Official Journal of the Asian Pacific Society of Respirology
Respirology

INVITED REVIEW SERIES:
MODERN STATISTICAL METHODS IN RESPIRATORY MEDICINE
SERIES EDITORS: RORY WOLFE AND MICHAEL ABRAMSON

Introduction to propensity scores

ELIZABETH J. WILLIAMSON[1,2,3,4] AND ANDREW FORBES[1,3]

[1]*School of Public Health & Preventive Medicine, Monash University, and* [2]*Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne,* [3]*The Victorian Centre for Biostatistics (ViCBiostat), Melbourne, Victoria, Australia, and* [4]*Farr Institute of Health Informatics Research, London, UK*

ABSTRACT

Although randomization provides a gold-standard method of assessing causal relationships, it is not always possible to randomly allocate exposures. Where exposures are not randomized, estimating exposure effects is complicated by confounding. The traditional approach to dealing with confounding is to adjust for measured confounding variables within a regression model for the outcome variable. An alternative approach—propensity scoring—instead fits a regression model to the exposure variable. For a binary exposure, the propensity score is the probability of being exposed, given the measured confounders. These scores can be estimated from the data, for example by fitting a logistic regression model for the exposure including the confounders as explanatory variables and obtaining the estimated propensity scores from the predicted exposure probabilities from this model. These estimated propensity scores can then be used in various ways—matching, stratification, covariate-adjustment or inverse-probability weighting—to obtain estimates of the exposure effect.

In this paper, we provide an introduction to propensity score methodology and review its use within respiratory health research. We illustrate propensity score methods by investigating the research question: 'Does personal smoking affect the risk of subsequent asthma?' using data taken from the Tasmanian Longitudinal Health Study.

Keywords: causal inference, confounding, environmental and occupational health and epidemiology, observational studies, statistics.

Abbreviations: ATE, average treatment effect; ATT, average treatment effect in the treated; CI, confidence interval; IPTW, inverse probability of treatment weighting; OR, odds ratio; SD, standard deviation; TAHS, Tasmanian Longitudinal Health Study.

INTRODUCTION

In respiratory health research, we often wish to estimate the causal effect of a particular exposure on a health outcome.[1] When the exposure is not randomly allocated, such analyses are inevitably affected by confounding. We typically address this by adjusting for measured confounding variables within a multivariable regression model for the outcome including the exposure as an explanatory variable.[2] Alternatively, we could perform a propensity score analysis.[3] Instead of fitting a regression model for the outcome, the propensity score approach involves fitting a model for the exposure; the predicted exposure probabilities from this model are the estimated propensity scores. Broadly speaking, by controlling for these propensity scores, we hope to 'balance out' the confounders among exposure groups, thus removing observed confounding. This can be achieved in a number of ways, for example, matching or stratification on the propensity score, inverse probability-of-treatment weighting and covariate adjustment using the propensity score.[3–5]

In this paper, we provide an introduction to propensity score methodology and review its use within respiratory health research. To illustrate these methods, we use propensity scoring to investigate the effect of personal smoking (a non-randomized exposure) on asthma remission among adults who reported asthma during childhood, using data taken from the Tasmanian Longitudinal Health Study (TAHS), a population-based longitudinal cohort study in Tasmania, Australia.

Correspondence: Andrew Forbes, School of Public Health & Preventive Medicine, The Alfred Centre, 99 Commercial Road, Melbourne, Vic. 3004, Australia. Email: andrew.forbes@monash.edu

The Authors: Dr Elizabeth J Williamson is a Biostatistician with research interests focusing on methods for causal inference. Professor Andrew B Forbes is head of the Biostatistics Unit, School of Public Health and Preventive Medicine, with research interests in biostatistical methodology applied to practical problems and collaborative epidemiological and clinical research.

Received 24 March 2014; Accepted: 30 March 2014.
Article first published online: 29 May 2014

ESTIMATING CAUSAL EFFECTS

Intuitively, the causal effect of personal smoking on subsequent asthma remission for an individual can be conceptualized by contrasting their asthma remission status under two possible scenarios: the asthma status that would have occurred if that person had smoked, and the asthma status that would have occurred if that person had not smoked. (For simplicity, we assume that a binary classification of smoking is reasonable). Suppose that this individual did, in fact, smoke. Their asthma status under the non-smoking scenario is a *counterfactual* outcome; it is contrary to fact.[6] Because we can only ever observe the outcome under one possible exposure status, we can never observe these causal effects. This has been called the *fundamental problem of causal inference*.[7] However, under certain assumptions, we can estimate the average causal effect for a population of individuals even though we cannot *observe* the causal effect for them.

We can quantify the causal effect of smoking in a population by the causal odds ratio, which is a hypothetical construct: the odds of asthma remission that we would have seen in the population if everyone in the population had smoked, divided by the odds of asthma remission that we would have seen in the population if no one had smoked. We could alternatively consider the causal risk ratio, or causal risk difference, defined analogously. For brevity, we consider only the causal odds ratio.

In order for the causal odds ratio to be a well-defined quantity, everyone in the TAHS data must have had the possibility of being in both exposure groups—in this case, of being either a smoker or a non-smoker (the 'positivity' assumption[8]). We assume that the effect of personal smoking for one individual does not depend on other individuals' smoking behaviour (the 'stable-unit-treatment-value' assumption[6]). The critical and usually most controversial assumption, required to estimate the desired causal effect, is that all confounders have been adequately measured (the 'exchangeability' assumption[6]). We note that the exchangeability assumption is also required in conventional outcome regression models, so it represents nothing new. Accessible discussions of these issues are given elsewhere.[6,9]

PROPENSITY SCORES

The propensity score is defined as the probability of being exposed given the values of measured confounding variables.[3] This can be estimated for each individual by fitting, for example, a logistic regression model where the exposure takes the place of the outcome variable, and the measured confounding variables are included as explanatory variables. The predicted exposure probabilities from this model are the estimated propensity scores, which by definition all lie between zero and one. Advanced computationally intensive methods, such as neural networks, recursive partitioning and boosting, have also been suggested as alternatives to the logistic regression propensity score model.[10–12] These provide superior performance in some situations. Logistic regression, however, remains the most popular model choice.

Balancing covariates using the propensity score

In a simple randomized trial, we expect the distribution of all prognostic factors to be—on average—the same in the control and intervention arms of the trial. This expected balance of prognostic factors, or 'exchangeability', means that the control arm provides an estimate of the average outcome that would have been observed in the intervention arm had this group instead been assigned to the control condition, and vice versa. This allows us to estimate the causal effect of the intervention by simply contrasting the average outcomes between the two arms.

Rosenbaum and Rubin[3] showed that on average, we expect the distribution of all the variables included in the propensity score model to be the same in the exposed and unexposed groups *at each value of the estimated propensity score*. This expected balance means that we can estimate the exposure effect simply by contrasting the outcome between exposure groups *at each value of the estimated propensity score*, provided that all confounders are included in the propensity score model. Thus, matching or stratifying on the propensity score, or adjusting for it in a regression model for the health outcome, are valid ways to proceed in estimating the causal effect of exposure. We discuss these analyses in more detail below. Figure 1 depicts the main steps involved in a propensity score analysis. Details of the steps are given in the remainder of this paper.

Propensity score matching

In propensity score matching, we create groups of exposed and unexposed individuals who all have similar estimated propensity scores. A popular way of selecting these matched groups is 1:1 nearest neighbour matching within a caliper. This involves selecting a single unexposed match for each exposed individual, provided that a match can be found with a sufficiently close propensity score (where 'closeness' is determined by the caliper); otherwise the exposed individual is discarded. There are many variations on the procedure used to select the matched sample.[13,14] Matching can be either without replacement, where an unexposed individual is used as a match for at most one exposed individual, or with replacement. If the latter is adopted, the multiple use of matches must be accounted for in the statistical analysis. An increase in precision of the estimated causal effect can often be obtained by selecting more than one untreated match for each exposed individual; typically such matching strategies need to be accounted for in the analysis, for example through weighting, as illustrated in our example below.[13,14]

Once a matched sample has been selected, health outcomes are directly compared between the exposed and unexposed individuals. How the matching should be accounted for in the analysis remains controversial (see Austin[15] and discussion).

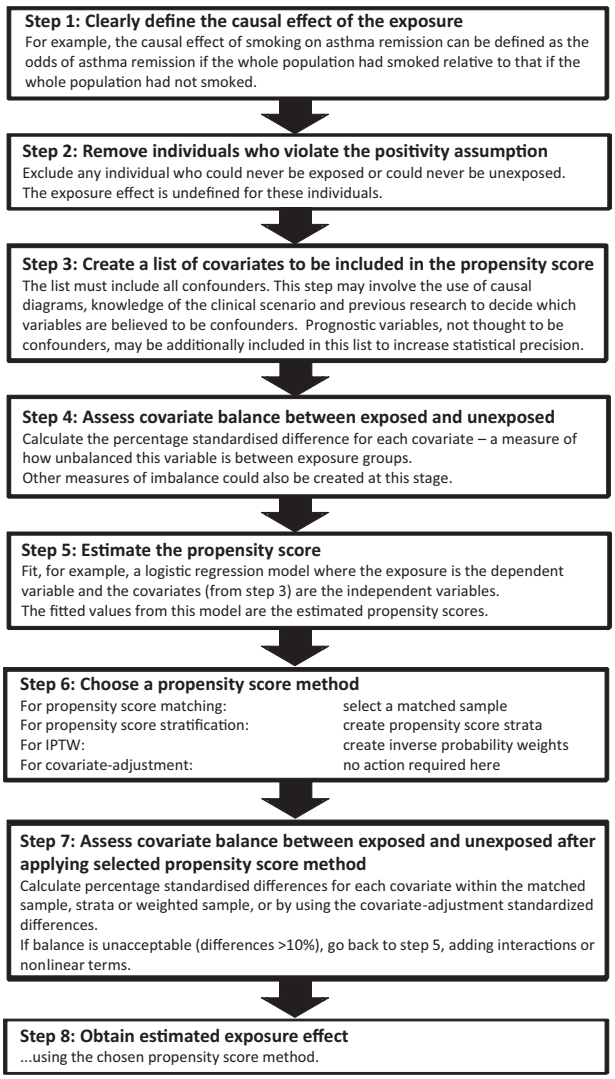

Figure 1 Overview of the key steps involved in a propensity score analysis. IPTW, inverse probability of treatment weighting.

Propensity score matching is often a good choice when the number of exposed individuals is greatly exceeded by the number of unexposed individuals. When this is not the case (as in our later data analysis), matching can result in many exposed individuals being discarded due to lack of a suitable unexposed match, or alternatively requires the use of matching with replacement.

A rather subtle issue with propensity score matching is that this procedure, as outlined earlier, estimates the causal effect of the exposure among the exposed individuals only, rather than in the whole population. This is often called the average treatment effect in the treated (ATT). Conceptually, this compares the health outcome of all exposed individuals with the outcome had they not been exposed. If the exposure effect is stronger among some patient subgroups, this can lead to discrepancies between estimates obtained from propensity score matching and other analysis methods.[16] Variations on the matching procedure can be used to estimate the causal exposure effect among the whole population (often called the average treatment effect (ATE)), if desired.[17]

Propensity score stratification

A simple method—propensity score stratification or subclassification—involves creating a number of strata, often five, based on the percentiles of the estimated propensity score. The within-strata exposure effects are estimated by contrasting the outcome between exposure groups within each stratum, and the mean of these across the strata is taken to obtain an overall estimate of the exposure's causal effect.[18] A variation on this method fits a regression model for the health outcome including both the propensity score strata and the exposure as explanatory variables. Additionally, including an interaction term between the strata and the exposure, and averaging the resulting strata-specific exposure effects across strata results in a similar estimate of the exposure's causal effect to the first stratified approach described.

Stratification is simple and intuitive. The main drawback of this method concerns residual confounding; small differences in the distribution of propensity scores between exposed and unexposed individuals may exist within the same propensity score stratum and could mean that some confounding remains. Increasing the number of strata—where sample size permits—can reduce residual confounding.

Covariate adjustment using the propensity score

The third, and probably most commonly used, propensity score method involves fitting a regression model for the outcome including the exposure and the estimated propensity score as explanatory variables, with the estimated propensity score treated as a continuous variable and often additionally adjusting for other measured confounders.[19]

This approach—which we will call covariate adjustment—is simple and easy to apply. It is related to the regression-based stratification approach described above where the strata, rather than the continuous propensity score, are included as explanatory variables. However, because the propensity score is modelled as a continuous variable, the covariate-adjustment method imposes a restrictive assumption about the relationship between the health outcome and the propensity score. The estimate of the exposure's effect will be valid only if this relationship is correctly modelled. This contrasts strikingly with the essence of the other propensity score methods where emphasis is on the propensity score model alone. Indeed, avoiding the need for correct specification of an outcome model is a key advantage of the propensity score approach, and therefore we, like others, do not recommend the use of covariate adjustment propensity score method.[3,20]

Inverse probability of treatment weighting

The final propensity score method we consider is called inverse probability of treatment weighting

(IPTW).[21] Unlike the other methods, it does not attempt to compare subgroups of individuals with the same value of the propensity score. Instead, it uses the estimated propensity scores to weight individuals in such a way as to create a 'pseudo-population'[9] in which the measured confounding variables are balanced between exposure groups, thereby effectively removing the confounding. This is achieved, with p representing the individual's estimated propensity score, by allocating a probability weight of 1/p to exposed individuals, and a weight of 1/(1-p) to unexposed individuals. For example, suppose 1/2 of the 200 males and 1/4 of the 100 females in a sample are exposed. Upweighting exposed men by a factor of 2, unexposed men also by a factor of 2, exposed women by a factor of 4 and unexposed women by a factor of 4/3 yields balance in gender across exposure groups.

To estimate the exposure effect, a suitable regression model for the outcome (linear for continuous outcomes, logistic for binary etc.) is fitted including the exposure as the sole explanatory variable and applying the probability weights described above. The estimated coefficient for exposure provides an estimate of the exposure's causal effect.

This method works well when the estimated propensity scores do not lie close to zero or one. Propensity scores near zero or one can result in extremely large weights, leading to very imprecise estimates of the exposure effect. Trimming the weights can alleviate this problem, although potentially at the cost of a small amount of bias in estimating the exposure's effect.[22] However, in the absence of large weights this method is easy to apply and mathematically elegant. We find that it is often our method of choice, particularly when the sizes of the exposed and unexposed groups are similar, diminishing the attraction of propensity score matching. However, thus far, this weighting approach is not as frequently used in practice as other propensity score methods.

Further considerations

We have focused on propensity score methods for binary exposures. These methods have been extended to categorical exposures[23] and some work has been done extending the methods to continuous exposures,[24] although theory is much less well developed in this area.

Choosing variables to include in the propensity score model can be important. The propensity score model must include *all* confounders (assuming they are measured), that is all variables that are believed to be associated with exposure and prognostic of the health outcome. Additionally including predictors of exposure but not of outcome typically decreases the precision of the exposure effect estimate, increasing the *P*-value and the width of the confidence interval, without decreasing the bias. Additionally including variables prognostic of outcome but not associated with exposure increases precision without increasing the bias.[25] Thus it may be helpful to include variables thought to be prognostic of outcome, even if they may not be confounders.

Recent work has combined traditional outcome regression modelling with each of the propensity score methods described in the previous section.[21,26] Such combinations can increase the precision of the estimated exposure effect and can offer some robustness against the possibility of mis-modelling the propensity score. However these are not yet routinely used, and are beyond the scope of this introductory paper.

Missing data in confounding variables can be dealt with by using multiple imputation[27] or by applying the simpler missing-data-category method, which, although a biased approach in standard regression analyses,[28,29] may be less biased within the propensity score context.[14,18]

Standard errors and *P*-values from propensity score analyses often do not take the estimation of the propensity score into account. Counter-intuitively, this produces conservative results; the confidence interval for the exposure effect will be too wide, and the *P*-value too large. We are presently aware of only one software package (Stata, version 13.0)[30] that makes appropriate corrections. This issue is most problematic with continuous outcomes; little precision is lost when the outcome is binary.

PROPENSITY SCORES IN RESPIRATORY HEALTH RESEARCH

We conducted a brief literature search using the Web of Knowledge database, of five respiratory health journals: the *American Journal of Respiratory and Critical Care Medicine* (*AJRCCM*), *Thorax*, *Chest*, *European Respiratory Journal* (*ERJ*), and *Respirology*. Articles published between 2009 and 2013, including the words 'propensity score(s)' or 'propensity scoring' were included. Abstracts and tutorial papers were excluded.

We identified 28 articles using propensity score methodology, 9 from *AJRCCM*,[31–39] 12 from *Chest*,[40–51] 5 from *ERJ*[52–56] and 1 each from *Thorax*[57] and *Respirology*,[58] with around 6 papers published in each of the 5 years considered. A summary of these studies can be found in the online supplement. In the majority of studies (22 (79%)) the exposure was binary, with one study considering a categorical exposure,[40] three studies assessing continuous exposures,[41,45,57] one with a mixture of binary and categorical exposures,[32] and the final study considering the dose of fluoroquinolone received[33] both dichotomized and as the original continuous measure.

Of the 24 studies including at least 1 binary exposure, 18 (75%) used a logistic regression model to estimate the propensity score, 1 used a generalized boosting model[58]—an alternative to logistic regression, which may account better for interactions between explanatory variables—and the remaining 5 studies did not give details of the model used. Multinomial regression models were used to estimate the propensity score in the two studies involving a categorical exposure. The models used for the continuous outcomes were unclear.

In 17 (61%) of the studies, the primary health outcome was binary, often 30-day mortality in hospital-based studies. Nine studies had a time-to-event

outcome, and the remaining two studies had a continuous outcome relating to healthcare costs.

In nine of the studies (32%), an analysis based on traditional outcome regression modelling was also presented. In these cases, the propensity score and outcome regression estimates were very similar. In a further eight studies (29%), several propensity score methods were applied to assess the robustness of findings to the methods used. The remaining 11 studies used a single propensity score method.

Including the propensity score as a continuous explanatory variable in a model for the health outcome was the most widely used method, appearing in 14 (50%) of the studies, closely followed by propensity score matching used in 12 (43%) of studies. Eight studies used some form of propensity score stratification. Inverse probability weighting was not used in any of the studies. The most popular matching method was 1:1 matching without replacement. The number of exposed individuals who were able to be matched, only reported, in 5 of the 12 relevant studies, ranged from 50%[36] to 100%.[49] For stratification, three, five and nine strata were used in different studies. Combinations of methods were occasionally used (e.g. Sadatsafavi *et al*. included the propensity score as a covariate in an outcome regression model fitted to a propensity-score matched sample[48]).

Generally, the studies using propensity score methodology were carefully performed and reported. However, several studies did not adequately, or at all, assess whether balance of the confounding variables between exposure groups had actually been achieved—the key diagnostic measure for evaluating the performance of a propensity score method. Our next section, therefore, focuses on propensity score diagnostics.

PROPENSITY SCORE DIAGNOSTICS

When using propensity score methods, the key diagnostic criterion is whether balance of the confounding variables between exposure groups (covariate balance) has been achieved either within the matched sample, within propensity score strata, after adjustment for the propensity score or within the weighted pseudo-population, for the four methods, respectively. An excellent and comprehensive review of various diagnostic measures to assess covariate balance is given by Austin.[59]

Hypothesis testing is often used to assess covariate balance. This is discouraged, particularly for propensity score matching, due to its dependence on sample size and its focus on statistical significance rather than magnitude of differences.[60] Standardized differences, described later, provide a useful way of assessing balance that avoids these pitfalls.

Standardized differences

For a continuous confounder, let $\bar{x}_{exp}$ and $\bar{x}_{un}$ represent the mean in the exposed and unexposed groups, respectively, and s_{exp} and s_{un} the standard deviations (SD). The percentage standardized difference in this confounder between exposed and unexposed individuals is defined as

$$100\times\frac{(\bar{x}_{exp}-\bar{x}_{un})}{SD_{pool}} \quad \text{where} \quad SD_{pool}=\sqrt{\frac{s_{exp}^2+s_{un}^2}{2}} \qquad (1)$$

For binary confounders, with $\bar{p}_{exp}$ and $\bar{p}_{un}$ representing the confounder's observed prevalence in the exposed and unexposed groups, the percentage standardized difference is

$$100\times\frac{\bar{p}_{exp}-\bar{p}_{un}}{SD_{pool}} \quad \text{where} \quad SD_{pool}=\sqrt{\frac{\bar{p}_{exp}(1-\bar{p}_{exp})+\bar{p}_{un}(1-\bar{p}_{un})}{2}} \qquad (2)$$

Categorical variables can be converted into a set of binary indicators, one for each non-reference level of the variable, and then a set of standardized differences defined. Other authors present variations on these definitions. Stuart,[14] for example, replaces our pooled SD with the SD from the exposed group only.

A value of 10% or greater in magnitude in the percentage standardized difference is often taken as an indication of meaningful imbalance for—and thus potential confounding by—that variable. However, the negative consequences of imbalance will depend also on how prognostic of the health outcome the imbalanced variable is; it is more important to achieve close balance for strongly prognostic variables.

The standardized difference for a covariate can be calculated in the original sample and after applying one of the propensity score methods described previously (see subsections below; Fig. 3 demonstrates this graphically for our TAHS example). As has been suggested elsewhere,[14] we recommend using the same pooled SD before and after applying the propensity score method. This ensures that reductions in the standardized difference reflect a real increase in balance rather than simply a change in scale due to varying SD.

Standardized differences after propensity score matching

An advantage of propensity score matching is the ease with which covariate balance can be assessed within the matched sample. Standardized differences can be calculated within the matched sample by replacing the means in each exposure group in Equations 1 and 2 with the means within the matched sample but leaving the value of SD_{pool} unchanged. Where a varying number of unexposed matches is used, these means should be replaced by weighted means, weighting the unexposed individuals in a matched group by the inverse of the number of unexposed matches in the group.[61]

Standardized differences after propensity score stratification

When using stratification, standardized differences can be calculated by replacing the means in each exposure group in Equations 1 and 2 with the difference in the within-stratum means averaged over the strata according to the fraction of the sample in each stratum, and leaving the value of SD_{pool} unchanged.

Standardized differences after covariate adjustment

It is much less clear how to assess post-adjustment balance when using covariate adjustment, although Austin[19] has suggested an approach to calculating the standardized differences in this scenario.

Standardized differences after IPTW

The post-weighting standardized differences are obtained by applying the inverse probability weights to obtain weighted means of the covariate in each exposure group. These are substituted into Equations 1 and 2, dividing by the pooled SD from the original (unweighted) sample.

Other model fit diagnostics

Although the area under the curve or C-statistic are often reported in propensity score analyses, along with measures of goodness of fit such as the Hosmer–Lemeshow statistic, these are not relevant in the propensity score context because they do not measure the degree of control of confounding achieved.[62]

ANALYSIS OF THE TASMANIAN LONGITUDINAL HEALTH STUDY (TAHS)

In this section, we apply propensity score methods to estimate the effect of personal smoking on asthma remission (no adult asthma) among TAHS participants who reported asthma during childhood.

The TAHS cohort

At study enrolment in 1968,[63] parents provided information on their child's respiratory health including asthma and bronchitis history, plus information on their own respiratory health, smoking history and occupation. In 2004, the participant's adult asthma status, smoking history and occupation (reflecting socio-economic status) were documented.

The original data have been used in an extensive investigation of risk factors for asthma remission; clinical interpretations of these analyses have been reported previously.[64] This analysis is for illustrative purposes only and uses a subsample of 194 participants from the TAHS who reported asthma during childhood.

Confounding

Because smoking status is not randomly allocated, confounding is likely to be present. In a companion paper,[1] we discuss the use of causal diagrams for confounder selection. Applying this approach to the current research question, we concluded that the following confounders must be controlled for in the statistical analysis: poor childhood lung function, chronic bronchitis, number of asthma attacks, gender, number of parents reporting smoking, and socioeconomic status. In our previous paper,[1] we adjusted for these characteristics via a multivariable regression model for the outcome. Alternatively, we can adopt a propensity score approach. We use both approaches below for comparison.

The estimated propensity score

The propensity score is the probability of being a smoker, conditional on the selected confounders. We estimated the propensity score using a logistic regression model for smoking including the confounders as explanatory variables including no interaction terms or nonlinear terms. The propensity score was estimated for each individual using the fitted values from this model. An iterative procedure is often used, where covariate balance is assessed after fitting the initial propensity score model and is modified if covariate imbalance remains.[18] However, our initial model achieved acceptable balance (standardized differences at most around 10% in magnitude) so we retained this simpler model.

Analysis methods

We first performed standard outcome regression modelling, by fitting a logistic regression model for asthma remission including smoking status and the selected confounders as explanatory variables.

We then applied four propensity score methods. Firstly, we applied propensity score matching. Because the smoking group was the larger group, we used matching with replacement. For each smoker, we selected up to three non-smokers with the closest estimated propensity scores, provided that these were within a distance of 0.14 on the log odds scale (calculated as 0.2 standard deviations of the log odds of the estimated propensity score). We estimated the odds ratio for smoking using conditional logistic regression which is a method to accommodate the smoker-non-smoker groupings created by the matching, with robust standard errors accounting for the re-use of individual non-smokers in multiple groupings, after giving the non-smoker matches in each grouping a weight of $1/n_m$, where n_m is the number of non-smokers in that grouping, and a weight of 1 to the single smoker in the matched grouping.

Secondly, we created five equally-sized strata, based on the quintiles of the estimated propensity score distribution. A greater number of strata was impractical due to the small sample size. We fitted a logistic regression model of asthma remission on smoking status and these strata, including interactions between the strata and smoking status. The five within-strata effects (log-odds ratios) of smoking were combined in an arithmetic mean, weighting each estimate by the fraction of the sample in that propensity score stratum.

Thirdly, we performed covariate adjustment using the propensity score by fitting a logistic regression model for asthma remission on smoking status and including the estimated propensity score as an explanatory variable with a linear effect. We also fitted a similar model additionally adjusting for the confounders listed above.

Finally, we created inverse probability weights using the estimated propensity score. A logistic regression model of asthma remission on smoking status only, applying these probability weights, was fitted; the odds ratio for smoking status from this model is the IPTW estimate of the exposure effect.

Table 1 Baseline characteristics of subsample of 194 participants from the Tasmanian Longitudinal Health Study

Characteristic	Never smoker (n = 75)	Smoker (n = 119)
Demographics		
Age (year) at 2004 survey; mean (standard deviation)	42.6 (0.5)	42.6 (0.4)
Male	49 (65%)	66 (56%)
Socioeconomic status:		
1 (Highest)	33 (44%)	30 (25%)
2	6 (8%)	11 (9%)
3	12 (16%)	25 (21%)
4	10 (13%)	24 (20%)
5 (Lowest)	14 (19%)	29 (24%)
1968 survey		
Bronchitis	30 (40%)	40 (34%)
Poor lung function	4 (5%)	5 (4%)
Number of asthma attacks in the last 12 months		
1	8 (11%)	12 (10%)
2–5	27 (36%)	46 (39%)
6–10	15 (20%)	32 (27%)
11–20	14 (19%)	17 (14%)
>20	11 (15%)	12 (10%)
Parental smoking		
Neither	26 (35%)	17 (14%)
One	29 (39%)	64 (54%)
Both	20 (27%)	38 (32%)

Covariate balance

Standardized differences were used—within the strata, the matched sample and in the weighted pseudo-population—to assess covariate balance. We did not attempt to assess balance for the covariate-adjusted propensity score method.

Results

In 194 participants, 119 (61%) were classed as ever-smokers. Fifty-one (68%) of the never-smokers achieved asthma remission, a slightly lower proportion than 86 (72%) of the ever-smokers.

Table 1 shows a summary of the selected confounding variables by smoking status. Women, participants in the lower socioeconomic groups, participants without childhood lung problems and with less severe (or no) childhood asthma were more heavily represented in the smoking group. A larger proportion of children whose parents smoked were themselves smokers.

In this case, the propensity score is the probability of smoking, given the variables shown in Table 1. The median (minimum, maximum) of the estimated propensity scores was 0.58 (0.17, 0.82) in the non-smoking group, and 0.69 (0.20, 0.95) in the smoking group. The distribution of the estimated propensity scores is shown in the histograms of Figure 2 by smoking status. Generally, individuals in the smoking group tend to have higher propensities of smoking (as expected). However, the range of propensity scores in the two groups is broadly similar. Were this not the case, it would be a crude indication of a violation of the positivity assumption—some individuals in the sample have no one comparable with them in the other exposure group—in which case it may be necessary to restrict the analysis to a more select group.[8]

For the propensity score matching, suitable matches were found for all but 2 (2%) of the 119 smokers. Figure 3 shows the propensity score distribution within the propensity score matched sample. As expected, the distribution in the smokers and non-smokers has become virtually identical.

For the IPTW approach, the median (minimum, maximum) probability weights were 2.39 (1.21, 5.59) for the non-smokers and 1.45 (1.18, 5.06) for the smokers.

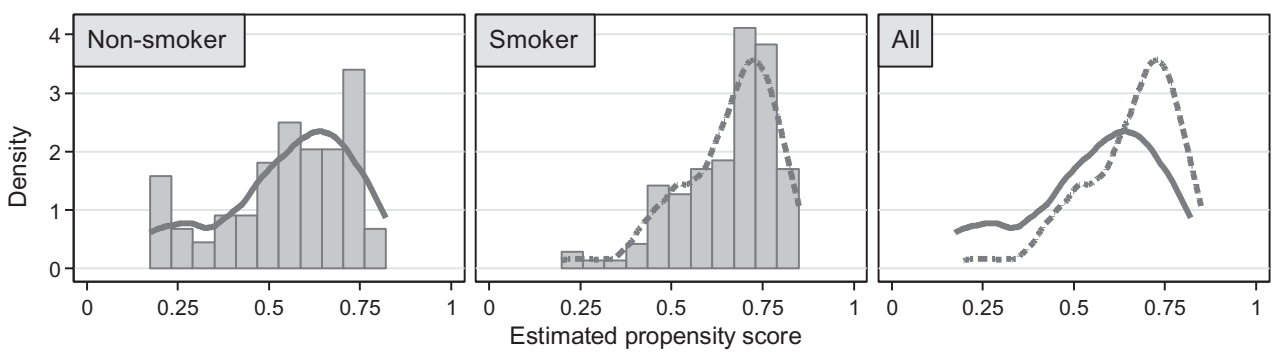

Figure 2 Histograms of the estimated propensity score by smoking status, with smoothed density estimates overlaid. Smoothed density: ——, non-smokers; - - -, smokers.

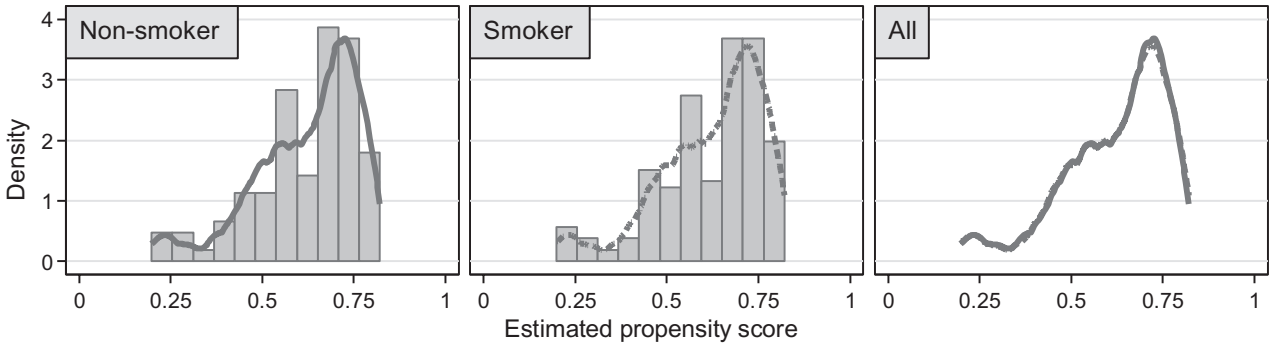

Figure 3 Estimated propensity scores within matched sample constructed to estimate the average treatment effect of personal smoking on subsequent asthma remission. Smoothed density: ▬, non-smokers; ▪▪▪, smokers.

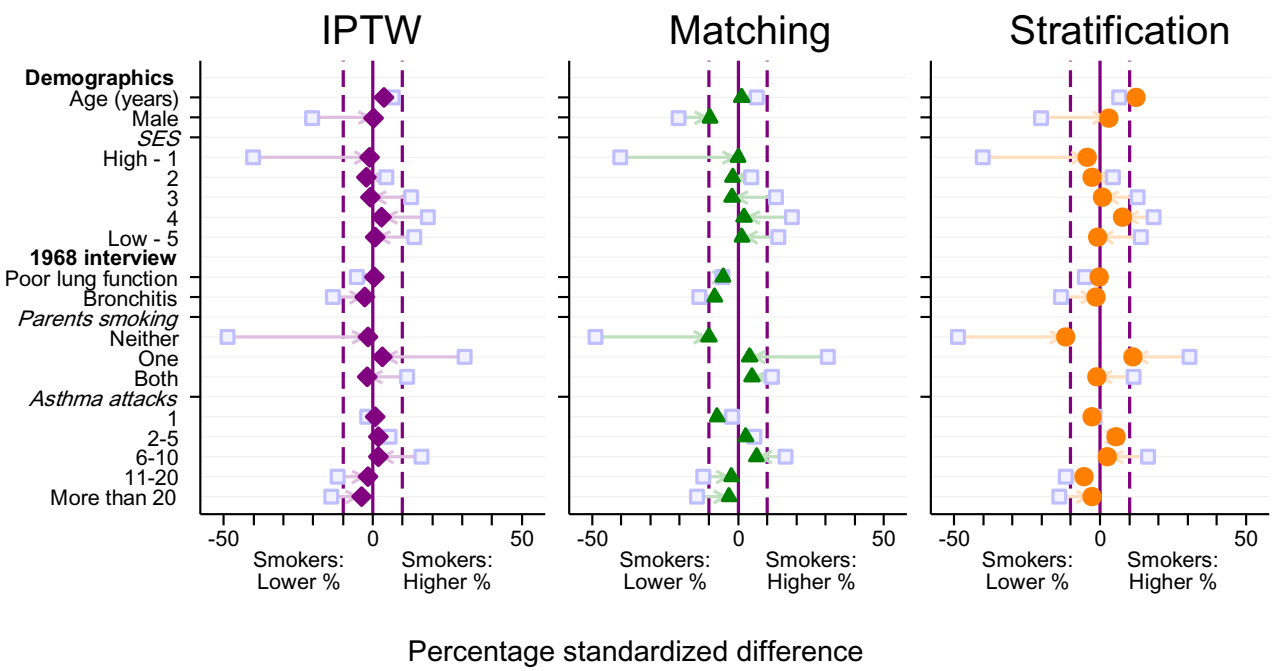

Figure 4 Percentage standardized differences before and after IPTW, matching and stratification on the estimated propensity score. The dashed vertical lines indicate the cut-off of 10%; values larger in magnitude are considered to represent substantial confounding. □, initial sample; ◆, weighted sample; ▲, matched sample; ●, within strata.

Figure 4 shows the percentage standardized differences of confounders between smoking groups, both in the original sample and in the matched sample, within the strata and in the weighted pseudo-population. We did not attempt to assess balance for the covariate-adjusted propensity score method. The initial sample had several differences of greater magnitude than 10% indicating moderate confounding. These standardized differences were all reduced to <10% by the IPTW approach. Matching greatly reduced the standardized differences, although a couple remained near 10%. Stratification, as expected, reduced the standardized differences the least, although confounder balance was still substantially reduced.

Table 2 shows estimates of the effect of smoking on subsequent asthma remission. All methods resulted in an estimated odds ratio less than the unadjusted odds ratio of 1.23 (95% confidence interval (CI): 0.65–2.30), $P = 0.53$, with the biggest reduction coming from stratification (OR = 0.99, 95% CI: 0.47–2.09) and the smallest reduction from covariate adjustment with additional adjustment for confounders (OR = 1.17, 95% CI: 0.57–2.37).

DISCUSSION

It is important to separate limitations of propensity score methods from limitations of the data sources. The latter are often of greater concern because analyses of non-randomized data frequently use existing data for which measurement of potential confounders is not within the control of the investigators. In

Table 2 Estimates of the effect of personal smoking on subsequent asthma remission using various statistical analysis methods

Analysis method	Odds ratio	95% CI	*P*-value
Unadjusted	1.23	(0.65, 2.30)	0.53
Adjusted using logistic regression	1.12	(0.56, 2.27)	0.75
Propensity score methods			
IPTW	1.13	(0.58, 2.21)	0.73
Stratification	0.99	(0.47, 2.09)	0.99
Matching (ATT)	1.06	(0.60, 1.86)	0.85
Covariate adjustment + Adjustment via logistic regression	1.10 1.17	(0.56, 2.14) (0.57, 2.37)	0.79 0.67

ATT, average treatment effect in the treated; CI, confidence interval; IPTW, inverse probability of treatment weighting.

consequence, critical measures may be collected poorly or not at all, thereby violating the key assumption that all confounders have been adequately measured. We note that this inhibits all forms of analysis of exposure–outcome relationships, whether using outcome regression, propensity scoring or otherwise.

Propensity scores allow the analysis to be conducted almost entirely without reference to the outcome variable, to a large extent avoiding the possibility of the chosen analysis approach being influenced by the results of the analysis. They are a particularly attractive analysis option where modelling the exposure is easier than modelling the outcome, for instance, where the outcome is rare but the exposure is common. Where many outcomes and few exposures are of interest, propensity scores can be useful because the one set of propensity scores can be applied (assuming common confounders across outcomes). However, where many exposures are of interest, it can be inefficient to model the propensity score for each exposure. Where outcome regression modelling and propensity score methods are both possible, they often give comparable estimates of exposure effect.[65]

Summary

In this paper, we have defined propensity scores, illustrated their use in a respiratory health context, presented diagnostic measures to validate balance after propensity score adjustment and discussed the broad advantages depending on the exposure and outcome data available. Propensity score methods form a useful addition to the medical researcher's toolkit. Even where propensity scoring is not the primary method of analysis, estimating and graphing the propensity score can be an invaluable tool in assessing the comparability of exposed and unexposed individuals. We encourage the further use of propensity scores in respiratory health research.

Acknowledgement

We thank the Tasmanian Longitudinal Health Study (TAHS) Steering Committee for providing us with a random subset of the data from the TAHS cohort which was funded by the National Health and Medical Research Council, Australia, ID#299901. This work was supported under a National Health and Medical Research Council Centre of Research Excellence grant, ID#1035261, to the Victorian Centre for Biostatistics (ViCBiostat).

REFERENCES

1 Williamson E, Aitken Z, Lawrie J, Dharmage S, Burgess J, Forbes A. An introduction to Causal Diagrams for confounder selection. *Respirology* 2014; **19**: 303–11.
2 Kasza J, Wolfe R. Statistical regression models: interpretation of commonly-used models. *Respirology* 2014; **19**: 14–21.
3 Rosenbaum P, Rubin D. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**: 41–55.
4 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**: 2265–81.
5 Williamson E, Morley R, Lucas A, Carpenter J. Propensity scores: from naive enthusiasm to intuitive understanding. *Stat. Methods Med. Res.* 2012; **21**: 273–93.
6 Hofler M. Causal inference based on counterfactuals. *BMC Med. Res. Methodol.* 2005; **5**: 28.
7 Holland PW. Statistics and causal inference. *J. Am. Stat. Assoc.* 1986; **81**: 945–60.
8 Westreich D, Cole SR. Invited commentary: positivity in practice. *Am. J. Epidemiol.* 2010; **171**: 674–7.
9 Hernan MA, Robins JM. Estimating causal effects from epidemiological data. *J. Epidemiol. Community Health* 2006; **60**: 578–86.
10 Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. *Stat. Med.* 2010; **29**: 337–46.
11 McCaffrey DF, Ridgeway G, Morral AR. Propensity score estimation with boosted regression for evaluating causal effects in observational studies. *Psychol. Methods* 2004; **9**: 403–25.
12 Setoguchi S, Schneeweiss S, Brookhart MA, Glynn RJ, Cook EF. Evaluating uses of data mining techniques in propensity score estimation: a simulation study. *Pharmacoepidemiol. Drug Saf.* 2008; **17**: 546–55.
13 Rassen JA, Shelat AA, Myers J, Glynn RJ, Rothman KJ, Schneeweiss S. One-to-many propensity score matching in cohort studies. *Pharmacoepidemiol. Drug Saf.* 2012; **21**: 69–80.
14 Stuart EA. Matching methods for causal inference: a review and a look forward. *Stat. Sci.* 2010; **25**: 1–21.
15 Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. *Stat. Med.* 2008; **27**: 2037–49.
16 Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, Robins JM. Results of multivariable logistic regression, propensity matching, propensity adjustment, and propensity-based weighting under conditions of nonuniform effect. *Am. J. Epidemiol.* 2006; **163**: 262–70.
17 Abadie A, Imbens GW. Large sample properties of matching estimators for average treatment effects. *Econometrica* 2006; **74**: 235–67.
18 Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J. Am. Stat. Assoc.* 1984; **79**: 516–24.
19 Austin PC. Goodness-of-fit diagnostics for the propensity score model when estimating treatment effects using covariate adjustment with the propensity score. *Pharmacoepidemiol. Drug Saf.* 2008; **17**: 1202–17.
20 Rubin DB. On principles for modelling propensity scores in medical research. *Pharmacoepidemiol. Drug Saf.* 2008; **17**: 1202–17.
21 Lunceford JK, Davidian M. Stratification and weighting via the propensity score in estimation of causal treatment effects: a comparative study. *Stat. Med.* 2004; **23**: 2937–60.

22 Lee BK, Lessler J, Stuart EA. Weight trimming and propensity score weighting. *PLoS ONE* 2011; **6**: e18174.
23 Imbens G. The role of the propensity score in estimating dose-response functions. *Biometrika* 2000; **87**: 706–10.
24 Imbens G, Hirano K. The propensity score with continuous treatments. In: Gelman A and Meng X (eds) *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*. John Wiley and Sons, West Sussex, 2004; 73–84.
25 Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. *Am. J. Epidemiol.* 2006; **163**: 1149–56.
26 Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. *J. Am. Stat. Assoc.* 2000; **95**: 573–85.
27 Qu Y, Lipkovich I. Propensity score estimation with missing values using a multiple imputation missingness pattern (MIMP) approach. *Stat. Med.* 2009; **28**: 1402–14.
28 Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. *Am. J. Epidemiol.* 1995; **142**: 1255–64.
29 Knol MJ, Janssen KJ, Donders AR, Egberts AC, Heerdink ER, Grobbee DE, Moons KG, Geerlings MI. Unpredictable bias when using the missing indicator method or complete case analysis for missing confounder values: an empirical example. *J. Clin. Epidemiol.* 2010; **63**: 728–36.
30 StataCorp. 2013. Stata Statistical Software: Release 13. College Station, TX: StataCorp LP.
31 Brun-Buisson C, Richard JC, Mercat A, Thiebaut AC, Brochard L. Early corticosteroids in severe influenza A/H1N1 pneumonia and acute respiratory distress syndrome. *Am. J. Respir. Crit. Care Med.* 2011; **183**: 1200–6.
32 Castleberry AW, Worni M, Osho AA, Snyder LD, Palmer SM, Pietrobon R, Davis RD, Hartwig MG. Use of lung allografts from brain-dead donors after cardiopulmonary arrest and resuscitation. *Am. J. Respir. Crit. Care Med.* 2013; **188**: 466–73.
33 Devasia RA, Blackman A, Gebretsadik T, Griffin M, Shintani A, May C, Smith T, Hooper N, Maruri F, Warkentin J, *et al*. Fluoroquinolone resistance in Mycobacterium tuberculosis: the effect of duration and timing of fluoroquinolone exposure. *Am. J. Respir. Crit. Care Med.* 2009; **180**: 365–70.
34 Ferrer R, Artigas A, Suarez D, Palencia E, Levy MM, Arenzana A, Perez XL, Sirvent JM. Effectiveness of treatments for severe sepsis: a prospective, multicenter, observational study. *Am. J. Respir. Crit. Care Med.* 2009; **180**: 861–6.
35 Kim SH, Hong SB, Yun SC, Choi WI, Ahn JJ, Lee YJ, Lee HB, Lim CM *et al*. Corticosteroid treatment in critically ill patients with pandemic influenza A/H1N1 2009 infection: analytic strategy using propensity scores. *Am. J. Respir. Crit. Care Med.* 2011; **183**: 1207–14.
36 Pham T, Combes A, Roze H, Chevret S, Mercat A, Roch A, Mourvillier B, Ara-Somohano C *et al*. Extracorporeal membrane oxygenation for pandemic influenza A(H1N1)-induced acute respiratory distress syndrome: a cohort study and propensity-matched analysis. *Am. J. Respir. Crit. Care Med.* 2013; **187**: 276–85.
37 Sellares J, Lopez-Giraldo A, Lucena C, Cilloniz C, Amaro R, Polverino E, Ferrer M, Menendez R *et al*. Influence of previous use of inhaled corticoids on the development of pleural effusion in community-acquired pneumonia. *Am. J. Respir. Crit. Care Med.* 2013; **187**: 1241–8.
38 Tegethoff M, Greene N, Olsen J, Schaffner E, Meinlschmidt G. Inhaled glucocorticoids during pregnancy and offspring pediatric diseases: a national cohort study. *Am. J. Respir. Crit. Care Med.* 2012; **185**: 557–63.
39 Wisnivesky JP, Halm E, Bonomi M, Powell C, Bagiella E. Effectiveness of radiation therapy for elderly patients with unresected stage I and II non-small cell lung cancer. *Am. J. Respir. Crit. Care Med.* 2010; **181**: 264–9.
40 Arabi YM, Khedr M, Dara SI, Dhar GS, Bhat SA, Tamim HM, Afesh LY. Use of intermittent pneumatic compression and not graduated compression stockings is associated with lower incident VTE in critically ill patients: a multiple propensity scores adjusted analysis. *Chest* 2013; **144**: 152–9.
41 Haque NZ, Zuniga LC, Peyrani P, Reyes K, Lamerato L, Moore CL, Patel S, Allen M *et al*. Relationship of vancomycin minimum inhibitory concentration to mortality in patients with methicillin-resistant Staphylococcus aureus hospital-acquired, ventilator-associated, or health-care-associated pneumonia. *Chest* 2010; **138**: 1356–62.
42 Kates M, Swanson S, Wisnivesky JP. Survival following lobectomy and limited resection for the treatment of stage I non-small cell lung cancer ≤1cm in size: a review of SEER data. *Chest* 2011; **139**: 491–6.
43 Kaw R, Pasupuleti V, Walker E, Ramaswamy A, Foldvary-Schafer N. Postoperative complications in patients with obstructive sleep apnea. *Chest* 2012; **141**: 436–41.
44 Miano TA, Reichert MG, Houle TT, MacGregor DA, Kincaid EH, Bowton DL. Nosocomial pneumonia risk and stress ulcer prophylaxis: a comparison of pantoprazole vs ranitidine in cardiothoracic surgery patients. *Chest* 2009; **136**: 440–7.
45 Mirsaeidi M, Peyrani P, Aliberti S, Filardo G, Bordon J, Blasi F, Ramirez JA. Thrombocytopenia and thrombocytosis at time of hospitalization predict mortality in patients with community-acquired pneumonia. *Chest* 2010; **137**: 416–20.
46 Ortiz G, Frutos-Vivar F, Ferguson ND, Esteban A, Raymondos K, Apezteguia C, Hurtado J, Gonzalez M *et al*. Outcomes of patients ventilated with synchronized intermittent mandatory ventilation with pressure support: a comparative propensity score study. *Chest* 2010; **137**: 1265–77.
47 Rineer J, Schreiber D, Katsoulakis E, Nabhani T, Han P, Lange C, Choi K, Rotman M. Survival following sublobar resection for early-stage non-small cell lung cancer with or without adjuvant external beam radiation therapy: a population-based study. *Chest* 2010; **137**: 362–8.
48 Sadatsafavi M, Fitzgerald M, Marra C, Lynd L. Costs and health outcomes associated with primary vs secondary care after an asthma-related hospitalization: a population-based study. *Chest* 2013; **144**: 428–35.
49 Thomas CP, Ryan M, Chapman JD, Stason WB, Tompkins CP, Suaya JA, Polsky D, Mannino DM *et al*. Incidence and cost of pneumonia in medicare beneficiaries. *Chest* 2012; **142**: 973–81.
50 Valles J, Peredo R, Burgueno MJ, Rodrigues de Freitas AP, Millan S, Espasa M, Martin-Loeches I, Ferrer R *et al*. Efficacy of single-dose antibiotic against early-onset pneumonia in comatose patients who are ventilated. *Chest* 2013; **143**: 1219–25.
51 Zilberberg MD, Nathanson BH, Sadigov S, Higgins TL, Kollef MH, Shorr AF. Epidemiology and outcomes of clostridium difficile-associated disease among patients on prolonged acute mechanical ventilation. *Chest* 2009; **136**: 752–8.
52 Attridge RT, Frei CR, Restrepo MI, Lawson KA, Ryan L, Pugh MJ, Anzueto A, Mortensen EM. Guideline-concordant therapy and outcomes in healthcare-associated pneumonia. *Eur. Respir. J.* 2011; **38**: 878–87.
53 Groenwold RH, Hoes AW, Hak E. Impact of influenza vaccination on mortality risk among the elderly. *Eur. Respir. J.* 2009; **34**: 56–62.
54 Malo de Molina R, Mortensen EM, Restrepo MI, Copeland LA, Pugh MJ, Anzueto A. Inhaled corticosteroid use is associated with lower mortality for subjects with COPD and hospitalised with pneumonia. *Eur. Respir. J.* 2010; **36**: 751–7.
55 Mounier R, Adrie C, Francais A, Garrouste-Orgeas M, Cheval C, Allaouchiche B, Jamali S, Dinh-Xuan AT *et al*. Study of prone positioning to reduce ventilator-associated pneumonia in hypoxaemic patients. *Eur. Respir. J.* 2010; **35**: 795–804.
56 Wisnivesky JP, Bonomi M, Lurslurchachai L, Mhango G, Halm EA. Radiotherapy and chemotherapy for elderly patients with stage I-II unresected lung cancer. *Eur. Respir. J.* 2012; **40**: 957–64.
57 Shaheen SO, Northstone K, Newson RB, Emmett PM, Sherriff A, Henderson AJ. Dietary patterns in pregnancy and

respiratory and atopic outcomes in childhood. *Thorax* 2009; **64**: 411–17.
58 Havstad SL, Johnson CC, Zoratti EM, Ezell JM, Woodcroft K, Ownby DR, Wegienka G. Tobacco smoke exposure and allergic sensitization in children: a propensity score analysis. *Respirology* 2012; **17**: 1068–72.
59 Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat. Med.* 2009; **28**: 3083–107.
60 Imai K, King G, Stuart E. Misunderstandings among experimentalists and observationalists about causal inference. *J. R. Stat. Soc. [Ser. A]* 2008; **171**(Pt 2): 481–502.
61 Austin PC. Assessing balance in measured baseline covariates when using many-to-one matching on the propensity-score. *Pharmacoepidemiol. Drug Saf.* 2008; **17**: 1218–25.
62 Westreich D, Cole SR, Funk MJ, Brookhart MA, Sturmer T. The role of the c-statistic in variable selection for propensity score models. *Pharmacoepidemiol. Drug Saf.* 2011; **20**: 317–20.
63 Gibson HB, Silverstone H, Gandevia B, Hall GJ. Respiratory disorders in seven-year-old children in Tasmania: aims, methods and administration of the survey. *Med. J. Aust.* 1969; **2**: 201–5.
64 Burgess JA, Matheson MC, Gurrin LC, Byrnes GB, Adams KS, Wharton CL, Giles GG, Jenkins MA *et al.* Factors influencing asthma remission: a longitudinal study from childhood to middle age. *Thorax* 2011; **66**: 508–13.
65 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**: 550–9.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Table S1 Papers identified that were published in 2013.

Table S2 Papers identified that were published in 2012.

Table S3 Papers identified that were published in 2011.

Table S4 Papers identified that were published in 2010.

Table S5 Papers identified that were published in 2009.