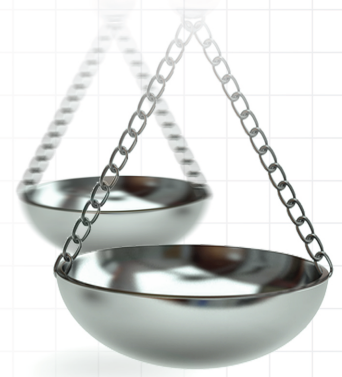




# R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 17

Journal of **Comparative Effectiveness Research**

Paul Arora<sup>1,2</sup> & Sreeram V Ramagopalan<sup>\*,3</sup>

<sup>1</sup>Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, Ontario, M5T 3M7, Canada

<sup>2</sup>Inka Health, Schwartz Reisman Innovation Campus, University of Toronto, Toronto, Ontario, M5G 0C6, Canada

<sup>3</sup>Centre for Pharmaceutical Medicine Research, King's College London, London, SE1 9NH, UK

\*Author for correspondence: [sreeram@ramagopalan.net](mailto:sreeram@ramagopalan.net)

In this update, we discuss a position statement from the National Institute of Health and Care Excellence (NICE) on the use of artificial intelligence for evidence generation and publications reviewing the use of real-world data as external control arms. Finally, we discuss a number of recent studies investigating the real-world effectiveness of glucagon-like peptide-1 receptor agonists and whether these studies are informative for reimbursement decision making.

First draft submitted: 23 October 2024; Accepted for publication: 4 November 2024; Published online: 27 November 2024

**Keywords:** artificial intelligence • evidence generation • external control arms • GLP-1RA • health technology assessment • natural language processing • NICE • obesity • real-world data • real-world evidence

The National Institute of Health and Care Excellence (NICE) recently released a position statement on the use of artificial intelligence (AI) methods in evidence generation and reporting for health technology assessment (HTA) [1]. The statement aims to provide clarity on how NICE will consider AI-generated evidence. Potential AI applications for evidence generation are outlined: for systematic reviews, AI could automate search strategies, screening, data extraction and synthesis. For clinical (trial) evidence, AI may be used to optimize trial design, adjust for covariates and generate synthetic data. In real-world data (RWD) analysis, natural language processing could be used to analyze unstructured data, or AI could assist with multimodal data integration. For cost-effectiveness evidence, AI might aid in health economic model conceptualisation and construction. Recent demonstrations have shown the potential of this approach [2,3]. NICE emphasizes that AI methods should only be used when there is demonstrable value, carefully balancing potential benefits against risks such as algorithmic bias and reduced transparency. Submitting organisations must clearly justify AI use and outline assumptions, with more explainable, robust methods presented in the first instance as compared with less transparent approaches. Justification of approaches used and assumptions made can use existing tools such as ISPOR's PALISADE [4]. Early engagement with NICE is encouraged when considering AI methods, and use should align with UK government AI regulation frameworks and ethical standards. The position statement stresses that AI should augment rather than replace human involvement, and confidence in the methods used needs to be built with careful technical and external validation required. The use of AI for estimating comparative treatment effects is classed as a 'high-risk' application, and here multiple sensitivity analyses need to be performed and results triangulated with clinical evidence. NICE plans to monitor AI use in evaluations and the position may be updated as significant new evidence emerges. This position statement broadly aligns with guidance from other bodies discussed previously in this series (e.g., ISPOR and ESMO) [3]. It reflects the growing recognition that AI will become increasingly important in evidence generation, but transparency, rigor and human oversight is needed. As AI methods evolve, manufacturers should stay informed of best practices to ensure high-quality evidence generation that will meet HTA standards.

Two recent articles have systematically examined the use of RWD as external controls. The first article by Hermans *et al.*, including FDA co-authors, conducted a systematic review of 32 studies comparing uncontrolled trials to

external RWD controls in hematological cancers [5]. They found that generally published RWD external controls had limitations. Issues identified included: eligibility criteria used in the trial were rarely applied fully to the RWD external control, differences in end point definitions between trial and RWD, potentially inappropriate handling of missing data in RWD, a lack of appropriate statistical methods to account for bias and confounding and finally limited pre-specification of statistical analysis plans (SAP). The authors conclude that the best practices outlined by regulator guidance for external controls [3,6], are not widely implemented. They recommend improvements like prospective RWD collection, applying complete trial eligibility criteria to RWD cohorts, defining SAPs prior to data analysis (preferably after discussion with regulators) and standardizing end points. The second article by Hogervorst *et al.* conducted a systematic literature review of analytical methods for comparing uncontrolled trials to external RWD controls, and compared findings to regulatory/HTA guidelines and reports [7]. The authors identified many advanced methods in the literature for comparing trials with RWD, such as g-computation for causal inference. Regulatory/HTA guidelines were generally aligned with the literature in recommending more sophisticated methods for causal questions (although tended not to cite the most advanced methods). Guidelines however lacked clear preferences for one methodology over another. Despite the methods publications in the literature and recommendations in guidelines, regulatory/HTA submissions often used simpler methods like naive comparisons or basic regression, especially with aggregate data. The authors propose 12 recommendations to improve trial–RWD comparisons. Some of these recommendations align with Hermans *et al.*, as well as other recent publications [8], for example, to use the target trial emulation approach and pre-specify analytical plans. Hogervorst *et al.* also suggest there is a need for guidance on when to use specific analytical methods and how a discussion is needed with stakeholders, including regulators and HTA bodies around what the optimal balance is between leveraging advanced methodologies and ensuring their interpretability and transparency by decision makers. These two articles align with recent findings showing that no RWD submission to a HTA agency to date has employed target trial emulation [9], and as such RWD submissions are falling short of best practice being called out in regulatory and HTA guidance. This may be because RWE guidance has only been around for about two years, perhaps meaning that manufacturers have not had enough time to employ the approach for submissions. There is perhaps some education needed for stakeholders to get more familiar with some of the advanced causal inference methods which the work of Hogervorst *et al.* and others will help with. RWD external control submissions should be tracked in the future to see if the best practice highlighted in guidance is being implemented and to understand reasons if not.

Obesity is one of the most pressing challenges facing healthcare systems and wider society today [10]. Recently, the glucagon-like peptide-1 receptor agonist (GLP-1RA) class of drugs has been shown to be an effective treatment for the disease. Not only do these drugs lead to weight loss, they may also have other clinical benefits which are being tested in clinical trial programs [10]. As trials tend to have relatively short follow-up (up to three years for example), any longer-term adverse effects of these treatments are also unknown. To better understand the efficacy and safety of these treatments, RWD may be informative. As such, a number of recent studies have attempted to look at efficacy and safety of GLP-1RAs. However, are these studies useful for reimbursement decision making (a thorny issue for obesity treatments given the potentially large budget impact [10])? Recent RWD studies have shown that the GLP-1RA semaglutide has been shown to be associated with lower weight loss as compared with another GLP-1RA, tirzepatide [11]; that semaglutide is associated with an increased risk of nonarteritic anterior ischemic optic neuropathy (NAION) [12]; and that the use of GLP-1RAs is associated with a reduced risk of cancer [13]. Historically, RWD studies have a poor reputation for causal inference, as described above – some RWD results have significantly differed from randomized-controlled trials (RCTs) results due to bias and confounding [14]. This has been a key reason why decision makers have been reluctant to accept RWD as a measure of treatment effectiveness. However, this poor reputation is largely attributable to poor choice of analytic methods rather than the RWD itself [15]. An indication that residual confounding is present in two of the above studies can be seen by the almost immediate separation of the Kaplan–Meier curve for the comparison of semaglutide with non-GLP-1RA treatments for the risk of NAION and for the cumulative incidence graph comparing GLP-1RAs and insulin for the risk of cancer [16]. It is biologically implausible that GLP-1RA treatments work immediately for these outcomes, and therefore the populations being compared are likely different in unmeasured ways. For example, GLP-1RA drugs are generally used for more advanced cases of diabetes where NAION would be more likely, and the analysis did not adjust for this. The results from these studies therefore need to be taken with caution. Rodriguez *et al.* compared weight loss between semaglutide and tirzepatide disregarding the fact that these medicines have different doses, and for semaglutide doses approved for different indications [11]. RCTs have demonstrated that weight loss is dose

dependent [17], and the approved semaglutide dose for the control of diabetes is lower than that for weight loss. While some may be interested in this finding, what is driving the difference in weight loss between treatments remains unclear (e.g., is it just because groups taking different doses for different purposes are being compared?). More interest may come from answering a causal question [18], for example, at a specific dose indicated for weight loss, and when these doses are adhered to (important given supply issues for GLP-1RAs), what is the comparative effectiveness of tirzepatide and semaglutide? Causal inference from RWD is possible [18]. As noted above, regulators and HTA agencies have endorsed the target trial emulation framework to enable this, which includes defining a causal question and being thoughtful in study design so bias is minimized [19]. They have also advocated for quantitative bias analysis to evaluate potential residual confounding [19]. None of the GLP-1RA RWD studies discussed employed these techniques. A dose–response analysis within the framework of target trial emulation would have offered insights into how the drugs compare when adjusted for appropriate dose levels and indications. The reimbursement of GLP-1RA treatments for weight loss has been challenging worldwide [10], and recently, the UK government has announced Eli-Lilly will be conducting a real-world study to assess the impact of tirzepatide on economic productivity [20]. It is important best practice is followed to inform decision makers on the true value of GLP-1RAs.

### Financial disclosure

SV Ramagopalan has received an honorarium from Becaris Publishing for the contribution of this work. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

### Competing interest disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties.

### Writing disclosure

No writing assistance was utilized in the production of this manuscript.

### Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <https://creativecommons.org/licenses/by-nc-nd/4.0/>

## References

1. NICE. Use of AI in evidence generation: NICE position statement. Available at: <https://www.nice.org.uk/about/what-we-do/our-research-work/use-of-ai-in-evidence-generation--nice-position-statement>
2. Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial intelligence to automate health economic modelling: a case study to evaluate the potential application of large language models. *Pharmacoecon. Open.* 8(2), 191–203 (2024).
3. Castanon A, Tsvetanova A, Ramagopalan SV. RWE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 16. *J. Comp. Eff. Res.* 13(8), e240095 (2024).
4. Padula WV, Kreif N, Vanness DJ *et al.* Machine learning methods in health economics and outcomes research – The PALISADE Checklist: A Good Practices Report of an ISPOR Task Force. *Value Health.* 25(7), 1063–1080 (2022).
5. Hermans SJF, van der Maas NG, van Norden Y *et al.* Externally controlled studies using real-world data in patients with hematological cancers: a systematic review. *JAMA Oncol.* 10, 1426–1436 (2024).
6. Bray B, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 11. *J. Comp. Eff. Res.* 12(5), e230008 (2023).
7. Hogervorst MA, Soman KV, Gardarsdottir H, Goettsch WG, Bloem LT. Analytical methods for comparing uncontrolled trials with external controls from real-world data: a systematic literature review and comparison with European regulatory and health technology assessment practice. *Value Health* S1098–3015(24), 02842–0 (2024).
8. Castanon A, Bray BD, Ramagopalan SV. R WE ready for reimbursement? A round up of developments in real-world evidence relating to health technology assessment: part 15. *J. Comp. Eff. Res.* 13(5), e240033 (2024).
9. Castanon A, Duffield S, Ramagopalan S, Reynolds R. Why is target trial emulation not being used in health technology assessment real-world data submissions? *J. Comp. Eff. Res.* 13(8), e240091 (2024).

10. Collins E, Beattie A, Ramagopalan SV, Pearson-Stuttard J. First in class, best in class or a wild card: who will dominate the anti-obesity medication market? *J. Comp. Eff. Res.* 13(7), e240044 (2024).
11. Rodriguez PJ, Goodwin Cartwright BM, Gratzl S *et al.* Semaglutide vs tirzepatide for weight loss in adults with overweight or obesity. *JAMA Intern. Med.* 184, 1056–1064 (2024).
12. Hathaway JT, Shah MP, Hathaway DB *et al.* Risk of nonarteritic anterior ischemic optic neuropathy in patients prescribed semaglutide. *JAMA Ophthalmol.* 142, 732–739 (2024).
13. Wang L, Xu R, Kaelber DC, Berger NA. Glucagon-like peptide 1 receptor agonists and 13 obesity-associated cancers in patients with Type 2 diabetes. *JAMA Netw. Open.* 7(7), e2421305 (2024).
14. Ramagopalan SV, Simpson A, Sammon C. Can real-world data really replace randomized clinical trials? *BMC Med.* 18(1), 13 (2020).
15. Hernán MA, Sauer BC, Hernández-Díaz S, Platt R, Shrier I. Specifying a target trial prevents immortal time bias and other self-inflicted injuries in observational analyses. *J. Clin. Epidemiol.* 79, 70–75 (2016).
16. Mohyuddin GR, Prasad V. Detecting selection bias in observational studies – when interventions work too fast. *JAMA Intern. Med.* 183(9), 897–898 (2023).
17. Jastreboff AM, Aronne LJ, Ahmad NN *et al.* Tirzepatide once weekly for the treatment of obesity. *N. Engl. J. Med.* 387(3), 205–216 (2022).
18. Hernán MA, Wang W, Leaf DE. Target trial emulation: a framework for causal inference from observational data. *JAMA* 328, 2446–2447 (2022).
19. Chen S, Tikhonovsky N, Dhanji N, Ramagopalan S. Emulating trials and quantifying bias: the convergence of health technology assessment agency real-world evidence guidance. *Value Health.* 27, 265–267 (2024).
20. The University of Manchester. New study to deepen understanding of a weight loss medication. Available at: <https://www.manchester.ac.uk/about/news/new-study-to-deepen-understanding-of-a-weight-loss-medication/>