American Journal of Epidemiology
Vol. 163, No. 3
DOI: 10.1093/aje/kwj047
Advance Access publication December 21, 2005

Original Contribution

Results of Multivariable Logistic Regression, Propensity Matching, Propensity Adjustment, and Propensity-based Weighting under Conditions of Nonuniform Effect

Tobias Kurth[1,2,3], Alexander M. Walker[3,4], Robert J. Glynn[1,5,6], K. Arnold Chan[3,4], J. Michael Gaziano[1,2,7], Klaus Berger[8], and James M. Robins[3,6]

[1] Division of Preventive Medicine, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.
[2] Division of Aging, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.
[3] Department of Epidemiology, Harvard School of Public Health, Boston, MA.
[4] i3 Drug Safety, Auburndale, MA.
[5] Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA.
[6] Department of Biostatistics, Harvard School of Public Health, Boston, MA.
[7] Massachusetts Veterans Epidemiology Research and Information Center, Boston VA Healthcare System, Boston, MA.
[8] Institute of Epidemiology and Social Medicine, University of Muenster, Muenster, Germany.

Received for publication February 21, 2005; accepted for publication October 4, 2005.

Observational studies often provide the only available information about treatment effects. Control of confounding, however, remains challenging. The authors compared five methods for evaluating the effect of tissue plasminogen activator on death among 6,269 ischemic stroke patients registered in a German stroke registry: multivariable logistic regression, propensity score–matched analysis, regression adjustment with the propensity score, and two propensity score–based weighted methods—one estimating the treatment effect in the entire study population (inverse-probability-of-treatment weights), another in the treated population (standardized-mortality-ratio weights). Between 2000 and 2001, 212 patients received tissue plasminogen activator. The crude odds ratio between tissue plasminogen activator and death was 3.35 (95% confidence interval: 2.28, 4.91). The adjusted odds ratio depended strongly on the adjustment method, ranging from 1.11 (95% confidence interval: 0.67, 1.84) for the standardized-mortality-ratio weighted to 10.77 (95% confidence interval: 2.47, 47.04) for the inverse-probability-of-treatment-weighted analysis. For treated patients with a low propensity score, risks of dying were high. Exclusion of patients with a propensity score of $<5\%$ yielded comparable odds ratios of approximately 1 for all methods. High levels of nonuniform treatment effect render summary estimates very sensitive to the weighting system explicit or implicit in an adjustment technique. Researchers need to be clear about the population for which an overall treatment estimate is most suitable.

causality; cerebrovascular accident; confounding factors (epidemiology); data interpretation, statistical; logistic models; models, statistical; observational study

Abbreviations: CI, confidence interval; IPTW, inverse-probability-of-treatment weighted; SMR, standardized mortality ratio; t-PA, tissue plasminogen activator.

Nonexperimental observational studies are never considered the "gold standard" for causal inference. However, when randomized trial data are unavailable, observational studies provide the only information about treatment effects.

Correspondence to Dr. Tobias Kurth, Division of Aging, Brigham and Women's Hospital, 1620 Tremont Street, Boston, MA 02120 (e-mail: tkurth@rics.bwh.harvard.edu).

Even when randomized trial data are available, rigorous patient inclusion and exclusion criteria may limit the generalizability of their results. A major methodological problem in observational studies is that investigators have no control over the treatments used by participants. To account for differences in observed covariates in the treatment groups, investigators must frequently carry out analytic adjustments to control confounding when estimating treatment effects (1, 2). Data on scores of potential confounders are often available for analysis, but the richness of this information does not always translate into a reliable analysis (3–6). In studies with a dichotomous outcome, the most common adjustment method is logistic regression of the outcome on treatment and a subset of the pretreatment covariates.

In 1983, alternative methods for control of confounding in observational studies based on the propensity score were proposed (7). The propensity score is the probability that an individual would have been treated based on that individual's observed pretreatment variables. Adjustments using the estimated propensity score tend to balance observed covariates that were used to construct the score. Several adjustment methods incorporating the estimated propensity score have been proposed, including matching (8, 9), regression adjustment (1, 10), and weighting (11–15).

One difficulty faced with all methods of confounder control (i.e., logistic regression or propensity score–based methods) is that if key predictors or important interactions are not included in the outcome regression model or in the propensity score model, then residual confounding due to the excluded covariates and interactions may be substantial. On the other hand, by including all available covariates and their lower-order interactions, the estimate of the treatment effect can be very imprecise and, in nonlinear models such as the logistic, may even be biased (3–5, 16). Thus, it is unclear which adjustment method is preferable in which situation.

To assess the utility of different techniques to adjust for confounding, we used data from a regional German stroke registry to compare estimates of the effect of treatment with tissue plasminogen activator (t-PA) on the in-hospital mortality of ischemic stroke patients. We chose this scenario because some observational studies have shown an increased risk of death associated with t-PA treatment (17–21), while randomized controlled trials demonstrated no causal association between t-PA treatment and death (22–25).

MATERIALS AND METHODS

Description of the data set

The Westphalian Stroke Registry is a regional data bank in northwestern Germany and has been described in detail previously (26, 27). The registry included all patients treated for stroke symptoms who were admitted to the participating 42 hospitals. Patient documentation was performed anonymously and included sociodemographic characteristics, cerebrovascular risk factors, comorbidities, stroke type and severity, and details regarding the treating institution, the mode of admission, diagnostic and therapeutic procedures, complications, and discharge conditions.

Between 2000 and 2001, data for 12,410 patients with stroke symptoms were entered in the registry. Of those patients, 8,208 were diagnosed with ischemic stroke, 2,794 with transient ischemic attacks, 793 with primary intracerebral hemorrhage, and 615 with stroke of unknown mechanism. For our analyses, all but the ischemic stroke cases were excluded. In addition, we excluded 1,880 patients from 19 centers that did not perform t-PA therapy during the time interval of investigation, 27 ischemic stroke cases with intraarterial lysis, and 32 ischemic stroke cases for whom no admission data were available, leaving 6,269 patients for this analysis.

Propensity score

The propensity score is the probability that an individual would have been treated based on that individual's observed pretreatment variables. To describe the propensity score, let the dichotomous (0,1) variable Z indicate treatment, and let $\mathbf{X}$ be the vector of available pretreatment covariates. The propensity score $e(\mathbf{X})$ for an individual is defined as the conditional probability of being treated given his or her covariates $\mathbf{X}$: $e(\mathbf{X}) = \Pr(Z = 1|\mathbf{X})$.

The propensity score is a one-dimensional variable that summarizes the multidimensional pretreatment covariates $\mathbf{X}$. Among persons with a given propensity score, the distribution of the covariates $\mathbf{X}$ is on average the same among the treated and untreated.

Propensity score model

The estimated propensity score, ($\hat{e}(\mathbf{X})$), for t-PA treatment was obtained from the fit of a logistic regression model for which we considered the following pretreatment variables: age (5-year increments), gender, Rankin scale (28) at the time of admission (1–3, 4–5, 6), time from event to hospital admission (<1 hour, 1–3 hours, >3 hours), paresis (monoplegia, hemiplegia, tetraplegia), state of consciousness (awake, somnolent, comatose), type of admitting ward (normal, stroke unit, intensive care unit), transportation to the hospital (emergency medical service, other qualified transport, private, other), aphasia, hypertension (defined as measured systolic blood pressure $\geq$140 mmHg or diastolic pressure $\geq$90 mmHg, or current treatment of hypertension regardless of admission blood pressure), diabetes mellitus (defined as pathologic glucose tolerance test, or two times serum glucose values of $\geq$140 mg/dl), atrial fibrillation, history of other cardiac illnesses, previous stroke, and the admitting clinical center. During the study period, the associations of age (<70 years vs. $\geq$70 years), time from symptoms to admission to the hospital (<1 hour, 1–3 hours, >3 hours), and Rankin scale (1–5, 6) with t-PA treatment changed in the last three 6-month periods compared with the first (29). We therefore added time-covariate interaction terms for these variables into the propensity score model.

Analytical approach

All analyses were performed by using SAS (version 8.2) software (SAS Institute, Inc., Cary, North Carolina). We compared patient pretreatment characteristics with respect to t-PA treatment by using Student's t test for continuous variables and the chi-square test for categorical variables. All p values are two sided.

Our goal was to estimate the effect of treatment with t-PA among ischemic stroke patients on in-hospital mortality by using different methods to control for confounding. Specifically, we compared five methods: multivariable logistic regression adjustment and four propensity score methods (matching, regression adjustment, and two weighted regression adjustments). In the absence of unmeasured confounding, the first of the weighting methods estimates the treatment effect in a population whose distribution of risk factors is equal to that found in all study subjects. This method is also known as inverse-probability-of-treatment-weighted (IPTW) estimator (11, 12). The second weighting method estimates the treatment effect in a population whose distribution of risk factors is equal to that found in the treated study subjects only. This method is known as the standardized mortality ratio (SMR)–weighted estimator (15). Both weighting methods can be interpreted as multivariable standardization methods that use different standard populations. IPTW uses as weights the inverse (estimated) propensity score, $1/\hat{e}(\mathbf{X})$, for treated patients and the inverse of 1 minus the propensity score, $1/(1 - \hat{e}(\mathbf{X}))$, for untreated patients. Thus, IPTW estimates a standardized effect measure with the total study group as the standard population (12, 30). SMR-weighted analyses use as weights the value 1 for the treated and the propensity odds for the untreated, $(\hat{e}(\mathbf{X})/(1 - \hat{e}(\mathbf{X}))$, and estimates a standardized effect measure that considers the exposed group as the standard population (15).

We compared the following five methods for control of confounding:

1. Multivariable-adjusted logistic regression model
2. Logistic regression analyses after matching on the propensity score in a range of ± 0.05
3. Logistic regression model adjusted for the propensity score (as a linear term and as decile categories)
4. IPTW logistic regression model (11, 12) of response on treatment with the weights $1/\hat{e}(\mathbf{X})$ for treated individuals and $1/(1 - \hat{e}(\mathbf{X}))$ for untreated individuals
5. SMR-weighted logistic regression model (15) of the response on treatment with weights of 1 for treated and $\hat{e}(\mathbf{X})/(1 - \hat{e}(\mathbf{X}))$ for untreated individuals

Matching procedure

We matched participants who did not receive t-PA treatment to those treated with t-PA based on a range of ± 0.05 of the propensity score. We chose the matching range of ± 0.05 because it is commonly used, provides reasonable balance of the included covariates, and does not lose many treated individuals as unmatchable. We examined a closer matching range (± 0.01) and did not find results that were meaningfully different from those reported below. To match participants, we used an automated matching procedure in the SAS software that randomly selected a treated individual and randomly selected an untreated individual (comparator) from the pool of potential comparators to determine whether he or she fulfilled the matching criterion. If the selected comparator was eligible, he or she was matched to the treated individual, and the pair was removed. This procedure was repeated until all treated patients were matched to one comparator or until no further comparators fulfilled the matching criteria.

Weighted models

For both weighting methods, we used SAS's GENMOD procedure with a logit link to weight participants and empirical (i.e., robust or sandwich) standard error estimation to calculate the 95 percent confidence intervals. Since our logistic model is saturated, estimates and standard errors based on our weights will be identical to those based on stabilized weights (11). Here, we additionally report the nonparametric 95 percent confidence interval for the IPTW analysis based on 10,000 bootstrap samples drawn with replacement from the study population.

For the unweighted multivariable-adjusted outcome models, we considered the same set of covariates as those considered in the logistic regression model to calculate the propensity score, including the interaction terms mentioned previously.

We evaluated whether the odds ratio between t-PA and death was modified by the propensity score. To do so, we compared a logistic model containing indicator variables for t-PA treatment and propensity score quintiles with a model that also included indicator variables for the product of t-PA and propensity score quintiles with a 4-df likelihood ratio test.

RESULTS

The 23 centers included in this analysis submitted usable data on 6,269 ischemic stroke patients, of whom 212 (3.4 percent) had been treated with t-PA. Of the patients treated with t-PA, 34 (16.0 percent) died during hospitalization. Of those not treated with t-PA, 327 (5.4 percent) died during hospitalization. The distribution of patients' pretreatment characteristics with respect to t-PA treatment is summarized in table 1. Patients who received t-PA treatment were younger, more often had hemiplegia and aphasia, had more severe stroke symptoms as measured by the Rankin scale, were more likely to report a history of atrial fibrillation, were less likely to have had previous strokes, and were more likely to have been living with their family. They were also more likely to have arrived earlier at the hospital and to have been initially admitted to an intensive care unit.

Propensity score model

The logistic model used to estimate the propensity score yielded a c-statistic of 0.94. The mean propensity to receive

TABLE 1. Characteristics of 6,269 ischemic stroke patients, before treatment with tissue plasminogen activator, registered in a German stroke registry between 2000 and 2001, according to treatment status*

Variable	Treatment		*p* value†
	Yes ($n = 212$)	No ($n = 6,057$)	
Age in years (mean (SD‡))	65.7 (12.5)	69.7 (13.0)	<0.01
Male gender	57.6	47.8	0.42
Paresis			
None/unknown	8.2	32.3	<0.01
Monoplegia	1.9	9.1	
Hemiplegia	89.6	57.4	
Tetraplegia	0.5	1.3	
Aphasia	49.5	23.1	<0.01
State of consciousness			
Awake	66.5	75.9	0.87
Somnolent	24.5	10.6	
Comatose	0.9	2.0	
Unknown	8.0	11.6	
Rankin scale			
1–3	5.2	32.9	<0.01
4–5	37.3	41.8	
6	51.4	18.8	
Unknown	6.1	6.5	
Hypertension	70.8	72.5	0.57
Atrial fibrillation	37.3	20.9	<0.01
Heart disease	26.9	26.0	0.76
Diabetes	25.0	31.0	0.06
Other comorbidities	27.8	30.8	0.36
Previous stroke	9.4	17.5	<0.01
Living situation			
Alone	10.9	19.5	0.02
Family	75.0	59.9	
Nursing home	1.4	4.1	
Unknown	12.7	16.5	
Transport to the hospital			
Private	4.7	25.3	<0.01
Emergency medical service	63.2	22.3	
Other qualified	28.3	39.3	
Other/unknown	3.8	13.1	
Time from symptom to admission			
<1 hour	54.3	35.5	
1–3 hours	40.6	15.7	<0.01
>3 hours	5.2	48.8	
Admitting ward			
Normal	1.4	20.7	<0.01
Stroke unit	75.5	63.5	
Intensive care unit	18.4	4.4	
Unknown	4.7	11.4	

* Except for age, all variables are expressed as percentages. Percentages may not sum to 100 because of rounding.

† From *t* test for continuous variables and chi-square test for categorical variables.

‡ SD, standard deviation.

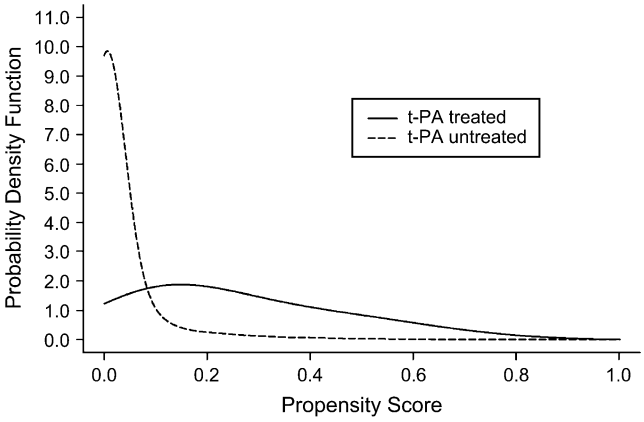

FIGURE 1. Probability density function of the propensity score for the 212 tissue plasminogen activator (t-PA)–treated and the 6,057 t-PA–untreated ischemic stroke patients registered in a German stroke registry between 2000 and 2001.

t-PA treatment for patients actually treated was 0.252 (standard deviation, 0.193) compared with 0.026 (standard deviation, 0.070) for patients not receiving t-PA treatment. The probability density functions of the propensity score for treated and untreated patients are summarized in figure 1. As expected, the distribution of the propensity score for treated patients shifted somewhat toward 1 and for the untreated group toward 0. The figure also illustrates that the overlap of the propensity score for the treated and untreated is limited to a narrow range.

Propensity-stratum-specific effects

For mortality, gradients across levels of the propensity score for the treated and untreated groups were strong and unexpectedly different. Table 2 summarizes information about the proportions of patients who died during the hospital stay in the treated and untreated groups according to percentiles of the propensity score. Several things are notable. First, below the 10th percentile of the overall propensity score, no individuals were in the group treated with t-PA. Second, among those not treated with t-PA, mortality increased with increasing propensity score, while, among those treated with t-PA, mortality decreased with increasing propensity score. As a consequence, the associated empirical odds ratio for in-hospital mortality increased from 0.25 for the 99th percentile of the propensity score to 25.11 for the 10–25th percentile (table 2). This difference was statistically significant. Specifically, for a 4-df test for homogeneity of the treatment odds ratio across quintiles of the propensity score, the p value was 0.008.

Matching

Almost all (96 percent) of the 212 patients who received t-PA treatment could be matched to comparator patients who did not receive t-PA by using a matching criterion of ± 0.05

TABLE 2. Proportion of deaths among 6,269 ischemic stroke patients registered in a German stroke registry between 2000 and 2001 who were treated or not treated with tissue plasminogen activator, according to percentiles of the propensity score for the entire study population

Percentile	Treated ($n = 212$)				Not treated ($n = 6,057$)				Empirical OR*
	Score†	No.	Deaths		Score†	No.	Deaths		
			No.	%			No.	%	
99 to 100	0.5809	36	3	8.3	0.5474	26	7	26.9	0.25
95 to <99	0.3143	73	13	17.8	0.2912	178	27	15.2	1.21
90 to <95	0.1393	55	8	14.6	0.1363	258	19	7.4	2.14
75 to <90	0.0585	31	3	9.7	0.0459	910	82	9.0	1.08
50 to <75	0.0115	10	4	40.0	0.0084	1,558	87	5.6	11.27
25 to <50	0.0017	5	2	40.0	0.0014	1,561	54	3.5	18.60
10 to <25	0.0004	2	1	50.0	0.000267	940	36	3.8	25.11
5 to <10		0	0	0	0.000066	313	6	1.9	
1 to <5		0	0	0	0.000027	251	8	3.2	
0 to <1		0	0	0	0.000007	62	1	1.6	
Overall	0.2521	212	34	16.0	0.0262	6057	327	5.4	3.35

* Propensity-stratum-specific-treatment–mortality odds ratio.
† Mean propensity score in percentile.

of the propensity score. In the groups matched on propensity score, the pretreatment characteristics did not differ meaningfully between treated and untreated patients (table 3). The means of the propensity scores in the matched population were 0.233 (standard deviation, 0.174) for treated and 0.220 (standard deviation, 0.177) for untreated patients.

Comparison of different methods to control for confounding

The crude odds ratio between t-PA treatment and death after ischemic stroke was 3.35 (95 percent confidence interval (CI): 2.28, 4.91). The effect estimates resulting from the five different methods to control for confounding were extremely different and are summarized in table 4. The SMR-weighted analysis yielded the smallest estimated odds ratio of 1.11 (95 percent CI: 0.67, 1.84), followed by the propensity-matched odds ratio of 1.17 (95 percent CI: 0.68, 2.00). Adjusting for the propensity score in a logistic regression model yielded estimated odds ratios ranging from a low of 1.53 (95 percent CI: 0.95, 2.48) when only a linear propensity term was included to a high of 1.96 (95 percent CI: 1.20, 3.20) when both dummy variables for deciles of the propensity score and the other pretreatment covariates were included in a multivariate logistic regression model. The logistic regression model including the individual covariates without the propensity score produced an estimated odds ratio of 1.93 (95 percent CI: 1.22, 3.06). The IPTW model yielded the extreme odds ratio estimate of 10.77, with 95 percent confidence intervals of 2.47, 47.04 as estimated from the robust or sandwich estimator and 1.21, 96.03 when estimated by bootstrapping.

Because of the extreme difference in the empirical t-PA–mortality odds ratio among low-propensity versus high-propensity strata shown in table 2, we report in table 5 analyses restricted to patients whose propensity scores were greater than or equal to 0.05. The crude odds ratio in this restricted population was 1.36 (95 percent CI: 0.84, 2.19). The ranges of estimated odds ratios resulting from the five different methods were no longer significantly different from each other or from the crude odds ratio. The SMR-weighted analysis again yielded the smallest estimated odds ratio of 0.82 (95 percent CI: 0.47, 1.44).

DISCUSSION

Using data on t-PA treatment and mortality from a regional stroke registry, we found that five different methods to control for confounding yielded extremely different treatment effect estimates, ranging from estimated odds ratios of 1.11 (95 percent CI: 0.67, 1.84) for the SMR-weighted estimator to 10.77 (95 percent CI: 2.47, 47.04) for the IPTW estimator. Randomized clinical trials showed no significant effect of t-PA treatment on death among patients with ischemic stroke (22–25). In a cumulative meta-analysis of several randomized trials, the pooled relative risk estimate for death due to t-PA treatment was 1.16 (95 percent CI: 0.95, 1.43) (31). The SMR-weighted estimate and the propensity-matched estimate did not differ significantly from the null value and were very close to the risk estimates obtained from the randomized trials. The effect estimates from outcome models that included the propensity score depended on incorporation of the score. When the score was included as a linear term, no statistically significant association was found. In contrast, all other methods suggested increased risk of in-hospital mortality.

We now argue that the variation we observed in effect estimates cannot be ascribed to the small numbers of subjects in the low-propensity strata and the variability of the

TABLE 3. Characteristics of 406 ischemic stroke patients, before treatment with tissue plasminogen activator, registered in a German stroke registry between 2000 and 2001, according to treatment status after matching on the propensity score*

Variable	Treatment		*p* value†
	Yes ($n = 203$)	No ($n = 203$)	
Age in years (mean (SD‡))	65.6 (12.6)	66.1 (12.9)	0.69
Male gender	57.1	59.1	0.71
Paresis			
None/unknown	8.4	14.3	0.08
Monoplegia	2.0	2.0	
Hemiplegia	89.2	82.8	
Tetraplegia	0.5	1.0	
Aphasia	48.3	50.3	0.69
State of consciousness			
Awake	66.5	64.0	0.54
Somnolent	24.6	25.6	
Comatose	1.0	1.0	
Unknown	7.9	9.4	
Rankin scale			
1–3	5.4	9.4	0.73
4–5	37.9	37.0	
6	51.2	50.3	
Unknown	5.4	3.5	
Hypertension	71.4	73.4	0.66
Atrial fibrillation	37.0	37.0	1.00
Heart disease	26.6	30.5	0.38
Diabetes	24.1	29.1	0.26
Other comorbidities	27.6	26.6	0.82
Previous stroke	9.9	11.3	0.63
Living situation			
Alone	10.3	16.8	0.50
Family	74.9	68.5	
Nursing home	1.5	2.0	
Unknown	13.3	12.8	
Transport to the hospital			
Private	4.9	7.4	0.50
Emergency medical service	62.1	63.1	
Other qualified	29.1	26.1	
Other/unknown	3.9	3.9	
Time from symptom to admission			
<1 hour	54.7	54.2	0.63
1–3 hours	39.9	37.9	
>3 hours	5.4	7.9	
Admitting ward			
Normal	1.5	4.9	0.45
Stroke unit	74.9	71.4	
Intensive care unit	18.7	18.2	
Unknown	4.9	5.4	

* Except for age, all variables are expressed as percentages. Percentages may not sum to 100 because of rounding.

† From *t* test for continuous variables and chi-square test for categorical variables.

‡ SD, standard deviation.

TABLE 4. Comparison of the estimated treatment effect of tissue plasminogen activator on death using multivariable logistic regression, propensity score–matched analysis, regression adjustment with the propensity score, inverse-probability-of-treatment-weighted, and standardized mortality ratio-weighted analyses for ischemic stroke patients registered in a German stroke registry between 2000 and 2001

	No.	OR*	95% CI*
Crude model	6,269	3.35	2.28, 4.91
Multivariable model†	6,269	1.93	1.22, 3.06
Matched on propensity score	406	1.17	0.68, 2.00
Regression adjusted with propensity score			
Propensity score, continuous	6,269	1.53	0.95, 2.48
Multivariable†	6,269	1.85	1.13, 3.03
Propensity score, deciles	6,269	1.76	1.13, 2.72
Multivariable†	6,269	1.96	1.20, 3.20
Weighted models			
IPTW*	6,269	10.77	2.47, 47.04
SMR* weighted	6,269	1.11	0.67, 1.84

* OR, odds ratio; CI, confidence interval; IPTW, inverse-probability-of-treatment weighted; SMR, standardized mortality ratio.

† Adjusted for age, gender, time from symptoms to hospital admission, Rankin scale, paresis, aphasia, state of consciousness, transportation to the hospital, admitting ward, admitting hospital, history of hypertension, diabetes, atrial fibrillation, other cardiac illnesses, previous history of stroke, and interaction terms for follow-up time and age, time from symptoms to admission to the hospital, and Rankin scale.

associated estimated odds ratios. Furthermore, we argue that this variation does not prove or even suggest that any one of the five methods is superior for controlling confounding. The analyses reported in table 4 instead answer different questions implicit or explicit in the adjustment method.

In the absence of unmeasured confounding, the SMR-weighted method estimates the average treatment effect in a population whose distribution of risk factors is equal to that for the t-PA-treated patients only (15). As shown in table 2, most of the treated patients were in the propensity strata with a low associated risk of death and had an empirical odds ratio of less than 2.2, so it is not surprising that the SMR-weighted odds ratio was 1.11. In contrast, IPTW estimates the average effect of treatment in the entire study population, that is, for patients who were and were not treated with t-PA. Given that 65 percent of the entire study population was in the three propensity strata associated with high empirical odds ratios (table 2), it is not surprising that the IPTW estimate is 10.77. Similarly, it is no surprise that the IPTW estimate decreases to 1.09 when the patients in these three strata were excluded by restricting the analysis to the subpopulation of treated and untreated patients whose propensity scores were greater than or equal to 0.05. Indeed, in the subpopulations dominated by patients whose propensity scores were high, the estimated treatment effect is approximately 1. Thus, limiting our study population to patients

TABLE 5. Comparison of the estimated treatment effect of tissue plasminogen activator on death using multivariable logistic regression, propensity score–matched analysis, regression adjustment with the propensity score, inverse-probability-of-treatment-weighted, and standardized mortality ratio-weighted analyses for ischemic stroke patients registered in a German stroke registry between 2000 and 2001, after restriction to participants whose propensity score is ≥0.05

	No.	OR*	95% CI*
Crude model	978	1.36	0.84, 2.19
Multivariable model†	978	1.30	0.74, 2.31
Matched on propensity score	338	0.89	0.49, 1.63
Regression adjusted with propensity score			
Propensity score, continuous	978	0.99	0.58, 1.68
Multivariable†	978	1.29	0.73, 2.29
Propensity score, deciles	978	1.24	0.75, 2.03
Multivariable†	978	1.31	0.74, 2.33
Weighted models			
IPTW*	978	1.09	0.62, 1.93
SMR* weighted	978	0.82	0.47, 1.44

* OR, odds ratio; CI, confidence interval; IPTW, inverse-probability-of-treatment weighted; SMR, standardized mortality ratio.

† Adjusted for age, gender, time from symptoms to hospital admission, Rankin scale, paresis, aphasia, state of consciousness, transportation to the hospital, admitting ward, admitting hospital, history of hypertension, diabetes, atrial fibrillation, other cardiac illnesses, previous history of stroke, and interaction terms for follow-up time and age, time from symptoms to admission to the hospital, and Rankin scale.

whose propensity scores were greater than or equal to 0.05 produced roughly comparable estimates for all methods, none of which differed significantly from 1 and from the results of the randomized trials.

When, as in the present example, the number of untreated subjects is many times larger than the number of treated subjects, propensity score matching will typically result in all or nearly all treated patients being successfully matched, while many untreated patients will remain unmatched and be excluded from the analysis (which may lead to slightly reduced efficiency). As a result, the distribution of covariates in the (successfully) matched subpopulation will be close to that in the treated study population. Thus, it is no surprise that the propensity-matched estimate is very close to the SMR-weighted estimate. Although the SMR-weighted and matched propensity analyses gave similar results in this particular data set, the SMR-weighted analysis has the theoretical advantages that 1) data from all patients are used, and 2) it is not affected by further uncontrolled confounding attributable to the inability to find an exact match for each treated subject (32).

In the present study, as in most studies of adverse drug effects, the death rate was low, and the ratio of untreated to treated subjects was large. Here, there was also strong and statistically significant effect modification. Under such circumstances, the estimated treatment effect from a logistic model that includes only the treatment indicator and the estimated propensity score (as decile indicators or continuous) is largely driven by the magnitude of the treatment effect in propensity strata with the greatest number of treated patients who died. The propensity stratum with the most treated deaths ($n = 13$) had an empirical odds ratio of 1.21 (table 2). Furthermore, among patients treated with t-PA, the fraction ($7/34 = 21$ percent) of all deaths that occurred in the three propensity strata with large empirical odds ratios exceeded the fraction ($17/212 = 8$ percent) of all nondeaths that occurred in these strata. Thus, it is no surprise that the estimate from the logistic model adjusted for the propensity score would somewhat exceed the SMR-weighted estimate, as confirmed by our results. Adjustment for covariates in addition to the propensity score made little difference.

Data sets in which a much larger untreated population has a strikingly different risk factor distribution than a small treated population are common in observational studies of drug effects. In such studies, the low-propensity strata are composed of those members of the population for whom most physicians regard treatment as inappropriate, and such patients would not meet the inclusion criteria of randomized clinical trials. The SMR-weighted and the propensity-matched estimates most closely approached results observed in the clinical trials that evaluated the association of t-PA treatment in ischemic stroke patients (22–25). Presumably, this finding reflects a focus in these analyses on individuals who would have been eligible to participate in clinical trials on the basis of their characteristics. Restriction to persons whose propensity scores are greater than or equal to 0.05 focuses even more sharply on this target population.

However, the similarity of the result obtained with the SMR-weighted and propensity score–matched analyses to the results of the randomized trials should not be taken as evidence that, compared with other multivariable outcome models, these two methods are a better tool to adjust for covariates in observational research. Indeed, once we restricted the analysis to subjects whose propensity score exceeded 0.05, all adjustment methods gave fairly similar results. In addition, in most studies in the literature, the effect estimates from multivariable regression models were quite close to the effect estimates derived from various implementations of the propensity score, as long as the number of outcome events was much larger than the number of potential confounders (16, 33–35). An apparent advantage in using the propensity score, however, may be that the strong effect modification in this clinical example is very obvious across propensity score strata, as shown in table 2. This effect modification may be difficult to unveil when evaluating individual risk factors.

A caution with regard to the use of weighted methods is that they can perform poorly when the weights for a few subjects are very large. Although some partial approximate fixes have been described (36, 37), there is no perfect solution for this problem. These few large weights imply that the population parameter estimated by the weighted method (e.g., the average effect of treatment in the entire population for the IPTW method) cannot be accurately derived from the data in the absence of additional a priori assumptions.

In this setting, the estimated standard-error-of-treatment effect, whether based on a robust variance estimator or the bootstrap, may underestimate the true difference between the weighted estimator and the population parameter it estimates.

With respect to this particular clinical example, two possible explanations should be considered for the high odds ratios found for the three low-propensity strata. First, the time period covered in our study began when t-PA treatment for patients with ischemic stroke was officially approved in Germany (August 2000) and thus reflects first experiences with t-PA treatment. Initiation of t-PA treatment for patients who had low propensities for this therapy may represent a last-ditch effort to salvage patients characterized by unrecorded situation-specific and individual risk factors. With this explanation, the high odds ratios in the low-propensity strata were due to unmeasured confounders (e.g., presence of extreme severity, physician or family attitudes) not recorded for data analysis. A second possibility is that patients who had a low propensity for treatment may have had contraindications to t-PA based on their recorded values for some covariates in the database. In that case, the high odds ratio for these patients would be solely attributable to effect modification by these covariates and not to confounding.

Seen in this light, the IPTW analysis gives at best (i.e., when the effect-modification rather than the confounding explanation is the case) the ''right'' answer to the possibly ''wrong'' question: What would be the effect of giving t-PA to every patient with an acute ischemic stroke? That question would be wrong if prior knowledge had already ruled out administering t-PA to subjects with certain contraindications. On the other hand, to develop new indications for treatment or to conduct surveillance of the appropriateness of the current indications and protocol, the effect of t-PA treatment in patients whose propensity scores are low may be of importance. It remains to be specified which factors are associated with increased risk of death for patients with a low propensity for t-PA treatment. Little clinical experience with t-PA may be one of these factors, as suggested by some studies (19–21, 29).

In summary, in the setting of a nonuniform treatment effect across covariate or propensity-score levels, much of the difference between the adjustment strategies resulted from explicitly or implicitly incorporating low-propensity patients, who were uncommon in the treated group, common in the untreated group, and radically different from other patients with respect to treatment-associated risk of death. High levels of nonuniform treatment effect render summary estimates very sensitive to the weighting system explicit or implicit in an adjustment technique. The divergent results conceivably may all be correct but are answers to different questions. The researcher needs to be clear as to the population for which an overall treatment estimate is most suitable.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the contributions of the following departments and participating hospitals in Germany to the Westphalian Stroke Registry, and their continuing cooperation:

Departments of Neurology: Bathildis Krankenhaus, Bad Pyrmont; Bergmannsheil Bochum; St. Josef-Hospital, Bochum; Knappschafts-Krankenhaus Bottrop; Evang. Krankenhaus, Castrop-Rauxel; Krankenhaus St. Elisabeth-Stift, Damme; Knappschaftskrankenhaus Dortmund; Städtische Kliniken, Dortmund; Franziskus Hospital, Dülmen; Hans-Susemihl-Krankenhaus, Emden; Evangelisches Krankenhaus Gelsenkirchen; Universitätsklinik Greifswald; St. Johannes Hospital, Hagen; St. Marien-Hospital Hamm; Frederikenstift, Hannover; Gemeinschaftskrankenhaus Herdecke; Klinikum Lippe-Lemgo; Klinikum Minden; Universität Münster; Klinikum Osnabrück; St. Vincenz Krankenhaus, Paderborn; Christliches Krankenhaus Quakenbrück; Knappschafts-Krankenhaus Recklinghausen; and Klinikum Wuppertal. *Departments of Internal Medicine:* St. Josef-Hospital Bochum-Linden; Marienhospital Arnsberg; Kreiskrankenhaus Diepholz; Kath. Krankenhaus West, Dortmund; Städtische Kliniken Nord, Dortmund; St. Josefs-Hospital, Dortmund; Krankenhaus Bethanien, Dortmund; Evangelisches Krankenhaus, Hamm; St. Barbara Klinik Heesen; Krankenhaus Lübbecke; Franziskus-Hospital, Münster; and Joseph-Hospital, Warendorf. *Departments of Geriatric Medicine:* Hüttenhospital, Dortmund; Marienhospital, Herne; Elisabeth-Krankenhaus, Recklinghausen; St. Marien Hospital, Vechta; and Marienhospital Wattenscheid.

Conflict of interest: none declared.

REFERENCES

1. D'Agostino RB, Jr. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. Stat Med 1998;17:2265–81.
2. Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. Am J Epidemiol 1986; 123:392–402.
3. Harrell FE Jr, Lee KL, Califf RM, et al. Regression modelling strategies for improved prognostic prediction. Stat Med 1984; 3:143–52.
4. Harrell FE Jr, Lee KL, Mark DB. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat Med 1996;15:361–87.
5. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996;49:1373–9.
6. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol 2000;151:531–9.
7. Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika 1983;70:41–55.
8. Rosenbaum PR, Rubin DB. Constructing a control group using multivariate matched sampling methods that incorporates the propensity score. Am Stat 1985;39:33–8.
9. Rubin DB, Thomas N. Matching using estimated propensity scores: relating theory to practice. Biometrics 1996;52: 249–64.

10. Rubin DB, Thomas N. Combining propensity score matching with additional adjustments for prognostic covariates. J Am Stat Assoc 2000;95:573–85.
11. Robins JM. Marginal structural models. In: 1997 Proceedings of the Section on Bayesian Statistical Science. Alexandria, VA: American Statistical Association, 1998:1–10.
12. Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology 2000;11:550–60.
13. Hirano K, Imbens GW. Estimation of causal effects using propensity score weighting: an application to data on right heart catheterization. Health Services Outcomes Res Methodol 2001;2:259–78.
14. Hirano K, Imbens GW, Ridder G. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 2003;71:1161–89.
15. Sato T, Matsuyama Y. Marginal structural models as a tool for standardization. Epidemiology 2003;14:680–6.
16. Cepeda MS, Boston R, Farrar JT, et al. 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.
17. Katzan IL, Furlan AJ, Lloyd LE, et al. Use of tissue-type plasminogen activator for acute ischemic stroke: the Cleveland area experience. JAMA 2000;283:1151–8.
18. Reed SD, Cramer SC, Blough DK, et al. Treatment with tissue plasminogen activator and inpatient mortality rates for patients with ischemic stroke treated in community hospitals. Stroke 2001;32:1832–40.
19. Heuschmann PU, Berger K, Misselwitz B, et al. Frequency of thrombolytic therapy in patients with acute ischemic stroke and the risk of in-hospital mortality: the German Stroke Registers Study Group. Stroke 2003;34:1106–13.
20. Heuschmann PU, Kolominsky-Rabas PL, Misselwitz B, et al. Predictors of in-hospital mortality and attributable risks of death after ischemic stroke: the German Stroke Registers Study Group. Arch Intern Med 2004;164:1761–8.
21. Heuschmann PU, Kolominsky-Rabas PL, Roether J, et al. Predictors of in-hospital mortality in patients with acute ischemic stroke treated with thrombolytic therapy. JAMA 2004;292:1831–8.
22. Tissue plasminogen activator for acute ischemic stroke. The National Institute of Neurological Disorders and Stroke rt-PA Stroke Study Group. N Engl J Med 1995;333:1581–7.
23. Hacke W, Kaste M, Fieschi C, et al. Intravenous thrombolysis with recombinant tissue plasminogen activator for acute hemispheric stroke. The European Cooperative Acute Stroke Study (ECASS). JAMA 1995;274:1017–25.
24. Hacke W, Kaste M, Fieschi C, et al. Randomised double-blind placebo-controlled trial of thrombolytic therapy with intravenous alteplase in acute ischaemic stroke (ECASS II). Second European-Australasian Acute Stroke Study Investigators. Lancet 1998;352:1245–51.
25. Clark WM, Wissman S, Albers GW, et al. Recombinant tissue-type plasminogen activator (Alteplase) for ischemic stroke 3 to 5 hours after symptom onset. The ATLANTIS Study: a randomized controlled trial. Alteplase Thrombolysis for Acute Noninterventional Therapy in Ischemic Stroke. JAMA 1999;282:2019–26.
26. Schmidt WP, Berger K, Taeger D, et al. Influence of institutional factors in neurological, medical and geriatric departments on length of stay in patients with stroke. (In German). Dtsch Med Wochenschr 2003;128:979–83.
27. Schmidt WP, Taeger D, Buecker-Nott HJ, et al. The impact of the day of the week and month of admission on the length of hospital stay in stroke patients. Cerebrovasc Dis 2003;16:247–52.
28. Berger K, Weltermann B, Kolominsky-Rabas P, et al. The reliability of stroke scales. The German version of NIHSS, ESS and Rankin scales. (In German). Fortschr Neurol Psychiatr 1999;67:81–93.
29. Berger K, Talbot D, Schmidt WP, et al. The learning curve—changes in hospital mortality of patients with ischemic stroke receiving thrombolysis with t-PA. (Abstract). Neurology 2003;5:S A429.
30. Hernan MA, Brumback B, Robins JM. Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. Epidemiology 2000;11:561–70.
31. Wardlaw JM, Sandercock PA, Berge E. Thrombolytic therapy with recombinant tissue plasminogen activator for acute ischemic stroke: where do we go from here? A cumulative meta-analysis. Stroke 2003;34:1437–42.
32. Rosenbaum PR, Rubin DB. The bias due to incomplete matching. Biometrics 1985;41:103–16.
33. Braitman LE, Rosenbaum PR. Rare outcomes, common treatments: analytic strategies using propensity scores. Ann Intern Med 2002;137:693–5.
34. Sturmer T, Schneeweiss S, Brookhart MA, et al. Analytic strategies to adjust confounding using exposure propensity scores and disease risk scores: nonsteroidal antiinflammatory drugs and short-term mortality in the elderly. Am J Epidemiol 2005;161:891–8.
35. Shah BR, Laupacis A, Hux JE, et al. Propensity score methods gave similar results to traditional regression modeling in observational studies: a systematic review. J Clin Epidemiol 2005;58:550–9.
36. Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for nonignorable drop-out using semiparametric non-response models. J Am Stat Assoc 1999;94:1096–120.
37. Potter FJ. The effect of weight trimming on nonlinear survey estimates. In: Proceedings of the Section on Survey Research Methods of the American Statistical Association, San Francisco, California, 1993. J Am Stat Assoc 1993:758–63.