

# LEGACY STUDY HARMONIZATION AND PATIENT-LEVEL INTEGRATION

From fragmented historical studies to a reusable  
integrated evidence dataset

## Technical report

Planimeter Kft.

Version 1.0.0

24 September 2026

**Methodological demonstration.** The six legacy studies are synthetic transformations sampled from one teaching dataset. They illustrate harmonization challenges; they are not six independent clinical trials.

## Executive summary

**Objective.** To demonstrate how heterogeneous legacy study data can be standardized into a common patient-level structure and how that structure enables analyses that extend beyond study-level meta-analysis.

**Data.** Six synthetic legacy breast-cancer study versions are generated from the gbsg teaching dataset in the R survival package. The versions intentionally differ in endpoint horizon, time unit, treatment coding and selected covariate representations.

**Key result.** Development outputs show fixed-effect meta-analysis HR 0.80 (95% CI 0.66-0.97), random-effects HR 0.78 (0.57-1.08), and patient-level pooled HR 0.78 (0.65-0.95). The pooled estimate therefore preserves the study-level evidence while the harmonized database supports common endpoint re-derivation and patient-level prediction.

**Interpretation.** The value of patient-level harmonization is not that it necessarily changes the headline treatment effect. Its main contribution is the creation of a reusable analytical asset from which common endpoints, adjusted models and future prediction tasks can be derived.

## 1. Demonstration design

### 1.1 Source population

The source population is the gbsg breast-cancer teaching dataset distributed with the R package survival. The released code accesses the dataset directly from the installed package, so no large external database is required for reproduction.

### 1.2 Six deliberately heterogeneous legacy studies

A fixed seed (20260610) is used to sample six study versions. Each version is then transformed into a different legacy-style structure. The heterogeneity is deliberate and mirrors common integration problems: incompatible field names, different units, alternative coding conventions, inconsistent endpoint horizons and lossy covariate representations.

| Study   | Generated design                  |
|---------|-----------------------------------|
| Study A | N=260; 5Y_RECURRENCE; time=Years  |
| Study B | N=230; 3Y_RECURRENCE; time=Months |
| Study C | N=210; 5Y_RECURRENCE; time=Days   |
| Study D | N=190; 5Y_DEATH_PROXY; time=Years |
| Study E | N=170; 3Y_RECURRENCE; time=Months |
| Study F | N=150; 5Y_RECURRENCE; time=Years  |

### 1.3 Common data model

All six source structures are mapped to a common analytical model containing study provenance, subject ID, age, treatment, tumor size, nodes, grade, analysis time and event status. Study-specific units are converted to years and treatment codes are mapped to Control/Treatment.

Two mappings are intentionally lossy to demonstrate a realistic legacy-data problem: Study C converts tumor-size categories to representative values (15/35/65 mm), and Study E converts node-count categories to 0/2/5. These are educational choices, not recommended clinical imputation rules.

### 1.4 Endpoint heterogeneity

The synthetic studies use 3-year recurrence, 5-year recurrence, and one 5-year death-proxy endpoint. In the demonstration source code, the death proxy is deliberately set equal to the recurrence indicator. This is not a mortality analysis; it exists solely to make endpoint non-equivalence visible.

## 2. Analytical framework

### 2.1 Study-level analysis and meta-analysis

Each legacy study is first analysed independently using a Cox proportional-hazards model with treatment as the predictor. Study-specific log hazard ratios and standard errors are then combined using an inverse-variance fixed-effect estimator and a DerSimonian-Laird-style random-effects estimator.

### 2.2 Patient-level pooled analysis

The harmonized patient-level database is analysed using a stratified Cox model with a common treatment coefficient and study-specific baseline hazards. This preserves study provenance while estimating a pooled treatment effect on a common scale.

### 2.3 Common endpoint re-derivation

After harmonization, common 3-year and 5-year event definitions are re-derived across all study records using the same algorithm. This is the central capability that study-level summary data alone cannot provide.

## 2.4 Predictive modelling

The harmonized data are also used to train two random-forest models for re-derived 3-year and 5-year endpoints. Predictors are study ID, age, tumor size, nodes, grade and treatment. A fixed 70/30 train-test split and 500-tree ranger models are used.

## 3. Results

### 3.1 Treatment-effect synthesis

| Method                        | Development-run result     |
|-------------------------------|----------------------------|
| Fixed-effect meta-analysis    | HR 0.80 (95% CI 0.66-0.97) |
| Random-effects meta-analysis  | HR 0.78 (95% CI 0.57-1.08) |
| Patient-level pooled analysis | HR 0.78 (95% CI 0.65-0.95) |

The patient-level pooled estimate is essentially aligned with the study-level synthesis. In this demonstration, harmonization does not manufacture a different treatment effect; it preserves the evidence while creating additional analytical options.

### 3.2 Re-derived endpoints

The integrated structure allows the same 3-year and 5-year event rules to be applied after pooling. This separates the source-study endpoint definitions from the common analytical endpoints and makes sensitivity to endpoint horizon directly examinable.

### 3.3 Prediction after harmonization

| Prediction target          | Development-run AUC |
|----------------------------|---------------------|
| Re-derived 3-year endpoint | approximately 0.84  |
| Re-derived 5-year endpoint | approximately 0.78  |

These results are methodological demonstrations rather than validated prognostic models. Their purpose is to show that patient-level integration creates a reusable modelling substrate, not to establish clinical predictive performance.

## 4. What patient-level harmonization adds beyond meta-analysis

| Capability                           | Study-level meta-analysis   Patient-level harmonization |
|--------------------------------------|---------------------------------------------------------|
| Common treatment-effect synthesis    | Yes   Yes                                               |
| Forest-plot evidence summary         | Yes   Yes                                               |
| Re-derived common 3-year endpoint    | No   Yes                                                |
| Re-derived common 5-year endpoint    | No   Yes                                                |
| Patient-level covariate modelling    | No   Yes                                                |
| Predictive modelling                 | No   Yes                                                |
| Reusable integrated evidence dataset | No   Yes                                                |

Meta-analysis and patient-level harmonization therefore answer different but complementary needs. Meta-analysis is an efficient evidence-synthesis framework when only study-level effects are required. Patient-level integration becomes valuable when the organization wants to re-derive endpoints, revisit covariate definitions, build adjusted models, develop prediction tools or reuse historical data for new questions.

## 5. Limitations and interpretation

- The six legacy studies are simulated transformations of one source dataset, not independent trials or cohorts.
- Treatment is derived from the source teaching variable `hormon`; causal treatment-effect interpretation is not warranted.
- Study D uses recurrence as a death proxy solely to demonstrate endpoint incompatibility.
- Study C tumor-size categories and Study E node categories are mapped to representative numeric values, deliberately illustrating information loss.
- The simple DerSimonian-Laird-style random-effects calculation is suitable for this demonstration but is not a universal recommendation for meta-analysis.
- The random-forest models use a single fixed train-test split and are not externally validated.
- Development-run numerical results are reported from the supplied source materials. The cleaned release scripts were not executed in the packaging environment because an R runtime was unavailable; one final local run is required before publication.

## 6. Reproducibility

The release is intentionally compact. The public `survival::gbsg` dataset is the only source data dependency. The first R script generates the six heterogeneous study files and the harmonized patient-level CSV. The second script reproduces the meta-analysis, pooled Cox model, endpoint re-derivation, random-forest prediction, calibration outputs and final figures. All paths are relative.

The package also contains compact specification and harmonization-map CSVs, plus reference development results for the headline hazard ratios and predictive AUCs. These reference values are intended as run checks after local execution.

## 7. AI use statement

Generative AI tools were used as a support tool during drafting, code refinement and package preparation. The analytical design, interpretation, final materials and responsibility for the released version remain with Planimeter Kft.

## References

- Therneau TM. `survival`: Survival Analysis. R package. Dataset: `gbsg` (German Breast Cancer Study Group teaching data).
- DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials*. 1986;7(3):177-188.
- Planimeter Kft. Legacy Study Harmonization and Patient-Level Integration. *Biostat Insight Series*.