Addressing Structured Missingness Challenges in Data Integration - An Application in Cystic Fibrosis Analysis

Turing-Roche Knowledge Share Event

Robin Mitra, **Eleni-Rosalina Andrinopoulou**

27 March, 2024

www.erandrinopoulou.com eandrinopoulou@erasmusmc.nl @ERandrinopoulou

Introduction: Motivation

A lot of information is available

➔ Electronic medical records

Introduction: Motivation

A lot of information is available

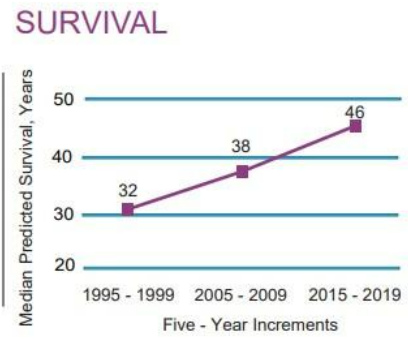
➔ Electronic medical records

Cystic Fibrosis: US Cystic Fibrosis Foundation Patient Registry

CYSTIC FIBROSIS FOUNDATION
ADDING TOMORROWS

46
YEARS
2015 - 2019

Among people with CF born between 2015 and 2019, half are predicted to live to 46 years old or more. This does not reflect individual variability in survival seen among people with CF.



Introduction: Motivation

Data

- ➔ 35153 patients, 1523406 observations
- ➔ Median baseline age: 9, 48% females
- ➔ Median follow-up duration: 10 years (IQR 5–17)

Different types of information

- ➔ Baseline characteristics: Sex, Age, genotype, SESlow, insuranse, infection (Pseudomonas aeruginosa)
- ➔ Biomarkers: FEV_1 % pred
- ➔ Nutritional status: BMI percentile

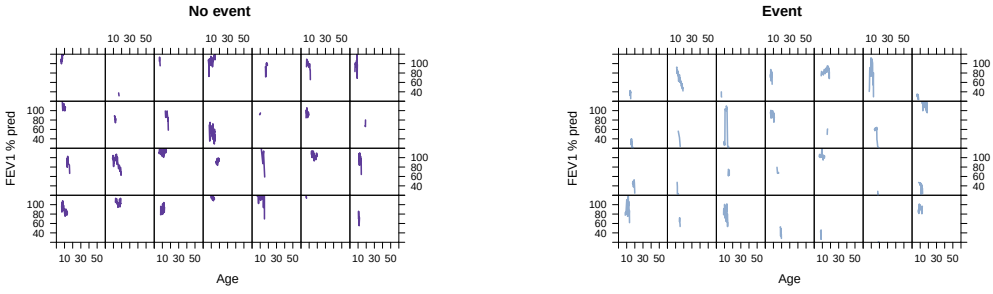
Research focus: Structured missingness

How do we handle patients that dropped-out due to death/lung-transplantation

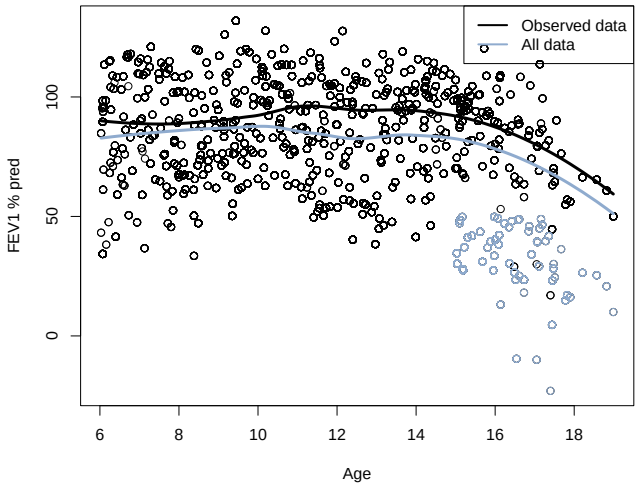
Research focus: Structured missingness

Different types of information

- ➔ Baseline characteristics: Sex, Age, genotype, SESlow, insuranse, infection (Pseudomonas aeruginosa)
- ➔ Biomarkers: FEV_1 % pred
- ➔ Nutritional status: BMI percentile
- ➔ Survival data: Time-to-death/lung transplantation



Research focus: Structured missingness



Research focus: Structured missingness

MNAR: joint distribution of the outcome and the missingness outcome

➔ FEV_1 % pred (BMI percentile) + time-to-death/lung transplantation

- ◇ selection models
- ◇ pattern mixture models
- ◇ shared parameter models

Joint models of longitudinal and survival data

Let y_i be the complete response

Joint distribution

$$p(y_i^o, y_i^m, r_i) = \int p(y_i^o, y_i^m \mid b_i) p(r_i \mid b_i) p(b_i) db_i$$

$$= \int p(y_i^o, \mid b_i) p(r_i \mid b_i) p(b_i) db_i$$

where

- ➔ r_i denotes the missing data indicator (1 is observed, 0 otherwise)
- ➔ y_o denotes the observed data
- ➔ y_m denotes the missing data

Research focus: Personalised dynamic predictions

Where is the best place to live with CF?

Research focus: Personalised dynamic predictions

Erasmus MC
University Medical Center Rotterdam

Personalized dynamic predictions

➔ We assume the following setting for a new patient l
- ◇ all baseline information
- ◇ available longitudinal outcomes (K) up to time t, $\tilde{Y}_{lk}(t) = y_{lk}(s), 0 \leq s < t$

➔ We are interested in **future longitudinal outcomes / events** in the medically relevant interval $(t, t + \Delta t]$

Based on the models we can get
- ◇ $E\{y_{lk}(t + \Delta t) \mid \tilde{Y}_{lk}(t), \boldsymbol{D}_n\}$
- ◇ $Pr\{T_l^* \geq t + \Delta t \mid T_l^* > t, \tilde{Y}_{lk}(t), \boldsymbol{D}_n\}$

Research focus: Personalised dynamic predictions

Measuring Predictive Performance

- **Longitudinal and survival outcomes**: the distance between the predicted outcome and the actual outcome (PE)
- **Survival outcomes**: how well can the longitudinal biomarker(s) discriminate between subject of low and high risk for the event (AUC)

Andrinopoulou, E. R., Eilers, P. H., Takkenberg, J. J., & Rizopoulos, D. (2018). Improved dynamic predictions from joint models of longitudinal and survival data with time-varying effects using P-splines. Biometrics, 74(2), 685-693.

Andrinopoulou, E. R., Harhay, M. O., Ratcliffe, S. J., & Rizopoulos, D. (2021). Reflection on modern methods: Dynamic prediction using joint models of longitudinal and time-to-event data. International Journal of Epidemiology, 50(5), 1731-1743.

Research focus: Personalised dynamic predictions

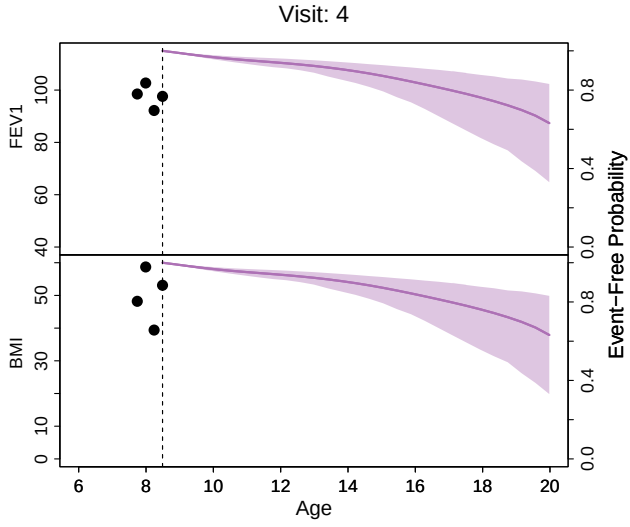
Joint models

- Outcomes:
  - **FEV_1**
  - **BMI**
  - **Weight**
  - **Height**
  - **Exacerbation**

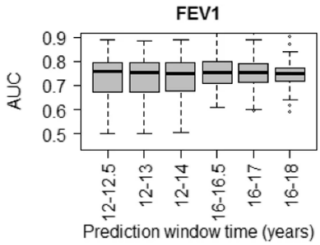
Aim: Including growth markers will yield gains in prediction performance

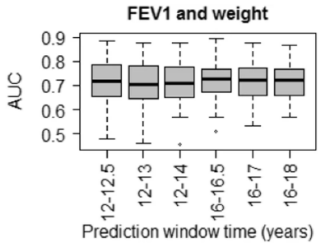
Andrinopoulou, E. R., Clancy, J. P., & Szczesniak, R. D. (2020). Multivariate joint modeling to identify markers of growth and lung function decline that predict cystic fibrosis pulmonary exacerbation onset. BMC pulmonary medicine, 20, 1-11.

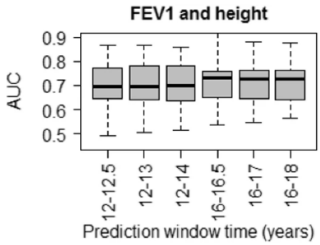
Research focus: Personalised dynamic predictions

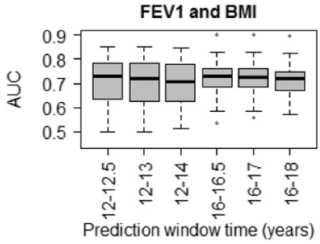


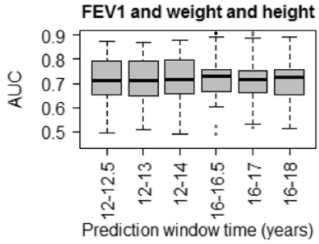
Research focus: Personalised dynamic predictions











www.erandrinopoulou.com eandrinopoulou@erasmusmc.nl @ERandrinopoulou

Research focus: Integrating data sources

Can we integrate different types of information?

Research focus: Integrating data sources

Different types of information

- ➔ Baseline characteristics: Sex, Age, F508del, SESlow, Enzymes
- ➔ Biomarkers: FEV_1 % pred
- ➔ Nutritional status: BMI percentile
- ➔ Survival data: Time-to-death/lung transplantation
- ➔ Geomarkers (environmental/community factors): Deprivation index
- ➔ Image (CT scans): structural airway abnormalities

More is less?

Research focus: Integrating data sources

Deprivation index

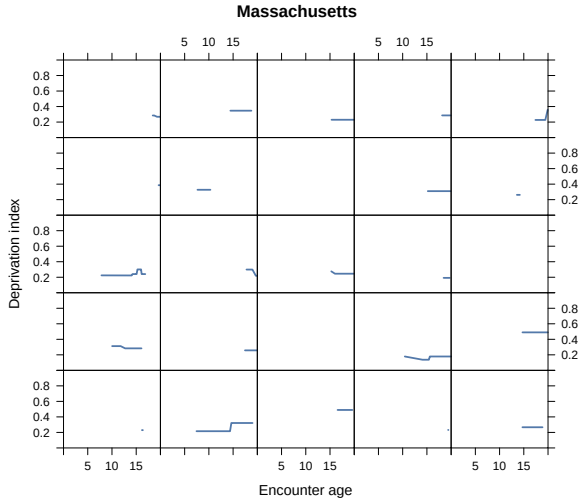
➔ Socioeconomic variables from the American Community Survey (ACS): capture "community deprivation"

- ◇ Principal components analysis of six different 2015 ACS measures
- ◇ "Deprivation Index": the first component explains over 60% of the total variance
- ◇ Rescaling and normalizing forces the index to range from 0 to 1, with a higher index being more deprived

Cole Brokamp, Andrew F. Beck, Neera K. Goyal, Patrick Ryan, James M. Greenberg, Eric S. Hall. Material Community Deprivation and Hospital Utilization During the First Year of Life: An Urban Population-Based Cohort Study. Annals of Epidemiology. 30. 37-43. 2019

Research focus: Integrating data sources

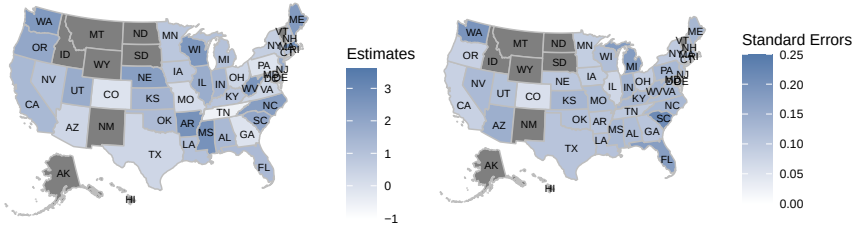
Deprivation index



Research focus: Integrating data sources

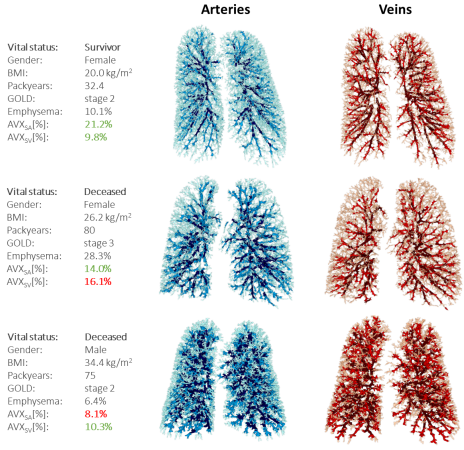
Relationship between community factors and lung function

For 0.1 unit increase in the normalized area of deprivation index



Research focus: Integrating data sources

Image data



Conclusions

- ➔ We have seen how Structured Missingness is a key challenge to address in Data Integration
- ➔ A simple simulation illustrates some approaches that could be considered
- ➔ Imputation may address this in some settings but is not a panacea
- ➔ In addition, in some settings imputation may not be plausible, so we need to explore other ways we can integrate information from different data sets
- ➔ In general, a combination of various different approaches is likely to be optimal, and dependent on the data or problem in question
- ➔ The example involving Cystic Fibrosis illustrates the scale of the challenge in practice, but also how addressing SM offers the potential to greatly improve how we use complex integrated data sources

Erasmus MC
University Medical Center Rotterdam

Thank you for your attention!

www.erandrinopoulou.com eandrinopoulou@erasmusmc.nl @ERandrinopoulou