PHARMACOEPIDEMIOLOGY AND DRUG SAFETY 2004; **13**: 841–853
Published online 23 April 2004 in Wiley InterScience (www.interscience.wiley.com). **DOI**: 10.1002/pds.969

ORIGINAL REPORT

Principles for modeling propensity scores in medical research: a systematic literature review†

Sherry Weitzen PhD[1]*, Kate L. Lapane PhD[1,2], Alicia Y. Toledano ScD[1,3], Anne L. Hume PharmD[4,5] and Vincent Mor PhD[1,2]

[1]*Department of Community Health, Brown Medical School, Providence, RI, USA*
[2]*Center for Gerontology and Health Care Research, Brown Medical School, Providence, RI, USA*
[3]*Center for Statistical Sciences, Brown Medical School, Providence, RI, USA*
[4]*Department of Pharmacy Practice, University of Rhode Island, Kingston, RI, USA*
[5]*Department of Family Medicine, Brown Medical School, Providence, RI, USA*

SUMMARY

Purpose To document which established criteria for logistic regression modeling researchers consider when using propensity scores in observational studies.
Methods We performed a systematic review searching Medline and Science Citation to identify observational studies published in 2001 that addressed clinical questions using propensity score methods to adjust for treatment assignment. We abstracted aspects of propensity score model development (e.g. variable selection criteria, continuous variables included in correct functional form, interaction inclusion criteria), model discrimination and goodness of fit for 47 studies meeting inclusion criteria.
Results We found few studies reporting on the propensity score model development or evaluation of model fit.
Conclusions Reporting of aspects related to propensity score model development is limited and raises questions about the value of these principles in developing propensity scores from which unbiased treatment effects are estimated.

KEY WORDS — propensity score; logistic regression; observational studies; confounding; bias; methods

BACKGROUND

Propensity score methods

Structural marginal models are a group of statistical methods used in observational studies to reduce bias due to confounding caused by non-random treatment assignment.[1] One technique within this broad class of methods is the propensity score. Although Rosenbaum and Rubin originally proposed the propensity score method in 1983,[2] it has only recently gained popularity in epidemiologic research.[3] The propensity score is the conditional probability of being assigned to a treatment group, given a set of pretreatment characteristics. The propensity score is a balancing score, such that the conditional distribution of the pretreatment characteristics given the propensity score is the same for the treated and untreated groups.[2] The propensity score is most commonly estimated in an observational study from patient and other background characteristics using a multivariable logistic regression model.[4]

Propensity model and logistic regression in epidemiologic research

Logistic regression modeling has become one of the most popular methods used to describe data, in

*Correspondence to: Dr S. Weitzen, Department of Community Health, Box G, Brown University, Providence, RI 02912, USA.
E-mail: Sherry_Weitzen@Brown.edu
†No conflict of interest was declared.

Received 12 February 2003
Revised 4 August 2003
Accepted 23 September 2003

general as well as for estimating propensity scores. However, even though these procedures are easy to use and interpret, scientists should not ignore important statistical properties and limitations of multivariable logistic regression models.[5] Indeed, in recent literature, established criteria for logistic model development and assessment have been recommended.[5–7] The propensity score is estimated from a multivariable logistic regression model. It is inherently a prediction model, used to estimate the probability of treatment selection given a set of observed confounders. The ultimate goal of the propensity score model is to ensure balance between the treatment groups on these confounders.[8] It is unclear which steps in developing a logistic model are necessary to ensure that balance between treatment groups is achieved.[5,7]

Study purpose

The purpose of this study is to systematically review the current literature in which the propensity score is used and to document the extent to which standards for developing logistic regression models are considered and evaluated in the propensity score's development. Although the role of the propensity score differs compared to the role of prediction models or models used to describe the relationship between exposure and outcome, it is unclear which, if any of the principles for logistic model development and assessment are necessary when using the propensity score method.

METHODS

Selection of articles for review

Studies in which the propensity score was used were identified through Medline and Science Citation. Initially, a keyword search was performed, identifying studies which included 'propensity score(s)', 'propensity analysis' or 'propensity matching' in the abstract or title, identifying 55 unduplicated references. In addition, we searched for articles that cited one of the critical propensity score methods articles,[2–4,9–11] identifying an additional 68 unduplicated references. The search was conducted on 23 August 2002 and limited to include studies published in 2001. There were no articles published in languages other than English. We excluded articles unrelated to medical research (20), methodological and statistical articles (30), studies that did not include analysis of data (11), meeting abstracts (1), review articles (1), editorials (7) and letters (2).

Information abstraction

To our knowledge, there are no established or recommended criteria for propensity score model development. In its absence, we used the steps for logistic model development and assessment recommended in the literature.[5–7] From each study, we abstracted information regarding eight aspects of the propensity score model development and assessment, as well as information on how the propensity score was used. This method of data collection may be imperfect since the information published is often limited by the editorial policies of journals. Therefore, each element of propensity score model development and assessment may have been by the authors but not reported in the published study.

The following sections provide a brief description and rationale of each element of model development and assessment we chose to include in our data collection.

Methods for variable selection

When building statistical models, the clinically relevant variables need to be considered. Once data are collected on the variables that are measured and considered important to the study, there are several methods used to decide which variables to include or exclude from the model. These methods include evaluating the univariate relationship between the variable and outcome to be modeled and making decisions based on statistical tests or using an algorithmic method, such as backward elimination, forward selection or stepwise selection, each of which relies on statistical tests for inclusion criteria. Some researchers choose to include all measured variables in a non-parsimonious model, which might affect the precision of the predicted probability. If an important variable is either unmeasured or not considered in the study hypothesis and therefore, not included in the model, then the predicted probability derived from the model could be inaccurate.[4,5,7,12] Therefore, it is important for readers to understand how variables were selected into the propensity score model.

Sufficient events per variable (EPV). We were interested in the number of EPV in the propensity score model. An event is defined as an observation in the less frequent of the two exposure groups, since the 'outcome' in this instance is being exposed or unexposed. The rule of thumb is that a logistic model can support 10 EPV before the precision and accuracy of the predicted probabilities may be compromised.[5,13,14]

Continuous variable conformity with linear gradient. If continuous variables were included in the propensity score model, then we sought information on whether the functional relationship of this variable and the exposure being modeled was considered. Often the functional form of a continuous variable is assumed to be linear when in fact a different functional form exists. The accuracy of the estimated propensity score could be compromised if estimated from a model with a misspecified variable or variables.[7,12,15]

Interactions. When the propensity score model included interactions between model variables, we looked for criteria for their inclusion. Adding an inappropriate interaction term could alter the estimated propensity score, possibly introducing bias to the estimate.[7] In theory, it might be important to consider the exclusion of an important interaction in the propensity score model, but this would be difficult to evaluate from published studies.

Collinearity. We looked for evidence of the assessment and correction for collinearity between variables in the propensity score. Collinearity occurs when two or more confounders and/or the exposure variable in the model are highly correlated with each other.[5] This indicates that they are likely to be measuring the same phenomenon. Some statisticians assert that in the presence of collinearity, the coefficients and standard errors for the correlated variables will be unrealistically large.[7] If this is true, inclusion of collinear variables in a propensity score model could result in both bias and imprecision of the estimated probabilities. However, others indicate that including collinear variables in the same model only affects the precision of the estimated coefficients.[8] If this is the case, then collinearity would not be an issue in the development of the propensity score model.

Assessment of model fit. After selecting a final logistic regression model, it is important to evaluate how well the model describes the data. Goodness of fit tests assess whether the distances between the observed (treatment—yes or no) and the predicted outcome from the model (propensity score) are small and unsystematic (i.e. among patients with propensity scores between 0.30 and 0.40, one would expect to see approximately 35% of those patients actually in the treatment group when the model's goodness of fit was adequate).[7,8,16] If the model does not accurately describe the data then estimated probabilities or propensity scores, generated from the model may be inaccurate. If using propensity scores estimated from a poorly fit model do not create balance between the treatment groups, this could lead to biased estimates of treatment effect. Therefore, some indication that model fit was assessed might be important to the interpretation of the results.[5,7,17] However, this has yet to be demonstrated empirically.

Discrimination of model. Model discrimination is defined as how well the predicted probabilities derived from the model classify patients into their actual treatment group. It is a measure of predictive accuracy. This is often measured by evaluating the area under the Receiver Operator Curve (ROC) or c-statistic. This measures a different concept than the goodness of fit of the model.[7]

Balance achieved. The goal of the propensity score is to create balance on the potential confounders included in the propensity score model between treatment groups.[2–4,11] Balance is often assessed by examining the differences in distributions of confounders between treatment groups, either after creating a matched sample or when stratifying the full sample based on the propensity score. We looked for evidence of this assessment in the articles, such as graphs or tables displaying the distribution of confounders for the treatment groups, either after matching or among propensity score strata. We also considered any statements indicating that balance between the groups was achieved before applying the propensity score to adjust treatment effect estimates.

Adjustment method. Finally, we collected information on how the propensity score was used to adjust treatment effect estimates. The potential methods included: (1) creating a matched sample of exposed and unexposed patients who had equal or similar propensity scores; (2) stratifying patients on their propensity score and estimating within strata treatment effects or (3) including the propensity score (as either a predicted probability, linear transformation or as a four design variables based on quintiles of the propensity score) with the exposure variable as covariates in a multivariable model of the outcome.[2,4]

RESULTS

Our key word search yielded 30 articles that employed the propensity score in the study of a health or medical related question. Twenty-one additional articles were found through a citation search of the significant methods articles written about the propensity score.[2–4,9–11] After further review, four studies were

excluded: one used the same data set and identical propensity score as a study by the same authors[18] which was included; two studies did not include any description of the propensity score used, each stating that details were described elsewhere (in articles published prior to 2001);[19,20] and one study estimated the propensity score with a polytomous logistic regression model.[21] The final study sample consisted of 47 articles.

Description of studies using the propensity score

Of the 47 articles reviewed, 30 (64%) were related to cardiovascular disease. Among the other 17 studies, 6 were relevant to cancer research;[22–27] 3 were in the area of mental illness;[28–30] 2 were related to postnatal outcomes in infants;[31,32] 2 in diabetes research;[33,34] and the other 4 included studies examining arthritis,[35] survival of transplant recipients,[36] antibiotic use,[37] and sinusitis.[38]

In Table 1, we list references, brief descriptions of exposures (or treatments) and outcomes under study, and the numbers of patients in each exposure group. The majority of studies were focused on quantifying the effect of a particular treatment on reducing the rate of mortality and morbidity (e.g. the use of hormone replacement therapy (HRT) among women who had a recent myocardial infarction (MI) on reducing the risk of stroke[39]). In several studies, treatments were examined for potential adverse effects (e.g. the use of selective serotonin reuptake inhibitors (SSRIs) on increasing the risk of developing heart valve regurgitation[40]). In a few studies, the authors used the propensity score to control for differences in exposure to a harmful substance (e.g. cigarette smoking[41]) or prognostic factor (e.g. bundle branch block[42]). Exposures listed in Table 1 were included as dependent variables in the propensity score models described by these studies.

Propensity score model development

In Table 2, the elements of the propensity score model development for each manuscript are listed including the variable selection method; the number of variables included in the model; the EPV (computed by dividing the number of patients in the smallest exposure group by the number of variables in the model); whether the functional form of any continuous variables was evaluated; the inclusion criteria for any interactions included in the model and whether authors mentioned assessment and correction of collinearity between model variables.

Variable selection method. In 24 of the 47 articles reviewed, information was not provided on what method was used to select variables to include in the propensity score. Among the remaining 23 articles, non-parsimonious models were used in six studies. Seven articles included information on significance testing at the univariate level, as the method by which they identified variables for the propensity score model. Algorithmic methods such as backward, forward or stepwise selection to select variables into the model were used in four studies. Four articles indicated *a priori* selection of variables as criteria for inclusion. In one article, the variables were selected for the propensity score based on goodness of fit tests of the model.[35]

EPV. Out of the 47 studies, 13 did not include information on the number of variables in the propensity score model. While eight of these studies had large sample sizes in both exposure groups, five had small numbers of patients in one or both of the treatment groups, therefore probably having fewer than 10 EPV in the propensity score. Among the 34 studies from which we could determine the number of variables included in the model, there were five in which the EPV was less than 10.

Functional form of continuous variables. In nearly all the studies reviewed, the scales of variables included in the propensity score model were not explicit. Of the three in which it was clear that continuous variables were used, only one mentioned that these variables' functional form was evaluated with respect to the exposure being modeled.

Interaction inclusion criteria. In most of the studies reviewed (30 out of 47), it was unclear whether interactions between variables were incorporated into the propensity score model. In 12 articles, the propensity score description clearly indicated that no interaction terms were included. Among the five articles in which it was clearly stated that interactions were in the propensity model, p values of the coefficients were used as criteria in three articles. Aronow *et al.*[43] used improvement in the discrimination of the propensity score model as criteria for interaction inclusion, while Shlipak *et al.*[44] considered improvement in the resulting balance between treatment groups for whether or not to include an interaction term.

Collinearity. None of the studies reported that collinearity between variables in the propensity score

Table 1. Study description

Reference	Exposure*	Outcome**	Total exposed[†]	Total unexposed**
Angeja[39]	HRT use	Stroke	7353	107 371
Aronow[43]	Lipid lowering therapy	180 day mortality	3653	17 156
Ascione[61]	On pump during CABG	Acute renal failure	51	202
Beuth[23]	Tx w/Oral enzyme	Symptoms of breast CA	239	410
Carmeli[37]	Ceftriaxone vs. ampicillin.	Isolation of pathogen	1308	2445
Cen[62]	Biologic vs. mechanical valve	10 year survival	495	644
Dammann[31]	Hypocarbia in 24 hours post birth	Echolucency	182	617
Earle[22]	Chemo for advanced lung CA	1 year survival	2012	4220
Ellis[63]	Beta-blockers	Creatine kinase levels	2926	3274
Feng[64]	Atrial fibrillation	Prothrombic state	47	3515
Ferrara[34]	HRT use	HBA1 levels	3852	11 583
Foody[41]	Smoking status	All cause mortality	2187	2166[‡]
Gillinov[49]	Mitral valve replace vs. repair	Post surgery survival	85	397
Grossi[46]	Mitral valve reconstruct vs. replace	30 day mortality	152	71
Gum[59]	Daily aspirin use	All cause mortality	2310	3864
Hayashi[65]	ACE inhibitor	Change in hematocrit	329	1884
Hesse[42]	Bundle branch block	All cause mortality	340	6733
Ioannidis[66]	BITA vs. SITA	In hospital mortality	867	830
Kachele[28]	Short vs. long term tx-eating disorders	Cure	353	324
Keating[24]	Physician discussion of options	Breast conserving therapy	412	380[§]
Kimmel[67]	Coronary stent	In hospital mortality	10 690	6121
Margolis[33]	Platelet relasate treatment	Healed wound	6252	20 347
Mast[40]	SSRI use	Valvular heart disease	292	5145
Mehta S[47]	PCI	Adverse events	2658	9904
Mehta R[68]	Increased left ventricular mass index	In hospital mortality	115	358
Mitra[25]	Intensive breast CA screening program	Stage of disease at diagnosis	58	3022
Mukamal[69]	Alcohol consumption prior to MI	All cause mortality	321	896
Normand[70]	Angiography	3 year survival	17 304	20 484
Novaro[71]	Statins	Change in aortic valve area	54	120
O'Day[27]	High dose tamoxifen for melanoma	Response to treatment	35	45
Osswald[72]	Complete vs. incomplete revascularization	180 day mortality	133	726
Peterson[73]	Heparin	30 day mortality	5576	1441
Piccirillo[38]	1st vs. 2nd line antibiotics for sinusitis	Response to treatment	17 329	11 773
Popiela[26]	Oral enzyme treatment for colon CA	Symptoms of colon CA	587	597
Rathore[48]	Reperfusion therapy	30 day mortality	171	1783
Rumsfield[74]	Multiple exposures	7 months status	NR	NR
Schnuelle[36]	Catcholamine in organ donors	Survival of recipient	1562[¶]	127
Sernyak[29]	Neuroleptic use for PTSD	Improvement in symptoms	67	715
Serynak[30]	Clozapine tmt for schizophrenics	Completed suicides	1415	44 502
Shlipak[44]	ACE inhibitors and beta-blockers	1 year survival	9108	5872
Shlipak[75]	HRT use	In hospital mortality	7353	107 371
Stenestrand[45]	Statins	1 year survival	5528	14 071
Suero[76]	Chronic total occlusion	In hospital mortality	2007	25 620
Tu[77]	MD pt. Volume	30 day mortality	27 494	70 700
Van Marter[32]	Antenatal glucocortoid tmt	Lung disease in babies	743	711
Welch[78]	Normal ECG (in MI pts)	In hospital mortality	30 759	222 875
Wiles[35]	DMARDs and/or steroids	Disability from arthritis	183	201

ACE, angiotensin converting enzyme; BITA, bilateral internal thoracic artery revascularization; CA, cancer; CABG, coronary artery bypass graft; DMARDs, disease-modifying antirheumatic drugs; ECG, electrocardiograph; HRT, hormone replacement therapy; MD, Physician; MI, myocardial infarction; NR, not reported; PCI, percutaneous coronary intervention; PTSD, post traumatic stress disorder; SITA, single internal thoracic artery revascularization; SSRI, selective serotonin reuptake inhibitor.

*Used as dependent variable in propensity score.

**Study outcome (not included in propensity score).

[†]Sample used in development of propensity score.

[‡]Current versus never smokers.

[§]Data for MA only (smaller of two separate samples).

[¶]Total of single and combined use groups.

Table 2. Propensity Score Development

Reference	Selection method*	Number of variables	EPV**	Functional form†	Interactions‡	Collinearity
Angeja	NR	NR	NR	NR	NR	NR
Aronow	*p* values	19	3653/19 = 192	Yes	Discrimination	NR
Ascione	NR	NR	NR	NR	NR	NR
Beuth	NR	16	239/16 = 15	NR	NR	NR
Carmeli	Algorithm	12	1308/12 = 109	NR	None	NR
Cen	*p* values	7	495/7=70	NR	None	NR
Dammann	*p* values	14	182/14 = 13	None	None	NR
Earle	NR	NR	NR	NR	NR	NR
Ellis	*p* values	10	2926/10 = 29	None	NR	NR
Feng	NR	6	47/6 = 8	NR	NR	NR
Ferrara	NR	7	3852/7 = 550	NR	NR	NR
Foody	Non-parsimonious	28	2166/28 = 77	NR	*p* values	NR
Gillinov	NR	NR	NR	NR	*p* values	NR
Grossi	*p* values	30	71/30 = 2	NR	NR	NR
Gum	Non-parsimonious	34	2310/34 = 68	NR	NR	NR
Hayashi	NR	10	329/10 = 33	NR	None	NR
Hesse	NR	30	340/30 = 11	NR	NR	NR
Ioannidis	Algorithm	17	830/17 = 49	NR	None	NR
Kachele	NR	32	324/32 = 10	NR	None	NR
Keating	NR	23	380/23 = 17	NR	NR	NR
Kimmel	NR	18	6121/18 = 340	NR	NR	NR
Margolis	*A priori*	14	6252/14 = 447	NR	NR	NR
Mast	Algorithmic	7	292/7 = 42	NR	None	NR
Mehta S	NR	NR	NR	NR	NR	NR
Mehta R	Algorithmic	21	115/21 = 5	NR	*p* values	NR
Mitra	*A priori*	29	58/29 = 2	NR	NR	NR
Mukamal	NR	18	321/18 = 18	NR	NR	NR
Normand	Non-parsimonious	102	17 304/102 = 170	NR	NR	NR
Novaro	NR	3	54/3 = 18	NR	None	NR
O'Day	*A priori*	8	35/8 = 4	NR	None	NR
Osswald	Non-parsimonious	NR	NR	NR	NR	NR
Peterson	Non-parsimonious	33	1441/33 = 44	NR	NR	NR
Piccirillo	*p* values	4	11 773/4 = 2943	NR	None	NR
Popeila	NR	8	587/8 = 73	NR	NR	NR
Rathore	NR	NR	NR	NR	NR	NR
Rumsfield	NR	NR	NR	NR	NR	NR
Schnuelle	NR	5	127/5 = 25	NR	None	NR
Sernyak	NR	17	1415/17 = 83	NR	NR	NR
Sernyak	NR	NR	NR	NR	NR	NR
Shlipak	Non-parsimonious	NR	NR	NR	Balance	NR
Shlipak	NR	NR	NR	NR	NR	NR
Stenestrand	Algorithmic	42	5528/42 = 132	NR	NR	NR
Suero	NR	7	2007/7 = 287	NR	NR	NR
Tu	*A priori*	11	27 494/11 = 2500	NR	NR	NR
Van Marter	*A priori*/*p* values	NR	NR	NR	NR	NR
Welch	NR	NR	NR	NR	NR	NR
Wiles	Goodness of fit	17	183/17 = 11	NR	None	NR

NR, not reported.

**A priori*, inclusion determined from previous research; algorithmic, forward, backward or stepwise selection method; goodness of fit, included variables that improved model goodness of fit; non-parsimonious, included all available variables; NR, not reported; *p* values, univariate statistical testing.

**EPV, events per variable, total in smallest exposure group/number of variables (rounded to nearest integer); NR, unable to compute EPV.

†Functional form, none; no continuous variables included in the model; NR, not reported; Yes, evidence that continuous variables evaluated for functional relationship with exposure.

‡Interactions, inclusion criteria for interactions; balance, included interactions when improved balance between exposure groups on confounders; discrimination, included interactions when improved model discrimination, None, no interactions included; NR, not reported; *p* values, statistical significance.

model was checked before estimating the probability of assignment to an exposure group.

Adequacy of propensity score model

Table 3 indicates which articles included evaluation of the adequacy of the propensity score model to describe the data.

Goodness of fit. Only 6 of the 47 studies considered the goodness of fit of the propensity model. In four of the six studies, goodness of fit was evaluated with the Hosmer–Lemeshow goodness of fit statistic and the authors provided the statistic and/or *p* value. Poor fit of the propensity score model was reported in two studies. Stenestrand and Wallentin[45] still used the estimates from this model to adjust treatment effects. However, Grossi *et al*., after finding their propensity score model poorly fit the data, decided not to use the propensity score to adjust their estimates.[46] Mehta *et al*.[47] used split sample validation to evaluate the fit of the propensity score model. Rathore *et al*. reported that 'the propensity model demonstrated appropriate. . .calibration',[48] p.517 but did not specify what methods they used.

Discrimination. Discriminatory ability of the propensity score model was reported in 18 of 47 studies. C-statistics or area under the ROC curve ranged from 0.52,[38] suggesting the model had no discriminatory ability, to 0.92[49] indicating almost perfect discrimination.[7]

Balance on variables. Nearly half (22 out of 47) of the studies included no information regarding whether the propensity score created balance between exposure groups on the characteristics considered in the propensity model. In one article, when balance was not achieved in three of five quintiles based on the propensity scores, patients in those quintiles were excluded from further analyses.[43]

DISCUSSION

In the past few years, the propensity score method has become a more common method used for confounder adjustment in observational studies.[3] Our intention was to document how propensity score models were developed and assessed from the available information in published studies. The purpose of this study was to illustrate the relative uncertainty of researchers as to what steps in the model development process are important to the propensity score, and not to judge the quality of the studies based on the standard logistic regression model criteria.

There are at least two purposes of logistic regression modeling in medical research in addition to developing propensity scores: predicting the outcome of an individual or group of individuals using their own values of the predictor variables, and quantifying the effect of an exposure on an outcome while simultaneously controlling for the influences of confounding variables.[5] Extensive research has been done on the effects of poorly developed and poorly fit logistic models when used for these purposes.

The accuracy and precision of estimates from logistic models used for either adjusting estimates of effect or predicting outcomes could be seriously affected by missing predictors or confounders, variable selection that relies solely on statistical significance,[7,13,50] misspecified continuous variables,[7,15] and inappropriate inclusion or exclusion of important interaction terms.[7,13,50] Models used to predict outcomes may be additionally influenced by inclusion of unimportant predictors, ignoring the inclusion of highly collinear variables and poor model fit.[5,7,8,51]

The propensity score model serves a different purpose than predictive and adjusted effect estimate models. As such, the extent to which some or all of the criteria are important to the purpose of the propensity score, that is balance between the treatment groups on confounders, is unknown. While the propensity score is a predicted probability resulting from the developed model, its main purpose is to control for multiple confounders simultaneously. To our knowledge, there is only one study, by Drake,[12] which examines the sensitivity of treatment effect estimates derived from propensity score models. Drake's analyses are limited to potential biases due to omitted confounders and misspecified continuous variables.

For the most part, we found that there is very little information regarding the development and validation of the propensity score models provided in published studies that employ propensity score methods to adjust treatment effect estimates. Over half the studies did not include information on variable selection, therefore making it difficult to assess whether all potential and available confounders were adjusted for using this tool. Several studies employed statistical significance testing or algorithmic methods to select variables into the propensity score model, even though these methods are discouraged for the evaluation of confounding.[7,13,50] When using the propensity score to adjust for confounders, assessment of balance between the treatment groups should be carefully considered when selecting variables into the propensity score model.[2,11,52]

Table 3. Adequacy of propensity score model

Reference	Goodness of fit	Discrimination	Balance check	Method for estimate adjustment*
Angeja	NR	NR	NR	Model covariate
Aronow	NR	c-stat = 0.87	Yes	Model covariate
Ascione	NR	NR	NR	Stratified analysis
Beuth	NR	NR	NR	Model covariate
Carmeli	NR	c-stat = 0.80	Yes	Model covariate
Cen	NR	NR	NR	Model covariate
Dammann	NR	NR	NR	Model covariate
Earle	NR	NR	Yes	Stratified analysis
Ellis	NR	c-stat = 0.62	NR	Model covariate
Feng	NR	NR	Yes	Matching
Ferrara	NR	NR	NR	Model covariate
Foody	NR	c-stat = 0.76	Yes	Model covariate
Gillinov	NR	c-stat = 0.92	NR	Model covariate
Grossi	Yes	NR	NR	NR
Gum	NR	c-stat = 0.83	Yes	Matching
Hayashi	NR	NR	Yes	Matching
Hesse	NR	c-stat = 0.76	NR	Model covariate
Ioannidis	NR	NR	NR	Model covariate
Kachele	NR	NR	Yes	Model covariate
Keating	NR	NR	NR	Stratified analysis
Kimmel	Yes	NR	NR	Stratified/model covariate
Margolis	NR	c-stat = 0.90	Yes	Stratified analysis
Mast	NR	c-stat = 0.82	NR	Stratified analysis
Mehta S	Yes	NR	NR	Model covariate
Mehta R	NR	NR	Yes	Stratified/model covariate
Mitra	NR	NR	Yes	Stratified/model covariate
Mukamal	NR	c-stat = 0.82	NR	Model covariate
Normand	NR	c-stat = 0.84	Yes	Matching
Novaro	NR	NR	NR	Model covariate
O'Day	NR	NR	Yes	Stratified analysis
Osswald	NR	c-stat = 0.81	Yes	Model covariate
Peterson	NR	c-stat = 0.69	Yes	Model covariate
Piccirillo	NR	c-stat = 0.52	Yes	Stratified analysis
Popiela	NR	NR	Yes	Matching
Rathore	Yes	NR	Yes	Model covariate
Rumsfield	NR	NR	NR	NR
Schnuelle	NR	NR	NR	Model covariate
Sernyak	NR	NR	Yes	Matching
Sernyak	NR	NR	NR	Model covariate
Shlipak	NR	NR	Yes	Stratified analysis
Shlipak	NR	NR	NR	Model covariate
Stenestrand	Yes	c-stat = 0.81	NR	Model covariate
Suero	NR	NR	Yes	Matching
Tu	NR	c-stat = 0.56	NR	Stratified analysis
Van Marter	NR	NR	NR	Model covariate
Welch	NR	c-stat = 0.82	Yes	Model covariate
Wiles	Yes	c-stat = 0.86	Yes	Stratified/model covariate

*How propensity score was subsequently used to adjust effect estimates; matching, created matched sample; model covariate, propensity score used as variable in multivariable model; stratified analysis, sample divided into quintiles based on propensity score and stratum specific estimates computed; stratified/model covariate, sample divided into quintiles and design variables for quintile membership included in multivariable model as covariates.

Other elements of model development, such as insuring that continuous variables were entered in the appropriate functional form, providing clear criteria for inclusion of interactions, and checking variables for potential collinearity, were generally not mentioned in these studies. If it is reported that balance between treatment groups was achieved by using the propensity score, then this information may not be important to the interpretation of the results. Collinearity may not be relevant at all, as long as the inclusion of

collinear variables does not inflate the estimated regression coefficients in the propensity score.[7,8]

In the following paragraphs, we discuss the potential impact of propensity scores measured with error, either by having too few EPV or poor model fit. Among the 34 studies that provided the number of variables included in the propensity score model, 5 had too few EPV to support the estimation of the propensity score. In a recent study, Cepeda *et al*. when the outcome is rare, and there are many confounders, fitting a logistic model with few EPV can lead to biased estimates of the treatment effect.[53] The authors concluded that the propensity score could be useful in these situations since it is a composite of all confounders, thereby reducing the number of covariates that need to be adjusted for in the second stage model. However, they did not examine the effect of modeling a rare exposure as a function of many confounders in the propensity score model.

Simulation studies imply that too few EPV in a logistic model can inflate the estimated regression coefficients, and potentially result in biased propensity scores.[14] If the regression coefficients are inflated in the propensity score model, then the propensity score may be estimated with error. Since the coefficients used to estimate the propensity score are the same regardless of individuals' treatment assignment, any measurement error in the propensity score estimate is non-differential with respect to treatment assignment. Likewise, since the study outcome is not considered in the estimation of the propensity score, errors in the propensity score estimates are unrelated to individuals' observed outcomes as well.

Only 6 of the 47 studies included any indication that the model fit was evaluated prior to using the estimated score to adjust the treatment effect. In general, poor model fit is a function of problems related to model development, such as continuous variable misspecification, inappropriate interaction terms or omitted confounders or interactions.[7,54] Each of these flaws in the model could lead to errors in propensity score estimates.[12] However, available goodness of fit tests have relatively low power to detect such problems in logistic regressions models, and may not be useful for assessing the propensity score model.[7,8,54,55]

Again, both few EPV and lack of model fit could lead to propensity scores estimated with error. We speculate the measurement error in the propensity score is non-differential with respect to treatment assignment. However, if balance between treatment groups is achieved within propensity score strata, then biased estimates of the propensity score may be irrelevant.[52] Additionally, the extent to which error biases estimates of treatment effect may be a function of how the propensity score is used to control for confounding (e.g. matching, stratification or model covariate). If the propensity score is used as a matching variable, error in its measurement has no effect on the estimates of treatment effect derived from the study. This is true as long as the individuals retain their appropriate rank by propensity score, because treated patients should be matched with the same untreated patients as they would when using a perfectly estimated propensity score. Therefore, treatment effect estimates adjusted by this method are robust with respect to propensity scores measured with error. Similarly, adjustment for confounding using a stratified analysis will also be robust with respect to propensity scores measured with error as long as the strata are defined using percentile cut-points.

The same may not be true for treatment effects estimated when the propensity score is used as a continuous covariate in a regression model. Over half the studies that we reviewed used this approach. D'Agostino states that overfitting the model from which the propensity score is derived is not a concern.[4] This is in conflict with epidemiologic studies demonstrating that adjustment for confounders measured with error as continuous variables, in this case, the propensity score, lead to biased estimates of effect due to residual confounding.[56–58]

Lastly, the c-statistic is an indication of the discriminatory power of the model, and how well the propensity score classified the patient into the observed treatment groups. In theory, a c-statistic of 0.50 indicates poor model discrimination, and therefore using the model to classify patients into groups is as good as flipping a coin. On the other hand, a c-statistic of 0.90 indicates excellent discrimination, meaning that 90% of the time the propensity score of a treated individual was higher when compared to the propensity score of an untreated individual.[7] In other words, there could be very little overlap between the treatment groups with respect to the distribution of the propensity score.* When this occurs, if the propensity score is used to match individuals who are receiving treatment to untreated individuals, there may be a large proportion of patients that would be lost from the sample due to the lack of an appropriate match. In fact, this was the case in a study by Gum *et al*.[59] The c-statistic for the propensity score was 0.83, indicating good

*Unpublished research by Robert Obenchain, presented at Fourth Annual Workshop on Pharmaceutical Outcomes Research, Newport, RI, 18 October 2002.

discrimination. The original study cohort included 2310 'treated individuals' and 3846 'untreated individuals'. The results from the propensity-matched analysis were based on 1351 pairs of treated and untreated patients, losing more than half the study population. As is often the case with clinical trials, when we eliminate a large portion of our sample because they do not meet inclusion criteria (i.e. heterogeneous with respect to patient characteristics), we potentially sacrifice the generalizability of the results for better internal validity of our study.

A propensity score with a high c-statistic could also be detrimental to a stratified analysis based on the score. In this case, there could potentially be none or very few untreated individuals in the highest strata, as well as none or very few of the treated individuals in the lowest strata, and therefore treatment effects cannot be estimated in these strata. In fact, that occurred in at least one study included in this report. Gillinov *et al*. comparing mitral valve repair (treatment) versus replacement (control), developed a propensity score model with a c-statistic of 0.92.[49] When stratifying the sample based on the propensity score, strata 3, 4 and 5 had fewer than five patients (total $n = 97$) who received mitral valve replacement, making it difficult, if not impossible to estimate the effect of repair versus replacement on post-surgical survival using stratified analysis. Excellent discrimination of the propensity score model could lead to little or no overlap of the estimated propensity score between the treatment groups. As a covariate used in the second stage regression model, this problem is difficult to detect before fitting the model, especially if not evaluated by stratifying the sample and examining the number of treated and untreated individuals in each subgroup. Many software packages will allow the model to be fit, but the coefficient estimates and corresponding standard errors will be biased, and it is likely that confounding by the propensity score will not be completely controlled for in the model from which the treatment effect is estimated. Therefore, the estimates of treatment effect could be biased when using a propensity score estimated from a model with very good or excellent discrimination.[7] When using a highly discriminating propensity score to create a matched sample, there would be very few exposed and unexposed individuals that could actually be matched by this method, leaving few individuals in the study sample. Therefore, it appears that a propensity score model that produces estimates that accurately classify individuals into their treatment groups may be undesirable for the purposes of using the propensity score to adjust for factors related to treatment assignment. However, from this study, it appears that propensity scores with large c-statistics are still being used to adjust treatment effect estimates.

Finally, the key purpose of the propensity score is generate probabilities of treatment assignment conditional on a set of variables that are both related to treatment and the outcome. Additionally every individual's probability of receiving treatment based on this model must be greater than zero.[60] The distributions of these confounders should be fairly equal between the treatment groups within strata of the propensity score in order to establish that propensity score will provide adequate control for these confounders when estimating the effect of treatment on the outcome.[2,4,52,60] However, nearly half the studies reviewed did not report that the propensity score balanced the treatment groups. These omissions may be attributable to space limitations for publication. Nevertheless, almost a third of these studies reported a c-statistic which would seem to be of less importance for the interpretation of the results.

There are several potential limitations of this study. As in most systematic reviews of the literature, it is difficult to evaluate what analyses were performed versus what was reported in the published article. Therefore, it is possible that much of the model development and assessment was performed by the researchers but not explicitly described in their articles. Another potential limitation of this study was the possible inability to obtain the entire population of studies that used the propensity score method with our search strategy. However, by expanding our search to articles that cited the key propensity score methods articles, we greatly reduced the likelihood of missing many published studies.

From this review of the literature, it is clear that the reporting of propensity score model development and model fit is variable. This may indicate that researchers are uncertain as to which criteria for logistic regression modeling used to generate predictions or for estimating adjusted treatment effect estimates are important with respect to estimating a propensity score. To our knowledge, there are no published guidelines or recommendations on what considerations are important to estimating unbiased propensity scores and ensuring balance is achieved between the treatment groups when using this method. Empirical studies need to be conducted in order to fully understand the role of each of the previously recommended criteria for logistic regression modeling in the estimation of useful propensity scores.

REFERENCES

1. Robins JM, Mark SD, Newey WK. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* 1992; **48**: 479–495.
2. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. *Biometrika* 1983; **70**(1): 41–55.
3. Joffe MM, Rosenbaum PR. Invited commentary: propensity scores. *Am J Epidemiol* 1999; **150**(4): 327–333.
4. D'Agostino RB, Jr. Tutorial in biostatistics: propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Stat Med* 1998; **17**: 2265–2281.
5. Bagley SC, White H, Golomb BA. Logistic regression in the medical literature: standards for use and reporting, with particular attention to one medical domain. *J Clin Epidemiol* 2001; **54**: 979–985.
6. Feinstein A. *Multivariable Analysis: An Introduction*. Yale University Press: New Haven, 1996.
7. Hosmer DL, Lemeshow S. *Applied Logistic Regression*, 2nd edn. John Wiley and Sons: New York, 2000.
8. Harrell FE, Jr. *Regression Modeling Strategies*. Springer-Verlag: New York, 2001. Springer Series in Statistics.
9. Rosenbaum PR, Rubin DB. Reducing bias in observational studies using subclassification on the propensity score. *J Am Stat Assoc* 1984; **79**(387): 516–524.
10. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *Am Statistician* 1985; **39**(1): 33–38.
11. Rubin DB. Estimating causal effects from large data sets using propensity scores. *Ann Intern Med* 1997; **127**: 757–763.
12. Drake C. Effects of misspecification of the propensity score on estimators of treatment effect. *Biometrics* 1993; **49**: 1231–1236.
13. Greenland S. Model and variable selection in epidemiologic analysis. *Am J Public Health* 1989; **79**(3): 340–349.
14. 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**(12): 1373–1379.
15. Royston P, Ambler G, Sauerbrei W. The use of fractional polynomials to model continuous risk variables in epidemiology. *Int J Epidemiol* 1999; **28**: 964–974.
16. Ruttimann UE. Statistical approaches to development and validation of predictive instruments. *Crit Care Clin* 1994; **10**(1): 19–35.
17. Hosmer DW, Taber S, Lemeshow S. The importance of assessing the fit of logistic regression models: a case study. *Am J Public Health* 1991; **81**(12): 1630–1635.
18. Sernyak MJ, Rosenheck R, Desai R, Stolar M, Ripper G. Impact of clozapine prescription on inpatient resource utilization. *J Nerv Ment Dis* 2001; **189**(11): 766–773.
19. Harrison G, Hopper K, Craig T, *et al*. Recovery from psychotic illness: a 15- and 25-year international follow-up study. *Br J Psychiatry* 2001; **178**: 506–517.
20. Kessler R, Almeida DM, Berglund P, Stang P. Pollen and mold exposure impairs the work performance of employees with allergic rhinitis. *Ann Allergy Asthma Immunol* 2001; **87**: 289–295.
21. Laine C, Hauck WW, Gourevitch MN, Rothman J, Cohen A, Turner BJ. Regular outpatient medical and drug abuse care and subsequent hospitalization of persons who use illicit drugs. *J Am Med Assoc* 2001; **285**: 2355–2362.
22. Earle CC, Tsai JS, Gelber RD, Weinstein MC, Neumann PJ, Weeks JC. Effectiveness of chemotherapy for advanced lung cancer in the elderly: instrumental variable and propensity analysis. *Am J Clin Oncol* 2001; **19**(4): 1064–1070.
23. Beuth J, Ost B, Pakdaman A, Rethfeldt E, Bock PR, Hanisch J, Schneider B. Impact of complementary oral enzyme application on the postoperative treatment results of breast cancer patients—results of an epidemiological multicentre retrolective cohort study. *Cancer Chemother Pharmacol* 2001; **47**(Suppl.): S45–S54.
24. Keating NL, Weeks JC, Landrum MB, Borbas C, Guadagnoli E. Discussion of treatment options for early-stage breast cancer: effect of provider specialty on type of surgery and satisfaction. *Med Care* 2001; **39**(7): 681–691.
25. Mitra N, Schnabel FR, Neugut AI, Heitjan DF. Estimating the effect of an intensive surveillance program on stage of breast carcinoma at diagnosis: a propensity score analysis. *Cancer* 2001; **91**(9): 1709–1715.
26. Popiela T, Kulig J, Hanisch J, Bock PR. Influence of a complementary treatment with oral ezymes on patients with colorectal cancers-an epidemiologic retrolective cohort study. *Cancer Chemother Pharmacol* 2001; **47**(Suppl.): S55–S63.
27. O'Day S, Boasberg PD, Kristedja TS, *et al*. High-dose tamoxifen added to concurrent biochemotherapy with decrescendo interleukin-2 in patients with metastatic melanoma. *Cancer* 2001; **92**(3): 609–619.
28. Kachele H, Kordy H, Richard M. Therapy amount and outcome of inpatient psychodynamic treatment of eating disorders in Germany: data from a multicenter study. *Psychother Res* 2001; **11**(3): 239–257.
29. Sernyak MJ, Kosten TR, Fontana A, Rosenheck R. Neuroleptic use in the treatment of post-traumatic stress disorder. *Psychiatr Q* 2001; **72**(3): 197–213.
30. Sernyak MJ, Desai R, Stolar M, Rosenheck R. Impact of clozapine on completed suicied. *Am J Psychiatry* 2001; **158**(6): 931–937.
31. Dammann O, Allred EN, Kuban KC, *et al*. Hypocarbia during the first 24 postnatal hours and white matter echolucencies in newborns $<$ or $=$ 28 weeks gestation. *Pediatr Res* 2001; **49**(3): 388–393.
32. Van Marter L, Allred EN, Leviton A, Paganon M, Parad R, Moore M, Neonatology Committee for the Development Epidemiology Network. Antenatal glucocorticoid treatment does not reduce chronic lung disease among surviving preterm infants. *J Pediatr* 2001; **138**: 198–204.
33. Margolis DJ, Kantor J, Santanna J, Strom BL, Berlin JA. Effectiveness of platelet releasate for the treatment of diabetic neuropathic foot ulcers. *Diabetes Care* 2001; **24**(3): 483–488.
34. Ferrara A, Karter AJ, Ackerson LM, Liu JY, Selby JV, the Northern California Kaiser Permanente Diabetes Registry. Hormone replacement therapy is associated with better glycemic control in women with type 2 diabetes. *Diabetes Care* 2001; **24**(7): 1144–1150.
35. Wiles NJ, Lunt M, Barrett EM, *et al*. Reduced disability at five years with early treatment of inflammatory polyarthritis: results from a large observational cohort, using propensity models to adjust for disease severity. *Arthritis Rheum* 2001; **44**(5): 1033–1042.
36. Schnuelle P, Berger S, de Boer J, Persijn G, van der Woude FJ. Effects of catecholamine application to brain-dead donors on graft survival in solid organ transplantation. *Transplantation* 2001; **72**(3): 455–463.

37. Carmeli Y, Castro J, Eliopoulos GM, Samore MH. Clinical isolation and resistance patterns of and superinfection with 10 nosocomial pathogens after treatment with cefriaxone versus ampicillin-sulbactam. *Antimicrob Agents Chemother* 2001; **45**(1): 275–279.
38. Piccirillo J, Mager DE, Frisse ME, Brophy RH, Goggin A. Impact of first-line vs. second-line antibiotics for the treatment of acute uncomplicated sinusitis. *JAMA* 2001; **286**(15): 1849–1856.
39. Angeja B, Shlipak MG, Go AS, *et al*. National Registry of Myocardial Infarction 3 Investigators. Hormone therapy and the risk of stroke after acute myocardial infarction in postmenopausal women. *J Am Coll Cardiol* 2001; **38**: 1297–1301.
40. Mast ST, Gersing KR, Anstrom KJ, Krishnan KR, Califf RM, Jollis JG. Association between selective serotonin-reuptake inhibitor therapy and heart valve regurgitation. *Am J Cardiol* 2001; **97**: 989–993.
41. Foody JM, Cole CR, Blackstone EH, Lauer MS. A propensity analysis of cigarette smoking and mortality with consideration of the effects of alcohol. *Am J Cardiol* 2001; **87**(6): 706–711.
42. Hesse B, Diaz LA, Snader CE, Blackstone EH, Lauer MS. Complete bundle branch block as an independent predictor of all-cause mortality: report of 7073 patients referred for nuclear exercise testing. *Am J Med* 2001; **110**: 253–259.
43. Aronow HD, Topol EJ, Roe MT, *et al*. Effect of lipid lowering therapy on early mortality after acute coronary syndromes: an observational study. *Lancet* 2001; **357**(9262): 1063–1068.
44. Shlipak M, Browner WS, Noguchi H, Massie B, Frances CD, McClellan M. Comparison of the effects of angiotensin converting-enzyme inhibitors and beta blockers on survival in elderly patients with reduced left ventricular function after myocardial infarction. *Am J Med* 2001; **110**: 425–433.
45. Stenestrand U, Wallentin L. Early statin treatment following acute myocardial infarction and 1-year survival. *J Am Med Assoc* 2001; **285**(4): 430–436.
46. Grossi E, Goldberg JD, LaPietra A, *et al*. Ischemic mitral valve reconstruction and replacement: comparison of long-term survival and complications. *J Thorac Cardiovasc Surg* 2001; **122**(6): 1107–1124.
47. Mehta S, Yusuf S, Peters RJG, *et al*. Clopidogrel in unstable angina to prevent recurrent events trial (CURE) investigators. Effects of pretreatment with clopidogrel and aspirin followed by long-term therapy in patients undergoing percutaneous coronary intervention: the PCI-cure study. *Lancet* 2001; **358**: 527–533.
48. Rathore SS, Gersh BJ, Weinfurt KP, Oetgen WJ, Schulman KA, Solomon AJ. The role of reperfusion therapy in paced patients with acute myocardial infarction. *Am Heart J* 2001; **142**(3): 516–519.
49. Gillinov AM, Wierup PN, Blackstone EH, *et al*. Is repair preferable to replacement for ischemic mitral regurgitation? *J Thorac Cardiovasc Surg* 2001; **122**(6): 1125–1141.
50. Rothman KJ, Greenland S. *Modern Epidemiology*, 2nd edn. Lippincott-Raven: Philadelphia, 1998.
51. Harrell FE, Jr, Lee KL, Mark DB. Tutorial in biostatistics: multivariable prognostic models: issues in developing models evaluating assumptions and adequacy, and measuring and reducing errors. *Stat Med* 1996; **15**: 361–387.
52. Wang J, Donnan PT. Propensity score methods in drug safety studies: practice, strengths and limitations. *Pharmacoepidemiol Drug Safe* 2001; **10**: 341–344.
53. 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**(3): 280–287.
54. Hosmer DW, Hosmer T, Le Cessie S, Lemeshow S. A comparison of goodness of fit tests for the logistic regression model. *Stat Med* 1997; **16**: 965–980.
55. Pulkstenis E, Robinson TJ. Two goodness-of-fit tests for logistic regression models with continuous covariates. *Stat Med* 2002; **21**: 79–93.
56. Ahlbom A, Steineck G. Aspects of misclassification of confounding factors. *Am J Ind Med* 1992; **21**: 107–112.
57. Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980; **112**(4): 564–569.
58. Marshall J, Hastrup JL. Mismeasurement and the resonance of strong confounders: uncorrelated errors. *Am J Epidemiol* 1996; **143**(10): 1069–1078.
59. Gum PA, Thamilarasan M, Watanabe J, Blackstone EH, Lauer MS. Aspirin use and all-cause mortality among patients being evaluated for known or suspected coronary artery disease: a propensity analysis. *J Am Med Assoc* 2001; **286**(10): 1187–1194.
60. Perkins SM, Tu W, Underhill MG, Zhou XH, Murray MD. The use of propensity scores in pharmacoepidemiologic research. *Pharmacoepidemiol Drug Safe* 2000; **9**: 93–101.
61. Ascione R, Nason G, Al-Ruzzeh S, Ko C, Ciulli F, Angelini GD. Coronary revascularization with or without cardiopulmonary bypass in patients with preoperative nondialysis-dependent renal insufficiency. *Ann Thorac Surg* 2001; **72**: 2020–2025.
62. Cen Y-Y, Glower DD, Landolfo K, *et al*. Comparison of survival after mitral valve replacement with biologic and mechanical valves in 1139 patients. *J Thorac Cardiovasc Surg* 2001; **122**(3): 569–577.
63. Ellis SG, Brener SJ, Lincoff AM, *et al*. Beta-blockers before percutaneous coronary intervention do not attenuate postprocedural creatine kinase isoenzyme rise. *Circulation* 2001; **104**(22): 2685–2688.
64. Feng D, D'Agostino RB, Silbershatz H, *et al*. Hemostatic state and atrial fibrillation (The Framingham Offspring Study). *Am J Cardiol* 2001; **87**: 168–171.
65. Hayashi K, Hasegawa K, Kobayashi S. Effects of angiotensin-converting enzyme inhibitors on the treatment of anemia with erythropoietin. *Kidney Int* 2001; **60**(5): 1910–1916.
66. Ioannidis JP, Galanos O, Katritsis D, *et al*. Early mortality and morbidity of bilateral versus single internal thoracic artery revascularization: propensity and risk modeling. *J Am Coll Cardiol* 2001; **37**(2): 521–528.
67. Kimmel S, Localio AR, Krone RJ, Laskey WK, Registry Committee of the Society for Cardiac Angiography and Interventions. The effects of contemporary use of coronary stents on in-hospital mortality. *J Am Coll Cardiol* 2001; **37**: 499–504.
68. Mehta R, Bruckman D, Das S, *et al*. Implications of increased ventricular mass index on in-hospital outcomes in patients undergoing aortic valve surgery. *J Thorac Cardiovas Surg* 2001; **122**(5): 919–928.
69. Mukamal K, Maclure M, Muller JE, Sherwood JB, Mittleman MA. Prior alcohol consumption and mortality following acute myocardial infarction. *J Am Med Assoc* 2001; **285**: 1965–1970.
70. Normand ST, Landrum MB, Guadagnoli E, *et al*. Validating recommendations for coronary angiography following acute myocardial infarction in the elderly: a matched analysis using propensity scores. *J Clin Epidemiol* 2001; **54**(4): 387–398.
71. Novaro G, Tiong IY, Pearce GL, Lauer MS, Sprecher DL, Griffin BP. Effect of hydroxymethylglutaryl coenzyme A reductase inhibitors on the progression of calcific aortic stenosis. *Circulation* 2001; **104**: 2205–2209.

72. Osswald BR, Blackstone EH, Tochtermann U, *et al*. Does the completeness of revascularization affect early survival after coronary artery bypass grafting in elderly patients? *Eur J Cardiothorac Surg* 2001; **20**(1): 120–125.
73. Peterson JG, Topol EJ, Roe MT, *et al*. Prognostic importance of concomitant heparin with epitifibatide in acute coronary syndromes. Pursuit Investigators. Platelet glycoprotein IIb/IIIa in unstable angina: receptor suppression using integrilin therapy. *Am Coll Cardiol* 2001; **87**(5): 532–536.
74. Rumsfield J, Magid DJ, Plomondon ME, *et al*. Predictors of quality of life following acute coronary syndromes. *Am Coll Cardiol* 2001; **88**: 781–784.
75. Shlipak MG, Angeja BG, Go AS, Frederick PD, Canto JG, Grady D, National Registry of Myocardial Infarction-3 Investigators. Hormone therapy and in-hospital survival after myocardial infarction in postmenopausal women. *Circulation* 2001; **104**: 2300–2304.
76. Suero J, Marso SP, Jones PG, *et al*. Procedural outcomes and long-term survival among patients undergoing percutaneous coronary intervention of a chronic total occlusion in native coronary arteries: a 20-year experience. *J Am Coll Cardiol* 2001; **38**: 409–414.
77. Tu J, Austin PC, Chan BTB. Relationship between annual volume of patients treated by admitting physician and mortality after acute myocardial infarction. *JAMA* 2001; **285**: 3116–3122.
78. Welch RD, Zalenski RJ, Frederick PD, *et al*. Prognostic value of a normal or nonspecific initial electrocardiogram in acute myocardial infarction. *JAMA* 2001; **286**(16): 1977–1984.

