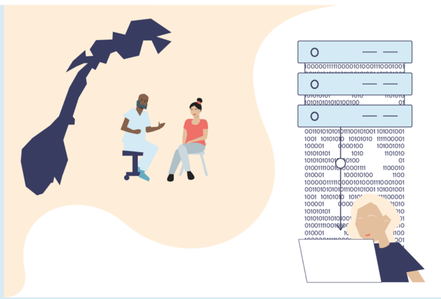
Meeting the challenges of reproducibility and transparency in analyses of cohort and registry data with open-source software: examples from the PsychGen Centre for Genetic Epidemiology and Mental Health

Alejandra Martinez Sanchez[1] and Laurie J. Hannigan[1,2]

[1] PsychGen Centre for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, [2] Psychiatric Genetic Epidemiology Group, Research Department, Lovisenberg Diaconal Hospital

Cohort and registry data in research



Modern epidemiological research increasingly relies on data from a combination of cohort studies and population registries:

- Cohort data collected over several years often contain detailed information on health, behavior and environment.
- Registry data offer high-quality and broad-coverage information collected across many years.

These data have vast research potential, as they can be reused to revisit or address new research questions without the need of new data collection. However, analyses of secondary data can present specific challenges for reproducibility and transparency:

- Repeated re-use of data by separate researchers can lead to unintended variations and discrepancies in the data preparation and handling process.
- Complex data preparation and analysis is rarely fully documented with well-commented, interoperable code.

Our approach

To realize the potential of these secondary data sources and mitigate risks from unwanted variability, researchers need access to tools and practices that facilitate their efficient, reproducible, and transparent preparation and usage.

We present two such tools: the `{phenotools}` and `{regtools}` R packages.

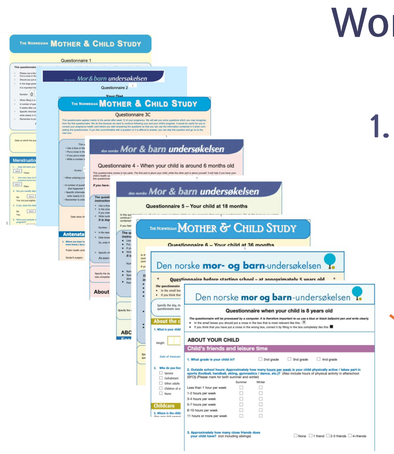
{phenotools}

Aims

GitHub repository

- Enable consistent, well-documented, and reproducible use of phenotypic data from the Norwegian Mother, Father, and Child Cohort study (MoBa) across projects.
- Automate of time- and/or resource-consuming and error-prone tasks.
- Encode specialist knowledge to facilitate high(er) quality interdisciplinary work.
- Facilitate transparent, robust re-analysis/revision, clear reporting and worthwhile analytic code-sharing.

Workflow and functionality



Helsedirektoratet

1. Create projects structured for reproducibility and transparency
2. Curate datasets across MoBa questionnaires and linked registry sources
3. Code variables reproducibly, based on domain-specific expertise

preg_id	Barn_nr	M_id	F_id	Dep_01_c
00001	1	003	101	2.5
00001	2	003	101	3.4
00002	1	005	103	1.3

phenotools *(Hannigan et al., 2021)* has been used in more than 15 published scientific projects

{regtools}

Aims

- Facilitate reproducible data preparation and descriptive epidemiological analyses based on Norwegian registry health data, supplemented with sociodemographic information from other registry sources (e.g. Statistics Norway).
- Provide "hands-on" guidance on how to work with individual-level registry data for epidemiological research.
- Aid with the efficient and consistent analysis and updating of public health reports.

Workflow and functions

GitHub repository

1. Read and validate data
2. Filter data by diagnostic codes and sociodemographic groups
3. Link health and administrative data
4. Analyze: incidence and prevalence rates
5. Visualize: line plots, bar charts, choropleth maps

- Logs that documents each function's internal data processing, warnings/errors, and corresponding outputs.
- Supports parquet files for the manipulation of larger-than-memory files.
- Helper functions: creation of synthetic registry-like data and harmonization of historical Norwegian municipality codes.

The workflow and functions of **regtools** have been used in the Norwegian Institute of Public Health's Thematic Issue on Mental Health of Children and Adolescents (Martinez Sanchez et al., 2025).

Impact and other considerations

- Open-source software can help standardize and reduce the time invested in data preparation and analysis of secondary data, such as registries and cohort data.
- Well-documented R packages and workflows promote reproducibility across different research projects using the same data sources.
- It is crucial to have a plan for continuous integration, development and maintenance of these tools.

References:

Martinez Sanchez, A., Pettersen, J., Bang, L., Bjuland, K., Scheiene, M., Aase, H., & Havdahl, A. (2025). Thematic Issue of the Public Health Report 2025 – Mental Health of Children and Adolescents (p. 85). Norwegian Institute of Public Health. https://www.fhi.no/contentassets/b5b3603ec4794c5cb0c8651589b359f8/temautgave-barn-og-unges-psykiske-helse_2025.pdf

Hannigan, L., Corfield, E., Askelund, A., Askeland, R., Hegemann, L., Jensen, P., Pettersen, J., Rayner, C., Ayorech, Z., & Bakken, N. (2021). phenotools: An R package to facilitate efficient and reproducible use of phenotypic data from MoBa and linked registry sources in the TSD environment. https://doi.org/10.17605/OSF.IO/6G8BJ

Contact:

Alejandra Martinez Sanchez
alejandra.martinez.sanchez@fhi.no
Laurie Hannigan
laurie.hannigan@fhi.no

Norwegian Institute of Public Health

DOI: https://doi.org/10.7557/5.8265