Health Serv Outcomes Res Method (2011) 11:27–43
DOI 10.1007/s10742-011-0070-x

Well-being and obesity of rheumatoid arthritis patients

Nicholas T. Longford · Catia Nicodemo · Montserrat Núñez · Esther Núñez

Received: 20 May 2010 / Revised: 25 October 2010 / Accepted: 24 January 2011 /
Published online: 11 February 2011

Abstract We apply the potential outcomes framework in the analysis of an observational study of rheumatoid arthritis patients, in which we compare the mean functional-health and well-being scores (SF–36) of patients who are overweight and who are not. We combine propensity score matching with multiple imputation for nonresponse. We assess the sensitivity of the conclusions with respect to the details of the propensity model and the definition of being overweight.

Keywords Body mass index · Multiple imputation · Potential outcomes · Propensity score matching · Rheumatoid arthritis

1 Introduction

Many phenomena in health care can be explored only in observational studies, because the values or levels of attributes we want to compare cannot be assigned to subjects by design.

Electronic supplementary material The online version of this article (doi:10.1007/s10742-011-0070-x) contains supplementary material, which is available to authorized users.

N. T. Longford (✉)
SNTL and Departament d'Economia i Empresa, Universitat Pompeu Fabra,
Ramon Trias Fargas 25–27, 08005 Barcelona, Spain
e-mail: NTL@sntl.co.uk

C. Nicodemo
Universitat Autònoma de Barcelona, Barcelona, Spain

C. Nicodemo
IZA, Frankfurt am Main, Germany

M. Núñez
Hospital Clínic, Universitat de Barcelona, Barcelona, Spain

E. Núñez
Institut Català de la Salut, Barcelona, Spain

For example, the quality of life of overweight and normal-weight patients can be compared either without any adjustment or by a regression in which we include covariates we deem or empirically find to be relevant. The former comparison is tainted by imbalance of the groups of patients with respect to the covariates (background information), and with both approaches we are likely to encounter various conflicts with (model) assumptions, such as normality and homoscedasticity. Another problem is that some patients in one or both groups have configurations of values of the covariates that are extreme and their records exercise unduly strong influence on the regression fit. Yet, such patients are least relevant to the comparisons we want to make.

In the extreme case, if there were only one relevant covariate and the two groups of patients had completely separated values of this covariate (if every subject in one group had values smaller than R and all other subjects values greater than R), then the two groups could be compared only after making some very strong modelling assumptions that might be difficult to confirm empirically. Any meaningful comparison should rely on the subgroups of patients that have similar profiles of the covariates. This intuition is supported by Rosenbaum and Rubin (1983), who proved that the propensity score is the coarsest balancing score for two groups. That is, if there are several covariates, so that the balance of the values for the two groups is difficult to arrange by selection based on the values of these covariates, it suffices to match the two groups on their (univariate) propensity scores. With such scores, the extent of balance is easy to assess and the lack of overlap easy to diagnose, although the only solution to an imbalance is to seek a more complex model for the propensity scores.

Propensity scores are commonly applied in causal analysis of observational studies (Rosenbaum 2002; Rubin 2005), because the balancing is essential for making comparisons of like with like. Balancing is no less important when the variable concerned is not associated with a cause. Being overweight or not cannot be regarded as a cause (treatment), because its value (Yes or No) could not be assigned (controlled) by design in any realistic setting. Nevertheless, it is meaningful to pose the question whether overweight patients would benefit more from a specific health-care treatment (for rheumatoid arthritis, RA) if their weights were normal, and some form of adjustment for the different profiles of the covariates is essential.

In the potential outcomes framework, each subject observed in a study is associated with two outcome variables; in our setting, one variable is for the normal-weight version and the other for the overweight version of the subject. Only the version that corresponds to the subject's genuine weight status is observed. Information about the outcome for the other status is sought by matching the subject with one who has the other weight status and has as similar a profile on the background variables as can be arranged. Propensity scoring is one method for arranging such matching (pairing).

In this article, we compare the functional health and well-being scores of overweight and normal-weight patients in ARQUALIS, a multicentre study of the quality of life of patients with RA conducted in five general practices and eight hospitals in Catalonia, Spain, between Aug 2004 and Jan 2005; Núñez et al. (2009). The study collected data on 812 patients about their background, height and body mass, from which their body mass index (BMI) is calculated, and responses to the SF–36 questionnaire (see http://www.sf-36.org), which are coded as scores on four variables each related to physical and mental health. They are composites of between two and ten questionnaire items. The variables and sample information related to them are listed in Table 1. In the analysis, we treat them as outcome variables.

Table 1 The SF–36 scales

	Scale (component)	Items	Missing values	Sample mean	Standard deviation	Zero score	Score 100	Unique values
Physical health								
PF	Physical functioning	10	20	46.2	27.1	36	13	25
RP	Physical role	4	24	44.0	43.2	316	240	5
BP	Bodily pain	2	22	43.2	26.6	55	46	24
GH	General health	5	24	40.9	18.9	9	2	36
Mental health								
VT	Vitality	4	21	42.9	24.3	40	10	21
SF	Social functioning	2	20	64.1	30.0	35	194	9
RE	Emotional role	3	23	61.3	44.7	239	416	4
MH	Mental health	5	23	59.3	23.4	4	29	28

Note: The questionnaire, comprising 36 items, contains one item that contributes to none of the scores

As indicated in the table, the scores are not recorded for 20–24 patients for each component. All eight scores are missing for 19 patients, and further 11 patients have some scores missing, 25 scores in total. We deal with this nonresponse by multiple imputation (MI); details are given in Sect. 3. As a variable, each of the eight scores has values in the range 0–100, although the number of possible values is limited, especially for Physical role, `RP`, and Emotional role, `RE`, which have only five (0, 25, 50, 75, and 100) and four possible values (0, 33.3, 66.7 and 100), respectively. The right-most column gives the number of distinct values that occur in the data for each outcome variable. The preceding two columns give the numbers of subjects with extreme scores, 0 and 100, to indicate the coarse and truncated nature of the variables. For each variable, higher score is associated with more desired states (fewer problems, better condition, less pain, and the like).

The background information is summarised in Table 2 which lists the variables considered in the analysis and gives details of their definitions. Except for age, all the variables are categorical. The originally recorded categories have been recoded so that none of the categories contains very few subjects. Further details of the variables available from Online Resource 1 (Electronic Supplementary Material). The propensity score models referred to in the right-most column of Table 2 are explained in Sect. 4.

2 Matched pairs and missing data

For the task of comparing overweight and normal-weight patients, we apply the potential outcomes framework (Holland 1986; Rubin 2005). For examples and arguments in favour of this framework over adjustment by regression for causal analysis, see Rubin (2006). A detailed exploration of these and related methods is made by Kurth et al. (2006).

Suppose each patient in a study *could* be subjected to at most one of the treatments denoted by u and w. We associate each patient i with potential outcomes $y_i^{(u)}$ and $y_i^{(w)}$ that would be recorded after the application of the respective treatments u and w. We qualify these outcomes as potential because only one of them can be realised. In the framework, these pairs of outcomes are fixed; that is, in a replication of the study with the same set of patients, but a different assignment of patients to treatments, the same pair of potential

Table 2 Background variables in the ARQUALIS study

	Variable	Categories	Frequencies	Missing values	In propensity models
Edu	Education (none/basic/sec. or higher)	3 (4)	106 + 342 + 347	17	ABC
MSt	Marital status (married/not married)	2 (4)	258 + 547	7	C
DoC	Domestic chores (doing some/doing none)	2 (2)	610 + 194	8	ABC
Occ	Occupation (at home/at work)	2 (12)	344 + 457	11	CD
Sit	Situation (econ. inactive/at work or unempl.)	2 (7)	627 + 169	16	ABCD
Opr	Other problems & chronic conditions (none/at least one)	2 (2 + 2)	396 + 407	9	ABCD
Inc	Income (≤1500/ >1500 Euro p. month)	2 (5)	425 + 332	55	A CD
Ind	Independence (income not sufficient/ sufficient)	2 (5)	397 + 383	32	ABC
EvA	Everyday assistance (not needed/ needed)	2 (3)	425 + 374	13	ABCD
Sat	Satisfaction with the assistance (less/very much)	2 (5)	466 + 337	9	CD
SEf	Secondary effects (none/some)	2 (2)	215 + 585	12	CD
Stf	Stiffness (none/some)	2 (5)	286 + 508	18	C
Art	Articulation (no joints affected/some joints affected)	2 (# + # + 2)	103 + 693	16	CD
Exc	Exercise (none/some)	2 (2)	226 + 574	12	CD
Cmb	Comorbidity (some/none)	2 (5)	548 + 264	0	CD
FuC	Functional class (moderate or none/substantial)	2 (4)	542 + 206	64	ABCD
Vs1	Symptoms 1 (none/RA factor positive or erosive)	2 (4 + 4)	401 + 405	10	CD
Vs2	Symptoms 2 (none/some)	2 (4 + 4)	536 + 265	11	C
Age	Age (in years)	–	19–89	15	ABC
Sex	Sex (M/F)	2 (2)	172 + 633	7	CD

Notes: The numbers of categories of the (original) variables from which the variables used in the analysis were defined by recoding are given in parentheses. The frequencies are given for the categories of the listed variable; the first number is for the reference category. Where a summation is given in parentheses, e.g., for Opr, it indicates that the variable is defined from several categorical variables, with their numbers of categories given as the summands. For example, Opr is defined from two categorical variables, each with two categories. # indicates counts. Age is the only continuous background variable; the range of values is given in the 'Frequencies' column. The right-most column indicates inclusion of the variable in the model for propensity scoring. The models A–D are referred to in Sects. 4 and 5

outcomes would apply for each subject, and exactly one of the pair would be realised, depending on the treatment assigned, and applied. In our analysis, this assumption has an entirely technical nature because the weight status of a subject cannot be manipulated (controlled). The essential element of the framework is the selection of a subsample for which there is a (near) balance of all the observed covariates across the two groups. This is accomplished by forming matched pairs of subjects, one subject from each treatment group in every pair. The matching is based on estimated propensity scores.

If the values of both outcomes were available, the treatment effect would be evaluated straightforwardly for both a subject, as $\theta_i = y_i^{(u)} - y_i^{(w)}$, and the entire study, as

$$\theta = \frac{1}{n}\sum_{i=1}^{n}\left(y_i^{(u)} - y_i^{(w)}\right), \tag{1}$$

where n is the relevant number of subjects. In our study, we ask whether the patients who are overweight would have higher scores (on average) if they had normal weight. Therefore, the average in (1) is over the overweight patients in the study. We refer to them as the focal group, and to patients with normal weight as the reference group. We do not assume that the individual-level differences $\theta_i = y_i^{(u)} - y_i^{(w)}$ are constant, so the mean difference (weight-status effect) θ depends on the group to which the average in (1) is applied. For this, we choose the patients classified as overweight. Instead of taking differences, another way of comparing values could be used. For example, the ratios $y_i^{(u)}/y_i^{(w)}$, the log-differences $\log(y_i^{(u)}) - \log(y_i^{(w)})$ or the differences on another scale could be averaged. Also, the median or another measure of the central tendency could be used instead of the mean.

We address the problem of not having observed both potential outcomes by treating it as a missing-data problem (Little and Rubin 2002; Rubin 2002; Longford 2008, Chap. 7). In its terminology, the unrealised outcomes $y_i^{(T)}$ (T = u or w, one per patient) comprise the missing data, and its union with the realised (incomplete) data, $2n$ values $y_i^{(T)}$, is the complete data. The hypothetical evaluation of (1) with the complete data is called the complete-data analysis. It is associated with no variation because θ would be evaluated with precision *if* the complete data, the two potential outcomes for each patient, were available.

The qualifiers 'missing', 'complete' and 'incomplete' are related to the problem at hand. We use MI for two problems, potential outcomes and nonresponse (missing values in the narrow sense). To avoid confusion, we add the prefix P or R to these qualifiers. Thus, the R-complete data have no missing values on any of the covariates, but have the outcomes recorded only for the realised weight status of each subject, while the P-complete data have the outcomes recorded for both weight states. By the prefix RP we refer to the combination of the problems (and solutions), dealing with nonresponse (R) first and then with potential outcomes (P).

We use propensity scoring to find matches for the focal patients. The propensity score is defined as the conditional probability of being in the focal group (receiving the focal treatment) given the background variables, $p(\mathbf{x}) = \mathrm{P}(T = \mathrm{w}|\mathbf{X} = \mathbf{x})$. Monotone transformations of the propensity score are for all purposes equivalent; we prefer to use the logit of $p(\mathbf{x})$. We estimate the mean weight-status effect θ from multiple sets of matched pairs of focal and reference patients. We fit a logistic regression model for the weight status in terms of the background variables. The outcome variables have no role in this model. The model fit defines a set of estimated propensity scores $s_i = \mathbf{x}_i\hat{\boldsymbol{\beta}}$, which are used for pairing each focal patient with a reference patient. This is done by setting cutpoints $-\infty = c_0 < c_1 < \cdots < c_K = +\infty$, which define K bands (c_{k-1}, c_k), and pairing each focal patient with score $s_i \in (c_{k-1}, c_k)$ with a reference patient in the same band. The pair (match) is selected at random, without replacement. This process is replicated M times, yielding M sets of matched pairs. As few as $K = 6$ bands may suffice because the propensity score categorized in this way retains most of its original variation (Rubin 2005).

A small advantage is gained by setting K higher, so long as each band contains enough units to represent the within-band variation in the background variables.

Estimation of the propensity scores entails some uncertainty; Rubin and Thomas (1992, 1996) show that this source of uncertainty can be ignored; in fact, matching on the estimated propensity scores, treating them as fixed is preferred. Setting the nonresponse problem aside, we apply the P-complete-data analysis given by (1) to each set of matched pairs, obtaining replicate P-completed-data estimates $\hat{\theta}^{(1)}, \ldots, \hat{\theta}^{(M)}$. The P-multiple-imputation (P-MI) estimator $\tilde{\theta}_{\mathrm{MI}}$ is equal to their average, $\tilde{\theta}_{\mathrm{MI}} = (\hat{\theta}^{(1)} + \cdots + \hat{\theta}^{(M)})/M$, and its sampling variance is estimated by

$$(\tilde{\sigma}^2 =)\widehat{\mathrm{var}}\left(\tilde{\theta}\right) = \frac{M+1}{M(M-1)} \sum_{m=1}^{M} \left(\hat{\theta}^{(m)} - \tilde{\theta}_{\mathrm{MI}}\right)^2. \tag{2}$$

Details of propensity score matching are given in Sect. 4. The theory in Rubin (2002) suggests that as few as five replicate sets suffice for (approximately) unbiased and nearly efficient estimation of θ. However, a much greater number is required for reliable estimation of the sampling variance of the P-MI estimator of θ. Assuming that $\hat{\theta}^{(m)}$ are normally distributed, $\tilde{\sigma}^2$ has a scaled χ^2_M distribution with expectation $\sigma^2 = \mathrm{var}(\tilde{\theta})$ and standard deviation $\sigma^2\sqrt{2/M}$. Without an upper bound on σ^2 known in advance we cannot identify a suitable number M.

3 R-imputation

As indicated in Tables 1 and 2, a small number of values is missing for the outcome and background variables (up to 8%). By listwise deletion we would discard the data for several patients, most of whom have only a few values missing. We avoid this by integrating MI for these missing values within the process of propensity score matching. Simply, we generate an (R-)completion of the set of background and outcome variables and a set of plausible matches for the focal group based on this completion (imputation). For R-completion, we assume that the values are missing at random (MAR). This assumption is difficult to verify or test. In our case, only a small fraction of the values is missing for any variable, and the main purpose of R-imputation is to include in the analysis all the focal patients. Apart from this inclusion, the impact of the imputed values (vis-à-vis listwise deletion) on the results is likely to be negligible.

We do not seek a completion for BMI, on which the classification of the patients as focal and reference is based, because we prefer to have a fixed focal group in the (replicate) RP-completed-data analyses. The value of BMI is calculated from the patient's height and body mass. It is missing for 29 patients; 27 of them have missing values for both height and body mass and two patients have their body mass recorded, but no height. Despite having a rich set of background variables, BMI cannot be predicted from them with any precision. In summary, $821 - 29 = 792$ subjects contribute to the analysis.

Multiple sets of plausible values for the (R-)missing data in the background and outcome variables are generated by the method of chained equations (van Buuren et al. 1999). First, provisional values are imputed for all the missing items. Replacements for the (originally) missing values of one variable are generated as plausible values from the fit of the observed values of this variable on all the other variables with the observed or provisionally imputed values. Plausible values are then generated for each variable in turn,

using the observed and (provisionally) imputed values for all the other variables as covariates. For a continuous variable, we use ordinary regression and for a categorical variable logistic regression, with its multinomial adaptation when the variable has more than two categories. The outcomes `RP` and `RE` are regarded as multinomial variables. We prefer not to take the ordering of their categories into account because of the special status of the extreme scores 0 and 100.

Generating one set of replacements for each variable is referred to as a cycle. The cycles are repeated until the distribution of the missing values is sufficiently close to stationarity. With only a small number of values missing for each variable, it is difficult to judge when this convergence is attained. We draw on our experience from other applications, which suggests that ten cycles usually suffice; van Buuren and Groothuis-Oudshoorn (2011; Sect. 2.3) give similar advice. We prefer to err on the side of more cycles and use 20 cycles throughout. They yield one set of plausible values for the missing items. Other sets are generated by replications.

The plausible values do not necessarily satisfy the same conditions as do the variables for which they are intended. Thus, the plausible values for an outcome variable may be outside the range (0, 100). Such values are truncated. We checked on several sets of plausible values that none of them are substantially smaller than zero or greater than 100. The plausible values for age are in the range 40–70 years with very few exceptions, even though their range in the data, 19–89, is much wider.

4 Propensity scoring—models and matching

The covariates for the propensity model are selected from the list of background variables given in Table 2. We show in Sect. 5 that it is preferable to retain all the background variables, but want to illustrate that a conventional procedure for model reduction is in our case not useful. This is established by simple diagnostic checks for the balance of the background variables across the weight-status groups. However, we show that even with models that inappropriately omit some variables the results are not altered substantially.

In model reduction, we use listwise deletion specific for the model, but check the appropriateness of the selection on several completions of the background information. We exclude variables in stages and very conservatively, not more than two at a time, and monitor the increase in the value of the deviance (–2 log-likelihood) in the process. In each instance, we exclude the variable(s) with the smallest value(s) of the absolute t ratio, $|\hat{\beta}/\widehat{\mathrm{sd}}(\hat{\beta})|$.

In one particular analysis we classify a patient as overweight if his or her BMI is greater than 32 kg/m^2 (79 patients), and as normal-weight if BMI < 30 kg/m^2 (632 patients). The 72 patients with BMI in the range (30, 32) are excluded from the analysis, so as to have a clear separation between the two groups; the values of BMI for the remaining ($812 - 79 - 72 - 632 =$) 29 patients are missing. In the model selection for this variable, we excluded in six steps the following background variables: `MSt` and `Cmb`, increasing the deviance from 362.156 to 362.162, followed by `Stf` and `Vs1` (deviance 363.90), `Sat` and `Vs2` (364.86), `Occ` and `SEf` (366.40), `Exc` and `Sex` (368.70), and finally `Art` (370.73). Each deviance in parentheses is reported for the same set of 647 patients for whom all covariates are recorded. The t ratio for every variable that was excluded was smaller than 1.0 in absolute value in the model fit concerned. Note that listwise deletion may yield more observations when some covariates are excluded; then the values of the deviance for a model and its submodel could not be compared.

By conventional criteria, it might be appropriate to exclude also Income (`Inc`); its t ratio in the final model is -1.3. However, exclusion of `Inc` is associated with an increase of the deviance by 3.87 (on the same set of 647 patients). The comparison with the critical values of the χ^2 distribution with one degree of freedom suggests that `Inc` should be retained in the model. As a form of sensitivity analysis, we consider both models, with and without `Inc`, denoted respectively by A and B in Table 2, and carry out each analysis with fitted propensity scores generated by both models. The two model fits are displayed in Table 3. In addition, we consider the model with all the background variables (model C).

Other methods of model selection may conclude with different models, but are equally mis-applied because the purpose of the propensity score modelling is neither attainment of a good fit nor efficient prediction or estimation of certain parameters, but solely the forming of sets of matched pairs which are balanced with respect to the covariates. Rubin (2005, 2006) emphasises that we should be liberal with the inclusion of covariates and their interactions in the propensity score model, and be satisfied with a model only when approximate balance has been achieved, as it would be in a study with randomisation.

The histograms of the fitted (estimated) propensity scores are drawn in Fig. 1 separately for the two groups, for the model that includes `Inc` (scores A) at the top and for the model without `Inc` (scores B) in the middle row. The vertical dashes are drawn at the deciles of the scores. The deciles are used for matching; see below. The lines highlight the uneven distribution of the propensity scores in the two groups; scores in the bottom decile are rare in the focal group and scores in the top decile are rare in the reference group. Both groups are represented in each decile, so the reference group has suitable matches for most (or maybe even all) observations in the focal group. In principle, a decile band may contain no subjects from a group; then the subjects in this band cannot be involved in any matched pairs.

The bottom two panels plot the sets of propensity scores derived by the two models, to indicate how much the fitted scores depend on the selected model. There appear to be two bands of points parallel with the identity line (dots). Income, the variable omitted from the propensity model B, is only weakly associated with the division into these two bands. The horizontal and vertical ticks are drawn at the deciles of the scores for (the union of) the two groups. The $(79 + 632 =)$ 711 pairs of propensity scores are in the same respective decile band for 395 patients (56%), in neighbouring bands for 271 patients (38%), are two bands apart for 44 patients (6%), and three bands apart in one case. In Sect. 5, we explore the impact of this difference on the estimate of the mean treatment effect θ. The tabulation of the differences is for a particular R-completion of the dataset. A different R-completion (imputation) yields a slightly different tabulation.

Table 3 Model fits for the propensity scores for distinguishing between patients with BMI > 32 kg/m^2 and those with BMI < 30 kg/m^2

	1	Edu.2	Edu.3	DoC	Sit	Opr	Ind	EvA	FuC	Age	Inc	Deviance
Estimate	−0.56	0.88	0.40	0.34	−0.82	−0.58	0.55	1.03	−0.74	−0.031	−0.52	370.73
St. error	1.02	0.56	0.59	0.32	0.43	0.29	0.29	0.32	0.34	0.011	0.40	
Without `Inc`												
Estimate	−0.54	0.82	0.34	0.37	−0.95	−0.63	0.43	0.98	−0.75	−0.029	0	374.57
St. error	1.01	0.56	0.58	0.31	0.42	0.29	0.28	0.31	0.34	0.011	0	

The second column, marked 1, corresponds to the intercept. `Edu.2` and `Edu.3` indicate the respective contrasts of the categories 2–1 and 3–1 of Education

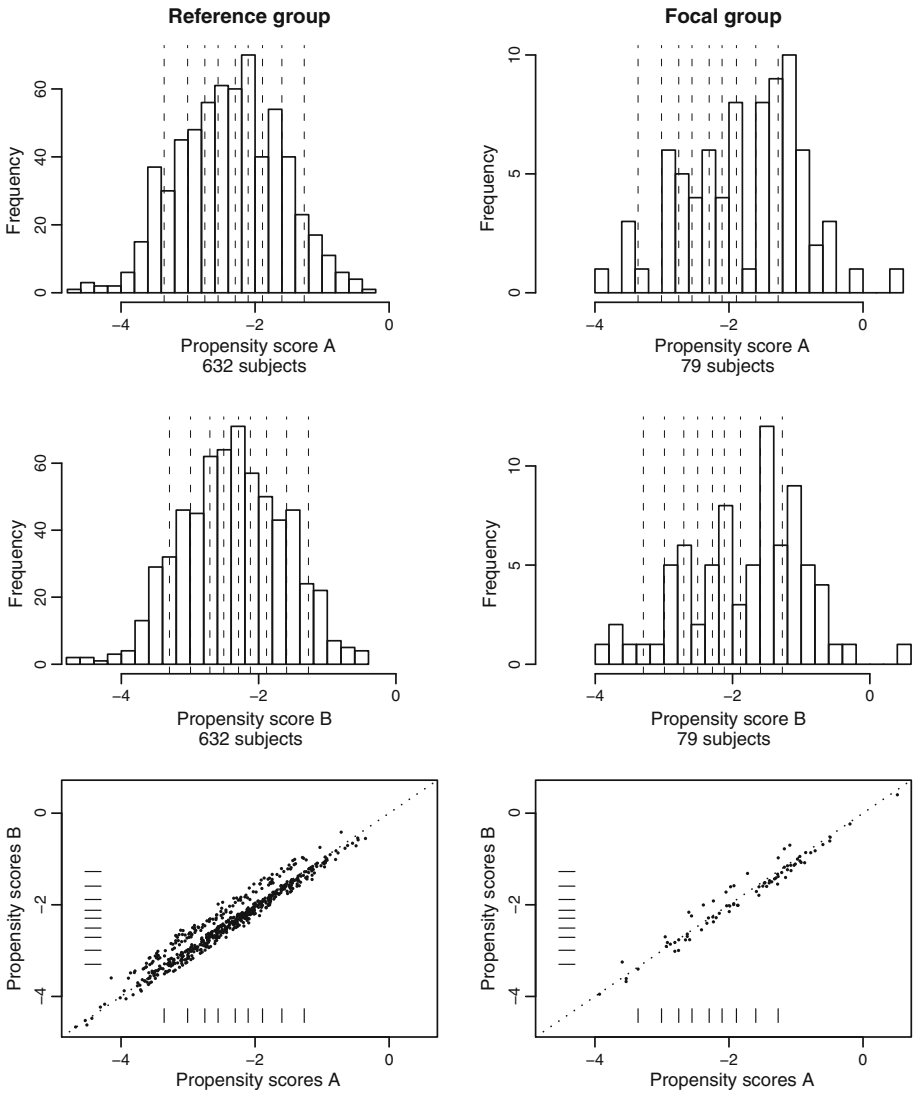

Fig. 1 The fitted (estimated) propensity scores based on the models with and without `Inc`

For a set of fitted propensity scores, we match each patient in the focal group with a patient in the reference group by the following process. The propensity scores are split into deciles, sets of patients of nearly equal size with scores in disjoint intervals. Within each decile, a patient who belongs to the focal group is matched with a patient from the (more numerous) reference group. The matching within each decile is completely at random and without replacement. If a decile contained more patients from the focal group, the assignment would be made in reverse (matches are found for the patients in the reference group), and some patients in the focal group would not be matched. To check that failure to find a match has a very small probability, we replicated the matching process 100 times, in addition to the $M = 25$ replications used in the analysis. A match was assigned to every

patient in the focal group in every one of the 125 replications. With a different definition of weight status, 151 patients, with BMI in excess of 30 kg/m^2, are classified as overweight and 518 patients have BMI smaller than 28 kg/m^2 (normal weight). There are sufficient matches for all the focal patients in the first nine decile bands, but in the highest band there are 10–13 more focal than reference patients, and so the number of pairs in the analysis is around 140.

The sets of matched pairs define the plausible (RP-completed) datasets, and each of them yields a RP-completed-data estimate of the mean treatment effect by applying (1). The RP-MI estimate of the treatment effect is defined as the average of these plausible estimates, and its sampling variance is estimated by (2).

5 Results

In the analysis, we compare patients classified as overweight with selected subsets of patients whose weight is classified as normal. The comparison is not symmetric, because we estimate the mean difference of the scores of overweight patients with the scores they would have if their weights were normal. That is, the mean treatment effect depends on the focal group. If we designate the normal-weight patients as the focal group (and ask how much lower their scores would be on average if they were overweight), then the mean treatment effect will not be equal to the negative of the mean treatment effect of the patients who are overweight.

The body mass index (BMI) is universally regarded as a suitable adjustment of the body mass of adults for height, but there is no agreed threshold which would separate overweight and normal-weight subjects. We address this issue by defining several separations and report how similar are the conclusions they yield. We are also concerned about the appropriateness of the model for propensity scoring, and so we consider the three models, A (selected by stepwise deletion, without `Inc`), B (as A, with `Inc` included) and C (all covariates), described in Sect. 4 and indicated in Table 2.

We give the details for the setting in which being overweight corresponds to BMI greater than 32 kg/m^2, and normal weight to BMI smaller than 30 kg/m^2. Patients with BMI between 30 and 32 kg/m^2 play no part in the analysis. Separate analyses are conducted for the eight outcome variables listed in Table 1. Each analysis is based on the same three sets of $M = 25$ sets of matched pairs. The three sets are generated using the models A–C and each set comprises 79 pairs. The 79 matches in each of them are selected from 632 normal-weight subjects.

First we check the extent of the balance of the two groups with respect to the background variables. This is summarised in the left-hand panel of Fig. 2. For each variable, the horizontal segment connects the sample difference of proportions with its negative, which corresponds to the same extent of imbalance. The lowercase symbols, a, b and c in the left-hand panel, mark the differences in the proportions for the last completed dataset for the respective propensity models A, B and C. For example, the difference in the proportions of women among the overweight and normal-weight in the 25th replicate R-completed dataset is 0.071, and the differences in the sets of matched pairs formed from this dataset are 0.025, 0.013 and –0.025 using the respective models A – C. For a few variables, the mean absolute differences in the set of matched pairs are greater than in the sample. This is not a sign of failure of the matching procedure, because with relatively few matched pairs, 79, such deviations can arise even when the underlying probabilities are equal. These probabilities are the average propensities (the propensity scores transformed to the

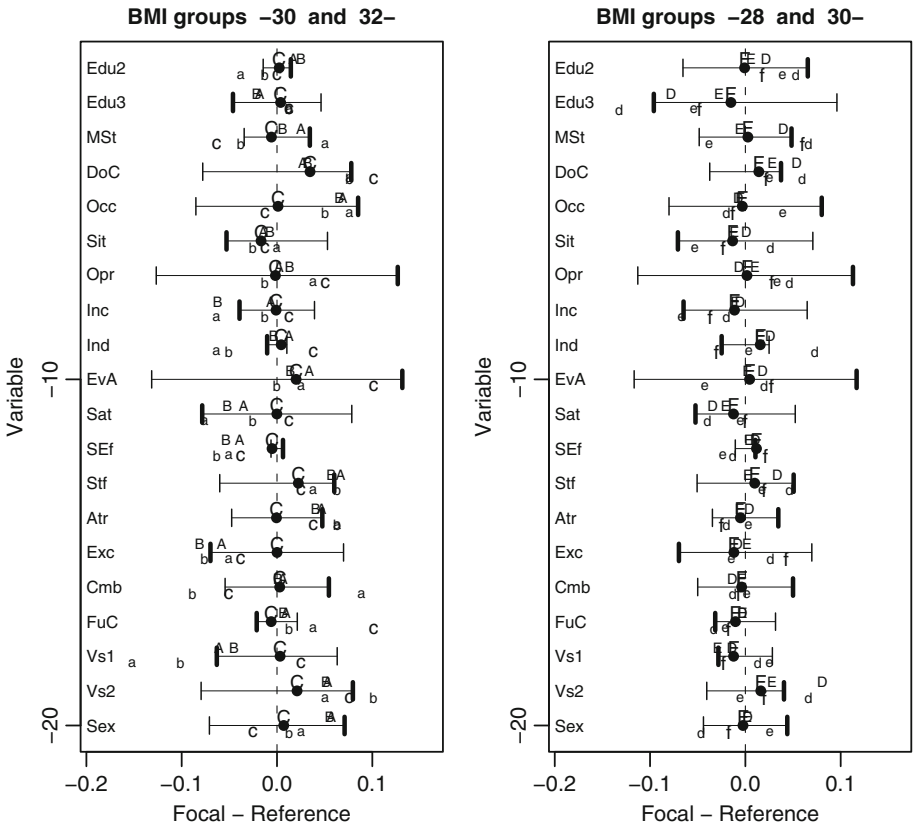

Fig. 2 The differences of the sample and matched-pairs proportions of the categorical background variables between the focal and reference patients. The *thick vertical ticks* indicate the differences of the proportions in a completed dataset and the *thin ticks* their negatives. The *capitals* mark the differences of the within-group mean propensities and the *lowercases* the mean differences in a set of matched pairs. The *black discs* mark the adopted models, *C* and *F*

probability scale), marked in the diagram by the symbols A, B and C, and the latter also by black discs. For `Sex`, these probabilities are 0.057, 0.055 and 0.007 for the respective models A–C; the imbalance with C is only slight. The balance of the probabilities is near perfect for model C, and for most variables it is superior to A and B. Thus, the matched pairs select a subsample that has the properties very close to what randomisation would arrange by design. Note that the conventional model selection would yield an inferior model A (or B); the richer model C attains a better balance of the groups.

The right-hand panel compares three propensity score models for the weight-status groups defined by the bounds of 28 and 30 kg/m^2 using a parsimonious model, D, the model with all the covariates, E (the same as model C, but used with a different assignment to groups), and model F, in which the interactions of `Exc` with `FuC` and of `Opr` with `Sat` are added to all the covariates. These two interactions have been identified by trial and error, seeking balance better than achieved by model E. The principal candidates for the interactions are variables with large estimated coefficients in model E. As the propensity model is altered, the improvement (or deterioration) of the balance is not uniform, so a judgement has to be made in the model selection.

The balance achieved for Age, the only continuous background variable, is checked similarly, but we inspect both the means and standard deviations. For a particular R-completed dataset, the sample means (standard deviations) are 58.29 (13.49) in the focal and 60.52 (14.54) in the reference group, so their differences are −2.23 and −1.05. The mean and standard deviation of the subsample of the reference group involved in the matched pairs is 58.82 (14.05), much closer to the summaries for the focal group. The sets of matched pairs are much closer to balance than the original sample or its completion.

The sets of estimates and standard errors for the mean effects are displayed in Tables 4 and 5 for physical and mental health, respectively. The tables show that we have evidence of a positive mean difference for all four physical-health and two mental-health variables (VT and SF) for the weight status defined by the limits of 30 and 32 kg/m^2. For the weight status defined by 28 and 30 kg/m^2, we have failed to find evidence of a positive mean difference also for RP. These statements are based on the propensity models C and F.

The results with the other models are displayed to illustrate the impact of the imbalance in the covariates. The differences (model A or B *vs*. C) are modest for most outcomes; they

Table 4 Estimates and standard errors (in parentheses) for the differences on SF–36 (*physical health*) between overweight and normal-weight patients; based on 25 replicate matched sets (C_{100} and F_{100}—100 matched sets)

Propensity model	PF	RP	BP	GH
	Normal weight: <30 kg/m^2		Overweight: >32 kg/m^2	
A	11.03 (2.45)	8.19 (3.19)	5.39 (2.50)	4.87 (1.92)
B	11.02 (2.59)	8.64 (3.91)	5.37 (2.64)	5.13 (1.61)
C	8.70 (2.43)	7.89 (3.60)	4.78 (2.29)	4.05 (1.79)
C_{100}	9.15 (2.56)	8.31 (3.88)	4.93 (2.47)	4.00 (1.61)
	Normal weight: <28 kg/m^2		Overweight: >30 kg/m^2	
D	8.94 (1.80)	5.17 (2.01)	4.59 (1.61)	4.02 (1.02)
E	8.28 (1.40)	5.19 (2.86)	4.15 (1.44)	3.29 (1.11)
F	8.65 (1.78)	3.77 (2.69)	3.08 (1.51)	2.92 (1.31)
F_{100}	8.67 (1.92)	4.03 (2.97)	2.67 (1.77)	3.00 (1.36)

Table 5 Estimates and standard errors (in parentheses) for the differences on SF–36 (*mental health*) between overweight and normal-weight patients; based on 25 replicate matched sets (C_{100} and F_{100}—100 matched sets)

Propensity model	VT	SF	RE	MH
	Normal weight: <30 kg/m^2		Overweight: >32 kg/m^2	
A	7.28 (2.08)	8.29 (2.60)	4.40 (4.22)	2.44 (2.20)
B	8.36 (2.09)	9.20 (2.38)	6.93 (3.87)	4.07 (1.79)
C	6.66 (1.97)	8.52 (2.91)	4.68 (4.13)	1.97 (1.90)
C_{100}	6.93 (2.03)	8.35 (2.82)	4.70 (3.81)	2.43 (1.85)
	Normal weight: <28 kg/m^2		Overweight: >30 kg/m^2	
D	6.06 (1.20)	8.41 (1.37)	5.06 (2.56)	3.46 (1.35)
E	5.55 (1.56)	8.16 (1.82)	4.21 (3.44)	2.16 (1.33)
F	4.80 (1.34)	7.45 (1.74)	3.78 (2.70)	2.15 (1.39)
F_{100}	5.19 (1.40)	7.82 (1.71)	4.21 (2.94)	2.52 (1.43)

are largest for PF, RE and MH. They happen not to alter any statements about the significance of the outcome variables for either weight-status variable.

In our perspective, the values of the eight mean status effects (but not their estimates) are determined by the overweight patients in the study who have fixed (patient-level) status effects. Estimation of the mean status effect entails uncertainty due to the randomness of the matching process and the ambiguity about the threshold values of BMI for the normal-weight and overweight status. We ignore the uncertainty about the propensity model, because its purpose is solely to arrange a balance of the two groups; we do not pursue the goal of matching on the underlying propensity scores associated with an ideal model. Sample selection as a source of uncertainty is discussed in Sect. 6.

Insights into the relative sizes of the identified sources of uncertainty can be gained by using alternative definitions of the weight status and a greater number of replicate sets of matched pairs. With the qualification BMI > 32 kg/m^2, we have only 79 overweight patients; ten of them have one missing background item each and two others have three each. They have no missing values on the outcome variables. With the more liberal qualification BMI > 30 kg/m^2, there are 151 focal patients, 125 of them with complete records on all background variables, but 10 – 13 of them are not matched in a P-complete-data analysis, because the top decile band contains too few reference patients. We cannot expect that the results for the two qualifications would coincide, but substantial differences in the results would indicate a lack of robustness of the conclusions with respect to how the weight status is defined.

The results are displayed in the bottom parts of Tables 4 and 5, for models D (see Table 2), E (all covariates) and F (E supplemented with two interactions). The estimated mean difference is altered substantially only for the outcome variable RP (estimates around 8.0 with model C and around 4.0 with model E). For all outcomes except MH, the estimate with model F is smaller than with model C, suggesting that the weight-status effect is diluted with a more liberal definition of the focal group. The estimated standard errors are smaller, but not uniformly $\sqrt{2}\doteq 1.4$ times, as the near-doubling of the size of the focal group would suggest. This is a consequence of having a different focal group that is more heterogeneous in some aspects and less in others, and to a lesser extent due to the stochastic nature of estimation of the standard errors, even with a given (incomplete) dataset.

We explored other reasonable definitions of the weight status, including those with greater and smaller difference between the upper and lower bounds for the normal-weight and overweight status. In most cases, the results deviate from those based on model C less than the results based on model F.

The RP-MI estimates and the associated standard errors are not deterministic functions of the data because they also depend on the plausible (imputed) values and pairings of the patients. The sets of results in Tables 4 and 5 differ because different propensity models and definitions of weight status were applied *and* because of the randomness intrinsic in MI. We can assess the impact of the latter source by repeating the analysis with an independent set of $M = 100$ R-completions and P-pairings. This is computationally not demanding; the three sets of results are obtained using only a few minutes of CPU time (with a customised code in R; R Development Core Team, 2009). This includes model fitting, generating plausible propensity scores, finding matched pairs, and evaluating the estimators for the eight outcome variables. The differences attributable to the propensity models persist, but the extent of randomness due to using a finite number M of replicates is reduced. The results for $M = 100$ are displayed in Tables 4 and 5 in rows C_{100} and F_{100}. These estimates and the associated standard errors differ somewhat from their counterparts obtained with $M = 25$

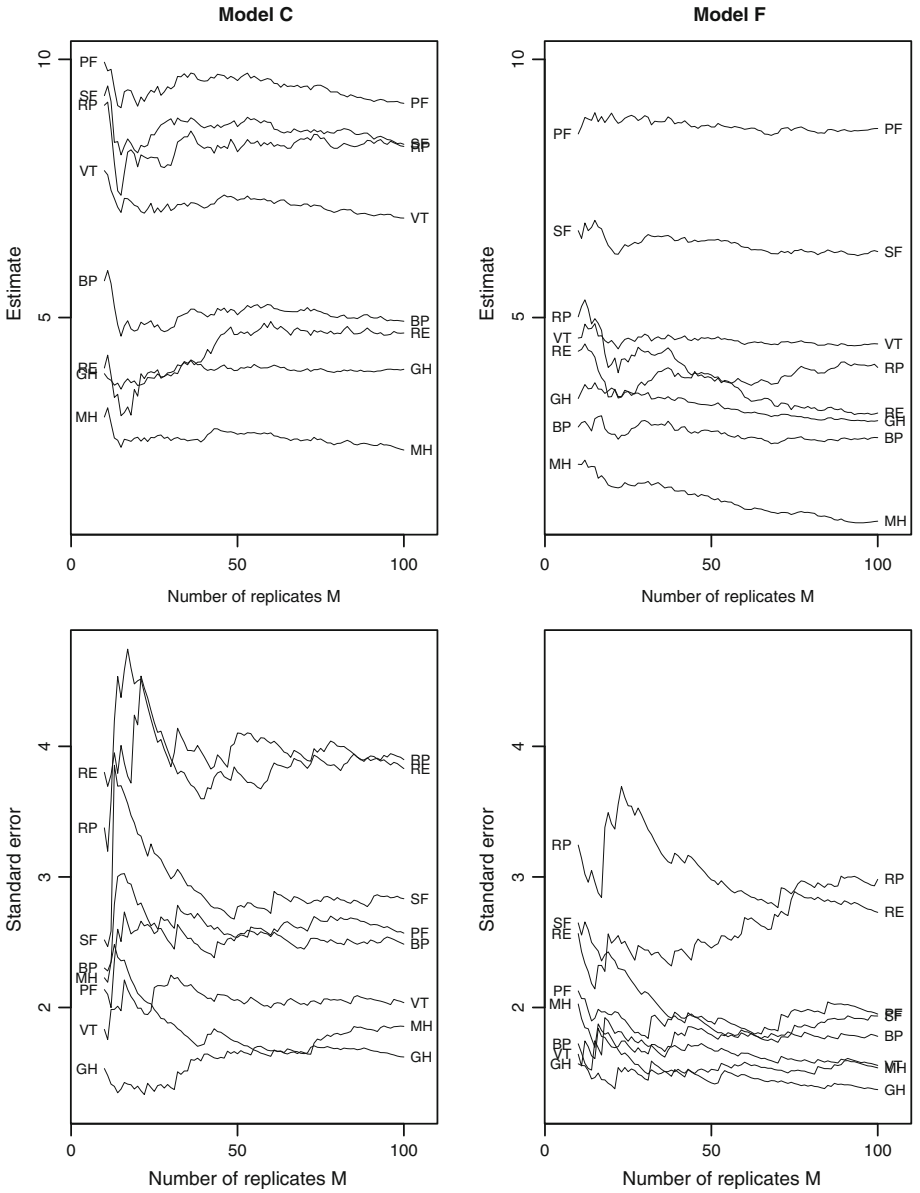

Fig. 3 Stability of the MI estimates and standard errors. Based on the first m completed datasets, $m = 10, 11, \ldots, 100$

replicates, but lead to similar conclusions. To illustrate the stability of the estimates and the estimated standard errors with increasing number of replicates, Fig. 3 displays the MI estimates and estimated standard errors based on the first m completed-datasets for $m = 10, 11, \ldots, 100$. It is obvious that five or ten replications, suggested for application of MI in other contexts, would not suffice for stable estimation of the standard errors, but that 25 replicates is the minimum necessary and 100 is more than sufficient.

6 Discussion and conclusion

We analysed an observational study in which we posed a question about two groups of patients (overweight and normal-weight) and applied the potential outcomes framework to answer it. The framework is devised for causal analysis, but the key feature of its implementation is the formation of (multiple) sets of matched pairs, which is ideal for comparing two treatments even when they are not associated with causes that could, at least in principle, be externally manipulated. The comparisons are not for the two groups as such, but for those who are overweight with their would-be outcomes if their weights were normal.

The results, summarised in Tables 4 and 5, require an assessment of their clinical relevance. Walters and Brazier (2003) suggested that the smallest change that could be regarded as 'clinically and socially relevant' on the scale of GH is by five points. Angst et al. (2001) concluded that the smallest important difference is in the range 3.3–5.3 points for the scale of PF and 7.2–7.8 points for BP in patients with osteoarthritis of the hip or knee. Hays and Morales (2001) stated, with some reservations, that the smallest clinically important difference for the SF–36 scales is 'typically in the range of 3–5 points'. In view of these findings and opinions, the estimates we have obtained correspond to differences that are clinically important and of non-trivial magnitude.

The method of inference based on the potential outcomes framework involves only minimal distributional assumptions. No pattern of the (within-subject) differences of the outcomes is assumed. In contrast, regression-based methods assume that these differences are constant after an appropriate adjustment for background variables. The principal modelling task is related to the propensity scores which characterise the (treatment) assignment process. In this modelling, we assume that given the background variables, used as covariates in the logistic regression for propensity scores, the assignment process is conditionally independent of the outcomes. This is an unverifiable assumption; in our study, we can merely claim that rich background information was collected and that it is implausible that there might be an unobserved background variable that is strongly associated with the assignment process and the outcome, even after accounting for the observed covariates. Recording the relevant background variables in the study is essential for the quality of the analysis.

No assumption of normality (or another distribution) of the outcomes is made. By presenting estimates and estimated standard errors, we indirectly refer to a t distribution. This distribution is for a contrast of means, for which normality is much more palatable than for individual outcomes.

The potential outcomes framework is an application of the missing-data principle. Generality is a strength of this principle; its original application is for dealing with non-response, but measurement error and misclassification, coarse data, inference with complex models that involve random coefficients, are other applications (Longford 2005; Part I). The principle can be seamlessly applied to more than one source of data incompleteness; we applied it to nonresponse combined with having observed only one of the two potential outcomes for each patient.

Sensitivity analysis is an important component of the analysis, exploring the impact of the settings used on the conclusions. In our case, the classification of the patients to being overweight or normal-weight entails some arbitrariness, which we address by estimating the average treatment effect (mean difference of the outcomes between the two groups) for several alternative classifications. Similarly, we explored the impact of using more replicate sets of paired patients. We grouped the propensity scores by their deciles, but explored

how the results would be altered if a coarser or more refined division (into fewer or more intervals) were used. The conclusions are robust with respect to these settings. By conventional criteria, we have found evidence that overweight RA patients tend to have poorer outcomes than if they had normal weights.

In the framework used, these conclusions formally refer to the groups of patients in the study, and their extrapolation to the population at large is contingent on the study patients being a good representation of the relevant population. The patients were recruited into the study at cooperating health-care institutions by a form of quota sampling and subject to informed consent, and so their good representation as a sample from the population of all RA patients (in Catalonia) cannot be uncritically assumed. However, our study contributes to the pool of evidence related to the treatment and management of RA; the mean differences inferred for the studied groups are so large that, at least for some of the outcome variables, they are unlikely to vanish or attain negative values in the population of all RA patients.

Acknowledgements Preparation of this manuscript was supported by the Grant No. SEJ2006–13537 from the Spanish Ministry of Science and Technology. The ARQUALIS Study was supported by the Foundation La Marató TV3 (Grant No. 30510).

References

Angst, F., Aeschlimann, A., Stucki, G.: Smallest detectable and minimal clinically important differences of rehabilitation intervention with their implications for required sample sizes using WOMAC and SF-36 quality of life measurement instruments in patients with osteoarthritis of the lower extremities. Arthritis Rheum. **4**, 384–391 (2001)

Hays, R.D., Morales, L.S.: The RAND-36 measure of health-related quality of life. Ann. Med. **33**, 350–357 (2001)

Holland, P.B.: Statistics and causal inference. J. Am. Stat. Assoc. **81**, 945–970 (1986)

Kurth, T., Walker, A.M., Glynn, R.J., Chan, A.K., Gaziano, J.M., Berger, K., Robins, J.M.: Results of multivariable logistic regression, propensity matching, propensity adjustment and propensity-based weighting under conditions of nonuniform effect. Am. J. Epidemiol. **163**, 262–270 (2006)

Little, R.J.A., Rubin, D.B.: Statistical Analysis with Missing Data, 2nd edn. Wiley, New York (2002)

Longford, N.T.: Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician. Springer-Verlag, New York (2005)

Longford, N.T.: Studying Human Populations. An Advanced Course in Statistics. Springer-Verlag, New York (2008)

Núñez, M., Núñez, E., Sanchez, A., Luis de Val, J., Bonet, M., Roig, D., Muñoz, D., and the ARQUALIS Study Group: Patients' perceptions of health-related quality of life in rheumatoid arthritis. Clin. Rheumatol. **28**, 1157–1165 (2009)

R Development Core Team: `R`: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2009)

Rosenbaum, P.R.: Observational Studies, 2nd edn. Springer-Verlag, New York (2002)

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

Rubin, D.B.: Multiple Imputation for Nonresponse in Surveys, 2nd edn. Wiley, New York (2002)

Rubin, D.B.: Causal inference using potential outcomes: design, modeling, decisions. 2004 Fisher Lecture. J. Am. Stat. Assoc. **100**, 322–331 (2005)

Rubin, D.B.: Matched Sampling for Causal Effects. Wiley, New York (2006)

Rubin, D.B., Thomas, N.: Characterizing the effect of matching using linear propensity score methods with normal distributions. Biometrika **79**, 797–809 (1992)

Rubin, D.B., Thomas, N.: Matching using estimated propensity scores: relating theory to practice. Biometrics **52**, 249–264 (1996)

van Buuren, S., Boshuizen, H.C., Knook, D.L.: Multiple imputation of missing blood pressure covariates in survival analysis. Stat. Med. **18**, 681–694 (1999)

van Buuren, S., Groothuis-Oudshoorn, K.: Multiple imputation by chained equations in R. J. Stat. Softw. **37**, forthcoming (2011)

Walters, S.J., Brazier, J.E.: What is the relationship between the minimally important difference and health state utility values? The case of the SF-6D. Health Qual. Life Outcomes **1**, 4 (2003)