JAMA Guide to Statistics and Methods

Using Propensity Score Methods to Create Target Populations in Observational Clinical Research

Laine Thomas, PhD; Fan Li, PhD; Michael Pencina, PhD

In a propensity score–matched cohort study published in the March 12, 2019, issue of *JAMA*, Zeng et al[1] found that prescription tramadol was associated with significantly greater 1-year mortality compared with nonsteroidal anti-inflammatory alternatives in adults with osteoarthritis. At baseline, patients receiving tramadol were different than those who received other analgesics in terms of demographics, medical comorbidities, medications, and prior hospital resource utilization. Zeng et al[1] used propensity score matching in an effort to account for differences between groups.[2] This matched sample corresponds to a unique target population.

Explanation of the Concept

What Is a Target Population?

The *target population* is an intended group of patients characterized by inclusion and exclusion criteria and described by baseline characteristics to whom the average treatment effect applies. Two samples derived from a cohort with the same inclusion and exclusion criteria can have different characteristics that represent different target populations because of variation in sites and the way patients were enrolled in the study (**Box**).

Why Is the Target Population Important?

Understanding a study's target population is important for knowing how study results apply to certain types of patients. Observational studies performed with propensity score methods can change the target population by shifting the distribution of patient characteristics that contribute to analysis. Thus, propensity score analyses are used to reduce bias in the comparison between a population that received treatment and a control population and effectively mimic different randomized clinical trials (RCTs) that examine different target populations.[3]

Limitations and Alternatives to Developing Target Populations by Propensity Methods

A propensity score is the probability that a patient would receive the treatment of interest based on characteristics of the patient, treating clinician, and clinical environment.[2] Propensity methods can be conducted in many ways, with different approaches that create different target populations.[4] Two common methods are propensity matching and propensity score weighting (Box).

Propensity Matching

The most common propensity matching method, used by Zeng et al,[1] uses 1:1 nearest-neighbor matching, known as *greedy matching*, in which each individual who received treatment A is assessed sequentially to find the closest propensity score match among remaining individuals who received treatment B, usually within a prespecified bound on the closeness of the propensity score. Patients for whom no match exists within the bound are excluded. For example, Zeng et al[1] began with 16 372 eligible patients who received tramadol and 21 675 control individuals who received diclofenac. Prior to matching, patients who received tramadol and diclofenac had large differences in their characteristics, including mean age (72.1 vs 67.5 years) and number of visits to a general practitioner (GP) in the past 2 years (14.3 vs 9.7). After propensity score matching only 38% of eligible patients were retained in 6512 pairs.

To illustrate how propensity score matching alters the target population of an observational study, a simulation of 50 patients representative of those studied by Zeng et al[1] is presented in the **Figure**, A. After matching, patients receiving tramadol and diclofenac who were included in the study were similar but the matched sample differed from the original sample, with a narrower distribution of age and GP visits (Figure, B). Patients with successful matches tended to have middle-range propensity scores, indicating that such patients often receive either treatment in routine practice (Figure, B). The result is aligned with the concept of equipoise in clinical trials,[5] whereby patients for whom treatment decisions remain uncertain are enrolled.

Propensity Score Weighting

Inverse probability of treatment weighting (IPTW) is a form of propensity score weighting. Typically, patients who received treatment A are weighted by 1/propensity score and those who received treatment B are weighted by 1/(1−propensity score), where the propensity score is the probability of receiving treatment A in practice.[2] Outcomes are analyzed on the weighted sample. A patient who received treatment A received a larger weight if their probability of being treated with treatment A (propensity score) was small and a smaller weight if their probability of being treated with treatment A (propensity score) was large. Intuitively, the weights make up for underrepresentation or overrepresentation of certain types of individuals in each treatment group. In Figure, C, IPTW is applied to the simulated patients that resemble those in the Zeng et al study.[1] The relative contribution of each patient, after weighting, is shown by the size of each bubble. No patients are excluded from

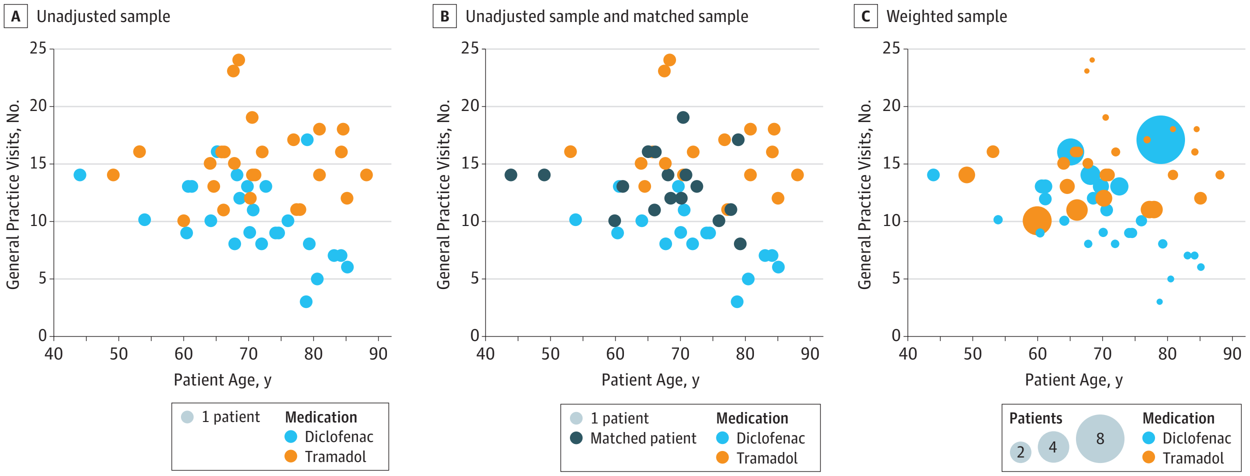
Box. Target Population, Propensity Score Weighting, and Propensity Score Matching

The target population, to whom the average treatment effect generalizes, depends not only on inclusion and exclusion criteria, but also on how patients enter the sample. In observational analyses this is further influenced by the choice of propensity score method.

Propensity score weighting, using the form of inverse probability of treatment weighting, applies to all sampled patients who received either comparator (A or B), but may excessively weight findings from patients who would typically receive only 1 of these options and are not good candidates for the other. Propensity score weighting addresses the question, "What if everyone in the sampled population received one treatment vs if everyone received the other treatment?"

Propensity score matching in observational studies can emulate the populations in randomized clinical trials by having relatively narrow patient characteristics that typify patients who are most eligible for either treatment being compared, but usually exclude a portion of the sample. Propensity score matching typically addresses the question, "What if everyone in the sampled population *who could be matched* received one treatment or the other treatment?"

Figure. Relative Contribution of 50 Simulated Patients With Different Ages and Number of General Practice Visits Within the Past Year



Simulated according to the distribution of the same variables in the Zeng et al study.[1] The bubble size reflects the relative contribution of each patient to analysis. A, Each patient represents only themselves. Patients receiving tramadol are older with more general practitioner visits. B, Each patient with a dark blue dot was successfully matched to similar patient receiving the other treatment and represents only themselves. Patients without a dark blue dot failed to find a match and are excluded, resulting in a smaller sample with a narrower range. C, After inverse probability of treatment weighting, some patients represent up to 8 other patients and others represent less than 1.

the population, but some represent up to 8 other patients while others represent less than 1 other patient.

The weighted sample mimics the potential population in an RCT drawn from the same target population as the original sample of treatment A plus B. In the study by Zeng et al[1] this would include all 38 407 patients. The clinical relevance of the corresponding RCT depends on whether the sampled population is representative of patients to whom results will be generalized. When the sample includes patients who already received treatment A or B with near certainty, they are up-weighted by IPTW to have larger influence on the results. In the simulated example, patients older than 75 years with more than 16 GP visits in the past 2 years almost always received tramadol. The single patient who received diclofenac and had these characteristics got a large weight (Figure, C). This up-weighting, particularly when it is extreme, can result in inflated variance.[6] There is substantial uncertainty about the average treatment effect for all patients when the target population is defined to include patients who nearly always receive 1 treatment. Other target populations can be studied by varying the definition of the weight.[3]

How Should the Propensity Analysis in Zeng et al Be Interpreted?

Zeng et al[1] created a matched sample of 13 024 patients, representing a relatively narrow target population. IPTW would have used all 38 407 patients and represented the population from which they were all sampled. In this study, another 66 261 patients received other analgesics and could have been candidates to receive tramadol or diclofenac, further broadening the potential population that might have been considered for an RCT. None of these populations is inherently better, but each can be characterized through a baseline characteristics table. After matching in the analysis by Zeng et al,[1] patients who received tramadol were younger (mean age, 70.3 vs 72.1) and had a lower mean (SD) number of GP visits (12.8 [9.7] vs 14.3 [12.8]). The reduction in SD suggests that the range of GP visits under evaluation narrowed, limiting the generalizability of the findings. To determine whether the target population that resulted from the application of propensity score methods is clinically relevant, investigators should undertake a comprehensive review of baseline characteristics, which goes beyond the simple comparison of means or medians.

ARTICLE INFORMATION

Author Affiliations: Biostatistics and Bioinformatics, Duke University, Durham, North Carolina (Thomas); Department of Statistical Science, Duke University, Durham, North Carolina (Li); Duke Clinical Research Institute, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina (Pencina).

Corresponding Author: Michael Pencina, PhD, Duke Clinical Research Institute, Department of Biostatistics and Bioinformatics, Duke University, 2400 Pratt St, Durham, NC 27705 (michal.pencina@duke.edu).

Section Editors: Roger J. Lewis, MD, PhD, Department of Emergency Medicine, Harbor-UCLA Medical Center and David Geffen School of Medicine at UCLA; and Edward H. Livingston, MD, Deputy Editor, *JAMA*.

Published Online: January 10, 2020. doi:10.1001/jama.2019.21558

Conflict of Interest Disclosures: Dr Pencina reported receiving grants from Sanofi/Regeneron and Amgen and personal fees from Merck and Boehringer Ingelheim outside the submitted work. No other disclosures were reported.

REFERENCES

1. Zeng C, Dubreuil M, LaRochelle MR, et al. Association of tramadol with all-cause mortality among patients with osteoarthritis. *JAMA*. 2019;321(10):969-982. doi:10.1001/jama.2019.1347

2. Haukoos JS, Lewis RJ. The propensity score. *JAMA*. 2015;314(15):1637-1638.

3. Li F, Morgan KL, Zaslavsky AM. Balancing covariates via propensity score weighting. *J Am Stat Assoc*. 2018;113(521):390-400.

4. Stuart EA. Matching methods for causal inference. *Stat Sci*. 2010;25(1):1-21.

5. London AJ. Equipoise in research: integrating ethics and science in human research. *JAMA*. 2017;317(5):525-526. doi:10.1001/jama.2017.0016

6. Li F, Thomas LE, Li F. Addressing extreme propensity scores via the overlap weights. *Am J Epidemiol*. 2019;188(1):250-257.