Contemporary Clinical Trials 32 (2011) 731–740

Contents lists available at ScienceDirect

Contemporary Clinical Trials

journal homepage: www.elsevier.com/locate/conclintrial

Review

Are propensity scores really superior to standard multivariable analysis?

Giuseppe Biondi-Zoccai [a,b,*], Enrico Romagnoli [b,c], Pierfrancesco Agostoni [b,d], Davide Capodanno [b,e], Davide Castagno [b,f], Fabrizio D'Ascenzo [b,f], Giuseppe Sangiorgi [b,g], Maria Grazia Modena [a]

[a] Division of Cardiology, University of Modena and Reggio Emilia, Modena, Italy
[b] Meta-analysis and Evidence based medicine Training in Cardiology (METCARDIO), Ospedaletti, Italy
[c] Division of Cardiology, Policlinico Casilino, Rome, Italy
[d] Division of Cardiology, University Medical Center Utrecht, Utrecht, The Netherlands
[e] Division of Cardiology, University of Catania, Catania, Italy
[f] Division of Cardiology, University of Turin, Turin, Italy
[g] Division of Cardiology, University of Tor Vergata, Rome, Italy

ARTICLE INFO

Article history:
Received 25 February 2011
Received in revised form 9 May 2011
Accepted 11 May 2011
Available online 16 May 2011

Keywords:
Bias
Cox proportional hazard analysis
Logistic regression
Multivariable analysis
Propensity score

ABSTRACT

Clinicians often face difficult decisions despite the lack of evidence from randomized trials. Thus, clinical evidence is often shaped by non-randomized studies exploiting multivariable approaches to limit the extent of confounding. Since their introduction, propensity scores have been used more and more frequently to estimate relevant clinical effects adjusting for established confounders, especially in small datasets. However, debate persists on their real usefulness in comparison to standard multivariable approaches such as logistic regression and Cox proportional hazard analysis. This holds even truer in light of key quantitative developments such as bootstrap and Bayesian methods. This qualitative review aims to provide a concise and practical guide to choose between propensity scores and standard multivariable analysis, emphasizing strengths and weaknesses of both approaches.

Contents

* Corresponding author at: Division of Cardiology, University of Modena and Reggio Emilia, Via Del Pozzo 71, 41124 Modena, Italy. Tel.: +39 059 422 57 83; fax: +39 059 422 37 14.
E-mail address: gbiondizoccai@gmail.com (G. Biondi-Zoccai).

1551-7144/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.cct.2011.05.006

The great advances in science usually result from new tools rather than from new doctrines
Freeman Dyson

1. Introduction: meet the bias

Clinical decision making is based on the appraisal of causality. In other words, we choose a given treatment instead of another one only if we are reasonably convinced that using it will cause specific effects (hopefully favorable ones) on our individual patient [1]. All statistical analyses are best viewed in this framework of causality assessment. Indeed, evidence of a statistically significant association (in frequentist terms) or highly probable association (in Bayesian terms) is just a piece of the puzzle to finally conclude that a given treatment or exposure is causing a given effect (e.g. aspirin prevents recurrent myocardial infarction events, or cigarette smoking causes lung cancer) [2]. These key criteria have been spelled out in 1965 by Sir Austin Bradford Hill, whose seminal paper makes an interesting reading even now: analogy, biological gradient, coherence, consistency, experiment, plausibility, specificity, strength, and temporality [1]. Thus, factors having causal effects should be clearly distinguished from variables exhibiting casual (i.e. random) association.

Whenever a treatment or a preventive means has a dramatic impact on outcomes, no statistical analysis is usually needed. A small case series or small cohort (i.e. group of patients followed thoroughly over time to capture retrospectively or prospectively outcomes of interest) usually suffice in such cases. For instance, few would argue against using a parachute when jumping out of an airplane high in the sky or administering systemic antibiotics in a patient with acute infective endocarditis [3]. However, most treatment alternatives in current clinical practice are unlikely to cause huge effects in a single individual. We must thus rely on large sample sizes and appraise mild or moderate effects. This was the reason behind the first formal randomized clinical trial conducted by the Streptomycin Tuberculosis Trials Committee in the 1940s [4], which aimed to appraise the role of antimycobacterials in patients with tuberculosis, typically a disease with low response rates occurring during long and variable periods of time. Randomized clinical trials, alone or combined with systematic reviews and meta-analyses, are the uppermost ladder in the hierarchy of evidence based medicine, and the vast majority (if not all) clinical decisions should be based on the integration of their results with the individual practice of medicine (thus encompassing cost and patient values) [5]. Unfortunately, randomized clinical trials might not be available at all to guide a specific diagnostic or treatment choice, or they might be of low internal or external validity. Specifically, their findings might be biased by issues in patient selection, subject attrition, event adjudication, or therapeutic performance.

Thus, we often rely on observational studies, mostly retrospective or based on administrative datasets, but also not uncommonly prospective studies (i.e. registries). However, extracting credible results from such studies requires sophisticated statistical methods, as univariate and bivariate analyses cannot adjust effect estimates taking into account confounding factors. A hypothetical example highlighting the issue of bias and confounding is the following. A cohort of apparently healthy subjects is followed prospectively over 20 years in order to identify risk factors for lung cancer. Bivariate analysis apparently demonstrates a statistically significant association between the number of matches (such as those required to light a cigarette) used by study participants over the years and the risk of lung cancer. Is such a "demonstration" of statistical significance convincing enough for us to ban the sale of matches nationwide to reduce the incidence of lung cancer? Further careful analysis shows actually that another variable, i.e. cigarette smoking, is also strongly associated with lung cancer. By appraising the independent impact of matches on the risk of lung cancer while simultaneously adjusting for cigarette smoking, we recognize that the apparent association between matches and lung cancer was only due to the biasing effect of a confounding factor (namely cigarette smoking).

Our present work aims to provide a concise review of current methods to perform multivariable analysis in order to adjust for the role of such covariates when estimating effects of risk factors or interventions, with a particular emphasis on a key recent item in the biostatistical armamentarium, propensity scores. The reader should however bear in mind that dozens of forms of bias can undermine a clinical study, and multivariable methods cannot take into account all of them as well as other unknown confounders, as they can only adjust for what is accurately measured and explicitly forced into the multivariable model [6].

2. Scope of multivariable approaches

Multivariable analysis aims to explore the relationship between a dependent variable (e.g. risk of death) and two or more independent variables (a.k.a. moderators, covariates, or factors; e.g. age and serum cholesterol level) appraised simultaneously. Thus, they estimate the independent impact of a given covariate on the dependent variable, by concomitantly adjusting for the contributions of all the other covariates present in the predictive model. Hundreds of examples of multivariable analyses are permeating clinical practice. For instance, the Framingham Heart Study and the Framingham Heart Score are largely based on multivariable logistic regression and Cox proportional hazard models [7].

Several types of multivariable analysis are available, from stratification and matching, to simple linear models to predict blood pressure or highly complex hierarchical (i.e. mixed-effect) models taking into account multiple levels and clusters in the data (Table 1). Specifically, matching and stratification have been historically the first methods to adjust for confounders, given the shortage of computational resources up to a few decades ago. However, they are significantly limited in their robustness and applicability by the fact that both stratification and matching can be performed only on a handful of covariates, thus being impractical whenever we have more than 4–5 confounders to take care of. Whereas discriminant analysis has also been historically crucial in the identification of independent predictors of events, such as for the seminal Norris coronary prognostic index [8], it is nowadays seldom used because of limitations in modeling capabilities, including appraisal of overall performance and interactions. Classification and regression tree (CART) analyses have similar strengths and weaknesses in comparison to discriminant analysis, but they

Table 1
Selected approaches to perform multivariable analysis.

Method	Key features
Bayesian methods	Computation-intensive methods able to generate posterior probability distributions appraising multiple parameters from simple as well as highly complex models using Markov chain Monte Carlo techniques (MCMC). Occasionally used. Implementation requires considerable expertise in Bayesian statistical methods and programming in WinBUGS or similar software packages.
Bootstrap	Computation-intensive resampling technique with replacement making no assumptions regarding underlying population distribution, able to generate inferences from simple as well as highly complex datasets. Usually combined with a standard multivariable approach such as logistic regression or Cox analysis. Occasionally used. Implementation requires considerable expertise in R or similar software packages.
Classification and regression tree (CART)	A recursive partitioning statistical method which builds classification and regression trees for predicting continuous dependent variables (regression) and categorical predictor variables (classification). Rarely used. Implementation requires some expertise in SPSS or similar software packages.
Cox proportional hazard analysis	A parametric statistical method to perform survival analysis including censored data and adjusting for covariates. Several assumptions must be met for valid results, including an event per variable ratio >8–10 and proportionality of hazard over time. Frequently used. Implementation requires some expertise in SPSS or similar software packages.
Discriminant analysis	A statistical technique to determine the class of an observation based on a set of variables known as predictors or input variables. Rarely used. Implementation requires some expertise in SPSS or similar software packages.
Exact methods	Computation-intensive methods to approximate the exact probability of a given statistical model, irrespective of its complexity, by means of Monte Carlo procedures. Rarely used. Implementation requires considerable expertise in SAS or similar software packages.
Instrumental variable analysis	Econometric method used to remove the effects of hidden bias in observational studies, based on two key characteristics: it is highly correlated with treatment and does not independently affect the outcome, so that it is not associated with measured or unmeasured patient health status. Rarely used. Implementation requires substantial expertise in Stata or similar software packages.
Logistic regression	A parametric statistical method to appraise the impact of independent variable(s) on a categorical dependent variable. Binary logistic regression is most commonly used in clinical research and focuses on a dichotomous dependent variable. Several assumptions must be met for valid results, including an event per variable ratio >8–10 and lack of collinearity or overfitting. Frequently used. Implementation requires some expertise in SPSS or similar software packages.
Matching	A method to combine one case with one or more controls selected according to specific matching criteria (e.g. difference in age <5 years) in order to adjust for key confounders. A major limitation is that it can usually be employed only for a few variables, as matching subsets increase exponentially in keeping with the number of matching factors (e.g., with n dichotomous matching variables, the final number of matching subsets is equal to 2^n). Occasionally used. Implementation of simple matching requires some expertise in Excel or similar spreadsheets, but more complex matching procedures require more sophisticated softwares and skills.
Mixed-effect methods	A statistical method to appraise a hierarchical model combining fixed and random-effect relationships between variables and enabling also the appraisal of complex clusters in the dataset. It is usually based on a generalized linear model with specific link functions. Occasionally used. Implementation requires considerable expertise in SPSS or similar software packages.
Propensity score	A score built according to the likelihood or propensity that a given treatment has been administered to a subject according to all pertinent covariates that influence this treatment choice. It thus acts as a proxy between treatment, confounders and, eventually, events. Several key assumptions must be met for the propensity score to be valid, and implementation requires combination with matching, stratification, or regression. Frequently used. Implementation requires some expertise in SPSS or similar software packages.
Stratification	A method to adjust an analysis by dividing the population in a given number of subsets, perform separate analyses in such subsets, and then obtain an average estimate of effect. Major limitations are that it is feasible and meaningful only when effects are consistent across strata, and that it can usually be employed only for few variables, as strata increase exponentially in keeping with the number of stratification factors. Rarely used. Implementation requires some expertise in SPSS or similar software packages.

have the immediate appeal of clarity and relevance for the busy clinical reader [9].

Binary logistic regression and Cox proportional hazard analysis are currently the most commonly used multivariable approaches in clinical research, as they provide clear guidance to identify predictors of binary outcomes such as death or rehospitalization, if a number of key assumptions are met. In addition, they maintain the flexibility for complex modeling of interaction terms and also the capability to appraise predictive power, discrimination (e.g. by means of c-statistic, i.e. area under the curve [AUC] of the receiver operator characteristic [ROC]), and calibration (e.g. by means of goodness of fit tests) [10–12].

A major advancement in the field of multivariable analysis has been the introduction in 1983 by Rosenbaum and Rubin of propensity scores, which can be defined as the conditional probability of being treated given the covariates [13]. Indeed, after having modeled the distribution of the treatment indicator variable given the observed covariates, the ensuing propensity score can be used to reduce selection bias through matching, stratification (i.e. subclassification), regression adjustment, or some combination of all three [14]. Propensity scores have recently become very successful among clinical researchers (Fig. 1), as they are proposed more and more often as the gold standard multivariable approach to adjust for selection bias and ensuing confounders in non-randomized studies.

Other more challenging methods which may help in the development or validation of complex statistical models include bootstrap [15,16], and other resampling approaches based on Monte Carlo simulations (which are however not per se alternatives to propensity scores) [17], exact logistic

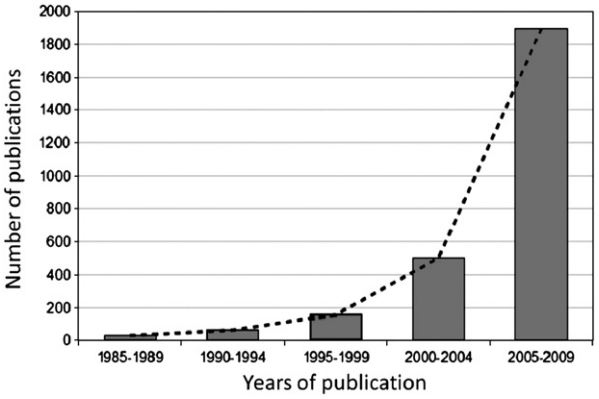

Fig. 1. Exponential increase in publications indexed in MEDLINE/PubMed and reporting on or exploiting propensity scores for multivariable adjustment. PubMed queried on 29 May 2010 with the following string: propensity AND (score* OR match*).

regression [18,19], instrumental variables [20], and Bayesian analyses [21]. Nonetheless, such approaches have so far been used less commonly in clinical research and further refinements in statistical methods and user-friendliness are needed to foster their wider adoption.

Despite the availability of so many different approaches, no single method in itself can be considered perfect or immune from key drawbacks [22]. In addition, as all such methods are usually applied in datasets stemming from finite samples with limited size (e.g. <2000 patients, <100 events, and/or excluding direct population estimates). Thus, only similar results from subsequent independent studies can effectively provide proof of external validity [5].

3. Standard multivariable analysis

Binary logistic regression is a statistical technique exploiting the logit function, defined as the ln(p/[1-p]), where ln is the natural logarithm and p is the probability of an event [10]. The logit function transforms a dependent variable ranging between 0 and 1 such as a probability of an event (e.g. the probability of dying after an acute myocardial infarction) into a variable stemming from $-\infty$ to $+\infty$. Thus, event probabilities can be appraised as a linear regression function in order to appraise the logit of the probability of an event (dependent variable) given one or more dependent variables (e.g. age of the patient, presence of diabetes mellitus, or usage of a novel coronary drug or device). In other words, binary logistic regression is a formidable tool to appraise multivariate predictors of dichotomous events, and evaluate the independent predictive role of one or more independent variables of interest [16]. Binary logistic regression is commonly used and reported in non-randomized trials, but possibly maintains a role also in small-size randomized clinical trials, to adjust for confounders [23]. Yet, the reader should bear in mind that logistic regression is different from linear regression and usually covariate adjustment does not necessarily increase power with control of covariates in such pilot randomized trials [23].

Despite such strengths, logistic regression should not be used or interpreted naively, as a number of key assumptions must be met to ensure that its results are valid [10]. First, an overfit model can be highly predictive in the dataset in which the model was developed, but not in one in which it is validated or tested. Multicollinearity, whereby covariates present in the model are unduly associated (e.g. Pearson correlation coefficient >0.9 or <−0.9), can be present but need to be controlled for in some way (e.g. interaction terms are always collinear with their component main effects). Outliers should not have an excessive weight on the analysis. Matching (e.g. clustering) features should not be present (otherwise a conditional logistic regression analysis or generalized estimating equations should be used). Residual variability follows a binomial distribution and what remain unexplained leads to overdispersion. However, it is very difficult to define and interpret residuals from binary outcome models. Finally, a crucial issue is maintaining an event per variable ratio >8–10, as logistic regression looses power and becomes biased whenever ≤8–10 events are included in the dataset for each independent variable or covariate forced in the initial model [24,25].

Variable selection for final entry in the model is also a pivotal aspect of model building for any multivariable analysis, including logistic regression, and can be performed in several ways, including automatic stepwise approaches such as the forward or backward elimination techniques. We recommend however to avoid in most cases automatic algorithms with stepwise selection, but rather rely on prior epidemiologic evidence (i.e. established association from prior well conducted experimental or clinical studies) and strong associations (e.g. $p<0.10$ or $p<0.05$ at bivariate analysis) stemming from the specific dataset of interest [22,26]. Nonetheless, we caution that use of arbitrary p-value cutoffs for variable selection without appropriate consideration of scientific principles can be problematic, particularly if there is a lot of data dredging and variable transformation tried. P-values are also highly sensitive to sample sizes. In addition, whenever performing a logistic regression analysis, attention should be paid to appraise prediction (e.g. by means of Nagelkerke R^2), which quantify the variability in the dependent variable explained solely by the model), discrimination (e.g. by means of c-statistic, i.e. area under the curve of the receiver operator characteristic), which describes the probability of correctly discriminating cases from non-cases, and calibration (e.g. by means of goodness of fit tests such as the Hosmer–Lemeshow test, which is however often fraught with low power in small sample sizes and rarely rejects the null hypothesis of good calibration) [10,22].

The other mainstay in multivariable analyses, Cox proportional hazard analysis, has strengths and weaknesses similar to logistic regression, with the notable advantage of addressing differences in follow-up duration and censored data [11,12]. The hazard function, which forms the basis of Cox analysis, is defined as the event rate at time t conditional on survival until time t or later. It is based on a semiparametric model whereby the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate, itself inversely proportional to the survival rate [27]. Censored patients are exploited to compute hazards and are assumed in the Cox model to fail at the same rate as the non-censored, but are not supposed to survive to the next time point. Being a member of the larger family of survival methods, Cox analysis is ideally suited to estimate and interpret survival and hazard curves, to compare two or more survival functions, and, in particular, to

assess the relationship of explanatory variables to survival time controlling for covariates and known confounders. However, the relationship between hazard rate (inherently conditional) and survival rate (inherently unconditional) remains complicated. Indeed, a hazard function is a statistical model describing the relationship between instantaneous hazard probability conditionally to the events at previous time points. Moreover, in addition to allowing time-varying covariates (i.e. predictors), the Cox model may be generalized to time-varying coefficients as well. Additional key assumptions, on top of those required for logistic regression, must be met to ensure that the Cox model is valid. In particular, a key requirement is the proportional hazards assumption, i.e. the assumption that covariates multiply hazard. A simple way to check this assumption is plotting the graph of log(−log(survival)) versus log of survival time, whereby log is the natural or base 10 logarithm. This should result in largely parallel lines if the hazard function is proportional. Another more general issue is non-informative censoring. To satisfy this assumption, the design of the study must ensure that the mechanisms giving rise to censoring of individual subjects are not related to the probability of an event occurring. In other words, censored and non-censored patients should have the same chance of failure, the chance of censoring should be independent of failure, censored patients should be representative of those at risk at censoring time, and censored patients should be supposed to survive to the next time point. Finally, similarly to what applies to logistic regression, an event per variable initially entered into the model ratio >8–10 is required to obtain stable, unbiased, and precise results [28,29].

4. Propensity scores

The propensity score is defined as the conditional probability of receiving an exposure or treatment given a vector of measured covariates, and can be used to adjust for selection bias when assessing causal effects in observational studies. In other words, propensity scores act as a proxy between cases and covariates influencing exposure, and thus can be used instead of such covariates to simplify the analysis plan and increase robustness. Use of propensity scores represents a quasi-empirical correction strategy attempting to reduce bias of treatment estimates in non-randomized studies [13,14,30]. Indeed, they have been referred as a "balancing variable", i.e. a proxy variable which may summarize different confounding factors into a single dimension, and thus can be exploited to achieve balance between 2 study groups. However, only variables acting well before or shortly after the beginning of the treatment exposure of interest should theoretically be used to generate propensity scores (e.g. a propensity score for coronary stenting versus balloon-only angioplasty should include as generating variables diabetes mellitus, but not whether or not the patient has undergone 6-month post-procedural angiographic follow-up). Thus, the propensity score does not usually use the outcome to identify the confounders, while a clinical risk score does.

Table 2 shows some key steps necessary to implement or appraise a propensity score, in keeping with recommendations from Weitzen et al [32]. First, the best way to generate a propensity score in our opinion is to conduct a logistic regression in a non-parsimonious fashion by calculating regression coefficients (i.e. beta coefficients) for all variables acting well before or shortly before the beginning of treatment exposure, and including key interaction terms. The outcome of interest should not be included in the model, and overfitting does not apply to this phase of the analysis. The results of this non-parsimonious logistic regression are then exploited to build the propensity score according to the following formula: propensity score $= 1/(1+\exp^{\text{model}})$, whereby the model has the form of $\text{alpha} + \text{beta}_1{}^{*}\text{x} + \text{beta}_2{}^{*}\text{y} + \ldots + \text{beta}_N{}^{*}\text{z}$.

Once the propensity score has been computed for each case and its performance has been appraised in terms of discrimination and calibration, we face the decision of how to use this balancing variable in our analysis. Weitzen et al. suggest that Hosmer–Lemeshow goodness of fit tests and c-statistic should not be used to appraise, respectively, calibration and discrimination of the propensity score, as

Table 2
Key items for the correct conduct and appraisal of propensity scores.

Item	Elaboration
Methods for variable selection	Which method for variable selection was used (e.g. non-parsimonious, parsimonious with backward stepwise algorithm, ….).
Event per variable ratio	Whether an 8–10 event per variable ratio could be achieved when developing the propensity score.
Balance	Whether balance on the potential confounders included in the propensity score model between treatment groups has been achieved. This particularly applies to matching and stratification procedures.
Collinearity	Whether two or more confounders and/or the exposure variable in the model are highly correlated with each other.
Continuous variable conformity with linear gradient	Whether propensity scores follow a linear gradient.
Interactions	How interaction terms were handled, as both inclusion of unnecessary interactions and exclusion of meaningful ones may bias the propensity score.
Assessment of model fit (calibration)	Whether the distances between the observed (treatment—yes or no) and the predicted outcome from the model (propensity score) are small and unsystematic. This is usually formally appraised with the Hosmer–Lemeshow goodness of fit test.
Discrimination of model	How well the predicted probabilities derived from the model classify patients into their actual treatment group. This is usually quantified with c-statistic, receiver operator characteristic, and area under the curve.
Adjustment method	How the propensity score is employed to adjust for confounders: matching, stratification into quantiles (e.g. quintiles), or regression modeling. Matching is nowadays considered the most effective means.
Matching method	Which specific matching approach was employed.

even biased scores may appear satisfactory [33]. However, this proof of lack of sensitivity is by no means a lack of specificity. We thus maintain our stance that both Hosmer–Lemeshow goodness of fit tests and c-statistic should be routinely computed and reported whenever propensity scores are used to appraise, respectively, calibration and discrimination. Nonetheless, researchers should bear in mind that if the discrimination is too good then the patients cannot be adequately matched or stratified in the two groups.

Despite this standard approach to propensity scores, some modifications have been recently proposed. It has been claimed that logistic regression methods are inefficient to define propensity scores and machine learning methods (e.g. boosted classification and regression trees) should be used instead [34]. Moreover, one of the foremost expert in the field, Peter C. Austin, has proposed to include in the variables used to derive the propensity score also those related to the outcome(s) of interest, in order to minimize the impact of residual confounding [30]. In other words, if I am interested in building a propensity score related to the use of drug-eluting versus bare-metal coronary stents for restenosis prevention in patients with coronary artery disease, I would include as covariates to generate the propensity score even those variables which come after my specific therapeutic choice (e.g. performance of angiographic follow-up). However, this shifts the focus of the propensity score from confounders determining selection bias up to treatment choice but not afterwards to confounders playing a role even after this event. Moreover, this change limits the application of the same propensity score to different outcome (i.e. dependent) variables that might be related to altogether different covariates. Finally, it may be argued that this makes propensity scores much more similar to clinical risk scores (such as the Framingham Heart Score or the Thrombolysis in Myocardial Infarction score).

Choosing the most appropriate way to exploit propensity scores and incorporate them into the analysis remains challenging, as several matching approaches are available (e.g., with calipers of width of 0.2 of the standard deviation of the logit of the propensity score, Mahalanobis metric matching, or greedy matching), as well as stratification (e.g. in quintiles or deciles), generation of inverse probability of selection weights, or incorporation into straightforward regression models [14,31,34–36]. We do not favor matching as it may discard several cases. Yet some evidence in support of the superiority of this approach over stratification and regression strategies has been provided [37,38]. In particular, at least when focusing on relative risks, matching may result in less bias than stratification on quintiles, even if stratification on quintiles or deciles provides more precise effect estimates [38]. It should also be borne in mind that if stratification is employed, each stratum should have a reasonably overlapping and symmetric composition of patients with similar propensity scores but treated with different strategies. If this comparability assumption is not met, stratification, matching and regression adjustment based on propensity scores may be biased and/or imprecise.

There is also a dark side of the moon. As Rubin poignantly stated, “it is important to keep in mind that even propensity score methods can only adjust for observed confounding covariates and not for unobserved ones.” [39]. In addition, most propensity scores in published studies have been so far poorly designed, analyzed, or reported. A comprehensive appraisal of 44 articles published in major cardiology journals between 2004 and 2006 and exploiting propensity scores was recently reported [40]. In this work, 45% of the studies did not provide adequate information on how the propensity score matched pairs were obtained, 32% did not report whether matching on the propensity score balanced baseline characteristics between matched treated and untreated subjects in the sample, only 9% reported appropriate means to compare baseline characteristics, only 25% used statistical methods appropriate for the analysis of matched data, and only 5% described the matching method used, assessed balance in baseline covariates by appropriate methods, and also used appropriate statistical methods to estimate the treatment effect and its significance. Indeed, whenever employing matched pairs, standard methods for non-clustered data (including chi-squared tests, Gossett/Student unpaired t tests, Wilcoxon rank sum tests, logistic regression, Kaplan–Meier survival curves, or Cox regression) are inappropriate as they fail to take into account the clustered structure of the data. Conversely, the following methods should be used: Gossett/Student paired t tests, Wilcoxon signed ranks tests, McNemar test for correlated binary outcomes, conditional logistic regression, logistic regression models based on generalized estimating equations, or Cox proportional hazards models stratified on the matched pairs [41].

Other major caveats involve the impact of missing data, which are often present in observational studies and limit model building by discarding several cases without complete data or require extensive data imputation, and the combination of covariates which impact outcomes as well as those not impacting outcomes. For instance, in an observational study comparing coronary artery bypass grafting and percutaneous coronary intervention for severe coronary artery disease, propensity scores are similarly modified by a variable impacting both on treatment choice and mortality (e.g. insulin-dependent diabetes mellitus which fares better with surgery but also has an unfavorable prognosis), and by a variable with minor or no effect on mortality but strongly associated with treatment exposure (e.g. nickel allergy, which contraindicates coronary stenting but bears no meaningful prognostic effect). The implications of such similar impact on propensity scores are unclear. In any case, a final cautionary statement involves the clinical interpretation of propensity scores, which should never be interpreted as clinical risk scores such as the Thrombolysis in Myocardial Infarction score [42].

5. And the winner is…

Of course in science there is never a single winner or a single loser. In fact, it is difficult to clearly identify the best method to adjust for confounders in non-randomized studies, as both standard multivariable methods and propensity scores have key limitations, and none is able to take into account unknown confounders (Fig. 2; Table 3). It is also clear that in many cases both approaches provide similar results [43,44].

For instance, as a typical blow to the effort of identifying any single method which is better than all the others, Stukel et al. compared 4 analytic methods for removing the effects of

selection bias in an observational study including 122,124 patients with acute myocardial infarction: multivariable regression, propensity score risk adjustment, propensity-based matching, and instrumental variable analysis were all employed and compared [45]. Results from Cox proportional hazards regression, propensity score risk adjustment and propensity score matching were all similarly and strongly in favor of cardiac catheterization. However, a different technique, instrumental variable analysis, showed a notably weaker impact on mortality. Even instrumental variable analysis has however its own caveats, and is actually more suitable for policy questions than specific clinical issues. Nonetheless, instrumental variable analysis, by attempting to adjust for unmeasured confounders, retains a key role in clinical research, given its superior performance in the identification/use of an instrument [20].

Some authors have actually suggested that propensity scores do outperform standard multivariable methods. Martens et al. have performed a careful simulation study comparing logistic regression and propensity scores based on quintiles, showing that the adjusted treatment effect in logistic regression is usually, and even in large datasets, further away from the true marginal treatment effect than the adjusted effect of quintiles of propensity scores [46]. Thus, they concluded that propensity score methods usually yield treatment effect estimates that are closer to the true marginal treatment effect than a logistic regression model in which all confounders are appraised.

Conversely, Cepeda et al. have conducted extensive Monte Carlo simulations to compare logistic regression with propensity scores in terms of bias, precision, empirical coverage probability, empirical power, and robustness when the number of events is low relative to the number of confounders initially entered in the model [25]. In this very interesting and carefully conducted work, bias in logistic regression decreased as the number of events per confounder increased, whereas with the propensity score bias decreased as the strength of the association of the

Table 3
Pros and cons of standard multivariable approaches and propensity scores.

Method	Pros	Cons
Binary logistic regression/Cox proportional hazard analysis	• Established approach with relatively straightforward interpretation • Exploitable to build clinical prediction models and tools • Enable complex model building approaches and iterations	• Several key assumptions must hold true (e.g. lack of collinearity and overfitting) • Biased and with low power if event per variable ratio >8–10
Propensity scores	• Succinctly synthesize several contributors of confounding into a single variable • Intuitive appeal as quasi-randomized adjustment method • Suitable even when event per variable ratio <8–10	• Application to multiple treatment comparisons not straightforward • Several key assumptions must hold true (e.g. acceptable discrimination and calibration) • Must be combined with other traditional multivariable methods, such as matching, stratification, or regression • Low power if event per variable ratio >8–10 • Interpretation difficult and often mistaken as risk score

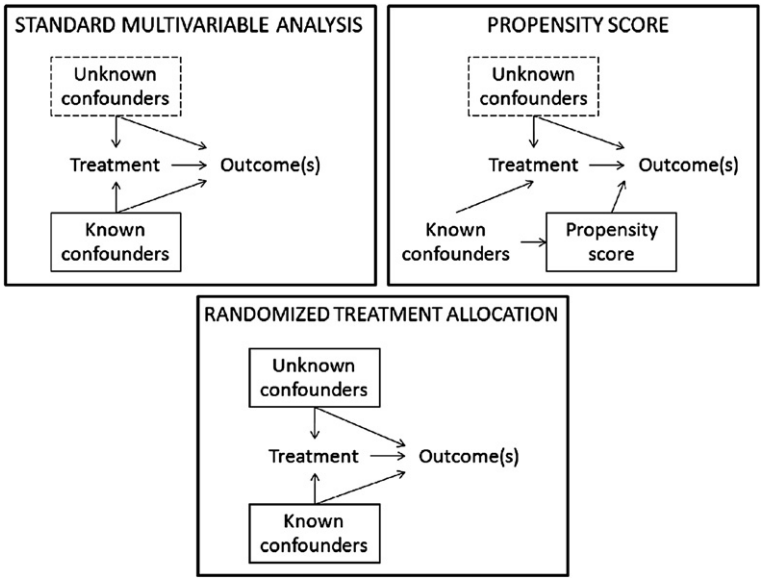

Fig. 2. Rationale underlying standard multivariable analysis for non-randomized studies (top left panel), propensity score adjustment for non-randomized studies (top right panel), and randomized treatment allocation with ensuing bivariate analysis (bottom panel). Confounders enclosed in continuous lines are adjusted for, whereas those enclosed in dashed lines are not adjusted for. Propensity score adjustment for non-randomized studies can also directly appraise known confounder (e.g. when fitted in regression models). Notably, only randomized treatment allocation with ensuing bivariate analysis also takes into account unknown or unmeasured confounders.

exposure with the outcome increased. Thus, propensity scores proved less biased, more robust, and more precise than logistic regression when there were ≤7 events per confounder, whereas the propensity score empirical coverage probability decreased after ≥8 events per confounder. This was at odds with the fact that logistic regression had an increasing empirical coverage probability as the number of events per confounder increased. In conclusion, these authors found that propensity scores are a good multivariable technique when there are ≤7 events per confounder, but with a suboptimal (35%–60%) empirical power. On the other hand, logistic regression (or Cox proportional hazard analysis) is the first choice approach when there are ≥8 events per confounder.

Similarly critical results challenging the hype surrounding propensity scores have been reported by Austin et al., who demonstrated that propensity scores developed using administrative data do not necessarily balance patient characteristics contained in clinical studies and that measures of treatment effectiveness were attenuated when obtained using clinical data compared to administrative data [47]. Accordingly, Mansson et al. have showed that propensity scores may also yield less precise and robust estimates when applied to case–control or case–cohort studies, where there might be artifactual effect modification of the odds ratio by level of propensity score [48].

It may truly appear as a paradox the fact that propensity scores, which are probably superior to standard multivariable methods only in small datasets, actually work better in larger samples, yet still missing the performance of logistic regression or Cox proportional hazard analysis. Indeed, in small observational studies, substantial imbalances of some covariates may be unavoidable despite thorough subclassification or matching using a sensibly estimated propensity score [39]. Conversely, the larger the study, the smaller are such imbalances, but the lower the statistical power of propensity scores in comparison to standard multivariable methods [25,49]. In conclusion, Shah et al. thoughtfully suggested that in medio stat virtus, by reviewing 43 studies including 78 exposure–outcome associations in which both propensity scores and traditional regression models were used [50]. They found similar findings in 90% cases, with all discrepancies due to a statistically significant association at regression analysis which was not observed with propensity scores, whose quality of implementation was however variable.

Our final recommendation is thus that, whatever method you employ, careful analysis and reporting are mandatory to enable appropriate appraisal of the reported findings and, whenever necessary, suitable replication. The bottom line is also that propensity score methods are not meaningfully superior to standard multivariable approaches when adequate assumptions for logistic regression and Cox proportional hazard analysis are met and, in particular, when the event per variable ratio is >8–10 (Fig. 3).

6. Any suitable alternative?

Discussing other analytical approaches to complex inference is beyond the scope of this review, but hierarchical (mixed-effect), exact regression, instrumental variable analyses, bootstrap, and Bayesian methods can all prove to be powerful, precise, and robust [12,15,18,20,21]. However, these approaches are not recommended for inexperienced researchers, as more assumptions are usually required for these complex methods. In addition, with the notable exception of the distribution-free bootstrap, assumptions on specific probability distributions complicate the application and validation of these methods. Last but not least, routine adoption of these analytical tools is slowed by hurdles in programming in the suitable statistical packages.

7. Future perspective

What does the future hold for statistical methods dedicated to non-randomized studies? Prognostication and divination are difficult tasks, but our informed guess is that data mining packages will become more and more common, enabling most researchers to conduct extensive analysis runs with complex hierarchical models. Indeed, bootstrap is likely

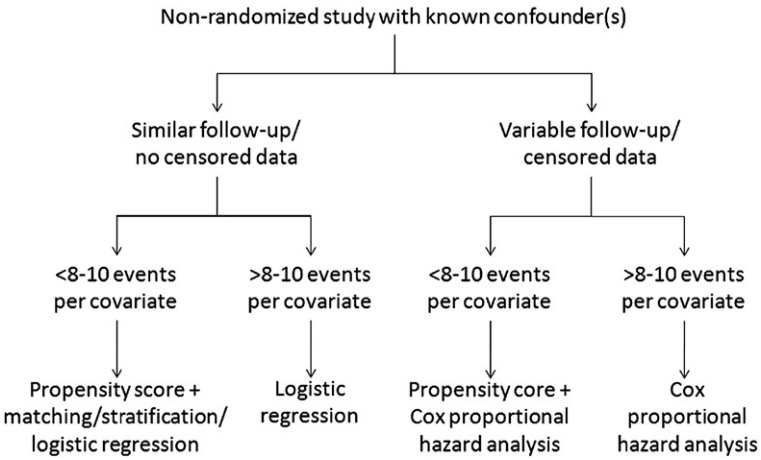

Fig. 3. Practical algorithm for multivariable analysis of non-randomized clinical studies. The event per variable ratio is proposed as the key criterion to justify the use of propensity scores. However, exact analytic methods (e.g. exact logistic regression) or resampling methods (e.g. bootstrap) can also be used instead of propensity scores when there are 7 or fewer events per variable.

to become mainstream in a few years, as several bootstrapping routines have already been added to the last version of SPSS (IBM, Armonk, NY, USA). Unfortunately, the incorporation of these advanced statistical techniques in user-friendly statistical programs may lead to further increase in inappropriate application of certain statistical analyses. Therefore, it remains important for not only researchers to learn appropriate statistical techniques but also for peer reviews to identify potential errors and recommend the correct analysis to be performed, and for readers to maintain a vigilant and critical stance whenever appraising non-randomized clinical studies.

A further development could be the increased user-friendliness of statistical packages to perform Bayesian analyses [51], which could also put into a larger context the current debate on the comparison of propensity scores and standard multivariable techniques, further showing that each approach has its own pros and cons, and that both should best be viewed as complementary rather than alternative.

8. Conclusions

Since their introduction, propensity scores have proved beneficial to adjust for confounders in small datasets of non-randomized studies, where they clearly appear less biased, more robust, and more precise than standard multivariable methods. Their performance in larger datasets with at least 8–10 events per variable is similar or even worse to that provided by logistic regression or Cox proportional hazard analyses, as demonstrated by Cepeda et al., [25]. Awaiting novel additional methods to adjust for known and unknown confounders, clinical researchers should be aware that currently available adjustment methods including propensity scores (yet with the notable exclusion of instrumental variable analysis), can only address observed confounders, but dangerously overlook unobserved ones.

Funding

None.

Conflict of interest

None.

Acknowledgements

None.

References

[1] Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965;58:295–300 *Key paper detailing the basic principles for the scientific appraisal of causality.
[2] Bayarri MJ, Berger JO. The Interplay of Bayesian and frequentist analysis. Stat Sci 2004;19:58–80.
[3] Smith GC, Pell JP. Parachute use to prevent death and major trauma related to gravitational challenge: systematic review of randomised controlled trials. BMJ 2003;327:1459–61.
[4] Streptomycin in Tuberculosis Trials Committee. Streptomycin treatment of pulmonary tuberculosis. A Medical Research Council investigation. Br Med J 1948;2:769–82.
[5] Guyatt G, Rennie D, Meade MO, Cook DJ. Users' guide to the medical literature: a manual for evidence-based clinical practice. 2nd edition. New York: McGraw-Hill; 2008. **Comprehensive textbook detailing the principles of evidence based medicine as well as clinical research.
[6] Sackett DL. Bias in analytic research. J Chronic Dis 1979;32:51–63.
[7] D'Agostino RB, Lee ML, Belanger AJ, Cupples LA, Anderson K, Kannel WB. Relation of pooled logistic regression to time dependent Cox regression analysis: the Framingham heart study. Stat Med 1990;9:1501–15.
[8] Norris RM, Brandt PWT, Caughey DE, Lee AJ, Scott PJ. A new coronary prognostic index. Lancet 1969;1:274–8.
[9] Kastrati A, Schömig A, Elezi S, et al. Predictive factors of restenosis after coronary stent placement. J Am Coll Cardiol 1997;30:1428–36.
[10] Hosmer DL, Lemeshow S. Applied logistic regression. New York: John Wiley and Sons; 2000. *Detailed textbook on logistic regression.
[11] Cox DR, Oakes D. Analysis of survival data. New York: Chapman & Hall; 1984.
[12] Harrell FE. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. New York: Springer; 2001. *Detailed textbook on regression and survival analysis.
[13] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
[14] D'Agostino Jr RB. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81 **Thorough review on the rationale and implementation of propensity scores.
[15] Efron E, Tibshirani R, Tibshirani RJ. An introduction to the bootstrap. New York: Chapman & Hill; 1993.
[16] Biondi-Zoccai GG, Agostoni P, Sangiorgi GM, Airoldi F, Cograve J, Chieffo A, et al. Real-world eluting-stent comparative Italian retrospective evaluation study investigators. Incidence, predictors, and outcomes of coronary dissections left untreated after drug-eluting stent implantation. Eur Heart J 2006;27:540–6.
[17] Mooney CZ. Monte Carlo simulation. Sage University Paper series on Quantitative Applications in the Social Sciences, 07–116. Thousand Oaks: Sage; 1997.
[18] Mehta CR, Patel NR. Exact logistic regression: theory and examples. Stat Med 1995;14:2143–60.
[19] Biondi-Zoccai GG, Sangiorgi GM, Chieffo A, et al. RECIPE (Real-world Eluting-stent Comparative Italian retrosPective Evaluation) Study Investigators. Validation of predictors of intraprocedural stent thrombosis in the drug-eluting stent era. Am J Cardiol Jun. 15 2005;95(12):1466–8.
[20] Rassen JA, Schneeweiss S, Glynn RJ, Mittleman MA, Brookhart MA. Instrumental variable analysis for estimation of treatment effects with dichotomous outcomes. Am J Epidemiol 2009;169:273–84.
[21] Spiegelhalter DJ, Abrams KR, Myles JP. Bayesian approaches to clinical trials and health-care evaluation. New York: John Wiley and Sons; 2004.
[22] Katz MH. Multivariable analysis: a practical guide for clinicians. New York: Cambridge University Press; 2006. **Concise manual on the practical aspects of multivariable analysis.
[23] Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev 1991;59:227–40.
[24] Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
[25] Cepeda MS, Boston R, Farrar JT, Strom BL. Comparison of logistic regression versus propensity score when the number of events is low and there are multiple confounders. Am J Epidemiol 2003;158:280–7 **Seminal paper comparing precision and accuracy of propensity scores versus logistic regression.
[26] Steyerberg EW, Eijkemans MJ, Habbema JD. Stepwise selection in small data sets: a simulation study of bias in logistic regression analysis. J Clin Epidemiol 1999;52:935–42.
[27] Agostoni P, Valgimigli M, Van Mieghem CA, Rodriguez-Granillo GA, Aoki J, Ong AT, et al. Comparison of early outcome of percutaneous coronary intervention for unprotected left main coronary artery disease in the drug-eluting stent era with versus without intravascular ultrasonic guidance. Am J Cardiol 2005;95:644–7.
[28] Concato J, Peduzzi P, Holford TR, Feinstein AR. Importance of events per independent variable in proportional hazards analysis. I. Background, goals, and general strategy. J Clin Epidemiol 1995;48:1495–501.
[29] Peduzzi P, Concato J, Feinstein AR, Holford TR. Importance of events per independent variable in proportional hazards regression analysis. II. Accuracy and precision of regression estimates. J Clin Epidemiol 1995;48:1503–10.
[30] Romagnoli E, De Servi S, Tamburino C, Colombo A, Burzotta F, Presbitero P, et al. I-BIGIS Study Group Milan, Italy. Real-world outcome of coronary bifurcation lesions in the drug-eluting stent era: Results from the 4,314-patient Italian Society of Invasive Cardiology (SICI-GISE) Italian Multicenter Registry on Bifurcations (I-BIGIS). Am Heart J 2010;160:535–42 e1.
[31] Austin PC, Grootendorst P, Anderson GM. A comparison of the ability of different propensity score models to balance measured variables between treated and untreated subjects: a Monte Carlo study. Stat Med 2007;26:734–53.

[32] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Principles for modeling propensity scores in medical research: a systematic literature review. Pharmacoepidemiol Drug Saf 2004;13:841–53 **Careful systematic appraisal of propensity scores, providing shortlist of items for quality check.
[33] Weitzen S, Lapane KL, Toledano AY, Hume AL, Mor V. Weaknesses of goodness-of-fit tests for evaluating propensity score models: the case of the omitted confounder. Pharmacoepidemiol Drug Saf 2005;14:227–38.
[34] Lee BK, Lessler J, Stuart EA. Improving propensity score weighting using machine learning. Stat Med 2010;29:337–46.
[35] Austin PC. Some methods of propensity-score matching had superior performance to others: results of an empirical investigation and Monte Carlo simulations. Biom J 2009;51:171–84.
[36] Tamburino C, Capodanno D, Di Salvo ME, et al. Routine versus selective coronary artery bypass for left main coronary artery revascularization: The appraise a customized strategy for left main revascularization (CUSTOMIZE) study. Int J Cardiol 2011;150(3):307–14.
[37] Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc 2000;95:573–85.
[38] Austin PC. The performance of different propensity-score methods for estimating relative risks. J Clin Epidemiol 2008;61:537–45.
[39] Rubin DB. Estimating causal effects from large data-sets using propensity scores. Ann Intern Med 1997;127:757–63.
[40] Austin PC. Primer on statistical interpretation or methods report card on propensity-score matching in the cardiology literature from 2004 to 2006: a systematic review. Circ Cardiovasc Qual Outcomes 2008;1:62–7.
[41] Austin PC. A critical appraisal of propensity-score matching in the medical literature between 1996 and 2003. Statist Med 2008;27:2037–49.
[42] Morrow DA, Antman EM, Snapinn SM, McCabe CH, Theroux P, Braunwald E. An integrated clinical approach to predicting the benefit of tirofiban in non-ST elevation acute coronary syndromes. Application of the TIMI Risk Score for UA/NSTEMI in PRISM-PLUS. Eur Heart J 2002;23:223–9.
[43] Kurth T, Walker AM, Glynn RJ, Chan KA, Gaziano JM, Berger K, et al. 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.
[44] Stürmer T, Joshi M, Glynn RJ, Avorn J, Rothman KJ, Schneeweiss S. A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. J Clin Epidemiol 2006;59:437–47.
[45] Stukel TA, Fisher ES, Wennberg DE, Alter DA, Gottlieb DJ, Vermeulen MJ. Analysis of observational studies in the presence of treatment selection bias: effects of invasive cardiac management on AMI survival using propensity score and instrumental variable methods. JAMA 2007;297:278–85.
[46] Martens EP, Pestman WR, de Boer A, Belitser SV, Klungel OH. Systematic differences in treatment effect estimates between propensity score methods and logistic regression. Int J Epidemiol 2008;37:1142–7.
[47] Austin PC, Mamdani MM, Stukel TA, Anderson GM, Tu JV. The use of the propensity score for estimating treatment effects: administrative versus clinical data. Stat Med 2005;24:1563–78.
[48] Månsson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case–control and case–cohort studies. Am J Epidemiol 2007;166:332–9.
[49] Winkelmayer WC, Kurth T. Propensity scores: help or hype? Nephrol Dial Transplant 2004;19:1671–3.
[50] 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 **Comprehensive comparison of results of propensity scores and traditional regression methods.
[51] Lunn DJ, Thomas A, Best N, Spiegelhalter D. WinBUGS — a Bayesian modelling framework: concepts, structure, and extensibility. Stat Comput 2000;10:325–37.