Balancing Confounding and Generalizability Using Observational, Real-world Data: 17-gene Genomic Prostate Score Assay Effect on Active Surveillance

Steven Canfield, MD,[1] Michael J. Kemeter, MSPAS,[2] Phillip G. Febbo, MD,[2] John Hornberger, MD, MS, FACP[2]
[1]University of Texas, Houston, TX; [2]Genomic Health, Inc., Redwood City, CA

Randomized, controlled trials can provide high-quality, unbiased evidence for therapeutic interventions but are not always a practical or viable study design for certain healthcare decisions, such as those involving prognostic or predictive testing. Studies using large, real-world databases may be more appropriate and more generalizable to the intended target population of physicians and patients to answer these questions but carry potential for hidden bias. We illustrate several emerging methods of analyzing observational studies using propensity score matching (PSM) and coarsened exact matching (CEM). These advanced statistical methods are intended to reveal a "hidden experiment" within an observational database, and so refute or confirm a potential causal effect of assignment to an intervention and study outcome. We applied these methods to the Optum™ Research Database (ORD; Eden Prairie, MN) of electronic health records and administrative claims data to assess the effect of the 17-gene Genomic Prostate Score® (GPS™; Genomic Health, Redwood City, CA) assay on use of active surveillance (AS). In a traditional multivariable logistic regression, the GPS assay increased the use of AS by 29% (95% CI, 24%-33%). Upon applying the matching methods, the effect of the GPS assay on AS use varied between 27% and 80% and the matched data were significant among all algorithms. All matching algorithms performed well in identifying matched data that improved the imbalance in baseline covariates. By using different matching methods to assess causal inference in an observational database, we provide further confidence that the effect of the GPS assay on AS use is statistically significant and unlikely to be a result of confounding due to differences in baseline characteristics of the patients or the settings in which they were seen.

[Rev Urol. *2018;20(2):69–76 doi: 10.3909/riu0799*]

KEY WORDS

Prostate cancer • Active surveillance • Evidence-based practice • Comparative effectiveness research • Genomic biomarker • Propensity score • Matching

Observational studies have been used extensively to assess associations between interventions and outcomes.[1] Such studies can often be performed more quickly and less expensively than randomized trials, while providing insights about real-world clinical practice. One critical challenge of observational studies is discerning whether the observed associations reflect a cause-and-effect relationship. Specifically, another variable, or set of variables referred to as confounders, may be the true cause of the effect seen in the intervention and in the outcome. Figure 1 illustrates a causal graph of the effects among a confounder, an intervention assignment, and an outcome.[2] In Figure 1(A), the confounder influences the intervention assignment and the outcome; however, there is no effect of the intervention assignment on the outcome. For example, older age or comorbidities may affect assignment to one or an other intervention and affect the outcome (eg, survival), but there exists no causal effect of the intervention assignment on the outcome. In Figure 1(B), the confounder still exists, but there also exists a causal effect of the intervention assignment on the outcome.

A gold-standard solution to the problem of confounding is to randomly assign the intervention in a randomized, controlled trial. With a sufficiently large number of trial participants, random intervention assignment is expected to provide balance in the empirical distribution of the baseline variables that may be confounders. Consequently, the causal effect of an intervention on an outcome can be more confidently inferred, regardless of the causal effects of any confounding variables.

A well-known limitation of randomized clinical trials is that the trial participants may not be representative of the population of interest for whom the findings are ultimately intended to be applied. For example, higher rates of trial participation are associated with personal factors (younger age, male sex, white race), social factors (higher education and income), and structural factors (simpler informed consent procedures, convenience to the trial center, lower costs of participation).[3-5] One approach to understand the generalizability of trial findings to real-practice settings is to compare the balance in the distributions of baseline characteristics between trial participants and the target population.

Another approach is to use large, observational datasets that are more likely to be representative of the target population than many clinical trials, and to apply advanced statistical methods to control for confounding variables. Multivariable regression analysis has been used for decades to control for imbalances in baseline covariates that could explain observed effects between the treated and control groups.[6-12] Propensity scoring analysis has been applied as an advanced method to refine multivariable regression analysis and reduce potential for bias even further. An even more recent refinement technique, coarsened exact matching (CEM), can reduce bias even further. Pre-processing matching of participants in a database has emerged as a method to obtain a dataset that might have resulted from a randomized experiment that is not transparent within an observational dataset. The intent of matching is to reveal the "hidden experiment" upon which to make contrast, and to have greater confidence in the revealed inferences. Operationally, the aim of matching is to reduce imbalance in the empirical distribution of the known pre-treatment confounders between the comparison groups.

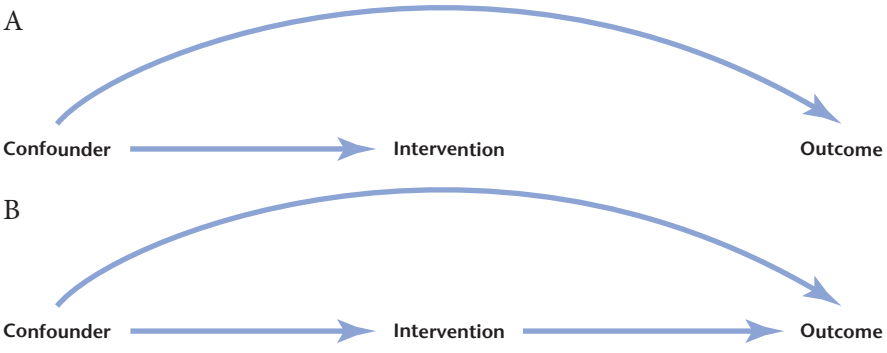

Figure 1. Causal graphs. **(A)** Associations between confounder, intervention, and outcome when confounder has effect on intervention and the outcome, but there is no causal effect of the intervention on the outcome. **(B)** Associations between confounder, intervention, and outcome when confounder has effect on intervention and the outcome, and there exists also a causal effect of the intervention on the outcome.

We herein illustrate the application of matching in analyzing the effect of the 17-gene GPS assay on active surveillance (AS) use in men with low-risk prostate cancer (PCa) who received GPS testing or were not tested, utilizing a large US payer system. This genomic assay has been commercially available since 2013, validated for multiple clinical endpoints, and shown to provide clinically meaningful results to newly diagnosed patients electing AS or definitive treatment.[13-15, 21-23] In our original analysis of this dataset, we performed multivariable logistic regression to control for variations in baseline covariates and found that at 6 months of follow-up, AS use was 31.0% higher (95% CI, 27.6%-34.5%; $P < 0.001$) for men receiving the GPS test only versus men who did not undergo GPS testing.[16] In the study described here, we contrast and compare the results of different matching methods with each other and with traditional multivariable regression analysis to appraise the robustness of observed causal effects of the GPS assay on AS use.

Methods

Data Sources, Patient Selection, and Study Measures

We used extracts of the Optum™ Research Database (ORD) of electronically stored medical records and administrative claims data linked to enrollment information and laboratory data from a large US health insurer offering both commercial and Medicare Advantage health plans. Details about the ORD and selection of patients were reported previously.[16]

Baseline patient characteristics included age at diagnosis, year of diagnosis, census geographical region, insurance status, and number of comorbidities. Genomic testing was recorded using the Common Procedural Terminology (CPT) codes 84999, 81479, and/or 81599. Other tests were based on the test's name and measurement. For example, "test_name = 'Prostate-specific antigen (PSA)', measurement type = 'Gleason'". PCa–related procedures were recorded for patients in the cohort at six- and 12-months of follow-up, and included radical prostatectomy, radiation therapy, brachytherapy, cryotherapy, or hormone therapy. A PCa-related procedure was reported if the patient had only one procedure during the follow-up period. A patient with more than one procedure over the interval was designated as having had multiple procedures. A patient with no recorded PCa–related procedures was designated as having been assigned to AS.

Data Analysis

We applied propensity scoring matching (PSM) and CEM.[9,17] For PSM, we applied the following methods: nearest-neighbor, genetic, optimal, and full. CEM uses an automatic binning algorithm to coarsen values of specified covariates (eg, "coarsened" number of diagnoses rather than exact number of diagnoses). It provides exact matched observations based on these coarsened values, dropping observations in both the treatment and control groups without an exact match.

The extracted data were aggregated by type of treatment, year, and whether the follow-up observation was at 6 or 12 months. We then restricted the analysis to the 6-month data and expanded the dataset by randomly assigning values based on means and standard deviation (SD) if continuous variables (eg, age, number of diagnoses, and number of selected diagnoses) and Dirichlet distributions if bracketed (insurance status and region of treatment). The final dataset included 300 GPS-tested patients and 7446 patients who had no testing (total n = 7746).

We performed all analyses using R implemented in RStudio (Version 1.1.414, © 2009-2018 RStudio, Inc). We applied MatchIt, designed to work in conjunction with the R programming language and statistical software R Development Core Team (2011). We initially assessed the record of using GPS Testing versus No Testing as a function of the baseline covariates, using logit link (default in MatchIt).

We applied all analyses to correct for imbalance among baseline covariates prior to estimating the mean treatment effects. We ran generalized linear modeling (GLM) using logit family on the matched samples, applying weights produced by the algorithms. All GLM models were run with the Zelig library in the R studio programming platform.[18]

Results

Balance of Covariates Before and After Matching

We first looked at the balance, measured by standardized mean difference for each covariate between the treated (GPS Testing) and control (No Testing) patients (Table 1). The largest imbalance was found for the covariate "number of selected diagnoses", followed by "number of all diagnoses" and "age". A higher proportion of patients undergoing GPS testing had commercial insurance than other insurances. The standardized mean difference declined for all covariates after matching, especially in the covariates with the largest standardized mean differences (Figure 2).

TABLE 1

Summary of Balance Between No Testing and GPS Testing for Raw Data and After Matching (Nearest Neighbor)

	Raw Data			Matched Data		
Covariate	No Testing (n = 7446)	GPS (n = 300)	Standardized Mean Difference[a]	No Testing (n = 7446)	GPS (n = 300)	Standardized Mean Difference[a]
Age, mean (SD)	66.5	63.9	−0.31*	63.6	63.5	0.04
Region, n (%)						
Northwest	12%	9%	−0.09	6%	8%	0.08
Midwest	52%	47%	−0.11	50%	53%	−0.03
South	26%	31%	0.13	31%	25%	−0.12
West	9%	12%	0.11	11%	14%	0.13
Other	1%	0%	−0.07	0%	0%	0.00
Insurance, n (%)						
Commercial	44%	52%	0.18*	54%	56%	0.03
Medicare	28%	23%	−0.09*	20%	22%	0.05
Medicaid	1%	0%	−0.11*	1%	0%	−0.06*
Uninsured	7%	7%	0.02*	7%	8%	0.02
Multiple	21%	17%	−0.09*	19%	15%	−0.08
Other	7%	6%	−0.06*	6%	6%	−0.08
Number of selected diagnoses, mean (SD)[b]	0.29	0.16	−3.36*	0.20	0.20	0.13*
Number of all diagnoses, mean (SD)[c]	0.16	0.14	−0.45*	0.16	0.16	−0.1

[a]The standardized difference compares the difference in means in units of the pooled standard deviation.
[b]Erectile dysfunction, incontinence, cystitis, prostatitis.
[c]History of malignant neoplasm or symptoms or involving respiratory system and other chest symptoms.
*$p < 0.05$.

The covariate-adjusted probability—*propensity score* estimated by logistic regression—for GPS testing among patients without testing (No Testing) was 2% (95% CI, 0%-2%) (Figure 3A). By contrast, the covariate-adjusted propensity of GPS testing among the 300 patients who received GPS testing was 61% (95% CI, 57%-65%). After nearest-neighbor matching, the propensity scores were substantially overlapping with an average probability of 50% (95% CI, 39%-60%) in those without testing and 50% (95% CI, 39%-61%) in those who had GPS testing (Figure 3B). We also applied CEM, a newer, nonparametric method that bins ("coarsens") levels of covariates intended to lower imbalance that may result from missing data and the dependency on the research analyst's specification of the regression model.[9]

Causal Effects of GPS Testing on AS After Matching

After matching, some covariates were dropped due to collinearities with other variables or small proportions of patients derived by the match, such as "Region–Other" or "Insurance–Multiple". When we tested the effect of No Testing versus GPS Testing on AS use by multivariable regression using our original published report and after applying various matching methods (Table 2), different matching algorithms selected different matched pairs. For example, the nearest-neighbor PSM algorithm found a match that included 132 subjects in both groups. CEM resulted in 981 untested patients and 99 GPS-tested patients. In the raw data of 7446 untested patients, the rate of AS use was 40%. AS use was similarly 40% to 41% among these patients using nearest-neighbor and optimal PSM algorithms. Untested patients matched by CEM or full PSM algorithms had an estimated propensity for GPS testing

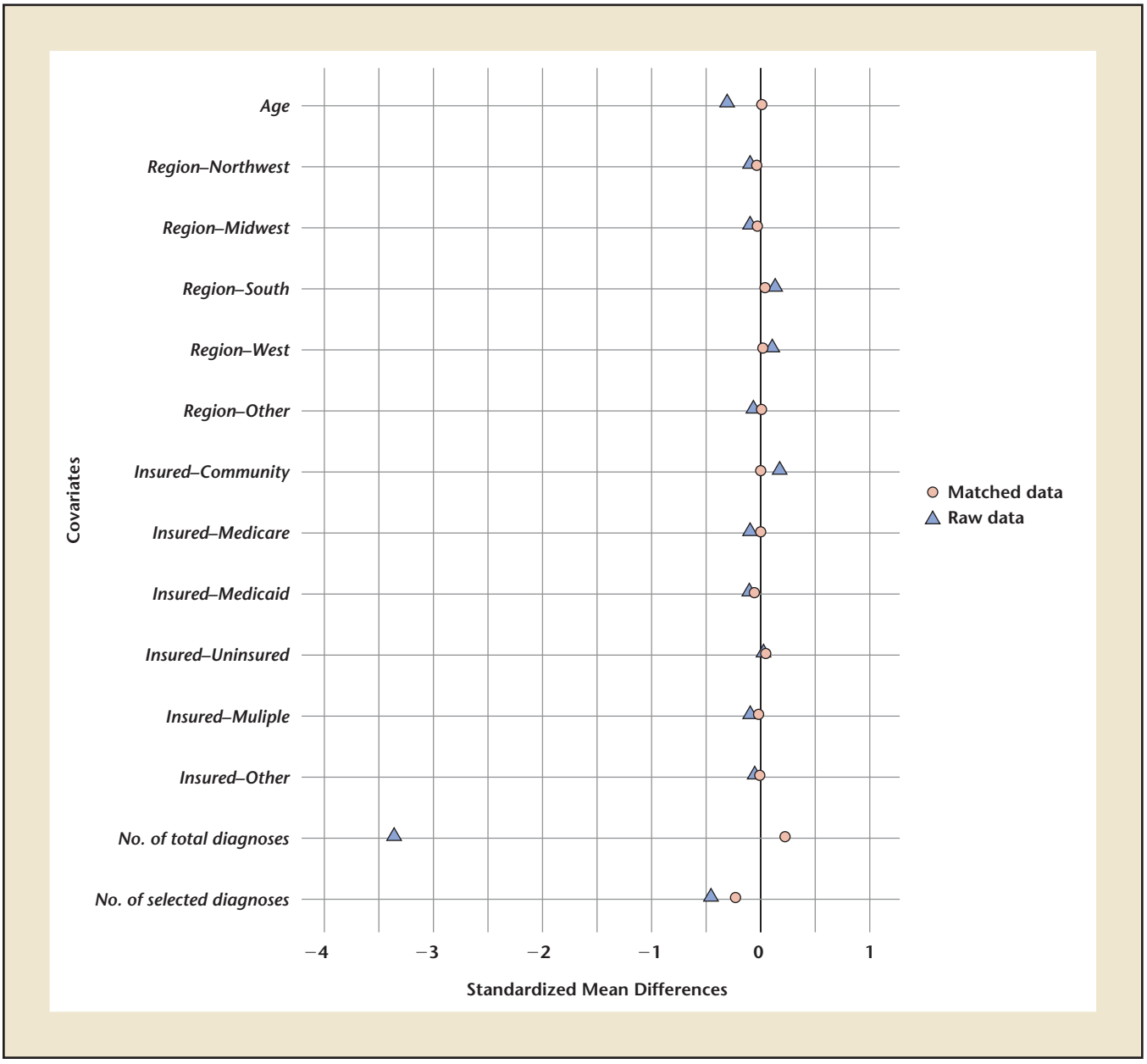

Figure 2. Effect of matching (nearest neighbor), relative to raw data, on the standardized mean difference between GPS Testing and No Testing for each covariate.

of 58% to 60%, respectively. The untested patients matched with the genetic PSM algorithm had the lowest propensity for GPS testing, equal to 13%.

In the multivariable logistic regression, GPS testing was associated with a 31% (95% CI, 24%-33%) mean difference in AS use (Table 2). Using PSM or CEM, the mean differences in AS use between GPS Testing and No Testing were approximately the same or higher than the differences using multivariable regression; all differences were statistically significant. Nearest neighbor, optimal, and full matching PSM algorithms all had mean differences of approximately 30%. The mean differences for CEM and genetic PSM algorithms were 80% (95% CI, 72%-86%) and 41% (37%-44%), respectively.

Discussion

In assessing the association between GPS testing and AS use, we sought to strike a fair and reasonable balance between (1) the generalizability of using a database, the ORD, with representation of the target population and (2) the application of state-of-the-art methods to control for baseline-confounding variables. The various methods to control for confounding bias revealed improved balance among baseline covariates, especially in those that had the greater imbalance without matching, such as age and the number of diagnoses. Regardless of the method for assigning matched sets, the effect of GPS Testing on AS use relative to no testing remained clinically

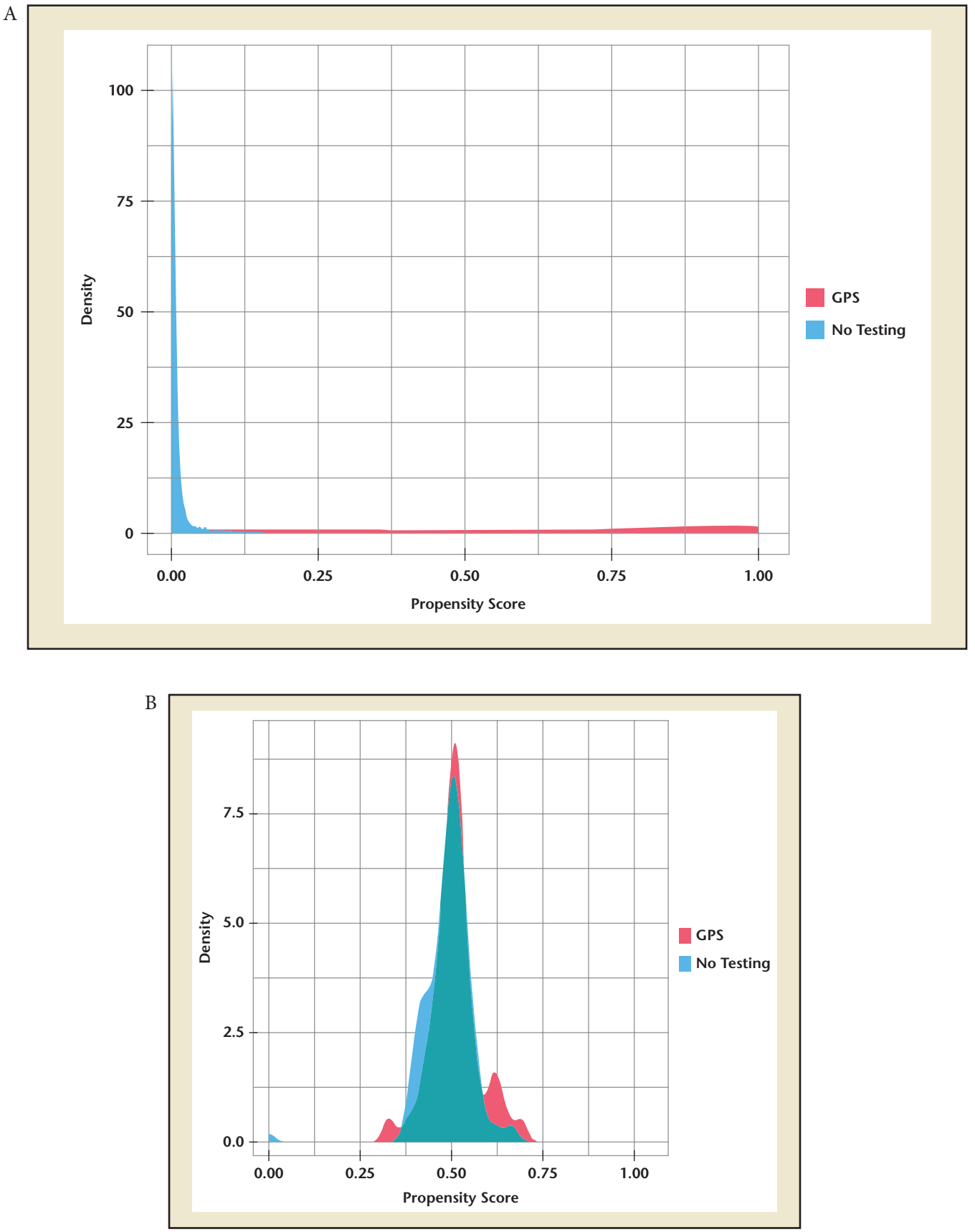

Figure 3. Distribution of propensity scores, by testing (No Testing and GPS Testing) and between data sets {raw data [Figure 2(A)] and matched data [Figure 2(B)]}. **(A)** Raw data (No Testing, 7446 patients; GPS testing, 300 patients). **(B)** After nearest-neighbor matching (No Testing, 132 patients; GPS Testing, 132 patients).

TABLE 2

Comparison of Matching Metrics and Causal Effects From Multivariable Regression, Propensity Score Matching (PSM), and Coarsened Exact Matching (CEM)

Balancing Method	Number of Patients per Group		AS Use in No Testing	Mean Difference in AS Use
	No Testing	GPS		GPS vs No Testing (95% CI)
Multivariable regression	7446	300	40.0%	29% (24%, 33%)
Propensity score matching				
Nearest neighbor	132	132	40.0%	33% (0.2%, 70%)
Optimal	300	300	41.0%	27% (0.1%, 80%)
Genetic	108	300	13.0%	80% (72%, 86%)
Full	7446	300	60.0%	29% (24%, 32%)
Coarsened exact matching	981	99	53.4%	41% (37%, 44%)

meaningful and statistically significant. In fact, all but one of the matching methods revealed a larger effect of GPS testing on AS use than did the original multivariable regression estimation.[16]

A well-known limitation with any study, randomized or not, is that only covariates that are measurable—or extractable—from the dataset can be examined explicitly for balance between the intervention-assignment groups. A large enough sample size in a randomized trial increases confidence but does not necessarily assure that balance has occurred in the potential, unmeasured confounders. Increasing the sample size extracted from a single observational, non-randomized dataset does not, however, have the same property unless there is substantial correlation between the measured and unmeasured variables.[19] Researchers have been exploring and developing methods to assess the influence of unmeasured confounding in observational studies; a long-standing qualitative approach is to assess the presence of a consistent direction and magnitude of effect across multiples studies, especially if they exhibit differences in aforementioned personal, social, and structural factors.[20] Notably, the findings herein are consistent with three previously reported studies in different settings showing that the GPS assay provides clinically meaningful results to newly diagnosed PCa patients electing AS or definitive treatment.[21-23]

When subjects are matched into blocks or strata, as in CEM, reductions in the sample population can occur. Unlike this dataset, this potential threat to causal inference is especially relevant when the size of the control group is small compared with the size of the treatment group. Exploring different matching approaches, as we have done in this study, allows a more explicit view into the trade-off in lost precision in causal effects from smaller samples in the original population (increased variance) versus potential bias reduction due to improved balance and homogeneity in the matched cohorts. Our use of several different methods for preprocessing matching to improve causal inference in observational data provides further confidence that the effect of GPS testing on AS use is statistically significant and unlikely due to confounding through differences in baseline characteristics of the patients or the settings in which they were seen. These findings complement existing reports showing that the GPS assay provides clinically meaningful results to newly diagnosed PCa patients electing AS or definitive treatment.[21-23]

AS in low-risk PCa patients is recommended by the American Urological Association and the National Comprehensive Cancer Network.[24-26] Despite the guidelines, however, substantial variability in AS use exists across many centers, illustrating well-known challenges to changing management patterns.[27,28] Practice-pattern variables have been attributed to clinical and non-clinical factors.[28,29] Among these, even physicians who have endorsed AS report barriers in persuading patients of the value of a well-validated approach for deferring immediate treatment.[30] How patients perceive and experience uncertainty in decisions is an evolving field, revealing that regret avoidance is one of many motivators for patients and physicians to desire additional objective, personalized data about future risks.[31,32] The increased rate of AS

use with GPS testing that we found in our study, regardless of the method for matching patients to mimic an experiment hidden in a real-world observational dataset, suggests that GPS testing is an efficient approach to align clinical decisions with guideline recommendations. ■

Funding for this research provided by Genomic Health, Inc. (Redwood City, CA).

The authors wish to acknowledge Bethann Hromatka, Michele Lee, Ruixiao Lu, and Donna Polizio for their copy-editing, support, and overall review of the manuscript.

References

1. Austin PC. The use of propensity score methods with survival or time-to-event outcomes: reporting measures of effect similar to those used in randomized experiments. *Stat Med*. 2014;33:1242-1258.
2. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. *Epidemiology*. 2004;15:615-625.
3. Bell JAH, Balneaves LG. Cancer patient decision making related to clinical trial participation: an integrative review with implications for patients' relational autonomy. *Support Care Cancer*. 2015;23:1169-1196.
4. Brown DR, Topcu M. Willingness to participate in clinical treatment research among older African Americans and Whites. *Gerontologist*. 2003;43:62-72.
5. Walter JK, Davis MM. Participate in clinical research. *IRB Ethics Hum Resour*. 2016;38:15-20.
6. Campbell G, Yue LQ. Statistical innovations in the medical device world sparked by the FDA. *J Biopharm Stat*. 2016;26:3-16.
7. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42:1-28.
8. Iacus SM, King G, Porro G. cem: software for coarsened exact matching. *J Stat Softw*. 2009;30(9). doi:10.18637/jss.v030.i09.
9. Iacus SM, King G, Porro G. Causal inference without balance checking: coarsened exact matching. *Polit Anal*. 2012;20:1-24.
10. Ho DE, Imai K, King G, Stuart EA. MatchIt: nonparametric preprocessing for parametric causal inference. *J Stat Softw*. 2011;42:1-28.
11. Little A. Zelig and matching in R with an application to conflict and leader tenure. New York: New York University Department of Politics; 2009.
12. Sainani KL. Propensity scores: uses and limitations. *PMRJ*. 2012;4:693-697.
13. Cullen J, Rosner IL, Brand TC, et al. A Biopsy-based 17-gene genomic prostate score predicts recurrence after radical prostatectomy and adverse surgical pathology in a racially diverse population of men with clinically low-and intermediate-risk prostate cancer. *Eur Urol*. 2015;68:123-131.
14. Klein EA, Cooperberg MR, Magi-Galluzzi C, et al. A 17-gene assay to predict prostate cancer aggressiveness in the context of gleason grade heterogeneity, tumor multifocality, and biopsy undersampling. *Eur Urol*. 2014;66:550-560.
15. Van Den Eeden SK, Lu R, Zhang N, et al. A Biopsy-based 17-gene genomic prostate score as a predictor of metastases and prostate cancer death in surgically treated men with clinically localized disease. *Eur Urol*. 2018;73:129-138.
16. Canfield S, Kemeter MJ, Hornberger J, Febbo PG. Active surveillance use among a low-risk prostate cancer population in a large US payer system: 17-Gene genomic prostate score versus other risk stratification methods. *Rev Urol*. 2017;19:203-212.
17. Austin PC. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behav Res*. 2011;46:399-424.
18. Imai K, King G, Lau O. Logit: logistic regression for dichotomous dependent variables. *Zelig Everyone's Stat Softw*. 2008. http://gking.harvard.edu/zelig.
19. Rubin DB. The design versus the analysis of observational studies for causal effects: parallels with the design of randomized trials. *Stat Med*. 2007;26:20-36.
20. Zhang X, Faries DE, Li H, Stamey JD, Imbens GW. Addressing unmeasured confounding in comparative observational research. *Pharmacoepidemiol Drug Saf*. 2018;27(4):373-382. doi:10.1002/pds.4394.
21. Dall 'Era MA, Maddala T, Polychronopoulos L, Gallagher JR, Febbo PG, Denes BS. Utility of the Oncotype DX® Prostate Cancer Assay in Clinical Practice for Treatment Selection in Men Newly Diagnosed with Prostate Cancer: A Retrospective Chart Review Analysis. *Urol Pract*. 2015;2:343-348. doi:10.1016/j.urpr.2015.02.007.
22. Albala D, Kemeter MJ, Febbo PG, et al. Health economic impact and prospective clinical utility of Oncotype DX® Genomic Prostate Score. *Rev Urol*. 2016;18:123-132.
23. Eure G, Germany R, Given R, et al. Use of a 17-gene prognostic assay in contemporary urologic practice: results of an interim analysis in an observational cohort. *Urology*. 2017;107:67-75.
24. National Comprehensive Cancer Network. Prostate Cancer Version 3.2018. https://www.nccn.org/professionals/physician_gls/PDF/prostate.pdf. Accessed June 25, 2018.
25. Sanda MG, Cadeddu JA, Kirkby E, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO Guideline. Part I: risk stratification, shared decision making, and care options. *J Urol*. 2018;199:683-690.
26. Sanda MG, Cadeddu JA, Kirkby E, et al. Clinically localized prostate cancer: AUA/ASTRO/SUO Guideline. Part II: recommended approaches and details of specific care options. *J Urol*. 2018;199:990-997.
27. Cary C, Odisho A, Cooperberg MR. Variation in prostate cancer treatment associated with population density of the county of residence. *Prostate Cancer Prostatic Dis*. 2016;19:174-179.
28. Dall'Era MA. Patient and disease factors affecting the choice and adherence to active surveillance. *Curr Opin Urol*. 2015;25:272-276.
29. Pang K, Fitch M, Ouellet V, et al. Describing perspectives of health care professionals on active surveillance for the management of prostate cancer. *BMC Health Serv Res*. 2018;18:430.
30. Ehdaie B, Assel M, Benfante N, et al. A systematic approach to discussing active surveillance with patients with low-risk prostate cancer. *Eur Urol*. 2018;71:866-871.
31. Michiels-Corsten M, Donner-Banzhoff N. Beyond accuracy: hidden motives in diagnostic testing. *Fam Pract*. 2018;35:222-227.
32. Benishek LE, Weaver SJ, Newman-Toker DE. The cognitive psychology of diagnostic errors. *Scientific American Neurology*. doi:10.2310/7900.6288.