The Journal of Open Source Software

regtools: facilitating manipulation, analysis and visualization of data from Norwegian health and population registers

Alejandra Martinez Sanchez [1], **Johanne Hagen Pettersen** [1,2], **Helga Ask** [1], **Alexandra Havdahl** [1,3], **and Laurie John Hannigan** [1,4]

1 PsychGen Centre for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health **2** Center for Precision Psychiatry, University of Oslo **3** PROMENTA Research Center, Department of Psychology, University of Oslo **4** Psychiatric Genetic Epidemiology Group, Research Department, Lovisenberg Diaconal Hospital

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor:

Submitted: 05 November 2025
Published: unpublished

Summary

The use of health and administrative registers, often in combination, is an essential component of modern epidemiological research. Among the Nordic countries, in particular, an array of registry sources offers high-quality, broad-coverage data collected across many years (Laugesen et al., 2021). The use of existing data from registers in research circumvents the data collection process, thus making research more cost and time effective (Thygesen & Ersbøll, 2014). Moreover, health and administrative registers offer enormous sample sizes and high representativeness that are much needed, particularly in epidemiological research. Although registers are rich sources of information, pre–processing and working with the large datasets they produce can be challenging and time-consuming – especially for researchers with limited programming experience – and the process is vulnerable to both unintended variations across projects and highly consequential errors.

The `regtools` R package is an open-source toolkit designed to aid researchers in performing efficient and well-documented manipulation, analysis and visualization of individual-level data from Norwegian health and population registers. With it, we aim to facilitate reproducible descriptive epidemiology based on Norwegian health data, supplemented with sociodemographic information, such as income and education information, from other registry sources. The package includes functions to validate, filter, and link health (diagnostic) and administrative (sociodemographic) data. For transparency, each function creates a log that documents the function's internal data processing, warnings/errors, and corresponding outputs. Finally, considering the extensive use of registers in epidemiological research, `regtools` includes functions intended to help users compute common descriptive epidemiology statistics, such as prevalence and incidence rates, and visualize the underlying data.

Statement of need

Due to their characteristics, Nordic registers are highly regarded for their unique potential in current epidemiological research (Jervelund & Montgomery, 2020; Maret-Ouda et al., 2017). In the last decades, epidemiological research in the Nordic countries has harnessed the advantages of registry data, such as primary and secondary health care registers (Miettunen et al., 2011; Thygesen & Ersbøll, 2014). In part, this is due to the introduction of personal identification numbers into the Nordic population-based health registers, which enables linkage to other data sources, allowing long-term, multi-dimensional follow-up of individuals in the population.

Registry data from national statistical institutes (NSIs) are a widely-used source of auxiliary information in this regard.

In Norway, the Norwegian Patient Registry (NPR) is used in a large variety of research projects (Bakken et al., 2020). As of 2025, more than 1000 research papers have been published based on data from the NPR (Norwegian Institute of Public Health, 2024). Statistics Norway (SSB) provides sociodemographic individual-level data on various topics, such as social welfare, education and income. For instance, between 2021 and 2024, SSB delivered around 900 individual-level data assignments to both public authorities and research institutes for analytic and research purposes (Statistics Norway, 2024, 2025). Despite their relatively widespread use in research, health and administrative registers are not designed with research or statistical purposes in mind. This creates numerous potential challenges, inefficiencies, and vulnerabilities in the process of carrying out epidemiological research using linked register data.

Considering the wide range of researchers using individual-level registers in Norway, it is highly likely that there are differences in the way researchers pre-process and prepare their data for analysis. Access to register individual-level data is regulated by strict confidentiality laws, which makes "hands-on" training or tutorials hard to access and standardize. The use of proprietary software to manipulate and analyze the data further hinders efforts to ensure reproducibility and transparency across research projects working with the same data (Mathur & Fox, 2023). In this context, we have identified the need for an open-source toolkit to assist researchers working with Norwegian individual-level registry data to prepare, manipulate, and analyze it in a robust, transparent, and reproducible way.

While other projects (e.g., `phenotools` (Hannigan et al., 2021), `csverse` (White, 2025)) have showcased the potential of using open-source R packages to assist researchers working with Norwegian survey and register data, the `regtools` package is the first to focus on larger individual-level data with descriptive epidemiological analyses in mind. As an example of a likely use case for `regtools`, the package has been successfully used to analyze time trends in autism diagnoses in Norway over recent years for a public health report (Martinez Sanchez et al., 2025). The functions included in the package are modular and operate independently from one another, which increases their possible application in various research projects. Given the potential of multinational registry-based cohort studies (Maret-Ouda et al., 2017), it is important to note that, while the package workflow is originally designed for Norwegian data sources, its flexibility may allow for use with other national registries.

One of the first challenges researchers working with population-based registers encounter is that of efficiently manipulating very large datasets into smaller and tidier datasets with which they can work analytically. The `regtools` package includes reading and filtering functions that support files in parquet format (Apache Parquet, 2025), which seamlessly enables users to efficiently work with larger-than-memory files in R without requiring deeper knowledge on the inner workings of parquet format objects. Furthermore, the logs created by each function can help researchers keep track of and document all manipulation or processing steps applied to their datasets. The package also includes functions that are particularly useful for descriptive epidemiology analyses, such as the computation of prevalence and incidence rates, along with a function for visualizing the results. There are some specific challenges related to Norwegian registry data that are addressed in the helper functions of `regtools`, such as harmonizing municipality codes and retrieving population counts from SSB's open data.

In addition to helping solve practical challenges associated with processing, manipulation, and analysis of Norwegian register data, `regtools` provides "hands-on" guidance on how to efficiently work with individual-level registry data for epidemiological research. The functions in the package are intended to serve as a loose framework that can be adapted by researchers working with similar data and research questions. The package includes a series of vignettes explaining the main functions and real-life examples of descriptive epidemiology. The vignettes and possibility of creating synthetic individual-level datasets (`synthetic_data()`) also allow research-groups to use the package as zero-risk training material for new members, and to plan and structure analytic projects prior to obtaining data access.

Acknowledgements

This work was performed on the TSD (Tjeneste for Sensitive Data) facilities, owned by the University of Oslo, operated and developed by the TSD service group at the University of Oslo, IT Department (USIT). (tsd-drift@usit.uio.no).

AMS was supported by the Research Council of Norway (#336085). JHP and HA were supported by the Research Council of Norway (#324620), and NordForsk(#156298). LJH was supported by the South-Eastern Norway Regional Health Authority (#2922083). AH was supported by the Research Council of Norway (#336085), the South Eastern Norway Regional Health Authority (#2020022), and the European Union's Horizon Europe Research and Innovation programme (FAMILY #101057529).

Thanks to Guido Biele and Lasse Bang for their assistance during the early stages of this project.

References

Apache Parquet. (2025). Documentation. In *Apache Parquet*. https://parquet.apache.org/docs/.

Bakken, I. J., Ariansen, A. M. S., Knudsen, G. P., Johansen, K. I., & Vollset, S. E. (2020). The Norwegian Patient Registry and the Norwegian Registry for Primary Health Care: Research potential of two nationwide health-care registries. *Scandinavian Journal of Public Health*, *48*(1), 49–55. https://doi.org/10.1177/1403494819859737

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

Jervelund, S. S., & Montgomery, C. J. D. (2020). Nordic registry data: Value, validity and future. *Scandinavian Journal of Public Health*. https://doi.org/10.1177/1403494819898573

Laugesen, K., Ludvigsson, J. F., Schmidt, M., Gissler, M., Valdimarsdottir, U. A., Lunde, A., & Sørensen, H. T. (2021). Nordic Health Registry-Based Research: A Review of Health Care Systems and Key Registries. *Clinical Epidemiology*, *13*, 533–554. https://doi.org/10.2147/CLEP.S314959

Maret-Ouda, J., Tao, W., Wahlin, K., & Lagergren, J. (2017). Nordic registry-based cohort studies: Possibilities and pitfalls when combining Nordic registry data. *Scandinavian Journal of Public Health*, *45*(17_suppl), 14–19. https://doi.org/10.1177/1403494817702336

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.

Mathur, M. B., & Fox, M. P. (2023). Toward Open and Reproducible Epidemiology. *American Journal of Epidemiology*, *192*(4), 658–664. https://doi.org/10.1093/aje/kwad007

Miettunen, J., Suvisaari, J., Haukka, J., & Isohanni, M. (2011). Use of Register Data for Psychiatric Epidemiology in the Nordic Countries. In *Textbook of Psychiatric Epidemiology* (pp. 117–131). John Wiley & Sons, Ltd. https://doi.org/10.1002/9780470976739.ch8

Norwegian Institute of Public Health. (2024). Vitskaplege publikasjonar basert på data frå NPR og KPR. In *Folkehelseinstituttet*. https://www.fhi.no/he/npr/vitenskaplege-publikasjonar-basert-pa-data-fra-norsk-pasientregister/.

Statistics Norway. (2024). Årsrapport 2023. In *SSB*. https://www.ssb.no/omssb/ssbs-virksomhet/planer-og-meldinger/statistisk-sentralbyras-arsrapport/arsrapport-2023.

Statistics Norway. (2025). Årsrapport 2024. In *SSB*. https://www.ssb.no/omssb/ssbs-virksomhet/planer-og-meldinger/statistisk-sentralbyras-arsrapport/arsrapport-2024.

Thygesen, L. C., & Ersbøll, A. K. (2014). When the entire population is the sample: Strengths and limitations in register-based epidemiology. *European Journal of Epidemiology*, *29*(8), 551–558. https://doi.org/10.1007/s10654-013-9873-0

White, R. (2025). *CSIDS - Consortium for Statistics in Disease Surveillance*. https://www.csids.no/.