

TRIAL-LIKE VS REAL-WORLD EVIDENCE

A reproducible synthetic-data demonstration of eligibility, population shift and transportability

Technical report

Planimeter Kft.

Version 1.0.0

23 September 2026

Methodological demonstration only. All patient records used in this analysis are synthetic. The results are not clinical estimates and should not be used for patient-level decision-making.

Executive summary

Objective. To demonstrate how trial-style eligibility criteria can reshape a broader healthcare population and to examine whether an illustrative prognostic association remains directionally consistent across trial-like and real-world populations.

Data. A 20,000-patient synthetic healthcare population derived from Synthea was transformed into patient-level eligibility features and a hypertension outcome cohort. The public release contains two compact analysis-ready CSV files rather than the full source tables.

Key result. The final eligibility rules retained 1,604 of 20,000 synthetic patients (8.0%). The hypertension outcome cohort contained 3,676 patients, including 650 trial-like and 3,026 RWE-only patients. Five-year diabetes incidence was 12.5% in hypertension-only patients and 15.8% in patients with baseline prediabetes. The unadjusted association was OR 1.32 (95% CI 1.09-1.60) in the full cohort, OR 1.69 (0.49-5.83) in the trial-like subset and OR 1.23 (1.01-1.50) in the RWE-only subset. The prediabetes-by-trial-eligibility interaction had $p = 0.622$.

Interpretation. The example illustrates two distinct issues: eligibility criteria alter the composition and absolute risk of the analyzed population, while an association may remain directionally similar across populations. The interaction result should be interpreted as absence of evidence for effect modification in this synthetic example, not as proof of equivalence or transportability.

1. Data and analytical framework

1.1 Synthetic source data

The source data consist of synthetic patient, condition, medication and encounter histories generated with Synthea. Patient-level analytical variables were derived by linking these sources. The distributed datasets use sequential public patient identifiers; original synthetic UUIDs are not required for the released analyses.

1.2 Reproducibility

The development script originally calculated age using the system date. In the public release, the analysis date is fixed to the observed data horizon: 9 June 2026, the date of the latest encounter in the supplied source data. This removes calendar-time dependence from trial-eligibility classification.

1.3 Trial-like eligibility

The trial-like cohort is illustrative rather than protocol-specific. Eligibility combines age, follow-up, cardiovascular-risk, comorbidity, medication-burden, major-disease exclusion and healthcare-utilization criteria. These rules are intended to demonstrate sequential population selection, not to emulate a named clinical trial.

Selection step	Patients	% of original
All patients	20,000	100.0%
Age window	9,558	47.8%
CV risk	5,867	29.3%
Comorbidity	2,757	13.8%
Medication burden	1,958	9.8%
No major exclusions	1,757	8.8%
Trial eligible	1,604	8.0%

1.4 Hypertension outcome cohort

The index date is the first recorded hypertension diagnosis. Patients aged 40-85 years at index were retained. Baseline prediabetes was defined as a recorded prediabetes condition on or before the hypertension index date. The outcome was a diabetes diagnosis after index and within five years, excluding prediabetes and gestational diabetes descriptions.

1.5 Statistical analysis

- Five-year diabetes incidence was summarized by baseline prediabetes status.
- Unadjusted logistic regression estimated the odds ratio for baseline prediabetes in the full hypertension cohort, the trial-like subset and the RWE-only subset.
- An interaction model included prediabetes x trial eligibility to examine evidence of association heterogeneity across the two population definitions.
- A secondary adjusted logistic model included age, gender, number of conditions and number of medications. It is retained in the released code as an exploratory analysis but is not the primary transportability result.

2. Results

2.1 Population selection

Sequential eligibility criteria reduced the 20,000-patient source population to 1,604 trial-eligible patients. The largest reductions occurred before the final exclusion and utilization steps, showing how multiple moderate filters can cumulatively create a substantially different analysis population.

2.2 Hypertension cohort and absolute risk

The hypertension outcome cohort contained **3,676** patients: **650 trial-like** and **3,026 RWE-only**. Five-year diabetes incidence was 12.5% among patients without baseline prediabetes and 15.8% among those with baseline prediabetes, an absolute difference of approximately 3.4 percentage points.

Group	N	Events	5-year incidence
Hypertension only	1,533	191	12.5%
Hypertension + prediabetes	2,143	339	15.8%

2.3 Association estimates

Population	N	OR	95% CI	p-value
Full HTN cohort	3,676	1.32	1.09-1.60	0.0043
Trial-like	650	1.69	0.49-5.83	0.406
RWE-only	3,026	1.23	1.01-1.50	0.0356

The point estimates were above 1 in all three populations, while precision differed substantially. The trial-like estimate had a wide confidence interval because the subset was smaller and contained fewer outcome events. Statistical significance therefore should not be used as the criterion for deciding whether the estimates are compatible across populations.

2.4 Interaction analysis

The prediabetes x trial-eligibility interaction had $p = 0.622$. In this synthetic example, the model therefore provided no evidence that the prediabetes association differed by trial-like eligibility status. This result does not establish equivalence; the interaction test is limited by the amount of information available in the smaller trial-like subgroup.

3. Interpretation

The analysis is designed to separate population composition from association estimation. Trial-like eligibility changes who remains in the cohort and can change baseline risk, event frequency and statistical precision. A relative association can nevertheless remain directionally similar across selected and broader populations. Demonstrating that similarity requires direct comparison of effects rather than comparison of separate p-values.

4. Limitations

- All records are synthetic. The numerical results are methodological examples and have no direct clinical interpretation.
- Eligibility rules are illustrative and are not derived from a specific trial protocol.
- Disease flags are based on text matching in synthetic condition descriptions and are therefore simplified phenotypes.

- The logistic models are intentionally simple. They do not address all sources of confounding, censoring, selection bias or outcome misclassification that arise in real RWE studies.
- The interaction analysis tests evidence of heterogeneity; a non-significant interaction is not proof that effects are identical or transportable.
- Odds ratios and absolute risks answer different questions. Population shifts may alter absolute risk even when relative associations appear similar.

5. Reproducibility package

The public package contains two analysis-ready CSV files, the complete R analysis script, a separate data-preparation script, validated output tables, a data dictionary and the GPL license text. All paths in the released code are relative.

File	Purpose
rwe_patient_features.csv	Eligibility funnel and patient-level trial-like criteria.
rwe_htn_outcome.csv	Risk, regression and transportability analyses.
rwe_analysis_planimeter_v1.0.0.R	Primary reproducible analysis and six figures.
rwe_data_preparation_planimeter_v1.0.0.R	Transformation from Synthea-derived source tables to public CSVs.
README.md / DATA_DICTIONARY.md	Execution notes, provenance and variable definitions.

6. License and AI use

The Planimeter R source code is released under the GNU General Public License v3.0 or later (GPL-3.0-or-later). Generative AI was used as a support tool during code development, review and validation. The analytical design, final code, interpretation and responsibility for the released materials remain with Planimeter Kft.

Source

Synthea - Synthetic Patient Population Simulator. The MITRE Corporation.

<https://github.com/synthetichealth/synthea>