

FROM SUBGROUPS TO PRECISION MEDICINE

Classical subgroup analysis, continuous effect
modification
and patient-level model contrasts

Technical report

Planimeter Kft.

Version 1.0.0

24 September 2026

Methodological demonstration only. Patient-specific scores are in-sample model-based relative hazard contrasts, not validated individualized causal treatment effects. The historical PBC treatment coding is documented explicitly below.

Executive summary

Objective. To demonstrate why classical one-variable-at-a-time subgroup estimates can be unstable, how dichotomizing continuous variables discards information, and how effect-modification models can represent heterogeneity on a continuous or patient-level scale.

Data. The analysis uses randomized participants from the Mayo Clinic primary biliary cholangitis (PBC) trial dataset distributed with the R survival package. After complete-case selection for the variables used in the demonstration, 310 observations and 124 deaths remain.

Key result. The overall analytical arm contrast was HR 0.962 (95% CI 0.676-1.369; $p=0.831$). Classical subgroup estimates were imprecise and none of the released subgroup-specific effects had nominal $p<0.05$. The continuous bilirubin interaction coefficient was 0.191 ($p=0.281$), illustrating a smooth modelled pattern but not strong evidence of effect modification.

Interpretation. The example supports a methodological shift from arbitrary cutoffs and repeated subgroup estimates toward explicit interaction modelling. The patient-level contrasts generated by the final model are exploratory model outputs; this analysis does not establish clinically actionable individualized treatment effects.

1. Data and analytical framework

1.1 Source data

The source is `survival::pbc`, a teaching dataset based on the Mayo Clinic randomized placebo-controlled trial in primary biliary cholangitis. The original dataset contains 312 randomized trial participants plus additional non-randomized observations. The case study retains participants with non-missing randomized treatment assignment and then removes rows with missing values among the selected modelling variables.

Characteristic	Value
Prepared analysis dataset	310 observations
Deaths used as events	124
Control analysis label	157
Treatment analysis label	153
Median bilirubin	1.4 mg/dL

1.2 Historical treatment coding

Important reproduction note. In the source `survival::pbc` documentation, `trt=1` denotes D-penicillamine and `trt=2` denotes placebo. The original Insight script used generic labels in the opposite semantic direction: `trt=1` was named Control and `trt=2` Treatment. Version 1.0.0 preserves those labels only to reproduce the published numerical contrast. They should not be read as the historical active-treatment and placebo names.

1.3 Outcome and covariates

Death is treated as the event (`original status=2`); transplantation is not counted as the event in this demonstration. Follow-up time is converted from days to years. Classical subgroup variables include median-split bilirubin, age and albumin, as well as edema and histologic stage. The continuous models use log bilirubin and additional baseline covariates.

2. From classical subgroups to continuous effect modification

2.1 Overall and subgroup-specific Cox estimates

The overall Cox model shows no clear average difference between the two generic analytical arm labels. Splitting the same dataset into subgroups widens uncertainty because each estimate is based on less information. The table below shows selected results from the released forest-plot dataset.

Population	N	HR	95% CI	p
Overall	310	0.96	0.68-1.37	0.831
bili_group: Higher bilirubin	156	0.95	0.63-1.43	0.804
bili_group: Lower bilirubin	154	1.25	0.58-2.67	0.566
age_group: Older	154	1.15	0.74-1.80	0.527
age_group: Younger	156	0.87	0.49-1.57	0.652
stage: 4	108	0.88	0.54-1.46	0.626
stage: 3	120	1.10	0.60-2.02	0.747
stage: 2	66	0.65	0.24-1.80	0.412

No subgroup-specific coefficient in the released classical analysis has a nominal p-value below 0.05. More importantly, the purpose of the example is not to search for a significant subgroup: it is to show how repeated categorical splits can produce unstable, disconnected effect estimates.

2.2 The cost of dichotomizing bilirubin

Bilirubin is first split at the sample median. The lower-bilirubin group yields HR 1.25 (95% CI 0.58-2.67), while the higher-bilirubin group yields HR 0.95 (95% CI 0.63-1.43). These two estimates are easy to display, but the cutoff treats nearby patients on opposite sides of the median as qualitatively different.

Median-split bilirubin group	N	HR	95% CI
Lower bilirubin	154	1.25	0.58-2.67
Higher bilirubin	156	0.95	0.63-1.43

2.3 Continuous bilirubin effect modification

The continuous Cox interaction model replaces the low/high split with treatment-by-log-bilirubin interaction. The estimated interaction coefficient is 0.191 (p=0.281). Across the central 94% of the observed bilirubin distribution, the modelled analytical arm HR increases smoothly from approximately 0.73 at bilirubin 0.4 mg/dL to approximately 1.50 at 17.3 mg/dL.

The continuous curve is a more information-preserving representation than a median split, but the wide confidence intervals and non-significant interaction remain important. A smooth curve should not be confused with evidence that the underlying biological interaction has been established.

3. Patient-level heterogeneous treatment-effect modelling

3.1 Multivariable interaction model

The final demonstration fits a Cox model containing treatment interactions with log bilirubin, age, albumin, edema and histologic stage. For each observation, the fitted linear predictor is calculated twice: once under the Control analysis label and once under the Treatment analysis label. Their difference is exponentiated to produce a patient-specific relative hazard contrast.

The released variable `benefit_score` is defined as $1 - \text{HR}$. This transformation is useful for visualization but is not a probability, an absolute risk reduction or a validated clinical benefit measure.

3.2 Distribution of model-based contrasts

Using the illustrative thresholds inherited from the development analysis, the 310 observations are partitioned as follows:

Illustrative model category	Definition	N
High predicted benefit	model HR < 0.75	125
Moderate predicted benefit	0.75 <= model HR < 1.00	67
Low or no predicted benefit	model HR >= 1.00	118

Interpretation boundary. These categories are in-sample outputs from a flexible interaction model fitted to 310 observations. They were not cross-validated or externally validated, and the source treatment labels are generic reproduction labels. They therefore cannot support individualized treatment recommendations.

3.3 What changes relative to a forest plot?

A traditional subgroup forest plot asks for separate treatment-effect estimates inside a series of categories. The interaction model instead specifies a structured relationship between baseline characteristics and the analytical arm contrast. This is closer to the logic of modern HTE modelling, but stronger modelling assumptions create stronger validation requirements.

- Continuous moderators avoid information loss from arbitrary cutoffs.
- Multivariable models can combine several baseline characteristics simultaneously.
- Patient-level model outputs can describe heterogeneity as a distribution rather than a set of disconnected subgroup estimates.
- Individualized interpretation requires validation, clinically meaningful effect scales and careful causal assumptions.

4. Interpretation and limitations

The case study is best read as a methodological progression, not as an efficacy analysis of the historical PBC trial and not as a precision-medicine model ready for use. Several limitations are material to interpretation:

- The complete-case analysis contains only 310 observations and 124 deaths, limiting the stability of interaction estimates.
- Classical subgroup estimates are exploratory and no multiplicity adjustment is applied across the repeated subgroup comparisons.
- Age, albumin and bilirubin median splits are deliberately simple and illustrate the arbitrariness of dichotomization.
- The continuous bilirubin interaction is not statistically compelling in this dataset ($p=0.281$).
- The multivariable interaction model is evaluated in-sample; there is no cross-validation, shrinkage assessment or external validation.
- The patient-specific score is a relative hazard contrast, not an absolute treatment effect. It does not directly quantify how many events would be prevented for an individual patient.
- Proportional-hazards and functional-form assumptions are not the focus of this demonstration and should be examined in a substantive survival analysis.
- The generic Control/Treatment labels intentionally preserve the original Insight calculation but do not match the historical D-penicillamine/placebo names in `survival::pbc`.

5. Reproducibility and validation

5.1 Released workflow

The release separates data preparation from analysis. The first script reconstructs the compact analysis dataset from `survival::pbc`. The second script reads that fixed dataset, reproduces the reported Cox models and patient-level contrasts, writes analysis tables and creates the six publication figures used in the Insight narrative.

File	Role
subgroup_hte_data_preparation_planimeter_v1.0.0.R	Rebuilds the compact PBC analysis dataset
subgroup_hte_analysis_planimeter_v1.0.0.R	Reproduces models, tables and figures
data/pbc_hte_analysis_dataset.csv	Fixed compact analysis dataset
reference/*.csv	Reference outputs from the development workflow
output/validation_checks.csv	Independent release-validation checks

5.2 Validation results

All 56 automated release checks passed. The checks cover dataset dimensions and completeness, event and arm counts, cross-file consistency, the algebra of the continuous bilirubin curve, the patient-level score transformation and the category rules.

As an independent numerical check, the classical Cox models were refitted in Python using statsmodels proportional-hazards regression. The resulting hazard ratios and confidence intervals agreed with the R reference outputs within the prespecified numerical tolerance. Small differences are expected across implementations because survival-model fitting and tie handling can differ slightly.

6. Reproducibility notes and references

6.1 Software and source data

The underlying data are distributed with the R survival package. The analysis code uses survival for Cox proportional-hazards modelling and standard tidyverse-style packages for data management and figure generation. Relative paths are used throughout the release package; no local machine directories are required.

6.2 Selected references

- Therneau T, Grambsch P. Modeling Survival Data: Extending the Cox Model. Springer, 2000. The survival::pbc documentation describes the Mayo Clinic PBC data and original treatment coding.
- Cox DR. Regression Models and Life-Tables. Journal of the Royal Statistical Society: Series B. 1972;34(2):187-220.
- Kent DM, Paulus JK, van Klaveren D, et al. The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. Ann Intern Med. 2020;172(1):35-45. doi:10.7326/M18-3667.
- Kent DM, Steyerberg E, van Klaveren D. Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. BMJ. 2018;363:k4245. doi:10.1136/bmj.k4245.

6.3 AI use statement

Generative AI tools were used as support tools during drafting, code refinement, documentation and validation. Planimeter Kft. retained responsibility for the analytical design, final code, interpretation and released materials.

6.4 Release status

Version 1.0.0 freezes the analytical logic and reference outputs used for the reproducibility package. The cleaned release code deliberately preserves the original numerical contrast while documenting the historical treatment-coding issue and the distinction between model-based patient-level contrasts and validated causal treatment effects.