sanofi

Evaluating ML-Based Causal Inference in Rare-Disease Trials

Konstantinos Tsingas, Ahrim Youn, Yuqian Shen, Qi Zhang

November 02, 2026

Complications of Statistical Analysis in Rare Disease Trials

> *The LEAP2MONO study — a Phase 3 trial of venglustat in Gaucher Disease Type 3 — enrolled only **43 patients**, exemplifying the challenge.*

Key Question:

Which causal inference methods remain valid and powerful when N is very small?

Rare diseases can affect fewer than 1 in 2,000 individuals, creating a fundamental tension between the need for rigorous evidence and the practical impossibility of large randomized trials.

- **Small sample sizes:**
  N < 100 is common; N < 50 is frequent, severely limiting statistical power.
- **Limited randomization:**
  Single-arm trials and external control arms are often necessary.
- **Confounding:** Non-randomized comparisons introduce selection bias.
- **Missing data:** Dropout and censoring disproportionately affect small trials.
- **Regulatory flexibility:**
  FDA and EMA accept adaptive designs but require rigorous bias control.

"In very rare disorders, it is important that every patient participating in a study contributes as much information as possible to make a benefit–risk assessment possible. Therefore, the well-planned use of the best available techniques to obtain and analyse information is crucial."
— EMA GUIDELINE ON CLINICAL TRIALS IN SMALL POPULATIONS

sanofi

Regulatory Recommendations on the Use of AI

U.S. FOOD & DRUG ADMINISTRATION

Home / Regulatory Information / Search for FDA Guidance Documents / Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products

GUIDANCE DOCUMENT

Considerations for the Use of Artificial Intelligence To Support Regulatory Decision-Making for Drug and Biological Products

Draft Guidance for Industry and Other Interested Parties

JANUARY 2025

U.S. FOOD & DRUG ADMINISTRATION

Home / News & Events / FDA Newsroom / Press Announcements / FDA Launches Framework for Accelerating Development of Individualized Therapies for Ultra-Rare Diseases

FDA NEWS RELEASE

FDA Launches Framework for Accelerating Development of Individualized Therapies for Ultra-Rare Diseases



EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

20 September 2022
EMADOC-1700519818-907465
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion for Prognostic Covariate Adjustment (PROCOVA™)

Draft agreed by Scientific Advice Working Party (SAWP)	10 February 2022
Adopted by CHMP for release for consultation	24 February 2022[1]
Start of public consultation	22 March 2022 [2]
End of consultation (deadline for comments)	03 May 2022[3]
Adopted by CHMP	15 September 2022

Keywords	Qualification of Novel Methodology, Statistical methodology, Prognostic Covariate Adjustment, Sample size estimation

Agencies have signaled interest in safe use of AI/ML for drug development, including evidence generation for rare diseases.

PROCOVA is already qualified by the EMA (2022) and FDA (2023) as an acceptable primary-analysis method.

sanofi

Confidential - Sensitive

Benchmarking Causal Estimators and ML-Based Causal Models with Complex Data

01 Method Performance Under Small Samples

Evaluate the performance (bias, variance, power, type I error) of **10 causal inference methods** in rare and small sample size scenarios for estimating treatment effect.

- Realistic rare disease simulation setting
- Focus on degradation points for each method

02 Robustness to Missing Data

At various sample sizes, evaluate methods with **missing at random (MAR)** and **missing not at random (MNAR)** data for the primary endpoint.

03 Leveraging Real-World Data

Incorporate **external control arm (ECA) data** across different borrowing strategies and evaluate performance depending on shift in patient population.

Study Initiative Evidence-based recommendations on method selection and minimum viable sample size for rare disease trials with repeated measures.

sanofi

Confidential - Sensitive

Ten Causal Inference Methods — From Classical to Machine Learning

Method	Full Name	Type	Doubly Robust	Longitudinal	Missing Data	R Package
MMRM	Mixed model (repeated measures)	Parametric	No	Yes	MAR (REML)	mmrm
P-TMLE	Point-treatment TMLE	Semi-parametric	Yes	No	MAR (IPCW)	tmle
L-TMLE	Longitudinal TMLE	Semi-parametric	Yes	Yes	MAR (IPCW)	ltmle
A-TMLE	Adaptive TMLE	Semi-parametric	Yes	No	MAR (IPCW)	atmle
Match	Full propensity-score matching	Matching	No	No	Complete case	MatchIt
G-Comp	Sequential G-computation	Parametric	No	Yes	Complete case	gfoRmula
IPTCW	Inverse treatment & censoring weighting	Weighting	No	Yes	MAR (IPCW)	WeightIt
Causal Forest	Generalized random forests	Non-parametric	Partial	No	Complete case	grf
PROCOVA	Prognostic covariate adjustment	Parametric	No	No	Complete case	Base R
PROCOVA-MMRM	Longitudinal prognostic covariate adjustment	Parametric	No	Yes	MAR (REML)	mmrm

sanofi

Comparative Strengths and Limitations Across Key Properties

Criterion	MMRM	P-TMLE	L-TMLE	A-TMLE	Match	G-Comp	IPTCW	Causal Forests	PROCOVA	PROCOVA-MMRM
Doubly Robust	✗	✓	✓	✓	✗	✗	✗	–	✗	✗
Longitudinal Data	✓	✗	✓	✗	✗	✓	✓	✗	✗	✓
Small N Stability	✓	✗	✗	✗	✗	–	–	✗	✓	✓
Heterogeneous Treatment Effect Estimation	✗	✗	✗	✗	✗	✗	✗	✓	✗	✗
Misspecification Robust	✗	✓	✓	✓	–	✗	–	✓	–	–
Regulatory Acceptance	✓	–	–	–	✓	–	✓	–	✓	✓
Handles Censoring	–	✓	✓	✓	✗	–	✓	✗	✗	–
Interpretability	✓	–	–	–	✓	✓	✓	–	✓	✓
Borrowing from External Data Sources	–	–	–	✓	✓	–	–	–	✓	✓
Computational Cost	Low	Med	High	High	Low	Med	Med	High	Med	Med

✓ Yes / Strong – Partial / Moderate ✗ No / Weak

Key Insight

No single method dominates across all dimensions. MMRM and PROCOVA may offer the best balance of stability and regulatory acceptance at small N; TMLE variants offer double robustness at the cost of small-sample instability.

sanofi

MMRM: Regulatory Baseline

Mixed Model for Repeated Measures

THE REGULATORY GOLD STANDARD

- Used in LEAP2MONO phase 3 primary endpoint analysis for efficacy of venglustat.
- Models the full covariance structure of longitudinal data.
- Handles missing data under Missing At Random (MAR) via REML.
- ***External data: pool RCT+ECA controls in one model and adjust for source***
- ***indicator S.***

$$Y_{ij} = \mu + \alpha_i \cdot A_i + \beta_j \cdot \text{visit} + \gamma_{ij} \cdot (A_i \cdot \text{visit}) + \delta \cdot \text{baseline} + \epsilon_{ij}$$

\+ **Pros:** Widely accepted by FDA/EMA (ICH E9); handles dropout under MAR.

\+ **Pros:** Efficient use of all observed data points.

– **Cons:** Assumes correct model specification; not doubly robust.

Confidential - Sensitive

Propensity Score (PS) Matching and Inverse Probability Weighting (IPW)

Propensity Score Definition

$$g(X) = P(A = 1 \mid X), \qquad \pi(X) = P(S = 1 | A = 0, X)$$

The probability of treatment A or source S given observed covariates. Balances all observed confounders simultaneously.

Assumption: positivity

- *Strictly* $0 < g(X), \pi(X) <$

PS Matching

- Each treated unit is matched to the most similar control unit(s) based on propensity score, with replacement.
- After matching, a regression model is fit taking into account the matched design
- ***External data:*** ***match RCT patients to their nearest external control arm (ECA) counterparts using PS.***

Inverse Probability Treatment and Censoring Weighting

- Creates a pseudo-population based on weights W_i
- Censoring is independent of potential outcomes, re-creating a randomized trial with complete follow-up.
- ***External data:*** ***weights ECA controls*** $(S = 0)$ ***by inverse odds of trial-selection, then fits on the pooled sample.***
- For participant i, missingness indicator R_i, source S_i, and data X_i:

$$W_i = \frac{\hat{\pi}(X_i)}{1 - \hat{\pi}(X_i)} \cdot \frac{1}{\hat{P}(R_i = 1 \mid A_i, X_i)}, 1 \text{ otherwise}$$

Rare Disease Considerations

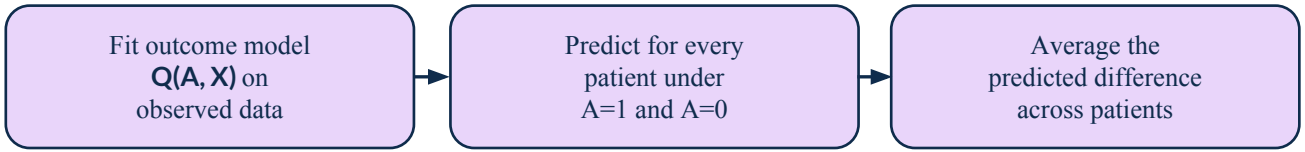
- **Positivity violation** is common in small samples when fitting $g(X)$ or $\pi(X)$.
- **Small-N instability** can lead to extreme weights in the IPW model.
- At very small N (< 30), matching may leave insufficient pairs for valid inference.

sanofi

G-Computation/G-Formula

Fit one outcome model Q on the observed data, then use it to predict what every patient's outcome would have been under each arm.

- Q can be fit using any flexible learner.
- Simulate each patient's potential outcome under both treatment arms.
- ***External data:* pool RCT+ECA controls in one model and adjust for source indicator S.**



$$\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}[\hat{Q}(1, X_i) - \hat{Q}(0, X_i)]$$

R Packages: mmrm, nlme, lme4

Confidential - Sensitive

Causal Forests: Machine Learning for Heterogeneous Treatment Effects

Beyond the ATE

While traditional methods estimate the Average Treatment Effect (ATE), Causal Forests estimate the **Conditional Average Treatment Effect (CATE)**:

$$\tau(x) = E[Y(1) - Y(0) \mid X = x]$$

This identifies subgroups that benefit most or least from treatment, driven by the data rather than pre-specified interactions.

The ATE is obtained by averaging the CATE.

The Algorithm (GRF)

1. Recursive Partitioning

Splits data to maximize differences in treatment effect between child nodes,

For each leaf L_T, estimate the treatment effect:

$$E(Y|A = 1, X \in L_T) - E(Y|A = 0, X \in L_T)$$

2. Honest Splitting

Uses one subsample to build tree structure and a separate subsample to estimate effects within leaves to prevent overfitting.

3. Ensemble Aggregation

Averages predictions across many trees to smooth variance.

External

Weights ECA controls by inverse odds of source S, then fits on pooled sample.

Rare Disease Reality Check

ML methods are data hungry. Causal Forests work well with larger samples (N > 500) but struggle to find stable splits in very small samples (N < 100). This effectively reduces the sample available for tree building.

sanofi

Confidential - Sensitive

PROCOVA / PROCOVA-MMRM: Prognostic Score Adjustment Boosts Power Without Increasing Sample Size

The Dimensionality Problem

In rare disease trials (N < 50), adjusting for multiple baseline covariates (age, sex, severity, history) consumes degrees of freedom, potentially reducing power or causing convergence issues.

The PROCOVA Solution

Prognostic Covariate Adjustment reduces dimensionality by training a model on *external historical data* to create a single composite score.

1. **Historical Control Data (RWD/Prior Trials)**
2. **Train Prognostic Mode**
   $\mu(X) = E[Y|X]$
3. **Generate Single Score for Current Trial** $\hat{\mu}_i(X_i$
4. **Adjust for Covariate**
   $Y_i = \alpha + \beta \cdot Trt_i + \gamma \cdot \hat{\mu}_i + \varepsilon_i$

Qualified by EMA (2022) & FDA (2023)

Formally qualified as a valid method for primary analysis in Phase 3 trials, provided the prognostic model is pre-specified and locked before unblinding.

PROCOVA-MMRM:

Generate **time-matched prognostic scores** at each visit and include them in an **MMRM** to boost precision for repeated measures:

$$Y_{it} = \alpha_t + \beta_t \cdot Trt_i + \gamma_t \cdot \hat{\mu}_i(t) + \varepsilon_{it}$$

Obtain μ using predictive AI/ML *applied to external dataset*

Super Learner: ensemble of learners, e.g.

- LASSO
- **Random Forests**
- **MARS**
- XG-Boost

sanofi

Confidential - Sensitive

TMLE: Double Robustness and Semiparametric Efficiency

Core Concept

Targeted Maximum Likelihood Estimation (TMLE) is a two-step semiparametric estimator that combines an initial outcome model with a propensity model and then targets the average treatment effect.

1. **Outcome Model (Q):** Initial estimate of E[Y|A,X]
2. **Propensity Model (g):** Estimate of P(A=1|X)

Double Robustness Property

Consistent if *either* the outcome model (Q) *or* the propensity score model (g) is correctly specified.

PROS: Double robust, semiparametric efficient, valid with ML models

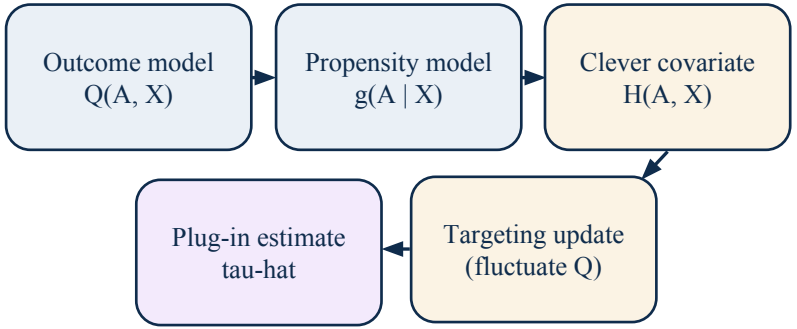
CONS: Requires large N, computationally intensive, extreme weights possible

Assumption: positivity

- *Strictly* $0 < g(X) < 1 \; \forall x$

P-TMLE (Point Treatment)

- Formulated for a single treatment decision. The algorithm follows 5 steps:



- **External data: pool RCT + ECA together and model the source indicator *S***

L-TMLE (Longitudinal Formulation)

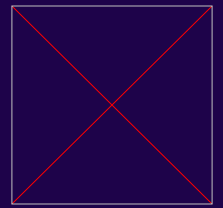
Extends TMLE to time-varying treatments and longitudinal outcomes using the sequential regression framework.

A-TMLE (Adaptive Formulation)

Extends TMLE to incorporate external control arm data. Decomposes the ATE as a pooled estimate minus a debiasing term depending on exchangeability between RCT and ECA data.

Simulation Study

DESIGN



Confidential - Sensitive

Building Synthetic Data Grounded in GD3 Biology

Endpoint

Change in RBANS Total at 52 (26) Weeks

Repeatable Battery for Neuropsychological Status.

(**Higher scores indicate better cognition.**)

Target Estimand: $\tau = E[Y(1) - Y(0) \mid S = \text{RCT}]$

Covariates (X)

- Baseline RBANS Score
- Age (Continuous)

*Adjusted for in original protocol analysis: "**Protocol MMRM**"*

- Age (Onset)
- Sex
- Weight and BMI
- Country of origin
- Race
- Baseline SARA Score
- Seizure history
- GBA mutation status
- Correlated covariate

Covariate Modeling

+ **Transform covariates:** for each factor variable, fit a random forest on the covariates to derive a continuous, subject-specific probability score.
+ **Normalize:** apply a normal transformation to continuous and transformed factor variables (X*).
+ **Add covariates**: investigate the impact of covariates prognostic for the outcome and simulate high ($r = 0.8$) and low ($r = 0.1$) correlation covariate.

Data Generation

+ Fit a mixture of multivariate normal (MVN) distributions to D = (Y, X*), covering the outcome at every timepoint.
+ Sample a large synthetic pool from that mixture, then invert the normal transform and convert probabilities to categories by sampling.
+ Conducted using Sanofi's ***gensynth*** package.

sanofi

Confidential - Sensitive

Sampling RCT and External Control Cohorts from the Synthetic Pool

Each replicate draws two independent cohorts from the same synthetic pool, so RCT and ECA patients share one common generative process.

RCT Sampling

- Randomly select N_1 synthetic patients into the RCT.
- Randomize to treatment $(A = 1)$ vs. control $(A = 0)$ in a $1:1$ ratio.
- For treated patients: replace Y_i with $Y_i + \beta \cdot A_i$, where beta is the true treatment effect.
- Alternative treatment effect $\beta = 8.32$, retrieved from LEAP2MONO phase 3 study.

External Control Sampling

- Select N_2 synthetic patients into the ECA with probability:

$$p_i = \text{logit}^{-1}(-c(F(X_i) - \overline{F(X)}))$$

- $F(X_i)$ is a prognostic score function; $c \in \{0,5\}$ is the selection-bias parameter.
- $c > 0$ preferentially selects patients with lower Y_i into the ECA, simulating an ECA population that differs systematically from the RCT.

sanofi

Confidential - Sensitive

Missing Data Patterns Reflect Real-World Dropout in Rare Disease Trials

In rare disease trials, missing data is rarely random. Patients drop out due to personal determinants, or lack of efficacy.

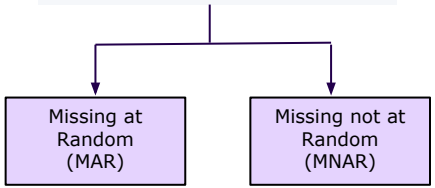
We simulate late dropout to evaluate methods when missing is random and not random:

Missing Data Pattern: Late Dropout

Missing at Week 52 only.

Completes intermediate visit but drops out before primary endpoint.

Context: Trial fatigue or perceived lack of benefit.



Missing Data Model

$$P(missing\ at\ week\ 52) = logit(30) + 1.5z_i$$

MAR: z_i = baseline RBANS

MNAR: z_i = change at week 52 from

Confidential - Sensitive

Simulation Design: Metrics and Parameters

Every scenario is scored on power, Type I error, and bias, across the parameter grid on the right.

Performance Metrics

+ Type I error: reject H_0 when $\beta = 0$ (**nominal target: 1-sided 5% / 2-sided 10%**).
+ Power: reject H_0 when $\beta = 8.32$ (**target > power of protocol MMRM**).
+ Bias: difference between the estimated and true treatment effect.
+ Variance: analytical and Monte Carlo standard errors for efficiency and precision of causal estimator.

Design Parameters

+ Sample size (N): 30 and 60, the rare-disease stress-test range.
+ Correlation (r): lowly (r=0.1) and highly (r=0.8) correlated prognostic covariate
+ Missingness mechanism: MAR and MNAR (30% of outcome data at week 52).
+ ECA selection bias (c): no difference in external and RCT populations (c=0) and non-exchangeable populations (c=5).

sanofi

Confidential - Sensitive

Simulation Scenario Prioritization

Scenarios were prioritized by relevance to the study's three objectives.

Core Priorities

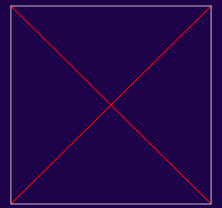
+ 1. Small-sample-size stress test: performance as N shrinks toward the rare-disease range.
+ 2. Missing-data impact: MAR vs. MNAR at each sample size.
+ 3. Treatment effect size: under the null, and estimated effect of venglustat from the LEAP2MONO study.

Secondary Priorities

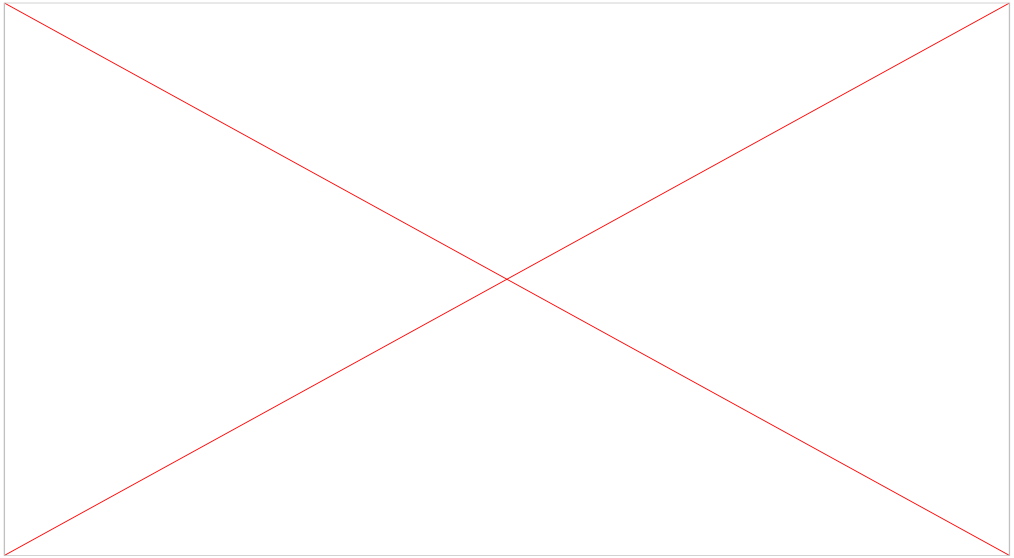
+ 3. High- vs. low-correlation impact on covariate-adjusted and prognostic-score methods.
+ 4. ECA methods comparison, with emphasis on PROCOVA and TMLE family’s borrowing behavior under selection bias.

Simulation Study

RESULTS



Small Sample Sizes: Most Methods Unbiased Under the Null



Robustness in larger sample sizes

Precision of treatment effect estimate increases as sample size increases, and all methods become more robust.

Small impact of missing data under the null

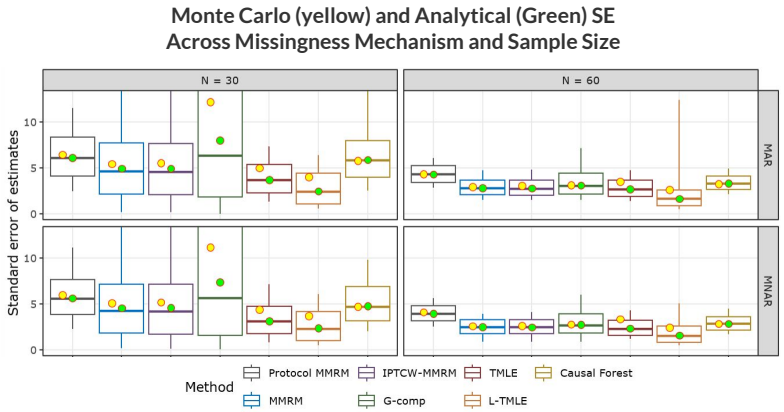
MAR and **MNAR** outcome distributions are similar under the null. Slight bias in small sample sizes across methods.

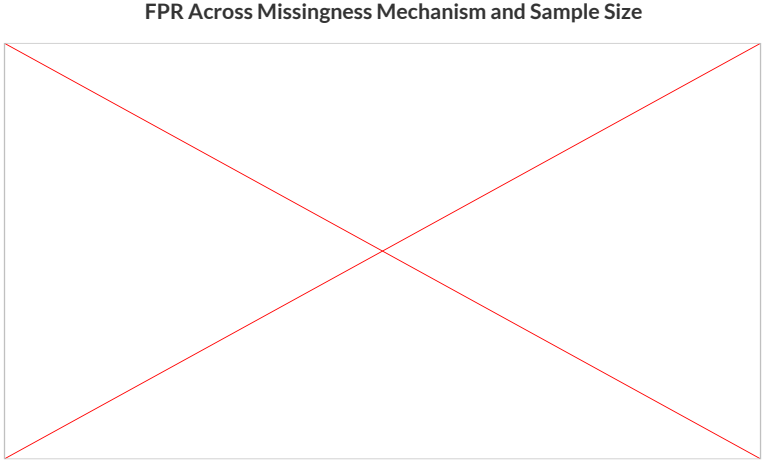
Large rare-sample instability

For N =30, **G-computation** is the least stable due to potential misspecification of conditional outcome model. Reducing covariates included in model helps calibration at small sample sizes.

Results from RCT-Only Analysis

TMLE Methods Exhibit Type I Error Inflation





Inefficiency and lack of precision in rare sample sizes

SE increases as N decreases. **G-computation** has high sampling variability. **TMLE** underestimates the analytic variance.

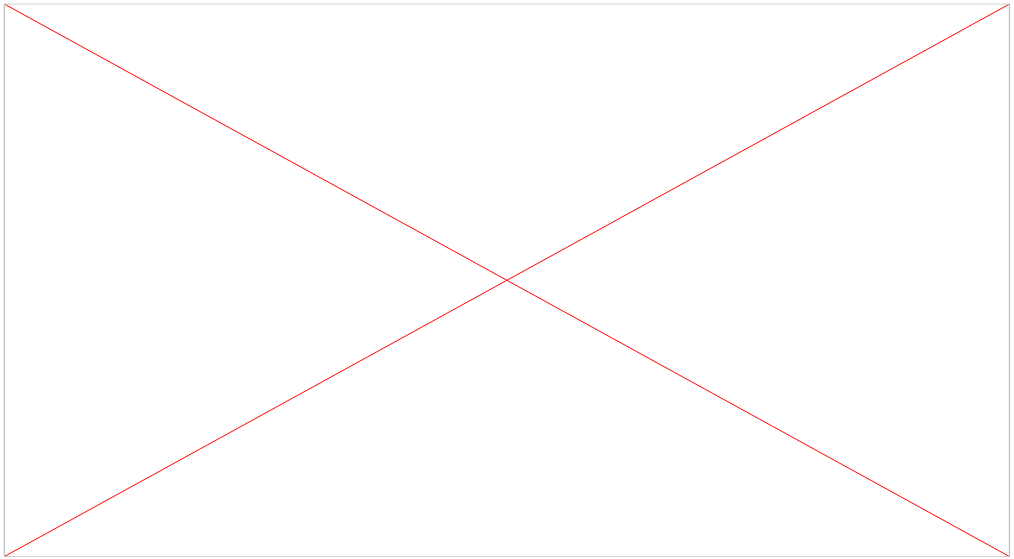
8-10% increased Type I error with TMLE methods

TMLE and **L-TMLE** have inflated false positive rate, regardless of sample size and missingness mechanism.

sanofi

Results from RCT-Only Analysis

All Methods Induce 10% - 50% Bias in Presence of MNAR Outcome Data



Increased precision under the alternative

All methods have better precision when estimating the true effect as opposed to the null. **G-computation** has the largest sampling variability.

Bias with TMLE and causal forest in MAR

TMLE family and **causal forest** approaches are not robust in MAR data, whereas other methods are unbiased.

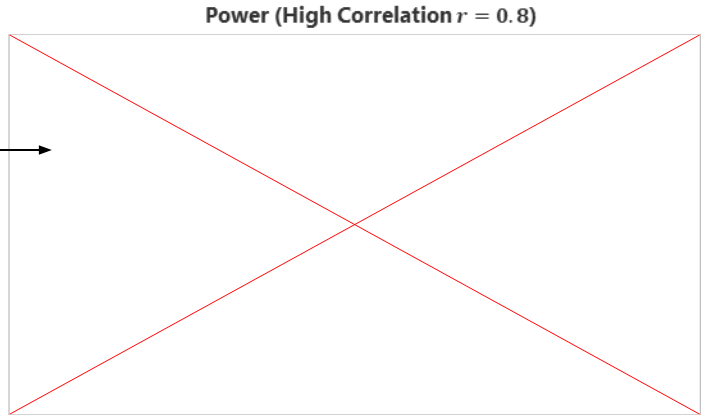
The true effect is underestimated by 10% -50% in the case of MNAR

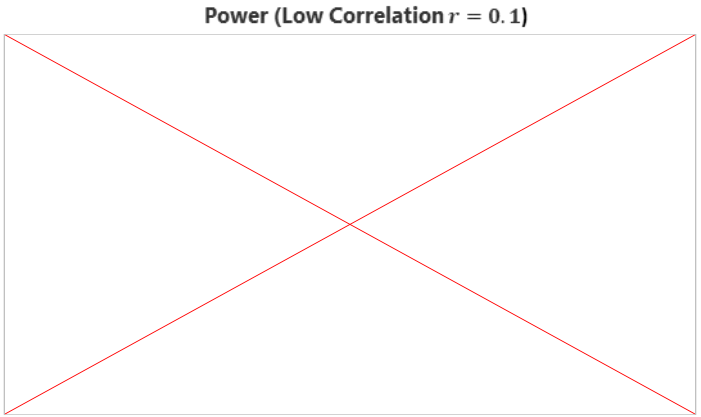
MNAR missingness mechanism depends on unobserved outcome information; all methods are biased.

Results from RCT-Only Analysis

Full Covariate-Adjusted MMRM Exhibits ~85% Power for N=60 – Similar to ML-Based Methods

Protocol MMRM Power in RCT-only





Full covariate-adjusted MMRM and causal forest improve power by 20-35% for N=60

MMRM: effective sample-size gain of 50-70% **(30–42 additional patients)**

Causal Forest: effective sample-size gain of 43-47% **(25–28 additional patients)**

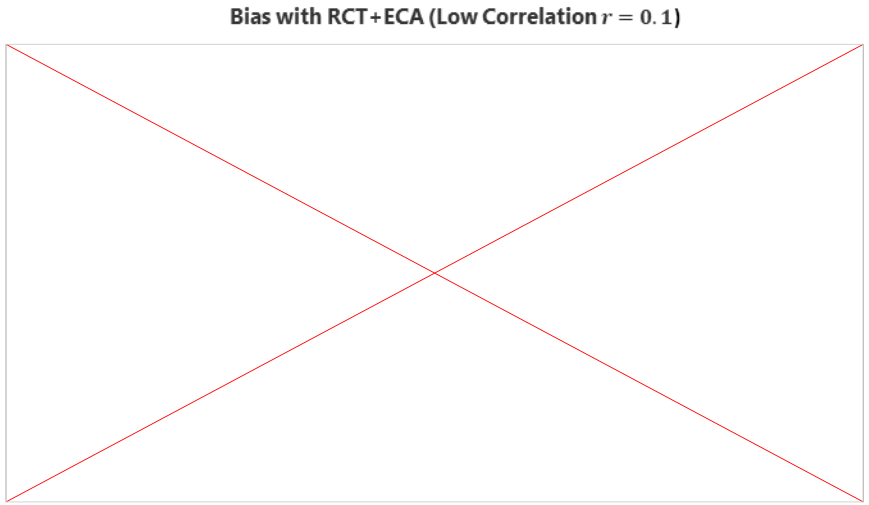
Adjusting for highly prognostic covariates increases power

Under weak correlation, **MMRM** and **Causal Forest** flatten to 20–40% effective sample-size gain (**12–24 additional patients**) at N=60.

sanofi

Results from RCT+ECA Analysis

Most Methods are Inconsistent when Borrowing from External Data with Selection Bias



Non-exchangeable external populations exhibit 10%-300% overestimation of treatment effect

MMRM, TMLE, L-TMLE, matching, and **weighting** are all biased when the ECA has a different distribution than the original RCT population.

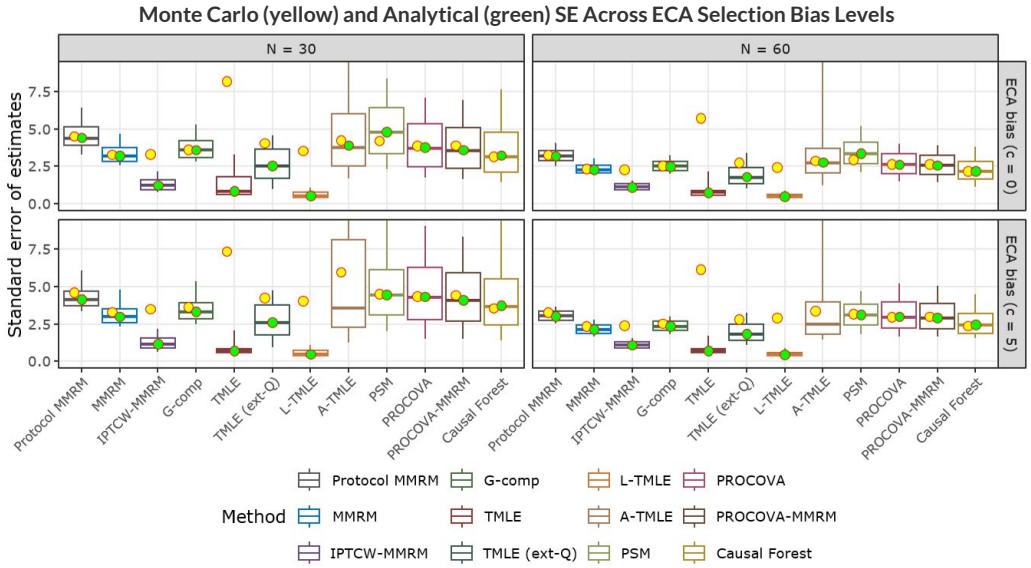
G-computation, A-TMLE, and PROCOVA methods are unbiased in all cases

Even when there is no highly prognostic covariate that may relate the RCT to the ECA, select methods are still unbiased.

sanofi

Results from RCT+ECA Analysis

Some Causal Methods Gain Efficiency and Precision in Presence of External Data



Some methods are more efficient and precise when compared to RCT-only data

G-computation is more stable in ECA scenario, and all methods have less variability with increased sample size.

Methods that involve inverse weighting of source *S* have high variability with ECA data

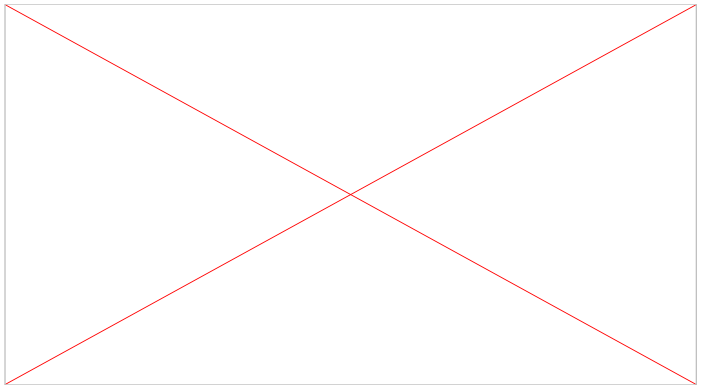
IPTCW and the TMLE family all have high Monte Carlo standard error and analytic standard error likely due to extreme weights. **A-TMLE** has cases where the analytic variance is undefined due to positivity violation in the debiasing step.

sanofi

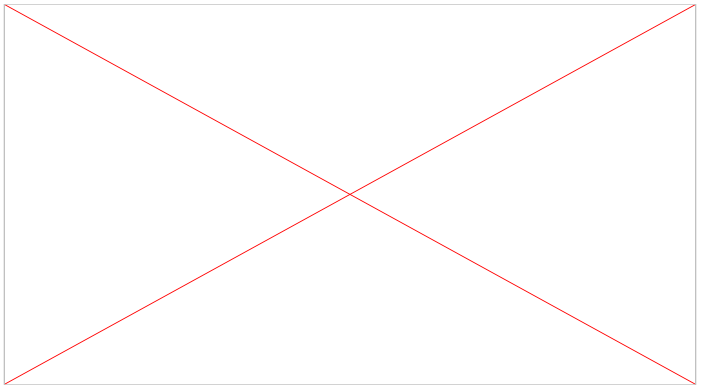
Results from RCT+ECA Analysis

PROCOVA and G-computation: Type I Error Control Across All Scenarios

FPR with RCT+ECA (High Correlation $r = 0.8$)



FPR with RCT+ECA (Low Correlation $r = 0.1$)



Some methods can integrate external sources, but not heterogeneous patient populations

When selection bias depends on highly prognostic determinants, **MMRM, PSM, and Causal Forest have 5% - 20% Type I error inflation**, but have nominal FPR with no selection bias.

PROCOVA, g-computation, and A-TMLE can control Type I error when there is selection bias for ECA depending on N

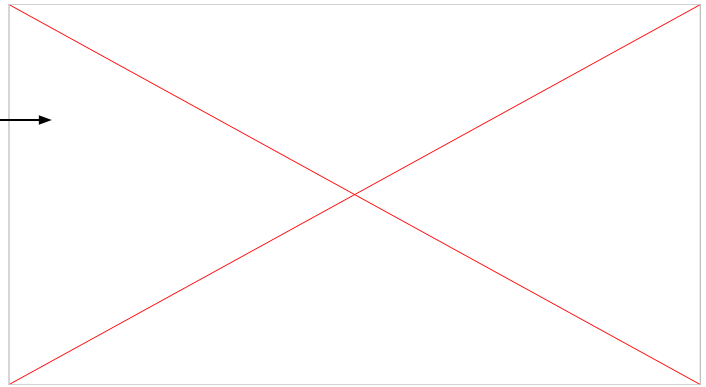
A-TMLE controls when N=60, but slightly error-prone when N=30. **PROCOVA and g-computation** can control Type I error when there is selection bias for ECA. **TMLE** and **L-TMLE** are extremely biased.

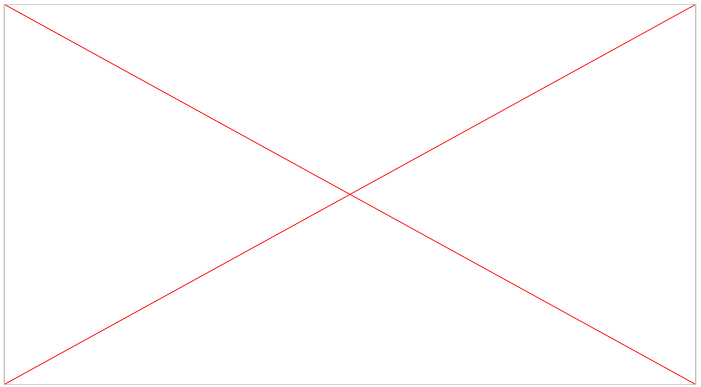
sanofi

Results from RCT+ECA Analysis

ML-Based Methods Improve Power over PROCOVA When Incorporating External Data in Small Samples

Protocol MMRM Power in RCT-only





A-TMLE, PROCOVA family, and G-computation show increased power with type I error control

A-TMLE (35% - 70%) and **G-computation** (48% - 105%) post effective sample-size gains at every scenario, while the **PROCOVA** family reaches 12% - 95%.

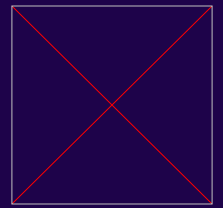
Causal Forest has largest power over all with low correlation + no selection bias

When there is no selection bias, **Causal Forest** controls Type I error and has effective sample-size gain of 85% - 110% (**51 to 66 additional patients**).

sanofi

Simulation Study

RECOMMENDATIONS



Confidential - Sensitive

Evidence-Based Recommendations for Rare Disease Causal Inference

Based on Type I error, power, and bias across the simulation.

RCT-Only

- ✓ **MMRM** -- regulatory standard, nominal Type I error, and can deal with MAR data; covariate adjustment improves power over protocol.
- ✓ **IPTCW/IPW-MMRM** or **Causal Forest** at N >= 60
- ✗ **TMLE / L-TMLE** alone -- inflated Type I error at N=30 and N=60
- ✗ **G-computation** below N=60 -- unstable SE
- ✗ Missingness mechanism needs to be modeled for MNAR data

RCT + ECA, No Bias Suspected

- ✓ **MMRM, PROCOVA / PROCOVA-MMRM, A-TMLE, Causal Forest, PSM, G-comp** -- all near nominal level
- ✗ **TMLE, L-TMLE, IPTCW/IPW-MMRM** -- inflated even with zero systematic bias

RCT + ECA, Bias Suspected

- ✓ **G-computation, PROCOVA / PROCOVA-MMRM or A-TMLE (large N only)** -- the sole methods holding nominal error under bias
- ✗ **Other methods not recommended**: MMRM/IPTW family (~31-44% FPR), TMLE (74%), L-TMLE (85-90%), PSM/Causal Forest (13-21%)

Minimum sample size

N=30 leaves the best methods under ~45% power; N=60 reaches ~75-82%.

Increasing the number of controls through external data improves power, but potentially at a trade-off depending on transportability of ECA and RCT.

sanofi

Confidential - Sensitive

Key Takeaways: Method Choice Matters When N is Small

01 **Type I Error Inflation Is Not Just a Small-N Problem**

TMLE and L-TMLE show inflated Type I error at both N=30 and N=60 -- the driver is anti-conservative variance estimation, not sample size alone.

02 **Simplicity Wins in Rare Disease**

Full covariate-adjusted MMRM outperforms complex ML and semiparametric methods in bias and efficiency for very small samples when analyzing RCT data alone.

03 **ECA Borrowing Is Safe as the Adjustment Strategy**

PROCOVA, A-TMLE, and g-computation hold nominal Type I error and high power under deliberate ECA selection bias; pooling RCT+ECA data - including in **MMRM, TMLE and L-TMLE** - does not.

04 **Missingness Mechanism Should Be Considered in Analysis and Study Design**

All methods are biased in RCT-only analysis with MNAR data; the missingness mechanism needs to be explicitly modeled, and missing data prevention should be considered as a first course before the trial is even conducted.

05 **AI/ML Causal Methods for Efficient Treatment Effect Estimation in Small-Samples**

G-computation and A-TMLE should be considered by agencies as viable alternatives when there are small to moderate sample sizes and non-exchangeable external populations.

sanofi

Future Directions Within and Beyond the Core Simulation

While frequentist adjustments (MMRM, PROCOVA) are the backbone of confirmatory analysis, additional frameworks can help address specific rare-disease and simulation challenges.



1. Bootstrap Variance

Stabilizes the variance estimate for G-computation, TMLE, and other causal methods in small sample sizes.



2. Bayesian Dynamic Borrowing

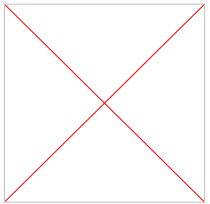
Incorporates historical data as a prior while down-weighting it if inconsistent with current trial data, protecting the Type I error rate.



3. Synthetic Control Arms

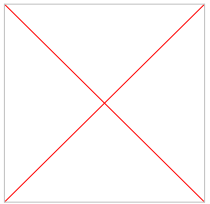
For fatal or ultra-rare conditions where placebo is infeasible, construct control arms from high-quality real-world data with rigorous matching.

Thank You – Q&A



sanofi

Supplementary Slides



sanofi

ECA Borrowing: Exchangeability, Transportability, and the Target Estimand

Borrowing external controls is only valid to the extent the positivity $(0 < P(S = \text{RCT}|A = 0, X) < 1)$ assumption and the assumption below holds.

Target estimand (RCT population)

$$\tau = E[Y(1) - Y(0) \mid S = \text{RCT}]$$

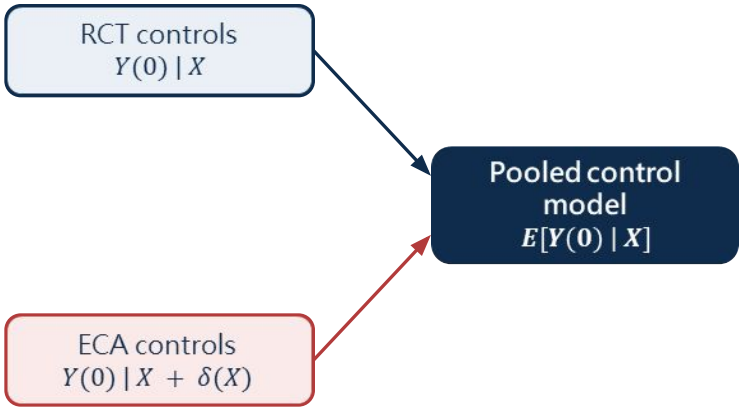
Exchangeability assumption

$$Y(0) \perp S \mid X$$

$$\iff E[Y(0) \mid X, S{=}\text{RCT}] = E[Y(0) \mid X, S{=}\text{ECA}] \quad \forall X$$

If violated: bias enters via $\delta(X)$

$$E[Y(0) \mid X, S{=}\text{ECA}] = E[Y(0) \mid X, S{=}\text{RCT}] + \delta(X)$$



Unbiased pooling requires $\delta(X) \approx 0$ for all X in the RCT's covariate range.

sanofi