V}|\mathbf{U} = \mathbf{u}] = \mathbf{M}.$$

Having specified the joint distribution of T and $\mathbf{X}$, let us now specify the distribution of the response R conditionally on $(\mathbf{X}, T)$. We shall consider the case where the response is a continuous quantity and set

$$R = \mu + \sigma\xi + A_0(\mathbf{X})(1 - T) + A_1(\mathbf{X})T, \tag{5.1}$$

where $\boldsymbol{\mu} \in \mathbb{R}$, $\sigma > 0$, A_0 and A_1 are real-valued functions defined on $\mathbb{R}^{p+q}$, and ξ is a standard normal random variable independent of $\mathbf{X}$ and T. Then the distribution of R conditionally on $T = t$ and $\mathbf{X} = \mathbf{x}$ is normal with mean $\mu + A_0(\mathbf{x})(1 - t) + A_1(\mathbf{x})t$ and variance σ^2:

$$E[R|\mathbf{X} = \mathbf{x}, T = t] = \mu + A_0(\mathbf{x})(1 - t) + A_1(\mathbf{x})t, \quad \mathrm{Var}[R|\mathbf{X} = \mathbf{x}, T = t] = \sigma^2.$$

Thus an individual's response is influenced by its 'characteristics' through A_1 if the individual is treated and through A_0 if the individual is not treated. The exponential of R could, for example, be interpreted as the individual's survival time.

In this setting, the (*mean*) *treatment effect* on an individual with characteristics $\mathbf{x}$ is

$$\mathrm{e}(\mathbf{x}) := E[R|\mathbf{X} = \mathbf{x}, T = 1] - E[R|\mathbf{X} = \mathbf{x}, T = 0] = A_1(\mathbf{x}) - A_0(\mathbf{x}),$$

and the *overall treatment effect* is

$$\boldsymbol{\epsilon} := E[\mathrm{e}(\mathbf{X})] = E[A_1(\mathbf{X})] - E[A_0(\mathbf{X})].$$

Integrating $E[R|\mathbf{X}=\mathbf{x},T=t]$ with respect to the distribution of $\mathbf{X}|T=t$ yields

$$E[R|T=t]=\mu+(1-t)E[A_0(\mathbf{X})|T=t]+tE[A_1(\mathbf{X})|T=t],$$

hence the *mean difference* between the responses in the treated and control groups is

$$\Delta:=E[R|T=1]-E[R|T=0]=E[A_1(\mathbf{X})|T=1]-E[A_0(\mathbf{X})|T=0].$$

The 'naive approach' (see the first remark of section 4, on p. 146) to estimating and testing for a treatment effect based on a random sample of vectors with the same distribution as $(\mathbf{X},T,R)$ would consist in estimating this Δ by the difference between the sample means of the responses in the treated and control groups. However, as pointed out earlier and is evident by comparing the expressions of $\boldsymbol{\epsilon}$ and Δ, we have $\Delta\neq\boldsymbol{\epsilon}$ unless, for example, $\mathbf{X}$ and T are independent—in which case $E[A_1(\mathbf{X})]=E[A_1(\mathbf{X})|T=1]$ and $E[A_0(\mathbf{X})]=E[A_0(\mathbf{X})|T=0]$.

As is evident from the very definition of $\boldsymbol{\epsilon}$, a correct approach is to estimate $\mathrm{e}(\mathbf{x})$ by *matching* or *stratifying on* $\mathbf{x}$—comparing individuals who differ with respect to treatment but have the same characteristics $\mathbf{x}$—and then average the estimates obtained across the different values of $\mathbf{x}$. In most studies this is probably the only correct approach to testing for a treatment effect; with few exceptions, assuming a regression model for the response as a function of $(\mathbf{X},T)$ and estimating the parameter(s) pertaining to T, as is often done to this day, is at best a 'sophisticatedly naive approach', as amply demonstrated by [9] and argued in chapter 14 of [30], for example.

Of course, the functions A_0 and A_1 can be almost anything. However, in section 6 we shall take them to be linear, namely of the form

$$A_t(\mathbf{x})=\mathbf{x}\cdot\boldsymbol{\beta}_t\quad(t=0,1,\mathbf{x}\in\mathbb{R}^{p+q}),\tag{5.2}$$

where $\boldsymbol{\beta}_t\in\mathbb{R}^{p+q}$. This is a convenient choice from a pedagogical point of view (though perhaps not a very realistic one) because it allows the explicit calculation of $\mathrm{e}(\mathbf{x})$, $\boldsymbol{\epsilon}$ and Δ, and an easy manipulation of these in terms of the parameters $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$. Thus, with this choice we have

$$\mathrm{e}(\mathbf{x})=\mathbf{x}\cdot(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)\quad\text{and}\quad\boldsymbol{\epsilon}=(\boldsymbol{\mu},\boldsymbol{\nu})\cdot(\boldsymbol{\beta}_1-\boldsymbol{\beta}_0)=(\boldsymbol{\mu},\boldsymbol{\nu})\cdot\boldsymbol{\beta}_1-(\boldsymbol{\mu},\boldsymbol{\nu})\cdot\boldsymbol{\beta}_0$$

(recall that $E[\mathbf{X}]=(E[\mathbf{U}],E[\mathbf{V}])=(\boldsymbol{\mu},\boldsymbol{\nu})$), and

$$\Delta=(\boldsymbol{\mu}_1,\boldsymbol{\nu}_1)\cdot\boldsymbol{\beta}_1-(\boldsymbol{\mu}_0,\boldsymbol{\nu}_0)\cdot\boldsymbol{\beta}_0,$$

where $\boldsymbol{\nu}_t:=E(\mathbf{V}|T=t)=E(\boldsymbol{\nu}+\mathbf{B}(\mathbf{U}-\boldsymbol{\mu})|T=t)=\boldsymbol{\nu}+\mathbf{B}(\boldsymbol{\mu}_t-\boldsymbol{\mu})$, $t=0,1$.

Comparing $\boldsymbol{\epsilon}$ and Δ, we see that the latter has $(\boldsymbol{\mu}_1,\boldsymbol{\nu}_1)$ and $(\boldsymbol{\mu}_0,\boldsymbol{\nu}_0)$ where the former has $(\boldsymbol{\mu},\boldsymbol{\nu})$. If $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are both equal to some $\boldsymbol{\beta}$ then $\boldsymbol{\epsilon}=0$ and there is no treatment effect; however, in this case the 'misleading parameter' Δ becomes $\Delta=(\boldsymbol{\mu}_1-\boldsymbol{\mu}_0,\boldsymbol{\nu}_1-\boldsymbol{\nu}_0)\cdot\boldsymbol{\beta}$, which will not be zero if $\boldsymbol{\mu}_1\neq\boldsymbol{\mu}_0$ or $\boldsymbol{\nu}_1\neq\boldsymbol{\nu}_0$ (i.e. if the covariates influence treatment assignment) and $\boldsymbol{\beta}\neq0$ (i.e. if the covariates influence the response).

We close this section by expressing our model in terms of equations (2.1) of the basic model. The covariates $\mathbf{X} = (\mathbf{U}, \mathbf{V})$ can be thought of as arising first, in two stages: from the generation of $\mathbf{U}$ according to a mixture of normal distributions (with means $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, covariance matrices Σ_0 and Σ_1, and mixture probabilities p_0, p_1) and then from the generation of $\mathbf{V}$ according to a normal distribution with mean $\boldsymbol{\nu} + \mathbf{B}(\mathbf{U} - \boldsymbol{\mu})$ and covariance matrix $\mathbf{M}$, say by

$$\mathbf{V} = \psi(\mathbf{U}, U_1),$$

for an appropriate function ψ and a uniform random variable U_1 independent of $\mathbf{X}$ (the particular choice of ψ being irrelevant). Next, the treatment assignment is determined by a Bernoulli random variable with 'success probability' that depends only on $\mathbf{U}$ and hence may arise as

$$T = \tau(\mathbf{U}, U)$$

for some function τ and a uniform random variable U independent of $\mathbf{X}$ and U_1 (the particular form of τ is irrelevant provided it is consistent with the expression for $f_{T|\mathbf{U}=\mathbf{u}}(t)$ given early in this section). Finally, the response is determined in terms of the covariates, of the treatment and of ξ by

$$\begin{aligned} R &= \mu + \sigma\xi + A_0(\mathbf{X})(1-T) + A_1(\mathbf{X})T \\ &=: \rho_1(\xi, \mathbf{X}, T) \equiv \rho_1(\xi, \mathbf{U}, \mathbf{V}, T) \\ &\equiv \rho_1(\xi, \mathbf{U}, \psi(\mathbf{U}, U_1), T) \\ &=: \rho_2(\xi, U_1, \mathbf{U}, T) \\ &=: \rho(V, \mathbf{U}, T), \end{aligned}$$

where V is a uniform random variable constructed from ξ and U_1 (e.g. by alternating the digits in the decimal expansions of their fractional parts), which therefore is independent of the U involved in the generation of T. Thus, both $(\mathbf{U}, T, R)$ and $(\mathbf{X}, T, R)$ satisfy the basic model (note that in the expression $R = \rho_1(\xi, \mathbf{X}, T)$ the role of the V in $R = \rho(V, \mathbf{U}, T)$ is assumed by ξ, which of course can be written as a function of a uniform, and that $T = \tau(\mathbf{U}, U)$ can be written as $T = \tau_1(U, \mathbf{X})$ for some τ_1 that is constant in its last q variables), so one can take care of confounding both by conditioning on $\mathbf{X}$ and by conditioning on $\mathbf{U}$, the latter being of course more efficient from a statistical point of view.

If $(\mathbf{X}_1, T_1, R_1), (\mathbf{X}_2, T_2, R_2), \ldots, (\mathbf{X}_N, T_N, R_N)$ arise independently and by the same mechanism as $(\mathbf{X}, T, R)$ does, then they clearly satisfy **A0**, **A0′**, **A1**, **A1′**, **A2**, **A3** and **A3′**—all the assumptions required by the various methods of stratification/matching on the $\mathbf{X}_i$s, on the $\mathbf{U}_i$s, or on the corresponding propensity scores. Our illustrations will be based on simulated samples of such vectors.

6. Some illustrations based on simulated data[12]

In order to simulate data from the model presented in section 5 it is convenient to fix the values of certain of its parameters once and for all. These parameters, which we call *primary parameters*, are those that, except for (p_0, p_1), determine the distribution of the vector of covariates $\mathbf{X} = (\mathbf{U}, \mathbf{V})$:

- $\boldsymbol{\mu}_0 = E[\mathbf{U}|T = 0]$ and $\boldsymbol{\mu}_1 = E[\mathbf{U}|T = 1]$;
- $\Sigma_0 = \mathrm{Var}[\mathbf{U}|T = 0]$ and $\Sigma_1 = \mathrm{Var}[\mathbf{U}|T = 1]$;
- $\boldsymbol{\nu} = E[\mathbf{V}]$ and $\mathrm{Var}[\mathbf{V}|\mathbf{U} = \mathbf{u}] = \mathbf{M}$;
- $\mathbf{B}$, the matrix determining $\mathbf{V}$ from $\mathbf{U}$ by $E[\mathbf{V}|\mathbf{U} = \mathbf{u}] = \boldsymbol{\nu} + \mathbf{B}(\mathbf{u} - \boldsymbol{\mu})$.

Having fixed the values of the primary parameters we may choose the *secondary parameters*—which determine the distribution of T conditionally on the value of $\mathbf{X}$ and the distribution of R conditionally on $(\mathbf{X}, T)$—differently in different examples:

- $p_0 = P(T = 0)$ and $p_1 = P(T = 1)$, which determine $\boldsymbol{\mu} := p_0\boldsymbol{\mu}_0 + p_1\boldsymbol{\mu}_1$;
- $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$, which determine the functions A_0 and A_1 of (5.2);
- μ and σ, which determine R conditionally on $(\mathbf{X}, T)$ by (5.1).

Although we do not wish to fashion our data sets after a particular real data set, it seems desirable to provide the distribution of $\mathbf{X}$ with a modicum of verisimilitude. Thus we take as values of the primary parameters the parameter estimates obtained by fitting the model for $\mathbf{X}$ to a subset of the `SAheart` or 'South-African Heart Disease' data set [32, 7]) consisting of the CHD (coronary heart disease) status and eight potential risk factors of 192 patients with family history of heart disease. More precisely, we identify positive and negative CHD statuses with the events $T = 1$ and $T = 0$, respectively, fit the model for $\mathbf{U}$ conditional on the event $T = 0$ to *transformed versions* of the variables[13]

```
systolic.blood.pressure,
cumulative.tobacco,
LDL.cholesterol,
adiposity,
type.A.behaviour,
age.at.onset,
```

in the subset of 96 patients not diagnosed with CHD and the model for $\mathbf{U}$ conditional on $T = 1$ to transformed versions of the same variables in the subset of 96 patients diagnosed with CHD, and this provides values for $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, Σ_0 and

[12] The results presented in this section may be reproduced by running `R` [23] scripts written by the author and which can be had upon request.

[13] The names of the variables we use are slightly different from—and more informative than—those used in the `SAheart` data set.

Σ_1.[14] Then we determine $\boldsymbol{\nu}$ and $\boldsymbol{B}$ by fitting the relationship

$$E[\mathbf{V}|\mathbf{U} = \mathbf{u}] = \boldsymbol{\nu} + \boldsymbol{B}(\mathbf{u} - E[\mathbf{U}])$$

by the least squares method to the transformed versions of the six variables just mentioned and the transformed versions of the variables

```
obesity,
alcohol,
```

which stand for $\mathbf{V}$, in the whole set of 192 patients.

The transformations involved in this procedure are one-to-one and, of course, aimed at making the marginal distributions of the data look normal. Appendix B shows histograms of the eight transformed variables in the group without CHD and in the group with CHD. These can be compared with the analogous plots in appendix C, obtained from a data set *simulated* from the fitted model, the group without CHD being now labelled as 'control' ($T = 0$) and the group with CHD as 'treated' ($T = 1$). As required, there is an overall similarity between real and simulated data. Also, the real data show no obvious signs of non-linearity in the relationships between pairs of variables (scatter plots are not shown).

In the original data, the assignment to 'treatment' depends mainly on the six variables represented by $\mathbf{U}$ and very little on the two variables represented by $\mathbf{V}$. This follows from a prediction analysis by the random forest algorithm [18], summarized in figure B.3 (appendix B), in which CHD status is predicted on the basis of the eight covariates. The variable importance plot of figure C.3 (appendix C), obtained by a prediction analysis of a *simulated* data set, gives an idea about the relative importance of the six variables represented by $\mathbf{U}$ to the determination of treatment assignment (see [18] for an explanation of variable importance). Comparison of this plot with the analogous plot of figure B.3, based on the real data, indicates an overall agreement between the model for $(\mathbf{X}, T)$ and the data set used to estimate its parameters (simulations suggest that the greater difference in importance between `age.at.onset` and `LDL.cholesterol` observed in the simulated data set can be attributed largely to sampling variation).

The following computer output provides the values of the primary parameters. The names of the variables are given in order to help fixing ideas; they actually refer to the *transformed versions* of the variables.

$\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$:

	mu.0	mu.1
systolic.blood.pressure	4.9097807	4.9599668
cumulative.tobacco	-0.4337722	0.1350606
LDL.cholesterol	1.4126311	1.6997419
adiposity	25.4662500	28.6969792
type.A.behaviour	52.8020833	54.4479167
age.at.onset	-0.5920987	0.1155519

[14] Although this is largely irrelevant for our purposes, the identification of treatment assignment with CHD status could be justified on the grounds that diagnosis of CHD in a patient should imply a special treatment. Of course, the response that we generate (for purposes of fixing one truth) has no correspondence in the `SAheart` data set, but it could be thought of as a function of age at death or as a measure of improvement in health after so many years.

Σ_0:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
systolic.blood.pressure	0.0128	0.0226	0.0085	0.3155	0.0848	0.0254
cumulative.tobacco	0.0226	0.9587	0.0594	1.5676	0.3252	0.4002
LDL.cholesterol	0.0085	0.0594	0.1447	1.0791	-0.2475	0.1157
adiposity	0.3155	1.5676	1.0791	55.3343	-4.5905	4.9472
type.A.behaviour	0.0848	0.3252	-0.2475	-4.5905	74.0552	-1.6215
age.at.onset	0.0254	0.4002	0.1157	4.9472	-1.6215	1.1272

Σ_1:

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
systolic.blood.pressure	0.0239	0.0182	-0.0009	0.1222	-0.2443	0.0340
cumulative.tobacco	0.0182	0.8575	-0.0129	0.5398	-0.1474	0.1534
LDL.cholesterol	-0.0009	-0.0129	0.1477	1.1354	0.4164	-0.0058
adiposity	0.1222	0.5398	1.1354	47.3851	-2.0867	1.5402
type.A.behaviour	-0.2443	-0.1474	0.4164	-2.0867	111.5130	-2.3206
age.at.onset	0.0340	0.1534	-0.0058	1.5402	-2.3206	0.5938

$\mathbf{B}^{\mathrm{T}}$ (i.e. the transpose of $\mathbf{B}$):

	obesity	alcohol
systolic.blood.pressure	0.0392153533	1.152472046
cumulative.tobacco |	0.0089462434	0.283561635
LDL.cholesterol	0.0007067134	-0.542370013
adiposity	0.0177522162	0.009332837
type.A.behaviour	0.0006056793	-0.001832629
age.at.onset	-0.0372922934	-0.203472847

$\mathbf{M}$:

	obesity	alcohol
obesity	0.02409144	0.01541203
alcohol	0.01541203	1.97418432

$\boldsymbol{\nu}$:

obesity	3.2697225
alcohol	-0.6140419

Subsections 6.1–6.6 illustrate several aspects of the methods presented in sections 3 and 4. The examples are based on simulated random samples of vectors $(\mathbf{X}_i, T_i, R_i)$ from the model of section 5; as said earlier, such samples satisfy the basic model as well as the assumptions necessary for the (approximate) validity of the methods.

6.1. Stratification in a situation where no treatment effect exists

We begin with a situation where treatment has no effect, setting

$$\boldsymbol{\beta} := \boldsymbol{\beta}_0 = \boldsymbol{\beta}_1 = (3, 7, 8, 4, 2, 8, 2, 10)/10.$$

We take $p_0 = 0.75$ and $p_1 = 0.25$ as the proportions of control and treated patients. Together with $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ these yield the following value for $\boldsymbol{\mu}$:

```
systolic.blood.pressure    4.9223272
cumulative.tobacco        -0.2915640
LDL.cholesterol            1.4844088
adiposity                 26.2739323
type.A.behaviour          53.2135417
age.at.onset              -0.4151861
```

These choices lead to $\Delta = 2.78$, so the naive approach will tend to conclude for a treatment effect, while we know that $e(\mathbf{x}) = 0$ for all $\mathbf{x} \in \mathbb{R}^8$ and hence $\boldsymbol{\epsilon} = 0$. Finally, we take $\mu = 0$ and $\sigma = 1$ in the error part of (5.1).

We simulate a sample of $N = 10,000$ independent random vectors following the model of section 5. The box plots of figure D.1 summarize the distributions of the response and of the eight covariates in the treated and control groups. With the possible exception of `alcohol`, the distributions of the covariates differ clearly in the two groups, which suggests that they have to be taken into account in testing for a treatment effect. However, we know that `obesity` (just as `alcohol`) has no influence on treatment assignment; so the apparent difference between treated and control groups regarding this covariate is a result of its dependence on the other six covariates, which do influence treatment.

Computing the difference between average responses in the treated and control groups we get $\hat{\Delta} = 2.66$ as an estimate of Δ; the usual approximate 95% confidence interval for Δ based on the normal approximation is $[2.46, 2.85]$. Thus, with this sample the naive approach would lead us to conclude that there is a clear treatment effect. In order to test for a treatment effect we stratify the sample on the covariates and use one of the tests described in subsection 3.2.[15]

The stratification of the sample is conveniently carried out by dividing the range of each of the eight covariates into intervals determined by quantiles. If quantiles of probabilities $1/i$, $2/i, \ldots, (i-1)/i$ are used then the range of each covariate is split into i intervals and the 8-dimensional range of the vector of covariates is partitioned into i^8 cells, each of which corresponds to a *potential* stratum (many of the strata will be empty). To get us started let us take $i = 4$, which leads to $i^8 = 65,536$ potential strata. With our sample, the resulting stratification contains 8565 non-empty strata and 389 useable strata (strata with at least one control and one treated unit). Of course, splitting the range of the variables into four intervals only may not be sufficient to reduce the bias in testing for a treatment effect, which always exists because the covariates are continuous and hence units within the same stratum are never fully comparable.

The following computer output shows the number of units, the number of treated units, the number of controls, and the sample means of the responses of treated and control units within a few strata. The strata, given in the rightmost column, are denoted by the intervals of the variables to which the units composing them belong; for instance, the stratum denoted by 4/3/3/3/4/2/4/4 consists of units with values of `systolic.blood.pressure` in the fourth interval, values of `cumulative.tobacco` in the third interval, values of `LDL.cholesterol` in the third interval, etc.

[15] The estimation method of section 4 is not applicable in our first two illustrations because stratification on the covariates uses only a small portion of the data.

sample.size	no.treated	no.controls	mean.of.treated	mean.of.control	stratum
2	1	1	27.12426	29.22979	4/3/3/3/4/2/4/4
5	3	2	29.15665	28.35064	4/4/4/4/3/4/4/1
4	1	3	26.02765	26.56545	4/2/3/4/2/3/4/3
2	1	1	19.59144	22.07207	4/1/4/2/3/1/2/1
3	1	2	15.87238	17.71164	3/1/1/1/3/1/2/3
2	1	1	31.72296	29.00652	4/2/4/4/3/2/4/4
2	1	1	30.86837	31.00467	3/4/4/4/3/4/3/3
2	1	1	30.59349	25.82433	2/2/3/4/4/4/4/1
2	1	1	27.25200	25.49833	4/1/4/4/4/4/4/1
2	1	1	25.38917	27.22410	3/2/3/4/1/4/2/4

The scatter plot of figure D.2 (appendix D) compares the sample means of the responses of the treated and control units within the 389 useable strata. Careful examination of the plot indicates that there are more points above the 45^0 line than below it, pointing to a treatment effect. And indeed, the scatter plots of figure D.4 indicate that the stratification has failed to remove enough bias. These plots represent pairs of the empirical versions of the conditional mean in (3.6) for $t = 0, 1$ computed per stratum and, despite a superficial agreement with the expectations, exhibit plenty of local structure around the 45^0 line. And, indeed, a conditional test based on this first stratification provides strong evidence for a treatment effect. Before presenting this result let us describe the conditional test in general.

Suppose that the stratification is based on splitting the range of each covariate into i intervals and consider the statistic

$$S(i) := \frac{1}{M(i)} \sum_{k=1}^{K(i)} S_k(i),$$

where $K(i)$ stands for the number of useable strata, $M(i)$ for the total number of treated units, and $S_k(i)$ for the sum of the responses of the treated units in stratum k. This is a version of the Mantel-Haenszel statistics of examples 3.9 and 3.10; the dependence on i emphasizes the fact that the strata involved in its computation change with the number of intervals used in the stratification. For each i the null distribution of $S(i)$ can be estimated as explained in section 3: the treatments are pseudo-randomly permuted within the strata and the corresponding value of $S(i)$ is computed a large number of times yielding a simulated random sample whose empirical distribution estimates the null distribution of $S(i)$. If the simulated random sample is large enough, the proportion of times in which its elements exceed the observed value of $S(i)$ can be taken as an estimate of the p-value of the test. For example, the histogram in figure D.3 provides an approximation, based on 10,000 simulations, to the null distribution of $S(i)$ when $i = 5$; an estimate of the p-value is obtained by integrating the histogram from the observed value of $S(i)$—indicated by the vertical dashed line—onwards.

In the stratification based on $i = 4$ the observed value of $S(i)$ is 26.36; the estimate of the p-value corresponding to it, based on 10,000 simulations, is 0.0142 (95% Wilson confidence interval of [0.0121, 0.0167]). This spurious evidence for a treatment effect is a consequence of the inappropriately coarse stratification and illustrates the need for checking and critically examining different stratifications.

TABLE 2
Numbers of strata and p-values obtained with stratifications of five data sets based on several values of i and on all eight covariates

i	Seed 2013		Seed 2014		Seed 2015		Seed 2016		Seed 2017	
	Strata	P-value	Strata	P-value	Strata	P-value	Strata	P-value	Strata	P-value
4	389	0.0142	355	0.7727	360	0.9200	362	0.0815	387	0.1587
5	98	0.0395	102	0.7435	94	0.6233	122	0.3225	103	0.6965
6	22	0.7100	25	0.8256	22	0.4339	39	0.0956	27	0.5241
7	6	0.0633	5	0.7542	12	0.2232	17	0.1751	10	0.9507

The numbers of strata and p-values resulting from stratifications based on $i = 4, 5, 6, 7$ are given in table 2 under the heading 'Seed 2013', which refers to the seed of the pseudo-random number generator used to simulate the sample. Although it appears from the scatter plots of figure D.5 that $i \geq 5$ yields a sufficient reduction of bias, it is only with $i = 6$ that a large p-value is obtained. Ideally, one should go on with the calculations for larger i, but with $i = 7$ we only get 6 useable strata and it is impossible to go farther than that with this data set. Unfortunately, the p-value estimated with $i = 7$ might leave one with doubts; however, we know that the occurrence of this small p-value must be attributed to chance, perhaps doing justice to the seed used to generate the first data set...

In order to get some idea about how the results vary with the choice of i we repeat the stratification and the testing procedures with different data sets of size $N = 10,000$. Table 2 also shows the numbers of strata and the p-values obtained with other four data sets, headed by the corresponding seeds of the pseudo-random number generator, and stratifications based on $i = 4, 5, 6, 7$.

The conditions under which the null hypothesis is tested—the strata, the sample sizes within them, the distributions of the responses within the strata—depend on i to some extent, so the p-values obtained with different stratifications do not necessarily have to agree. However, for an application of the methods of section 3 to be successful the stratification must exhibit a *good balance* between the covariates in the different treatment groups—witnessed by scatter plots such as those of figures D.5 and D.6—and yield consistently 'large' or 'small' p-values over a few 'large' values of i; for only then can one have some confidence that the p-values are practically unbiased. The results based on the second data set (seed 2014) seem quite clear, all p-values being unanimous in providing no evidence for a treatment effect, and the same can be said of the results based on the third and on the last data set (seeds 2015 and 2017). The results obtained with the fourth data set could raise some doubts: the relatively small p-value at $i = 6$ might suggest some evidence for a treatment effect, while the small sample size could be blamed for the larger p-value at $i = 7$.

Things can become much clearer if the sample size is sufficiently large relative to the number of covariates. To illustrate this we perform the procedure of stratification and testing on the five data sets based only on the six variables that influence treatment assignment—recall that the last two variables, `alcohol` and `obesity`, do not influence treatment assignment and hence need not be considered in the stratification (even though that could have not been surmised

TABLE 3
Numbers of strata and p-values obtained with stratifications of five data sets based on several values of i and on the six covariates that influence treatment assignment

i	Seed 2013		Seed 2014		Seed 2015		Seed 2016		Seed 2017	
	Strata	P-value	Strata	P-value	Strata	P-value	Strata	P-value	Strata	P-value
7	195	0.7090	196	0.9864	210	0.8616	206	0.6092	213	0.8211
8	113	0.9591	103	0.2537	98	0.9648	91	0.2277	126	0.9998
9	63	0.4566	52	0.1236	49	0.9531	59	0.7717	61	0.8998
10	25	0.8344	25	0.5030	32	0.9060	36	0.1426	34	0.9979

from the box plots of figure D.1, it is quite clear from the prediction analyses and hence has some justification). Table 3 shows the numbers of strata and the p-values obtained with stratifications based on $i = 7, 8, 9, 10$. The p-values are now unanimous in providing no evidence for a treatment effect and, with the support of scatter plots such as those of figure D.7 (appendix D), which suggest the correctness of stratifications based on $i = 7$, leave little room for doubts.

To give an idea of the numbers of units that are actually used in testing based on stratification, let us mention that the 195 useable strata obtained from the first data set when $i = 7$ and when only the six genuine confounders are considered contain a total of 435 units (218 treated and 217 controls); when $i = 6$, which (though not shown in table 3) also seems to yield a reasonably good balance and a correct conclusion, the number of strata is 401 and the number of units 1006 (487 treated and 519 controls). Thus, by stratification, the data actually used in testing can easily be as few as 10% of the whole sample.

As a last exercise of this subsection we consider stratifications that ignore the covariate with the greatest influence on treatment assignment, `age.at.onset`. Interestingly, with the present choice of parameters the method of stratification is quite robust to the omission of the principal confounder, often providing no evidence for a treatment effect, at least for the sample size of $N = 10,000$. However, if p_0 and p_1 are interchanged then the omission of `age.at.onset` usually leads to the wrong conclusion. For example, if we simulate a new data set of size $N = 10,000$ (with seed 2018) with $p_0 = 0.25$ and perform stratifications based on $i = 10, 11, \ldots, 14$ we get 175, 123, 83, 58 and 30 useable strata and corresponding p-values of 0.257, 0.029, 0.054, 0.009 and 0.064, which provide evidence for a treatment effect; the scatter plots of figures D.8–D.11 (appendix D) indicate no lack of balance in the distribution of the covariates, though they do indicate the existence of a treatment effect (top-left plots).

6.2. Stratification in a situation where treatment effect exists

We take $p_0 = 0.75$ and $p_1 = 0.25$ as in most of the previous subsection, thus getting the same value of $\boldsymbol{\mu}$, but now set

$$\boldsymbol{\beta}_0 = (3, 7, 8, 4, 2, 8, 2, 10)/10, \quad \boldsymbol{\beta}_1 = 1.2\,\boldsymbol{\beta}_0.$$

The treatment effect resulting from these choices is $\boldsymbol{\epsilon} = 4.66$, whereas the mean difference between the mean responses in the treated and control groups is $\Delta =$

7.87; thus in this case the naive approach only runs the risk of overestimating the evidence for a treatment effect.

As in the preceding subsection we take $\mu = 0$ and $\sigma = 1$ and the first seed used there (2013) to simulate a random sample of $N = 10,000$ vectors. For convenience we leave out `obesity` and `alcohol` (which have no influence on treatment assignment). If we compute the difference between average responses in the treated and control groups we get $\Delta = 7.72$ as an estimate of Δ. The corresponding 95% confidence interval, $[7.50, 7.94]$, is far to the right of the true treatment effect.

Guided by the results of the preceding subsection we test for a treatment effect on the basis of stratifications obtained with $\imath = 7, 8, 9, 10$; the scatter plots in appendix E indicate that these stratifications are appropriate. The resulting numbers of useable strata are 195, 113, 63 and 25 (cf. first column of results in table 3), and the p-values are practically equal to 0 (e.g. Wilson 95% confidence interval of [0, 0.00038]), a clear indication of the existence of a treatment effect.

Because $\Delta > \boldsymbol{\epsilon}$, it is clear that the evidence for a treatment effect can only become stronger if an important confounder such as `age.at.onset` is not considered in the stratification.

6.3. Stratification on the propensity score in a situation where no treatment effect exists

We consider the same situation as in subsection 6.1, but instead of stratifying the sample in terms of all the covariates we stratify it according to the estimated propensity score. Actually, in order to illustrate what one may expect in favourable circumstances, we begin by stratifying on the *true* propensity scores, namely on $f_{T|\mathbf{U}=\mathbf{u}}(t)$ with $t = 1$ and $\mathbf{u} = \mathbf{U}_i$ (cf. the beginning of section 5):

$$\lambda(\mathbf{X}_i) := \frac{p_1|\Sigma_1|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{U}_i-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{U}_i-\boldsymbol{\mu}_1)}}{p_0|\Sigma_0|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{U}_i-\boldsymbol{\mu}_0)^{\mathrm{T}}\Sigma_0^{-1}(\mathbf{U}_i-\boldsymbol{\mu}_0)} + p_1|\Sigma_1|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{U}_i-\boldsymbol{\mu}_1)^{\mathrm{T}}\Sigma_1^{-1}(\mathbf{U}_i-\boldsymbol{\mu}_1)}},$$

where $\mathbf{X}_i = (\mathbf{U}_i, \mathbf{V}_i)$.

Split the interval $[0, 1]$ (the theoretical range of $\lambda(\mathbf{X}_i)$) into $1/\ell$ subintervals of length ℓ. If $\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_{1/\ell}$ denote these intervals (ordered from left to right on the real line) then the k-th stratum is defined as the set of units i for which $\lambda(\mathbf{X}_i) \in \mathcal{I}_k$. As we have seen, the testing and estimation procedures of sections 3 and 4 should be approximately valid (provided ℓ is neither too large nor too small) when applied to the resulting strata. Alternatively, one may split $[0, 1]$ into intervals determined by the quantiles of probabilities ℓ, 2ℓ, 3ℓ, etc., as is often done in the literature (e.g. [29]). Both methods lead to similar results in our case, but we find the first method somewhat more natural: a difference between stratifying on the covariates and stratifying on the propensity score is that the latter typically uses most of the data, and it seems somewhat unnatural to force all strata to contain approximately the same number of observations (as they will if quantiles are used) even though the range of the propensity score is not uniformly populated.

TABLE 4

Numbers of useable strata and total numbers of observations in them, p-values of the two tests for treatment effect, and estimates and 95% confidence intervals (CI) for ϵ obtained by stratifying the data on the true propensity score using three values of ℓ

ℓ	No. strata	No. obs.	P-value 1	P-value 2	Estimate of ϵ (95% CI)
0.010	94	9873	0.6289	0.3208 (75)	-0.127 (-0.378, 0.124)
0.025	39	9941	0.6574	0.9190 (35)	-0.022 (-0.442, 0.399)
0.050	20	10,000	0.7250	0.4116 (18)	0.127 (-0.176, 0.431)

Table 4 shows the results of testing for and estimating the treatment effect obtained by stratifying the first sample used in subsection 6.1 (indicated by 'Seed 2013' in table 2) with three values of ℓ: $0.01, 0.025, 0.05$. The number of useable strata (second column of the table) ranges from 94 to 20, and the total number of useable observations (third column) is always close to $N = 10,000$. These numbers already suggest that stratification based on the propensity score may represent an enormous gain in efficiency. In particular, the fact that only a few units are discarded suggests that the estimation method of section 4 can be used to the full to estimate the overall treatment effect. And, indeed, the estimates of ϵ (with 95% confidence intervals) and the p-values of the tests of $\epsilon = 0$ against $\epsilon \neq 0$ (which consist of rejecting the null if and only if the interval does not contain 0), shown in the last two columns of table 4, agree with the expectations for all ℓ. These estimates and p-values are based on strata containing more than five controls and more than five treated units, the number of such strata being indicated between parentheses after the p-value in the penultimate column; still, most units are used, and most strata possess bigger numbers of treated and control units. The other p-values in table 4, under the heading of 'P-value 1', are the p-values of the Mantel-Haenszel test of subsection 6.1; they are consistent with the p-values based on the tests of $\epsilon = 0$ for all values of ℓ.

As before, the correctness of results such as these has to be judged on the basis of how balanced the distributions of treated and control units are within the strata. Because much more data are involved in the present analysis than in the analyses of subsections 6.1 and 6.2, it is convenient to check for balance by testing the equality of the distributions of the covariates in the control and treated groups per stratum, obtaining a p-value per covariate and stratum, and then checking the uniformity of the p-values per covariate. In this procedure the tests are not to be interpreted formally—if only because the covariates generally have slightly different distributions in the two groups (due to stratification on continuous variables)—but rather as indicators of how good the balance between control and treated units is. If the balance is good, then the p-values of the many two-sample tests comparing the two groups will be *approximately* uniformly distributed.

The probability plots in appendix F serve to assess the uniformity of the p-values obtained from Anderson-Darling tests (`R` package `kSamples` of Scholz and Zhu [34]) comparing the distribution of the covariates in the treated and control groups. Figures F.1–F.3 in appendix F, based on the stratifications with $\ell = 0.01, 0.025, 0.05$, indicate an overall agreement with the expected uniform

pattern (except perhaps regarding `obesity`), and thus support the results of table 4. The p-values shown above the probability plots come from Kolmogorov-Smirnov tests of uniformity carried out per covariate; they should be interpreted as *relative* measures of 'uniformity' since in general (certainly when the propensity scores are estimated from the data) the p-values whose uniformity is being tested are somewhat dependent and we know that they are *not* exactly uniformly distributed.

Having seen that the method of stratification on the propensity score must work in favourable situations—when the propensity score is known exactly or very approximately—let us turn to the more realistic situation where the propensity score function is unknown (although, as always, the possible confounders are known and hence the study is free of hidden bias) and therefore must be estimated. If it is known that the propensity score is well approximated by a function depending only on a small number of parameters then these parameters can in principle be estimated from the data and should lead to a good balance in the distribution of the covariates in the treated and control groups and hence to the correct conclusions. Thus, if we estimate the p_t, $\boldsymbol{\mu}_t$ and Σ_t from the simulated data and use them to estimate $\lambda(\mathbf{X}_i)$ then the results (not shown here, but see the next subsection) are excellent and entirely consistent with those obtained with the true propensity scores, which is not surprising since as shown in the first plot of figure F.4 the propensity score estimates obtained in this way are rather close to the true propensity scores.

It is doubtful, however, that the propensity score function will always be well approximated by a simple function depending on a few parameters. For example, if $\Sigma_0 \neq \Sigma_1$ then the use of a logistic regression model—which, as pointed out by Hade and Lu [11], is very often adopted as the model for the propensity score—in place of our $\lambda(\mathbf{X}_i)$ may (and will with the present values of the parameters) lead to glaring errors. Still, there appears to be a belief in the literature (cf. the articles cited in [11]) that the logistic regression model should provide a sufficiently accurate approximation to the true propensity score function as long as all the relevant covariates and their powers, the interactions between them, and the interactions between the treatment indicator and the covariates and their powers, are included in it. Even if this were true (we think it is not), it does not seem obvious that 'overfitting' the propensity score function by fitting a logistic model with many parameters to the data will always lead to the correct results.

Except in situations where a relatively simple parametric model manifestly leads to good balance between the distributions of the covariates in the treated and control groups, one should at least try to estimate the propensity score function by a non-parametric, consistent estimator. For the present analyses we have considered two types of estimators: random forest classifiers as implemented by Liaw and Wiener [18] in the `R` package `randomForest`, and non-parametric regression estimators as implemented by Hayfield and Racine [13] in the `R` package `np`. While non-parametric regression estimators are known to be consistent under very general conditions (including those of our simulation), it is not known under what conditions random forests are consistent. However, random forests

TABLE 5

Numbers of useable strata and total numbers of observations in them, p-values of the two tests for treatment effect, and estimates and 95% confidence intervals (CI) for ϵ obtained by stratifying the data on the estimated propensity score using different values of ℓ

ℓ	No. strata	No. obs.	P-value 1	P-value 2	Estimate of ϵ (95% CI)
0.010	87	9961	0.7192	0.7572 (72)	0.036 (-0.195, 0.268)
0.025	36	9974	0.6816	0.6133 (32)	0.069 (-0.197, 0.334)
0.050	19	9998	0.5615	0.1207 (17)	0.230 (-0.060, 0.520)

often perform better (in terms of classification error, sensitivity and specificity) than most classifiers, and a simpler variant of random forests has been shown to be universally consistent in [5], which suggests that random forests are consistent in many situations. In our case the random forest does perform better than the non-parametric regression estimator, whose estimates of the propensity score are rather inaccurate; we shall only describe results based on random forests.

The optimal values of error rate (probability of an incorrect classification), sensitivity and specificity, namely those of the Bayes rule associated with the model for $(T, \mathbf{U})$ (the rule that classifies a new unit with covariates $\mathbf{U} = \mathbf{u}$ as treated if and only if $f_{T|\mathbf{U}=\mathbf{u}}(1) > 1/2$) can be estimated by simulation as 0.188, 0.427 and 0.939, respectively. The corresponding parameters of the random forest classifier are estimated as 0.200, 0.372 and 0.942, so in terms of classification errors the performance of the random forest is not far from being optimal. Fortunately, the good performance in terms of prediction accuracy also translates into reasonably balanced strata.

Indeed, figures F.5–F.7 (based on stratifications with $\ell = 0.01, 0.025, 0.05$, as above) exhibit a rather good balance between treated and control units regarding all the important covariates. Accordingly, all the results of testing for and estimating the treatment effect, shown in table 5, conform to the expectations, except for the fact that the confidence intervals for ϵ are somewhat narrower than they should be, as follows by comparing them with those of table 4 and as we had anticipated in section 4.

Despite these good results, the experience we have gained from various simulations with the present model suggests that non-parametric propensity score estimates are sometimes so inaccurate that no satisfactory balance between treated and controls is achieved and the tests and estimates lead to wrong conclusions. That such estimates are always relatively inaccurate if compared to parametric ones—even when they lead to good balance and correct conclusions—is seen by figure F.4. Also, the settings of the random forest (or those of any other predictor) that yield the best propensity score estimates, and hence the best balance of the covariates, are somewhat dependent on the particular data set.[16] It may therefore be necessary to produce estimates of propensity scores under different settings and assess their quality in terms of the balance they provide.

[16] In the present simulation we have set the parameter `ntree` equal to 20,000 and used the default value of `mtry` in the `randomForest` implementation of [18].

Remark. In a simulation study, Lee, Lessler and Stuart [16] concluded that the use of random forests and other non-parametric predictors to estimate propensity scores provides estimates of the treatment effect that are almost unbiased. However, the simulation scenarios considered by these authors were based on generating covariates from a normal distribution and then generating the treatment conditionally on the covariates. In our setting the confounders are generated from a mixture of normal distributions and the method of estimation (based on the weighted estimators mentioned in the remark preceding subsection 4.1) used in [16] is often inapplicable (because many propensity score estimates are 0 or 1) and when it is applicable it yields estimates with very large variances. The simulation results of [11] indicate that the results of [16] are somewhat optimistic and can be explained in terms of the overlap between the distributions of the confounders in the treated and control groups. □

6.4. Stratification on the propensity score in a situation where treatment effect exists

We take up the situation of subsection 6.2 and the sample of 10,000 units used there. Thus we have an overall treatment effect of $\boldsymbol{\epsilon} = 4.66$; the difference between average responses in the treated and control groups, $\hat{\Delta} = 7.72$, is close to the theoretical $\Delta = 7.87$; and the corresponding 95% confidence interval for Δ, $[7.50, 7.94]$, lies well to the right of the treatment effect.

The plan of analysis is the same as in the preceding subsection, but the presentation of the results will be shortened, because many of the outcomes—in particular the figures in appendix F—are unaffected by the introduction of the treatment effect. Table 6 shows the results based on stratifications on the true propensity score, on the random forest estimate of it, and on the correct parametric estimate of it (that is, on $\lambda(\mathbf{X}_i)$ with the parameters $p_0, p_1, \boldsymbol{\mu}_0, \boldsymbol{\mu}_1$, etc., replaced by estimates computed from the sample), with $\ell = 0.025$. All three stratifications yield overwhelming evidence for a treatment effect, good estimates and plausible confidence intervals for $\boldsymbol{\epsilon}$. As might be expected, the less accurate estimate of $\boldsymbol{\epsilon}$ is the one obtained with the random forest. And, again, the interval obtained with the random forest estimates of $\lambda(\mathbf{X}_i)$ is unrealistically narrow, a reflection of the fact that our estimate of $\mathrm{SE}(\hat{\boldsymbol{\epsilon}})$ ignores the variability in the estimation of the propensity score function and in the subsequent stratification on it. Figure G.1 in appendix G indicates an overall good balance in the distribution of the covariates in the treated and control groups when $\lambda(\mathbf{X}_i)$ is estimated parametrically and stratified with $\ell = 0.025$; and it is plausible that the poorer balance in terms of `adiposity` and `type.A.behaviour`, which are among the weaker determinants of treatment, hardly contributes any bias to the corresponding estimate of $\boldsymbol{\epsilon}$.

In order to illustrate the consequences of overlooking important confounders let us see what happens if we estimate the propensity scores by a random forest without taking account of `LDL.cholesterol`, the *second* most important confounder. Figure G.2 reveals an extreme lack of balance with respect to the

TABLE 6

Numbers of useable strata and total numbers of observations in them, p-values of the two tests for treatment effect, and estimates and 95% confidence intervals (CI) for ϵ obtained by stratifying the data on the true propensity score, on the propensity score estimated by a random forest, and on the propensity score with its parameters estimated parametrically (i.e. based on the correct model), using $\ell = 0.025$

Propensity score	No. strata	No. obs.	P-value 1	P-value 2	Estimate of ϵ (95% CI)
True	39	9941	0.0000	0.0000 (35)	4.635 (4.145, 5.125)
Random forest	36	9974	0.0000	0.0000 (32)	4.759 (4.452, 5.066)
Parametrically	39	9949	0.0000	0.0000 (35)	4.564 (3.992, 5.137)

omitted confounder. This is as expected and not particularly significant because in an actual analysis where the data set misses an important confounder one cannot examine the balance with respect to it, but it does show how the comparison of treated and control units within strata can be biased. More significant is the concomitant lack of balance with respect to the most important confounder, `age.at.onset` (`systolic.blood.pressure`, in terms of which balance is just as bad, is less important as a confounder): in an actual application, such lack of balance would warn us not to trust the corresponding estimates of treatment effect.

With stratifications based on $\ell = 0.010, 0.025, 0.050$, these estimates are 5.283 (CI of [5.035, 5.531]), 5.269 (CI of [5.004, 5.534]), and 5.424 (CI of [5.152, 5.696]), respectively, all of them suggesting an effect greater than the actual $\epsilon = 4.66$. (The number of strata and the number of observations that result from these stratifications are very similar to those of table 5.) Interestingly, the bias is present in nearly every stratum, as seen by figure G.3 where the strata are ranked according to the propensity score estimate. This must correspond with the fact that in the plot of `age.at.onset` of figure G.2 only a few points approach the straight line.

The bias in the estimates must of course be attributed to the inaccuracy of the propensity score estimates; indeed, figure G.4 illustrates what the omission of the second most important confounder does to those estimates (compare it with figure F.4). In an actual application it may be difficult to determine whether the lack of balance observed with respect to important covariates is due to the omission of a confounder or to the inaccuracy of the predictor of treatment status (which may itself be the result of small sample size or of an incorrect choice of prediction model). In this connection one may wonder whether it is sometimes possible to fit a parametric model to a data set not containing an important confounder in such a way as to *force balance* with respect to the confounders considered; does the omitted confounder then become automatically balanced, guaranteeing the correctness of the treatment effect estimate?

6.5. Matching on the Mahalanobis distance

A more involved but sometimes better method than stratification is *matching on a distance between units*, such as the Mahalanobis distance between their

covariates—often applicable (possibly after transformation of variables) when the covariates are numeric, and especially when they represent continuous measurements. Chapter 10 of [26] gives a good introduction to matching methods, and the `R` package `optmatch` of Hansen and Klopfer [12] provides a user-friendly implementation of *full matching*, a method of creating matched sets that is optimal in a certain sense. Although the role of matching in our analyses is completely analogous to that of stratification, matching is really more complicated than stratification; in the following subsection we describe the ideas behind it very briefly, with just enough detail to make the illustration fully intelligible—readers who are interested mainly in the numerical results may prefer to go straight to subsection 6.5.2.

6.5.1. Matching

In essence, matching consists of computing distances $D(\mathbf{X}_i, \mathbf{X}_j)$ between the covariates of treated units i and controls j and forming *matched sets* of treated and control units at a small distance of each other. Thus, given a distance function D, matching requires a way of quantifying the closeness between the treated units and the control units in a set. If a generic matched set—or potential matched set—is represented by $M = (I, J)$, where I is a set of indices of treated units and J a set of indices of control units, the *average distance of* M,

$$D_M = \frac{1}{\#I \times \#J} \sum_{i \in I, j \in J} D(\mathbf{X}_i, \mathbf{X}_j),$$

provides a measure of how close or well-matched the elements of M are. In realistic situations where treated and controls differ systematically with respect to the confounders, D_M will tend to increase with the size of M, and hence large matched sets will tend to have bigger average distances than small matched sets.

Since good matched sets can be characterized by small average distances, one might think that the forming of matched sets is best achieved by matching first the treated and control units that are closest, then matching the treated and control units that are closest among the remaining units, and so on. But this process, an example of 'greedy matching', normally does not lead to the best results. For example, if (i, j) is the pair of treated and control units at the smallest distance, say $D(\mathbf{X}_i, \mathbf{X}_j) = d$, and (i', j') are such that

$$D(\mathbf{X}_{i'}, \mathbf{X}_{j'}) = d + 2\varepsilon \quad \text{and} \quad D(\mathbf{X}_i, \mathbf{X}_{j'}) = d + \frac{\varepsilon}{2} = D(\mathbf{X}_{i'}, \mathbf{X}_j),$$

then the pairing of i with j and of i' with j' will yield average distances of d and $d + 2\varepsilon$, and hence an *average* average distance of $d + \varepsilon$, while the pairings of i with j' and of i' with j will both yield an average distance of $d + \varepsilon/2$ (bigger than the minimum distance d) but a smaller average average distance: $d + \varepsilon/2 < d + \varepsilon$. In other words, and more generally formulated, choosing the early matches to be as close as possible often entails poorer matches later on, leading to matches of variable quality and of lower average quality as well.

For this reason, it is better to determine a *collection* of matched sets in terms of a global measure of distance than to determine matched sets by minimizing their average distances on a sequential basis. By a collection of matched sets we mean a set $\mathcal{M} = \{M_1, M_2, M_3, \ldots\}$ of elements $M_i = (I_i, J_i)$, I_i being a set of indices of treated units and J_i a set of indices of control units, such that $I_i \cap I_j = \emptyset$ and $J_i \cap J_j = \emptyset$ whenever $i \neq j$. The average of the average distances of such a collection $\mathcal{M}$, called simply the *distance of* $\mathcal{M}$, is defined by

$$\mathbf{D}(\mathcal{M}) = \sum_{M \in \mathcal{M}} w_M D_M,$$

where the w_Ms are positive *weights* ($\sum_M w_M = 1$), and is an overall measure of the quality of a set of matched sets which, when minimized with respect to $\mathcal{M}$, leads to more balanced and overall better sets of matched sets. Interestingly, the choice of the weights has little influence on the *structure* of the matched sets formed: if the w_Ms are *neutral* in the sense that $w_{(I,J)} = w_{(I\setminus\{i\},J\setminus\{j\})} + w_{(\{i\},\{j\})}$ (where $i \in I$, $j \in J$) then the optimal collection of matched sets is a *full matching*, that is, a collection in which each matched set consists either of one control and at least one treated unit or of one treated unit and at least one control (proposition 30, p. 310, of [26]).[17]

The determination of the optimal full matching by minimization of $\mathbf{D}(\mathcal{M})$ over a specified class of collections $\mathcal{M}$ is formally equivalent to a type of well-studied optimization problems (minimization of cost flows in networks) for which convergent and efficient algorithms exist and have been implemented in statistical packages (for references and further information on the algorithms and on packages implementing them see chapter 10 of [26]). Moreover, in these algorithms the class of collections $\mathcal{M}$ over which $\mathbf{D}(\mathcal{M})$ is to be minimized can be specified indirectly by requiring that the I_i and J_i making up the matched sets $M_i = (I_i, J_i)$ be contained in fixed subsets $\mathcal{I}$ and $\mathcal{J}$ of indices of treated and control units. In this way, by considering different choices of $\mathcal{I}$ and $\mathcal{J}$, one has the possibility of discarding 'undesirable units' (such as a treated unit that is too distant from all controls) from the minimization process.

6.5.2. Illustration

For our illustration we make use of the function `fullmatch` of the `R` package `optmatch` [12] with the *matrix of Mahalanobis distances* as input in order to get a full matching of the sample of subsection 6.2. With the rows representing labels of treated units and the columns representing labels of control units, and writing as usual $\mathbf{X}_i = (\mathbf{U}_i, \mathbf{V}_i)$, we define the (i, j)-entry of the matrix of Mahalanobis distances by

$$D(\mathbf{X}_i, \mathbf{X}_j) = \sqrt{(\mathbf{U}_i - \mathbf{U}_j)^{\mathrm{T}} \hat{\Sigma}^{\mathrm{T}} (\mathbf{U}_i - \mathbf{U}_j)},$$

[17] For example, weights $w_{(I,J)}$ proportional to $\#I + \#J$ are neutral.

where $\hat{\Sigma} = (\hat{\Sigma}_0 + \hat{\Sigma}_1)/2$ and $\hat{\Sigma}_0$ and $\hat{\Sigma}_1$ are the sample covariance matrices of the $\mathbf{U}_i$ vectors of the control and treated units, so that only the genuine confounders are used in the matching. The matching problem associated with this 2500×7500 matrix is actually too large to be handled by the `fullmatch` function, and we need to set most of its entries to ∞ in order to preclude the inclusion in the algorithm of matches which most probably should not be part of an optimal solution anyway—i.e. matches with large average distances. Thus, we redefine the (i, j)-entry of the matrix as ∞ unless

$$D(\mathbf{X}_i, \mathbf{X}_j) \leq C \cdot D(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\mu}}_1),$$

where C is a positive constant and $\hat{\boldsymbol{\mu}}_0$, $\hat{\boldsymbol{\mu}}_1$ are the sample means of the $\mathbf{U}_i$ vectors of the control and treated units, and then enter the redefined matrix as input in `fullmatch`. Different values of C will typically lead to different optimal full matchings; the smaller C, the smaller the average distances involved and the smaller the number of matched units tends to be.

The matched sets obtained by full matching look very much like the strata obtained in subsections 6.1 and 6.2, except that *all* matched sets have either at most one treated unit or at most one control unit; the following output summarizes 10 strata obtained with a full matching based on $C = 0.5$:

sample.size	no.treated	no.controls	mean.of.treated	mean.of.control	stratum
3	1	2	29.82007	25.18421	1.131
2	1	1	14.49381	16.22594	1.346
5	1	4	28.29192	24.02718	1.350
4	1	3	31.31730	28.16151	1.101
3	1	2	27.86204	21.63332	1.145
5	1	4	23.79148	19.80277	1.174
4	1	3	30.71736	25.29307	1.195
2	1	1	34.04405	30.86613	1.305
2	1	1	36.64107	23.65283	1.525
2	1	1	36.76584	31.43874	1.287

The number of matched sets is 526 and the total number of units contained in them is 1269, so the full matching based on $C = 0.5$ uses somewhat more data than do the stratifications of subsections 6.1 and 6.2 based on the same set of confounders and $i = 6, 7$.

The quality of the full matching can be appreciated as in stratification: the scatter plots in appendix H indicate good balance with respect to the six confounders when $C = 0.5$; a critical look at figure H.2, however, will show that much bias remains in many of the strata if $C = 1$. Accordingly, in the situation where there is no treatment effect the p-value of the Mantel-Haenszel test of subsection 6.1 is 0.4308 if $C = 0.5$, but practically 0 if $C = 1$, while when there is a treatment effect the p-value is practically 0 with both choices of C.

6.6. Estimation of the overall treatment effect by predicting 'counterfactuals'

As an illustration of the method described in subsection 4.1 we consider using the $\hat{\boldsymbol{\epsilon}}_N$ of (4.3) to estimate the treatment effect in the situation of subsection 6.2 and based on the sample used in subsections 6.2–6.5. Taking $\Pi(x, t)$ as a

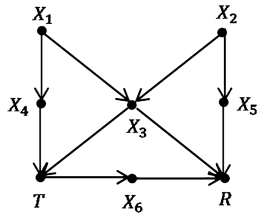
random forest we get $\hat{\boldsymbol{\epsilon}}_N = 4.68$. This is remarkably close to $\boldsymbol{\epsilon} = 4.66$, but is actually a fortunate result because other simulations indicate that the average of $\hat{\boldsymbol{\epsilon}}_N$ is closer to 4.8 or 4.9, so that $\boldsymbol{\epsilon}_N$ suffers from a bias of about 0.2. If we split the sample into five subsets and compute (4.3) with each, we get treatment effect estimates of 4.61, 4.56, 4.42, 4.39 and 4.90. From these we get 0.102 as an estimate of the standard error of $\hat{\boldsymbol{\epsilon}}_{2000}$, which leads to the nominal (conservative) 95% confidence interval of $[4.47, 4.88]$ for $\boldsymbol{\epsilon}$. This is narrower than the intervals presented in table 6.5; however, the putative smaller variance of $\hat{\boldsymbol{\epsilon}}_N$ is probably offset by a somewhat bigger bias. At any rate, in this case the method based on predicting counterfactuals seems to provide clear evidence for a treatment effect, and its estimate of $\boldsymbol{\epsilon}$ is consistent with the one obtained by stratifying on the estimated propensity score.

Acknowledgements

I should like to thank my colleague Albert Wong for interesting conversations on the subjects treated here. An associated editor suggested the need for some sort of assumption on the joint distribution of variables from different units other than assumption **A1**, prompting the formulation of **A0**; I am indebted to the editor for this and for other suggestions that helped improve the manuscript.

Appendix A: Selection of covariates from a causal diagram

Pearl ([21], p. 114) uses the following example to illustrate the selection of sets of confounders based on causal diagrams:



This scheme is really a condensed form of writing a rather particular model, namely

$$X_1 = \varphi_1(U_1), \quad X_2 = \varphi_2(U_2), \quad X_3 = \varphi_3(X_1, X_2, U_3),$$

$$X_4 = \varphi_4(X_1, U_4), \quad X_5 = \varphi_5(X_2, U_5), \quad X_6 = \varphi_6(T, U_6),$$

$$T = \tau(U, X_3, X_4), \quad R = \varrho(V, X_3, X_5, X_6),$$

where $U_1, U_2, \ldots, U_6$, U and V are independent standard uniform random variables and $\varphi_1, \varphi_2, \ldots, \varphi_6$, τ and ϱ are given functions.

Given a subset $\mathbf{X}$ of the variables X_1, X_2,..., X_6, one can check by inspection of $\mathcal{L}(R|\mathbf{X} = \mathbf{x}, T = t)$ whether or not the assumptions of the basic model hold. For example, starting with $\mathbf{X} = X_3$, which acts both on R and on T, we have for $x \in \mathbb{R}$

$$\mathcal{L}(R|\mathbf{X} = x, T = t) =$$
$$\mathcal{L}\left(\varrho(V, X_3, \varphi_5(X_2, U_5), \varphi_6(T, U_6))\middle|X_3 = x, \tau(U, X_3, X_4) = t\right) =$$
$$\mathcal{L}\left(\varrho(V, x, \varphi_5(X_2, U_5), \varphi_6(t, U_6))\middle|\varphi_3(X_1, X_2, U_3) = x, \tau(U, x, \varphi_4(X_1, U_4)) = t\right).$$

Since the law of X_2 conditional on the equations $\varphi_3(X_1, X_2, U_3) = x$ and $\tau(x, \varphi_4(X_1, U_4), U) = t$ generally depends on t (because X_1 is involved in both), the law of R conditional on these equations also generally depends on t through $\varphi_5(X_2, U_5)$, and not only through $\varphi_6(t, U_6)$, so the basic model cannot hold. Equivalently, the basic model is not satisfied because R and T are functions of variables other than X_3 which are not independent conditionally on X_3—namely the variables X_1 and X_2 which enter into T and R through X_4 and X_5, respectively, and conditionally on $X_3 = x$ satisfy the equation $\varphi_3(X_1, X_2, U_3) = x$.

If instead we consider $\mathbf{X} = (X_3, X_4)$, then the basic model holds: writing $\mathbf{x} = (x_3, x_4)$, we have

$$\mathcal{L}(R|\mathbf{X} = \mathbf{x}, T = t) =$$
$$\mathcal{L}\left(\varrho(V, X_3, \varphi_5(X_2, U_5), \varphi_6(T, U_6))\middle|\mathbf{X} = \mathbf{x}, \tau(U, X_3, X_4) = t\right) =$$
$$\mathcal{L}\left(\varrho(V, x_3, \varphi_5(X_2, U_5), \varphi_6(t, U_6))\middle|\mathbf{X} = \mathbf{x}, \tau(U, x_3, x_4) = t\right) =$$
$$\mathcal{L}\left(\varrho(V, x_3, \varphi_5(\varphi_2(U_2), U_5), \varphi_6(t, U_6))\middle|\mathbf{X} = \mathbf{x}, \tau(U, x_3, x_4) = t\right) =$$
$$\mathcal{L}\left(\varrho(V, x_3, X_5, \varphi_6(t, U_6))\middle|(X_3, X_4) = (x_3, x_4)\right),$$

so the comparison of $\mathcal{L}(R|\mathbf{X} = \mathbf{x}, T = t)$ with $\mathcal{L}(RX = x, T = t')$ for $t \neq t'$ allows the determination of the treatment effect (which in this case happens to act through φ_6, though that is irrelevant for our purposes).

Similarly, the basic model holds with $\mathbf{X} = (X_3, X_5)$, since

$$\mathcal{L}(R|\mathbf{X}=\mathbf{x}, T=t)=\mathcal{L}(\varrho(V, X_3, X_5, \varphi_6(T, U_6))|\mathbf{X}=(x_3, x_5), \tau(U, X_3, X_4)=t) =$$
$$\mathcal{L}(\varrho(V, x_3, x_5, \varphi_6(t, U_6))|\mathbf{X} = (x_3, x_5), \tau(U, x_3, \varphi_4(\varphi_1(U_1), U_4)) = t) =$$
$$\mathcal{L}(\varrho(V, x_3, x_5, \varphi_6(t, U_6))|\mathbf{X} = (x_3, x_5)) =$$
$$\mathcal{L}(\varrho(V, x_3, x_5, \varphi_6(t, U_6))),$$

which depends on t only through the last variable of ϱ, and it also holds with $\mathbf{X} = (X_1, X_3)$:

$$\mathcal{L}(R|\mathbf{X} = \mathbf{x}, T = t) =$$
$$\mathcal{L}(\varrho(V, X_3, X_5, \varphi_6(T, U_6))|\mathbf{X} = (x_1, x_3), \tau(U, X_3, X_4) = t) =$$
$$\mathcal{L}(\varrho(V, x_3, \varphi_5(\varphi_2(U_2), U_5), \varphi_6(t, U_6))|\mathbf{X} = (x_1, x_3), \tau(U, x_3, \varphi_4(x_1, U_4)) = t) =$$
$$\mathcal{L}(\varrho(V, x_3, \varphi_5(X_2, U_5), \varphi_6(t, U_6))|X_1 = x_1, \varphi_3(x_1, X_2, U_3) = x_3) =$$
$$\mathcal{L}(\varrho(V, x_3, X_5, \varphi_6(t, U_6))|X_1 = x_1, X_3 = x_3).$$

Appendix B: Figures illustrating the South-African heart disease data set

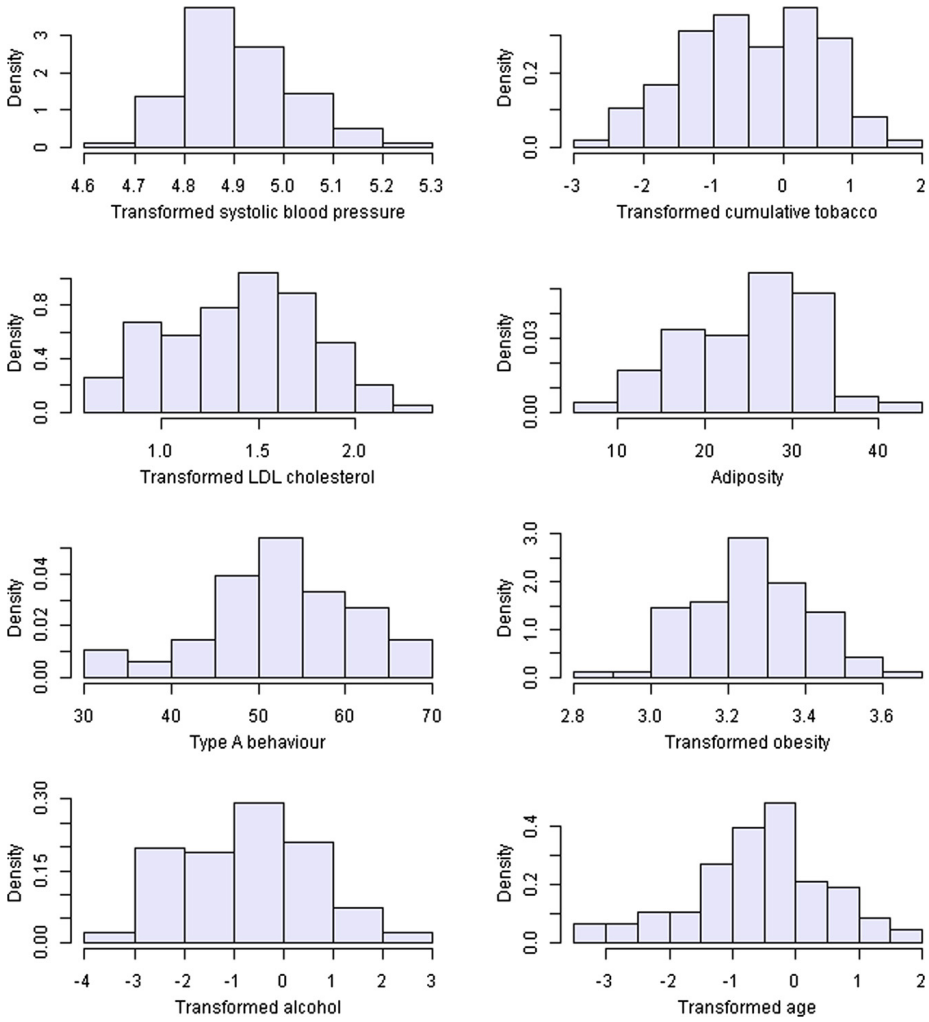


FIG B.1. *Histograms of the transformed variables of men without CHD.*

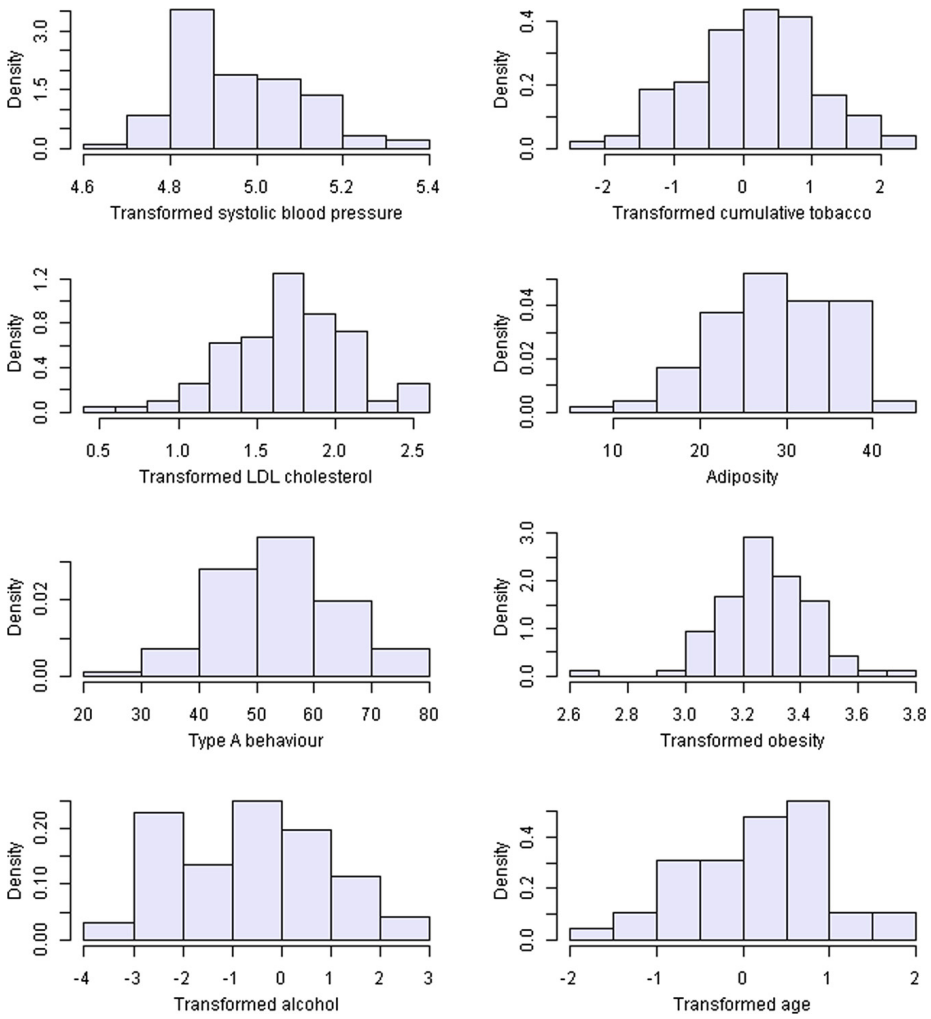


FIG B.2. *Histograms of the transformed variables of men with CHD.*

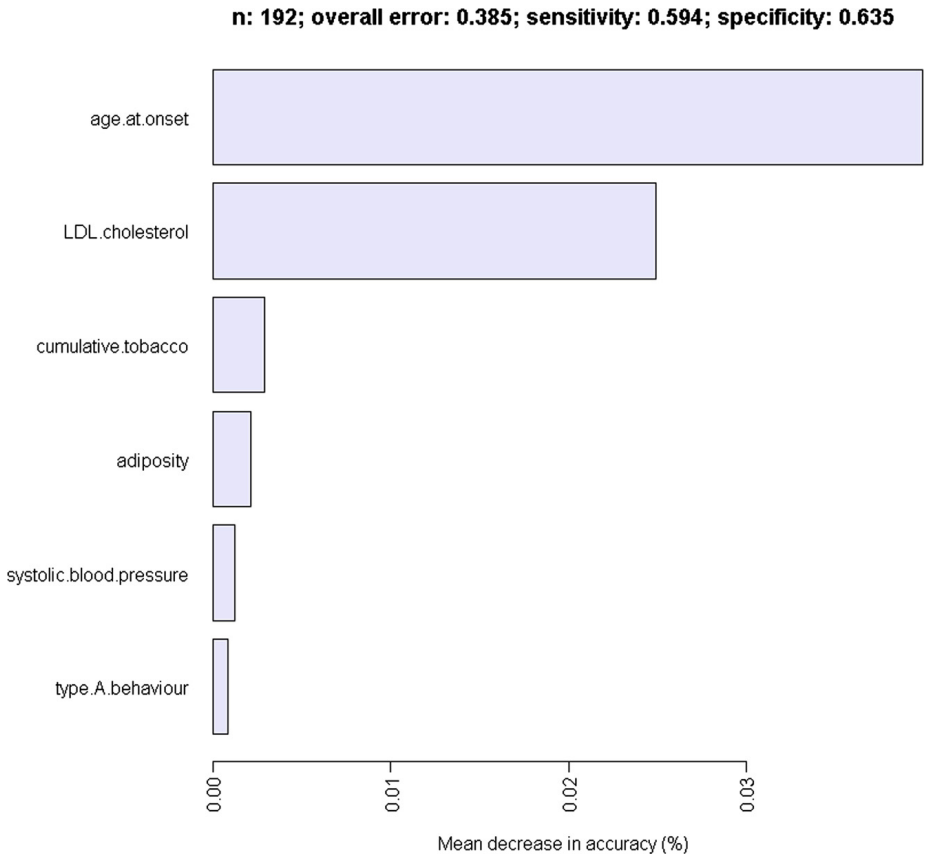


FIG B.3. *Prediction analysis based on a random forest.*

Appendix C: Figures illustrating a simulated data set

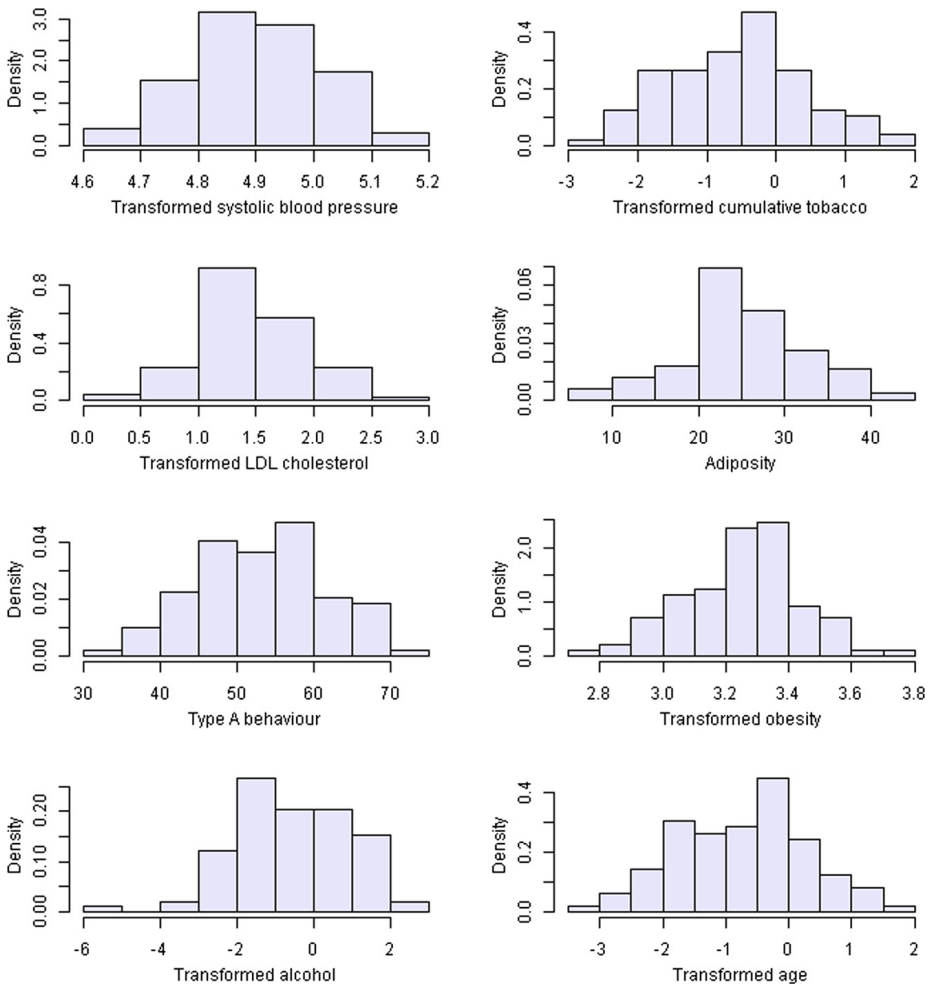


FIG C.1. *Histograms of the covariates of the sample of controls.*

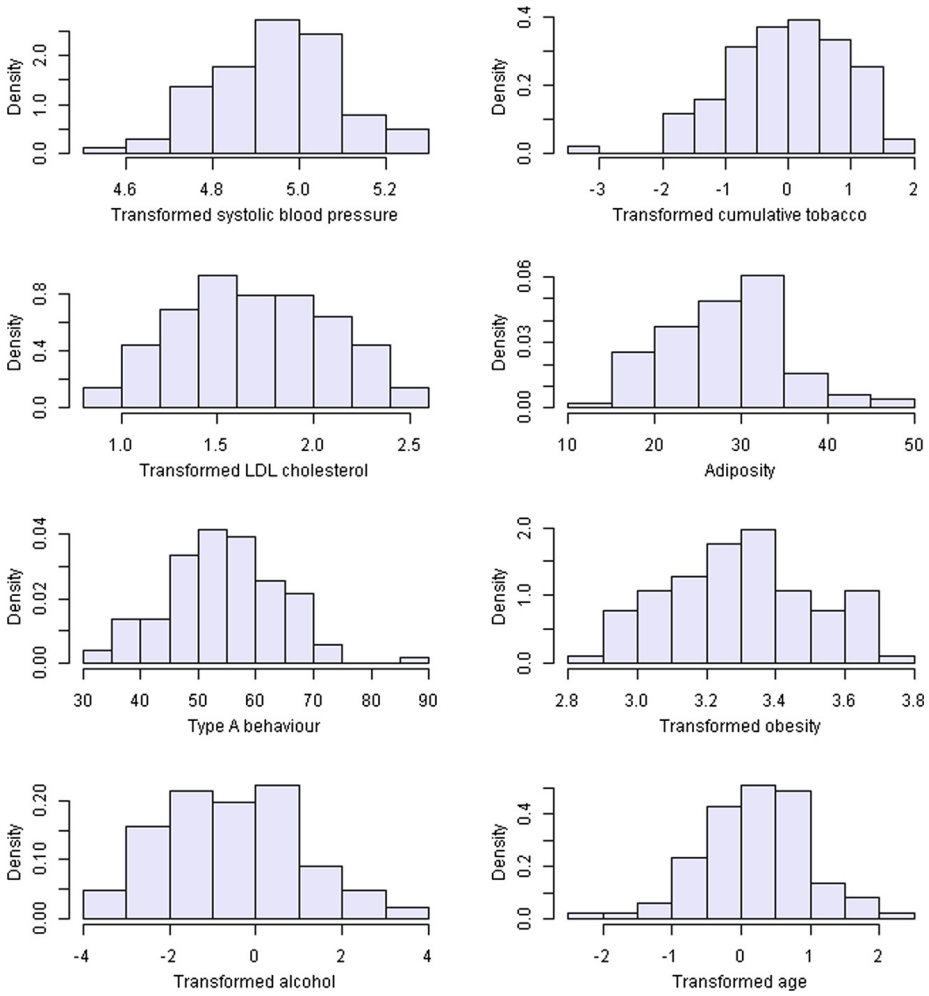


FIG C.2. *Histograms of the covariates of the sample of treated.*

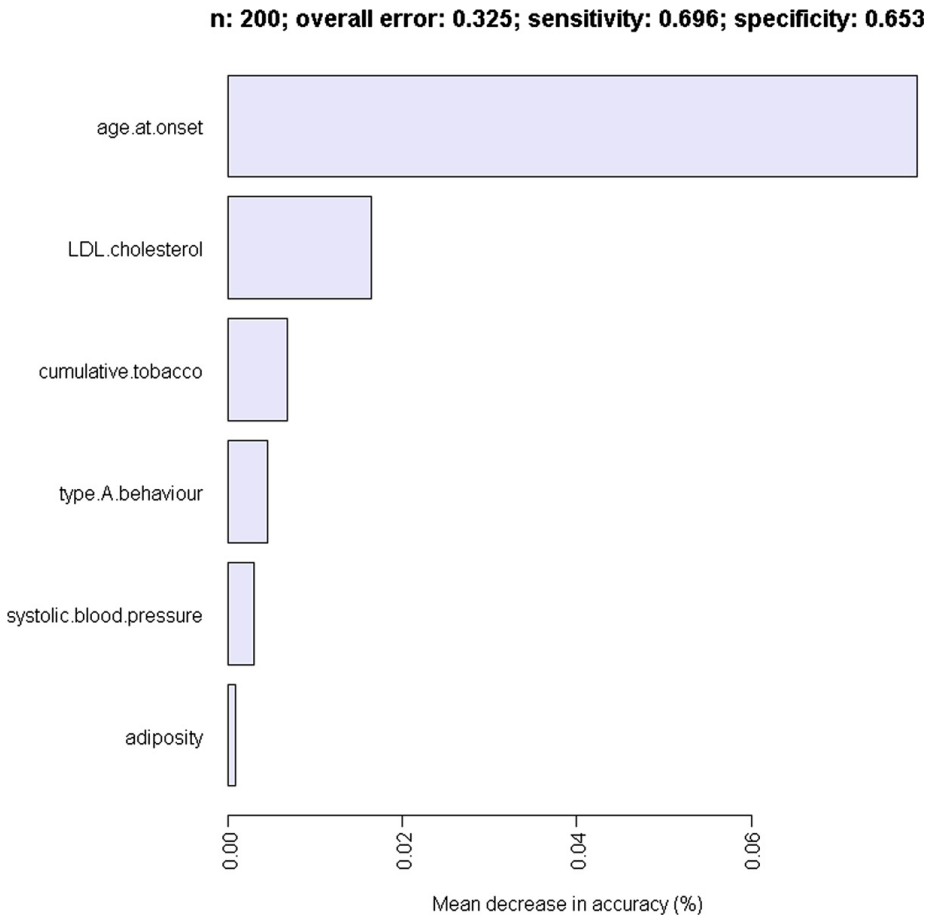


FIG C.3. *Prediction analysis based on a random forest.*

Appendix D: Figures pertaining to subsection 6.1

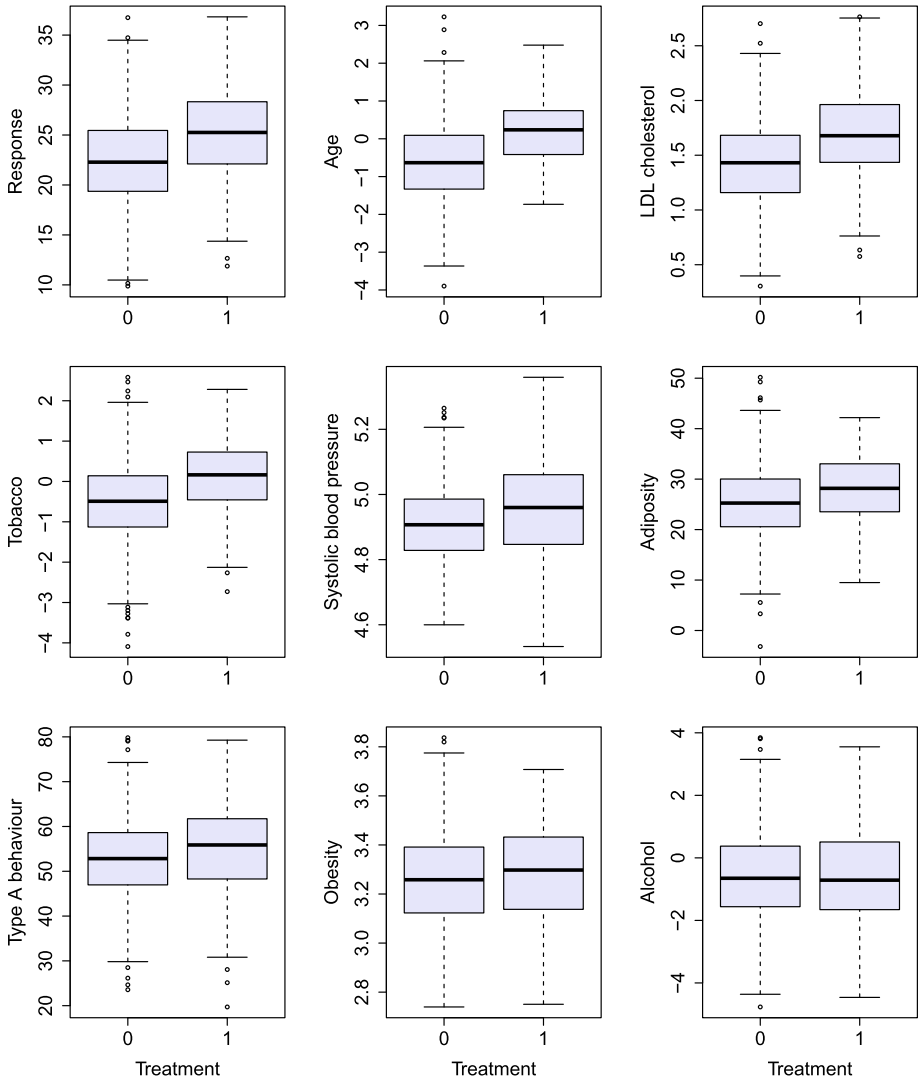


FIG D.1. *Box plots of a subset of the simulated data by treatment group.*

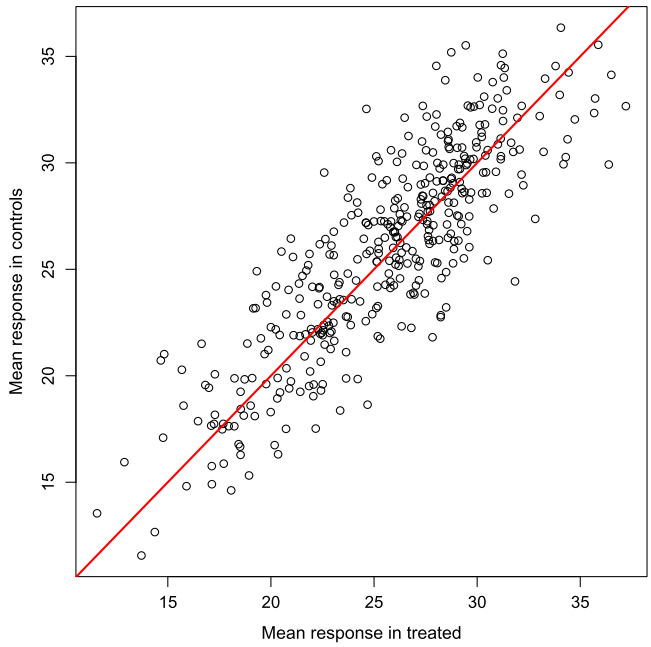


FIG D.2. *Scatter plot comparing the mean response between treated and control units within strata in a stratification based on* $i = 4$.

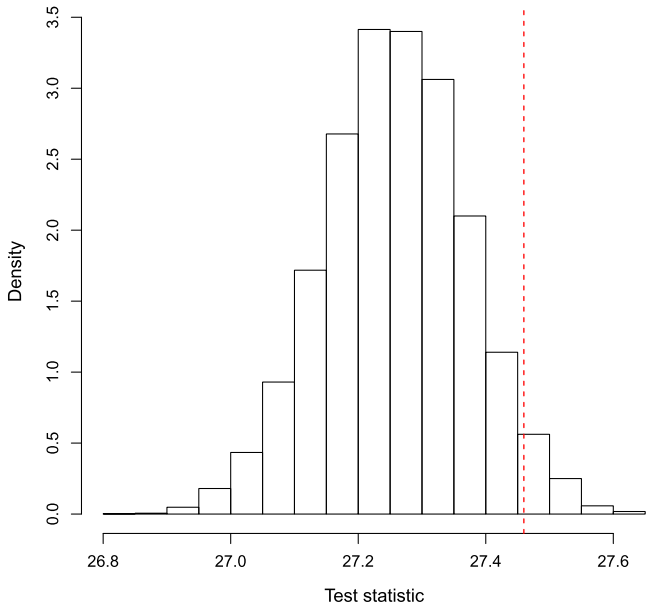


FIG D.3. *Histogram of a sample from the null distribution of the test statistic corresponding to a stratification based on* $i = 5$.

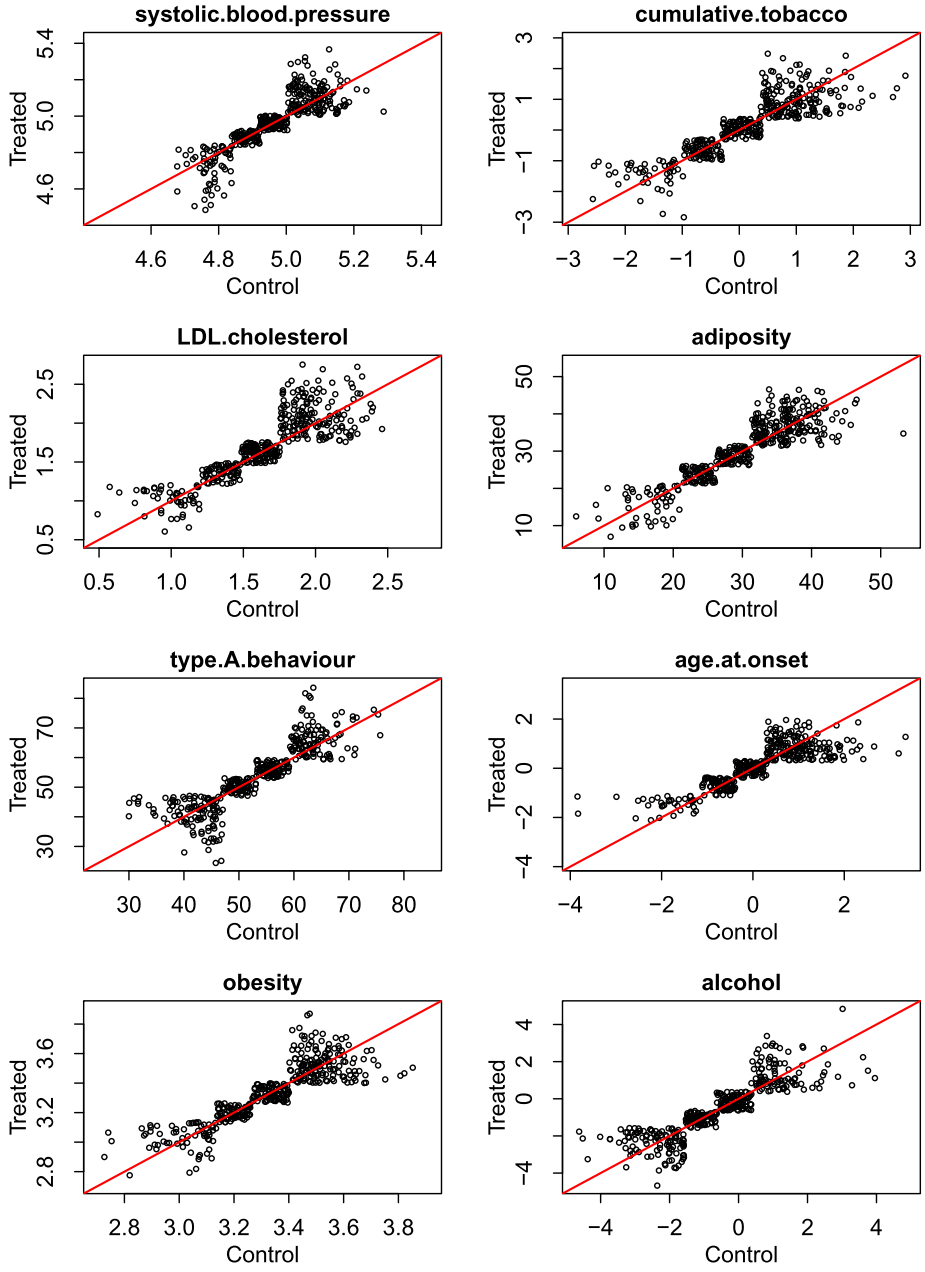


FIG D.4. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on* $i = 4$.

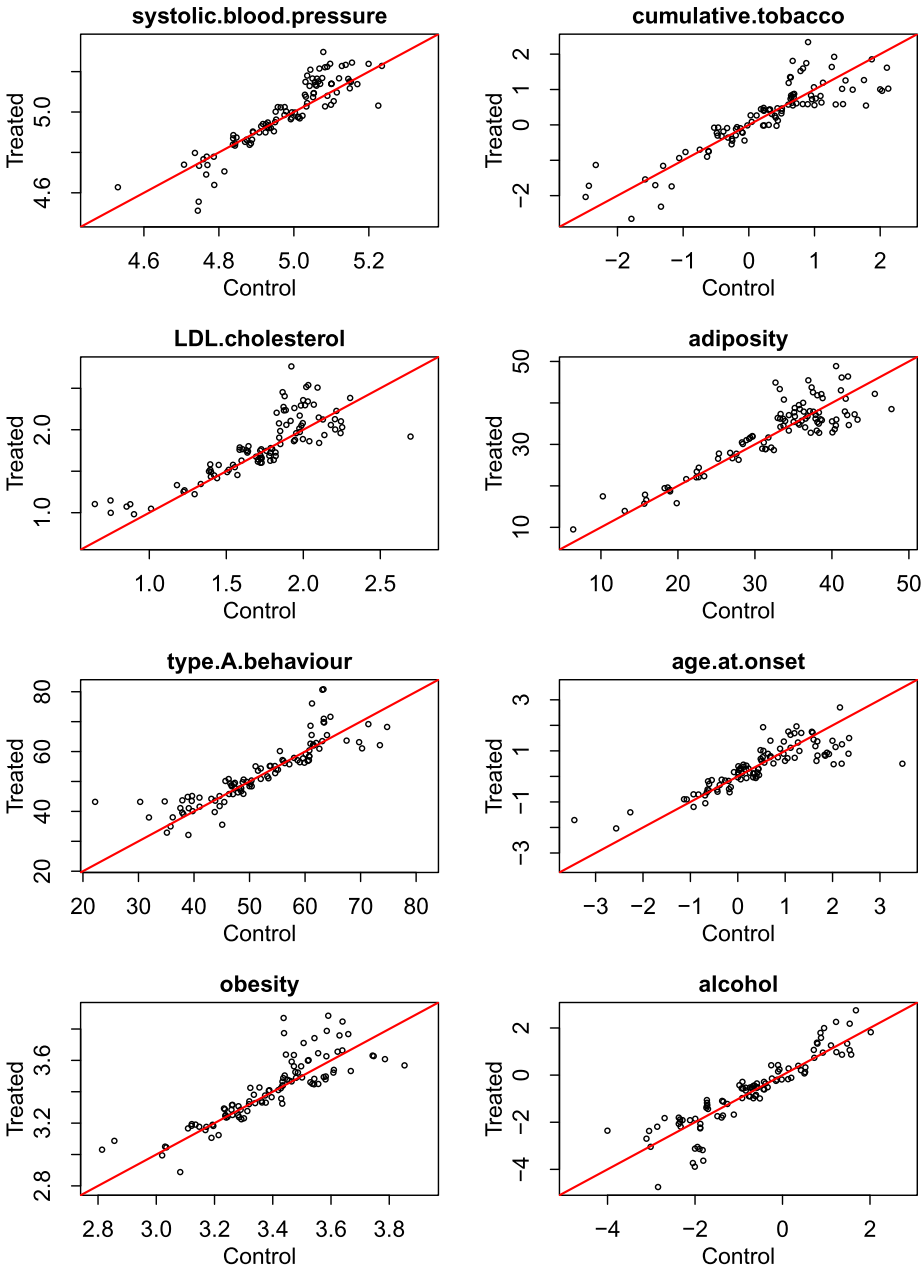


FIG D.5. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on* $i = 5$.

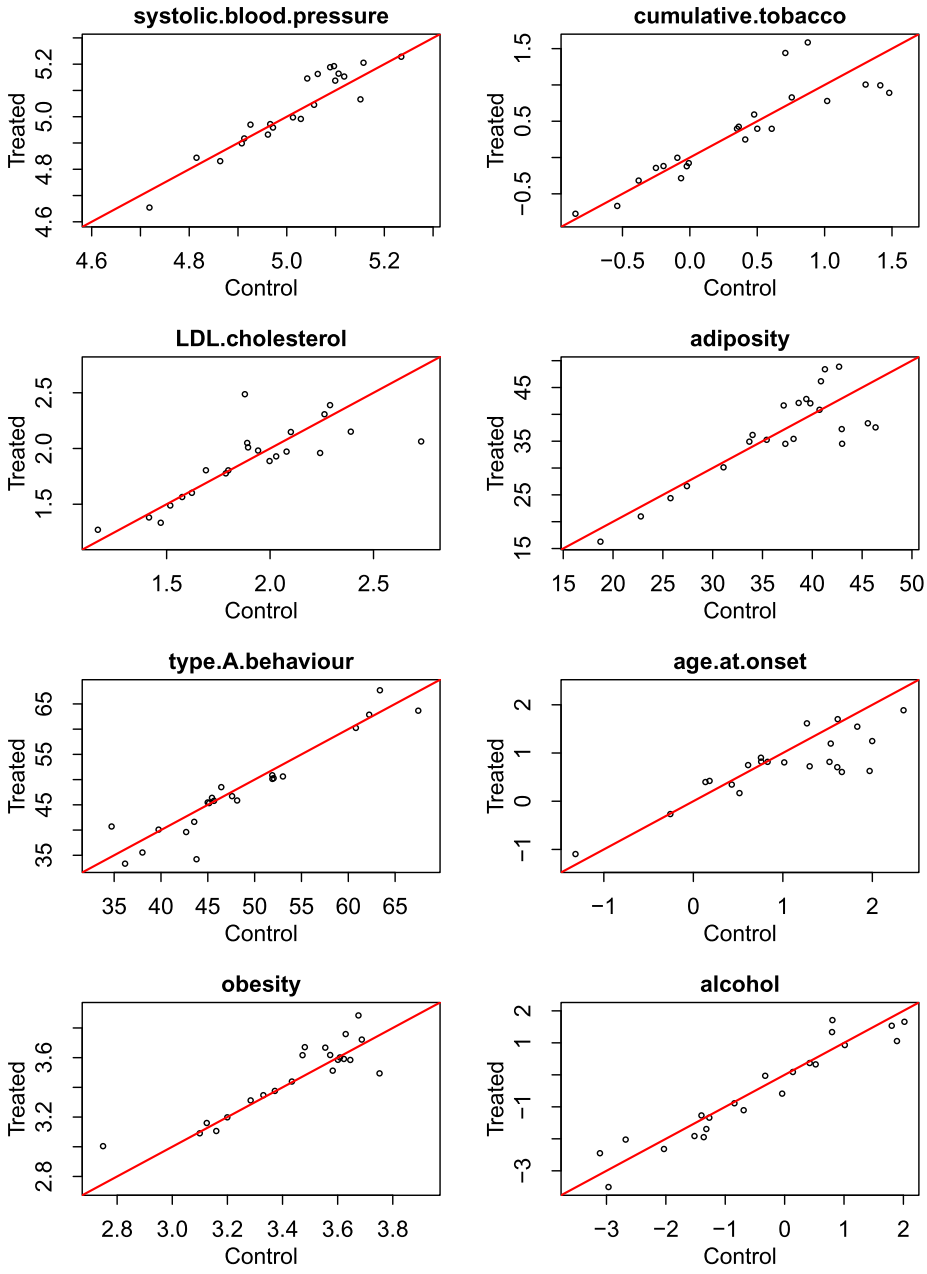


FIG D.6. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on* $i = 6$.

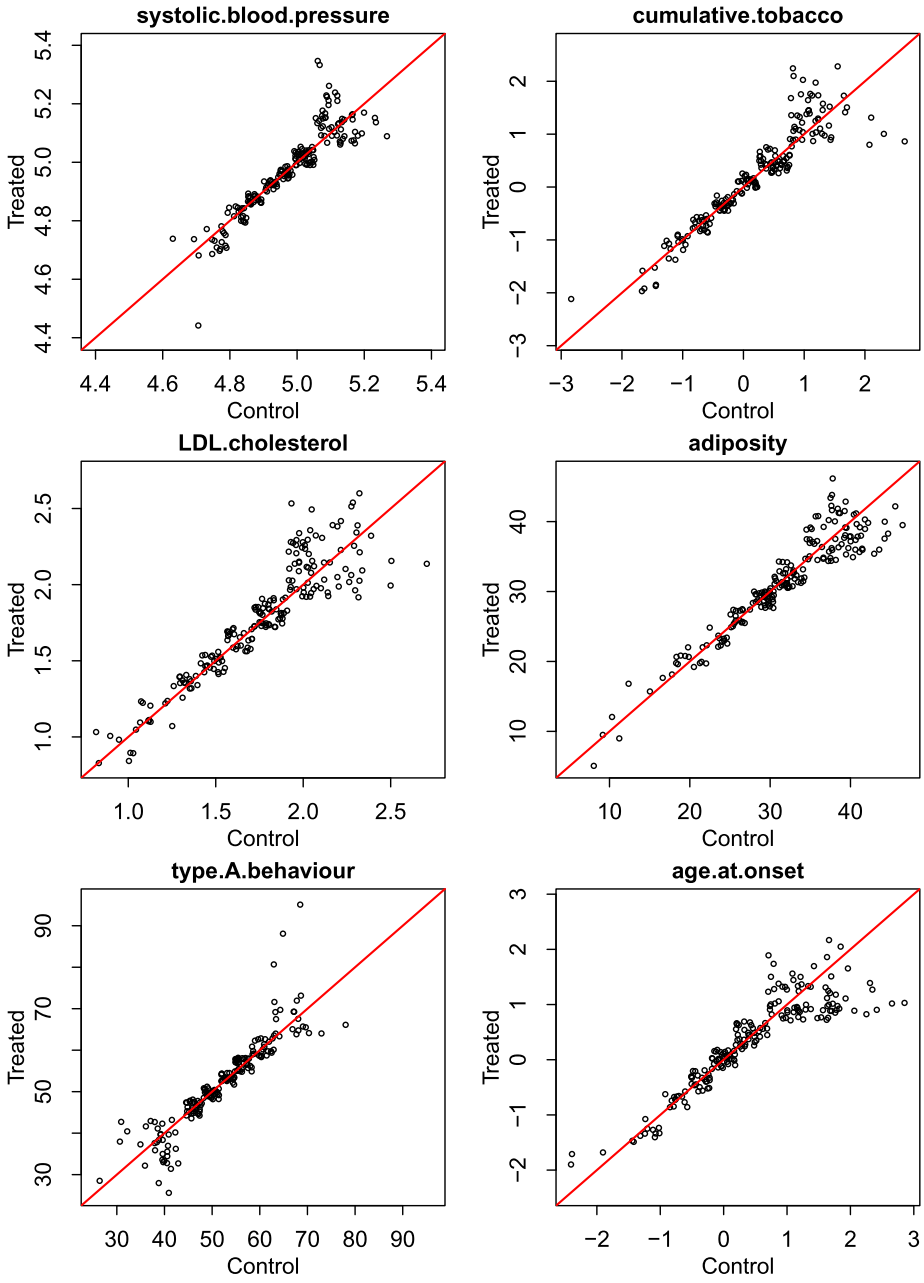


FIG D.7. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 7$ and on the six covariates influencing treatment.*

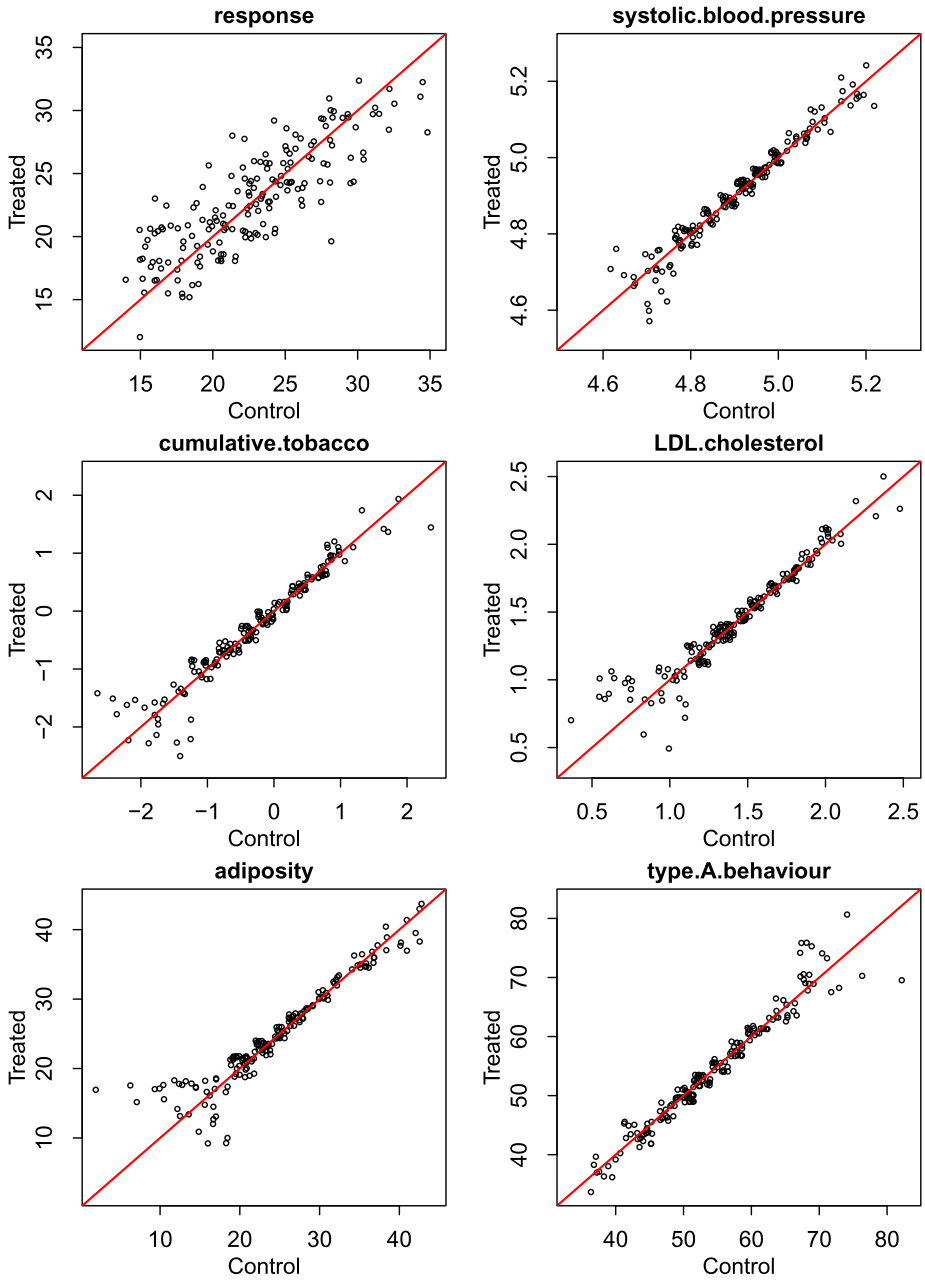


FIG D.8. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 10$ and on five of the covariates influencing treatment.*

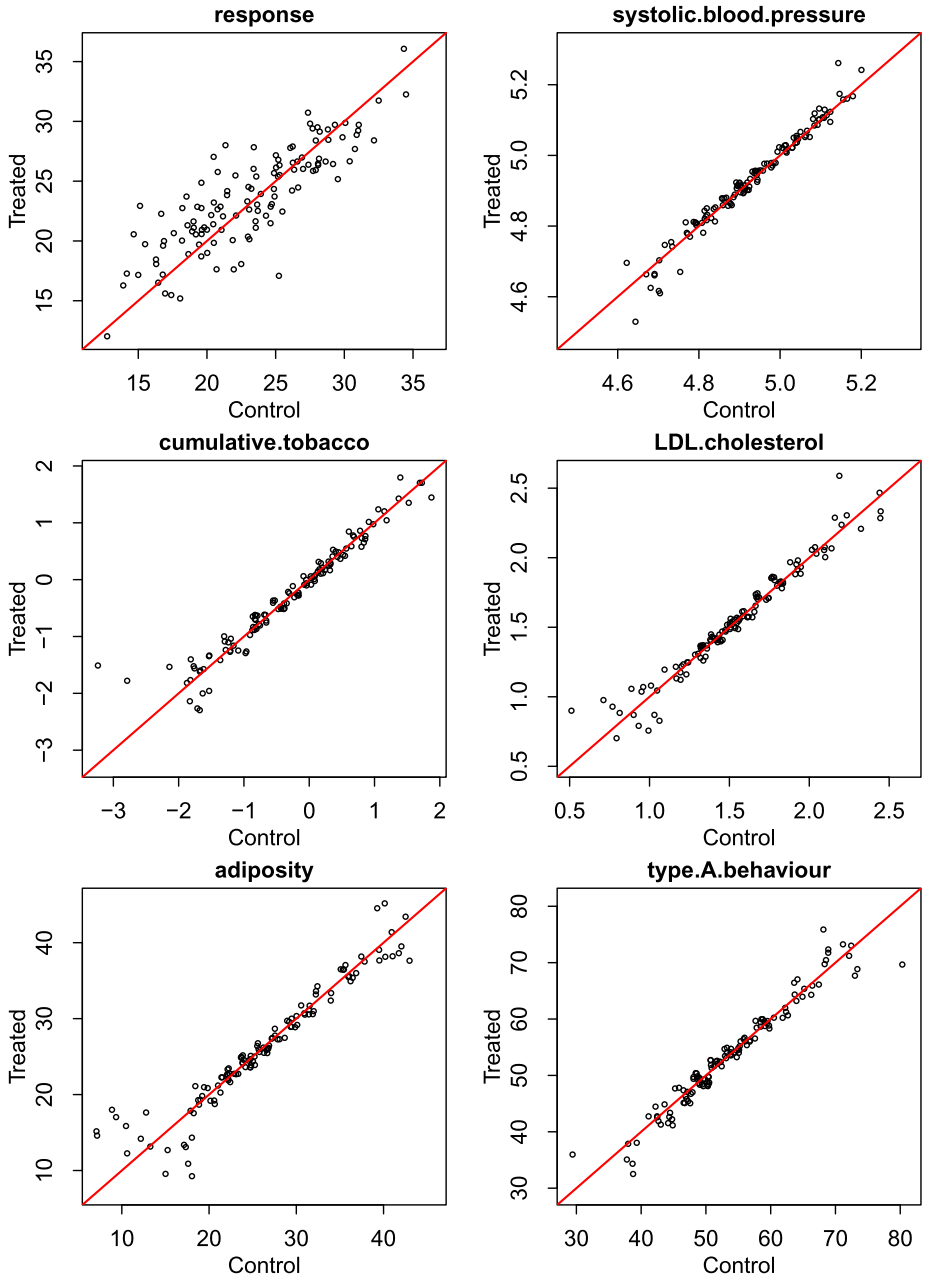


FIG D.9. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 11$ and on five of the covariates influencing treatment.*

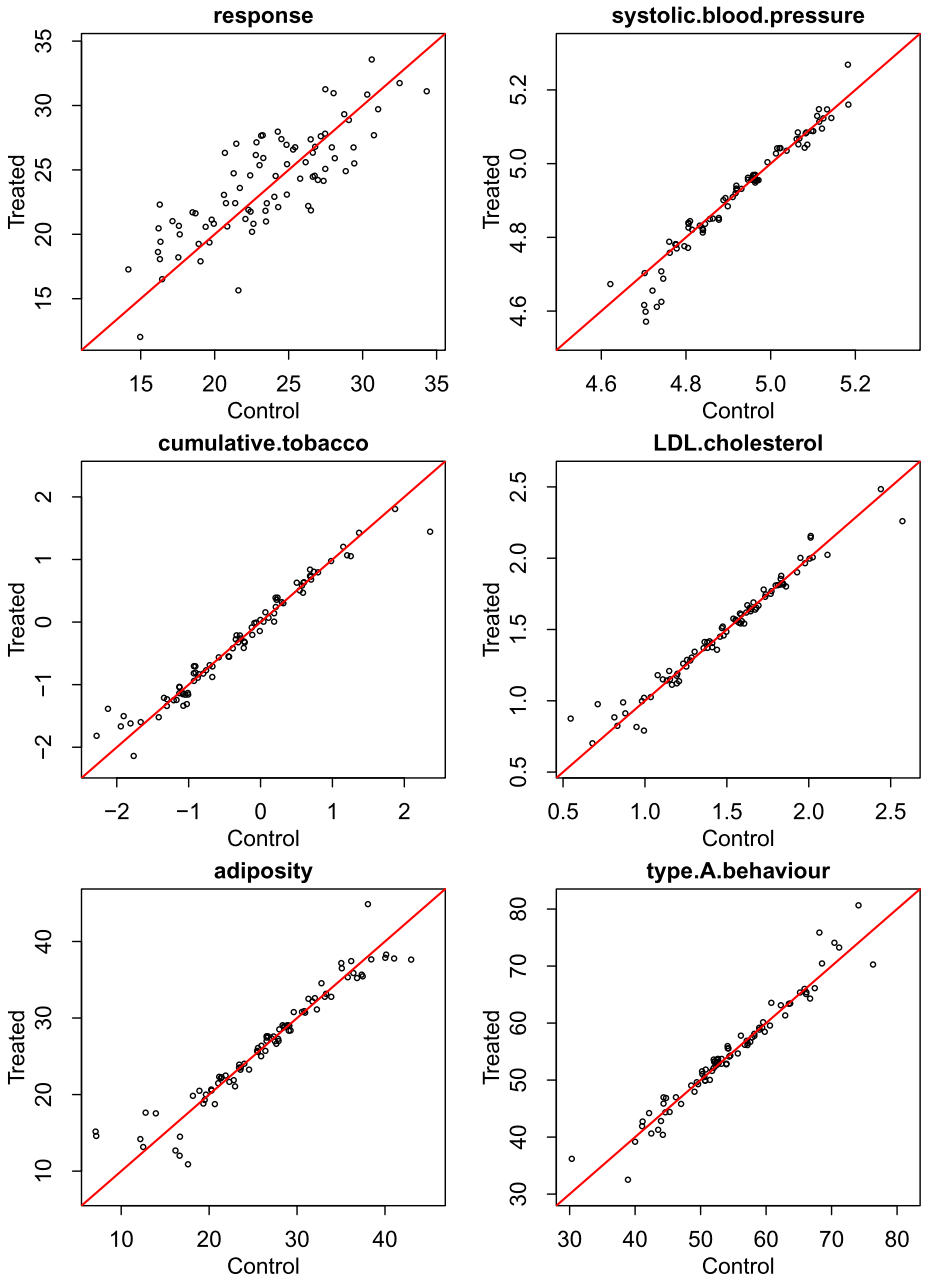


FIG D.10. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 12$ and on five of the covariates influencing treatment.*

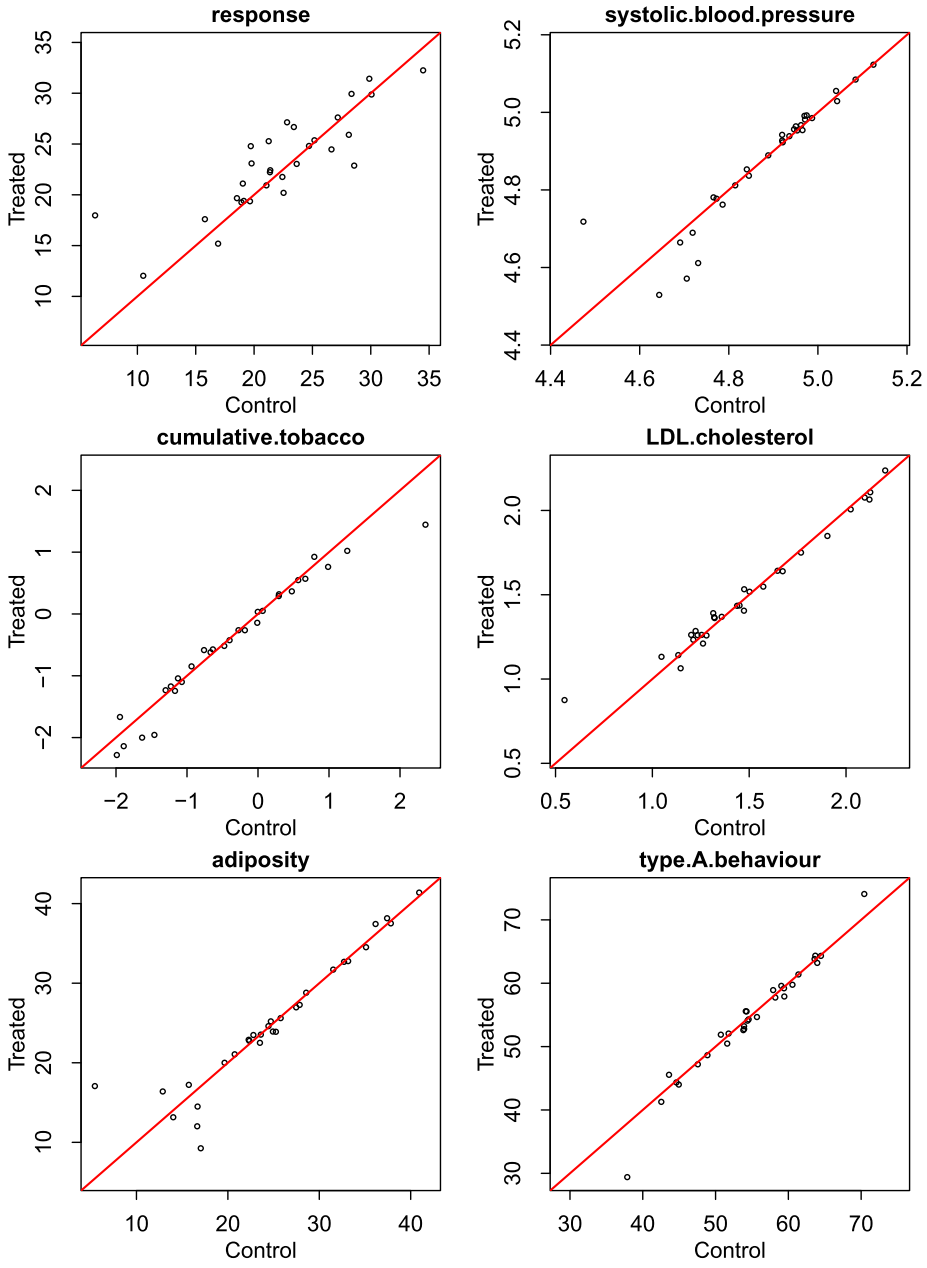


FIG D.11. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 14$ and on five of the covariates influencing treatment.*

Appendix E: Figures pertaining to subsection 6.2

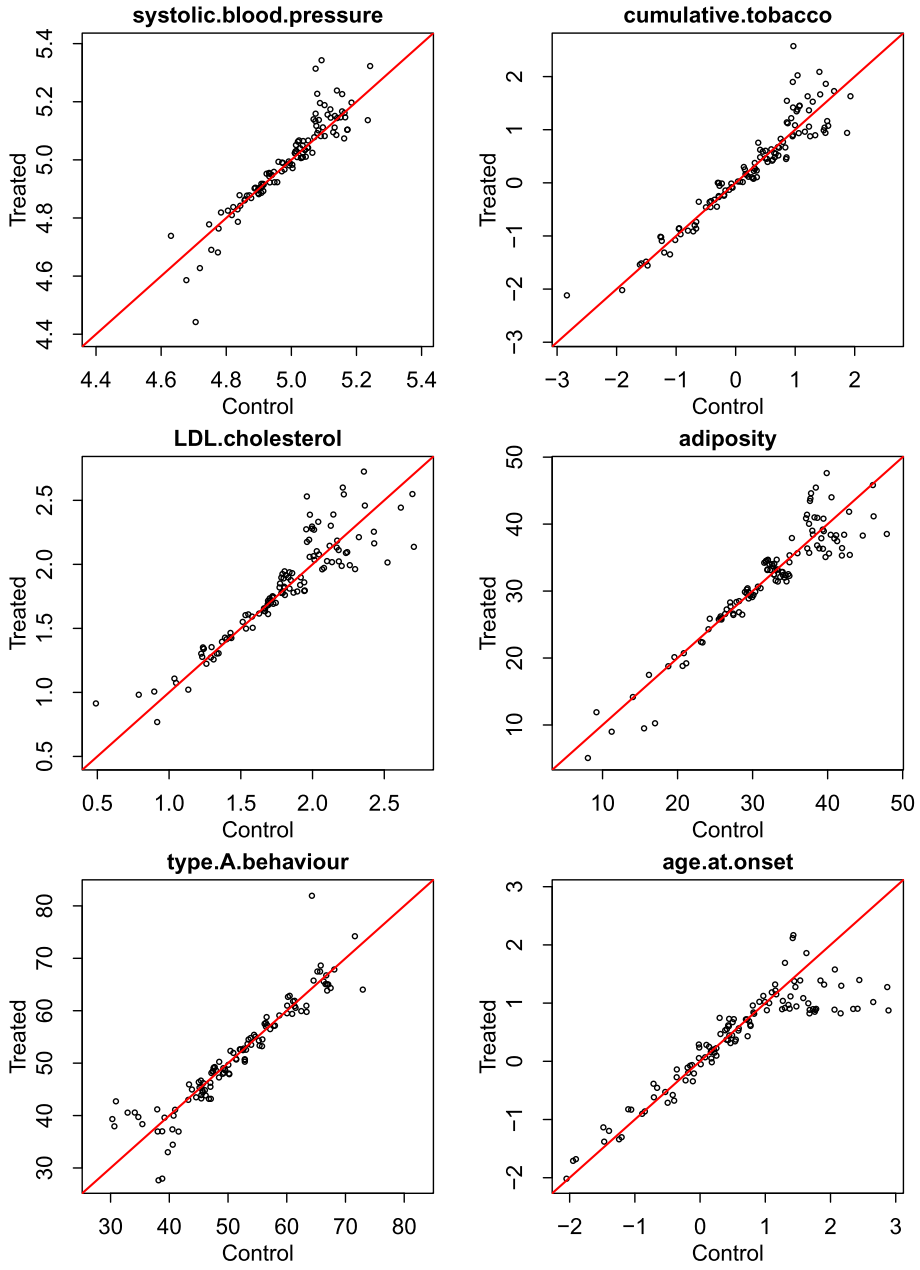


FIG E.1. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 8$ and on the six covariates influencing treatment.*

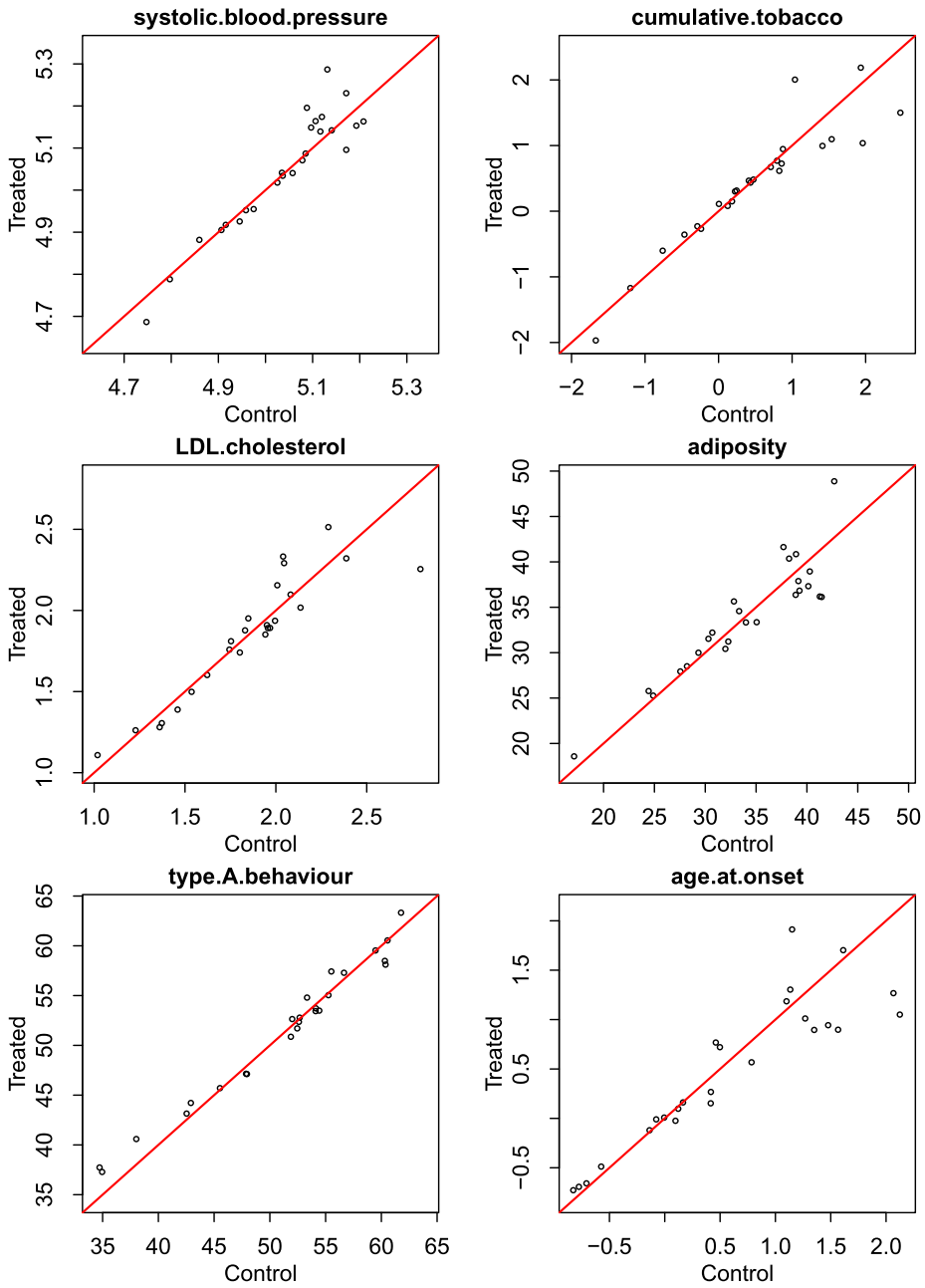


FIG E.2. *Scatter plots comparing the covariates in the treated and control groups per stratum in a stratification based on $i = 10$ and on the six covariates influencing treatment.*

Appendix F: Figures pertaining to subsection 6.3

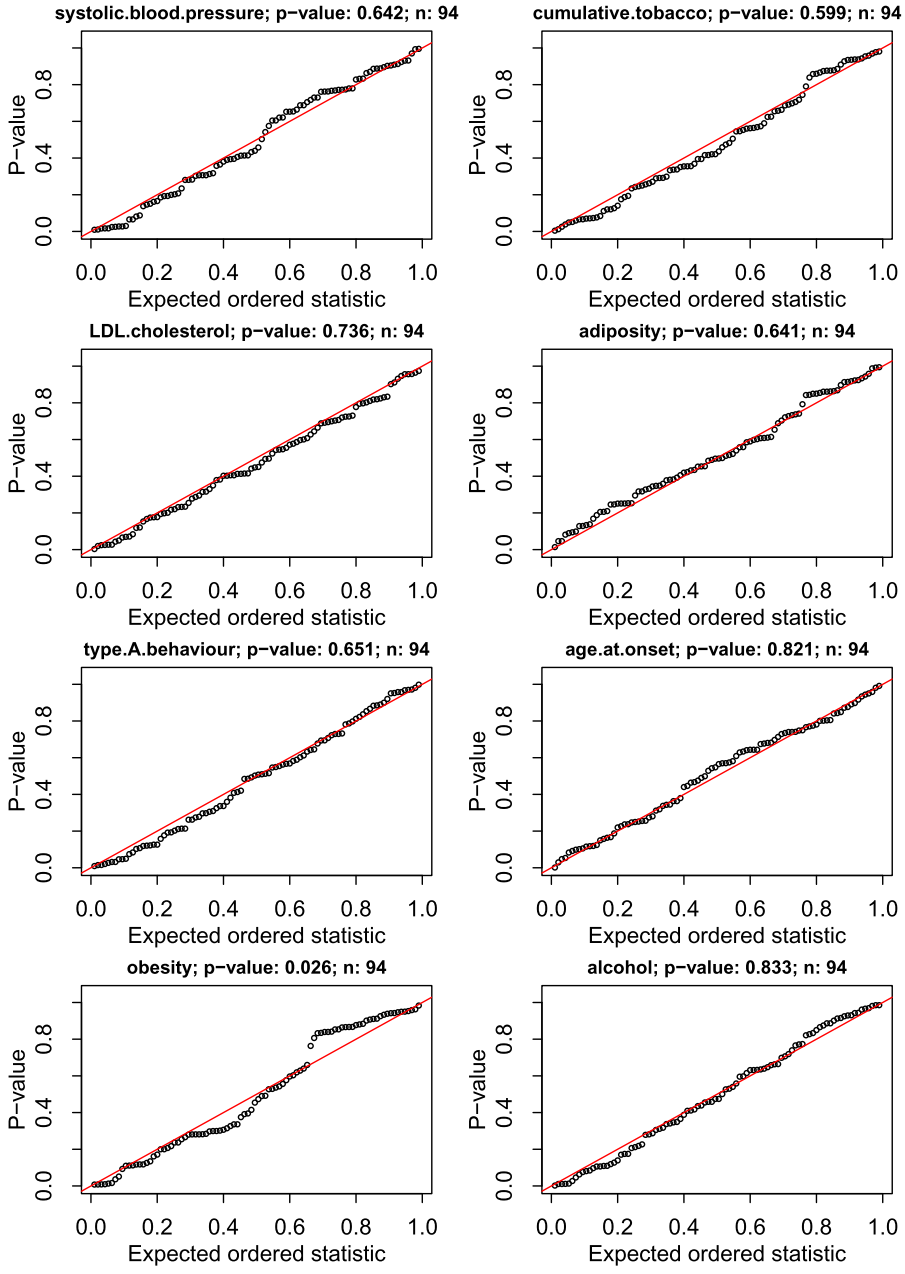


FIG F.1. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the true propensity score based on* $\ell = 0.01$.

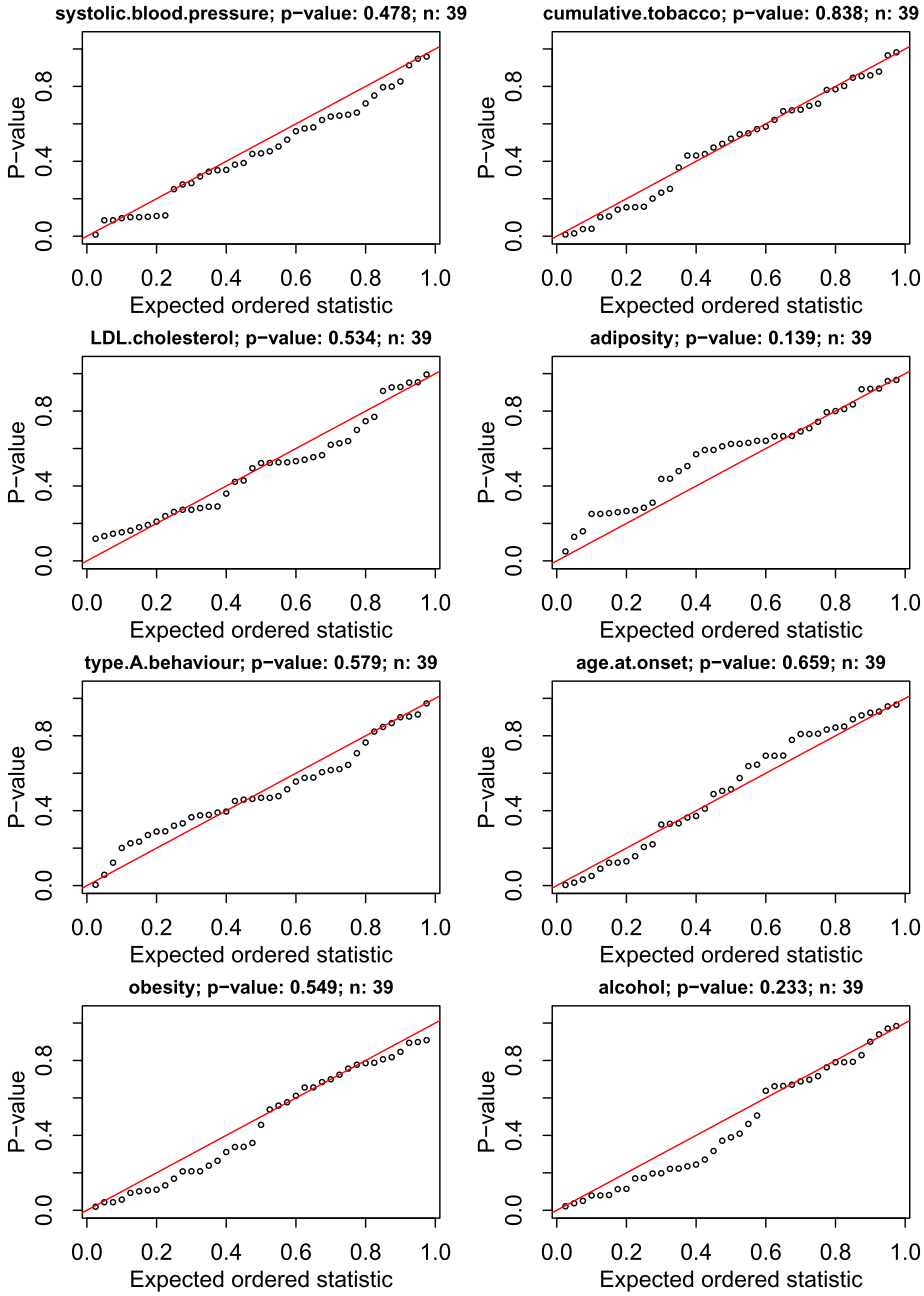


FIG F.2. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the true propensity score based on* $\ell = 0.025$.

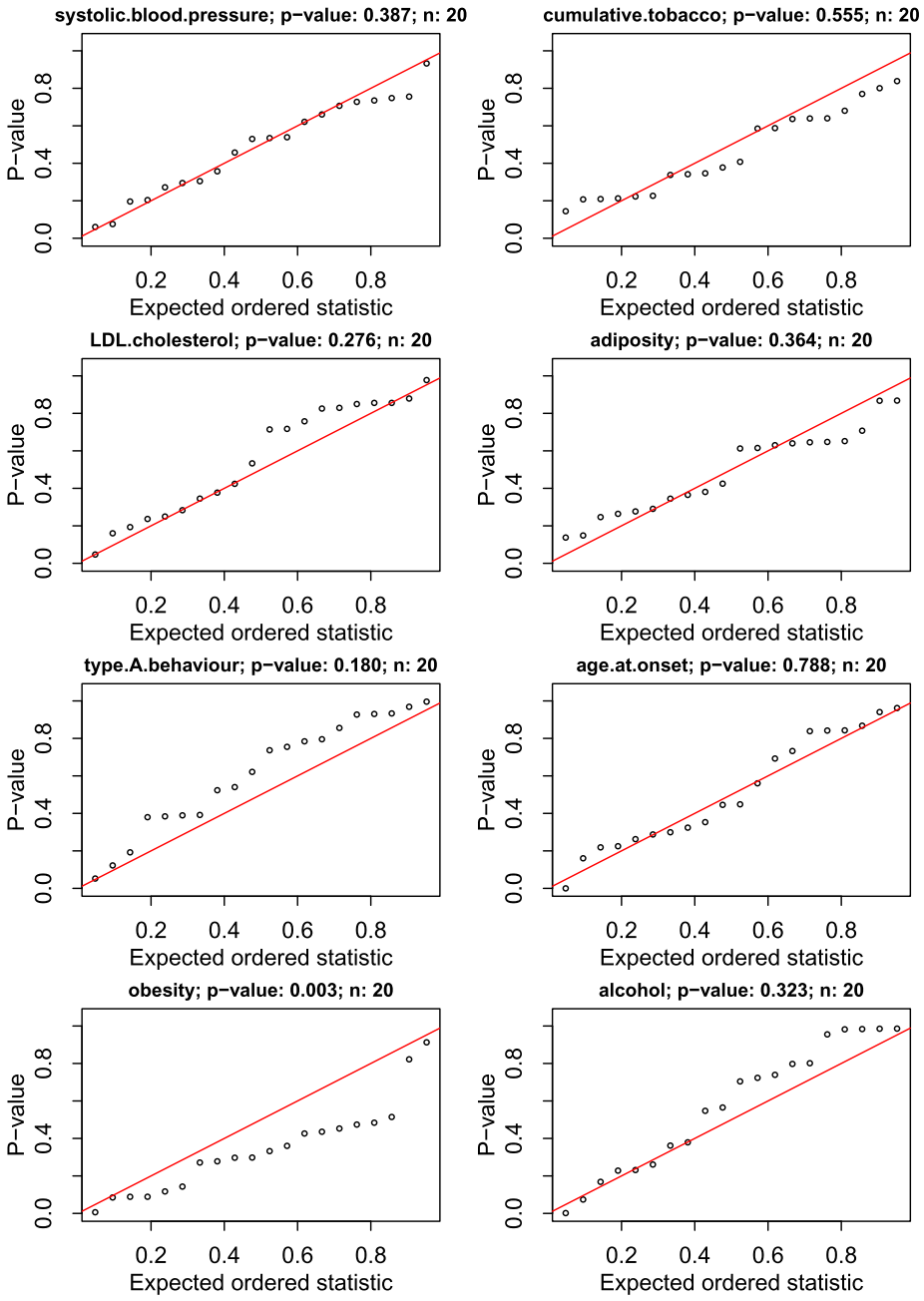


FIG F.3. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the true propensity score based on* $\ell = 0.05$.

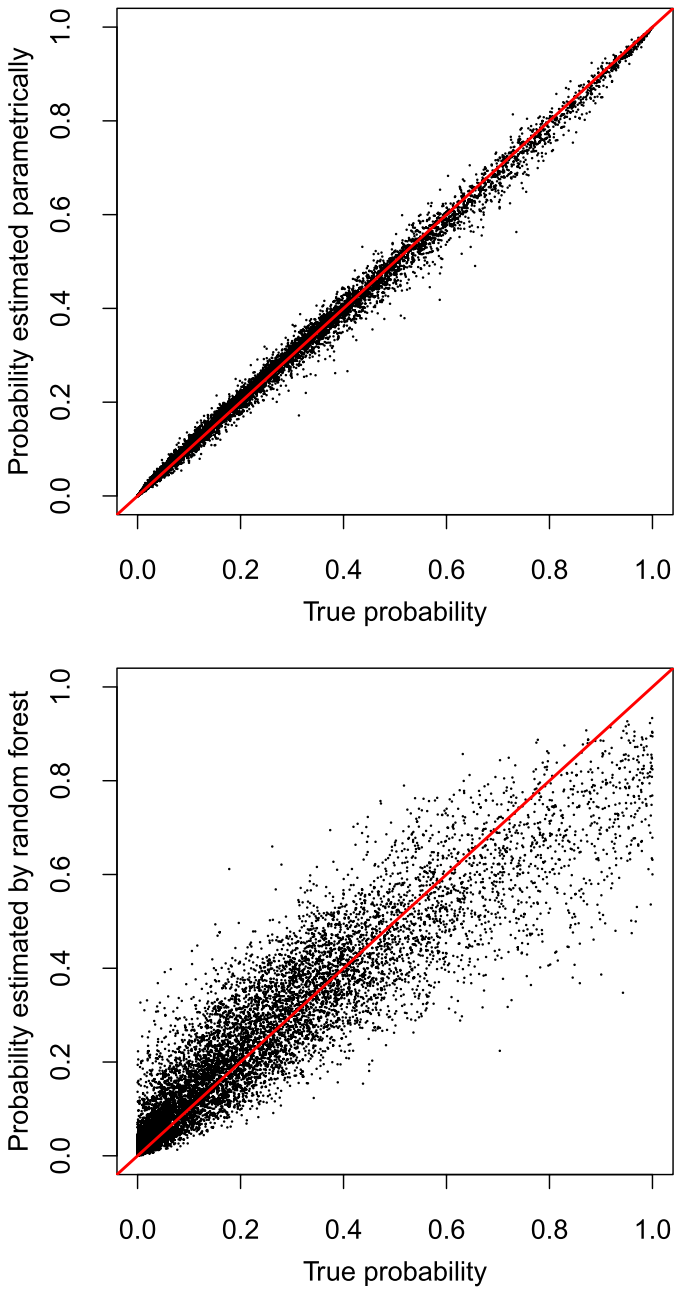


FIG F.4. *Scatter plots comparing the true propensity score with estimates of it obtained parametrically (by fitting the correct model to the data) and by a random forest classifier.*

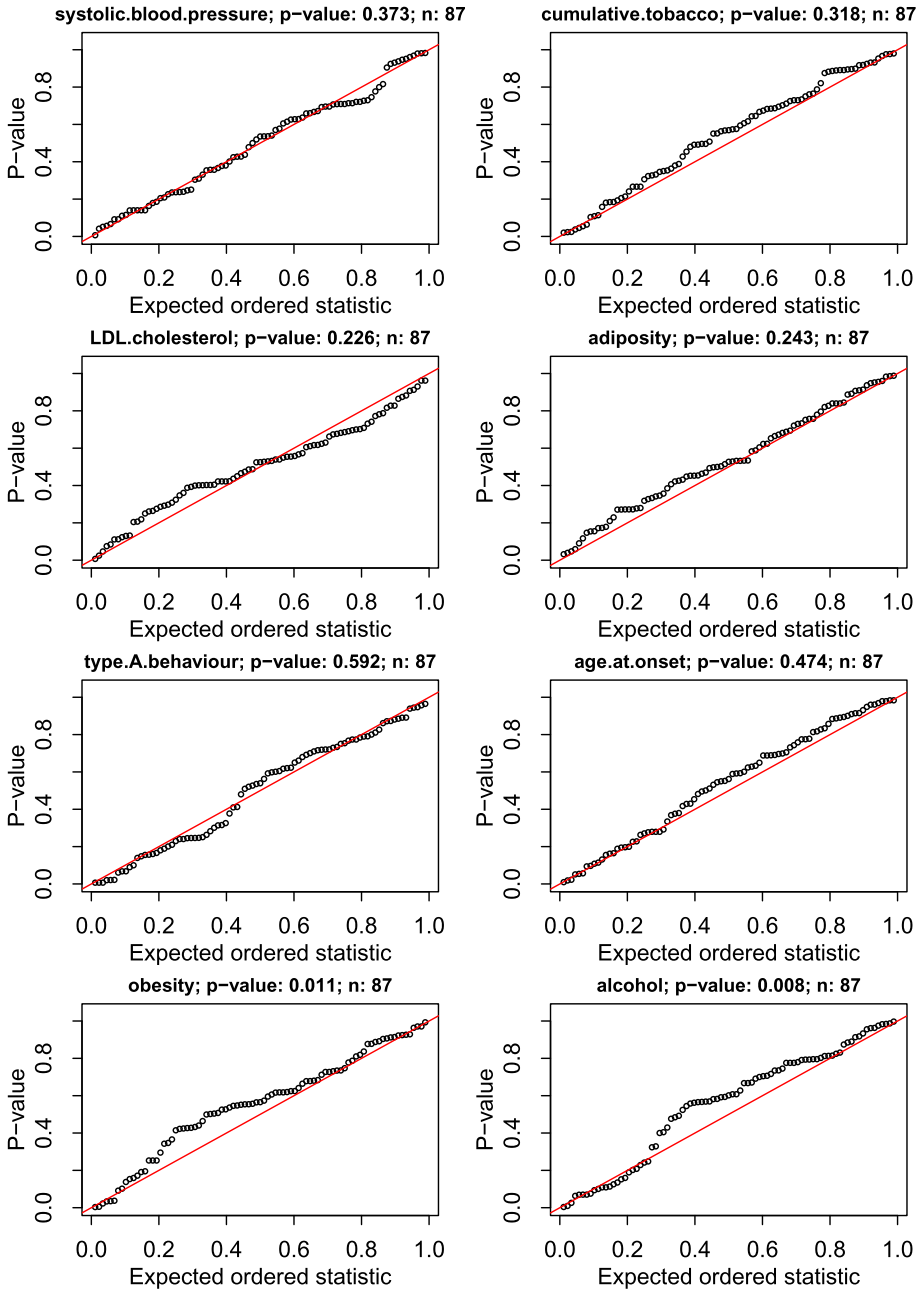


FIG F.5. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the estimated propensity score based on* $\ell = 0.01$.

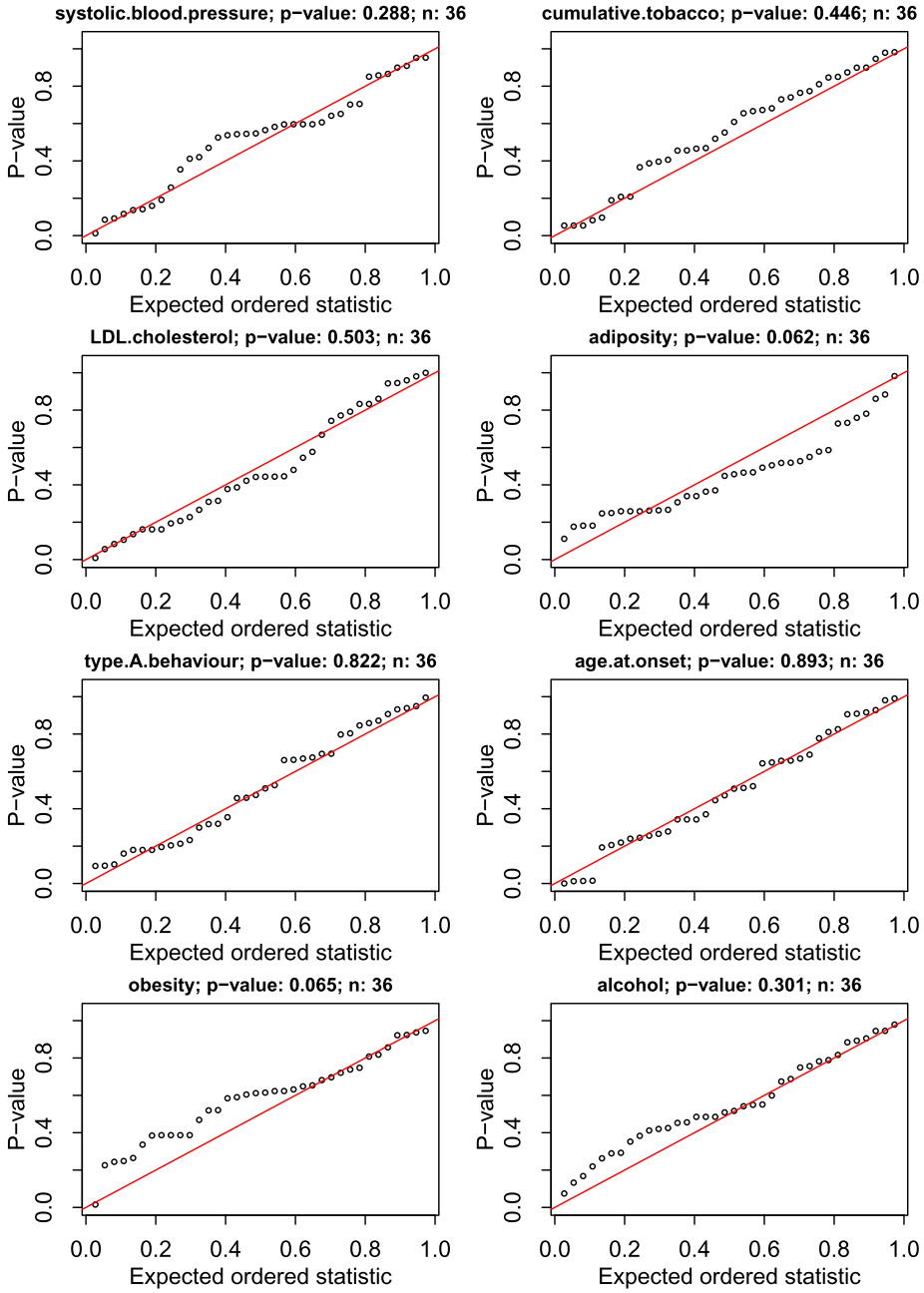


FIG F.6. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the estimated propensity score based on* $\ell = 0.025$.

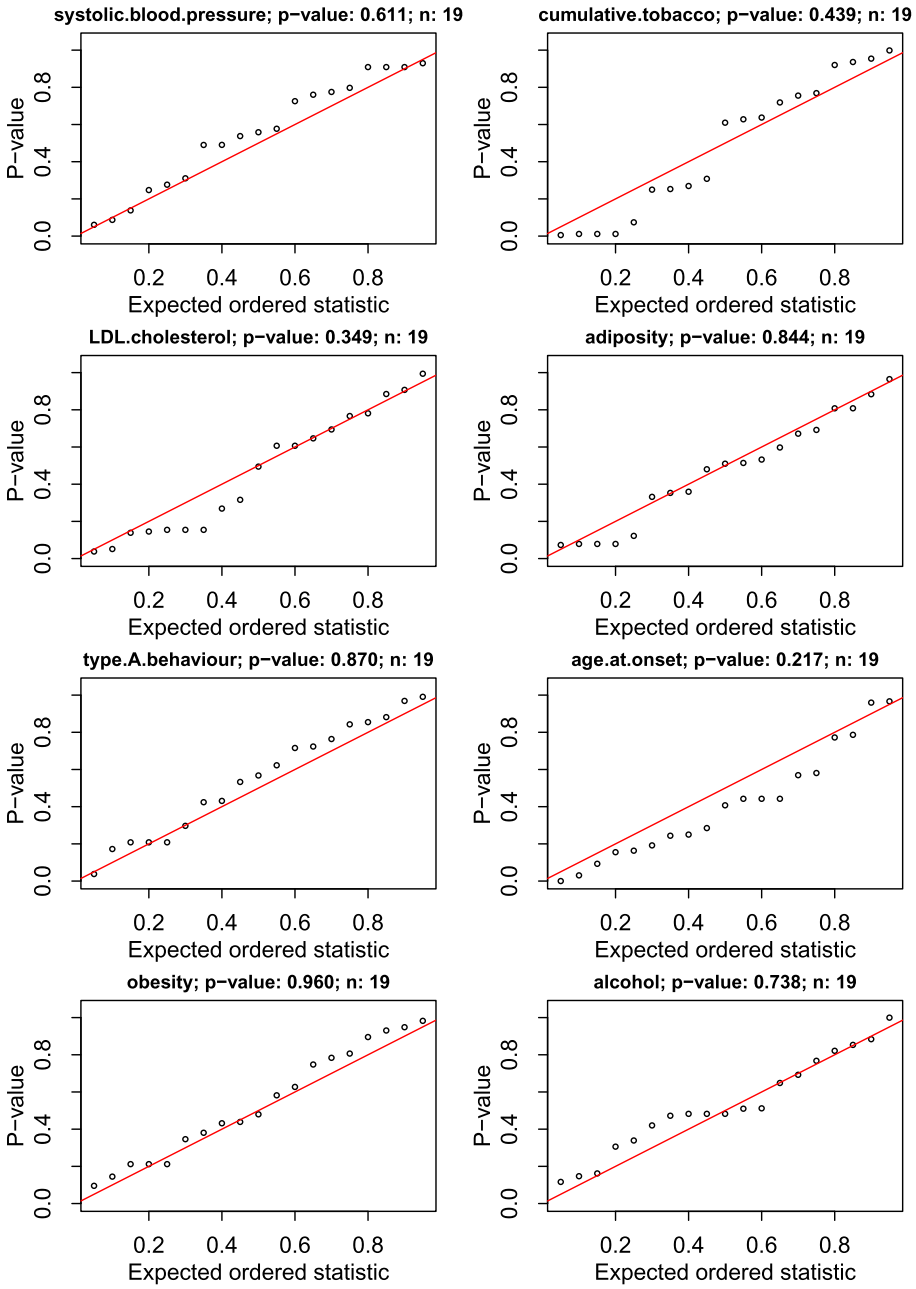


FIG F.7. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the estimated propensity score based on* $\ell = 0.05$.

Appendix G: Figures pertaining to subsection 6.4

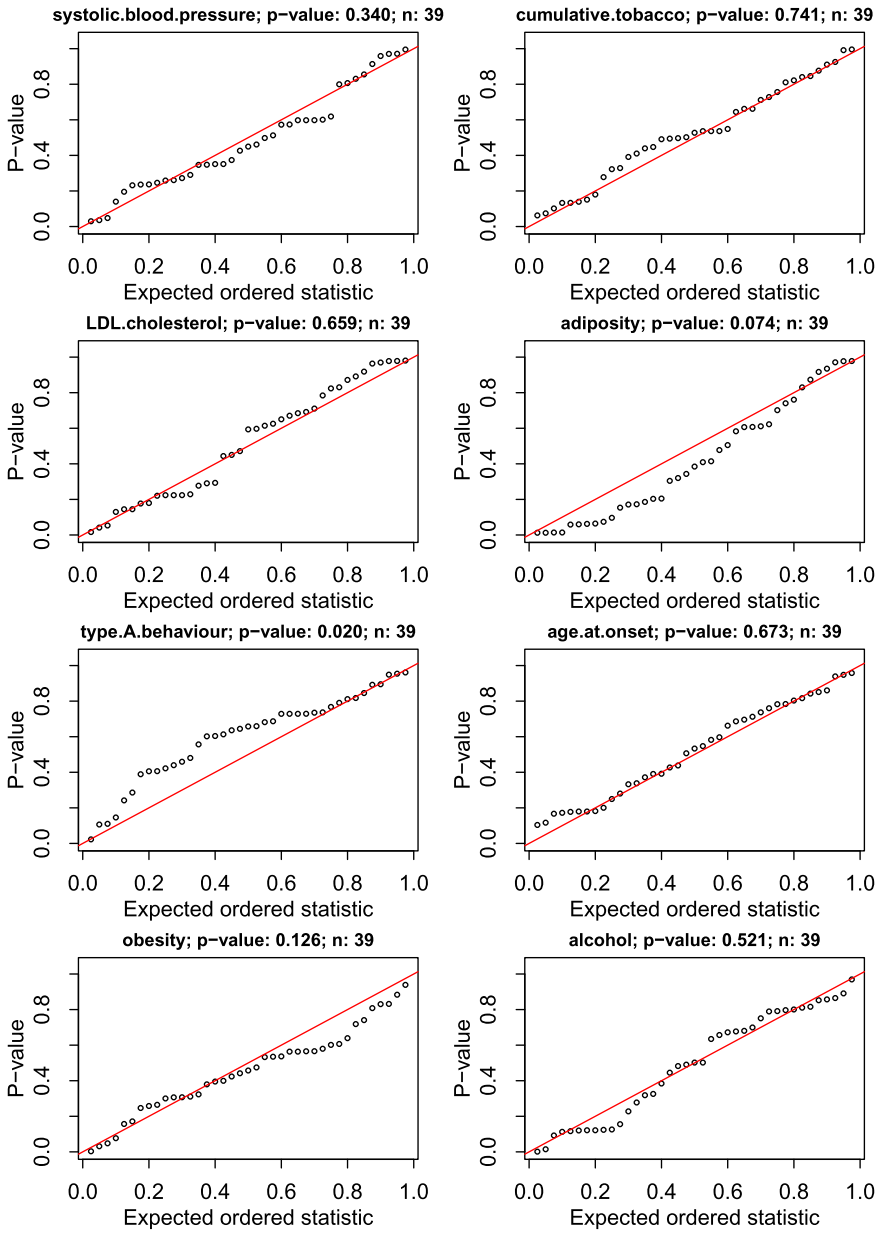


FIG G.1. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification on the parametric estimate of the propensity score based on* $\ell = 0.025$.

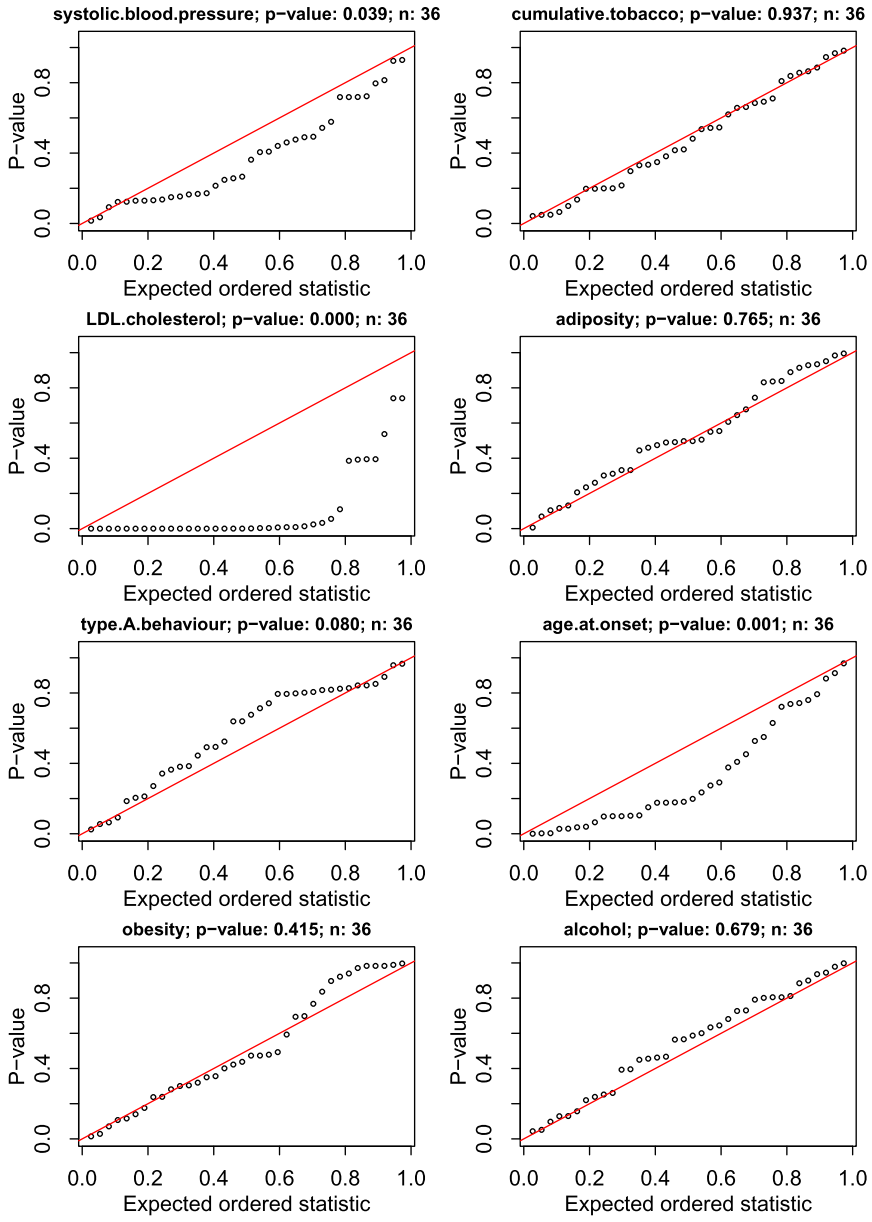


FIG G.2. *Probability plots assessing the uniformity of the p-values obtained from Anderson-Darling tests comparing the distribution of the covariates in the treated and control groups per stratum in a stratification (based on $\ell = 0.025$) on the propensity score estimated without* `LDL.cholesterol`.

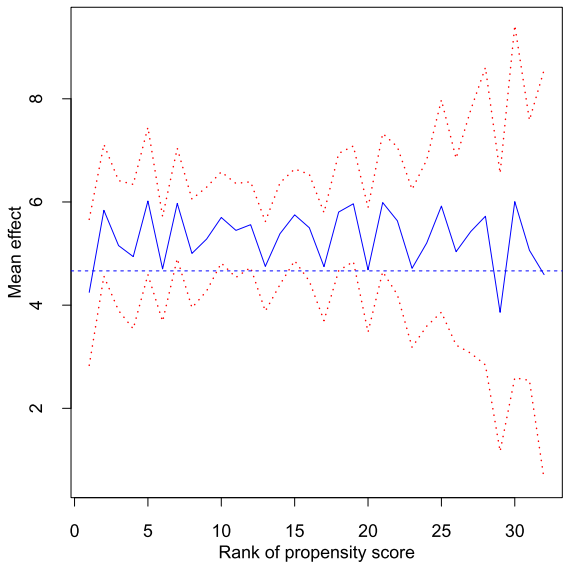


FIG G.3. *Estimates of treatment effect per stratum obtained from a stratification (based on* $\ell = 0.025$*) on the propensity score estimated by a random forest without* `LDL.cholesterol` *(the strata are ranked in increasing order of the propensity score estimate, the dashed horizontal line represents the true overall treatment effect, and the dotted lines represent* 95% *confidence intervals per stratum).*

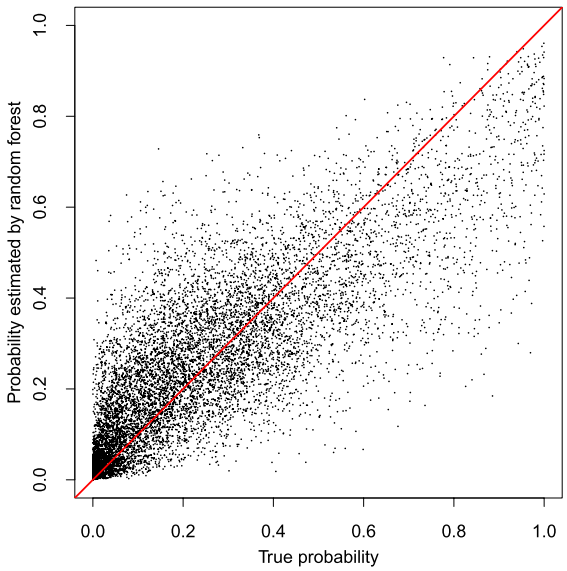


FIG G.4. *Scatter plot comparing the true propensity score with a random forest estimate of it that does not make use of* `LDL.cholesterol`.

Appendix H: Figures pertaining to subsection 6.5

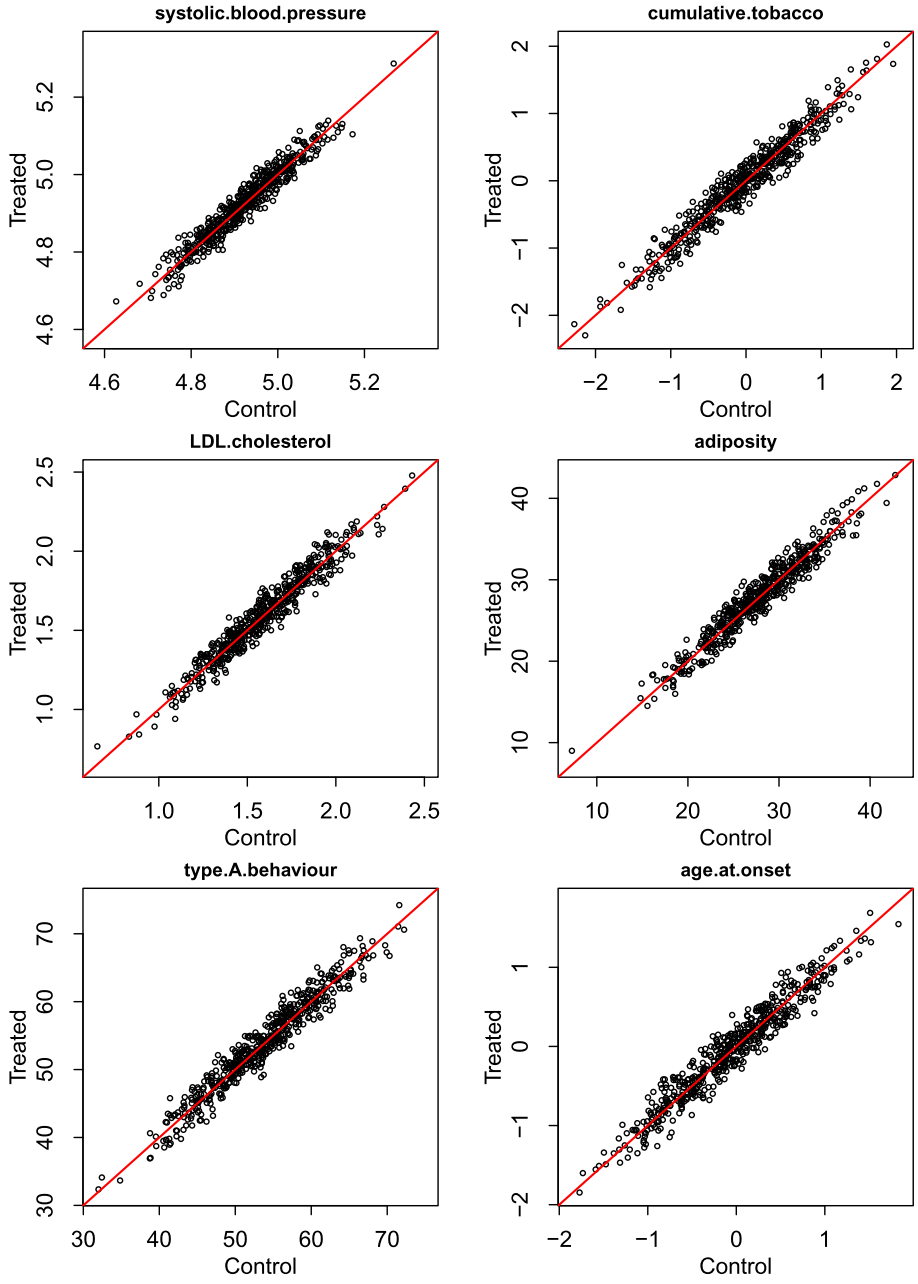


FIG H.1. *Scatter plots comparing the covariates in matched sets of treated and control groups obtained by full matching based on the Mahalanobis distance and* $C = 0.5$.

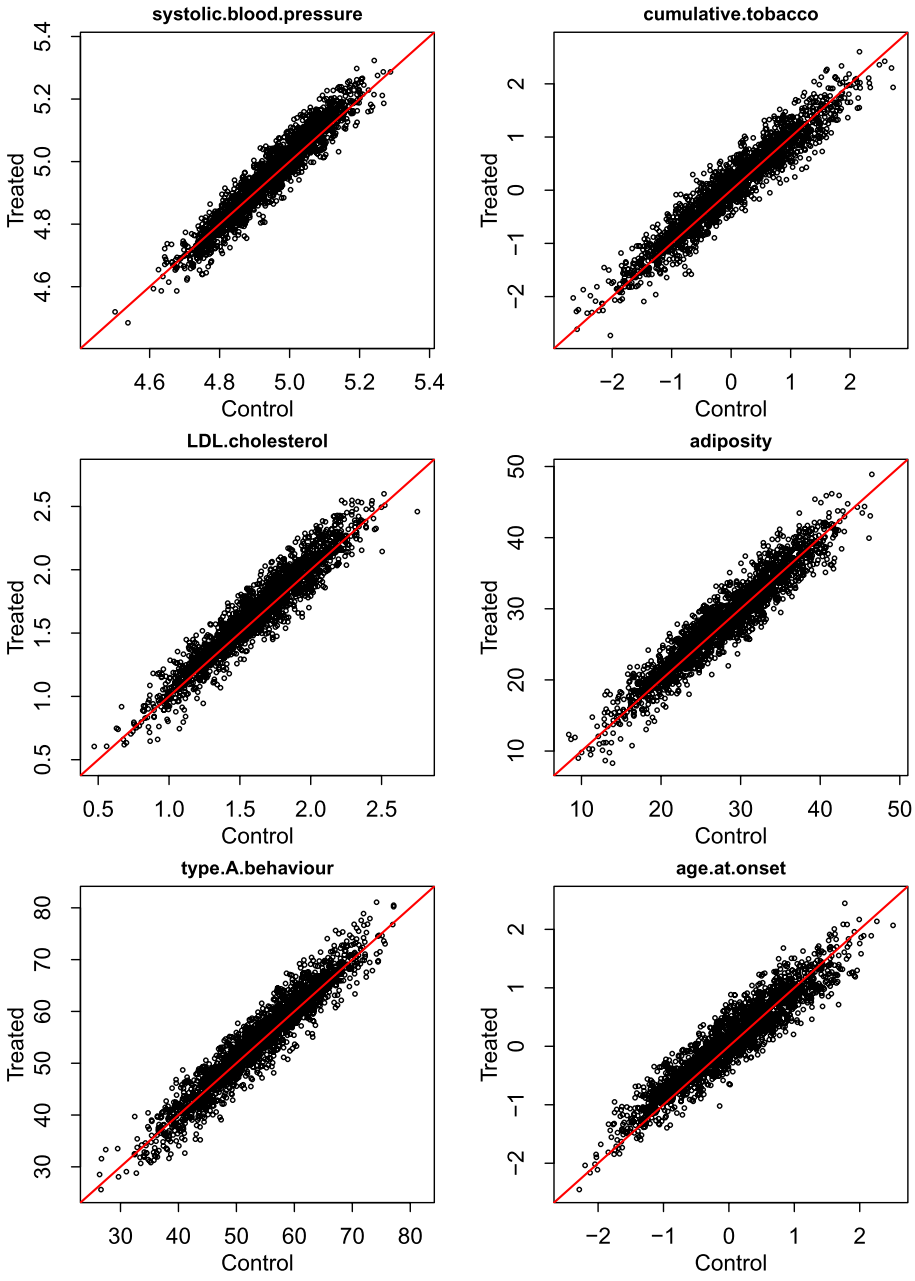


FIG H.2. *Scatter plots comparing the covariates in matched sets of treated and control groups obtained by full matching based on the Mahalanobis distance and* $C = 1$.

References

[1] Abadie, A. and Imbens, G.W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76** 1537–1557. MR2468559

[2] Arjas, E. and Parner, J. (2004). Causal reasoning from longitudinal data. *Scandinavian Journal of Statistics* **31** 171–187. MR2066247

[3] Arjas, E. (2012). Causal inference from observational data: a Bayesian predictive approach. *Causality: Statistical Perspectives and Applications*, edited by C. Berzuini, A.P. Dawid and L. Bernardinelli, pp. 71–84. Chichester, UK.

[4] Austin, P.C. (2012). Using ensemble-based methods for directly estimating causal effects: an investigation of tree-based G-computation. *Multivariate Behavioral Research* **47** 115–135.

[5] Biau, G., Devroye, L. and Lugosi, G. (2012). Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research* **9** 2015–2033. MR2447310

[6] Cornfield, J.(1951). A method for estimating comparative rates from clinical data. Applications to cancer of the lung, breast, and cervix. *Journal of the National Cancer Institute* **11** 1269–1275.

[7] ElemStatLearn (2012). Data sets, functions and examples from the book "The Elements of Statistical Learning, Data Mining, Inference, and Prediction" by T. Hastie, R. Tibshirani and J. Friedman. *Material from the book's webpage* (http://CRAN.R-project.org/package=ElemStatLearn) *and* `R` *port and packaging by Kjetil Halvorsen.*

[8] Ferguson, T.S. (1996). *A Course in Large Sample Theory.* Chapman and Hall. MR1699953

[9] Freedman, D.A. (2009). *Statistical Models: Theory and Practice*, revised ed. Cambridge University Press. MR2489600

[10] Freedman, D.A. (2010). *Statistical Models and Causal Inference: A Dialogue with the Social Sciences*, edited by D. Collier, J.S. Sekhon and P.B. Stark. Cambridge University Press. MR2668307

[11] Hade, E.M. and Lu, B. (2014). Bias associated with using the estimated propensity score as a regression covariate. *Statistics in Medicine* **33** 74–87. MR3141554

[12] Hansen, B.B. and Klopfer, S.O. (2006). Optimal full matching and related designs via network flows. *Journal of Computational and Graphical Statistics* **15** 609–627. MR2280151

[13] Hayfield T. and Racine J.S. (2008). Nonparametric econometrics: The `np` package. *Journal of Statistical Software* **27** (5).

[14] Imbens, G.W. (2004). Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* **86** (1) 4–29.

[15] Imbens, G.W. and Rubin, D.B. (2006). *Causal Inference in Statistics, Social and Biomedical Sciences: An Introduction.* Cambridge University Press. MR3309951

[16] Lee, B.K., Lessler, J. and Stuart, E.A. (2010). Improving propensity score weighting using machine learning. *Statistics in Medicine* **29** 337–346. MR2750549

[17] Lehmann, E.L. (1986). *Testing Statistical Hypotheses*, 2nd Edition. Wiley. MR0852406

[18] Liaw, A. and Wiener, M. (2002). Classification and regression by `randomForest`. *R News* **2** (3) 18–22.

[19] Lunceford, J.K. and Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Statistics in Medicine* **23** (1) 2937–2960.

[20] Pearl, J. (2009). *Causality: Models, Reasoning and Inference*, 2nd ed. Cambridge University Press. MR2548166

[21] Pearl, J. (2009). Causal inference in statistics: an overview. *Statistics Surveys* **3** 96–146. MR2545291

[22] Pearl, J. (2009). Myth, confusion, and science in causal analysis. *Technical Report* R-348, University of California, Los Angeles, CA.

[23] R Core Team (2014). `R`: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* Vienna, Austria. http://www.R-project.org

[24] Rice, J.A. (1995). *Mathematical Statistics and Data Analysis*, 2nd ed. Duxbury Press.

[25] Robins, J. (1986). A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling* **7** (9-12) 1393-1512. MR0877758

[26] Rosenbaum, P.R. (2002). *Observational Studies*, 2nd ed. Springer. MR1899138

[27] Rosenbaum, P.R. (2010). *Design of Observational Studies.* Springer. MR2561612

[28] Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974

[29] Rosenbaum, P.R. and Rubin, D.B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79** (387) 516–524.

[30] Rothman, K.J. (1986). *Modern Epidemiology.* Little, Brown & Co.: Boston/Toronto.

[31] Rothman, K.J. (2002). *Epidemiology: An Introduction.* Oxford University Press.

[32] Rousseauw, J., du Plessis, J., Benade, A., Jordaan, P., Kotze, J., Ferreira, J.(1983). Coronary risk factor screening in three rural communities. *South African Medical Journal* **64** (1) 430–436.

[33] Rubin, D.B. (2006) *Matched Sampling for Causal Effects.* Cambridge University Press. MR2307965

[34] Scholz, F. and Zhu, A. (2012). `kSamples`: K-sample rank tests and their combinations. `R` package: http://CRAN.R-project.org/package=kSamples.

[35] Shrier, I. (2009). Letter to the editor. *Statistics in Medicine* **27** 2740–2741. MR2440067
[36] Silvey, S.D. (1975). *Statistical Inference*. Halsted Press. MR0381045
[37] Snowden, J.M., Rose S. and Mortimer, K.M. (2011). Demonstration of G-computation on simulated data: demonstration of a causal inference technique. *American Journal of Epidemiology* **173** (71) 731–738.
[38] Stuart, E.A.(2010). Matching methods for causal inference: a review and a look forward. *Statistical Science* **25** (1) 1–21. MR2741812
[39] van der Vaart, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press. MR1652247