Create a **matched cohort** to provide **context** for your large-scale characterisation

Context: The value of matched sampling for large-scale characterisation during phenotype diagnostics

Background: Large-scale characterisation aims to summarise all clinical events recorded in the database for a specific cohort in some window around their index date. This key step enables the researchers to have a comprehensive overview of the cohort and determine if it represents the target population. However, it usually results difficult to establish if there are unexpected imbalances, as the population in each database can be very specific.

We present two distinct ways to generate a comparator cohort: Using Achilles tables and *CohortConstructor* R package.

Results

Example: Large-scale characterisation of a lupus cohort in CPRD GOLD database (N_c = 7,381).

Figure 1. Proportion of individuals with each health condition within each cohort. The top 10 conditions with the highest prevalence in the target cohort are shown, as well as the top 10 conditions with the highest ASMD between the target and the Achilles/Matched cohorts.

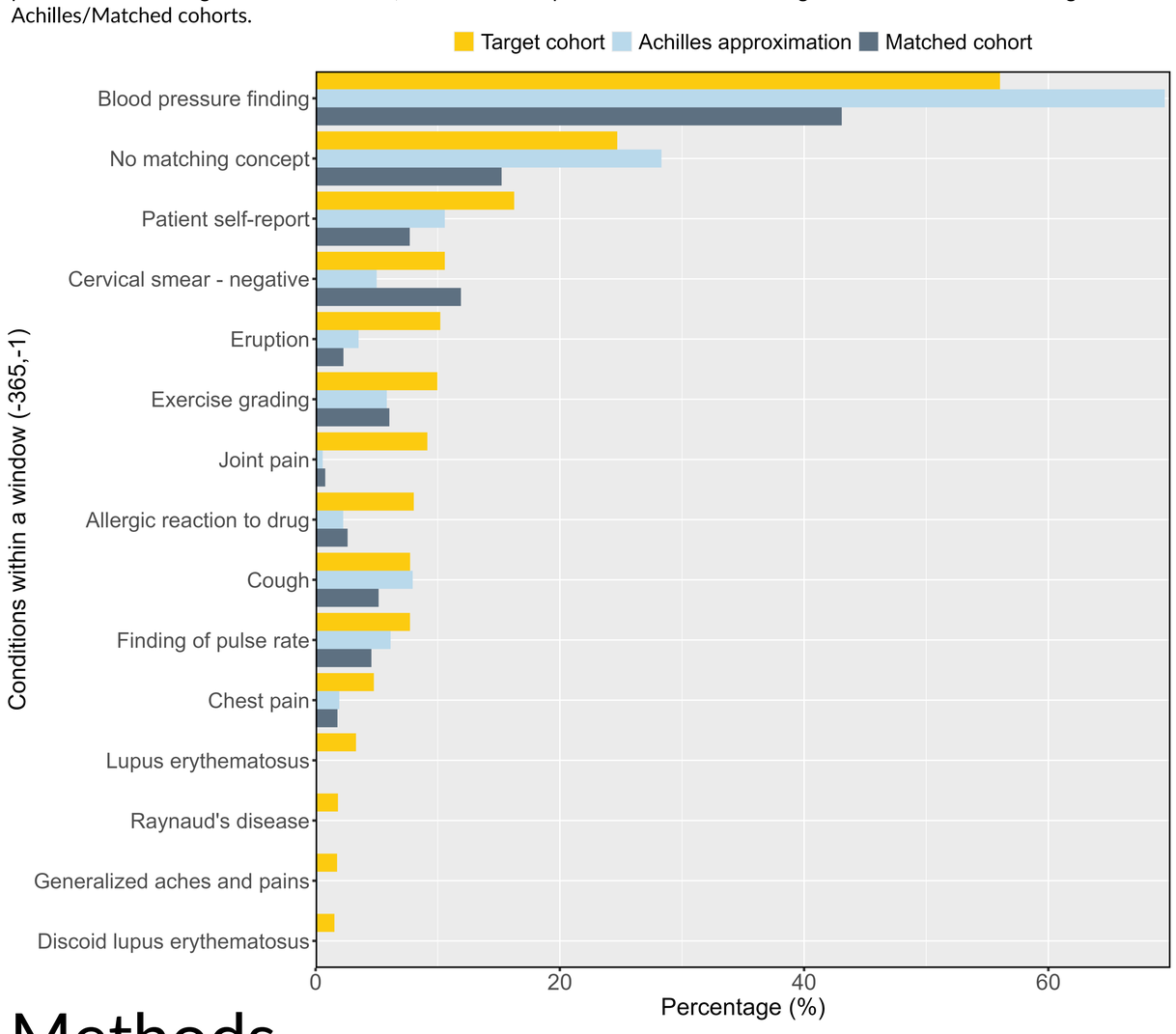
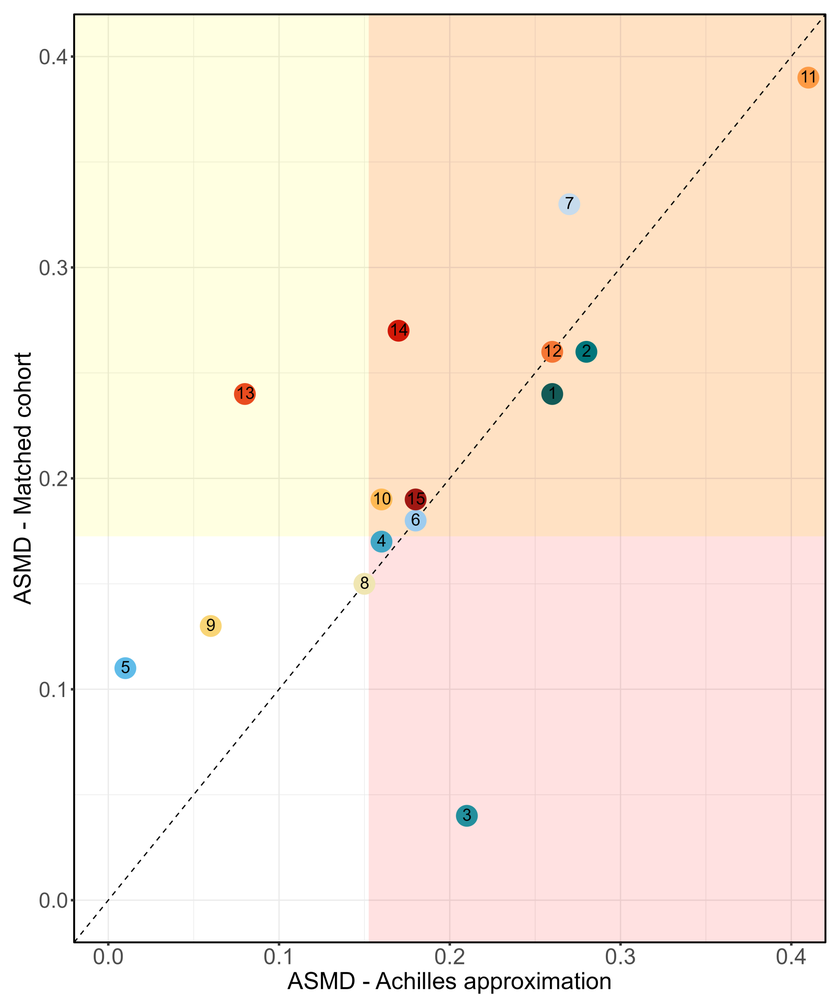


Figure 2. Comparison of ASMD between the Achilles and target cohorts (horizontal axis) and the Matched and target cohort (vertical axis). Top 10 conditions with the highest ASMD in each group are highlighted within the respective colored areas.



See all the results in the Shiny App!

Methods

Large scale characterisation of our target cohort

- Assume we have a database with N_p participants and a cohort with N_c individuals.
- We perform a large-scale characterisation of our cohort using the R package *PatientProfiles*.
- Only conditions that have an occurrence higher than 0.5% in our cohort within a time window w_1 are included.

Contextualisation using a matched cohort

- Create a matched cohort with individuals that (1) are not included in the target cohort and (2) have the same year of birth and sex:

```
cdm[["matched_cohort"]] <- CohortConstructor::matchCohorts(
                           cohort    = cdm$cohort,
                           name      = "matched_cohort",
                           matchSex = TRUE,
                           matchYearOfBirth = TRUE,
                           ratio     = 1)
```

Contextualisation using ACHILLES tables

- Create a pseudo-population within which the conditions are distributed uniformly across the database.
- Events of each condition expected in our cohort of N_c individuals within a window w_1:

$$N_{expected} = \frac{N_{events}}{person-days} \cdot N_c \cdot w_1$$

- Expected percentage of events in our cohort within a window w_1:

$$\%_{expected} = \frac{N_{expected}}{N_c} \cdot 100$$

- Assumptions:
  (1) No repetitive events in the same window
  (2) All individuals are observed in the window

Conclusion

- Proportions based on Achilles provide estimated counts if conditions are distributed uniformly, but otherwise can led to spurious context.
- *matchCohorts()* can be used to create a comparable matched cohort based on age and sex.

Marta Alcalde-Herraiz[1], Edward Burn[1], Mike Du[1], Yuchen Guo[1], Núria Mercadé-Besora[1], Albert Prats-Uribe[1], Daniel Prieto-Alhambra[1,2], Martí Català[1]

[1]Centre for Statistics in Medicine, Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences, University of Oxford, UK
[2]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, the Netherlands