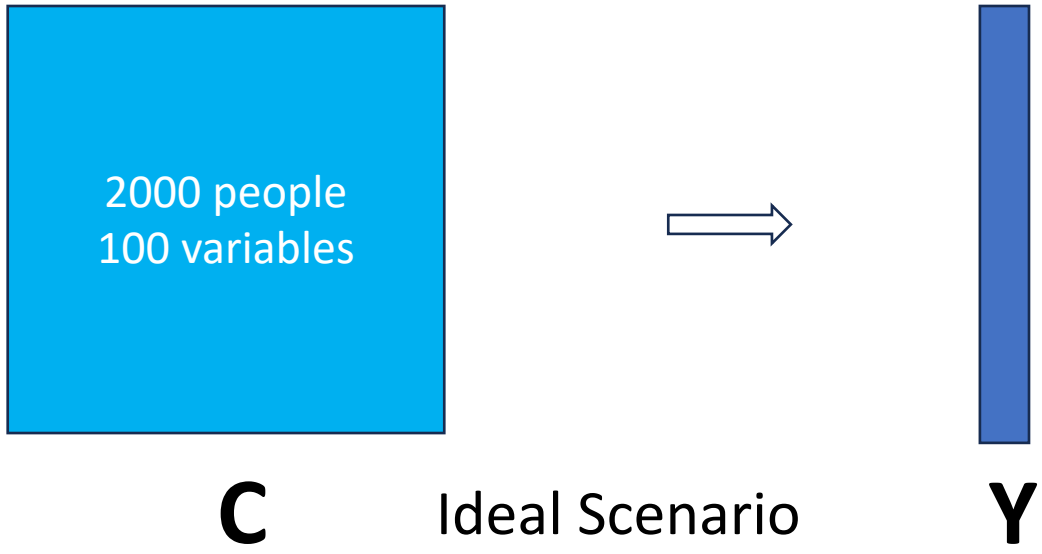
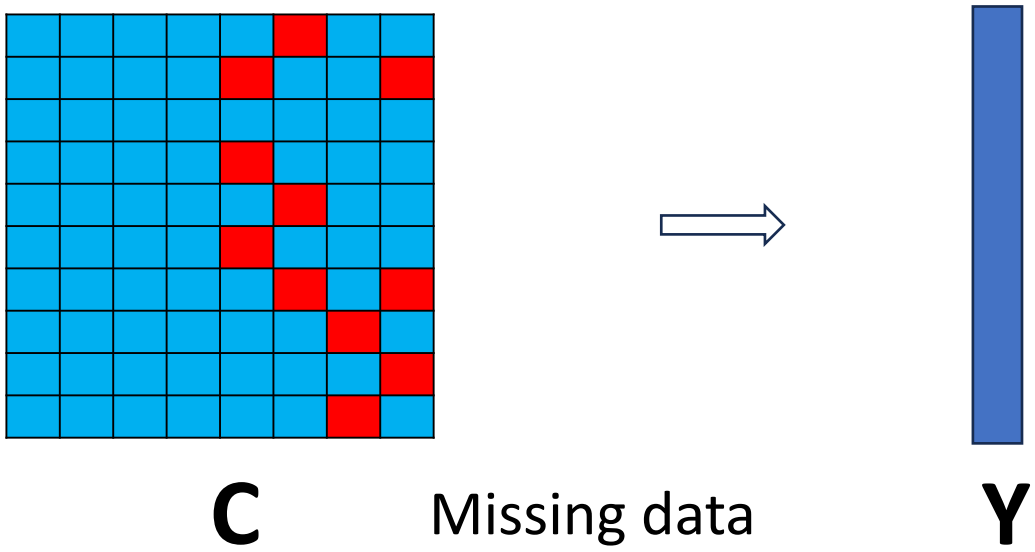
Addressing Structured Missingness Challenges in Data Integration

Robin Mitra, Eleni-Rosalina Andrinopoulou,

Motivation to deal with Missing data



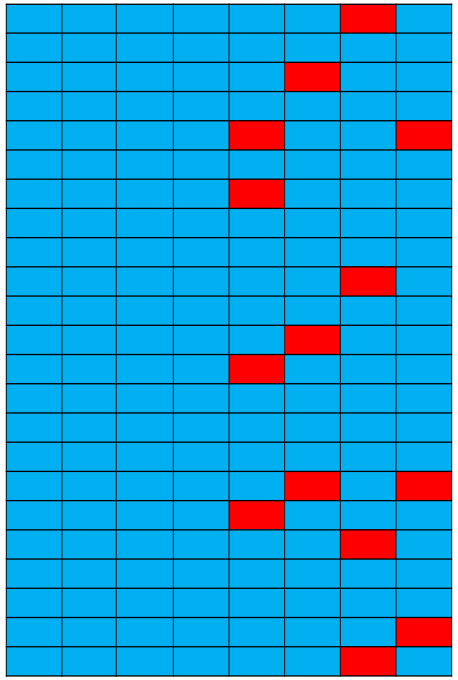
C Ideal Scenario **Y**



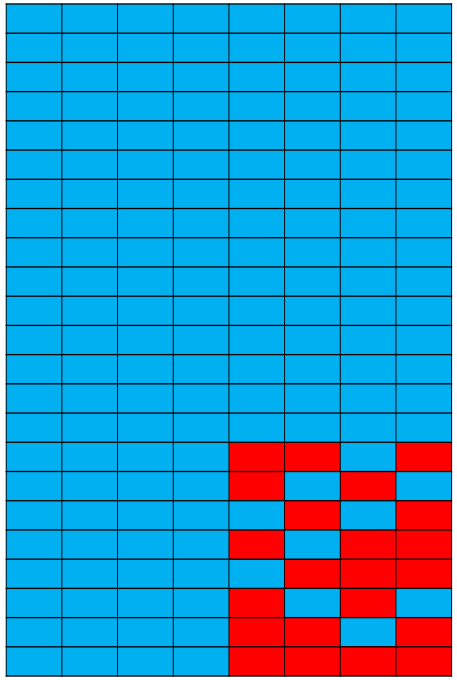
C Missing data **Y**

- Good quality data is essential in tackling important research questions
- E.g. suppose we wish to use clinical observations (C) to build a predictive model for an outcome of interest (Y)
- These data sets often comprise a lot of variables and require sophisticated model building methods
- Missing values are also a typically encountered problem and impede data analysis
- Fundamental work on missing data tends to focus on univariate instances but there is an increasing need to tackle multivariate challenges

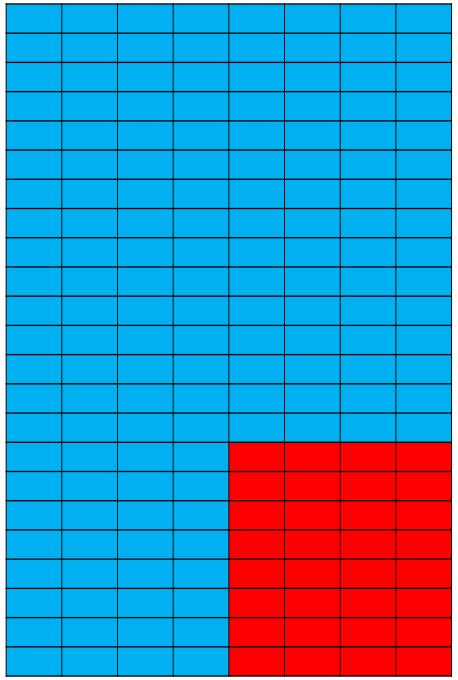
Multivariate missing data - Structured Missingness



Unstructured



Almost block missingness



Block missingness

Weak SM

Strong SM

- The previous illustration shows how missing data can complicate model building
- However, this is not the only way missing data can arise in multivariate settings
- There may be relationships or “structure” present within the missing data
- This gives rise to the idea of “Structured Missingness” (SM)
- SM poses some specific challenges that require careful consideration
- We can consider a simple illustration

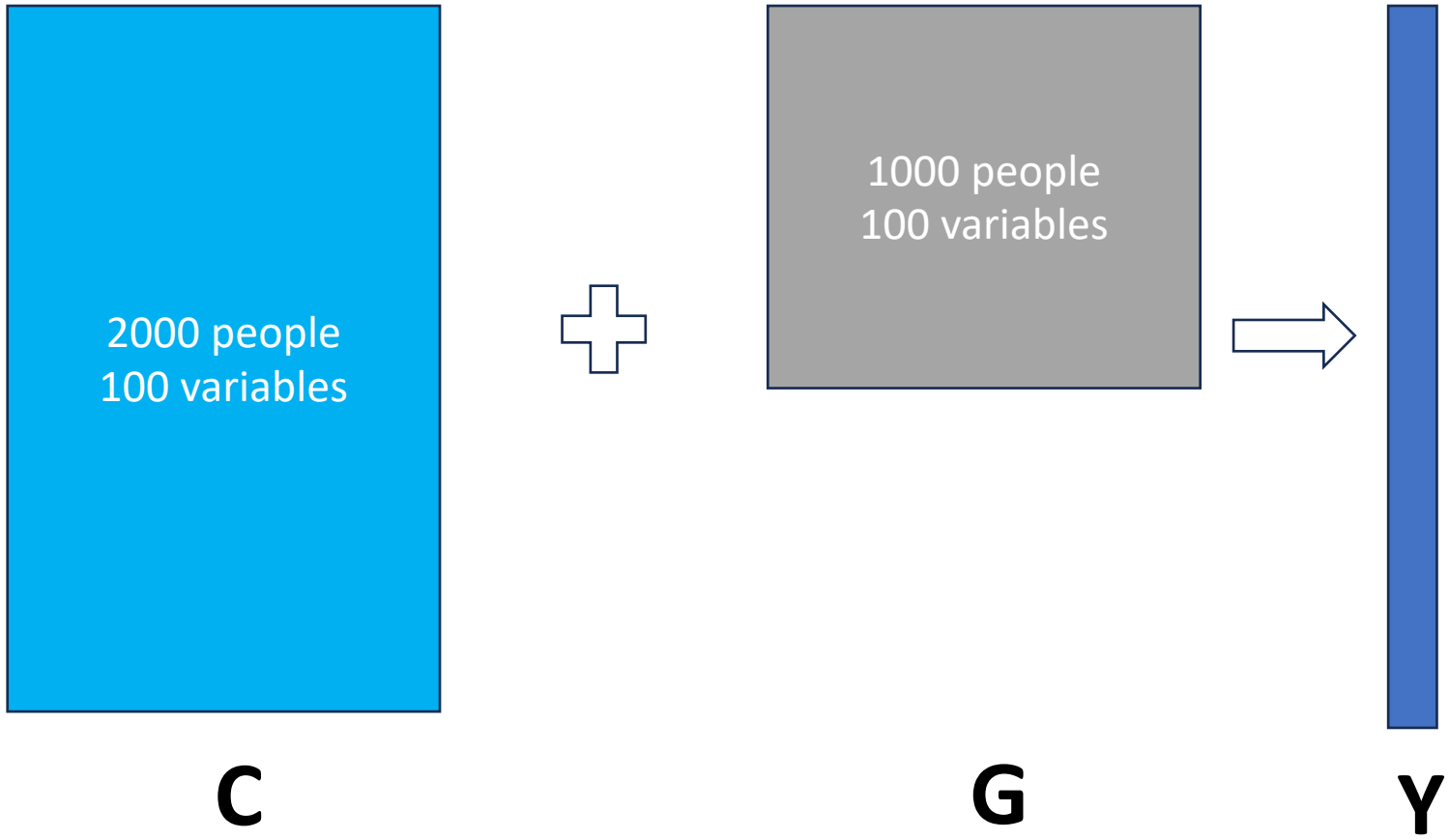
Simple illustration of SM challenges

PSA	Race	Gender	Sexual dysfunction	Age
2.1	Asian	M	N	28
?	White	F	Y	58
?	?	M	?	81
?	Black	F	N	?
4.3	?	M	?	?
?	White	F	N	48
?	Asian	?	?	76
?	White	F	N	?
2.3	White	M	N	23
5.6	Black	M	?	68
?	White	F	Y	62

Does it make sense to impute all missing values?

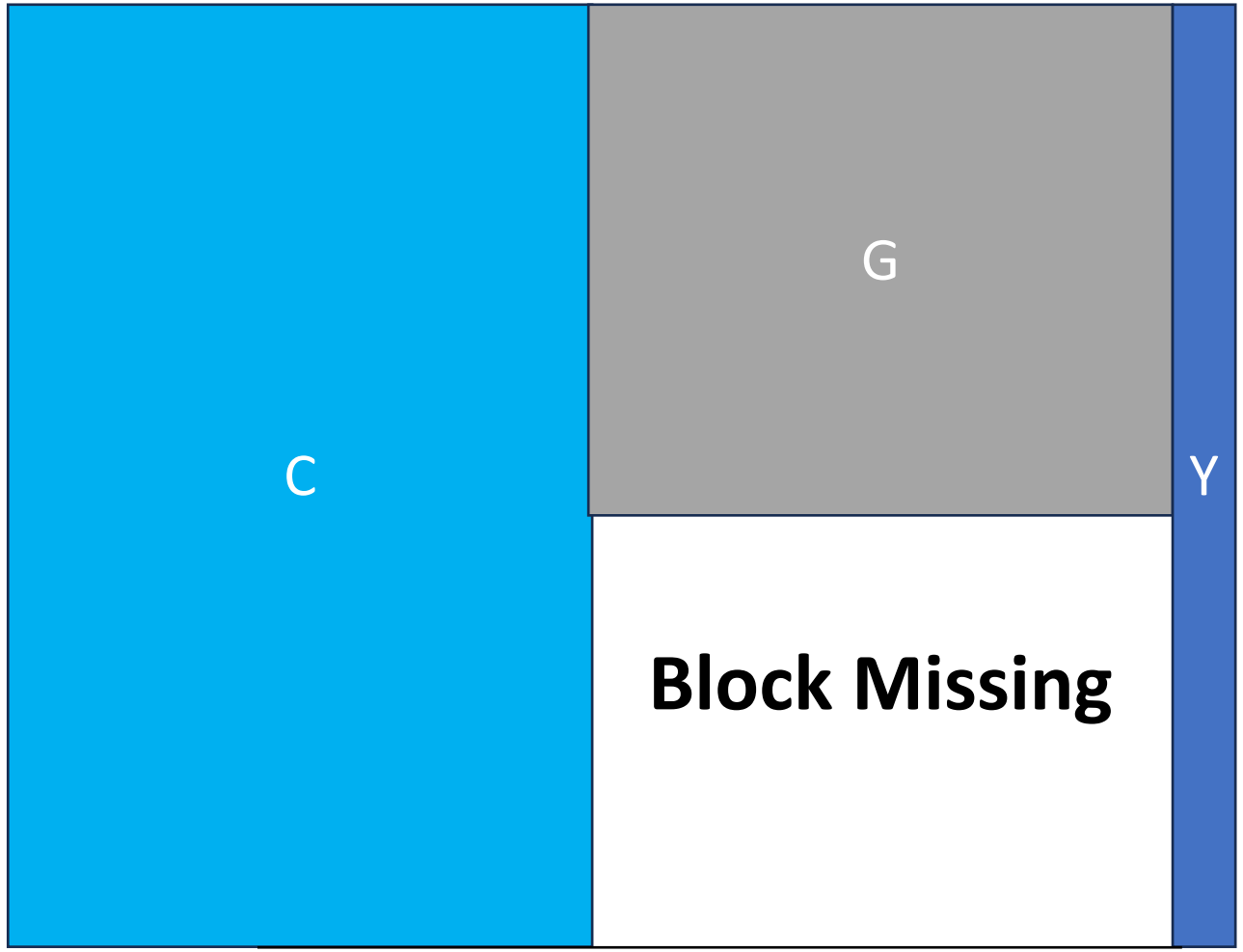
- In this simple example we can see some challenges posed by SM.
- Certain values are “supposed” to be missing.
- E.g. female individuals should not have a PSA measurement
- Thus care should be taken over which values to impute
- However, what should we do when someone is missing both gender and PSA?
- How can we best utilise the “imputable” and “non-imputable” missing values in our model development
- There are also practical settings where SM arises in data integration

Structured Missingness and Data Integration - illustration



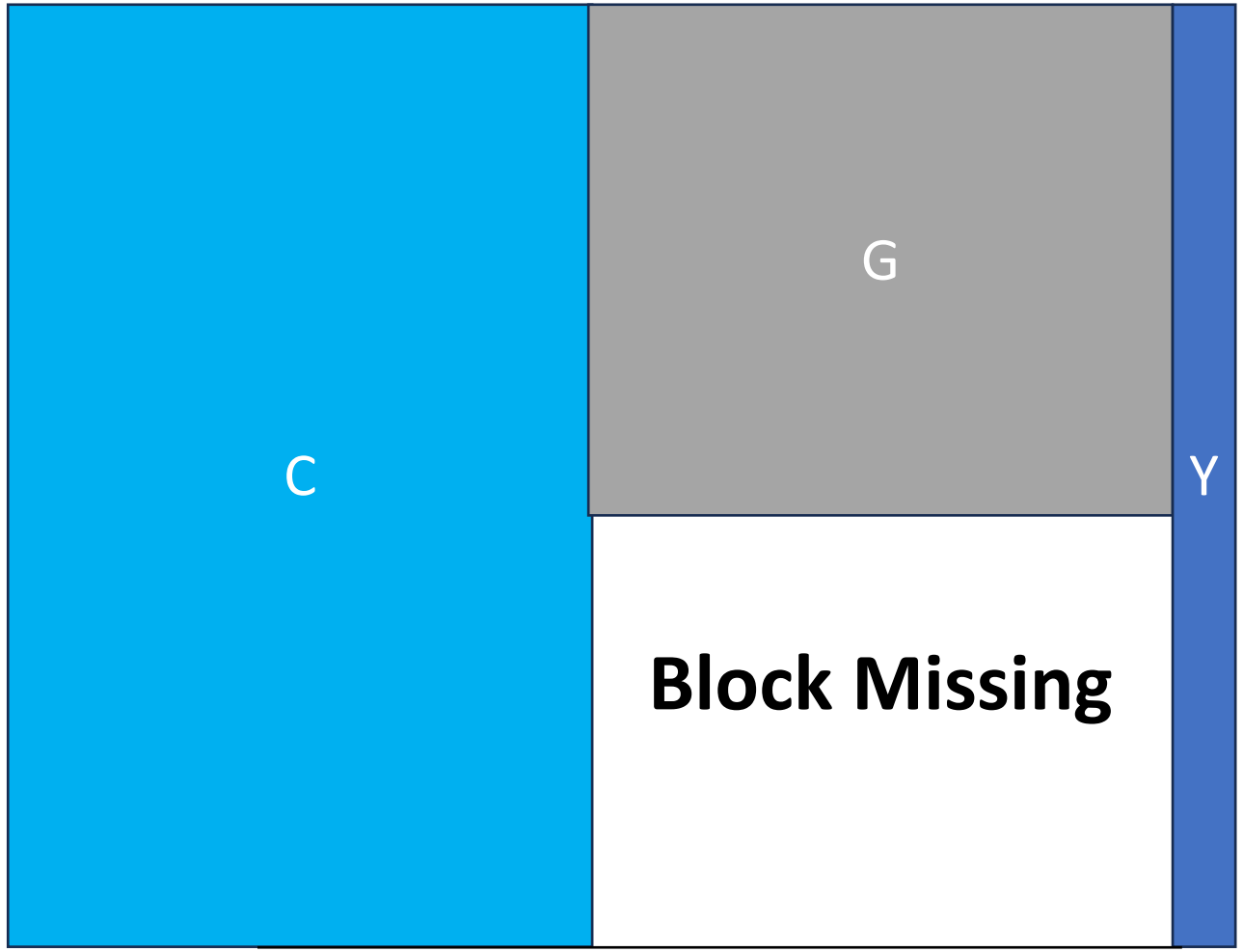
- Recall we want to use C to build a predictive model for Y
- But now suppose we also have some genomic measurements on a subset of the individuals in C
- We want to integrate the information present in G to build better predictions for Y
- We can invert this problem to view it as one of SM

Structured Missingness and Data Integration



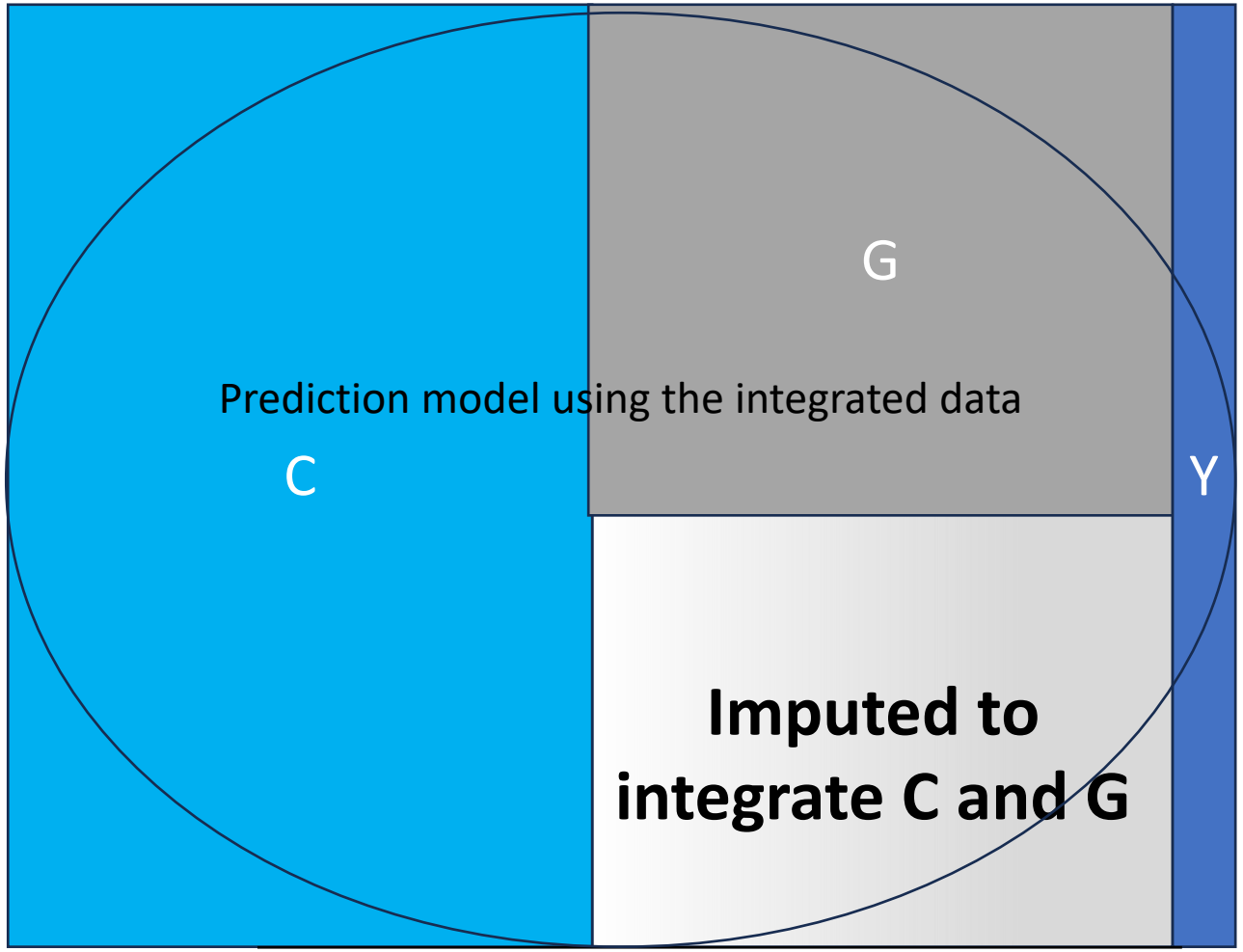
- The unobserved portion of G can be seen to be missing in a block
- The missingness here is highly structured
- Patients are either measured on the G variables or not measured at all
- This missingness could also be MCAR/MAR/MNAR
- MCAR: Units missing G are a random sample from units in C
- MAR: Units are more or less likely to be missing G, depending on variables in C
- Missing data could also be present in the C and G parts as well

Simulation – data generation



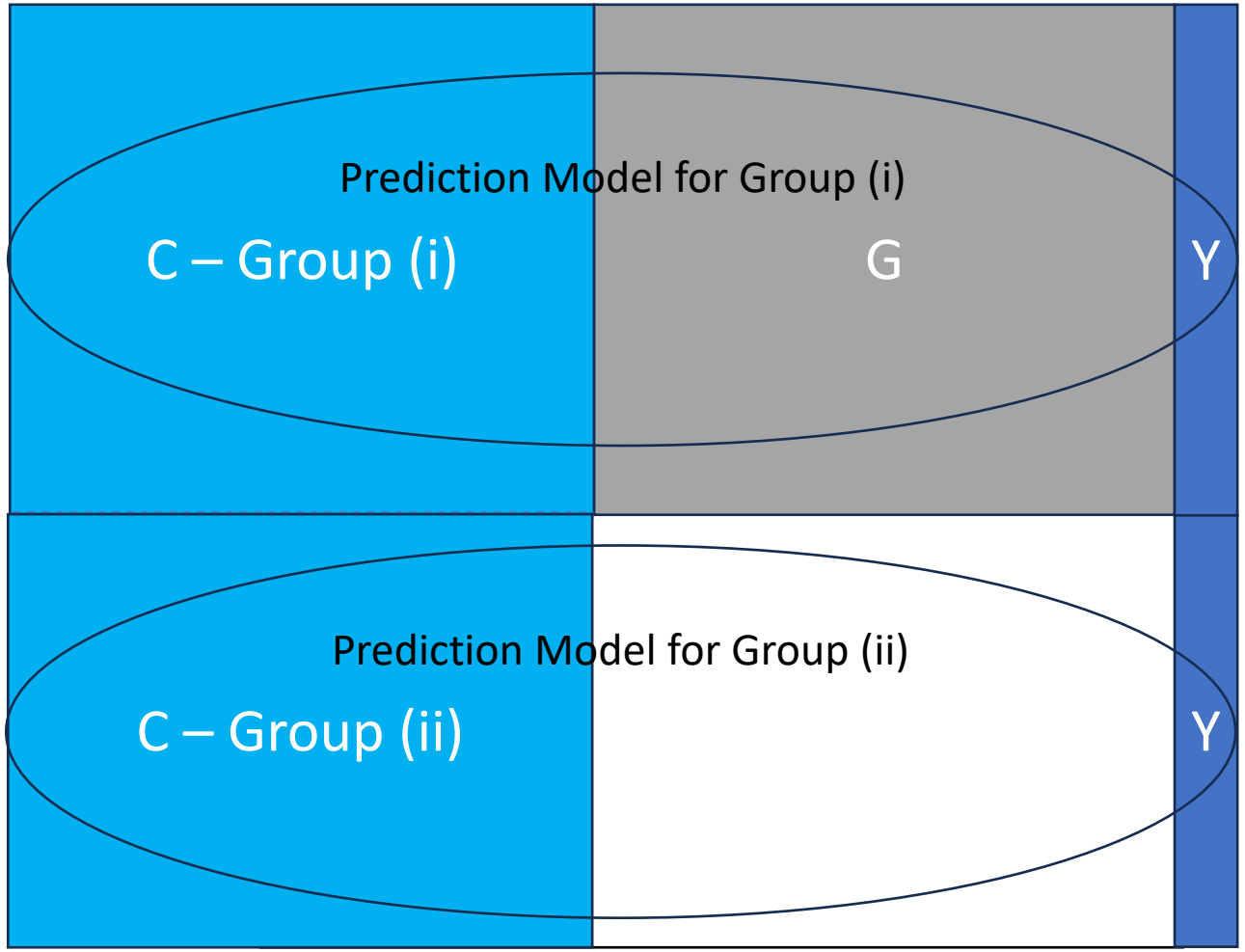
- (C,G) follow t-distributions with 3 degrees of freedom and variance 1
- Correlations between any 2 C variables or any 2 G variables are 0.4
- Correlations between any C and G variable are 0.2
- Y is normally distributed conditional on the first 10 C variables and first 70 G variables
- Also assume 50% MCAR data in C - structured via a copula
- No missing data in G

Simulation – integrated predictions via imputation



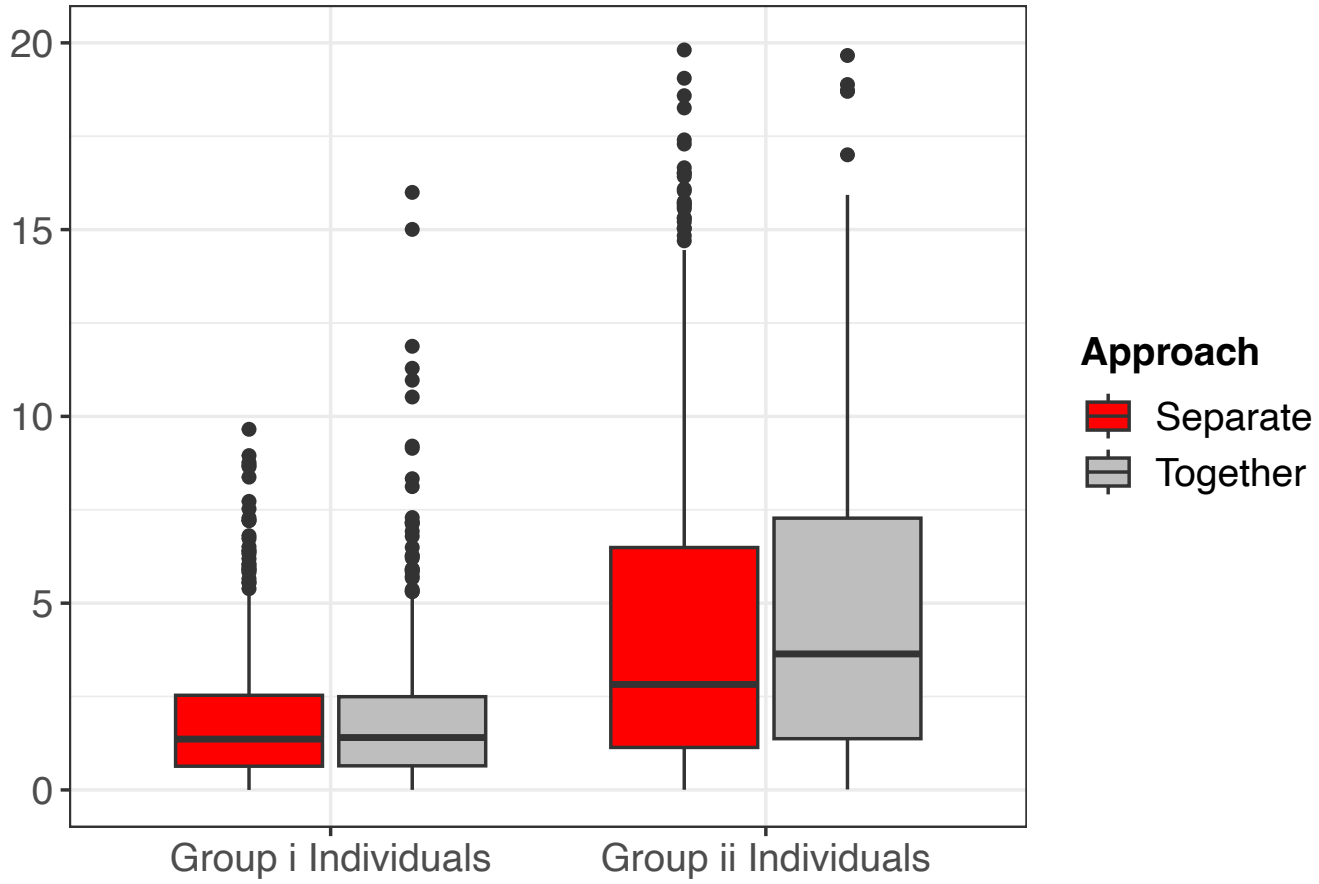
- We impute the missing data using the entire data source, i.e. C,G and Y
- This is done using the MICE package in R
- From the imputed data source we build a predictive model for Y
- Predictions are thus made using both C and G jointly, with imputation helping to integrate C and G into the prediction model
- To build our predictive model for Y we consider a LASSO model
- We measure predictive performance through the mean absolute difference between predicted and actual Y in a test set.

Simulation – non-integrated predictions



- We also compare the integrated imputation approach with an alternative non-integrated approach:
  - Split the units into two groups
  - Group (i) comprises units measured on both C and G
  - Group (ii) comprises units measured on C only
  - Impute and predict each group separately using LASSO
- Predictive performance measured in the same way as before

Simulation – data generation



- We see that for Group (i) units similar performance between both prediction approaches
- However for Group (ii) units there is a slight advantage in predicting using a model for C only. I.e. the imputed G data do not seem to help here
- There are likely to be other scenarios where integrating data sources through prediction offer advantages over fitting separate models.
- Our intention is to develop a range of methods (including non-imputation approaches) to best integrate multiple data sources when developing prediction models

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