Article:
Dunger-Baldauf, C., Jen, M.-H., Ren, S. et al. (2025) A landscape assessment of key evidence needs in study design and statistical methodologies for HTA submissions. Statistics in Biopharmaceutical Research. ISSN: 1946-6315

https://doi.org/10.1080/19466315.2025.2581122

Statistics in Biopharmaceutical Research

ISSN: 1946-6315 (Online) Journal homepage: www.tandfonline.com/journals/usbr20

A landscape assessment of key evidence needs in study design and statistical methodologies for HTA submissions

Cornelia Dunger-Baldauf, Min-Hua Jen, Shijie Ren, Xiang Zhang, Shahrul Mt-Isa, Tae Hyun Jung, Valentina Bayer, Liang Chen, Dai Feng, Yingyi Liu, Julia Ma & Weili He

To cite this article: Cornelia Dunger-Baldauf, Min-Hua Jen, Shijie Ren, Xiang Zhang, Shahrul Mt-Isa, Tae Hyun Jung, Valentina Bayer, Liang Chen, Dai Feng, Yingyi Liu, Julia Ma & Weili He (10 Nov 2025): A landscape assessment of key evidence needs in study design and statistical methodologies for HTA submissions, Statistics in Biopharmaceutical Research, DOI: 10.1080/19466315.2025.2581122

To link to this article: https://doi.org/10.1080/19466315.2025.2581122

A landscape assessment of key evidence needs in study design and statistical methodologies for HTA submissions

Cornelia Dunger-Baldauf[1], Min-Hua Jen[2], Shijie Ren[3], Xiang Zhang[4], Shahrul Mt-Isa[5], Tae Hyun Jung[6], Valentina Bayer[7], Liang Chen[8], Dai Feng[9], Yingyi Liu[9], Julia Ma[10], and Weili He[10,*]

[1]Advanced Methodology and Data Science, Novartis, Basel, Switzerland

[2]Global Statistical Sciences, Eli Lilly and the company, Bracknell, UK

[3]University of Sheffield, Sheffield, South Yorkshire, UK

[4]Medical Affairs and HTA Statistics, Quantitative Clinical Sciences and Reporting, CSL Behring, King of Prussia, Pennsylvania, USA

[5]Biostatistics and Research Decision Sciences (BARDS) HTA Statistics, MSD, Zurich, Switzerland

[6]Office of Biostatistics, Center for Drug Evaluation and Research, FDA, Silver Spring, Maryland, USA

[7]Global Biostatistics & Data Sciences, Boehringer Ingelheim Pharmaceuticals, Inc., Ridgefield, CT, USA

[8]Real World Statistics & Analytics, Biometrics & RWE, Vertex, Boston, Massachusetts, USA

[9]MA&HTA Statistics, AbbVie, North Chicago, Illinois, USA

[10]MA&HTA Statistics, AbbVie, Florham Park, New Jersey, USA

*Correspondence to: Dr. Weili He, AbbVie, Florham Park, New Jersey. Email: weili.he@abbvie.com

Abstract

Health Technology Assessment (HTA) evaluations play a crucial role in informing decisions related to the adoption, reimbursement, and utilization of healthcare technologies. To ensure robust and reliable outcomes, HTA requires a diverse range of evidence, which may vary depending on the specific technology under evaluation, the questions to be answered, and the available data sources. It is imperative to design and conduct studies that generate high-quality and pertinent evidence to facilitate effective HTA evaluations. Furthermore, sophisticated and appropriate statistical methodologies are often necessary to analyze and interpret the collected data in HTA assessments. Recognizing the lack of discussion and best practice recommendations to fulfill the HTA needs, the American Statistical Association (ASA) Biopharmaceutical Section (BIOP) Health Technology Assessment (HTA) Scientific Working Group (SWG) has undertaken an initiative to assess the HTA landscape in major global markets. We aim to offer strategic considerations for evidence planning related to HTA, alongside specific statistical methodologies commonly used in delivering clinical evidence and demonstrating value. Our targeted audience includes statisticians working in clinical development who may not be familiar with the intricacies and specific needs of HTA. This paper focuses specifically on study designs and statistical methods. This paper sheds light on the challenges that persist in study design and analytic approaches concerning HTA evidence requirements and discusses potential opportunities and mitigations. By bridging the knowledge gap in HTA needs and offering practical guidance on study designs and statistical methods, this research advances the field of statistics within HTA.

Key words: Health technology assessment, study design, statistical methodologies, reimbursement

1. Introduction and the importance of evidence/data in Health Technology Assessment

Health Technology Assessment (HTA) is a multidisciplinary process that uses transparent and principled methods to evaluate the value and impact of health technology at various stages throughout its lifecycle. The purpose is to inform decision-making to promote an equitable, efficient, and high-quality health system. In recent years, notable advances and innovations in HTA have taken place. For example, the new EU HTA regulation, known as Regulation (EU) 2021/2282 (EU, 2024), was adopted in December 2021, and it establishes a framework for HTA cooperation among the EU member states. In the new process by the EU HTA regulation, the HTA dossier addresses the evidence needs of medicine regulators alongside those of HTA bodies to enable a joint clinical assessment (JCA) by the European Medicines Agency (EMA) and EU HTA. Generating a single set of evidence for these stakeholders and EU healthcare payers should make decisions on pricing and reimbursement quicker and easier (EMA, 2019). The companion paper (Jen et al., 2025) provides more details on the HTA framework and EU HTA regulation.

The new process, alongside a growing trend to consider HTA requirements prior to marketing authorization (Hampson et al., 2022), and the application of advanced statistical methodologies to meet these requirements, will necessitate biopharmaceutical industry statisticians not only understand these advanced methodologies but also acquire knowledge of the new process. Recognizing the current knowledge gap, the American Statistical Association (ASA) Biopharmaceutical Section (BIOP) Health Technology Assessment (HTA) Scientific Working Group (SWG) has undertaken efforts to assess the HTA landscape in major markets worldwide focusing on statistical perspectives. The goal of the SWG is to introduce the conceptual framework of HTA evaluations, review key HTA guidance documents from major markets, discuss the evidence needs of major health authorities, address the challenges and opportunities in this field, and provide guidance on how statisticians can contribute to the field in a precompetitive space. The SWG has established two workstreams for this effort. Workstream 1 (WS1) focuses on understanding the current landscape of HTA submission requirements and identifying key challenges. Workstream 2 (WS2) conducts a landscape assessment of study design and statistical

methodologies for HTA submissions, presented here. The outputs from the two workstreams are presented in two papers. These provide complementary views of the current state of HTA evaluations and offer insights into addressing the challenges faced by statisticians in this field.

The value of a new technology (a healthcare intervention such as drug, device, or service) is assessed within the context of the healthcare system (HCS) in the country or region. This assessment considers the available technology options and informs decision making on the adoption, pricing and reimbursement of the new technology. Factors considered include comparative effectiveness, safety, cost-effectiveness, ease of use for a user or in the HCS, and long-term perspectives. Subgroups for whom the new technology might be particularly effective or represent high value for money will be accounted for as well as prevalence of the disease, disease burden and patient preferences. In addition, evidence needs may change over time, for example if further technologies become available while evaluations are ongoing.

These multifaceted evidence needs will rarely be coverable by a single source. Usually, evidence from several sources (for example randomized controlled trials (RCTs), observational studies, or clinical practice) needs to be synthesized. Specific study designs which combine evidence from several sources can be considered, or evidence synthesis methods allowing for formal statistical synthesis of results from existing studies can be used. When outcomes, populations, time spans or other features are not consistent across sources, statistical methods of population adjustment, extrapolation, mapping, or data transportability methods may be required. A comprehensive evidence base will be fundamental to meet the payers' needs, ultimately facilitating decision-making within the HCS and improving patients' outcomes.

Study designs which aim to fill evidence gaps not covered in the submission to obtain marketing authorization are discussed in Section 2. Recent developments in statistical methodology for the areas of comparative effectiveness, cost-effectiveness and utility assessment involving patient reported outcomes

(PROs), treatment switching, extrapolation, and subgroup analysis are reviewed in Section 3. We close with a summary of findings and main implications, methodologic challenges and research needs, and highlight the importance of early planning.

2. Early integration of payer requirements in trial design for enhanced market access

2.0 Early HTA in evidence planning at study design stage

Some countries require little additional evidence to successfully achieve reimbursement other than demonstrating a favorable benefit-risk profile to regulatory authorities. In the United States, for instance, market authorization allows immediate product launch readiness. However, many countries require additional evidence of clinical efficacy and safety, and/or favorable cost-effectiveness for drug reimbursement and patient access after regulatory approval. To optimize market access campaign, it is essential to consider HTA and payer requirements early in the clinical trial design. By engaging stakeholders (i.e., patients, clinicians, payers and other relevant parties), designing studies with relevant endpoints, incorporating real-world evidence (RWE), and aligning with HTA guidelines, Health Technology Developers (HTDs) can increase the likelihood of positive reimbursement decisions and successful market access. This proactive approach aims to ensure that new therapies reach patients faster, thereby potentially improving public health outcomes.

2.1 Identifying potential evidence gaps during study design to meet HTA requirements

Tunis and Turkelson (2012) conducted a comprehensive analysis using HTA as a tool to identify evidence gaps and inform the design of comparative effectiveness research. This study underscores the critical role of HTA in setting priorities for future health research, highlighting how systematic reviews often reveal significant knowledge gaps in common and critical clinical areas. Key evidence gaps identified by the authors include the insufficient duration of studies, which may not allow for adequate follow-up,

especially in diseases requiring long-term observation. Additionally, they pointed out the mismatch between internal and external validity, where highly selective patient populations in studies may not accurately reflect real-world clinical practice. Moreover, some outcomes assessed in studies may lack clinical relevance (Tunis & Turkelson, 2012). Most recently, a guidance titled "*Guidance on Validity of Clinical Studies*" was adapted by European HTA Member State Coordination Group on HTA in September 2024. This document introduces three key concepts for assessing the certainty of relative effectiveness results from clinical studies. It also discusses the strengths, weaknesses, and recommendations for various study designs. Although these discussions and recommendations do not apply to evidence syntheses from multiple sources, the publication of this guidance underscores the importance of considering different design options and selecting appropriate ones to meet HTA requirements, particularly in the context of the upcoming Joint Clinical Assessment (JCA).

To address these evidence gaps effectively, we utilize the PICO (Population, Intervention, Comparator, Outcome) framework to discuss specific considerations in study design. In the context of the EU HTA process, PICOs consolidated across the 27 member states of the European Union will be used to formulate HTA requirements (EU, 2024), which will be communicated to the HTDs before the JCA submission.

2.1.1. Populations

Due to local and historical variations in clinical practices, it is essential for HTDs to engage proactively with a broad range of stakeholders. This engagement will help identify the patient populations with unmet medical needs and assess the extent of these needs. Understanding the consensus among stakeholders regarding which patient populations should be prioritized is crucial for aligning study designs with payer expectations (Faulkner et al., 2021). Local variations in clinical practices can significantly influence the

selection of patient populations for studies. HTDs should carefully consider these differences and ensure that their study designs are adaptable to regional contexts. This might involve stratifying patient populations or developing region-specific study protocols. Identifying populations with unmet medical needs requires a thorough evaluation of the current treatment landscape. HTDs should assess the severity and scope of unmet needs across different regions and ensure that these needs are adequately addressed in their study designs (Sharma, 2015; Vreman et al., 2019).

2.1.2 Intervention

For interventions, it is crucial to consult with stakeholders. These consultations will help determine how to incorporate uncertainties related to dosing into both clinical and economic evaluations, ensuring that the intervention is appropriately assessed in different contexts or requirements specific to a country. Novel health technologies often present unique challenges, such as the integration into existing clinical practices and managing uncertainties with regard to uptake, dosing, and/or off-label use. HTDs should work closely with clinical practitioners and key opinion leaders to ensure that these challenges are addressed in the study design to generate the most appropriate evidence, allowing for a more accurate assessment of the intervention's effectiveness. Given the variability in how novel health technologies are utilized across regions, HTDs should ensure that their evidence generation strategies are adaptable to different local contexts. This may involve tailoring evidence packages to meet the specific needs and requirements of different regions (Fontrier et al., 2022).

2.1.3 Comparator

In disease areas where no standard of care (SOC) exists, choosing the right comparator for clinical trials becomes a critical challenge. This decision should be made in consultation with clinical development, key

opinion leaders (KOLs), and commercial teams to ensure that the chosen comparator is relevant and accepted across all target markets. HTDs need to clearly demonstrate the incremental benefits of their intervention beyond the SOC comparator in each market. This requires a deep understanding of the existing treatment landscape in each region and early collaboration between clinical and commercial teams to develop a coherent strategy. Nevertheless, the SOC may change, or a new competitor may become available at the time of HTA evaluation.

In situations where direct comparative data is lacking, HTDs may need to design studies that specifically address these gaps. To provide a more comprehensive comparison of treatment strategies, this could involve performing indirect treatment comparisons to another trial with the appropriate comparator or by leveraging RWE, or in the extreme scenario, conducting head-to-head trials to generate the evidence.

2.1.4 Outcomes

While regulatory submissions typically focus on outcomes like efficacy and safety, HTA bodies also consider other outcomes in their decisions, such as health-related quality of life, cost-effectiveness, and the wider societal and economic implications of the treatment in the longer term. It is important to note that EU HTA highlight the importance of considering all outcomes or endpoints with the same level of evidence, without ranking them. This approach aims to ensure that the evaluation of treatments is comprehensive and takes into account all relevant outcomes, regardless of their position in the hierarchy of endpoints defined in the trial. HTDs must recognize these differences and ensure that their study designs address both regulatory and HTA requirements (Wang et al., 2018). The choice of outcomes is therefore crucial. EU HTA and individual HTA agencies place great importance on patient-relevant endpoints that reflect a technology's therapeutic impact on patient experiences and outcomes that are important to the patients. These measures can provide valuable insights into the real-world effectiveness of the intervention.

Additionally, early filings such as at the first interim database lock, overall survival data from oncology trials is often lacking, but it is commonly a key endpoint of interest to HTA agencies as it provides long-term evidence from the treatment. In such cases, if the requested patient-relevant outcome is not available or if a HTA submission includes a surrogate/intermediate outcome , the probability of HTA/reimbursement success becomes low. Some statistical analyses may be planned to address this gap, but the uncertainty will be large. It is also important to consider the derivation of additional endpoints specifically for HTA use. These additional endpoints should be pre-specified in the study protocol to demonstrate to HTA (and regulatory) evaluators that they are not results-driven.

In addition to clinical outcomes, HTDs should carefully consider how evidence related to healthcare resource utilization (HCRU) (e.g., medical visits, pharmacy usage) and costs will be generated and analyzed, and the timing for such data collection. These types of information are often critical for HTA assessments and can provide a more comprehensive picture of the intervention's impact on healthcare systems.

Collecting PROs can be particularly challenging when different instruments or versions are required by different HTA agencies. Due to operational complexities and to avoid burden to study participants, it is often not possible to include all instruments to meet every need. In this case, HTDs may need to develop mapping strategies to an instrument required by the local markets, ensuring that the data remains meaningful, relevant and comparable for decision-making. However, mapping of PRO instruments could only provide an approximation of participants' experience and should not be systematically used in place of a proper data collection.

2.1.5. Other aspects of study design

The phenomena of treatment switching is becoming increasingly observed in clinical trials, particularly in long-term studies. However, collecting the necessary data to support robust analysis of treatment switching requires careful planning and execution within clinical trials (Latimer et al., 2016); more details appear in Section 3.3. HTDs should ensure that their study designs include mechanisms to proactively collect detailed data on treatment switching, including relevant baseline variables, timing, reasons, and subsequent outcomes. This data is crucial for accurately assessing the real-world effectiveness of the intervention and its comparative benefits and risks.

The duration of data collection is crucial, particularly in chronic diseases or conditions requiring long-term observation. HTDs should carefully plan the duration of follow-up and critical data to be collected as well as the timing of data collection to ensure that they align with both regulatory and HTA requirements. This may involve designing studies to capture long-term outcomes.

In summary, identifying and addressing potential evidence gaps at the study design stage is a critical component of meeting both regulatory and HTA requirements. By understanding HTA criteria, designing robust protocols, conducting gap analyses, and ensuring transparent reporting, researchers can develop studies that provide the comprehensive evidence needed to support the assessment and adoption of new health technologies without compromising the integrity of the trials.

2.2 Trial Design Considerations in Filling Evidence Gaps

This section aims to explore several study design concepts that can be utilized to generate early evidence in support of HTA. Table 1 provides an overview on these study designs.

Table 1. Study designs to generate early evidence in support of HTA

Design	Relevance for HTA	Advantages	Potential challenges and mitigations
FACTIVE	Provides HTA evidence on the technology concurrently and alongside regulatory evidence for MA. Enables causal inference by appropriate randomisation.	Early HTA evidence generation, concurrent with regulatory evidence, bridging the efficacy-effectiveness gap and enabling assessment of external validity.	The full design is complex, however FACTIVE is suggested to be used as a flexible toolbox of which design parts can be used to target specific evidence needs.
Pre-submission randomized pragmatic trials	Provides evidence on effectiveness of treatments in RW clinical practice.	Combines the real-world nature of an observational study with the scientific rigor of a randomized trial.	Evidence generation under clinical practice conditions could lead to high data variation, a high degree of missingness and inconsistency of data. Training of clinical practice personnel could enhance the availability and reliability of data.
SAT with ECA	Contextualizes information on a technology's effectiveness and safety where direct comparisons are not available.	Provides evidence on relative effectiveness and safety not available otherwise.	Population-adjustment approaches would usually be required and need to assume that there are no unobserved confounders. Where feasible, randomisation could be considered, or a target trial approach.
RWE/D in HTA	Complements evidence from traditional RCTs to provide a more comprehensive evaluation of technologies.	Provides insight into long-term effectiveness /safety, and support cost-effectiveness or utility evaluation	Inconsistent or missing data. Potential for bias (e.g.selection, information bias) and confounding. The use of registries, randomization where feasible e.g. pragmatic trials, and modelling could be considered.

FACTIVE = Flexible Augmented Clinical Trial for Improved evidence generation; MA = Marketing authorization; SAT = Single arm trial; ECA = External control arm; RWE/D = Real world evidence/data

2.2.1 FACTIVE Study Design

To facilitate simultaneous and concurrent evidence generation for regulators and payers, Flexible Augmented Clinical Trials for Improved eVidence gEneration (FACTIVE), a new class of trial designs was developed (Dunger-Baldauf et al., 2023). FACTIVE envisions flexible augmentation of confirmatory RCTs with concurrent and close-to-real-world elements. Certain well-defined treatment effects are estimated in the confirmatory part (core RCT) and other complementary treatment effects in a concurrent real-world part. High quality data are generated for both parts under one single protocol. The use of randomization ensures rigorous statistical inference and interpretation within and between the different parts of the trial. This enables payers to access their required evidence before marketing authorization, thereby supporting earlier patient access. FACTIVE designs can be tailored to the evidence needs of the technology, as illustrated in Yateman (2022). With early and comprehensive planning, FACTIVE could be designed to fill evidence gaps which might exist in available sources, for example for evidence synthesis. While the proposed augmented design offers various opportunities to increase the value of a technology, evidence needs related to the practical usability of a technology in a health care system will still need to be addressed post-marketing authorization, albeit likely with fewer studies, as some evidence typically collected in post-approval studies would, through the application of the FACTIVE design, be available pre-approval.

2.2.2 Pragmatic randomized clinical trials (PrCTs)

Pre-submission Pragmatic randomized clinical trials (PrCTs) play a crucial role in HTA by providing evidence on the effectiveness of treatments in real-world clinical practice. These trials encompass a diverse and representative population, allowing for broader inclusion and exclusion criteria, thus reflecting the heterogeneity of patients encountered in routine healthcare settings. Pragmatic trials "combine the real-world nature of an observational study with the scientific rigor of a randomized trial

and thereby give better answers to questions that are relevant to day-to-day clinical practice" (Zuidgeest et al., 2017).

The key design elements of PrCTs revolve around randomization, population, setting (primary care instead of research sites; fewer scheduled visits; drugs procured from pharmacies instead of HTDs supplying them), comparators (usual care instead of placebo), a variety of data sources (e.g., case report forms, electronic health records, insurance claims, mobile apps data) and outcomes (including HCRU, costs and PROs). PrCTs often have a simpler trial design with increased generalizability, but may be subject to pitfalls, such as selection bias, lack of treatment blinding, more missing data and nonadherence at treatment, and often require a large sample size (Le-Rademacher et al., 2023). The PRECIS-2 tool (Loudon et al., 2015) helps trial teams design PrCTs, by adjusting how explanatory (i.e., run under controlled conditions, on a homogenous population) or pragmatic a trial is, based on nine domains, with the goal of making the trial results more relevant to stakeholders. There is a continuum between explanatory trials (with high internal validity) and pragmatic trials (with high external validity), and a trial may contain elements of both categories. A recent trend is to incorporate pragmatic trial elements into an RCT, rather than have a full-fledged PrCT.

The questions addressed by PrCTs should be discussed with both regulators and payers upfront, considering a potentially smaller treatment effect in a real-world population, and the scarcity of HTA guidelines for PrCTs. The Salford lung study (Leather et al., 2020) is an example of a PrCT, phase III, which evaluated the effectiveness of a pre-licensed treatment. Its design and endpoints were discussed with regulators and the National Institute of Health and Care Excellence (NICE) as the main stakeholders.

2.2.3 Single-Arm trial using external control arm

The use of external control arms (ECAs) in single-arm trials (SATs) is becoming increasingly prevalent in HTA studies as traditional RCTs are facing practical and ethical concerns (ICON, 2021). In cases where a direct comparison group is unavailable, an ECA—constructed from clinical trial data or real-world data (RWD), whether retrospective or prospective—is crucial for contextualizing information regarding a product's clinical efficacy, safety, and cost-effectiveness. The reliability of findings depends on the comparability of the ECA and the SAT, which should encompass several key aspects (Appiah et al., 2024; Curtis et al., 2023; Patel et al., 2021; Sola-Morales et al., 2023).

First, addressing potential heterogeneity is essential for improving comparability, as failure to do so can introduce bias and undermine the reliability of results. This requires clearly defining patient populations and ensuring that external control patients are drawn from the same source population as those in the reference trial. Second, consistent outcome selection and definition between the ECA and SAT are also essential. Any variation in how outcomes are selected or defined can lead to misleading comparisons and obscure the true effects of the treatment. Third, any discrepancies in data collection methods or definitions across different trial settings require clear explanations to maintain transparency. This is because differences in how data is gathered, measured, or recorded can significantly impact the interpretation of results. Fourth, both geographical and temporal factors must be carefully considered when selecting data sources. Variations in healthcare systems and SOC practices across regions and over time can significantly influence treatment outcomes. Therefore, it's crucial to ensure alignment with the specific healthcare setting of the HTA submission and minimize temporal discrepancies between the ECA and SAT data. If a SAT and an external pragmatic trial are planned to be conducted concurrently, randomization of patients to either part could be considered to minimize confounding in technology assignment (application of the FACTIVE design). When using historical data, careful evaluation of potential changes in standard of care practices and their impact on comparability is necessary.

The HTA submission must clearly articulate the rationale for selecting the ECA data source and acknowledge any associated limitation. Such transparency is essential for establishing credibility and facilitating informed decision-making by HTA bodies. The use of patient-level data is highlighted as important for thoroughly evaluating comparability, data quality, and controlling for confounding variables. Employing sensitivity analyses can further evaluate the robustness of results in the face of data limitations (Curtis et al., 2023; Sola-Morales et al., 2023). Section 3 provides details on the statistical methodologies employed for such analysis.

While single arm trials alone may provide limited value in determining the relative effectiveness of the evaluated health technology against its comparator, data from SATs coupling with external sources as controls could allow comparative analysis to be performed. Under this circumstance, the target trial (Hernan and Robins, 2016; Hampson et al, 2023) emulation framework should be considered to properly formulate the causal inference questions, and thereby increase the internal validity and statistical precision of the comparative effective analysis.

2.2.4 Use of RWE/D in HTA

The scientific evidence generated from RWD and RWE are becoming increasingly important in HTA, complementing traditional RCTs to provide a more comprehensive evaluation of health technologies (Makady et al., 2017). While RCT remains the primary source of evidence for evaluating drug efficacy due to their rigorous design and ability to demonstrate causality, it has limitations in external validity. The strict eligibility criteria and controlled environments of RCTs often differ significantly from routine clinical practice, creating an efficacy-effectiveness gap (Eichler et al., 2011). RWD and RWE play a crucial role in bridging this gap by capturing treatment performance in real-world clinical settings, either

post-market or as part of pivotal trials, thereby enhancing the generalizability of HTA findings across diverse patient populations. They provide valuable insights into long-term outcomes, safety profiles, and cost-effectiveness—factors essential for informed healthcare decision-making (Akehurst et al., 2023; Makady et al., 2017).

Recognizing these benefits, HTA bodies are increasingly incorporating RWD/E into their assessment processes as a complement to RCTs, with a consistent emphasis on rigorous study design to ensure the validity of generated evidence (de Pouvourville et al., 2023; Makady et al., 2018). Key considerations for robust RWD/E studies include identifying appropriate high-quality RWD sources, employing robust statistical methods to minimize confounding and establish causality, and adequately accounting for various biases at the design stage. HTA agencies emphasize the particular importance of fit-for-purpose data that demonstrates both reliability and relevance. Various study designs can maximize the utility of RWD/E in HTA, offering viable alternatives when RCTs are infeasible or unethical. Pragmatic trials discussed in Section 2.2.3 is one type of real-world studies that can evaluate the relative effectiveness of a new intervention in a population representative of real-world patients. It is crucial to select study designs that align with the specific objectives of the HTA assessment, carefully considering all design elements, including Estimand. Section 2 in the Companion article provided a discussion on the comparison and contrast of the Estimand Framework for a study vs. PICO for a HTA evaluation (Jen et al., 2024).

However, one significant challenge in using RWD/E in HTA is the potential for biases and confounding factors that can compromise the validity of the findings (Akehurst et al., 2023; Makady et al., 2017; Zisis et al., 2024). RWD is subject to selection bias, information bias, and confounding due to unmeasured variables. To address these challenges, various methodological approaches can be employed at the design stage (Makady et al., 2017; Zisis et al., 2024). Additionally, HTDs should implement a multi-faceted approach, including validation studies, multiple data sources for cross-verification, advanced statistical

methods, and comprehensive sensitivity analyses. Section 3 provides more details on statistical methodologies for real-world studies.

2.3 Challenges, considerations, recommendations

The evolving evidentiary standard to secure market access and favorable reimbursement decision in countries/regions is posing new challenges for clinical development programs and observational studies utilizing RWD. To address those challenges, one key aspect is to properly integrate payers' need into the study design. Such integration requires careful consideration on factors such as the complexity, operational feasibility, cost and speed of clinical development program, the regulatory requirements and plausible regulatory strategies, and anticipatory evidence need from payers including HTA agencies. Navigating the intersection of those factors is a multifaceted endeavor and often needs involvement of multiple stakeholders within a HTD in strategic planning. As HTDs prepare the launch of their products, they should realize that clinical development and product launch are not isolated efforts. Rather, those efforts are interconnected, and involving the commercial teams (e.g., market access) early in the development program is crucial. A study design that balances both regulatory needs and HTA requirements ensures a smoother transition from approval to market access.

While this early evidence integration approach would offer numerous benefits, we cannot underestimate the challenges.

- First and foremost, lack of awareness and knowledge of HTA requirements (especially the new requirements under JCA) within the clinical development program could lead to insufficient consideration of evidence needs for HTAs. Furthermore, lack of communication and coordination between clinical development and market access/HEOR (Health Economic Outcome Research)

teams could cause delay in clinical study design and execution, and conflict with timely regulatory submissions.

- Secondly, while the PICO concept is not completely new to HTDs, extensive simulations that predict PICOs need to be conducted to inform the study design and pre-specified statistical analysis (e.g., subgroup analysis) in preparation to the JCA, and optimizing trial design using those consolidations requires an adaption in both decision making and operational models (EFPIA, 2024; EUnetHTA, n.d.).
- Thirdly, given the tight submission timeline of JCA (EU, 2024), HTDs are likely to initiate early the activities, e.g., indirect treatment comparison and pre-specified subgroup analysis, before the readout of pivotal trial results.

In addition, innovative approaches applied to clinical development may have undesired consequences in the HTA evaluation. For instance, in rare disease areas, SAT is not uncommon for regulatory approval. However, the SAT designs may not always be viewed as adequate in providing compelling evidence for access/reimbursement decision (Institute for Quality and Efficiency in Health Care (IQWiG), 2023; Jaksa et al., 2022). Furthermore, products approved based on a SAT will require a standard of care (SOC) on individual patient data (IPD) in the Health Technology Assessment Regulation (HTAR) submission and this requirement may limit the choice of data from clinical practice to constitute external control arm, if not all data is accessible to HTAR (IQVIA, 2022). Even if the IPD are available and accessible to HTAR, additional burden on the operation side (e.g., extra efforts to anonymize data) is not ignorable, and additional resource need to be planned for. Sometimes, implementing novel designs for early HTA evidence generation may limit the options to use such a specific study for evidence synthesis. For instance, retrospectively combining phase 3 RCTs with a pragmatic trial will be challenging in the best case, up to impossible if there are additionally differences in terms of inclusion/exclusion criteria, data sources, study setting, between the RCTs and pragmatic trial. Furthermore, the use of RWE/D in clinical

development and regulatory submissions comes with challenges, such as data quality issues, data relevance to the research questions at hand, and potential biases in study design or data analysis.

To address those challenges, HTDs should take efforts both externally and internally.

- Externally, HTDs should proactively engage with local HTA bodies (e.g., NICE, Federal Joint Committee, Germany (G-BA)) and other local stakeholders (e.g., physicians) to better understand the evidentiary requirements for benefit categorization. To prepare for JCA, it would be beneficial for HTDs to solicit feedback from both EMA and JSC through parallel consultation. These proactive engagements could potentially streamline evidence generation plans, addressing market authorization and access/reimbursement simultaneously.
- Internally, close collaboration among clinical development, regulatory and market access teams is the key to ensure alignment and optimized evidence generation from clinical trials. A few possible tasks that could be conducted under such collaboration include: to explore the possibility to influence trial eligibility criteria and make target population closer to payer-relevant populations; to understand the relationship between payer-relevant populations in each market and the label population; to probe if there exist heterogeneous treatment effect and safety profiles of different subgroups in the trial and if they exist, their impact on the HTA evaluation. And harmonize the potential subgroup analysis from each market and/or plan necessary analysis to extrapolate clinical evidence from label population to more generalized payer-relevant populations.
- Lastly, while it is critical and essential to call for closer collaboration between clinical development teams and commercial teams to incorporate as many elements required in HTA into trial design or to generate evidence from other sources, once the study design is finalized, commercial teams will need to discuss the implications of proposed trial design to HTAs and propose any mitigation plan (sometimes at risk) to address gaps in evidence need from the current clinical trials.

In summary, successful navigation of these complexities requires collaboration, adaptability, and a scientific-driven, systematic approach by HTDs. By integrating clinical development, regulatory and HTA strategies, HTDs could optimize evidence generation and achieve successful market access/reimbursement decisions.

3. Analysis methods to fulfill different HTA needs

A comprehensive evaluation of clinical effectiveness, cost-effectiveness, and quality of life is paramount to fulfill HTA needs beyond clinical trial evidence. This endeavor necessitates the employment of various evidence synthesis methods to integrate clinical evidence from diverse sources. Indirect treatment comparisons (ITCs), such as the Bucher method, network meta-analysis (NMA), matching-adjusted indirect comparison (MAIC), simulated treatment comparison (STC), and multilevel network meta-regression (ML-NMR), play a pivotal role in addressing variations of comparative study populations when direct head-to-head trials are lacking.

Clinical trials often have limited durations, necessitating the use of extrapolation techniques to estimate longer-term outcomes to demonstrate the value of a product to payers. Standard or flexible parametric modelling, Bayesian methods, and leveraging external data with longer-term follow-up can be employed to address this limitation. Additionally, treatment switching, a now common occurrence in clinical trials, can introduce bias and must be addressed through appropriate methods. Adjustment for treatment switching phenomena allows for a “cleaner” treatment effect to be estimated, yet still accounts for actuality in the real-life clinical practice.

More so in HTA than in regulatory approvals, health utility and PRO evidence are essential components in the context of health economics and quality of life (QoL). Their analyses also vary in importance, approaches and complexities based on different jurisdictions. Additionally, subgroup analyses play crucial roles for identifying potential heterogeneity and uncertainty in treatment effects across different patient populations to better inform national decision-making.

This section continues to address these topics in more detail.

3.1 Evidence synthesis method

Evidence synthesis plays a pivotal role in integrating data from diverse sources. This section delves into specific evidence synthesis methods.

3.1.1 Meta-analysis and network meta-analysis

Evidence-based healthcare decision-making in HTA requires comparison of all relevant competing treatments. However, robustly designed RCTs that simultaneously compare all treatments of interest are rarely available. Therefore, evidence synthesis including both direct and indirect evidence plays a critical role in decision-making by HTA agencies, providing useful information on the comparative effectiveness of multiple treatments.

Meta-analysis (MA) is a statistical method for combining or pooling results from multiple studies. It could be a pairwise meta-analysis where there are two interventions of interest, or Network MA (NMA) where multiple treatments are of interest. NMA compares multiple treatments by using both direct comparisons of interventions within randomized controlled trials and indirect comparisons across trials via a common comparator (Dias et al., 2018). In MA/NMA, studies refer to RCTs unless otherwise specified. Both pairwise MA and NMA can be conducted using either frequentist or Bayesian framework

with the Bayesian framework being more commonly adopted in practice. For the NMA, in particular, due to the advantages of being able to incorporate prior knowledge or expert opinion through the use of prior distribution; directly providing probabilistic statements about efficacy and ranking of treatment; and being highly flexible and able to handle complex models and small sample size.

Pairwise MA and NMA share a common assumption of homogeneity. NMA has additional assumptions related to homogeneity, transitivity and consistency due to combining direct and indirect evidence across a network of studies. This requires the distribution of treatment effect modifiers to be balanced between studies (Phillippo et al., 2018). Pairwise MA could be considered as a special case of NMA where there are only two treatments in the network. Either fixed effect or random effects model could be used in MA/NMA. The two types of models differ in assumptions and their interpretations. A fixed effect model assumes that there is a common treatment effect, and the variation is due to sampling error. Heterogeneity is expected in evidence synthesis because it combines studies that may have clinical and methodological heterogeneity. A random effects model would be preferred as it allows for heterogeneity in the treatment effects by assuming exchangeability of treatment effects. In the case of limited data, an informative prior could be used to help the estimation of the heterogeneity parameter between the studies (Ren et al., 2018).

Various guidance is available for conducting a Bayesian NMA (Dias et al., 2013; Hoaglin et al., 2011). They cover the details of analysis for various types of endpoints including categorical, binary, or continuous. Frequentist NMA approach can be found in Lumley (2002). Woods et al. illustrated how to conduct NMA of survival outcomes by assuming that the proportional hazards (PH) assumption holds (Woods et al., 2010). The model can incorporate data reported using hazard ratio (HR), median survival as well as count data (i.e., number of events and sample size). A limitation of the Woods approach, however, was the fact that it does not account for violation of the PH assumption, which could potentially introduce bias to the results. PH assumption should be assessed for time to event outcomes source publications by reconstruction patient-level survival data by digitalizing Kaplan-Meier curves when

reported (Guyot et al., 2012). The fractional polynomial model (FP) (Jansen, 2011) and Royston-Parmar model (Freeman & Carpenter, 2017) could be considered when any included studies have shown departures from the PH assumption.

3.1.2 Indirect treatment comparison

Indirect treatment comparison (ITC) refers to a comparison of different treatments that have not been directly compared with each other in a head-to-head (H2H) trial. It is often used when there is no evidence or insufficient evidence from H2H trials or when more than two medical interventions are of interest. Traditional ITC methods include the Bucher method, NMA, or population-adjusted indirect comparison (PAIC) methods such as MAIC or STC (Phillippo et al., 2018). Novel ITC methods have also been developed to overcome challenges of traditional ITC methods, for example, ML-NMR (Phillippo, Dias, Ades, Belger, et al. (2020). The earliest technique for adjusted ITC was introduced by Bucher et al. (1997). The Bucher method is based on the odds ratio (OR) as the measure of treatment effect. The Bucher method can be applied in star-shaped networks to obtain indirect comparisons of each pair of treatments via a shared comparator. It is also applicable to more complex networks including closed loops, but only in the form of pairwise comparison (Tingle et al., 2024). The Bucher method assumes the relative treatment effects are constant across included trials. It has the advantage of preserving the within-study randomization. However, the Bucher method may lead to biased results when the distribution of effect modifiers is imbalanced between trials. In addition, the Bucher method is not applicable for multi-arm trials as it assumes independence of pairwise comparisons (Bucher et al., 1997).

When the assumption of transitivity is violated, traditional methods of indirect comparisons, such as NMA and Bucher's method, can produce biased results (HTA Coordination Group, 2024; Phillippo et al., 2016). MAIC and STC with relaxed assumptions were developed. These methods adjust for imbalances in baseline covariates between studies to provide unbiased estimates of treatment effects in a setting when

IPD from one study are available alongside aggregated data (AgD) from a published study (HTA Coordination Group, 2024; Phillippo et al., 2016). MAIC uses IPD to match relevant baseline characteristics reported in a comparator's trial with AgD, subsequently re-weighting outcomes to facilitate comparison with the published outcomes (Signorovitch et al., 2012). In contrast, STC involves fitting outcome regression models using IPD to predict outcomes for a comparator's population, and then comparing these predicted outcomes with those reported in the publication (Phillippo et al., 2018).

MAIC and STC can be applied in two forms: anchored, for a connected network, and unanchored, for a disconnected network without a common comparator. The anchored forms assume conditional constancy of relative effects, meaning that the relative treatment effect is balanced across all effect modifiers. On the other hand, the unanchored forms require the much stronger assumption of conditional constancy of absolute effects, implying that the absolute treatment effect is balanced across all effect modifiers and prognostic factors. This assumption is generally considered unrealistic to meet (Phillippo et al., 2016). MAIC and STC have become more commonly used in HTA submissions. However, their appropriateness must be carefully justified. EU HTA (EUnetHTA, n.d.) highlights that when decisions are based on MAIC and STC, there should be a sufficiently large treatment effect to ensure that the observed effect is not solely due to unmet assumptions, such as missing covariates. Unanchored methods should only be considered when no connected network exists, such as in single-arm trials (Phillippo et al., 2018).

The ML-NMR extends the standard NMA framework by synthesizing both IPD and AgD through a connected network of multiple studies and treatments (Phillippo, Dias, Ades, Belger, et al., 2020). As this method leverages IPD, unlike standard NMA, it relaxes the assumption that effect modifiers are balanced across populations and derives population-adjusted indirect comparisons. When no covariates are included, the ML-NMR model simplifies to a standard NMA. Conversely, when IPD are available from all studies, it becomes an IPD network meta-regression, which is ideal but often difficult to achieve. In the ML-NMR analysis, once the individual-level likelihoods are specified, they are integrated over the target

population to create an aggregate-level likelihood. Studies providing only AgD are fitted by integrating the individual-level model over the covariate distributions, linking the individual- and aggregate-levels of the model, thereby avoiding aggregation bias. A quasi-Monte Carlo approach has been used for integration due to its flexibility and efficiency. (Phillippo, Dias, Ades, Belger, et al., 2020). Details of the mathematical framework and implementation of ML-NMR is introduced elsewhere (Phillippo, Dias, Ades, Belger, et al., 2020). In addition, a case study motivated by HTA is provided.

Several studies have demonstrated the advantages of ML-NMR. In a simulation study for anchored ITC, ML-NMR and STC performed similarly well under correct assumptions, effectively reducing bias, while MAIC often increased it (Phillippo, Dias, Ades, & Welton, 2020). The study reported that MAIC underperformed in all scenarios considered, with issues related to sample size and population overlap. Compared to STC, ML-NMR exhibited greater flexibility in handling larger treatment networks and could derive estimates for any target population. In a real-life application involving four clinical trials, ML-NMR substantially reduced the uncertainty of the population-average relative effect estimates compared to random-effects NMA by accounting for both within- and between-study variation (Phillippo, Dias, Ades, Belger, et al., 2020). These findings align with the perspectives of multiple HTA agencies. The NICE Decision Support Unit (DSU) report states that ML-NMR is the preferred method for population adjustment in indirect treatment comparisons, favoring it over MAIC and STC (Abrams, 2020). The report specifies that MAIC should not be used under any circumstances, while STC is suitable for scenarios involving two studies. Other HTA agencies generally support its use, provided that proper implementation minimizes bias and accurately represents covariate distribution. Although EU HTA noted that the current ML-NMR method has limitations, particularly its inapplicability to time-to-event outcomes, an upcoming publication is expected to address this issue (Phillippo et al., 2024). The extension to general likelihoods, including survival outcomes, will significantly enhance the method's applicability and effectiveness.

3.1.3 Non-randomized studies

Sometimes data from RCTs may not be available for an intended technology comparison, but comparative observational studies with patient level data (IPD) might be available. Methods addressing the potential confounding bias for comparative observational studies are discussed in the NICE DSU technical Support Document (TSD) 17 (Faria et al., 2015). These can be broadly distinguished by whether only observed confounders are controlled for (assuming ignorability of technology conditional on a set of observed confounders, "selection on observables" in the guidance) or can control for unobserved confounders ("selection on unobservables").Methods assuming selection on observable covariates include regression adjustment, inverse probability weighting (IPW), doubly robust methods, and regression on the propensity score and matching. Methods assuming selection on unobservable covariates include instrumental variable methods.

A wide variety of matching procedures have been proposed in the literature and, currently, there is no consensus on how exactly matching ought to be done and how to measure the success of the matching procedure (Sekhon, 2011). Matching based on multivariable regression methods is detailed in existing literature (Gelman & Hill, 2006), with further developments in PSM methods (Sekhon, 2011) and visual assessment of matching also discussed (Pruzek & Helmreich, 2009). Multivariable regression uses IPD from two data sets to predict which data set a patient is likely to belong to. Propensity score matching (PSM) methods have become popular, although methodological review papers have suggested that they have little advantage over traditional multivariable regression methods (Stürmer et al., 2006). Propensity score matching uses IPD from one data set to produce weights to match to another data set. Additionally, applications of bootstrapping to PSM methods have been introduced (Pan & Bai, 2015).
Recent advancements include doubly robust methods, such as targeted maximum likelihood estimator survival TMLE, all utilizing the survtmle package in R (Benkeser & Hejazi, 2017, Chen et al 2023)..

TMLE makes use of IPD from two data sets to fit regression-based models simultaneously to both arms. The survival TMLE package in R can use a Cox regression to model the covariates. It also leverages ensemble machine learning techniques to estimate parameters in a flexible manner. The procedure uses cross-validation to select the best-performing estimator from a library of candidate estimators.

NICE DSU TSD 17 recommends that the sensitivity of the results should be explored by estimating alternative models that rely on different assumptions (Faria et al., 2015). Hence it is not uncommon for statisticians to therefore suggest in a HTA statistical analysis plan.that all these methods should be performed and the results be compared

3.2 Extrapolation

The economic evaluation of a new health technology is crucial within HTA, serving as the cornerstone for decisions regarding market access and reimbursement. It plays a vital role in guiding policymakers on the allocation of limited healthcare resources, aiming to optimize their utilization. As required in HTAs, cost-effectiveness models (CEM) and budget-impact models (BIM) constitute the economic evaluation. According to various HTA guidelines (e.g., NICE and the Canadian Agency for Drugs and Technologies in Health(CADTH)), the recommended time horizon for assessing cost-effectiveness should comprehensively capture all relevant differences between the compared health technologies. Often, a lifetime horizon is deemed most suitable, particularly in chronic disease contexts where interventions can exert long-term impacts on patients. However, this necessitates extrapolating beyond the typically limited duration of clinical trials to thoroughly evaluate treatments, especially for periods extending beyond the available data. It's imperative not to underestimate the uncertainty entailed in such extrapolation. Various approaches exist, ranging from scientifically arguing for the expected durable effect of new health technology (Institute for Clinical and Economic Review (ICER), 2023) to simple extrapolation, (Rheault et al., 2023) extending observed outcomes trajectories to patients' lifetimes, and employing more complex

statistical models, (Shah et al., 2023) including AI/ML-based analytical methods, to predict long-term outcomes.

However, these extrapolations have limitations; for example, simple extrapolation assumes that observed trends will indefinitely continue, which may not always be scientifically plausible. Complex statistical models rely on certain assumptions and are influenced by factors like data volume and completeness, where even slight violations in assumptions or variations in data can drastically alter extrapolated values. Moreover, extrapolating lifetime or long-term outcomes from studies with limited follow-up leads to highly variable extrapolations as time extends beyond available data. Hence, it's crucial to conduct sensitivity analyses to quantitatively assess uncertainties associated with CEMs. In assessing treatments aimed at improving survival, extrapolation is commonly used, notably in oncology trials where survival data often undergo censoring at trial completion. Extending survival impact beyond observed data is essential for evaluating complete benefits. Economic evaluations prefer estimating mean effects on time-to-event for assessing incremental quality-adjusted life years (QALY) gains, favoring parametric models for survival extrapolation (NICE DSU TSD 14 (Latimer, 2011)). NICE DSU TSD 14 identifies six parametric models, each based on different hazard function assumptions, as standard methods.(Latimer, 2011). Model selection depends on fit to observed data and plausibility for unobserved data, with an emphasis on systematic assessment and summarizing methods. Similar procedures are endorsed by other agencies like CADTH. However, standard models have limitations, relying on plausible hazard function assumptions for specific scenarios, with inaccurate estimates resulting from implausible assumptions. More flexible methods for survival analysis such as flexible parametric survival, mixture, landmark, piecewise, cure and excess mortality models together with methods for incorporating external information can be found in NICE DSU TSD 21 (Rutherford et al., 2020). However, the complexity of these more advanced methods doesn't guarantee better outcomes.

This challenge would be even more prominent with the growing wave of developing new therapeutics based on technologies that modify a patient's gene (gene therapies) or technologies that transplant human cells to replace or repair damaged tissue and/or cells of a patient (cell therapies). In many instances, gene and cell therapies are viewed to be "curative" but clinical trials usually do not have long enough follow up data when the associated products receive regulatory approvals. Therefore, the durable effect of the new therapy remains largely unknown when HTA agencies and other payor-oriented organizations assess clinical and economic benefits of the products. This uncertainty could cast doubts on the determination of the effect of gene/cell therapies in HTA evaluations (ICER, 2022, 2023; NICE, 2023). Furthermore, gene and cell therapies predominantly address rare diseases characterized by considerable unmet medical needs, often leading to clinical trials with limited sample sizes. This further limits the information available to conduct scientifically sound extrapolation, and the development of new statistical methods is needed to address this challenge. Recently, Pan et al. developed a new Bayesian data selection approach that is able to select and integrate data outside the clinical trials of an investigational gene therapy, to help reduce the variability of the predicted outcome in the long run (Pan et al., 2024).

3.3 Treatment switching

Treatment switching poses challenges in estimating survival outcomes, which are crucial for economic evaluations. Various methods have been proposed to handle treatment switching. Careful consideration of trial characteristics, underlying assumptions, and fit-for-purpose data is essential for selecting an appropriate method and interpreting its results (Latimer & Abrams, 2014).

Simple methods like excluding or censoring switchers can introduce bias. The Rank Preserving Structural Failure Time Model (RPSFTM, Robins et al, 1991) estimates the counterfactual survival time, representing the treatment effect a patient would have experienced if they had not switched treatments. It assumes a "common treatment effect", where the magnitude of the treatment effect is the same regardless

of when a patient receives the experimental treatment. The Iterative Parameter Estimation (IPE) algorithm also assumes a "common treatment effect" and requires suitable statistical distribution models for survival times (Latimer & Abrams, 2014). The Inverse Probability of Censoring Weights (IPCW) assigns weights to patients based on their likelihood of switching treatments, aiming to create a hypothetical group representing what would have happened without switching. It relies on the assumption of no unmeasured confounders, meaning that all factors influencing both treatment switching and survival are accounted for in the analysis (Robins & Finkelstein, 2000). Two-stage estimation (TSE) estimates counterfactual survival times after a secondary baseline. Compared to the IPCW, it has an advantage of not requiring data to be collected on time-dependent covariates except those at the secondary baseline. The RPSFTM, IPE, and TSE could apply recensoring at an earlier time-point for switchers who survive during the study to address the issue of informative censoring. However, recensoring leads to loss of long-term information, which is a major concern in economic evaluation. TSE with IPCW addresses the informative censoring and can outperform the TSE with recensoring, but it also relies on the assumption of no unmeasured confounders (Latimer et al., 2019). In Ying and Tchetgen (2023), a structural cumulative survival model (SCSM) is proposed, which uses randomization as an instrumental variable to account for selection bias in switching. Furthermore, it accommodates unmeasured confounding by leveraging initial randomization as an instrumental variable. The SCSM is further enhanced by developing a doubly robust estimator, relying on a model for the randomized arm and another model for the hazards of death (Michiels et al., 2024). It remains unbiased even if one of the two models is mis-specified.

Treatment switching adjustment methods have assumptions and limitations. Conducting sensitivity analyses using different methods and exploring various assumptions can provide a comprehensive understanding of the uncertainty surrounding estimated treatment effects and cost-effectiveness results. Collecting suitable data at baseline and over time is crucial for several methods. Additionally, it is also important to use methods that can accommodate extrapolation beyond trial period to study lifetime benefits of treatment.

3.4 Patient-reported outcomes and health utilities

PROs and utilities both are important concepts in healthcare research and decision-making, particularly in HTA and cost-effectiveness analyses. They provide insights into patients' experiences, preferences, and the value they place on different health states, and ensure that patient values and preferences are considered alongside clinical and economic evidence. PROs refer to any report of a patient's health condition, symptoms, or quality of life directly from the patient themselves (FDA Guidance, 2009, 2018). PROs can include measures of physical functioning, symptom severity, emotional well-being, general health perceptions, and overall quality of life. They are typically collected through self-reported questionnaires or interviews, enabling patients to express their own perspectives on their health status and treatment outcomes. PROs are valuable in HTA and healthcare decision-making because they provide a patient-centered perspective, by capturing aspects of health that may not be captured by clinical measures alone and allowing for a more comprehensive assessment of the impact of a healthcare intervention on patients' lives. The 36-Item Short Form Survey (SF-36), 12-Item Short Form Survey (SF-12), and the EuroQol 5-Dimension Questionnaire (EQ-5D) are commonly used generic PRO instruments to assess health-related QoL (HRQoL) and health status. There are also disease-specific PRO measures that are designed to assess the unique aspects and challenges associated with a particular disease or health condition.

Health utilities, also known as preference-based measures or health-related QoL utilities, on the other hand, are numerical values that are used to quantify the value or desirability of different health states or outcomes. Utilities are typically represented on a scale from 0 to 1, where 0 represents a state equivalent to death or worst possible health and 1 represents perfect health (Wailoo et al., 2023). Utilities reflect individual preferences and can be used to compare the relative value of different health states or

interventions. Estimates of health utility can be obtained by either direct or indirect methods. With direct methods, utilities are often derived through preference-elicitation methods, such as time trade-off (TTO), standard gamble (SG), or visual analog scale (VAS). These methods involve asking individuals to make trade-off decisions or express their preferences for different health states relative to each other (Morimoto & Fukui, 2002). These preferences are then used to assign utility values to specific health states. Health utilities can also be indirectly elicited using generic preference-based instruments such as EQ-5D, Health Utility Index (HUI), or Short Form 6-dimension (SF-6D), an abbreviated variation of SF-36, which have been developed and validated for use across different health conditions and populations. These instruments provide a standardized method for measuring health utilities, which can facilitate comparisons across different health technologies and interventions.

However, one of the potential limitations of generic preference-based instruments is that they may lack sensitivity in specific disease contexts. In some cases, disease-specific instruments may be more appropriate for capturing the unique aspects of a particular disease or condition. These instruments may be designed to capture symptoms or functional limitations that are not captured by generic instruments and may be more sensitive to changes in health status in specific patient populations. When a disease-specific utility measure is not available, mapping the descriptions from a disease-specific instrument to the utility algorithm of a generic instrument is a potential alternative approach. This involves establishing a statistical relationship between the scores on the disease-specific instrument and the scores on the generic instrument, which can then be used to estimate health utilities for the disease-specific instrument (Wailoo et al., 2023). This approach can be useful when a disease-specific instrument has been implemented in a study, but health utilities need to be estimated for economic evaluations or comparisons across different interventions.

Utilities are often used in cost-effectiveness analyses to calculate QALYs, a measure that combines both quantity and quality of life. QALYs allow for comparisons of the health benefits of different interventions

and help inform resource allocation decisions. By incorporating health utilities into cost-effectiveness analyses, decision-makers can assess the value for money of different healthcare interventions or treatments. Utilities help in comparing the benefits and costs of interventions across different disease areas and can inform resource allocation decisions in healthcare systems. Many HTA agencies make specific statements about their preferences for health utility measures, but the guidelines may change over time (RTI Health Solutions, 2022). It is important to stay informed about the current guidelines and any changes made by HTA agencies.

3.5 Subgroup analyses

Subgroup analysis involves examining the effects of a medical intervention on specific subgroups of patients within a larger population. These subgroups may be defined by factors such as age, gender, disease severity, genetic characteristics, or other relevant criteria. Subgroup analysis plays a crucial role in HTA and helps to explore the heterogeneity of treatment effects across different subgroups of patients. By examining the outcomes and effectiveness of an intervention within specific subgroups, HTA agencies can provide more targeted and personalized recommendations for its use. In addition, subgroup analysis can explore potential differences in safety and cost-effectiveness within different subpopulations. This information is critical for HTA agencies to understand the overall value and impact of an intervention across various patient groups.

Payers, including national health systems and insurance agencies, are increasingly demanding subgroup analyses to inform their coverage and reimbursement decisions for healthcare technologies (Aggarwal & H, 2013). In addition to informing coverage and reimbursement decisions, subgroup analysis can help payers optimize their resource allocation by targeting the most cost-effective interventions to the patients who are most likely to benefit. Subgroup analyses are discussed in many HTA guidance documents (e.g., NICE and IQWiG). Researchers should consult the specific guidelines and recommendations provided by

their respective country's HTA agency or relevant professional societies for more detailed and context-specific guidance. The principles and best practice for conducting subgroup analyses for HTA purpose are outlined in the literature (Paget et al., 2011).

The PICO framework can be used to guide subgroup definition by helping researchers identify relevant subgroups based on the characteristics of the population, intervention, comparison, and outcomes being studied. Performing adequate subgroup analysis can be difficult due to several statistical and methodological challenges such as pre-specification, type I errors, multiplicity problems, lack of power and ecological bias (Wijn et al., 2019). The interpretation and recommendations based on subgroup analyses can also be complex and controversial. Open dialogue and careful consideration are necessary to ensure that HTA recommendations for subpopulations are clear, transparent, and ultimately serve the best interests of patients and healthcare systems.

3.6 Challenges, consideration and recommendations

MA and NMA are essential statistical methods for evidence synthesis and decision-making in HTA, allowing for the comparison of multiple treatments by combining direct and indirect evidence. While both pairwise MA and NMA can be conducted using frequentist or Bayesian frameworks, the Bayesian approach is more commonly adopted for NMA due to its advantages. Random-effects models are preferred over fixed-effect models to account for heterogeneity. Recommended methods for analyzing time-to-event outcomes under the proportional hazards assumption are well-documented in the literature (Woods et al., 2010). Alternative models like fractional polynomial or spline models should be considered if this assumption is violated.

Traditional methods like Bucher's and NMA assume transitivity, which can lead to bias if violated. MAIC and STC overcome this shortfall by adjusting for covariate imbalances when IPD are available from one

study, but assumptions like conditional/absolute constancy of effects must be justified, which is often unrealistic in the unanchored analyses. The ML-NMR extends NMA and relaxes the balanced effect modifiers assumption. Multiple studies demonstrate ML-NMR's advantages over MAIC and STC in reducing bias and uncertainty. NICE in the UK prefers ML-NMR over MAIC and STC, while it cautions on the use of MAIC. These views are also supported by EUnetHTA, CADTH, and IQWiG.

In non-randomized trials when IPD are available, NICE DSU recommends a selection of methods to estimate the treatment effects based on observed and unobserved variables to perform indirect comparisons. Given the lack of consensus on the best methods, the NICE DSU TSD 17 recommends performing sensitivity analyses using multiple adjustment methods requiring different assumptions and comparing the results to assess the robustness of findings.

In HTA, economic evaluations often require extrapolating survival data beyond clinical trial durations to capture full treatment benefits over a lifetime horizon. While standard parametric survival models are commonly used, they rely on assumptions about hazard functions that may not hold, which may lead to inaccurate extrapolations. Other flexible methods described in NICE DSU TSD 21 can improve extrapolations, but the increased complexity does not guarantee better performance. Bayesian methods can also help to reduce uncertainty in long-term extrapolations by incorporating external information and is particularly useful in gene/cell therapies. Regardless of the extrapolation method used, conducting extensive sensitivity analyses is crucial to quantify the uncertainty associated with survival projections informing cost-effectiveness models.

The presence of treatment switching in clinical trials is common, which can bias the survival analysis and compromise economic evaluations relying on survival estimates. While it is known and well-accepted that simple censoring leads to bias, more advanced statistical methods proposed to adjust for switching come with various assumptions and limitations. Adjustment method with re-censoring has been criticized and

should not be a method of choice when it leads to a considerable loss of information. Given the varying assumptions, conducting sensitivity analyses using multiple switching adjustment methods is recommended.

PROs and health utilities are critical concepts in HTA and economic evaluations for incorporating patients' perspectives on their health status, quality of life, and preferences for different health states. While generic PRO and utility instruments allow standardization (especially when a comparison across disease indications is needed), disease-specific measures may better capture unique aspects of particular conditions. Disease-specific PRO and utility instruments are generally preferred by HTA agencies, though mapping on a generic instrument may be used in their absence. HTA agency guidelines on PRO/utility frequently evolve over time, and therefore should be regularly consulted.

Subgroup and subpopulation analyses are vital for understanding the overall value proposition across different patient populations in HTA; with increasing demand for subgroup and subpopulation analyses to optimize coverage decisions and resource allocation towards subgroups who benefit the most. However, conducting robust subgroup analyses poses statistical challenges like multiplicity issues, lack of power, and ecological bias. Pre-specification of subpopulations and subgroup analyses using frameworks like PICO could reinforce the statistical rigor and allows transparent interpretation.

4. Summary and Conclusions

While it may not be exhaustive, this article provides reviews and discussions on how different study design and statistical methods could help in filling evidentiary gaps in HTA submissions based on our collective expertise. We have discussed the importance of following the PICO framework and have delved into the specific evidence needs from population, intervention, comparator, and outcome

perspectives, thereby noting that PICO does not handle treatment switching, in contrast to the estimand framework. Furthermore, the statistical challenges of using RWD in HTA could only be briefly described in this overview article. PICO and other research structuring frameworks in HTA and RWD use in HTA are included in the research topics the SWG is working on.

Recognizing the need for early evidence generation for HTA, we have also explored various study design concepts that allow for parallel planning of evidence collection alongside regulatory submission planning. We have highlighted the challenges and opportunities associated with planning studies that address HTA requirements. Additionally, we have provided a detailed overview of complex statistical methods that are essential for HTA submission. By outlining these methods and their underlying assumptions, we aim to guide practitioners on when and how to utilize them in their own applications. We have also extensively discussed the challenges and opportunities in the application of these statistical methods.

We aim to offer strategic considerations for early evidence planning related to HTA, alongside specific statistical methodologies commonly used in delivering clinical evidence and demonstrating value. Our targeted audience includes statisticians working in clinical development who may not be familiar with the intricacies and specific needs of HTA. While we covered a wide range of topics, we did not delve deeply into any particular area. Our goal is to raise awareness of HTA needs and strategic considerations for those who have not worked in this field, but who may need to understand HTA requirements to generate evidence in clinical development while serving HTA needs after regulatory submission. The SWG plans to explore a few areas in greater depth, such as the use of RWE in HTA and comparing and contrasting estimands and PICO.

In a companion paper (Jen et al., 2025), we introduced the HTA framework, outlined the requirements from different HTA bodies, and addressed operational challenges. Furthermore, we have shared our insights on how to tackle these challenges effectively.

As quantitative scientists involved in medical product development, we acknowledge the significant potential for statisticians to assume leadership roles and contribute to the HTA submission process. This article, focusing on study design and statistical methodologies, serves as evidence of the critical roles statisticians can play in the HTA process, ranging from strategy development and methodology design to evidence generation. Along with the companion paper (Jen et al., 2025), we aim to provide a comprehensive overview of the current state of HTA evaluations and offer insights into addressing the challenges faced in this field.

ACKNOWLEDGEMENTS

We express our gratitude to the American Statistical Association Biopharmaceutical Section for supporting the Health Technology Assessment Scientific Working Group, which contributed to the development of this manuscript. We also extend our thanks to the two reviewers and the Associate Editor for their constructive comments, which significantly enhanced the manuscript's content.

Disclosure Statement

The authors report there are no competing interests to declare. The views expressed are those of the authors and not necessarily those of their affiliations.

Reference

Abrams, K. (2020). CHTE2020 Sources and Synthesis of Evidence; Update to Evidence Synthesis Methods. *NICE Decision Support Unit, Sheffield, UK*.

Aggarwal, S., & H, T. (2013). *Role of Subgroup Analyses for Health Technology Assessments* (Value in Health Issue.

Akehurst, R., Murphy, L. A., Sola-Morales, O., Cunningham, D., Mestre-Ferrandiz, J., & de Pouvourville, G. (2023). Using Real-World Data in the Health Technology Assessment of Pharmaceuticals: Strengths, Difficulties, and a Pragmatic Way Forward. *Value Health*, *26*(4S), 11-19. https://doi.org/10.1016/j.jval.2023.01.010

Appiah, K., Rizzo, M., Sarri, G., & Hernandez, L. (2024). Justifying the source of external comparators in single-arm oncology health technology submissions: a review of NICE and PBAC assessments. *J Comp Eff Res*, *13*(2), e230140. https://doi.org/10.57264/cer-2023-0140

Benkeser, D. C., & Hejazi, N. S. (2017). *survtmle: Targeted minimum loss-based estimation for survival analysis in R.* . https://github.com/benkeser/survtmle.URL

Bucher, H. C., Guyatt, G. H., Griffith, L. E., & Walter, S. D. (1997). The results of direct and indirect treatment comparisons in meta-analysis of randomized controlled trials. *Journal of clinical epidemiology*, *50*(6), 683-691.

Chen, D., Petersen, M. L., Rytgaard, H. C., Grøn, R., Lange, T., Rasmussen, S., Pratley, R. E., Marso, S. P., Kvist, K., Buse, J., & van der Laan, M. J. (2023). Beyond the Cox Hazard Ratio: A Targeted Learning Approach to Survival Analysis in a Cardiovascular Outcome Trial Application. *Statistics in Biopharmaceutical Research*, *15*(3), 524-539. https://doi.org/10.1080/19466315.2023.2173644

Curtis, L. H., Sola-Morales, O., Heidt, J., Saunders-Hastings, P., Walsh, L., Casso, D., Oliveria, S., Mercado, T., Zusterzeel, R., Sobel, R. E., Jalbert, J. J., Mastey, V., Harnett, J., & Quek, R. G. W. (2023). Regulatory and HTA Considerations for Development of Real-World Data Derived External Controls. *Clinical Pharmacology & Therapeutics*, *114*(2), 303-315. https://doi.org/10.1002/cpt.2913

de Pouvourville, G., Armoiry, X., Lavorel, A., Bilbault, P., Maugendre, P., Bensimon, L., Beziz, D., Blin, P., Borget, I., Bouee, S., Collignon, C., Dervaux, B., Durand-Zaleski, I., Julien, M., de Leotoing, L., Majed, L., Martelli, N., Sejourne, T., & Viprey, M. (2023). Real-world data and evidence in health technology assessment: When are they complementary, substitutes, or the only sources of data compared to clinical trials? *Therapie*, *78*(1), 81-94. https://doi.org/10.1016/j.therap.2022.11.001

Dias, S., Ades, A. E., Welton, N. J., Jansen, J. P., & Sutton, A. J. (2018). *Network meta-analysis for decision-making*. John Wiley & Sons.

Dias, S., Sutton, A. J., Ades, A., & Welton, N. J. (2013). Evidence synthesis for decision making 2: a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical decision making*, *33*(5), 607-617.

Dunger-Baldauf, C., Hemmings, R., Bretz, F., Jones, B., Schiel, A., & Holmes, C. (2023). Generating the Right Evidence at the Right Time: Principles of a New Class of Flexible Augmented Clinical Trial Designs. *Clinical Pharmacology & Therapeutics*, *113*(5), 1132-1138. https://doi.org/10.1002/cpt.2869

EFPIA. (2024). *EU HTA Regulation for oncology medicines: Learnings from a simula-on on the impact of proposed EUnetHTA21 methods* European Federation of Pharmaceutical Industries and Associations. https://www.efpia.eu/media/qrjah2ij/efpia-evidera-research-on-eunethta21-methods.pdf

Eichler, H. G., Abadie, E., Breckenridge, A., Flamion, B., Gustafsson, L. L., Leufkens, H., Rowland, M., Schneider, C. K., & Bloechl-Daum, B. (2011). Bridging the efficacy-effectiveness gap: a regulator's perspective on addressing variability of drug response. *Nat Rev Drug Discov*, *10*(7), 495-506. https://doi.org/10.1038/nrd3501

EMA. (2019). *Authorization of medicines* European Medicines Agency. Retrieved 5.16.2024 from https://www.ema.europa.eu/en/about-us/what-we-do/authorisation-medicines#who-makes-decisions-on-patient-access-to-medicines-10944

EU. (2024). *Draft regulation and its appendix on the joint clinical assessment framework*. European Commission. Retrieved Aug 15 from https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/13708-Health-technology-assessment-joint-clinical-assessments-of-medicinal-products_en.

EUnetHTA. (n.d.). *PICO FAQ.* Retrieved August 15 from https://www.eunethta.eu/pico/

European Commission Directorate-General for Health and Food Safety (2024). *Guidance on the validity of clinical studies for joint clinical assessments*. European Commission Retrieved from https://health.ec.europa.eu/publications/guidance-validity-clinical-studies-joint-clinical-assessments_en

Faria, R., Hernandez Alava, M., Manca, A., & Wailoo, A. J. (2015). *NICE DSU Technical Support Document 17: The use of observational data to inform estimates of treatment effectiveness for Technology Appraisal: Methods for comparative individual patient data*. http://www.nicedsu.org.uk

Faulkner, S. D., Sayuri Ii, S., Pakarinen, C., Somers, F., Jose Vicente Edo, M., Prieto Remon, L., Diaz Ponce, A., Gove, D., Ferrer, E., Nafria, B., Bertelsen, N., Boudes, M., Brooke, N., Moutet, A., & Fahy, N. (2021). Understanding multi-stakeholder needs, preferences and expectations to define effective practices and processes of patient engagement in medicine development: A mixed-methods study. *Health Expect*, *24*(2), 601-616. https://doi.org/10.1111/hex.13207

FDA Guidance. (2009). *Patient-Reported Outcome Measures: Use in Medical Product Development to Support Labeling Claims*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-reported-outcome-measures-use-medical-product-development-support-labeling-claims

FDA Guidance. (2018). *Patient-Focused Drug Development: Collecting Comprehensive and Representative Input*. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/patient-focused-drug-development-collecting-comprehensive-and-representative-input

Fontrier, A.-M., Visintin, E., & Kanavos, P. (2022). Similarities and differences in health technology assessment systems and implications for coverage decisions: evidence from 32 countries. *PharmacoEconomics-open*, 1-14.

Freeman, S. C., & Carpenter, J. R. (2017). Bayesian one-step IPD network meta-analysis of time-to-event data using Royston-Parmar models. *Res Synth Methods*, *8*(4), 451-464. https://doi.org/10.1002/jrsm.1253

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press. https://doi.org/DOI: 10.1017/CBO9780511790942

Guyot, P., Ades, A. E., Ouwens, M. J., & Welton, N. J. (2012). Enhanced secondary analysis of survival data: reconstructing the data from published Kaplan-Meier survival curves. *BMC Med Res Methodol*, *12*, 9. https://doi.org/10.1186/1471-2288-12-9

Hampson, L. V., Bornkamp, B., Holzhauer, B., Kahn, J., Lange, M. R., Luo, W. L., Cioppa, G. D., Stott, K., & Ballerstedt, S. (2022). Improving the assessment of the probability of success in late stage drug development. *Pharm Stat*, 21(2), 439-459. https://doi.org/10.1002/pst.2179

Hampson, L. V., Chu, J., Zia, A., Zhang, J., Hsu, W. C., Parzynski, C. S., … Degtyarev, E. (2023). Combining the Target Trial and Estimand Frameworks to Define the Causal Estimand: An Application Using Real-World Data to Contextualize a Single-Arm Trial. Statistics in Biopharmaceutical Research, 16(1), 1–10. https://doi.org/10.1080/19466315.2023.2190931

Hernan , H, Robin, J. Using big data to emulate a target trial when a randomized trial is not available (2016). *American Journal of Epidemiology,* 183(8), 758-764

Hoaglin, D. C., Hawkins, N., Jansen, J. P., Scott, D. A., Itzler, R., Cappelleri, J. C., Boersma, C., Thompson, D., Larholt, K. M., Diaz, M., & Barrett, A. (2011). Conducting Indirect-Treatment-Comparison and Network-Meta-Analysis Studies: Report of the ISPOR Task Force on Indirect Treatment Comparisons Good Research Practices: Part 2. *Value in Health*, *14*(4), 429-437. https://doi.org/https://doi.org/10.1016/j.jval.2011.01.011

HTA Coordination Group. (2024). *Practical Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons*.

ICER. (2022). *An assessment of gene therapies, including etranacogene dezaparvovec and valoctocogene roxaparvovec*. Institute for Clinical and Economic Review. https://icer.org/assessment/hemophilia-a-and-b-2022

ICER. (2023). *Gene Therapies for Sickle Cell Disease*. Institute for Clinical and Economic Review. https://icer.org/wp-content/uploads/2023/08/ICER_SCD_Final_Report_FOR_PUBLICATION_082123.pdf

ICON. (2021). *Ensuring scientific rigor in external control arms: Methodologies to support HTA decision making*.

IQVIA. (2022). *The future of EU HTA - How pharmaceutical companies can prepare for the new process*. IQVIA. Retrieved April 15 from https://www.iqvia.com/library/white-papers/the-future-of-eu-hta

IQWiG. (2023). *Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation*. IQWiG. Retrieved August 15 from

Jaksa, A., Louder, A., Maksymiuk, C., Vondeling, G. T., Martin, L., Gatto, N., Richards, E., Yver, A., & Rosenlund, M. (2022). A Comparison of Seven Oncology External Control Arm Case Studies: Critiques From Regulatory and Health Technology Assessment Agencies. *Value Health*, *25*(12), 1967-1976. https://doi.org/10.1016/j.jval.2022.05.016

Jansen, J. P. (2011). Network meta-analysis of survival data with fractional polynomials. *BMC Med Res Methodol*, *11*, 61. https://doi.org/10.1186/1471-2288-11-61

Jen, M, Mt-Isa, S, Chu, H, Bridges, I, Cai, Z, Dunger-Baldau, C, Guan J, Ma, J, Ren, M., Skipka, G., Wang, Z., He, W (2024). The current landscape of HTA framework and key challenges. Submitted to SBR

Latimer, N. R. (2011). *NICE DSU Technical Support Document 14: Undertaking survival analysis for economic evaluations alongside clinical trials - extrapolation with patient-level data*. http://www.nicedsu.org.uk

Latimer, N. R., & Abrams, K. R. (2014). *NICE DSU Technical Support Document 16: Adjusting survival time estimates in the presence of treatment switching*. www.nicedsu.org.uk

Latimer, N. R., Abrams, K. R., & Siebert, U. (2019). Two-stage estimation to adjust for treatment switching in randomised trials: a simulation study investigating the use of inverse probability weighting instead of re-censoring. *BMC Medical Research Methodology*, *19*(1), 69. https://doi.org/10.1186/s12874-019-0709-9

Latimer, N. R., Henshall, C., Siebert, U., & Bell, H. (2016). Treatment Switching: Statistical and Decision-Making Challenges and Approaches. *Int J Technol Assess Health Care*, *32*(3), 160-166. https://doi.org/10.1017/S026646231600026X

Le-Rademacher, J., Gunn, H., Yao, X., & Schaid, D. J. (2023). Clinical Trials Overview: From Explanatory to Pragmatic Clinical Trials. *Mayo Clin Proc*, *98*(8), 1241-1253. https://doi.org/10.1016/j.mayocp.2023.04.013

Leather, D. A., Jones, R., Woodcock, A., Vestbo, J., Jacques, L., & Thomas, M. (2020). Real-World Data and Randomised Controlled Trials: The Salford Lung Study. *Adv Ther*, *37*(3), 977-997. https://doi.org/10.1007/s12325-019-01192-1

Loudon, K., Treweek, S., Sullivan, F., Donnan, P., Thorpe, K. E., & Zwarenstein, M. (2015). The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ*, *350*, h2147. https://doi.org/10.1136/bmj.h2147

Lumley, T. (2002). Network meta-analysis for indirect treatment comparisons. *Statistics in Medicine*, *21*(16), 2313-2324. https://doi.org/https://doi.org/10.1002/sim.1201

Makady, A., Ham, R. T., de Boer, A., Hillege, H., Klungel, O., Goettsch, W., & GetReal, W. (2017). Policies for Use of Real-World Data in Health Technology Assessment (HTA): A Comparative Study of Six HTA Agencies. *Value Health*, *20*(4), 520-532. https://doi.org/10.1016/j.jval.2016.12.003

Makady, A., van Veelen, A., Jonsson, P., Moseley, O., D'Andon, A., de Boer, A., Hillege, H., Klungel, O., & Goettsch, W. (2018). Using Real-World Data in Health Technology Assessment (HTA) Practice: A Comparative Study of Five HTA Agencies. *Pharmacoeconomics*, *36*(3), 359-368. https://doi.org/10.1007/s40273-017-0596-z

Michiels, H., Vandebosch, A., & Vansteelandt, S. (2024). Adjusting for time-varying treatment switches in randomized clinical trials: the danger of extrapolation and how to address it. *Statistics in Biopharmaceutical Research*, 1-13.

Morimoto, T., & Fukui, T. (2002). Utilities measured by rating scale, time trade-off, and standard gamble: review and reference for health care professionals. (0917-5040 (Print)).

NICE. (2023). *Etranacogene dezaparvovec for treating moderately severe or severe haemophilia B*. https://www.nice.org.uk/guidance/gid-ta10699/documents/draft-guidance

Paget, M. A., Chuang-Stein, C., Fletcher, C., & Reid, C. (2011). Subgroup analyses of clinical effectiveness to support health technology assessments. *Pharm Stat*, *10*(6), 532-538. https://doi.org/10.1002/pst.531

Pan, T., Zhang, X., Shen, W., & Ye, T. (2024). A Bayesian Approach for Selecting Relevant External Data (BASE): Application to a study of Long-Term Outcomes in a Hemophilia Gene Therapy Trial. *arXiv preprint arXiv:2403.13260*.

Pan, W., & Bai, H. (2015). Propensity score interval matching: using bootstrap confidence intervals for accommodating estimation errors of propensity scores. (1471-2288 (Electronic)).

Patel, D., Grimson, F., Mihaylova, E., Wagner, P., Warren, J., van Engen, A., & Kim, J. (2021). Use of External Comparators for Health Technology Assessment Submissions Based on

Single-Arm Trials. *Value Health*, *24*(8), 1118-1125. https://doi.org/10.1016/j.jval.2021.01.015

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., & Welton, N. J. (2016). *NICE DSU Technical Support Document 18: Methods for population-adjusted indirect comparisons in submission to NICE*. http://www.nicedsu.org.uk

Phillippo, D. M., Ades, A. E., Dias, S., Palmer, S., Abrams, K. R., & Welton, N. J. (2018). Methods for population-adjusted indirect comparisons in health technology appraisal. *Medical decision making*, *38*(2), 200-211.

Phillippo, D. M., Dias, S., Ades, A. E., Belger, M., Brnabic, A., Schacht, A., Saure, D., Kadziola, Z., & Welton, N. J. (2020). Multilevel network meta-regression for population-adjusted treatment comparisons. *J R Stat Soc Ser A Stat Soc*, *183*(3), 1189-1210. https://doi.org/10.1111/rssa.12579

Phillippo, D. M., Dias, S., Ades, A. E., & Welton, N. J. (2020). Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study. *Stat Med*, *39*(30), 4885-4911. https://doi.org/10.1002/sim.8759

Phillippo, D. M., Dias, S., Welton, N. J., & Ades, A. (2024). Multilevel network meta-regression for general likelihoods: synthesis of individual and aggregate data with applications to survival analysis. *arXiv preprint arXiv:2401.12640*.

Pruzek, R. M., & Helmreich, J. E. (2009). Enhancing dependent sample analyses with graphics. *Journal of Statistics Education*, *17*(1).

Ren, S., Oakley, J. E., & Stevens, J. W. (2018). Incorporating Genuine Prior Information about Between-Study Heterogeneity in Random Effects Pairwise and Network Meta-analyses. (1552-681X (Electronic)).

Rheault, M. N., Alpers, C. E., Barratt, J., Bieler, S., Canetta, P., Chae, D.-W., Coppock, G., Diva, U., Gesualdo, L., & Heerspink, H. J. (2023). Sparsentan versus irbesartan in focal segmental glomerulosclerosis. *New England Journal of Medicine*, *389*(26), 2436-2445.

Robins, J. M., & Finkelstein, D. M. (2000). Correcting for noncompliance and dependent censoring in an AIDS Clinical Trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics*, *56*(3), 779-788. https://doi.org/10.1111/j.0006-341x.2000.00779.x

RTI Health Solutions. (2022). *Summary of Guidance on Health-Utility Measures by Selected Health Technology Assessment* https://www.rtihs.org/content/guidelines-methods-measurement-health-utilities

Robins, J. M., & Tsiatis, A. A. (1991). Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics - Theory and Methods, 20(8), 2609–2631. https://doi.org/10.1080/03610929108830654.

Rutherford, M., Lambert, P., Sweeting, M., Pennington, R., Crowther, M., Abrams, K., & Latimer, N. (2020). *NICE DSU Technical Support Document 21. Flexible Methods for Survival Analysis*. http://www.nicedsu.org.uk

Sekhon, J. S. (2011). Multivariate and Propensity Score Matching Software with Automated Balance Optimization: The Matching package for R. *Journal of Statistical Software*, *42*(7), 1 - 52. https://doi.org/10.18637/jss.v042.i07

Shah, J. A.-O., Kim, H. A.-O., Sivamurthy, K. A.-O., Monahan, P. A.-O., & Fries, M. A.-O. (2023). Comprehensive analysis and prediction of long-term durability of factor IX activity following etranacogene dezaparvovec gene therapy in the treatment of hemophilia B. *Current Medical Research and Opinion 39*(1473-4877 (Electronic)), 227-237. https://doi.org/10.1080/03007995.2022.2133492

Sharma, N. S. (2015). Patient centric approach for clinical trials: Current trend and new opportunities. *Perspect Clin Res*, *6*(3), 134-138. https://doi.org/10.4103/2229-3485.159936

Signorovitch, J. E., Sikirica, V., Erder, M. H., Xie, J., Lu, M., Hodgkins, P. S., Betts, K. A., & Wu, E. Q. (2012). Matching-adjusted indirect comparisons: a new tool for timely comparative effectiveness research. *Value in Health*, *15*(6), 940-947.

Sola-Morales, O., Curtis, L. H., Heidt, J., Walsh, L., Casso, D., Oliveria, S., Saunders-Hastings, P., Song, Y., Mercado, T., Zusterzeel, R., Mastey, V., Harnett, J., & Quek, R. G. W. (2023). Effectively Leveraging RWD for External Controls: A Systematic Literature Review of Regulatory and HTA Decisions. *Clinical Pharmacology & Therapeutics*, *114*(2), 325-355. https://doi.org/10.1002/cpt.2914

Stürmer, T., Joshi M Fau - Glynn, R. J., Glynn Rj Fau - Avorn, J., Avorn J Fau - Rothman, K. J., Rothman Kj Fau - Schneeweiss, S., & Schneeweiss, S. (2006). A review of the application of propensity score methods yielded increasing use, advantages in specific settings, but not substantially different estimates compared with conventional multivariable methods. (0895-4356 (Print)).

Tingle, S. J., Kourounis, G., Liu, F., Wilson, C. H., Richardson, R., & Sandercock, J. (2024). Simple tool implementing Bucher's method for indirect treatment comparisons in meta-analyses. *BMJ Evidence-Based Medicine*, bmjebm-2024-113103. https://doi.org/10.1136/bmjebm-2024-113103

Tunis, S. R., & Turkelson, C. (2012). Using health technology assessment to identify gaps in evidence and inform study design for comparative effectiveness research. *J Clin Oncol*, *30*(34), 4256-4261. https://doi.org/10.1200/JCO.2012.42.6338

Vreman, R. A., Heikkinen, I., Schuurman, A., Sapede, C., Garcia, J. L., Hedberg, N., Athanasiou, D., Grueger, J., Leufkens, H. G., & Goettsch, W. G. (2019). Unmet medical need: an introduction to definitions and stakeholder perceptions. *Value in Health*, *22*(11), 1275-1282.

Wailoo, A., Hernandez Alava, M., & Pudney, S. (2023). *NICE DSU Technical Support Document 22 Mapping to estimate health state utilities*. http://www.nicedsu.org.uk

Wang, T., McAuslane, N., Liberti, L., Leufkens, H., & Hövels, A. (2018). Building synergy between regulatory and HTA agencies beyond processes and procedures—can we effectively align the evidentiary requirements? A survey of stakeholder perceptions. *Value in Health*, *21*(6), 707-714.

Wijn, S. R. W., Rovers, M. M., Le, L. H., Belias, M., Hoogland, J., IntHout, J., Debray, T., & Reitsma, J. B. (2019). Guidance from key organisations on exploring, confirming and interpreting subgroup effects of medical treatments: a scoping review. *BMJ Open*, *9*(8), e028751. https://doi.org/10.1136/bmjopen-2018-028751

Woods, B. S., Hawkins N Fau - Scott, D. A., & Scott, D. A. (2010). Network meta-analysis on the log-hazard scale, combining count and hazard ratio statistics accounting for multi-arm trials: a tutorial. (1471-2288 (Electronic)).

Yateman, N. (2022). *Challenges in developing next generation CAR-T* Basel Biometrics Seminar, https://baselbiometrics.github.io/home/docs/talks/20220715/3_Yateman.pdf

Ying, A., & Tchetgen, E. J. T. (2023). Structural cumulative survival models for estimation of treatment effects accounting for treatment switching in randomized experiments. *Biometrics*, *79*(3), 1597-1609. https://doi.org/10.1111/biom.13704

Zisis, K., Pavi, E., Geitona, M., & Athanasakis, K. (2024). Real-world data: a comprehensive literature review on the barriers, challenges, and opportunities associated with their inclusion in the health technology assessment process. *J Pharm Pharm Sci*, *27*, 12302. https://doi.org/10.3389/jpps.2024.12302

Zuidgeest, M. G. P., Goetz, I., Groenwold, R. H. H., Irving, E., van Thiel, G., Grobbee, D. E., & GetReal Work, P. (2017). Series: Pragmatic trials and real world evidence: Paper 1. Introduction. *J Clin Epidemiol*, *88*, 7-13. https://doi.org/10.1016/j.jclinepi.2016.12.023