Title

Are Observational, Real-World Studies Suitable to Make Cancer Treatment Recommendations?

Journal

JAMA Network Open, 3(7)

ISSN

2574-3805

Authors

Banerjee, Rahul
Prasad, Vinay

Publication Date

2020-07-01

DOI

10.1001/jamanetworkopen.2020.12119

JAMA Network Open

Invited Commentary | Oncology

Are Observational, Real-World Studies Suitable to Make Cancer Treatment Recommendations?

Rahul Banerjee, MD; Vinay Prasad, MD, MPH

+ Related article

Author affiliations and article information are listed at the end of this article.

Kumar et al[1] added a new chapter to a decades-long debate about whether observational studies—nonrandomized comparative effectiveness research (CER)—can replace randomized clinical trials (RCTs) to assess the efficacy of therapies. Their work is timely. The 21st Century Cures Act has empowered the US Food and Drug Administration to use real-world evidence beyond controlled trials to support drug approvals.[2] Retrospective analyses of observational registries are used to justify a wider range of treatments, including the delivery of radiotherapy and surgery.[3] However, a key question remains: when a physician relies on an observational study to make a therapeutic recommendation, how often is that recommendation correct?

To answer this question, Kumar and colleagues[1] collected 141 RCTs used in national cancer treatment guidelines that cover 8 tumor types and make recommendations regarding the use of drugs, radiation, and surgical interventions. For each trial, the authors performed their own observational CER study, using the National Cancer Database (NCDB) registry, which captures more than 70% of all US cancer cases with data from more than 1500 contributing sites.[4] They created patient cohorts within the NCDB that match RCT study populations with respect to age, diagnosis, and specific therapies. The authors did not limit the number of patients in each NCDB cohort, and the median observational study cohort size was more than 15-fold that of the RCT.

Their findings are discouraging. Propensity-weighted hazard ratios for overall survival from CER-based analyses fell outside the 95% CIs of their RCT counterparts 36% of the time (with 64% falling within). Furthermore, observational studies led to a different inference regarding therapeutic efficacy 55% of the time (ie, point estimates that were either in a different direction, nonsignificant in CER vs significant in RCT or significant in CER but nonsignificant in RCT).

The findings of Kumar et al[1] differ substantially from 2 studies[5,6] published 20 years ago in the *New England Journal of Medicine.* These analyses investigated clinical questions for which both observational and randomized studies had been published.[5,6] Both studies found largely similar results and size of treatment effects between the 2 study designs, and readers of the *New England Journal of Medicine* received a double-barreled warning against discounting the results of large observational studies purely on the basis of lack of randomization. A larger study in 2001 by Ioannidis and colleagues[7] in *JAMA* also found an association between the results of nonrandomized and randomized studies. However, Ioannidis et al[7] found that of the 7 clinical questions (16% of their analysis) for which CER-based odds ratios fell outside the 95% CI of their RCT-based counterparts, RCT-based odds ratios were closer to 1 (suggesting a smaller treatment effect) in all but 1 of the cases.

More recently, Soni et al[8] analyzed matched observational CER-RCT pairs within oncology. One notable difference was that Soni et al[8] looked for studies where both observational and randomized trials were published for the same clinical question. Only 62% of CER-based studies demonstrated overall survival hazard ratios within the 95% CI of corresponding RCTs, a proportion similar to that found by Kumar et al.[1] Of 350 CER-RCT pairs analyzed by Soni et al,[8] only 40% showed concordance with regard to the presence or absence of statistical significance between arms. However, Soni et al[8] acknowledge that only a minority of the observational studies in their analysis used propensity-score matching or instrumental variables to adjust for possible confounders.

Can better methods improve the accuracy of nonrandomized comparative effectiveness research? Kumar et al[1] tackle this question by performing multivariable and propensity score

analyses. The authors account for a wide number of potential confounders in these analyses, including Charlson Comorbidity Index and median income. However, neither method was able to move the needle substantially on CER-RCT concordance. Kumar et al[1] found that 44% of unadjusted NCDB analyses yielded overall survival hazard ratios outside of the 95% CI for their RCT counterparts. Incorporation of multivariable regression or propensity score–weighting decreased this proportion modestly to 30% and 36%, respectively. The fact that propensity score–weighting seems inadequate is consistent with prior research.[9] Notably, to our knowledge, no study to date has used the most sophisticated observational method, a target trial simulation,[10] and we encourage future researchers to examine this question.

When it comes to therapeutic inferences—that is, the question of whether a systemic therapy, radiation course, or surgery will benefit or harm my patient—55% of NCDB analyses yielded discordant results from RCTs. This varied depending on the take-home message of the observational study (see eTable 2 in Kumar et al[1]). Propensity-weighted observational studies generally found a more aggressive or invasive treatment regimen was beneficial (55% of results). When this occurred, only 40% of the time was the finding validated in the RCT. Observational studies less often found that a less aggressive strategy was superior (11%) or failed to find a benefit (34%). When the less-aggressive treatment was found to be preferable, this finding was supported by the RCT 67% of the time. These data are invaluable for a clinician making a treatment decision based solely on observational CER data. An aggressive therapy that looks favorable in observational data turns out to be beneficial less than half the time. Confounding by indication, a type of selection bias that means that we reserve aggressive therapies for the healthiest candidates, likely explains this finding. Aggressive therapies often look favorable in observational data not because they usually work, but because we preferentially deploy them in patients who are healthier than average.

Encouragingly, although 9% of CER-RCT pairs analyzed by Soni et al[8] demonstrated what might be called extreme discordance (with the 2 studies showing statistically significant differences in opposite directions), Kumar et al[1] report this occurs 5% of the time in propensity-weighted analyses. However, neither group was able to identify any trial or disease characteristics that were predictive of concordance between CER-based and RCT-based research.[1,8]

Limitations to Kumar et al's analysis[1] include missing data elements (eg, patient performance status) that are not captured in the NCDB registry but may be available in other data sets such as the electronic medical record, a data source leveraged by companies such as Flatiron. Adjusting for this variable may improve concordance. Of course, neither propensity score nor instrumental variable approaches can correct for the risk of confounding from variables that were not measured or those that researchers do not completely understand. Whether the analytic approach of target trials[10] is able to overcome the deficiencies here remains a hypothesis to be tested.

The major strength of the current work is that although some discrepancies between CER-based and RCT-based conclusions can be attributed to chance or larger sample sizes, the overall findings by Kumar et al[1] make it clear that even the most well-designed CER study will not necessarily deliver results similar to RCT-based research. This finding has immediate implications for the US Food and Drug Administration and cancer practitioners.

Is there a role for retrospective CER studies in oncology in light of these findings? We believe the answer is yes. Observational studies may clarify issues of prognosis, patterns of real-world usage, rare adverse events, and glaring disparities in cancer care delivery. However, when it comes to establishing the fundamental efficacy of therapeutic interventions, caution is warranted, and propensity score–based methods are not a panacea. Ultimately, adding the real-world rhetorical flourish to the title of a CER abstract based on its patient population understates the problematic elements of observational studies: unmeasured confounders, problems defining time 0, and selection bias (confounding by indication).

The holy grail of medicine is to develop a system where we can make reliable inferences regarding the effectiveness of therapies as fast as possible, as cheaply as possible, with the least

number of patients exposed to less effective regimens. Although many believe observational, real-world data will someday fill this niche, the work of Kumar and colleagues[1] reminds us that for the time being randomization remains the reference standard in cancer research.

ARTICLE INFORMATION

Published: July 30, 2020. doi:10.1001/jamanetworkopen.2020.12119

Correction: This article was corrected on November 20, 2020, to fix 6 places in the text where the terms *propensity score–matched* and *propensity score–matching* were used incorrectly. These terms have been replaced by *propensity-weighted* and *propensity score–weighting*, respectively.

Corresponding Author: Vinay Prasad, MD, MPH, Department of Epidemiology and Biostatistics, University of California San Francisco, 550 16th St, San Francisco, CA 94158 (vinayak.prasad@ucsf.edu).

Author Affiliations: Division of Hematology/Oncology, Department of Medicine, University of California San Francisco, San Francisco (Banerjee, Prasad); Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco (Prasad).

Conflict of Interest Disclosures: Dr Prasad reported receiving research funding from Arnold Ventures; royalties from Johns Hopkins Press; honoraria from Medscape, universities, medical centers, nonprofit organizations, and professional societies (for grand rounds and lectures); consulting fees from UnitedHealthcare; and speaking fees from Evicore. Dr Prasad's podcast *Plenary Session* has Patreon backers. No other disclosures were reported.

REFERENCES

1. Kumar A, Guss ZD, Courtney PT, et al. Evaluation of the use of cancer registry data for comparative effectiveness research. *JAMA Netw Open*. 2020;3(7):e2011985. doi:10.1001/jamanetworkopen.2020.11985

2. Raphael MJ, Gyawali B, Booth CM. Real-world evidence and regulatory drug approval. *Nat Rev Clin Oncol*. 2020;17(5):271-272. doi:10.1038/s41571-020-0345-7

3. Visvanathan K, Levit LA, Raghavan D, et al. Untapped potential of observational research to inform clinical decision making: American Society of Clinical Oncology research statement. *J Clin Oncol*. 2017;35(16):1845-1854. doi:10.1200/JCO.2017.72.6414

4. Mallin K, Browner A, Palis B, et al. Incident cases captured in the National Cancer Database compared with those in U.S. population based central cancer registries in 2012-2014. *Ann Surg Oncol*. 2019;26(6):1604-1612. doi:10.1245/s10434-019-07213-1

5. Benson K, Hartz AJ. A comparison of observational studies and randomized, controlled trials. *N Engl J Med*. 2000;342(25):1878-1886. doi:10.1056/NEJM200006223422506

6. Concato J, Shah N, Horwitz RI. Randomized, controlled trials, observational studies, and the hierarchy of research designs. *N Engl J Med*. 2000;342(25):1887-1892. doi:10.1056/NEJM200006223422507

7. Ioannidis JP, Haidich AB, Pappa M, et al. Comparison of evidence of treatment effects in randomized and nonrandomized studies. *JAMA*. 2001;286(7):821-830. doi:10.1001/jama.286.7.821

8. Soni PD, Hartman HE, Dess RT, et al. Comparison of population-based observational studies with randomized trials in oncology. *J Clin Oncol*. 2019;37(14):1209-1216. doi:10.1200/JCO.18.01074

9. Dahabreh IJ, Kent DM. Can the learning health care system be educated with observational data? *JAMA*. 2014;312(2):129-130. doi:10.1001/jama.2014.4364

10. Emilsson L, García-Albéniz X, Logan RW, Caniglia EC, Kalager M, Hernán MA. Examining bias in studies of statin treatment and survival in patients with cancer. *JAMA Oncol*. 2018;4(1):63-70. doi:10.1001/jamaoncol.2017.2752