# Case study: Diagnosing bacteremia

**Michael Kammer,** Georg Heinze, Aliaksandr Hubin, Mariana Nold

*michael.kammer@meduniwien.ac.at*

# Research question in one sentence

**We adress a prediction task:**

**Build a diagnostic model** for the presence of bacteria in the blood stream (i.e. bacteremia) using a dataset comprising missing data and estimate the **out-of-sample performance reflecting all sources of uncertainty**.

Ratzinger et al (2014), 'A risk prediction model for screening bacteremic patients: a cross sectional study', PLoS One 9(9), e106765.

# Research question: some background

**Bacteremia is a serious clinical condition in susceptible patients**

- High mortality rate (often due to sepsis, 14% – 37%)
- Gold standard for diagnosis: blood culture, which takes time and is costly
- Decision to conduct a blood culture not trivial:
  - False positive rate (e.g. contamination) is not negligible
  - Cost-effectivness an important consideration

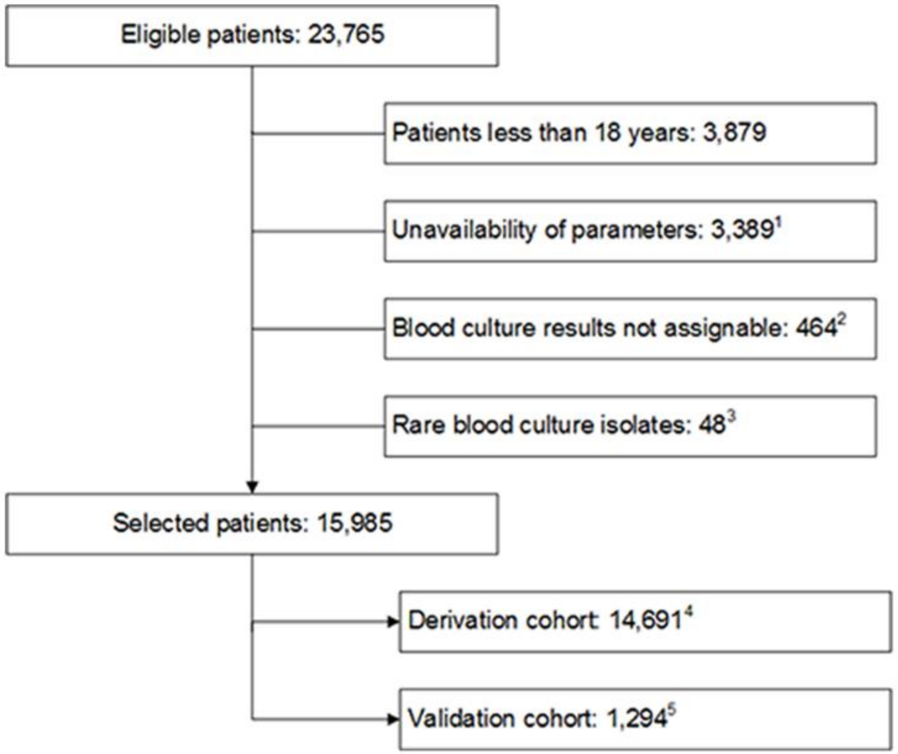
**A diagnostic model as pre-test may help**

Ratzinger et al (2014), 'A risk prediction model for screening bacteremic patients: a cross sectional study', PLoS One 9(9), e106765.

# The data

**Available at https://zenodo.org/records/7554815**

- In- and outpatients from General Hospital Vienna
  - Between Jan 2006 – Dec 2010
  - Clinical suspicion of bacteremia
- Total derivation cohort available 14691
  - **Note: we use only 4000 here for simplicity!**
- Zenodo data slightly modified for privacy




Ratzinger et al (2014), 'A risk prediction model for screening bacteremic patients: a cross sectional study', PLoS One 9(9), e106765.

# The data: some prior knowledge

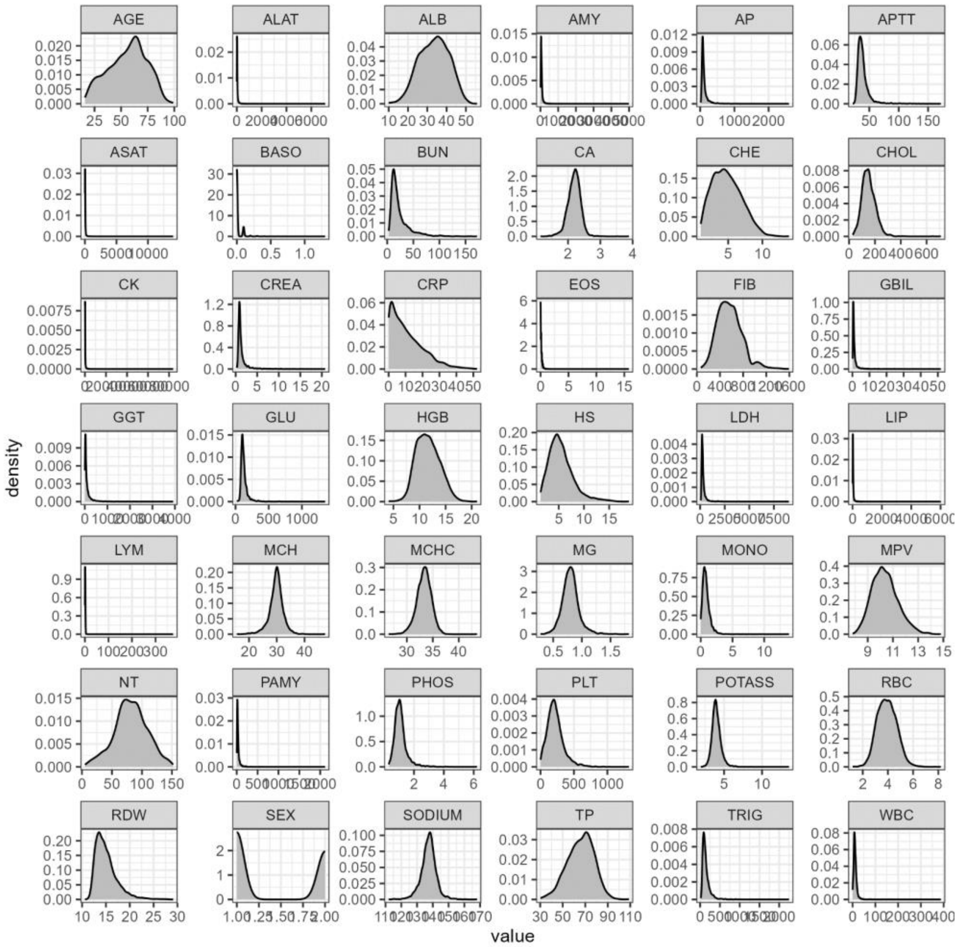
**51 variables, except sex all laboratory data (and age) are continuous**

- Some variables have known biological links (used to exclude them)
  - Leucocytes have different types and are measured in different ways
- Some variables known to be important (but we mostly ignore that here)
  - Several known from literature
  - Several more from discussion with clinical partners
- The outcome is binary: **bacteremia yes (~8%) / no**

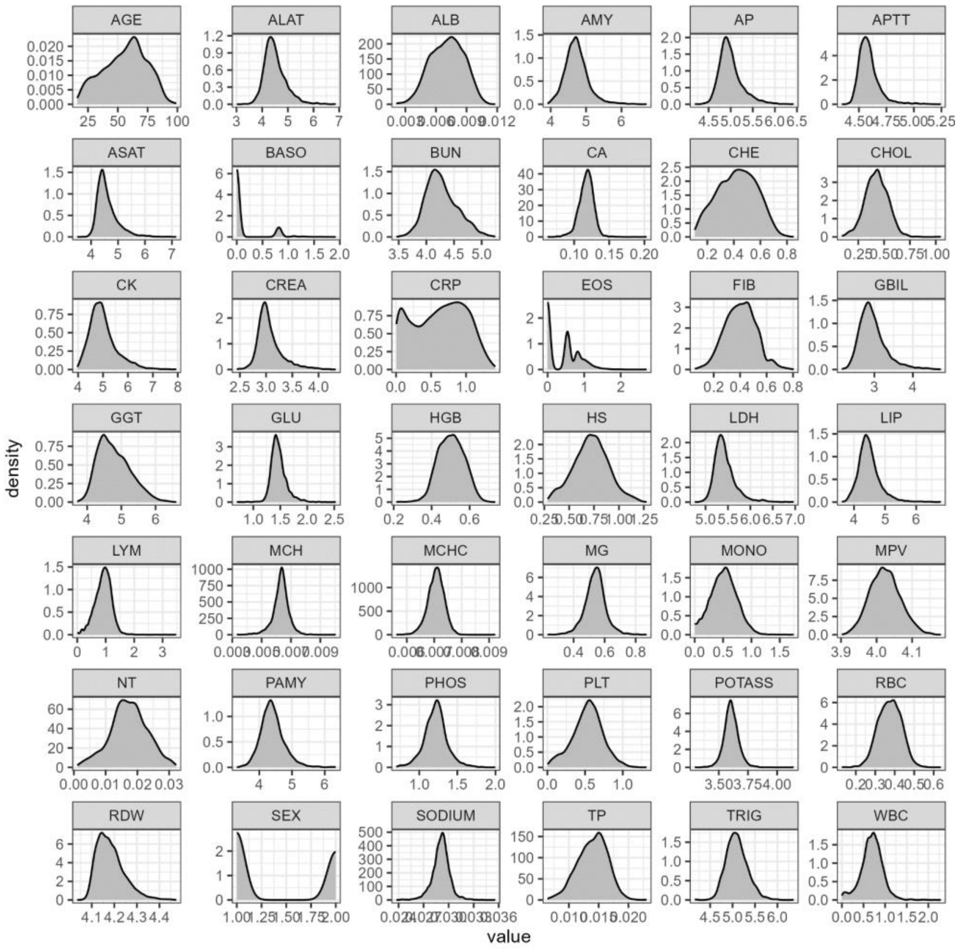
Ratzinger et al (2014), ‘A risk prediction model for screening bacteremic patients: a cross sectional study’, PLoS One 9(9), e106765.

# Initial data analysis

Original

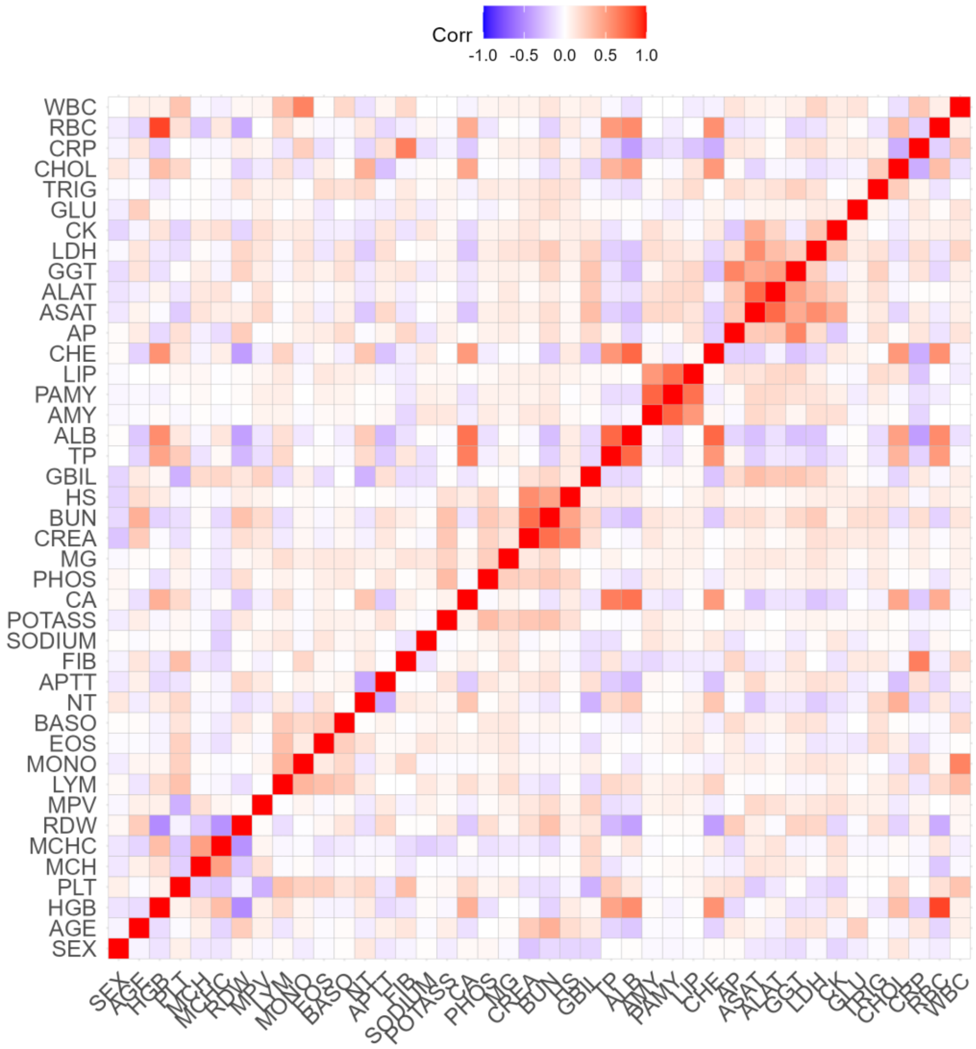


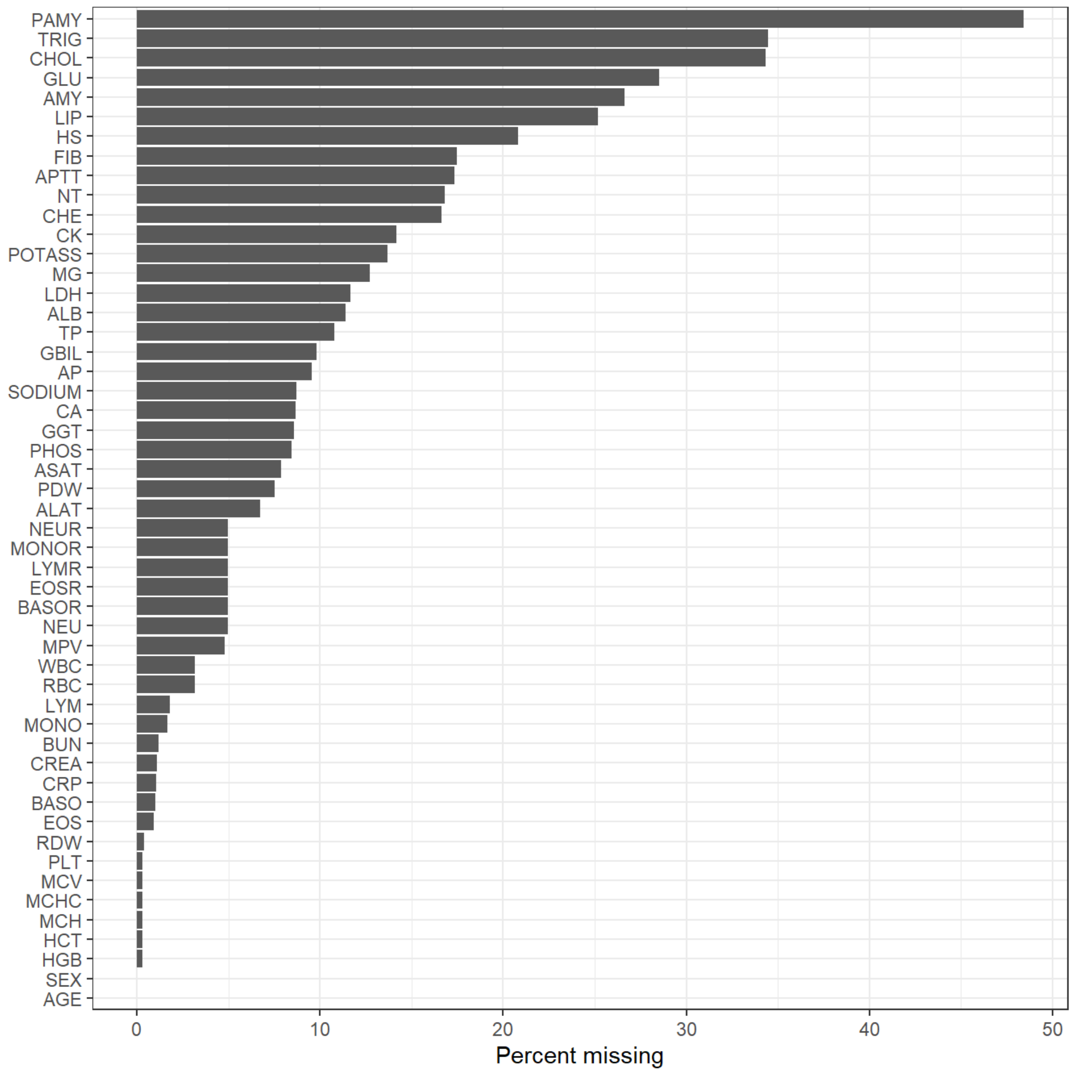

Pseudolog $\frac{\operatorname{asinh}\left(\frac{x}{2\sigma}\right)}{\log(b)}$




Heinze G et al. Regression without regrets –initial data analysis is a prerequisite for multivariable regression. BMC Medical Research Methodology. 2024;24(1):178.

# Initial data analysis



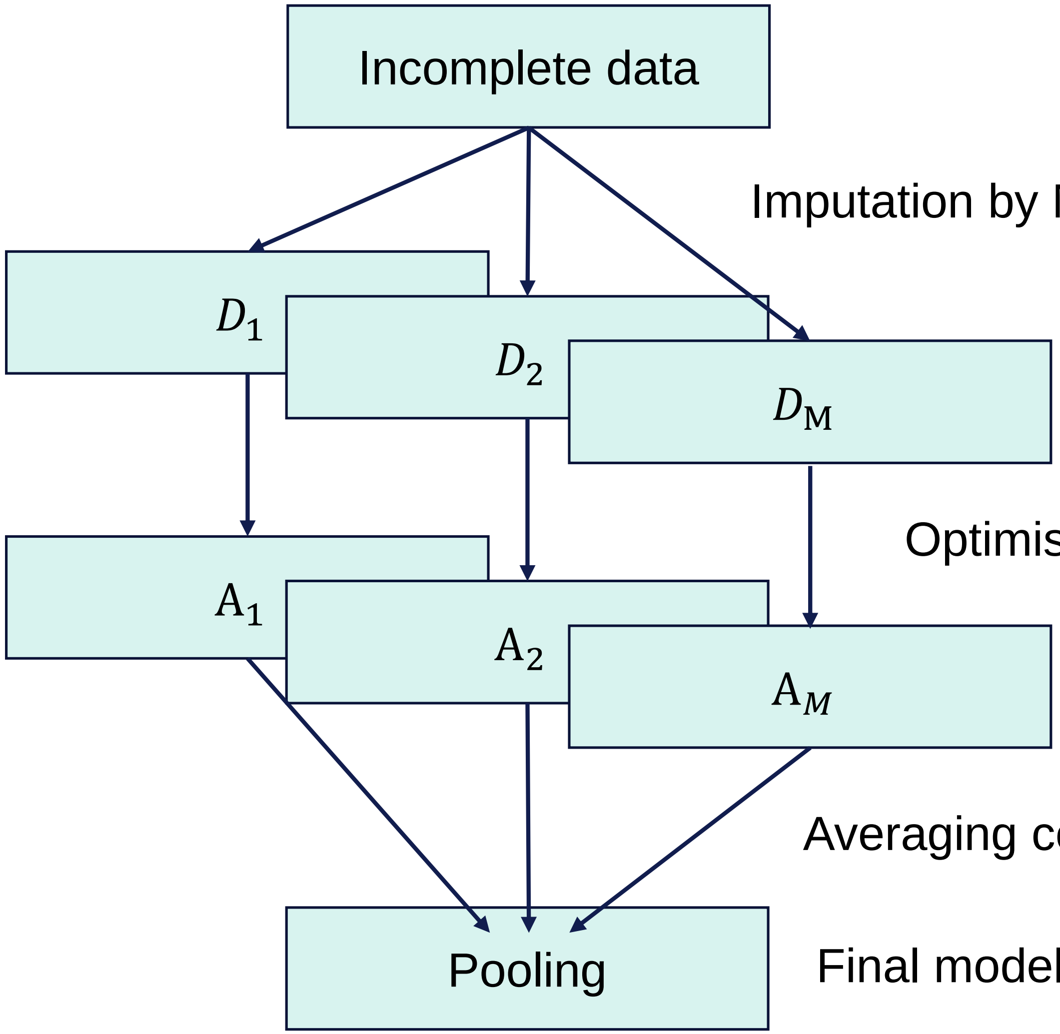



# Statistical analysis plan: sounds simple…

**We adopt a pragmatic Frequentist point of view**

- Use multiple imputation to handle missing data (i.e. to estimate the increased *uncertainty of the results* due to missing data)
- Use Logistic Ridge regression to fit diagnostic model for bacteremia
  - Optimise AUC via 5-fold cross-validation
  - Use all available covariates, linear effects only, no interactions
- Compute out-of-sample performance via bootstrap

# Statistical analysis plan: sounds simple…




Imputation by MICE using a single *imputation* model

M multiply imputed datasets for analysis

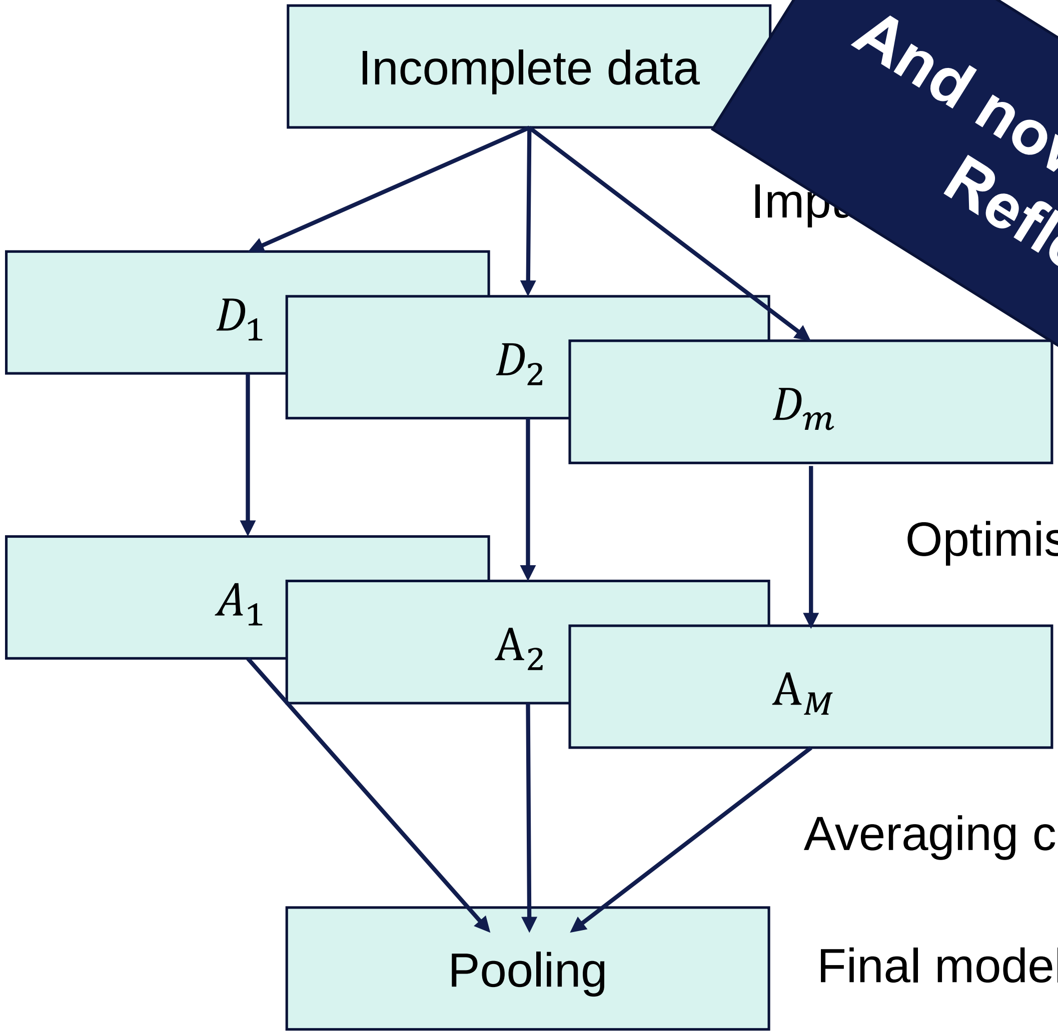
Optimisation of „*substantive*“ model (analyse $D_i$ separately)

M fitted Logistic Ridge regression models

Averaging coefficients (and standard errors)

Final model incorporating uncertainty due to missingness

# Statistical analysis plan: sounds simple…




Imp… a single *imputation* model

m… ets for analysis

Optimisation of „*subst*… $D_i$ separately)

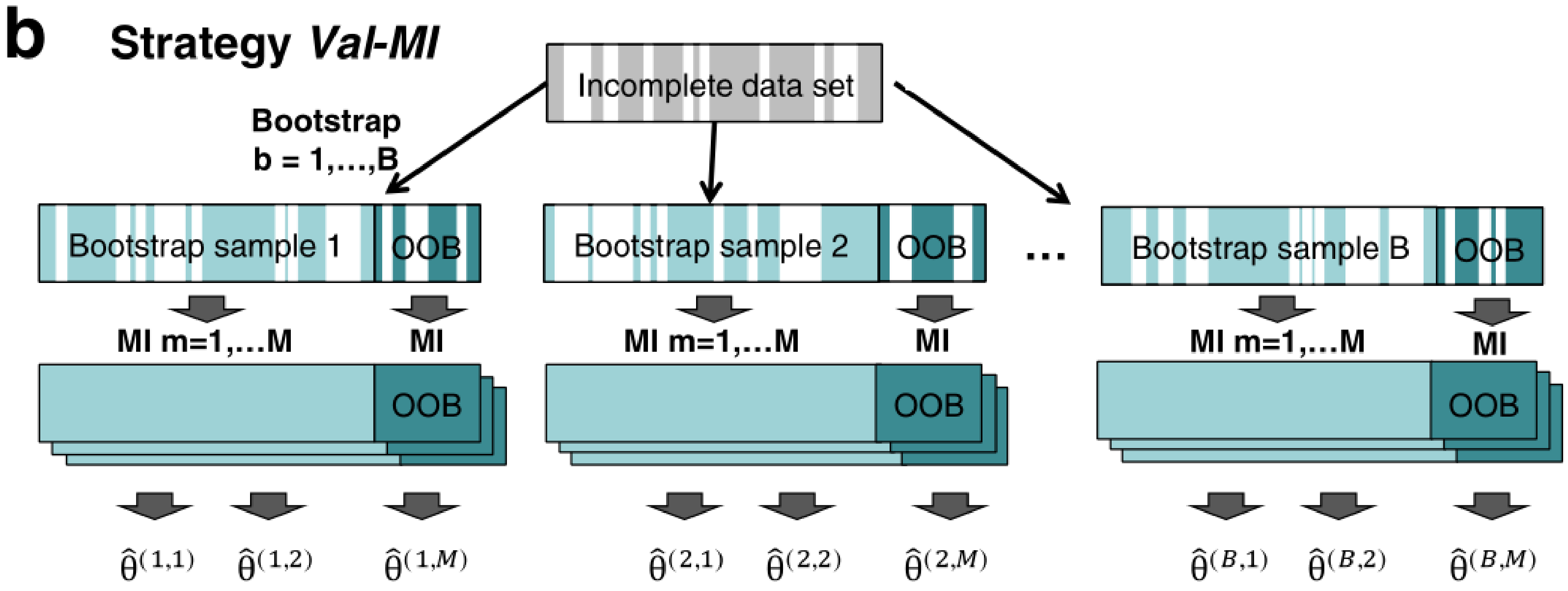
M fitted Logistic Ridge re…

Averaging coefficients (and standard errors)

Final model incorporating uncertainty due to missingness

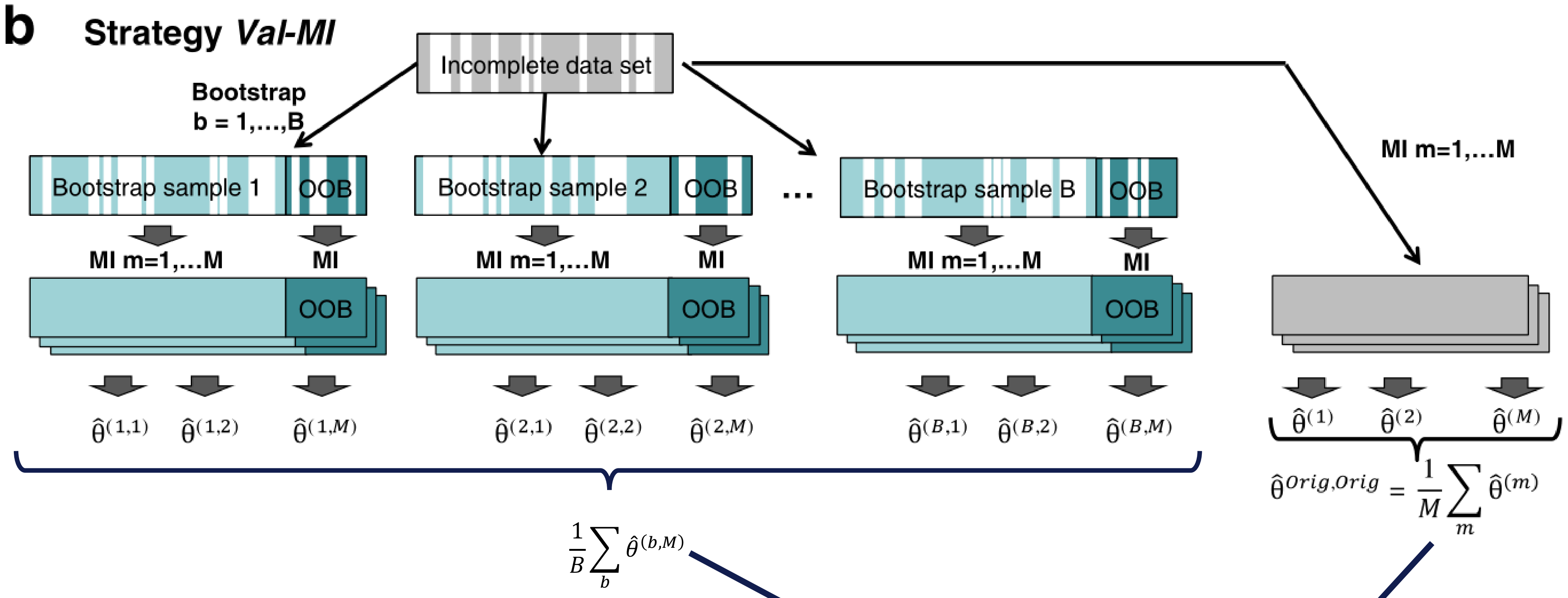
**And now: evaluate performance using bootstrap???**
**Reflecting „all sources of uncertainty“???**

# Statistical analysis plan: …it gets complicated




Wahl S et al. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Med Res Methodol. 2016;16(1):144.

# Statistical analysis plan: …it gets complicated

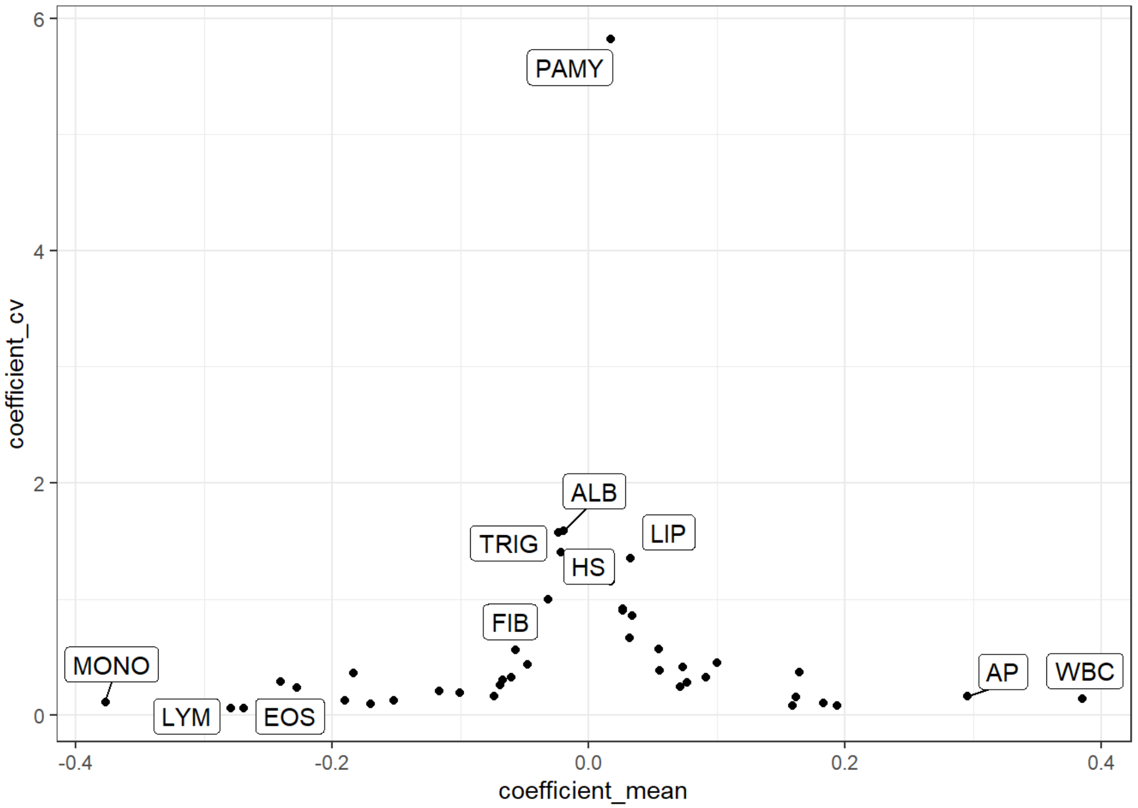



**0.632+ Bootstrap estimate by weighting bootstrap and apparent performance**

Wahl S et al. Assessment of predictive performance in incomplete data by combining internal validation and multiple imputation. BMC Med Res Methodol. 2016;16(1):144.

# Results from a Frequentist analysis

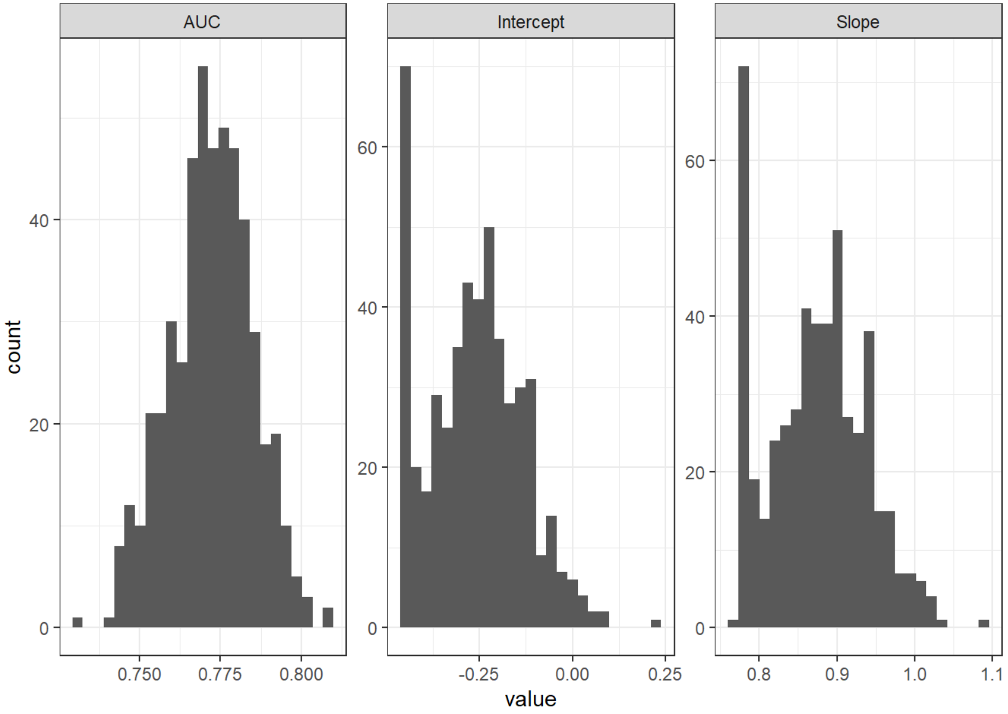
- Model checks: imputation model seems to converge fine (trace plots)
- Model coefficients show importance of known predictors



# Results from a Frequentist analysis

## Model evaluation using 500 bootstrap resamples

| Statistic | 0.632+ | 95% CI |
| --- | --- | --- |
| AUC | 0.757 | 0.746 – 0.796 |
| Calibration slope | 0.915 | 0.775 - 0.995 |
| Calibration intercept | -0.174 | -0.454 – 0.000 |



# Bayesian robustness check (Aliaksandr)

- **Approach:** Nonlinear Bayesian model with bootstrap validation using FBMS
  *https://cran.r-project.org/web/packages/FBMS*
- Handles imputations with corrections as will be discussed in the talk by **Florian Frommlet** in the **GS-8: Bayesian Modelling** section tonight.

# Bayesian robustness check: Dataset and methodology

- **Dataset**: 36 predictors (e.g., AGE, CRP, LYM, SODIUM) after removing 8 columns (e.g., WBC, MCV).
- **BloodCulture** converted to 0/1.
- **FBMS**: Automated inputting and correcting of imputations missing values during model fitting.
- **Model**: (Non)linear logistic regression (gmjmcmc.parallel or mjmcmc, Jeffreys prior) with transformations (sigmoid, sin, cos, exp dbl).
- **Evaluation**: C-index (AUC), calibration slope, calibration intercept, Brier scores, .632+ bootstrap estimates, 95% CIs (1000 bootstrap iterations).

# Bayesian robustness check: Results

- Bootstrap: 1000 iterations, estimates

| Model | Metric | Full train | Full test | .632+ | 95% CI |
|---|---|---|---|---|---|
| **BGNLM** | AUC | 0.742 | 0.727 | **0.758** | **0.735 – 0.801** |
| | Calibration slope | 1.019 | 0.927 | **1.035** | **0.881 – 1.213** |
| | Calibration intercept | 0.041 | -0.187 | **0.069** | **-0.295 – 0.440** |
| **BLR** | AUC | 0.744 | 0.746 | **0.715** | **0.661 – 0.744** |
| | Calibration Slope | 1.015 | 1.013 | **0.797** | **0.445 – 1.004** |
| | Calibration intercept | 0.031 | -0.035 | **-0.385** | **-1.262 – 0.044** |

- Discrimination: Both models show good C-index (>0.70); nonlinear achieves slightly higher optimism-corrected performance.
- Calibration slope: Nonlinear model remains close to 1, while linear shows drift in bootstrap estimates.
- Calibration intercept: Near zero for both, but linear model tends to underestimate risk under resampling

**Overall BGNLM seems to be better than BLR!**

# Multiple imputation as a bridge between worlds

**Multiple imputation as approximate Bayesian inference**

- Want: Bayesian posterior for regression coefficient

$$P(\beta|X_{obs}, R) = \int P(\beta|X_{obs}, X_{miss})P(X_{miss}|X_{obs}, R)dX_{miss}$$

- Rubins key insight (1970s): approximate by $\frac{1}{M}\sum P\left(\beta\middle|X_{obs}, X_{miss}^{(m)}\right)$, where $X_{miss}^{(m)}$ is sampled from $P(X_{miss}|X_{obs}, R)$ (i.e. by using a model)
- Often first two moments are sufficient to describe posterior, whence **„Rubin‘s rules“** where formulated to approximate posterior mean and variance for standard Frequentist inference
- But note: these simple approximations do not work well when posteriors are badly behaved (e.g. due to problems with imputation model)

# Conclusions for this case study

**Good agreement between Frequentist and Bayesian analysis**

- Combining multiple imputation with model evaluation is tricky and affects results
  - Proper evaluation incorporating all uncertainties is hard for Frequentists...
- Bayesian workflow well aligned with uncertainty transmission due to imputation
  - But model convergence and computational demand require sophisticated diagnostics and sampling algorithms

**Robustness checks bridging both worlds helps to be confident in the results!**