Documentation simulated data MM trajectories v1

Caterina Gregorio, Valentina Manzoni

April 2025

Data and organization

Schematic description of the scenarios

Scenario number	Folder name*	Number of disease patterns	Number of individuals	Observation scheme	Under reporting	Released
1	schema_xa_3000	2	3 000	Regular, 2 years (A)	No	✓
2	schema_xb_3000	2	3 000	Semi-regular, 3/5 years (B)	No	✓
3	schema_xc_3000	2	3 000	Irregular (C)	No	✓
4	schema_xa_under_3000	2	3 000	Regular, 2 years (A)	Yes	✓
5	schema_xb_under_3000	2	3 000	Semi-regular, 3/5 years (B)	Yes	✓
6	schema_xc_under_3000	2	3 000	Irregular (C)	Yes	✓
7	schema_xa_1000	2	10 000	Regular, 2 years (A)	No	✓
8	schema_xb_1000	2	10 000	Semi-regular, 3/5 years (B)	No	✓
9	schema_xc_1000	2	10 000	Irregular (C)	No	✓
10	schema_xa_under_1000	2	10 000	Regular, 2 years (A)	Yes	
11	schema_xb_under_1000	2	10 000	Semi-regular, 3/5 years (B)	Yes	
12	schema_xc_under_1000	2	10 000	Irregular (C)	Yes	

*100 datasets corresponding to different pseudo-random generations of the data are contained in each folder.

Dataset structure

The structure of the datasets are coherent across scenarios. Each dataset is organized in a long-format i.e. each row corresponds to a different visit. Subjects can have different number of rows depending on the scenario and time in the study. Below the variables contained in each dataset are described:

- `dataset_id` : identification number of the dataset (from 1 to 100).

- `subject_id`: identification number of the subject (from 1 to the number of subjects according to the scenario.
- `age_baseline`: age at the entry of the study in years.
- `age_exit`: age at the end of the observation period (death, loss to follow up or administrative end of the study) in years.
- `dth`: binary variable indicating whether the subject died (0: alive at the end of the observation period; 1 dead at the end of the observation period).
- `time_in_study`: time in years from the study entry until the end of the observation (`age_exit-age_entry`) in years.
- `cov1`, `cov2`, `cov3`: binaries variables indicating three different exposures that can be used as possible predictors of transitioning between the multimorbidity states and death. They are assumed to be fixed and measured at the study entry.
- `visit_number`: identification number for the visit.
- `age`: age at the study visit in years.
- `ndis`: number of prevalent diagnoses at the study visit.
- from `anemia` to `peripheral_vascular_dis`: 60 binary variables indicating a specific diagnosis at the study visit. All diseases are assumed to be chronic and irreversible (they can't be "turned off").

Description of the observation schemes

- **Schema A- Regular 2 years:** visits are every two years from baseline until death/ end of the observation due to loss-to-follow-up/administrative end of the study. The median number of visits per subject is 5 (IQR: 3-7).
- **Schema B- Semi-Regular similar to SNAC-K:** visits are every 6 years from baseline if the subject is 60-78 and every 3 years if they are 78+ until death/end of the observation due to loss-to-follow-up/administrative end of the study. The median number of visits per subject is 2 (IQR: 2-3).
- **Schema C - Irregular:** visits are at irregular intervals and they are different among subjects. The median number of visits per subject is 6 (IQR: 3-10).

Definition of under-reporting

Underreporting was defined as a random selection of diagnoses for five pre-specified diseases (Chronic Kidney Disease, Dementia, Deafness/Hearing Loss, Depression, and Osteoarthritis/Other Degenerative Joint Diseases) that were not detected during a visit. For simplicity, the probability of non-detection was kept constant across both time and diseases.

Data generating Mechanism

For each dataset n ($n = 1, 2, \ldots, N$) and for each subject k ($k = 1, 2, \ldots, N_{\text{sim}}$):

Population composition

1. Draw `cov1` from a binomial distribution:
   $cov1_k \sim \text{Binomial}(p = 0.45)$.

2. Draw `cov2` from a binomial distribution:
   $cov2_k \sim \text{Binomial}(p = 0.15)$.

3. Draw age at entry from a truncated gamma distribution with shape 0.9 and rate 0.15, constrained between 60 and 96:
   $A_k \sim \text{Gamma}(\alpha = 0.9, \beta = 0.15), \quad 60 \leq A_k \leq 96.$

   MM patterns, chronic diseases and survival

4. Simulate cluster at entry in the study and then simulate the diseases at baseline conditioned on the clusters to which they belong. For each disease, draw from a binomial distribution with probability p from the latent class model (i.e., the probability of developing a certain disease given a multimorbidity cluster and the age at entry).

5. Simulate latent multimorbidity cluster trajectories using a multi-state model with a Gompertz hazard, adjusted for the three binary covariates .

6. Simulate the prevalent diseases (among those the patient has not yet developed), conditioned on the latent cluster towards which the patient is transitioning. If the next state of transition is Death, then diseases are simulated based on the current state.

7. Simulate the age of onset for each developed disease from a truncated beta distribution based on disease-specific parameters. The distribution is truncated so that the age of onset falls between the transition from the previous state to the following. If the computed age exceeds the age of death, the corresponding simulated disease is discarded.

8. Simulate rare diseases independently of the states to which patients belong but dependently on the patient's age. Rare diseases are drawn from a binomial distribution with parameter p equal to the prevalence of such diseases stratified by age, as reported in Appendix A.

9. Remove subjects who do not present multimorbidity at baseline.

   End of the observation period

10. Draw right-censoring time from a uniform distribution:
    $T_k \sim \text{Uniform}(0.5, 20)$.

11. Compute the age of exit from the study as the minimum between the age of death and the sum of the age at entry and right censoring time.

12. Eliminate data after the age of exit (in the case of patients who leave the study before dying).

    Study design

Based on the data generation mechanism described, the underlying "true" datasets (ground truth) are obtained. Then, study design schemes are applied to derive the observed data through the study.

Following a specific observation schema, visit times are recorded for each subject, and the exact onset times of diseases are replaced by binary variables indicating whether a diagnosis was observed at the visit.

For schema A and schema B, visit times are deterministically simulated starting from the age of entry into the study.

In contrast, for schema C, visit times are simulated using a Weibull distribution with parameters shape = 5 and scale = 0.4, starting from the age of entry.

Examples:

- **Subject A**: Diabetes onset at age 60, study entry at age 65. The diabetes diagnosis variable is set to 1 at the baseline visit (first visit in the dataset).

- **Subject B**: Diabetes onset at age 65, study entry at age 62. The diabetes diagnosis variable is set to 1 at the closest visit after age 65 in a scenario without under-reporting. In a scenario with under-reporting, the diabetes diagnosis might appear later if an under-reported case is simulated for Subject B at the closest visit.