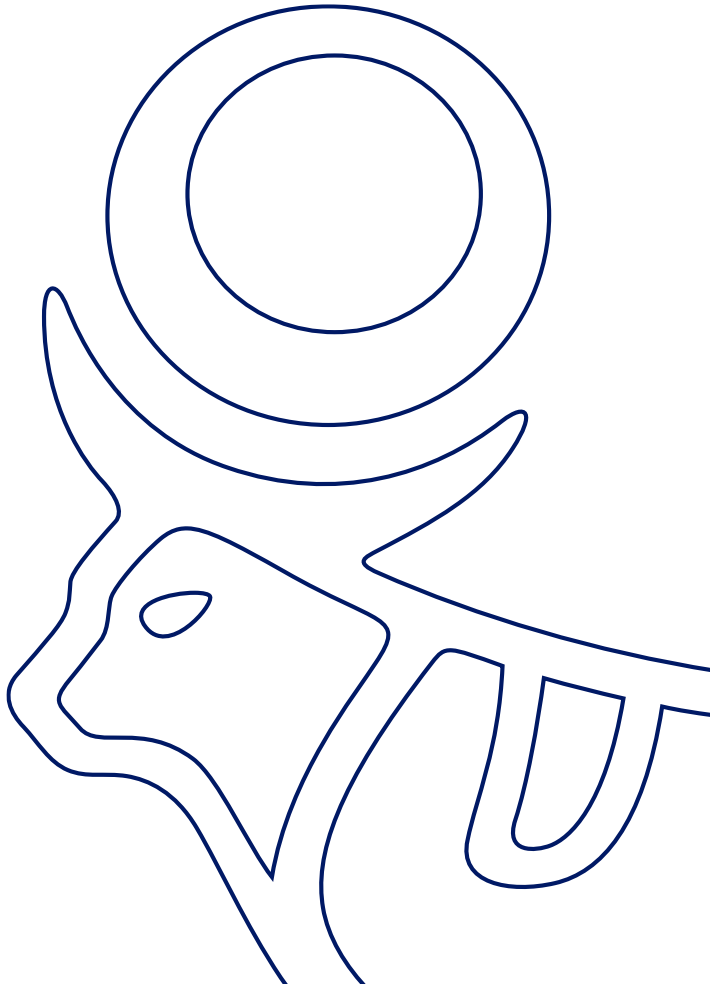
UNORDINARY DRIVES CHANGE

novo nordisk®

Data Fusion for Indirect Treatment Comparisons in Health Technology Assessment



Antonio Remiro-Azócar, PhD

Medical and Translational Science, Novo Nordisk

Joint Initiative for Causal Inference Seminar

2nd December 2025

Disclaimers

The views, findings and conclusions expressed in this presentation are solely those of the presenter, who is responsible for the contents of the presentation

The views, findings and conclusions in this presentation are not necessarily those of Novo Nordisk A/S

No statement in this presentation should be construed as an official position of Novo Nordisk A/S

Agenda

Topic	
Part 1: Health technology assessment and indirect treatment comparisons	15 min
Part 2: Link to external control arms	15 min
Part 3: Link to transportability with example	15 min
Part 4: Moving forward	5 min
Discussion, Q&A audience	10 min

Health technology assessment

Health technology assessment (HTA)

What is **health technology assessment**?

- A **health technology** is an intervention used to promote health, e.g., a pharmaceutical product or a medical device
- An **assessment** is required to inform policy decision-making
- The assessment is **multidisciplinary**, involving clinical, social, economic, organizational and ethical aspects

Regulatory versus HTA statistics

REGULATORY

- Focus on safety, clinical efficacy, quality, benefit-risk
- Focus on hypothesis testing
- Estimands
- Time-horizon is the trial follow-up; does not typically require extrapolation
- Relies mostly on “pivotal” Phase III clinical trial data as the primary source of evidence
- **Relies on a direct comparison to placebo or standard of care in a head-to-head trial**

HTA

- Focus on the “fourth hurdle”: clinical effectiveness (value) and cost-effectiveness (value-for-money)
- Focus on estimation
- PICOs
- Long-term or “lifetime” horizon; may require extrapolation beyond the trial follow-up period
- Requires use of secondary data sources beyond the “pivotal” clinical trial(s)
- **Comparators are all treatment options in clinical practice; indirect comparisons are required**

The PICO framework

- In HTA, the PICO framework is used to translate local policy questions into research questions
- Relevant PICO question(s) are specified in the HTA scoping process
- Health technology developers submit an evidence dossier to HTA agencies addressing the PICO question(s) in the scope

EU Joint Clinical Assessment "Guidance on the scoping process"

1 Introduction

1.1 The assessment scope

The basis of a Health Technology Assessment (HTA) is a set of defined research questions that are to be answered by the assessment and that together define the assessment scope. In the context of the Joint Clinical Assessment (JCA), the assessment scope reflects policy questions from the different healthcare systems in which the JCA will be used. The PICO framework provides a standard format for specifying research questions, detailing the following parameters:

- P (population),
- I (intervention),
- C (comparator[s]),
- O (outcomes).

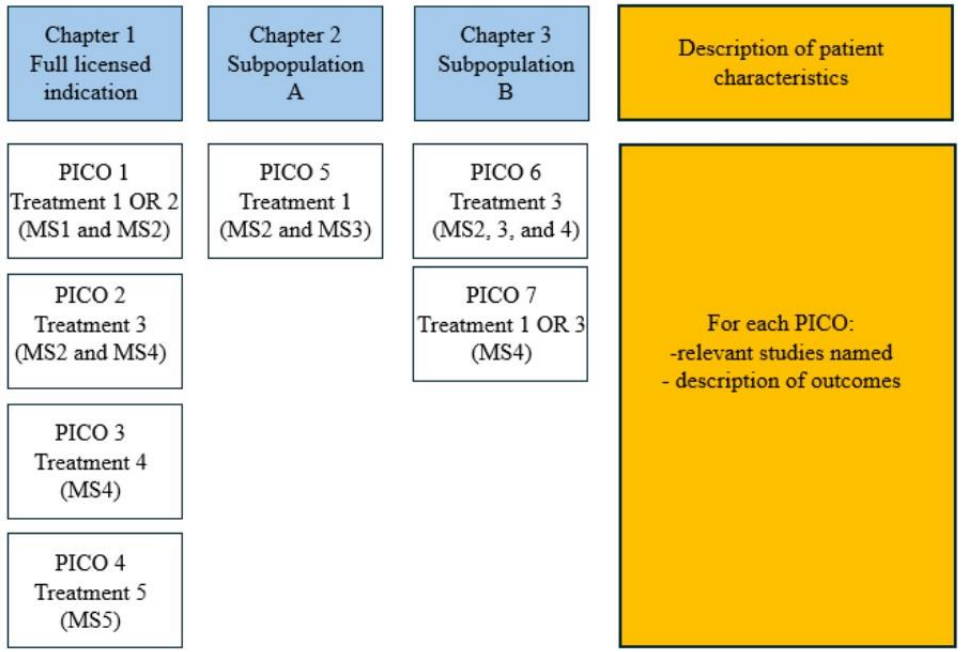
The new EU HTA Regulation (2025)

- The EU HTA Regulation (coming into place 2025-30) introduces a centralized framework for the Joint Clinical Assessment (JCA) of new medicines
- The JCA scoping process is inclusive and aims to meet the diverse evidence needs (PICOs) of all 27 EU member states simultaneously
- There is often limited consensus on **comparators** and **(sub) populations** across states, due to variation in clinical practice and reimbursement
- A multitude of PICOs is anticipated

EU Joint Clinical Assessment "Guidance on the scoping process"

Table 9: Consolidated PICOs based on Member State requests

	PICO 1	PICO 2	PICO 3	PICO 4	PICO 5	PICO 6	PICO 7
P	Full claimed indication	Full claimed indication	Full claimed indication	Full claimed indication	Subpopula-tion A	Subpopula-tion B	Subpopula-tion B
C	Treatment 1 OR Treatment 2	Treatment 3	Treatment 4	Treatment 5 - Individualised treatment	Treatment 1	Treatment 3	Treatment 1 OR Treatment 3
O	All outcomes	All outcomes	All outcomes	All outcomes	All outcomes	All outcomes	All outcomes



MS: Member State; PICO: Population, Intervention, Comparator(s), Outcomes.

Figure 5: Example of data presentation according to PICO(s)

A multitude of PICOs

- van Engen et al (2024) follow draft EU JCA guidance to determine the number of PICOs for two hypothetical products in two common oncology indications
- There are many target (sub) populations
- There is high variability in PICOs
- A substantial number of PICOs **require indirect treatment comparisons due to unavailable head-to-head comparisons**, some relying on small sample sizes and sparse networks

Table 3. Summary of PICO results

	1L NSCLC		3L MM	
	Base case (EU HTA reports)	**Base case + NICE report**	**Base case (EU HTA reports)**	**Base case + NICE report**
Populations	EMA label + 8 subpopulations	EMA label + 10 subpopulations	EMA label + 4 subpopulations	EMA label + 6 subpopulations
Comparators	9	9	8	9
Outcomes per PICO	28	28	45	45
Number of PICOs (% requested by single country)	10 (50%)	14 (50%)	16 (75%)	18 (78%)
PICOs requiring ITC	5	8	11	12
Number of analyses requested (% indirect analyses)	280 (50%)	392 (57%)	720 (69%)	810 (67%)

Abbreviations: European Medicines Agency; EU: European Union; ITC: indirect treatment comparison; HTA: health technology assessment; MM: multiple myeloma; NICE: National Institute for Health and Care Excellence; NSCLC: non-small cell lung cancer; PICO: population, intervention, comparator, outcome.

van Engen, A., Krüger, R., Parnaby, A., Rotaru, M., Ryan, J., Samaha, D. and Tzelis, D., 2024. The impact of additive population (s), intervention, comparator (s), and outcomes in a European joint clinical health technology assessment. Value in Health, 27(12), pp.1722-1731.

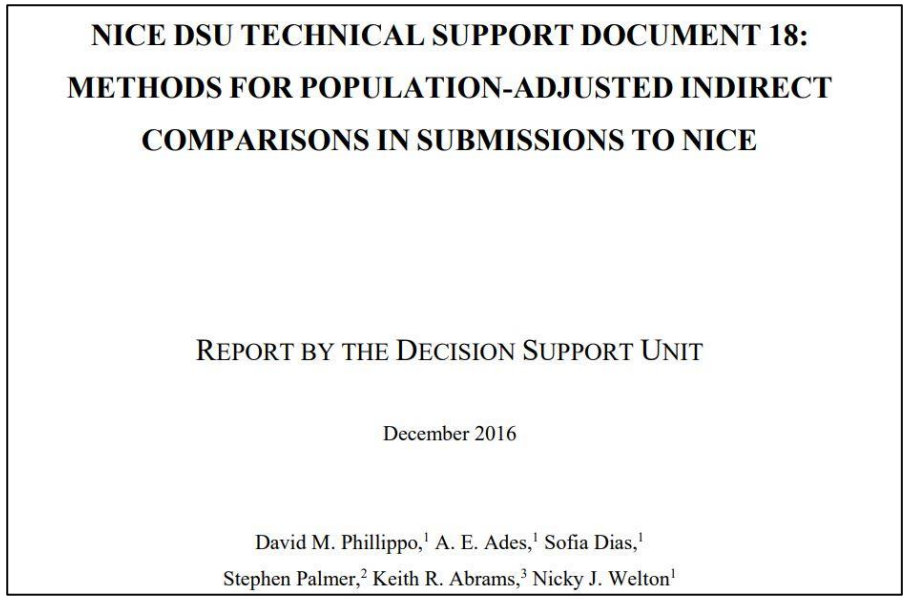
Indirect treatment comparisons

Indirect treatment comparisons (ITCs)

HTA requires comparisons versus all treatment options in routine clinical practice

- The scope of assessments depends on the policy question and is not always driven by the available data
- RCTs cannot have all desired treatment arms, given the number of jurisdictions and variations in clinical practice
- Some therapeutic areas evolve rapidly, with a changing comparator landscape and no single accepted standard-of-care

Indirect treatment comparisons are required


NICE DSU TECHNICAL SUPPORT DOCUMENT 18:
METHODS FOR POPULATION-ADJUSTED INDIRECT COMPARISONS IN SUBMISSIONS TO NICE

REPORT BY THE DECISION SUPPORT UNIT

December 2016

David M. Phillippo,[1] A. E. Ades,[1] Sofia Dias,[1]
Stephen Palmer,[2] Keith R. Abrams,[3] Nicky J. Welton[1]

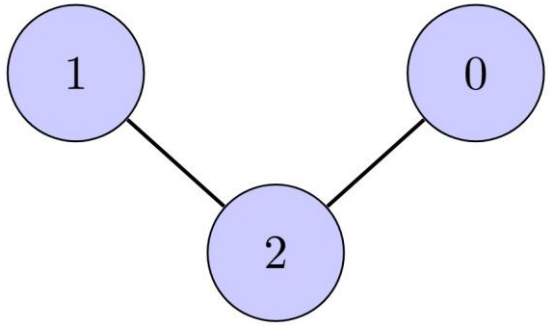

Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons

Adopted on 8 March 2024 by the HTA CG pursuant to Article 3(7), point (d), of Regulation (EU) 2021/2282 on Health Technology Assessment

Indirect treatment comparisons

ITCs can be **anchored** or **unanchored**

ANCHORED COMPARISON



$$\Delta_{10} = \Delta_{12} - \Delta_{02}$$

UNANCHORED COMPARISON

$$\Delta_{10} = g(\mu_1) - g(\mu_0)$$

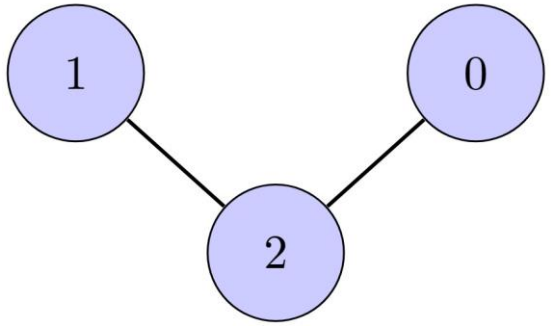
Anchored comparison:

- ”Anchors” the ITC via a common comparator arm: comparison of treatment effects across trials
- Subject to bias where treatment effects are heterogeneous over *effect modifiers* that vary in distribution across trials

Indirect treatment comparisons

ITCs can be **anchored** or **unanchored**

ANCHORED COMPARISON



$$\Delta_{10} = \Delta_{12} - \Delta_{02}$$

UNANCHORED COMPARISON

$$\Delta_{10} = g(\mu_1) - g(\mu_0)$$

Unanchored comparison:

- No compatible comparator for ”anchoring”: comparison of mean absolute outcomes across trials
- Subject to bias where there are cross-trial differences in factors that are *prognostic* of treatment outcomes

The need for covariate adjustment

Unadjusted ITCs…

- Rely on a very strong assumption: unconditional exchangeability across trials
- Produce bias with cross-trial imbalances in effect modifiers and/or prognostic variables
- Do not explain heterogeneity or explicitly produce estimates in any specific target population
- Ignore uncertainty due to cross-trial differences in baseline covariates

Covariate-adjusted ITCs…

- Relax the exchangeability assumption by conditioning on baseline covariates
- Can reduce bias due to cross-trial imbalances in effect modifiers and/or prognostic variables
- Explicitly produce estimates in specific target samples or populations
- Can account for uncertainty due to differences in baseline covariates across trials

Covariate adjustment is desirable for ITCs

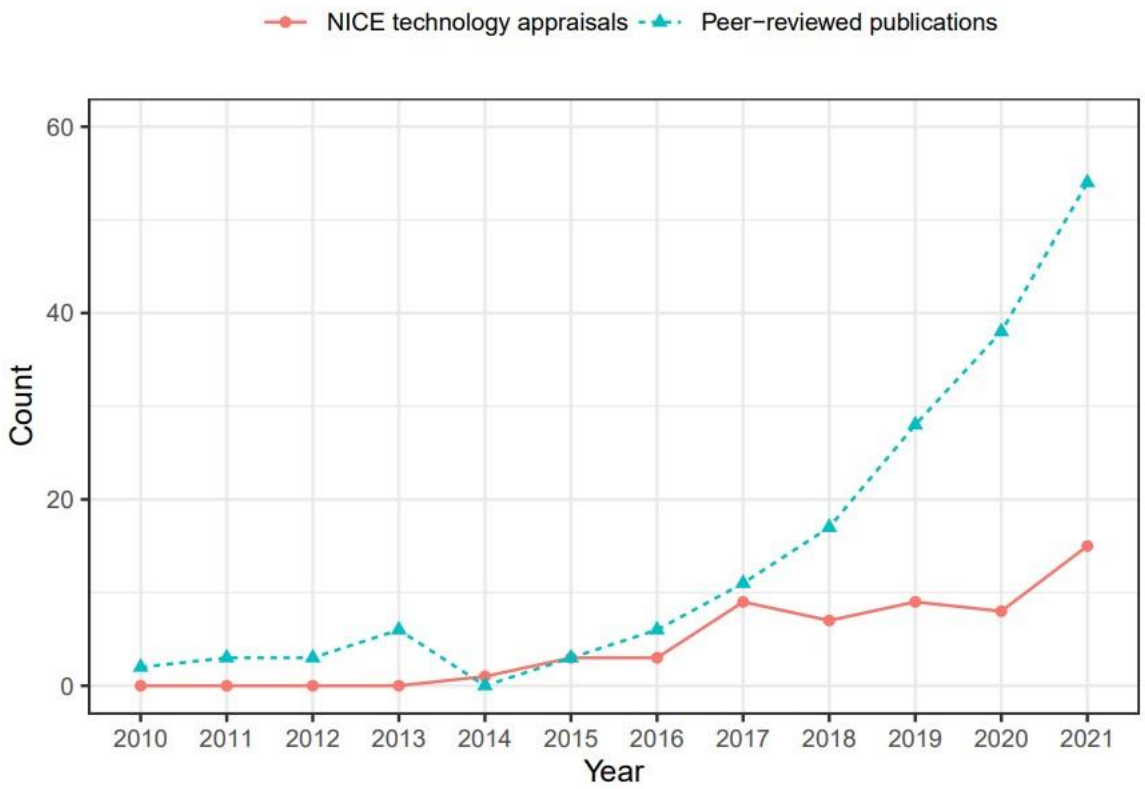
Currently available methodologies

Odds-weighting (entropy balancing)

- Denoted matching-adjusted indirect comparison (MAIC)

Outcome modelling

- Denoted simulated treatment comparison (STC)



Limitations

- Both approaches are singly robust (in most cases) and based on parametric modelling, relying on the correct specification of a single parametric model

Weighting versus outcome modelling

Outcome modelling is more precise and efficient than weighting when assumptions hold, but there are caveats…

ODDS-WEIGHTING (MAIC)

- Does not extrapolate; more “honest” uncertainty quantification
- MAIC is more “bias-robust” than than the standard “inverse weighting” modelling approaches
- Bias easier to diagnose, MAIC (entropy balancing) directly enforces balance in covariate moments
- Extreme weights explicitly manifest high uncertainty
- Feasible weighting solutions may not exist where there is limited covariate overlap

OUTCOME MODELLING (STC)

- Relies on model-based extrapolation to improve precision and efficiency
- Susceptible to bias when extrapolating a mis-specified outcome model
- Bias difficult to detect; outcome model that seems correct in the observed covariate space may extrapolate poorly
- Extrapolation uncertainty not typically accounted for
- Can produce the treatment effect estimates that are required for HTA where there is limited overlap

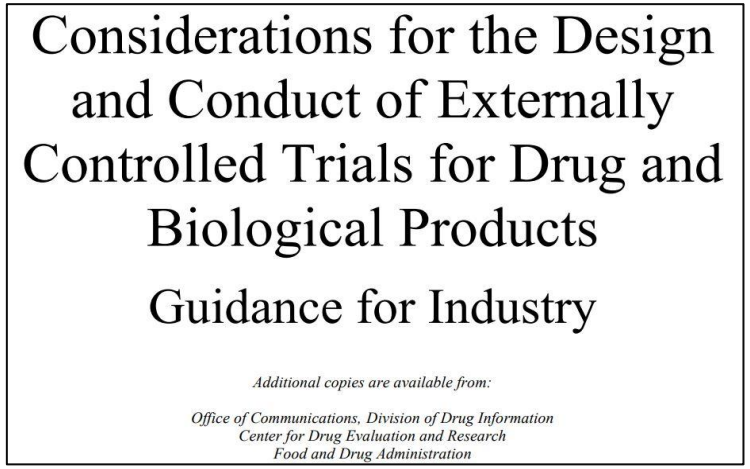
External control arms

Single-arm trials (SATs)

Conducting RCTs might not be possible:

- Where recruitment to RCTs is unfeasible due to small populations, e.g., rare diseases with orphan designation
- For life-threatening conditions with high unmet need and no standard of care, e.g., last-line of therapy in solid tumour oncology
- Where enrolling patients to placebo is unethical, e.g., paediatric trials for treatments with proven efficacy in adults

Regulators recognize that externally controlled SATs might be required in special circumstances



Regulatory submissions featuring externally controlled SATs are rising, mainly through accelerated approval pathways

A common framework

Unanchored ITCs are externally controlled SATs with two “special” characteristics:

- The external control is a competitor-sponsored historical trial
- There may be limited access to subject-level data for the external control, only aggregate-level data from publications

But: “In any situation with non-randomized data, such as observational evidence and single-arm trials, (…) complete access to the individual patient data is required” (Methodological Guideline for Quantitative Evidence Synthesis, EU HTA Coordination Group, 2024)

Different **estimands** (or **summary measures** or **causal contrasts**) can be targeted:

- **Average treatment effect (ATE)** among the combined SAT and external control…somewhat ambiguous here

$$\mathrm{ATE} = g\left(\mathrm{E}\left(Y^1\right)\right) - g\left(\mathrm{E}\left(Y^0\right)\right)$$

- **Average treatment effect in the treated (ATT)** among those participating in the SAT

$$\mathrm{ATT} = g\left(\mathrm{E}(Y^1 \mid S = 1)\right) - g\left(\mathrm{E}\left(Y^0 \mid S = 1\right)\right)$$

- **Average treatment effect in the control (ATC)** among those in the external control group

$$\mathrm{ATC} = g\left(\mathrm{E}(Y^1 \mid S = 0)\right) - g\left(\mathrm{E}(Y^0 \mid S = 0)\right)$$

ATT or ATC?

Difference between the summary measures is driven by them targeting different (sub) populations or “analysis sets”

Average treatment effect in the treated (ATT)

$$\mathrm{ATT} = g\left(\mathrm{E}(Y^1 \mid S = 1)\right) - g\left(\mathrm{E}\left(Y^0 \mid S = 1\right)\right)$$

Attractive for the regulatory context…

- Consistent with the emulation of a randomized comparison in the registrational SAT
- The external control would aim to “mimic” the internal control arm of “pivotal” clinical trial
- Compatible with the mean absolute outcome targeted by the SAT, preserving the original SAT results
- Typically, the primary estimand for externally controlled SATs seeking drug approval in the regulatory environment

Nevertheless…

- Potentially unappealing where generalizability to routine clinical practice is a priority
- SAT populations are often highly selected and may lack representativeness with respect to “real-world” populations

ATT or ATC?

Average treatment effect in the control (ATC)

$$\text{ATC} = g\left(\text{E}(Y^1 \mid S = 0)\right) - g\left(\text{E}(Y^0 \mid S = 0)\right)$$

Typically, the target estimand in HTA…due to necessity as subject-level data have often been unavailable for the external control

Potentially more desirable for external validity…

- External controls based on RWD or natural history studies: broad inclusion criteria targeting heterogeneous populations

…but not necessarily so…

- Historical controls from past clinical trials may not reflect the current standard of care
- RWD-derived external controls based on a single country are not necessarily transferable across jurisdictions

Statistical considerations (effective sample size, precision) may also play a role in the estimand choice, e.g., when weighting

We shall assume ATC is the target estimand for exposition (see back-up slides for ATT as the target estimand)

Four critical assumptions (Zhou et al 2025)

1. No direct effect of trial participation

- Trial participation does not affect the outcome except through treatment assignment itself (no Hawthorne effects)

2. Stable unit treatment value (SUTVA)

- No interference between subjects and treatment variation irrelevance (one well-defined version of each treatment)

$$Y_i = Y_i^1 T_i + Y_i^0 (1 - T_i)$$

3. Conditional ignorability of data source assignment

- ATT: Conditional on covariates, potential outcomes under the control are independent of the data source

$$Y_i^0 \perp S_i \mid \mathbf{X}_i$$

- ATC: Conditional on covariates, potential outcomes under the active intervention are independent of the data source

$$Y_i^1 \perp S_i \mid \mathbf{X}_i$$

4. Overlap or positivity

- ATT: Support of the covariates in the SAT is contained within that of the external control $0 < \Pr(S = 0 \mid \mathbf{X} = \mathbf{x}) < 1, \forall \mathbf{x}\ f(\mathbf{x} \mid S = 1) > 0$
- ATC: Support of the covariates in the external control is contained within that of the SAT $0 < \Pr(S = 1 \mid \mathbf{X} = \mathbf{x}) < 1, \forall \mathbf{x}\ f(\mathbf{x} \mid S = 0) > 0$

STRONG IGNORABILITY

These assumptions are unverifiable and potentially unreasonable in many practical scenarios…proceed with care (Senn 2025)

Contextual differences

EXTERNALLY CONTROLLED SATs (REGULATORY)

- Modelling-based approach to odds weighting
- Outcome modelling: G-computation
- Doubly robust (DR) methods are well established: augmented approaches, TMLE, etc.
- Use of data-adaptive (machine learning) estimators has been explored within a DR framework
- Methodologies assume full access to subject-level data
- Target is typically the ATT

UNANCHORED ITCS (HTA)

- MAIC: entropy balancing-based approach to odds-weighting
- Outcome modelling: Simulated treatment comparison
- Doubly robust augmented approaches, TMLE, yet to be leveraged
- Reliance on the correct specification of a single parametric model
- Methods developed with limited access to subject-level data
- Target is typically the ATC

Weighting: modelling versus balancing

MODELING

- Explicitly models the propensity score as a function of baseline covariates
- Propensity scores are estimated by maximizing the fit of a logistic regression
- Estimated weights do not produce adequate balance if the propensity score model is mis-specified
- Even a correctly specified propensity score model does not guarantee balance in finite samples
- Propensity score predictions that are close to zero produce extreme weights, which lead to imprecision
- Limited applicability with unavailable subject-level covariates for the external control

ENTROPY BALANCING

- Does not explicitly model the propensity score, but implicitly assumes a logistic propensity score model
- Covariate balance viewed as a convex optimization problem
- Less susceptible to bias by directly enforcing covariate balance
- Weights constrained to be positive and sample-bounded (interpolation as opposed to extrapolation)
- Minimally dispersed weights, which translates into larger effective sample sizes and precision
- Applicable where aggregate-level covariate moments are available for the external control

Modelling approach to weighting

Inverse-odds weights (IOW) defined as: $w_i = \frac{(1-e_i)S_i}{e_i} + (1-S_i),$ with $e_i = e(\mathbf{X}_i) = \Pr(S_i = 1 \mid \mathbf{X}_i)$

Weights are inverse conditional odds of SAT participation (conditional odds of external control participation)

A logistic regression is fitted to the concatenated SAT and external control subject-level data, typically using maximum-likelihood estimation, to estimate model-based propensity scores

$$\text{logit}(e_i) = \alpha_0 + \mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\alpha}$$

$$\hat{e}_i = \text{logit}^{-1}\left(\hat{\alpha}_0 + \mathbf{c}(\mathbf{X}_i)^\top \hat{\boldsymbol{\alpha}}\right)$$

Propensity score predictions are plugged into the weight equation to derive weight estimates

The weighted average of observed outcomes under the active intervention is contrasted with the unweighted average of observed outcomes for the external control

$$\widehat{\text{ATC}} = \underbrace{g\left(\frac{\sum_{i=1}^{n_1} \hat{w}_i Y_i}{\sum_{i=1}^{n_1} \hat{w}_i}\right)}_{\hat{\mu}_0^1} - \underbrace{g\left(\frac{1}{n_0}\sum_{i=n_1+1}^{n} Y_i\right)}_{\hat{\mu}_0^0}$$

weights normalized to sum to one to improve finite sample properties and provide more stable and precise estimation

Entropy balancing approach to weighting

Propensity score is not explicitly modelled, but logistic model for ”data source assignment” is assumed

$$\ln(\omega_i) \propto \ln\left(\frac{(1-e_i)}{e_i}\right) = \gamma_0 + \mathbf{c}(\mathbf{X}_i)^\top \boldsymbol{\gamma}$$

Weights proportional to the inverse conditional odds of SAT participation

”Method of moments” to estimate the model while enforcing covariate balance constraint

$$\frac{\sum_{i=1}^{n_1} \exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right) \mathbf{c}^*(\mathbf{X}_i)}{\sum_{i=1}^{n_1} \exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)} = \mathbf{0} \qquad \mathbf{c}^*(\mathbf{X}_i) = \mathbf{c}(\mathbf{X}_i) - \hat{\boldsymbol{\theta}}$$

Solve by minimizing objective function using convex optimization algorithm $Q(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n_1} \exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)$

Weights for SAT estimated as

$$\hat{\omega}_i = \frac{\exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)}{\sum_{i=1}^{n_1} \exp\left(\mathbf{c}^*(\mathbf{X}_i)^\top \hat{\boldsymbol{\gamma}}\right)}$$

ATC estimated as

$$\widehat{\text{ATC}} = g\underbrace{\left(\sum_{i=1}^{n_1} \hat{\omega}_i Y_i\right)}_{\hat{\mu}_0^1} - g\underbrace{\left(\frac{1}{n_0}\sum_{i=n_1+1}^{n} Y_i\right)}_{\hat{\mu}_0^0}$$

Bias-robustness considerations

Modelling approach is consistent if…

- The propensity score model for data source assignment is correctly specified
- That is, the logit of the propensity score (conditional probability of SAT participation) varies linearly with the covariate balance functions

Entropy balancing is consistent if…

- The logit of the conditional probability of external control participation (or SAT participation) **OR** the conditional outcome expectation under the active intervention varies linearly with the covariate balance functions
- For instance, mean-balancing ensures consistency if $\text{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha}$ OR $\mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}$
- Mean- and variance-balancing ensure consistency if $\text{logit}(e_i) = \alpha_0 + \mathbf{X}_i^\top \boldsymbol{\alpha_1} + (\mathbf{X}_i^2)^\top \boldsymbol{\alpha_2}$ OR $\mathrm{E}(Y_i^1 \mid \mathbf{X}_i) = \beta_0 + \mathbf{X}_i^\top \boldsymbol{\beta}_1 + (\mathbf{X}_i^2)^\top \boldsymbol{\beta}_2$

Entropy balancing is **linear doubly robust**:

- “Doubly robust with respect to linear outcome regression and logistic propensity score regression” (Zhao and Percival, 2017)

Entropy balancing claimed more bias-robust as it is consistent under a greater number of distinct data-generating mechanisms

Is entropy balancing doubly robust?

It is rarely plausible that outcomes vary linearly with the covariates

Standard balancing strategies do not allow conjecturing an implicit outcome model flexible enough for double robustness

One could consider balancing other non-linear covariate transformations and interactions, but this is rarely feasible:

- Increasing balancing constraints → more likely that covariate moments fall outside the convex hull of the observed covariate space
- Namely, feasible weighting solutions to the convex optimization problem do not exist (no set of positive weights can enforce balance)
- Increasing balancing constraints → further reductions in effective sample size and precision
- Aggregate data beyond means and variances (e.g., higher-order moments and means of transformed covariates) rarely reported

This motivates the explicit augmentation of the weighting estimators, allowing for a less restrictive outcome model

Augmented entropy balancing

Campbell and Remiro-Azócar (2025)

Postulate a model for the conditional outcome expectation under the active intervention and fit it to the SAT

$$q\left(\mathrm{E}(Y_i^1 \mid \mathbf{X}_i; \boldsymbol{\beta})\right) = m\left(\mathbf{X}_i; \boldsymbol{\beta}\right)$$

Predict potential outcomes for the active intervention for all subjects in the SAT and the external control

$$\hat{Y}_i^{\,1} = q^{-1}\left(m(\mathbf{X}_i; \hat{\boldsymbol{\beta}})\right)$$

The G-computation estimator is augmented with a weighted average of residuals, but using **entropy balancing weights**; the weighted average is the "one-step" correction term for the potential bias of G-computation

$$\begin{aligned}\hat{\mu}_0^1 &= \sum_{i=1}^{n_1} \hat{\omega}_i \left(Y_i - \hat{Y}_i^{\,1}\right) + \frac{1}{n_0} \sum_{i=n_1+1}^{n} \hat{Y}_i^{\,1} \\ &= \sum_{i=1}^{n_1} \hat{\omega}_i \epsilon_i^1 + \frac{1}{n_0} \sum_{i=n_1+1}^{n} \hat{Y}_i^{\,1},\end{aligned}$$

Estimator for the ATC:

$$\widehat{\mathrm{ATC}} = g\underbrace{\left(\sum_{i=1}^{n_1} \hat{\omega}_i \epsilon_i^1 + \frac{1}{n_0} \sum_{i=n_1+1}^{n} \hat{Y}_i^{\,1}\right)}_{\hat{\mu}_0^1} - g\underbrace{\left(\frac{1}{n_0} \sum_{i=n_1+1}^{n} Y_i\right)}_{\hat{\mu}_0^0},$$

Weighted G-computation

Another augmented estimator often claimed to be doubly robust consists of G-computation based on the predictions of a weighted outcome model

$$\hat{\mu}_0^1 = \frac{1}{n_0} \sum_{i=n_1+1}^{n} \hat{Y}_i^1 = \frac{1}{n_0} \sum_{i=n_1+1}^{n} q^{-1}\left(m(\mathbf{X}_i; \hat{\boldsymbol{\beta}}_v)\right)$$

$$\widehat{\mathrm{ATC}} = g\left(\hat{\mu}_0^1\right) - g\underbrace{\left(\frac{1}{n_0} \sum_{i=n_1+1}^{n} Y_i\right)}_{\hat{\mu}_0^0}$$

Note: this is only doubly robust where the outcome model is a GLM with canonical link function! (Gabriel et al 2024)

Results suggest asymptotic equivalence and similar finite-sample performance to the augmented weighting estimator described in the previous slide **for GLMs with canonical link functions** (Gabriel et al 2024, Słoczyński et al 2025)

Simulation study

Data-generating mechanisms

KS1: propensity score and outcome model correctly specified

KS1: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{X}_i, T_i) = \text{expit}(X_{1i} - 1.50X_{2i} + 0.5X_{3i} - 0.5X_{4i} + 1.50T_i - 0.50T_iX_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{X}_i) = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.5X_{i4}).$$

KS2: only propensity score model correctly specified

KS2: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{Z}_i, T_i) = \text{expit}(Z_{1i} - 1.50Z_{2i} + 0.5Z_{3i} - 0.5Z_{4i} + 1.50T_i - 0.50T_iZ_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{X}_i) = \text{expit}(-X_{i1} + 0.5X_{i2} - 0.25X_{i3} - 0.5X_{i4}).$$

$$Z_{i1} = \text{scale}(\exp(X_{i1}/2)),$$
$$Z_{i2} = \text{scale}(X_{i2}^2),$$
$$Z_{i3} = \text{scale}((X_{i1}X_{i3} + 0.6)^3),$$
$$Z_{i4} = \text{scale}((X_{i2} + X_{i4} + 20)^2)$$

KS3: only outcome model correctly specified

KS3: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{X}_i, T_i) = \text{expit}(X_{1i} - 1.50X_{2i} + 0.5X_{3i} - 0.5X_{4i} + 1.50T_i - 0.50T_iX_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{Z}_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.5Z_{i4}).$$

KS4: propensity score and outcome model incorrectly specified

KS4: Y_i is generated from a Bernoulli distribution with

$$\Pr(Y_i = 1 \mid \mathbf{Z}_i, T_i) = \text{expit}(Z_{1i} - 1.50Z_{2i} + 0.5Z_{3i} - 0.5Z_{4i} + 1.50T_i - 0.50T_iZ_{1i})$$

where $T_i = S_i$, and S_i is generated from a Bernoulli distribution with

$$\Pr(S_i = 1 \mid \mathbf{Z}_i) = \text{expit}(-Z_{i1} + 0.5Z_{i2} - 0.25Z_{i3} - 0.5Z_{i4}).$$

Target estimand will be the ATC

Variance estimation for all methods using non-parametric bootstrap

KS1: both models correctly specified

- The naïve estimator is biased
- All covariate-adjusted estimators are virtually unbiased under *n*=1000
- Some small-sample bias, even for theoretically consistent estimators, under *n*=200
- G-computation exhibits the greatest precision, but augmented weighting estimators are almost as precise

Method	Bias	ESE	95% CI coverage	Average 95% CI width
$n = 200$				
1. The naïve estimator	0.618	0.328	0.539	1.292
2. IOW with weights from modeling	0.024	0.528	0.939	1.978
3. IOW with normalized weights from modeling	0.049	0.456	0.944	1.763
4. MAIC	0.033	0.420	0.959	2.241
5. G-computation	0.016	0.350	0.955	1.430
6. DR with "modeling" IOW weights	0.029	0.421	0.954	1.667
7. DR with normalized "modeling" IOW weights	0.029	0.414	0.948	1.610
8. DR with MAIC weights	0.029	0.412	0.953	1.713
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.027	0.404	0.940	1.583
10. Augmented "weighted G-computation" with MAIC weights	0.026	0.406	0.943	1.740
$n = 1000$				
1. The naïve estimator	0.604	0.143	0.009	0.561
2. IOW with weights from modeling	0.009	0.205	0.950	0.806
3. IOW with normalized weights from modeling	0.010	0.196	0.941	0.750
4. MAIC	0.006	0.171	0.942	0.659
5. G-computation	0.003	0.150	0.949	0.592
6. DR with "modeling" IOW weights	0.005	0.174	0.946	0.675
7. DR with normalized "modeling" IOW weights	0.005	0.174	0.946	0.666
8. DR with MAIC weights	0.005	0.169	0.941	0.651
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.004	0.169	0.942	0.648
10. Augmented "weighted G-computation" with MAIC weights	0.004	0.169	0.940	0.649

DR: doubly robust; IOW: inverse-odds weighting: MAIC: matching-adjusted indirect comparison

KS2: PS model correctly specified

- G-computation exhibits bias
- Non-augmented and augmented weighting estimators are unbiased for *n*=1000 (weight normalization improves precision)
- Some small-sample bias, even for theoretically consistent weighting estimators, under *n*=200
- Outcome model misspecification does not induce a loss of precision for the augmented estimators compared to their non-augmented counterparts

Method	Bias	ESE	95% CI coverage	Average 95% CI width
$n = 200$				
1. The naïve estimator	0.223	0.322	0.910	1.275
2. IOW with weights from modeling	0.022	0.665	0.929	2.386
3. IOW with normalized weights from modeling	0.052	0.515	0.938	1.954
4. MAIC	0.052	0.501	0.960	2.747
5. G-computation	0.081	0.436	0.951	1.731
6. DR with "modeling" IOW weights	0.043	0.542	0.947	2.100
7. DR with normalized "modeling" IOW weights	0.043	0.520	0.941	2.003
8. DR with MAIC weights	0.039	0.490	0.950	2.044
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.049	0.480	0.936	1.875
10. Augmented "weighted G-computation" with MAIC weights	0.028	0.480	0.946	2.198
$n = 1000$				
1. The naïve estimator	0.220	0.141	0.665	0.556
2. IOW with weights from modeling	0.012	0.277	0.946	1.077
3. IOW with normalized weights from modeling	0.007	0.221	0.938	0.837
4. MAIC	0.006	0.205	0.934	0.777
5. G-computation	0.067	0.188	0.936	0.734
6. DR with "modeling" IOW weights	0.005	0.226	0.941	0.865
7. DR with normalized "modeling" IOW weights	0.006	0.224	0.937	0.848
8. DR with MAIC weights	0.005	0.205	0.936	0.775
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.009	0.202	0.935	0.778
10. Augmented "weighted G-computation" with MAIC weights	0.005	0.203	0.935	0.769

DR: doubly robust; IOW: inverse-odds weighting: MAIC: matching-adjusted indirect comparison

KS3: Outcome model correctly specified

- Non-augmented weighting estimators exhibit bias; including the MAIC (entropy balancing) approach
- MAIC (entropy balancing) is not doubly robust with a logistic outcome model
- Augmented weighting estimators are generally more precise than their non-augmented weighting counterparts
- G-computation exhibits the greatest precision, but augmented weighting estimators are almost as precise

Method	Bias	ESE	95% CI coverage	Average 95% CI width
$n = 200$				
1. The naïve estimator	-0.033	0.301	0.952	1.208
2. IOW with weights from modeling	0.132	0.568	0.961	2.212
3. IOW with normalized weights from modeling	-0.032	0.386	0.951	1.534
4. MAIC	0.121	0.346	0.955	1.518
5. G-computation	0.007	0.284	0.959	1.175
6. DR with "modeling" IOW weights	0.018	0.340	0.961	1.398
7. DR with normalized "modeling" IOW weights	0.017	0.328	0.957	1.335
8. DR with MAIC weights	0.015	0.310	0.956	1.290
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.012	0.302	0.954	1.242
10. Augmented "weighted G-computation" with MAIC weights	0.010	0.302	0.958	1.283
$n = 1000$				
1. The naïve estimator	-0.040	0.134	0.937	0.528
2. IOW with weights from modeling	0.117	0.228	0.968	0.918
3. IOW with normalized weights from modeling	-0.046	0.165	0.938	0.645
4. MAIC	0.104	0.146	0.889	0.573
5. G-computation	0.004	0.122	0.952	0.487
6. DR with "modeling" IOW weights	0.007	0.142	0.947	0.559
7. DR with normalized "modeling" IOW weights	0.007	0.140	0.946	0.549
8. DR with MAIC weights	0.006	0.132	0.946	0.517
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.005	0.127	0.949	0.505
10. Augmented "weighted G-computation" with MAIC weights	0.005	0.128	0.948	0.504

DR: doubly robust; IOW: inverse-odds weighting: MAIC: matching-adjusted indirect comparison

KS4: Dual model misspecification

- All approaches are biased
- Augmentation via an outcome model does not protect against the simultaneous misspecification of two models
- There is no bias or variance amplification for the augmented estimators under dual model misspecification!

Method	Bias	ESE	95% CI coverage	Average 95% CI width
$n = 200$				
1. The naïve estimator	0.519	0.338	0.684	1.329
2. IOW with weights from modeling	0.800	0.783	0.908	2.763
3. IOW with normalized weights from modeling	0.573	0.475	0.757	1.832
4. MAIC	0.608	0.469	0.787	2.013
5. G-computation	0.532	0.383	0.745	1.544
6. DR with "modeling" IOW weights	0.576	0.459	0.767	1.831
7. DR with normalized "modeling" IOW weights	0.571	0.443	0.750	1.753
8. DR with MAIC weights	0.513	0.414	0.774	1.664
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.541	0.420	0.761	1.678
10. Augmented "weighted G-computation" with MAIC weights	0.540	0.429	0.781	1.780
$n = 1000$				
1. The naïve estimator	0.497	0.146	0.071	0.574
2. IOW with weights from modeling	0.788	0.359	0.334	1.425
3. IOW with normalized weights from modeling	0.520	0.199	0.249	0.765
4. MAIC	0.552	0.192	0.165	0.738
5. G-computation	0.516	0.162	0.104	0.635
6. DR with "modeling" IOW weights	0.542	0.188	0.178	0.729
7. DR with normalized "modeling" IOW weights	0.541	0.185	0.170	0.716
8. DR with MAIC weights	0.487	0.173	0.188	0.665
9. Augmented "weighted G-computation" with normalized "modeling" IOW weights	0.530	0.178	0.148	0.710
10. Augmented "weighted G-computation" with MAIC weights	0.539	0.183	0.153	0.708

DR: doubly robust; IOW: inverse-odds weighting: MAIC: matching-adjusted indirect comparison

Simulation study: concluding remarks

- We hypothesized that entropy balancing weights can lead to more stable and precise ATC estimation than inverse-odds modelling weights
- This is confirmed for the non-augmented estimators in the simulation study; entropy balancing exhibits greater precision than (normalized or non-normalized) modelling weighting approaches in all scenarios
- The precision gains have been inherited by the augmented approaches; estimators using entropy balancing weights generally display enhanced precision compared to those using modelling weights
- The augmented “weighted G-computation” estimators are also doubly robust for the ATC, noting that the logistic outcome model has a canonical link function
- The augmented “weighting G-computation” estimators offer similar performance than our proposed doubly robust augmented estimator with entropy balancing weights (these are the least biased and most precise estimators)

Transportability

A transportability problem

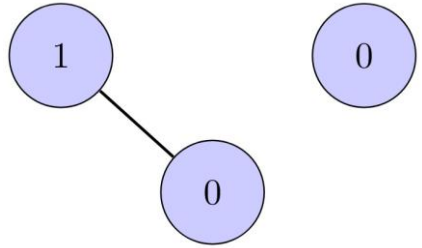
Unanchored ITC:

- Relies on transporting $\mu_1 = E(Y^1)$ from *S=1* to *S=0*
- Conditional constancy/transportability of absolute outcomes
- Conditional on covariates, potential outcomes under *T=1* are independent of data source *S*

$$Y^1 \perp S \mid \mathbf{X}$$

Externally controlled trial, e.g., RCT with hybrid control:

- Relies on transporting $\mu_0 = E(Y^0)$ from *S=0* to *S=1*
- Conditional constancy/transportability of absolute outcomes
- Conditional on covariates, potential outcomes under *T=0* are independent of data source *S*

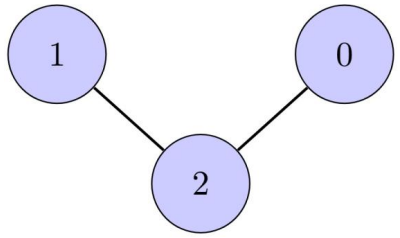


$$Y^0 \perp S \mid \mathbf{X}$$

Anchored ITC:

- Relies on transporting the treatment effect of an active intervention versus a common comparator
- **Conditional constancy/transportability of treatment effects**, considered easier to meet

Anchored ITC



What is the target estimand/summary measure/causal contrast?

- Average treatment effect in the *S=0* population?
  Requires transporting Δ_{12} from *S=1 to S=0* $\Delta_{12} \perp S \mid \mathbf{X}$

$$\Delta_{10}^{(0)} = \Delta_{12}^{(0)} - \Delta_{02}^{(0)}$$

- Average treatment effect in the *S=1* population?
  Requires transporting Δ_{02} from *S=0 to S=1* $\Delta_{02} \perp S \mid \mathbf{X}$

$$\Delta_{10}^{(1)} = \Delta_{12}^{(1)} - \Delta_{02}^{(1)}$$

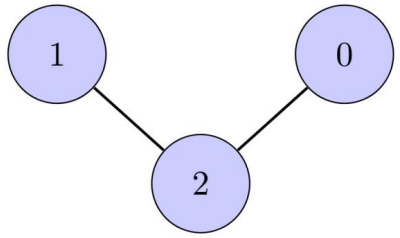
- Average treatment effect in some external target population *S=s?*
  Requires transporting Δ_{12} from *S=1 to S=s* AND transporting Δ_{02} from *S=0 to S=s*

$$\Delta_{10}^{(s)} = \Delta_{12}^{(s)} - \Delta_{02}^{(s)}$$

$$\Delta_{12}, \Delta_{02} \perp S \mid \mathbf{X}$$

Note: $\Delta_{tt'}^{S=s} = g(E[Y^t|S=s]) - g(E[Y^{t'}|S=s])$

Anchored ITC – typical setting



The following setting is common in HTA:

- The health technology developer has conducted an RCT *(S=1)* comparing its own novel intervention *T=1* versus placebo or standard-of-care *T=2*
- Another sponsor has conducted an RCT *(S=0)* comparing a “competing“ intervention *T=0* versus *T=2*
- No head-to-head trials between *T=1* and *T=0*
- Access to individual patient data (IPD) for *S=1* but not for *S=0* (only aggregate-level data from publications)

Issues:

- With no IPD for *S=0,* most covariate adjustment approaches restricted to contrast treatments in *S=0*!!!

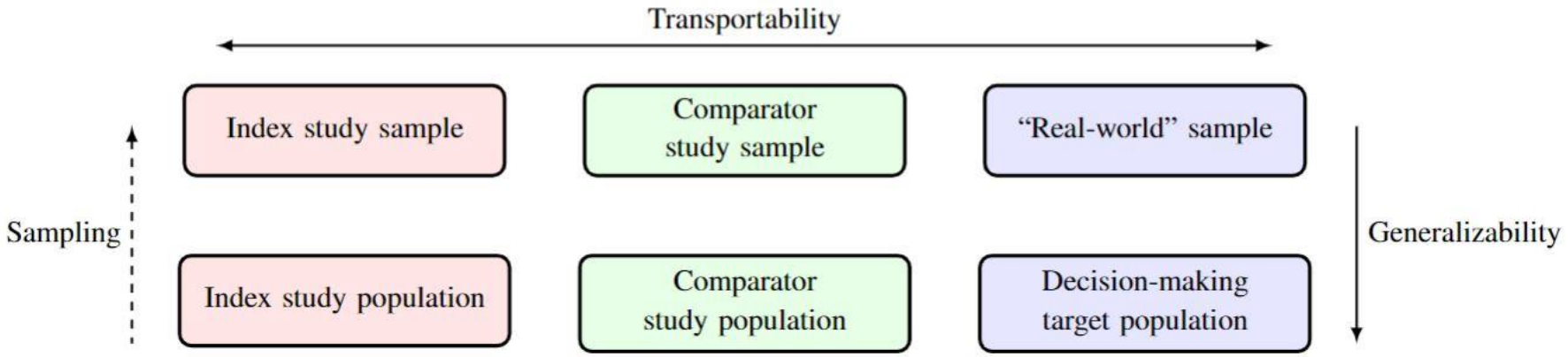
$$\Delta_{10}^{(0)} = \Delta_{12}^{(0)} - \Delta_{02}^{(0)}$$

- *S=0* sample may not be representative of the original target population of eligible patients for the trial, and may not be representative of the target population of clinical practice in the decision-makers’ jurisdiction
- Implicit acknowledgement that the competitor’s *S=0* population is more relevant to the decision than *S=1*

External validity

ITCs are used to inform HTA decisions, which are made for specific patient populations

We require treatment effects that are maximally relevant to policy decision-making and have high **external validity** with respect to the target population



Remiro-Azócar, A., 2022. Target estimands for population-adjusted indirect comparisons. *Statistics in Medicine*, *41*(28), pp.5558-5569.

"Index study": *S=1*; "Comparator study": *S=0*

If we had full access to individual patient data, we could…

- Compare covariate-adjusted estimates that have been transported to the same external target population
- Perform network meta-regression followed by standardization over any desired target population

Multilevel network meta-regression (ML-NMR)

Phillippo et al (2020), Phillippo et al (2021)

Bayesian network meta-regression approach adapted to the setting with limited individual patient data

Define an individual-level regression model (IPD meta-regression)

$$y_{ijk} \sim \pi_{\text{Ind}}\left(\theta_{ijk}\right)$$
$$\mathrm{g}\left(\theta_{ijk}\right) = \eta_{jk}\left(\boldsymbol{x}_{ijk}\right) = \mu_j + \boldsymbol{x}_{ijk}^{T}\left(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}\right) + \gamma_k$$

Average (integrate) over the target population to form the aggregate-level model

$$y_{\cdot jk} \sim \pi_{\text{Agg}}\left(\theta_{\cdot jk}\right)$$
$$\theta_{\cdot jk} = \int_{\mathrm{X}} g^{-1}\left(\eta_{jk}\left(\boldsymbol{x}\right)\right) f_{jk}\left(\boldsymbol{x}\right) d\boldsymbol{x}$$

Parametrized on individual-level conditional effects; estimates a conditional effect

$$\begin{aligned} d_{ab(P)} &= \int_{\boldsymbol{x}} \left(\mu_{(P)} + \boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,b}\right) + \gamma_b\right) f_{(P)}(\boldsymbol{x}) d\boldsymbol{x} - \int_{\boldsymbol{x}} \left(\mu_{(P)} + \boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,a}\right) + \gamma_a\right) f_{(P)}(\boldsymbol{x}) d\boldsymbol{x} \\ &= \bar{\boldsymbol{x}}_{(P)}^{\mathsf{T}}\left(\boldsymbol{\beta}_{2,b} - \boldsymbol{\beta}_{2,a}\right) + \gamma_b - \gamma_a, \end{aligned}$$

But a marginal (population-average) effect can be obtained via integration

$$\bar{p}_{k(P)} = \int_{\boldsymbol{x}} g^{-1}\left(\mu_{(P)} + \boldsymbol{x}^{\mathsf{T}}\left(\boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2,k}\right) + \gamma_k\right) f_{(P)}(\boldsymbol{x}) d\boldsymbol{x},$$
$$\Delta_{ab(P)} = \mathrm{g}(\bar{p}_{b(P)}) - \mathrm{g}(\bar{p}_{a(P)}).$$

Multilevel network meta-regression (ML-NMR)

Phillippo et al (2020), Phillippo et al (2021)

In the anchored scenario, ML-NMR can produce treatment effects in any specified target population:

- In any of the trials in the evidence base
- In an external dataset generated from real-world data, registries or observational studies

Subject to **two strong additional identifying assumptions!**

- **Shared effect modifier assumption**: same set of individual-level effect modifiers for the “aggregate-level data” treatments versus a common comparator, and treatment-covariate interactions are identical

  Can be relaxed, but this requires at least a moderate-to-large number of trials

- **Correct specification of the target population** (if IPD lacking): distributional assumptions about correlations, distributional forms…particularly for non-collapsible measures, which depend on joint distribution of prognostic variables

Example

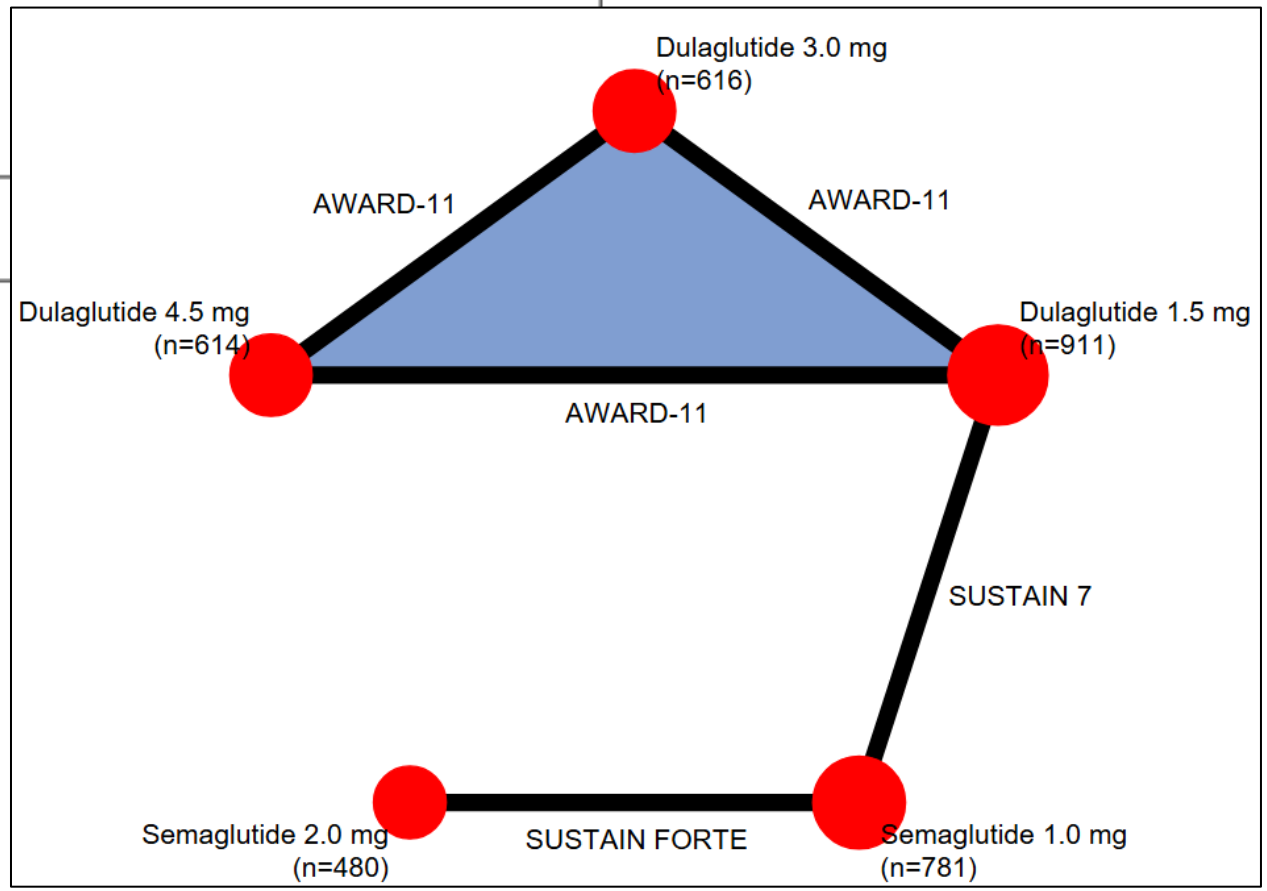
Illustrative example: anchored case (ML-NMR)

Lingvay et al (2022)

Population	Subjects with Type 2 diabetes on a background treatment of metformin
Intervention	Semaglutide 2.0 mg QW
Comparators	Dulaglutide 4.5 mg QW Dulaglutide 3.0 mg QW *Dulaglutide 1.5 mg QW* *Semaglutide 1.0 mg QW*
Outcomes	Change from baseline in HbA1c Change from baseline in body weight

No head-to-head data available comparing semaglutide 2.0 mg with dulaglutide 3.0 mg or 4.5 mg in patients with Type 2 diabetes (T2D)

An ITC on glycated hemoglobin (HbA1c) and body weight was performed



Evidence base

Trial	AWARD-11[60,66,67]		SUSTAIN 7[57,62,63]		SUSTAIN FORTE[59,64,65]	
Estimand name	Efficacy	Treatment regimen	De-jure	De-facto	Hypothetical (trial product)	Treatment policy
Population	Subjects with Type 2 diabetes inadequately controlled with metformin		Subjects with Type 2 diabetes on a background treatment with metformin		Subjects with Type 2 diabetes on a background treatment of metformin with or without sulphonylurea treatment	
Treatments	Subcutaneous dulaglutide 4.5 mg QW Subcutaneous dulaglutide 3.0 mg QW Subcutaneous dulaglutide 1.5 mg QW		Subcutaneous semaglutide 1.0 mg QW Subcutaneous semaglutide 0.5 mg QW Subcutaneous dulaglutide 1.5 mg QW Subcutaneous dulaglutide 0.75 mg QW		Subcutaneous semaglutide 2.0 mg QW Subcutaneous semaglutide 1.0 mg QW	
Variables (endpoints)	Primary: change from baseline to week 36 in HbA1c (%-points) Key secondary: change from baseline to week 36 in body weight (kg)		Primary: change from baseline to week 40 in HbA1c (%-points) Confirmatory secondary: change from baseline to week 40 in body weight (kg)		Primary: change from baseline to week 40 in HbA1c (%-points) Confirmatory secondary: change from baseline to week 40 in body weight (kg)	
Population-level summary measure	Mean difference in change from baseline		Mean difference in change from baseline		Mean difference in change from baseline	
Intercurrent event strategies	**Hypothetical** strategy for initiation of anti-diabetic rescue medication or premature treatment discontinuation	**Treatment policy** strategy for initiation of anti-diabetic rescue medication or premature treatment discontinuation	**Hypothetical** strategy for initiation of anti-diabetic rescue medication or premature treatment discontinuation	**Treatment policy** strategy for initiation of anti-diabetic rescue medication or premature treatment discontinuation	**Hypothetical** strategy for initiation of anti-diabetic rescue medication or premature treatment discontinuation **Treatment policy** strategy for change in treatment dose	**Treatment policy** strategy for initiation of anti-diabetic rescue medication or premature treatment discontinuation **Treatment policy** strategy for change in treatment dose

Remiro-Azócar, A. et al.., 2025. Incorporating estimands into meta-analyses of clinical trials. arXiv preprint arXiv:2510.15762.

Evidence base

Table 1. Overview of study characteristics and inclusion criteria of included trials

Study name (primary reference) NCT number	Study design	Inclusion criteria				Duration of follow-up	Randomized treatment	No. of patients randomized	Completion date
		HbA_{1c}	BMI	T2D duration	Background medication				
SUSTAIN FORTE(14) NCT03989232	Phase 3b, double-blind, RCT Multinational	8.0%-10.0%	No restriction	At least 6 mo	Metformin ≥ 1500 mg/d (or a maximal tolerated dose) ≥ 90 d with or without sulphonylurea (≥ half the maximum approved dose according to local label or maximum tolerated or effective dose)	40 wk	Semaglutide 2.0 mg Semaglutide 1.0 mg	480 481	November 2020
SUSTAIN 7(4)[a] NCT02648204	Phase 3b, open-label, RCT Multinational	7.0%-10.5%	No restriction	Not specified	Metformin ≥ 1500 mg/d (or a maximal tolerated dose) ≥ 90 d	40 wk	Semaglutide 1.0 mg Dulaglutide 1.5 mg	300 299	May 2017
AWARD-11(11) NCT03495102	Phase 3, double-blind, RCT Multinational	7.5%-11.0%	≥25 kg/m²	At least 6 mo	Metformin ≥ 1500 mg/d for ≥ 3 months	36 wk	Dulaglutide 3.0 mg Dulaglutide 4.5 mg Dulaglutide 1.5 mg	616 614 612	May 2019

Lingvay, I., Bauer, R., Baker-Knight, J., Lawson, J. and Pratley, R., 2022. An indirect treatment comparison of semaglutide 2.0 mg vs dulaglutide 3.0 mg and 4.5 mg using multilevel network meta-regression. The Journal of Clinical Endocrinology & Metabolism, 107(5), pp.1461-1469.

There are some differences in the trials' inclusion criteria…target patient populations are not identical

Ultimately leading to differences in the distribution of baseline characteristics between "analysis sets"

ML-NMR was performed to adjust for potential effect modifiers:

- Baseline HbA1c
- Baseline BMI
- Baseline T2D duration

Lingvay, I., Bauer, R., Baker-Knight, J., Lawson, J. and Pratley, R., 2022. An indirect treatment comparison of semaglutide 2.0 mg vs dulaglutide 3.0 mg and 4.5 mg using multilevel network meta-regression. The Journal of Clinical Endocrinology & Metabolism, 107(5), pp.1461-1469.

Results

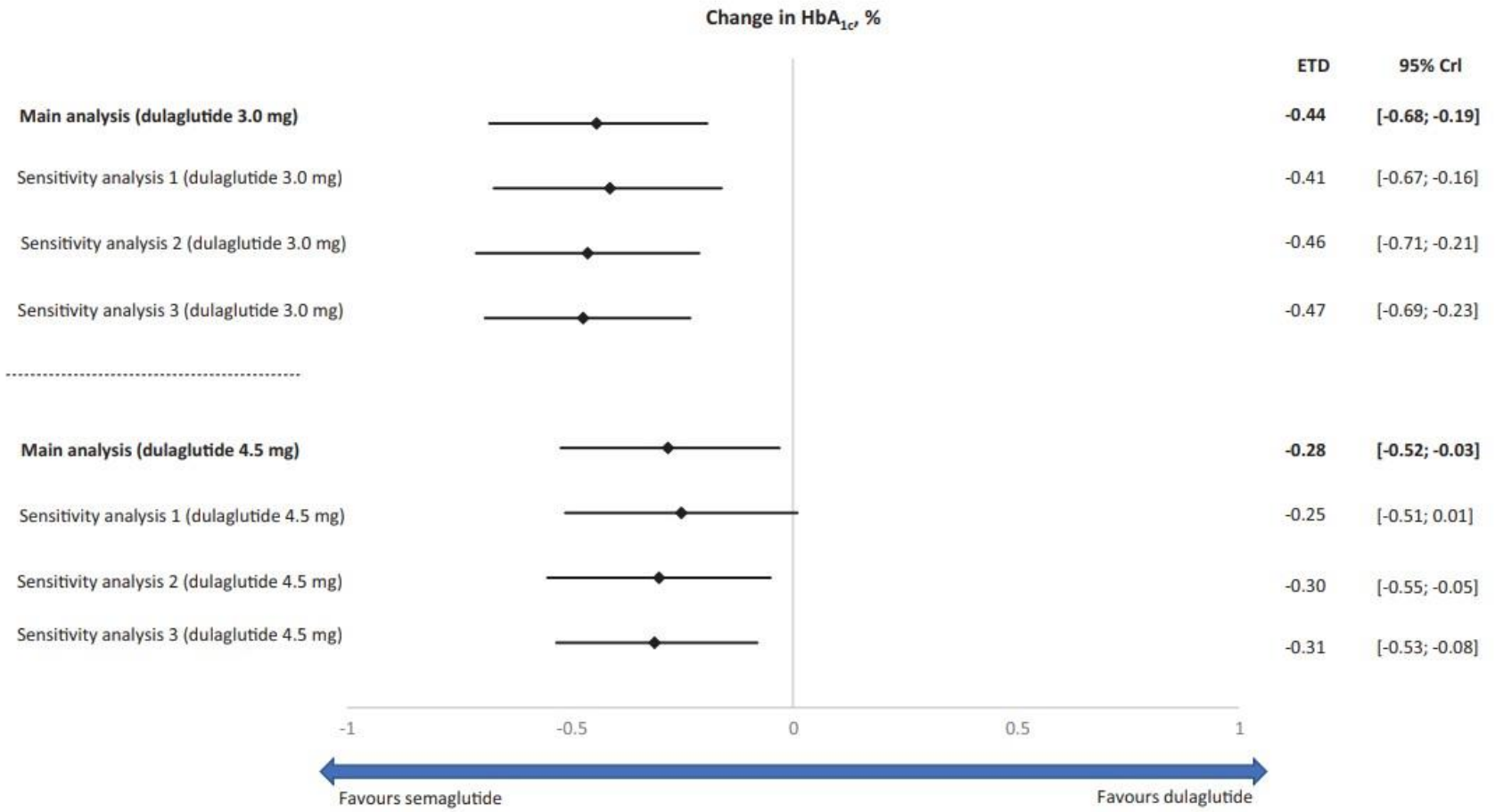
HbA1c

Main analysis: ML-NMR with AWARD-11 target population

Sensitivity analysis 1: ML-NMR with SUSTAIN 7 target population

Sensitivity analysis 2: ML-NMR with SUSTAIN FORTE target population

Sensitivity analysis 3: unadjusted Bayesian network meta-analysis



Conclusion: No appreciable impact of the confirming no appreciable impact of effect modification and the target population on the results of the analysis for HbA1c

Limitation: shared effect modifier assumption required for all dose levels of dulaglutide versus semaglutide

Results

Body weight

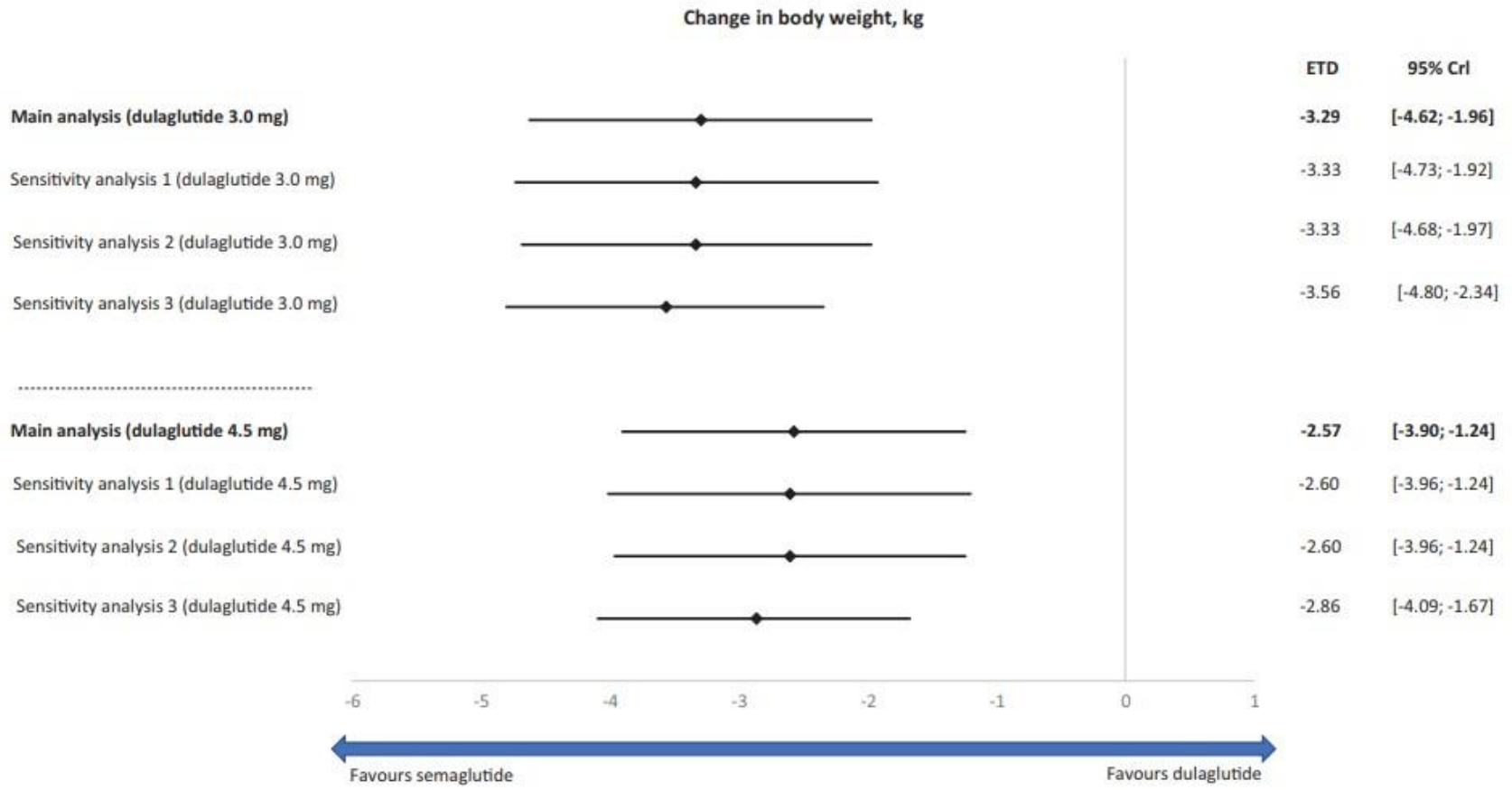
Lingvay, I., Bauer, R., Baker-Knight, J., Lawson, J. and Pratley, R., 2022. An indirect treatment comparison of semaglutide 2.0 mg vs dulaglutide 3.0 mg and 4.5 mg using multilevel network meta-regression. The Journal of Clinical Endocrinology & Metabolism, 107(5), pp.1461-1469.

Main analysis: ML-NMR with AWARD-11 target population

Sensitivity analysis 1: ML-NMR with SUSTAIN 7 target population

Sensitivity analysis 2: ML-NMR with SUSTAIN FORTE target population

Sensitivity analysis 3: unadjusted Bayesian network meta-analysis



Conclusion: No appreciable impact of the confirming no appreciable impact of effect modification and the target population on the results of the analysis for body weight

Limitation: shared effect modifier assumption required for all dose levels of dulaglutide versus semaglutide

Moving forward

Current methodological approaches have limitations

- Most covariate adjustment approaches used in the context of ITCs and HTA are singly robust
- Most approaches are based on parametric modelling, relying on the correct specification of a single parametric model
- Strong parametric assumptions are often unsubstantiated and fail to reflect the complexities of the real data
- If the parametric model is incorrectly specified, the singly robust estimator is subject to bias and this bias does not decrease with sample size, at any rate
- Doubly robust estimators should be less prone to model misspecification bias than singly robust estimators; they offer two opportunities for valid adjustment instead of one
- Nevertheless, they are still subject to bias where the two working models are parametric: both parametric models may be incorrect!

Need for more bias-robust methodologies

The use of more modern causal inference methods, involving data-adaptive estimation within a doubly robust framework, remains underexploited in ITCs and HTA, and could provide:

- Fewer structural assumptions about data generating-mechanisms, reducing the risk of model misspecification bias
- Precision/efficiency…while avoiding unreasonable extrapolation in poor overlap situations
- Good finite-sample performance…by allowing for the use of slower-converging models
- Valid statistical inference and uncertainty quantification…by sample splitting (”cross-fitting”) to relax the Donsker condition, weaking the restrictions on the algorithms that can be used

It is worth considering these in combination with entropy balancing approaches to weighting, which have desirable properties regardless of IPD availability

Integration with larger networks of treatments and studies and limited IPD likely to be a challenge

References

References

- Campbell, H. and Remiro-Azócar, A., 2025. Doubly robust augmented weighting estimators for the analysis of externally controlled single-arm trials and unanchored indirect treatment comparisons. arXiv preprint arXiv:2505.00113.
- van Engen, A., Krüger, R., Parnaby, A., Rotaru, M., Ryan, J., Samaha, D. and Tzelis, D., 2024. The impact of additive population (s), intervention, comparator (s), and outcomes in a European joint clinical health technology assessment. Value in Health, 27(12), pp.1722-1731.
- FDA guidance for industry: considerations for the design and conduct of externally controlled trials for drug and biological products. https://www.fda.gov/media/164960/download; 2023. Accessed: 11-08-2025.
- Gabriel, E.E., Sachs, M.C., Martinussen, T., Waernbaum, I., Goetghebeur, E., Vansteelandt, S. and Sjölander, A., 2024. Inverse probability of treatment weighting with generalized linear outcome models for doubly robust estimation. Statistics in Medicine, 43(3), pp.534-547.
- Guidance on the scoping process. https://health.ec.europa.eu/publications/guidance-scoping-process_en; 2024. Accessed: 11-08-2025.
- Lingvay, I., Bauer, R., Baker-Knight, J., Lawson, J. and Pratley, R., 2022. An indirect treatment comparison of semaglutide 2.0 mg vs dulaglutide 3.0 mg and 4.5 mg using multilevel network meta-regression. The Journal of Clinical Endocrinology & Metabolism, 107(5), pp.1461-1469.
- Methodological Guideline for Quantitative Evidence Synthesis: Direct and Indirect Comparisons. https://health.ec.europa.eu/latest-updates/methodological-guideline-quantitative-evidence-synthesis-direct-and-indirect-comparisons-2024-03-25_en; 2024. Accessed: 11-08-2025.
- MHRA draft guideline on the use of external control arms based on real-world data to support regulatory decisions. https://assets.publishing.service.gov.uk/media/6825bab1a4c1a40fde4e63e5/Draft_MHRA_Guideline_on_Studies_with_RWD_ECA_May2025.pdf. Accessed: 11-08-2025.

References

- Phillippo, D., Ades, T., Dias, S., Palmer, S., Abrams, K.R. and Welton, N., 2016. NICE DSU technical support document 18: methods for population-adjusted indirect comparisons in submissions to NICE.
- Phillippo, D.M., Dias, S., Ades, A.E., Belger, M., Brnabic, A., Schacht, A., Saure, D., Kadziola, Z. and Welton, N.J., 2020. Multilevel network meta-regression for population-adjusted treatment comparisons. Journal of the Royal Statistical Society. Series A (Statistics in Society), 183(3), p.1189.
- Phillippo, D., Dias, S., Ades, A.E. and Welton, N.J., 2021. Target estimands for efficient decision making: Response to comments on “Assessing the performance of population adjustment methods for anchored indirect comparisons: A simulation study”. Statistics in Medicine, pp.2759-2763.
- Reflection paper on establishing efficacy based on single-arm trials submitted as pivotal evidence in a marketing authorisation application. https://www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-establishing-efficacy-based-single-arm-trials-submitted-pivotal-evidence-marketing-authorisation-application_ en.pdf; 2024. Accessed: 11-08-2025.
- Remiro-Azócar, A., 2022. Target estimands for population-adjusted indirect comparisons. Statistics in Medicine, 41(28), pp.5558-5569.
- Remiro-Azócar, A. et al.., 2025. Incorporating estimands into meta-analyses of clinical trials. arXiv preprint arXiv:2510.15762.
- Senn, S., 2025. Causal estimates for external controls. How reasonable are the assumptions?. Journal of the Royal Statistical Society Series A: Statistics in Society, p.qnaf127.
- Słoczyński, T., Uysal, D. and Wooldridge, J.M., 2025. Covariate balancing and the equivalence of weighting and doubly robust estimators of average treatment effects.
- Zhao, Q. and Percival, D., 2017. Entropy balancing is doubly robust. Journal of causal inference, 5(1), p.20160010.
- Zhou, X., Zhu, J., Drake, C. and Pang, H., 2025. Causal estimators for incorporating external controls in randomized trials with longitudinal outcomes. Journal of the Royal Statistical Society Series A: Statistics in Society, 188(3), pp.791-818.

Thank you! Discussion/Q&A