University of Glasgow | School of Computing Science

THE AWARDS 2020 | UNIVERSITY OF THE YEAR

Imputation Strategies

Dr. Fani Deligianni,

fani.deligianni@glasgow.ac.uk

Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

https://www.gla.ac.uk/schools/computing/staff/fanideligianni

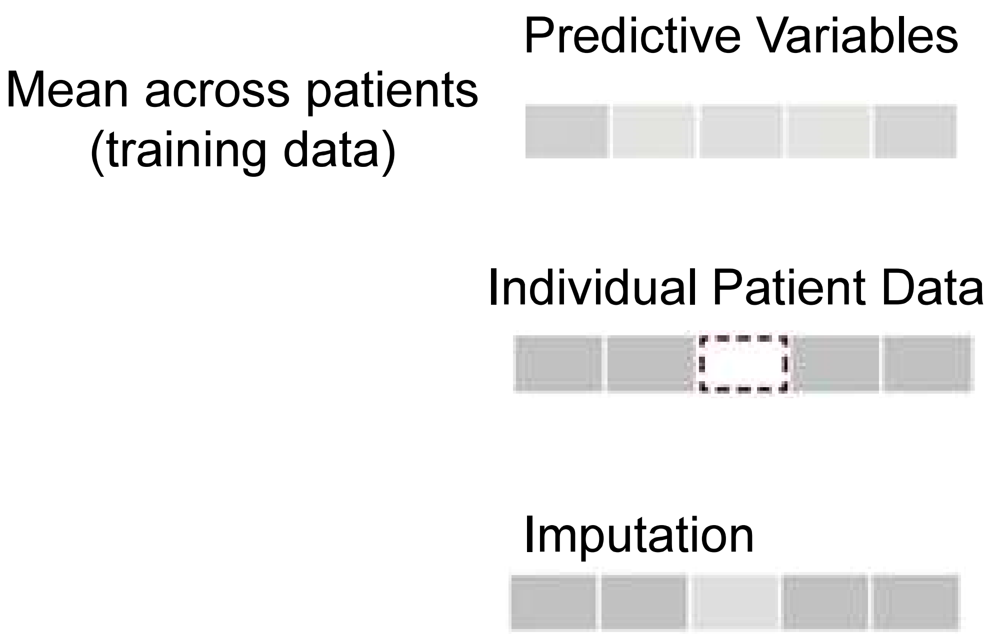
WORLD CHANGING GLASGOW

Missing Values

- Most prediction models cannot be used when predictive variables have missing values
- Population characteristics such as mean and covariance can be used to generate imputations specific to an individual

Mean Imputation

Missing values are replaced by the sample mean



Step 1: Estimate means of all predictors using only the training data

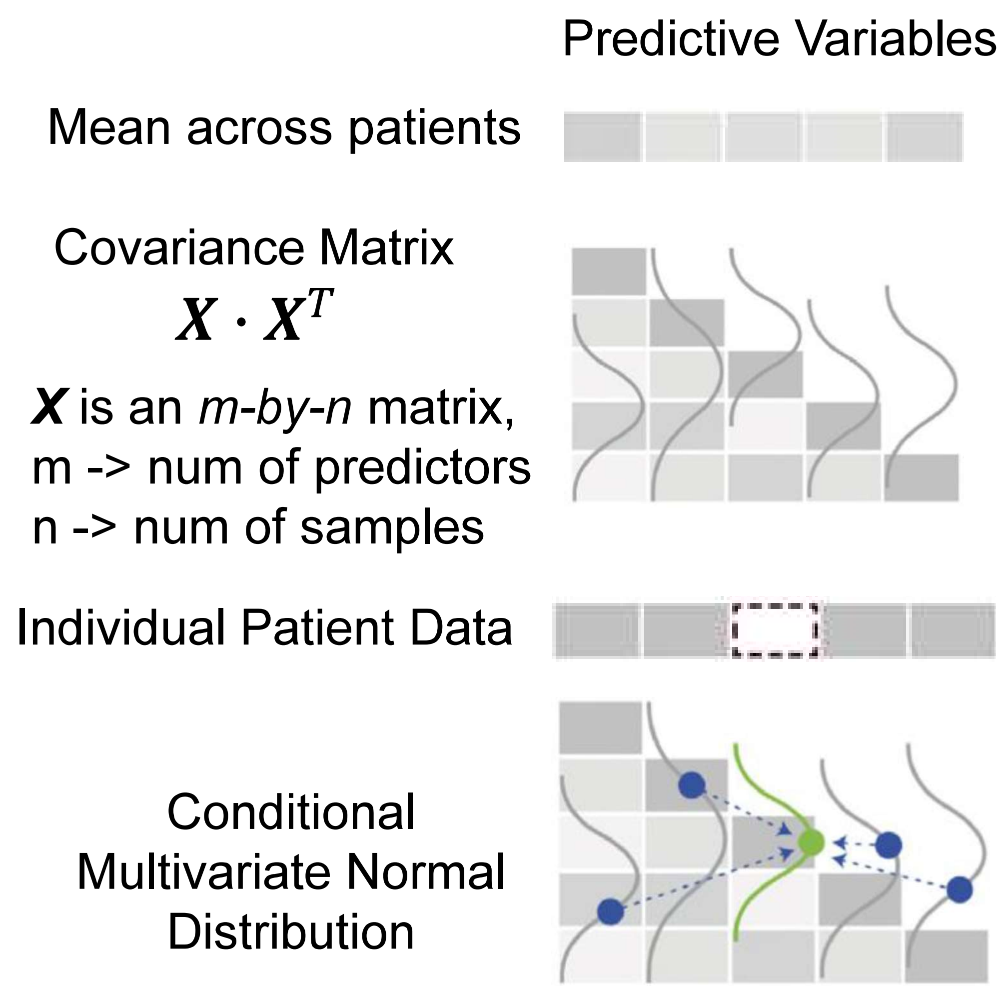
Step 2: Identify missing values

Step 3: Use means across patient data to fill in missing values

Mean Imputation - Limitations

- Mean imputation might be inadequate when the predictive variable with missing values is a strong predictor, or it has high variability
- Mean imputation does not distinguish between patients
- Mean imputation makes uncertainties about the imputed values unclear

Joint Modeling Imputation



Step 1: Estimate means of all predictors using only the training data

Step 2: Estimate covariance matrix of training data

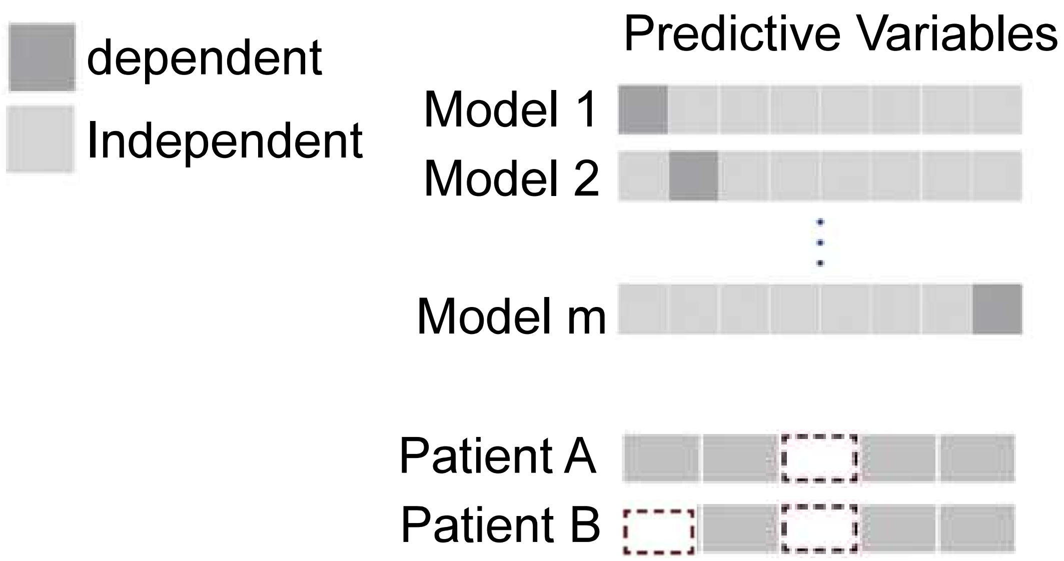
Step 3: Identify missing values

Step 4: Exploit derived distribution to generate imputation for missing values

Joint Modeling Imputation - Limitations

- Better than mean imputation as it considers the interaction between predictors
- It only requires population statistics in order to be computed
- It assumes that the predictor variables are normally distributed

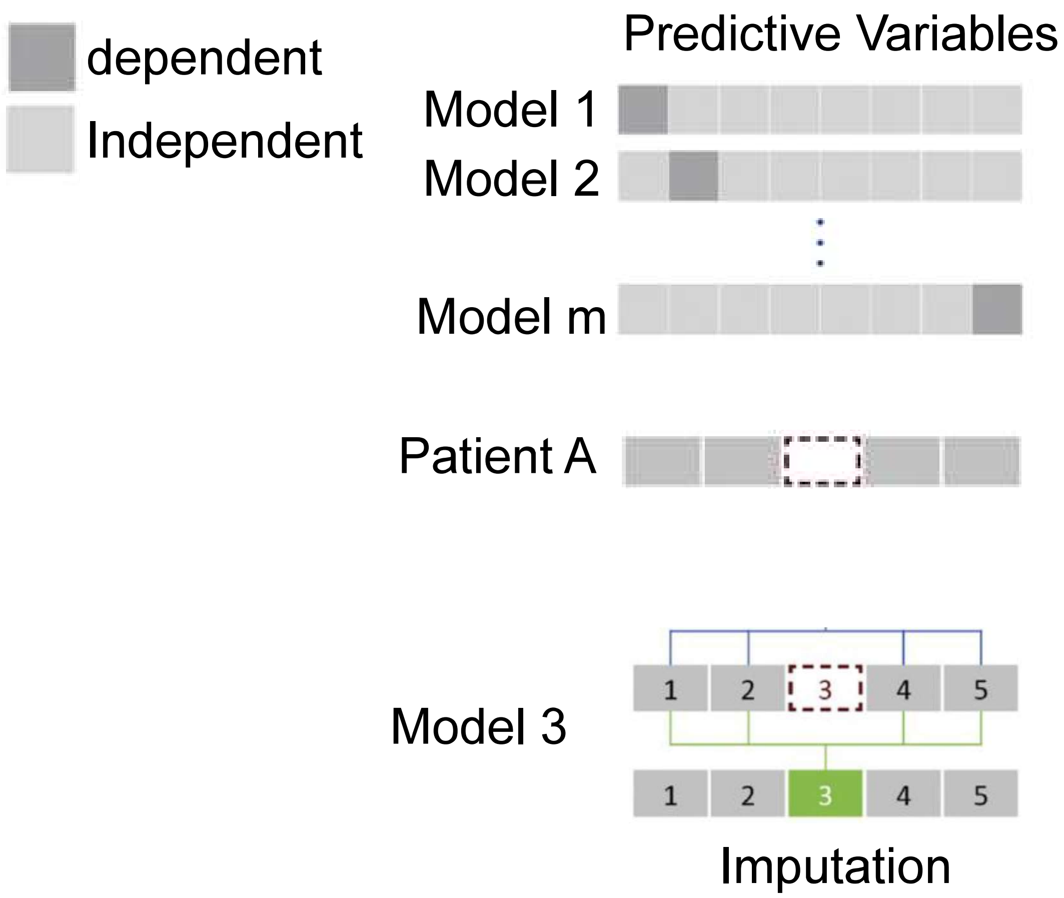
Conditional Modeling Imputation



Step 1: Derive a regression/prediction model for each predictor

Step 2: Identify if a patient has one or more values missing

Conditional Modeling Imputation

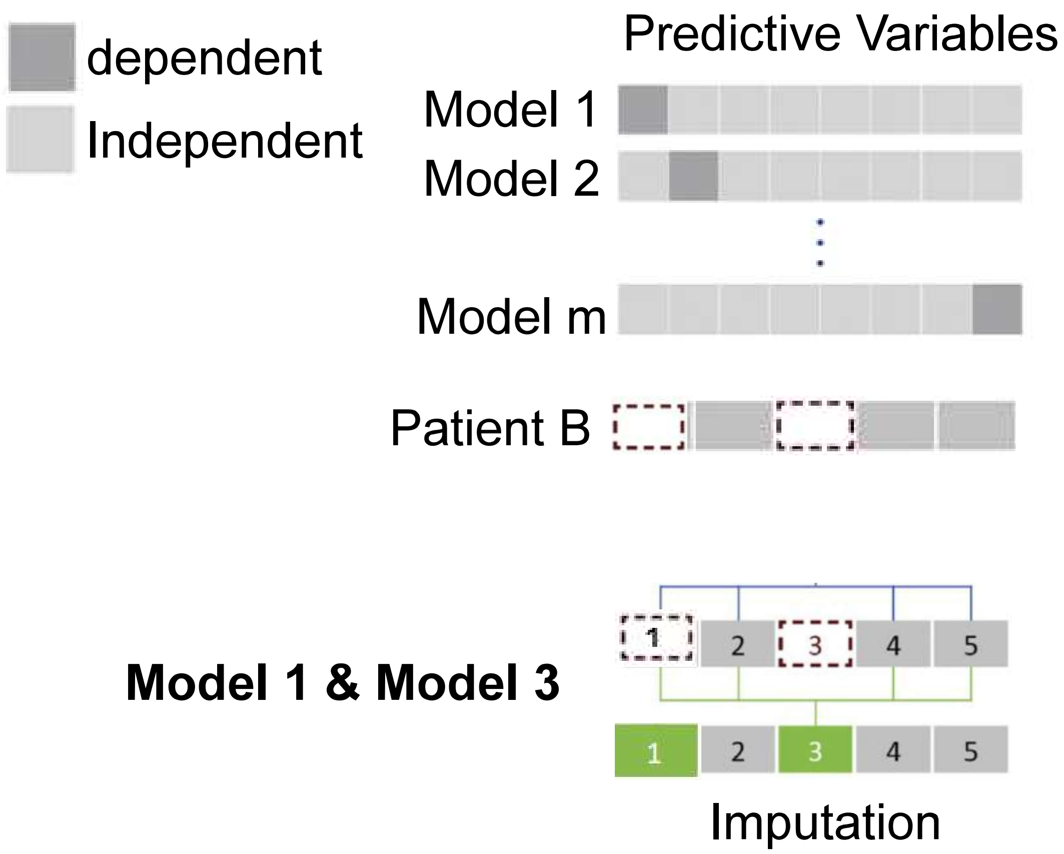


Step 1: Derive a regression/prediction model for each predictor

Step 2: Identify if a patient has one or more values missing

Step 3: When a single predictor has a missing value use directly the corresponding model to fill the gap

Conditional Modeling Imputation



Step 1: Derive a regression/prediction model for each predictor

Step 2: Identify if a patient has one or more values missing

Step 3: When multiple predictors are missing, the fitted regression models are combined via Markov Chain Monte Carlo Sampling

Evaluation of Imputation

- Leave-One-Out Cross Validation
- Root Mean Squared Error (RMSE) between the average of the multiple imputed predicted values and the true, original value (missing values selected at random)
- RMSE accumulates errors due to bias and variability
- Assessed confidence intervals around the imputed predictor variables
- Prediction performance with the actual values compared to imputed values

Summary

- Mean imputation underestimate the risk in high-risk patients
- The difference between mean imputation and both JMI and CMI is larger in high-risk patients
- Mean imputation is considered insufficient when strong predictors are missing

References

- Nijman et al. ‘Real-time imputation of missing predictor values improved the application of prediction models in daily practice’, Journal of Clinical Epidemiology, 2021.
- Carreras et al. ‘Missing not at random in end-of-life care studies: multiple imputation and sensitivity analysis on data from the ACTION study’, BMC Medical Research Methodology, 2021