Editorial

Lost in Translation?

Evaluating the Generalizability of Randomized Controlled Trial Findings to Broad Clinical Practice

Mark A. Clapp, MD, MPH

Randomized controlled trials (RCT) are often considered the gold standard for establishing causality between an intervention and an outcome in clinical medicine.[1] Ideally, through the process of randomization, investigators can control for unmeasured confounders (ie, factors affecting the relationship of interest that may not be known, cannot be seen, or are difficult to measure) such that the only difference between the groups is the exposure of interest. In a well-designed, well-conducted RCT, this fundamental principle gives us (clinicians, patients, researchers, policymakers) a degree of certainty in our understanding of the intervention's effects on the outcome.

The ability to ascertain causality through an RCT is not without trade-offs, many of which may affect the trial's generalizability.[2,3] For one, conventional RCTs often are designed to investigate the effect of one intervention on a primary outcome in a single-study population. This study population, defined by preset inclusion criteria, is shaped by an expected effect of the intervention and the study's feasibility (eg, number of patients expected, planned length of recruitment, resources required, and cost). Direct conclusions from the trial can be applied to the same intervention in the same study population with reasonable certainty; the external validity of the results is not guaranteed when extrapolating the findings to similar, but not identical, interventions or to other patient populations. Even in well-designed trials, RCTs that randomize individual participants are required to inform them of the study protocol and then obtain their voluntary consent before being included. Participants who agree to participate may be, and often are, different from the general population. Thus, it is important to question how the findings from an RCT can be applied to and translate into every day clinical practice.

Studies that evaluate the translation of RCT results into the "real world" (not the controlled environment of a trial) are valuable in shaping our understanding of a study's generalizability and providing us with additional evidence to better inform our clinical decision making and patient counseling. As with an RCT, it is similarly important to critically evaluate these observational studies, which may be more prone to bias.

This month's issue of the *Obstetrics & Gynecology* features two studies (see pages 239 and 242) examining the effects of the ARRIVE (A Randomized Trial of Induction Versus Expectant Management) trial on obstetric practice and outcomes.[4–6] Published in 2018, the ARRIVE trial randomized more than 6,000 nulliparous patients to induction of labor between 39 0/7 and 39 4/7 weeks of gestation compared with expectant management and showed that induction of labor did not increase adverse neonatal outcomes (primary outcome). As secondary outcomes,

See related articles on pages 239 and 242.

Mark A. Clapp is from the Department of Obstetrics & Gynecology at the Massachusetts General Hospital, Boston, Massachusetts; mark.clapp@mgh.harvard.edu.

Financial Disclosure
The author did not report any potential conflicts of interest.

ISSN: 0029-7844/23

individuals randomized to induction had 16% lower risk of cesarean delivery and 36% lower risk of hypertensive disorders of pregnancy (HDP) than those in the expectant management group. To achieve the planned enrollment, more than 50,000 patients had to be screened for their eligibility and interest in participation, which has raised some concerns about the external validity of the trial.[7]

In the study by Futterman et al,[5] the authors examine for the associations of induction on HDP in nulliparous patients before and after the ARRIVE trial using U.S. vital statistics data. They demonstrate that the rates of HDP among individuals undergoing 39-week inductions decreased after the trial. As the authors note, their findings likely represent an increase in non–medically indicated inductions at 39 weeks of gestation after the trial, thereby lowering the relative number of individuals being induced for HDP at the same gestation. The study design and the limited granularity of birth certificate data (eg, unknown timing of HDP in relation to induction) prevents us from inferring causality about whether the increased use of induction of labor resulted in less HDP. In contrast, the authors report that HDP rates among all nulliparous individuals delivering at 39 weeks of gestation or later (not just inductions at 39 weeks) were higher than expected in the post–ARRIVE period.[8] It remains possible that factors other than the changing practice of induction may have been influencing HDP rates during the study period (eg, the coronavirus disease 2019 [COVID-19] pandemic), thereby obscuring the true association with induction of labor after the ARRIVE trial.[9]

Nethery et al[6] also examine for population effects of the ARRIVE trial in 13 hospitals participating in a perinatal quality collaborative. In contrast to the Futterman et al study, which used a nationwide sample, there were many fewer patients included; however, the Nethery et al study is strengthened by the use of granular, chart-abstracted clinical data and its quasi-experimental study design. Quasi-experimental designs, like the interrupted time series model used in this study, conceptualize the introduction of a new policy or practice as a natural experiment and can account for unmeasured confounders that may bias traditional observational study designs. Nethery et al report that the ARRIVE trial did change clinical practice by increasing inductions in nulliparous patients, including an increase in elective inductions from 3.6% to 10.8%. However, there were no changes in the rates of cesarean delivery or HDP. The authors note several possibilities for their findings compared with the RCT results, including the relatively low rate of elective inductions overall, the increase in elective inductions among individuals outside the original trial's eligibility criteria, and the differences in baseline prevalence of the outcomes between the study sites and this patient population (eg, observed cesarean delivery rate of 27% vs 22% in the study).

These studies attempt to shed light on the generalizability of the ARRIVE trial results using different sources of population data and should prompt clinicians to consider how they routinely counsel their patients on the benefits of non–medically indicated induction of labor. To date, the best evidence of benefit (reduction in risk of cesarean delivery and HDP) for these inductions remains in nulliparous patients between 39 0/7 and 39 4/7 weeks of gestation. Although the Futterman et al and Nethery et al studies do not refute the RCT's findings, they do not convincingly demonstrate the same benefits of induction outside of the study population. Neither study examines other considerations that may go into the decision making and counseling around induction of labor, such as the patient's experience or preference for spontaneous labor–topics that remain important to understand and are challenging to ascertain from population-based data.

By design, conventional RCTs often investigate a narrowly focused research question in a specific population. Post–dissemination and implementation studies, especially those that use robust designs and minimize bias in observational data, are one tool that can give us insight into the generalizability of a trial's results when applied to populations both similar to the that in the original study and those in which the findings have been extrapolated. Studies such as these by Futterman et al and Nethery et al have yet to arrive at a consensus on whether and how the benefits of non–medically indicated induction of labor translate more broadly to clinical practice.

REFERENCES

1. Hariton E, Locascio JJ. Randomised controlled trials–the gold standard for effectiveness research: study design: randomised controlled trials. BJOG 2018;125:1716. doi: 10.1111/1471-0528
2. Malmivaara A. Generalizability of findings from randomized controlled trials is limited in the leading general medical journals. J Clin Epidemiol 2019;107:36–41. doi: 10.1016/j.jclinepi.2018.11.014
3. Averitt AJ, Weng C, Ryan P, Perotte A. Translating evidence into practice: eligibility criteria fail to eliminate clinically significant differences between real-world and study populations. NPJ Digit Med 2020;3:1–10. doi: 10.1038/s41746-020-0277-8
4. Grobman WA, Rice MM, Reddy UM, Tita ATN, Silver RM, Mallett G, et al. Labor induction versus expectant management

in low-risk nulliparous women. N Engl J Med 2018;379:513–23. doi: 10.1056/NEJMoa1800566

5. Futterman ID, Gilroy LC, Silver M, Minkoff H, Al-Kouatly HB, McLaren RA Jr. Changes in rates of hypertensive disorders of pregnancy among nulliparous patients after the ARRIVE (A Randomized Trial of Induction Versus Expectant Management) trial. Obstet Gynecol 2023;142:239–41. doi: 10.1097/AOG.0000000000005239
6. Nethery E, Levy B, McLean K, Sitcov K, Souter VL. Trial on elective induction and obstetric outcomes in term nulliparous patients. Obstet Gynecol 2023;142:242–50. doi: 10.1097/AOG.0000000000005217
7. Carmichael SL, Snowden JM. The ARRIVE trial: interpretation from an epidemiologic perspective. J Midwifery Womens Health 2019;64:657–63. doi: 10.1111/jmwh.12996
8. Gilroy LC, Al-Kouatly HB, Minkoff HL, McLaren RA. Changes in obstetrical practices and pregnancy outcomes following the ARRIVE trial. Am J Obstet Gynecol 2022;226:716.e1–12. doi: 10.1016/j.ajog.2022.02.003
9. Papageorghiou AT, Deruelle P, Gunier RB, Rauch S, García-May PK, Mhatre M, et al. Preeclampsia and COVID-19: results from the INTERCOVID prospective longitudinal study. Am J Obstet Gynecol 2021;225:289.e1–17. doi: 10.1016/j.ajog.2021.05.014 9