Beth L Stuart, Louise EN Grebel, Christopher C Butler, Kerenza Hood, Theo J M Verheij and Paul Little

Comparison between treatment effects in a randomised controlled trial and an observational study using propensity scores in primary care

Abstract

Background
Although randomised controlled trials (RCTs) are considered 'gold standard' evidence, they are not always feasible or appropriate, and may represent a select population. Observational studies provide a useful alternative to enhance applicability, but results can be biased due to confounding.

Aim
To explore the utility of propensity scores for causal inference in an observational study.

Design and setting
Comparison of the effect of amoxicillin on key outcomes in an international RCT and observational study of lower respiratory tract infections.

Method
Propensity scores were calculated and applied as probability weights in the analyses. The adjusted results were compared with the effects reported in the RCT.

Results
Groups were well balanced in the RCT but significantly imbalanced in the observational study, with evidence of confounding by indication: patients receiving antibiotics tended to be older and more unwell at baseline consultation. In the trial duration of symptoms (hazard ratio 1.06, 95% CI = 0.96 to 1.18) and symptom severity (–0.07, 95% CI = –0.15 to 0.007) did not differ between groups. Weighting by propensity score in the observational study resulted in very similar estimates of effect: duration of symptoms (hazard ratio 1.06, 95% CI = 0.80 to 1.40) and difference for symptom severity (–0.07, 95% CI = –0.34 to 0.20).

Conclusion
The observational study, after conditioning on propensity score, echoed the trial results. Provided that detailed information is available on potential sources of confounding, effects of interventions can probably be assessed reasonably well in observational datasets, allowing them to be more directly compared with the results of RCTs.

Keywords
antibiotics; observational study; primary health care; propensity score; randomised controlled trial; respiratory tract infection.

INTRODUCTION

Randomised controlled trials (RCTs) are considered to be the 'gold standard' study design for identifying the true effects of an intervention. However, RCTs may suffer from selection bias. This may be due to extensive exclusion and inclusion criteria, or because patients who decline to be randomised differ systematically from those who accept randomisation.[1] The treatment effects observed in a trial context therefore may not generalise to the wider population. There are other circumstances in which observational studies may be important. For example, ethical or practical considerations may prevent initiating an RCT.[2]

The disadvantage of an observational cohort study is that patients are not randomised, but get treatment according to usual clinical practice. The treated and untreated patients may differ systematically on key covariates that influence outcomes.[3] The randomisation process in an RCT creates groups that are balanced, ensuring that the intervention and control groups can be directly compared and used to establish causal effects.[4] In contrast, observational studies are at greater risk of confounding by indication, that is, the treated group may differ systematically from those who are not treated.[5] In the context of antibiotic prescribing, observational studies are particularly at risk of confounding by indication, as clinicians' decisions to issue a prescription are based on factors such as the severity of clinical signs and symptoms in the initial consultation, which in turn impact on the outcome measures of interest.

There are various statistical methods to adjust for confounders. These adjust the observed crude association for identified potential confounders.[2] In the 1980s, Rosenbaum and Rubin[6] introduced the propensity score, which is intended to address confounding by indication and its use has increased in recent years. The propensity score represents the probability of receiving the intervention and is calculated for each individual patient. The score can then be used to adjust outcomes using inverse probability weighting, stratification, or matching.[7,8] The propensity score balances the dataset on observed covariates. By creating a dataset balanced on observed covariates, similar to the structure of an RCT, it should be possible to make accurate causal inferences in the observational study population. If the results obtained in an RCT therefore represent the true treatment effect in the general population, it could be assumed that the same treatment effect would be seen in an observational study balanced by propensity score, assuming there is no residual unmeasured confounding.

In some situations, propensity scores may give similar results to traditional methods of controlling for confounding.[9]

BL Stuart, PhD, senior research fellow; **P Little**, FRCGP, professor, Primary Care and Population Sciences Division, University of Southampton, UK. **LEN Grebel**, MSc, medical student; **TJM Verheij**, PhD, professor, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, the Netherlands. **CC Butler**, FMedSci, professor of primary care, Nuffield Department of Primary Care Health Sciences, University of Oxford, UK. **K Hood**, PhD, professor, Centre for Trials Research, Cardiff University, UK.

Address for correspondence
Beth Stuart, University of Southampton, Aldermoor Health Centre, Aldermoor Close, Southampton SO16 5ST, UK.

Email: bls1@soton.ac.uk

Submitted: 12 September 2016; **Editor's response:** 18 November 2016; **final acceptance:** 23 February 2017.

This is the full-length article (published online 1 Aug 2017) of an abridged version published in print. Cite this version as: **Br J Gen Pract 2017; DOI: https://doi.org/10.3399/bjgp17X692153**

How this fits in

There have long been discussions about the benefits and disadvantages of randomised controlled trials versus observational studies, especially in primary health care, with higher risk of confounding being the main disadvantage of observational studies. This study shows that observational studies using the propensity score to adjust for confounding can allow accurate inferences about treatment effect to be made and can therefore sometimes be an acceptable alternative to randomised trials.

However, they are considered to have some methodological advantages.[10] Unlike traditional methods of controlling for confounding, the propensity score approach provides balance diagnostics, allowing examination of whether the model has been adequately specified. The propensity score is also developed independently of the analysis of the relationship between exposure, or treatment, and outcome, so the researcher avoids any temptation to continue adjusting the regression model until the desired effect is achieved.[11] There may also be more flexibility in studies where the outcome is rare but the exposure is common, For example, propensity scores have been used to explore the rare outcome of cardiovascular events in patients with diabetes according to the more common exposure of treatment with thiazolidinedione.[12.]

It might not be possible to include all the baseline confounders for a rare outcome — at least 10 events per covariate is often recommended.[13] But if the treatment is more common, there may be more flexibility in including these confounders in the calculation of the propensity score.[14]

Discussion of findings from RCTs compared with observational studies is often hampered by differing designs, settings, inclusion criteria, and outcome measures. The GRACE studies gathered data on patients seen in general practice for acute lower respiratory tract infection in 12 European countries. They offer a unique opportunity to compare and analyse the differences between outcomes from observational studies and an RCT using the same inclusion criteria, similar settings, and the same follow-up measurements. This sub-study aimed to compare estimates of the effect of antibiotic treatment in patients with lower respiratory tract infections in an RCT and prospective observational cohort. Analyses were performed that *did* and *did not* take propensity scores into account to see if this approach resulted in similar estimates of treatment effect in studies using both an observational and a randomised design.

METHOD

Study design and participants

This sub-study used data from an observational study and a RCT conducted within the GRACE Network of Excellence (Genomics to combat Resistance Against Antibiotics in Community-Acquired lower-respiratory-tract infection in Europe, a European funded project in 14 countries). Patients in the observational study[15] and the trial[16] were recruited concomitantly with the same inclusion criteria between November 2007 and April 2010 in 16 primary care research networks in 12 countries (Belgium, England, France, Germany, Italy, the Netherlands, Poland, Slovakia, Slovenia, Spain, Sweden, and Wales). Patients who required initial antibiotics (for example, those with a clinical diagnosis of community-acquired pneumonia) or those who declined randomisation were asked to contribute to the observational study.

Eligible patients were aged ≥18 years, consulting with an illness where an acute or worsening cough was the main dominant symptom (≤28 days' duration), or had a clinical presentation that suggested lower respiratory tract infection. All included patients gave written consent. Exclusion criteria were immunosuppression, pregnancy and breastfeeding, and those not able to fill in the study material.

Treatment

In the trial, patients were allocated to amoxicillin 1 g three times a day or placebo. For the observational study, the case report form was reviewed to determine whether or not patients were prescribed antibiotics.

Although the trial standardised antibiotic prescribing, with all participants receiving either amoxicillin or a placebo, no restriction was placed on the prescribing practice of clinicians in the observational study. Amoxicillin was the most frequently prescribed antibiotic, but amoxicillin/clavulanic acid (co-amoxiclav) was often prescribed in a number of countries, as were doxycycline and macrolides. These different types of antibiotics have a different working spectrum and hence might influence the outcome. In order to provide a direct comparison with the trial, the 'treated' arm of the observational study was limited for this sub-study to patients who were prescribed amoxicillin.

Outcomes

The primary outcome for all datasets was the duration of symptoms rated by the patient as 'moderately bad' or worse after initial presentation. Symptom severity and re-consultation with new or worsening symptoms were secondary endpoints. Symptom severity was measured as the mean diary score for all symptoms rated from 0 (normal/not affected) to 6 (as bad as it could be) during days 2–4 after the index consultation. Data on re-consultation was defined as a return to the physician with worsening symptoms, new symptoms or signs, or illness necessitating admission to hospital within 4 weeks after the first consultation (established from reviews of patients' notes).

Statistical analysis

Propensity score. The propensity score is the conditional probability that a patient receives treatment, given a set of observed covariates.[10] This score can be used in further analyses in a number of ways including as a covariate in a regression model, as a probability weight, and in propensity score matching.[17,18] In this study, the propensity score was used as a population overlap weight.[19] The population overlap weight weights each unit proportional to its assignment to the alternative group and is designed to balance the distribution of covariates between comparison groups.

The variables included in the calculation of the propensity score were chosen on the basis of their association with the study outcomes (for a full set of variables see Box 1) but did not include instrumental variables (that is, those associated only with the exposure).[20–23] The selected covariates were used in a logistic regression model to predict the probability of receiving an antibiotic prescription, creating a unique propensity score for each individual. The probability of receiving a prescription varies both by clinician and country.[24] The GP and network were therefore included in the propensity score model as random effects.[25]

The predicted probabilities from this mixed logistic regression model were used to calculate the population overlap weights. The resulting propensity scores were then checked to make sure they adequately corrected for covariate imbalance in all covariates measured at the baseline consultation. Covariate balance was assessed by examining the standardised mean differences and a difference of 0.10 was taken to indicate substantial imbalance.[26]

Box 1. List of variables used to calculate the propensity score

- Age
- Duration of illness before consultation
- Duration of cough before consultation
- Breaths per minute
- Pulse rate
- Abnormalities at auscultation
- Low blood pressure
- Temperature
- Phlegm colour
- Lung comorbidity
- Heart disease
- Cough (yes/no)
- Wheeze (yes/no)
- Crackles (yes/no)
- Rhonchi (yes/no)
- Runny nose (yes/no)
- Chest pain (yes/no)
- Muscle aches (yes/no)
- Headache (yes/no)
- Disturbed sleep (yes/no)
- Confusion (yes/no)
- Illness interferes with normal activities (yes/no)
- Feeling generally unwell (yes/no)

Analyses of effects of antibiotics. Analyses of trial data were performed blind to treatment allocation and were based on an intention-to-treat analysis. For the observational data, analyses were based on whether a patient received antibiotics, recorded by the GP on the case report form at the initial consultation. A proportional hazards model was used to model the duration of symptoms, a linear regression model for symptom severity, and a logistic regression model for new or worsening symptoms. In the trial there was no evidence of clustering at the GP or country level.[16] However, there was evidence of clustering at both levels in the observational study,[24] and therefore all models of the observational data controlled for clustering at the country and GP levels as random effects. For the observational study data, the propensity score was calculated as described above and used as a probability weight in all models.

In line with the analysis of the trial, this analysis included patients for whom complete outcome data were available. Stata (version 14) was used for all analyses.

RESULTS

Baseline characteristics

Data on 780 patients were available from the observational study (233 in the amoxicillin group and 547 in the no antibiotics group). In the RCT, 2061 patients were randomly assigned (1038 to the amoxicillin group and 1023 to the placebo group). Table 1 shows that the two studies have broadly similar profiles. Observational study participants

Table 1. Baseline characteristics of participants and covariate balance[a]

	Trial (n = 2061)			Observational study (n = 780)			Observational study before and after applying propensity score weights	
Variable	Amoxicillin	Placebo	Total	Amoxicillin	No antibiotics	Total	Standardised mean difference	Standardised mean difference
Females	624/1038 (60.1)	600/1023 (58.7)	1224/2061 (59.4)	144/233 (61.8)	336/546 (61.5)	480/779 (61.6)	0.008	0.045
Age, years; mean (SD)	48.6 (16.7)	49.3 (16.4)	49.0 (16.5)	54.6 (15.7)	48.6 (16.9)	51.5 (17.1)	0.383	0.022
Non-smoker (past or present)	477/1037 (46.0)	483/1022 (47.3)	960/2059 (46.6)	148/233 (63.5)	289/545 (53.0)	437/778 (56.2)	0.383	0.006
Illness duration before index consultation, days; mean (SD)	9.5 (8.0)	9.3 (7.2)	9.4 (7.6)	9.1 (6.3)	9.8 (7.8)	9.5 (7.1)	0.212	0.066
Respiratory rate, breaths per minute; mean (SD)	16.9 (3.3)	16.9 (3.3)	16.9 (3.3)	17.9 (3.9)	16.9 (4.2)	17.1 (4.0)	0.262	–0.004
Temperature, °C; mean (SD)	36.7 (3.3)	36.8 (3.3)	36.8 (3.3)	36.8 (0.6)	36.7 (0.6)	36.7 (0.6)	0.145	0.001
Lung disease[b]	163/1037 (15.7)	147/1023 (14.4)	310/2060 (15.0)	67/233 (28.8)	86/545 (15.8)	153/778 (19.7)	0.320	–0.034
Mean severity score (all symptoms);[c] mean (SD)	2.1 (0.5)	2.1 (0.5)	2.1 (0.5)	2.3 (0.5)	2.0 (0.5)	2.1 (0.5)	0.652	–0.042
Sputum production	814/1036 (78.6)	824/1021 (80.7)	1638/2057 (79.6)	205/233 (88.0)	415/546 (76.0)	620/779 (79.6)	0.267	–0.013
Discoloured sputum[d]	481/968 (49.7)	468/957 (48.9)	949/1922 (49.4)	120/233 (51.5)	250/547 (45.7)	370/780 (47.4)	0.082	–0.008
Abnormalities at auscultation[e]	348/1029 (33.8)	340/1018 (33.4)	688/2047 (33.6)	172/230 (74.8)	177/542 (32.7)	349/772 (45.2)	0.952	–0.087
Disturbed sleep	638/1035 (61.6)	642/1022 (62.8)	1280/2057 (62.2)	172/232 (74.1)	324/546 (59.3)	496/778 (63.7)	0.347	–0.007
Crackles	63/1033 (6.1)	63/1018 (6.2)	126/2051 (6.1)	67/232 (28.9)	38/542 (7.0)	105/774 (13.6)	0.572	–0.049
Rhonchi	143/1032 (13.9)	138/1018 (13.6)	281/2050 (13.7)	76/230 (33.0)	80/542 (14.8)	156/772 (20.2)	0.471	–0.059
Runny nose	770/1035 (74.4)	734/1022 (71.8)	1504/2057 (73.1)	165/233 (70.8)	359/546 (65.8)	524/779 (67.3)	0.129	–0.006
Chest pain	474/1034 (45.8)	468/1021 (45.8)	942/2055 (45.8)	118/233 (50.6)	244/546 (44.7)	362/779 (46.5)	0.136	–0.009
Muscle ache	519/1035 (50.1)	524/1022 (51.3)	1043/2057 (50.7)	132/233 (56.7)	237/546 (43.4)	369/779 (47.4)	0.295	–0.010
Headache	572/1035 (55.3)	593/1023 (58.0)	1165/2058 (56.6)	126/233 (54.1)	281/546 (51.5)	407/779 (52.2)	0.079	0.002
Confusion	31/1035 (3.0)	52/1022 (5.1)	83/2057 (4.0)	12/233 (5.2)	15/546 (2.8)	27/779 (3.5)	0.144	0.002
Heart disease	56/1036 (5.4)	50/1023 (4.9)	106/2059 (5.1)	27/233 (11.6)	39/545 (7.2)	66/778 (8.4)	0.136	–0.006

[a]Data are n/N (%) unless otherwise stated. [b]Chronic obstructive pulmonary disease, asthma, or other lung disease. [c]Severity of symptoms: 1 = no problem, 2 = mild problem, 3 = moderate problem, 4 = severe problem. [d]Yellow, green, or bloodstained. [e]Represents the clinician's report of hearing anything they judged to be abnormal on auscultation. SD = standard deviation.

were more likely to be smokers, have a heart or lung condition, and to have crackles, rhonchi, and abnormalities on auscultation.

As expected, the groups were well balanced in the RCT, while there was significant imbalance in the observational study, with evidence of confounding by indication. Taking a threshold of 0.10 as indicating substantial imbalance, 18/20 (90%) of the key covariates showed evidence of imbalance in the observational study. After applying the propensity score weights, the standardised mean difference was below 0.10 for all covariates and all but three were below a threshold of 0.01, suggesting that the dataset was now well balanced on observed covariates and that the propensity score weights were successful in making the groups more directly comparable.

Effects of amoxicillin

The main results of the trial are presented in Table 2, but full results can be found in Little *et al*.[16] Table 2 also sets out the results for the observational study for comparison. No significant results were found, although the confidence intervals are wide, likely because of insufficient sample size.

The point estimates of the effect in the observational study adjusting only for baseline severity indicated a slightly longer duration of symptoms, higher symptom severity, and slightly increased risk of re-consultation. Because those who received amoxicillin were likely to be more unwell at baseline, this is as expected.

Adjusting for known confounders had little impact on the result for duration of symptoms compared with simply including baseline severity in the model. But the result for symptom severity now showed no difference and re-consultation was now less likely in the amoxicillin group but, again, this was not statistically significant.

Adjusting using the propensity score also gave non-significant results. The point estimates were similar in direction and magnitude to the trial results, but with wide confidence intervals. The hazard ratio for duration of symptoms was 1.06 (95% CI = 0.96 to 1.18) in the trial and 1.06 (95% CI = 0.80 to 1.40) in the observational study. The difference in the severity score was –0.07 (95% CI = –0.15 to 0.007) in the trial and –0.07 (95% CI = –0.34 to 0.20) in the observational study. The odds ratio for re-consultation was 0.97 (95% CI = 0.63 to 0.99) in the trial and 0.91 (95% CI = 0.48 to 1.66) in the observational study.

Table 2. Outcomes for the RCT and adjusted and unadjusted outcomes for the observational study

		Duration of symptoms[a] Hazard ratio (95% CI)	Symptom severity[b] Mean difference (95% CI)	New/worsening symptoms[c] Odds ratio (95% CI)
RCT	Results controlling for baseline severity	1.06 (0.96 to 1.18)	–0.07 (–0.15 to 0.007)	0.97 (0.63 to 0.99)[d]
Observational study	Results controlling for baseline severity	0.92 (0.76 to 1.11)	0.06 (–0.11 to 0.23)	1.04 (0.64 to 1.67)
	Results controlling for baseline severity and confounding variables	0.92 (0.73 to 1.16)	–0.01 (–0.19 to 0.17)	0.85 (0.48 to 1.51)
	Results using propensity score weight	1.06 (0.80 to 1.40)	–0.07 (–0.34 to 0.20)	0.91 (0.48 to 1.66)

[a]Resolution of symptoms rated 'moderately bad' or worse in treatment versus no treatment group. [b]Difference of mean symptom severity score on days 2–4 after consultation between groups. [c]Worsening of illness in the treatment group versus no treatment. [d]Significant at P<0.05. CI = confidence interval. RCT = randomised controlled trial.

DISCUSSION

Summary

There was no statistically significant benefit of amoxicillin for either symptom duration or symptom severity in the observational study. Although the observational study represented a slightly different population from those who agreed to randomisation in the trial, the estimates of treatment effect were non-significant after controlling for confounders both in the traditional method and after weighting by propensity score.

Strengths and limitations

The format of the GRACE studies created a unique opportunity to compare outcomes from an observational study with an RCT with similar setting and inclusion criteria.

It is likely that there is no true effect of the intervention in this setting. The chances of finding similar negative results in both studies, regardless of the method used to control for confounding, was therefore high. Although propensity scores have methodological advantages, in this context it is not possible to say that this method provided superior control for confounding by indication when compared with traditional methods. Ideally, this analysis should be repeated in two studies where the trial has shown a statistically significant effect. However, the similarity in the magnitude and direction of the estimates obtained in the trial and observational study, both after controlling for confounding using the traditional approach and after weighting by propensity score, makes it more likely that the true effect in this population has been correctly estimated.

Although- the randomisation process

assures that RCTs are balanced both on observed and unobserved factors, propensity score methods can only account for measured confounders in observational data.[27] It is possible that observational data may still suffer from residual confounding. Given the similarity of the estimates in the observational study to the unconfounded estimates in the RCT, it is less likely that the results suffer from residual confounding.

The use of different antibiotic classes in the observational study compared with the trial might be a potential limitation. In this sub-study the analysis was limited to patients who were prescribed amoxicillin. However, this has reduced the available population for analysis, leading to wider confidence intervals.

The use of propensity scores as a weight in a regression model may give confidence intervals that are too narrow, as uncertainty surrounding the estimation of the propensity score is not accounted for in the model.[28] In this study, the confidence intervals were wide and the results non-significant. However, this may be an issue in larger datasets, and solutions such as Bayesian propensity score analysis[29] and bootstrapping[30] should be considered.

Comparison with existing literature

To the best of the authors' knowledge, this is the first study to compare the same outcomes from different study designs with such similar settings, inclusion criteria, and outcome measures. This, combined with the detailed information available from the patients, created an ideal opportunity to explore the utility of propensity score weights to enable causal inference from observational data and to explore the generalisability of the RCT findings.

Implications for research and practice

This study shows that it is possible to obtain estimates of treatment effect in observational data that are comparable with estimates from RCTs. In general, observational studies are expected not only to yield different results from trials because of confounding, but also because of a different setting or context, and differing behaviour of both health professionals and patients. The GRACE study design, however, made it possible to look solely at whether it is possible to account for confounding by indication sufficiently to allow causal inferences in observational data. This study is therefore a contribution to the assessment of the merits of observational and randomised studies: observational studies using appropriate methods to control for confounding by indication can sometimes be an acceptable alternative to RCTs.

It also confirms the RCT result showing a lack of benefit of amoxicillin can be replicated in a population of similar patients who were unwilling to be randomised. This suggests that the result is generalisable beyond the trial population.

Funding

Funding was from the European Commission Framework Programme 6 (LSHM-CT-2005-518226): Eudract-CT 2007-001586-15 UKCRN Portfolio ID 4175 ISRCTN52261229 FWO G.0274.08N. The researchers are independent of all funders. The work in the UK was also supported by the National Institute for Health Research. In Barcelona, the work was supported by: 2009 SGR 911, Ciber de Enfermedades Respiratorias (Ciberes CB06/06/0028); the Ciberes is an initative of the ISCIII. In Flanders (Belgium), this work was supported by the Research Foundation — Flanders (FWO; G.0274.08N). The South East Wales Trials Unit is funded by the National Institute for Social Care and Health Research.

Ethical approval

The study was approved by ethics committees in all participating countries. The competent authority in each country also gave their approval. Patients who fulfilled the inclusion criteria were given written and verbal information on the study and provided written informed consent. For full ethics statements, see the original articles of Little *et al*[16] and Hamoen *et al*.[15] The trial is registered with EudraCT (2007-001586-15), UKCRN Portfolio (ID 4175), ISRCTN (52261229), and FWO (G.0274.08N).

Provenance

Freely submitted; externally peer reviewed.

Competing interests

The authors have declared no competing interests.

Acknowledgements

The authors gratefully acknowledge all participating patients, GPs, and other professional participants who contributed to the GRACE study.

Discuss this article

Contribute and read comments about this article: **bjgp.org/letters**

REFERENCES

1. McKee M, Britton A, Black N, *et al.* Interpreting the evidence: choosing between randomised and non-randomised studies. *BMJ* 1999; **319(7205):** 312.
2. Grootendorst DC, Jager KJ, Zoccali C, Dekker FW. Observational studies are complementary to randomized controlled trials. *Nephron Clin Pract* 2010; **114(3):** c173–c177.
3. Greenland S, Morgenstern H. Confounding in health research. *Annu Rev Public Health* 2001; **22:** 189–212.
4. Sedgwick P. Randomised controlled trials: balance in baseline characteristics. *BMJ* 2014; **349:** g5721.
5. Klungel OH, Martens EP, Psaty BM, *et al.* Methods to assess intended effects of drug treatment in observational studies are reviewed. *J Clin Epidemiol* 2004; **57(12):** 1223–1231.
6. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70(1):** 41–55.
7. Guo S, Fraser MW. *Propensity score analysis: statistical methods and applications*. Thousand Oaks, CA: Sage Publications, 2010.
8. Williamson EJ, Forbes A. Introduction to propensity scores. *Respirology* 2014; **19(5):** 625–635.
9. Shah BR, Laupacis A, Hux JE, Austin PC. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. *J Clin Epidemiol* 2005; **58(6):** 550–559.
10. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res* 2011; **46(3):** 399–424.
11. Rubin DB. Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Serv Outcomes Res Methodol* 2001; **2(3):** 169–188.
12. Hajage D, Tubach F, Steg PG, *et al.* On the use of propensity scores in cases of rare exposure. *BMC Med Res Methodol* 2016; **16:** 38. DOI: 10.1186/s12874-016-0135-1.
13. Peduzzi P, Concato J, Kemper E, *et al.* A simulation study of the number of events per variable in logistic regression analysis. *J Clin Epidemiol* 1996; **49(12):** 1373–1379.
14. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. *Ann Intern Med* 2002; **137(8):** 693–695.
15. Hamoen M, Broekhuizen BD, Little P, *et al.* Medication use in European primary care patients with lower respiratory tract infection: an observational study. *Br J Gen Pract* 2014; DOI: https://doi.org/10.3399/bjgp14X677130.
16. Little P, Stuart B, Moore M, *et al.* Amoxicillin for acute lower-respiratory-tract infection in primary care when pneumonia is not suspected: a 12-country, randomised, placebo-controlled trial. *Lancet Infect Dis* 2013; **13(2):** 123–129.
17. McMahon AD. Approaches to combat with confounding by indication in observational studies of intended drug effects. *Pharmacoepidemiol Drug Saf* 2003; **12(7):** 551–558.
18. Rosenbaum PR. Model-based direct adjustment. *J Am Stat Assoc* 1987; **82(398):** 387–394.
19. Li F, Morgan KL, Zaslavsky AM. *Balancing covariates via propensity score weighting*. Working paper available at arXiv:14041785. 2014.
20. Brookhart MA, Schneeweiss S, Rothman KJ, *et al.* Variable selection for propensity score models. *Am J Epidemiol* 2006; **163(12):** 1149–1156.
21. D'Agostino RB Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17(19):** 2265–2281.
22. Shadish WR, Galindo R, Wong VC, *et al.* A randomized experiment comparing random and cutoff-based assignment. *Psychol Methods* 2011; **16(2):** 179–191.
23. Weitzen S, Lapane KL, Toledano AY, *et al.* Principles for modeling propensity scores in medical research: a systematic literature review. *Pharmacoepidemiol Drug Saf* 2004; **13(12):** 841–853.
24. Francis NA, Gillespie D, Nuttall J, *et al.* Antibiotics for acute cough: an international observational study of patient adherence in primary care. *Br J Gen Pract* 2012; DOI: https://doi.org/10.3399/bjgp12X649124.
25. Li F, Zaslavsky AM, Landrum MB. Propensity score weighting with multilevel data. *Stat Med* 2013; **32(19):** 3373–3387.
26. Normand ST, Landrum MB, Guadagnoli E, *et al.* Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol* 2001; **54(4):** 387–398.
27. Harder VS, Stuart EA, Anthony JC. Propensity score techniques and the assessment of measured covariate balance to test causal associations in psychological research. *Psychol Methods* 2010; **15(3):** 234–249.
28. An W. Bayesian propensity score estimators: incorporating uncertainties in propensity scores into causal inference. *Sociol Methodol* 2010; **40(1):** 151–189.
29. McCandless LC, Gustafson P, Austin PC. Bayesian propensity score analysis for observational data. *Stat Med* 2009; **28(1):** 94–112.
30. Williamson EJ, Morley R, Lucas A, Carpenter JR. Variance estimation for stratified propensity score estimators. *Stat Med* 2012; **31(15):** 1617–1632.