The Journal of Open Source Software

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

Alejandra Martinez Sanchez [1,2], **Johanne Hagen Pettersen** [1,3], **Helga Ask** [1], **Alexandra Havdahl** [1,4], **and Laurie John Hannigan** [1,5]

1 PsychGen Centre for Genetic Epidemiology and Mental Health, Norwegian Institute of Public Health, Oslo, Norway **2** Department of Psychology, University of Oslo, Oslo, Norway **3** Center for Precision Psychiatry, University of Oslo, Oslo, Norway **4** PROMENTA Research Center, Department of Psychology, University of Oslo, Oslo, Norway **5** Psychiatric Genetic Epidemiology Group, Research Department, Lovisenberg Diaconal Hospital, Oslo, Norway

DOI: 10.21105/joss.10240

Software

- Review
- Repository
- Archive

Editor: Julia Romanowska

Reviewers:

- @ilovemane
- @turkulerc

Submitted: 05 November 2025
Published: 03 September 2026

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 might create both unintended variations across projects and errors.

The `regkit` 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 the Norwegian health and population registers. With `regkit`, we aim to facilitate reproducible descriptive epidemiology based on Norwegian health data, supplemented with sociodemographic information, such as income and education information, from other registries. 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, `regkit` includes functions intended to help users compute common descriptive epidemiological statistics, such as prevalence and incidence rates, and visualize the underlying data.

Statement of need

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 showcased many of the advantages of these registry data (Bakken et al., 2020; Laugesen et al., 2021; Ludvigsson et al., 2016; Viippola et al., 2023). In part, Nordic registers' suitability for research is due to the existence of personal identification numbers, which enable linkage of data for each person from various registries, allowing long-term, multi-dimensional follow-up of individuals in the population.

Sociodemographic data from national statistical institutes are a widely-used source of auxiliary information in epidemiological research thanks to this possibility of cross-registry linkage.

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. 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 difficult 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 them in a robust, transparent, and reproducible way.

State of the field

Considering the confidential nature of health and administrative register data, a large proportion of the code used in research is not publicly available. Nonetheless, some few open-source software solutions have been developed to deal with the processing and analysis of Norwegian survey and register data. For instance, `phenotools` (Hannigan et al., 2021) is an R package that aims to facilitate efficient and reproducible use of survey information from Norwegian cohort data. Similarly, `niphr` (White, 2026) is a suite of packages focusing on the preparation and analysis of Norwegian disease surveillance and linked register data. Finally, the Norwegian statistics bureau (Statistics Norway) has developed a handful of packages addressing statistical disclosure control, retrieval of publicly available data, classification standards and survey weighting and estimation (Jentof et al., 2026; Langsrud & Bruusgaard, 2026; Langsrud & Lupp, 2025).

In summary, these existing projects have showcased the potential of using open-source R packages to assist researchers working with Norwegian survey and register data. The justification for building `regkit` as a new package, rather than contributing to existing packages, is that currently available software focuses on relatively narrow, well-defined use-cases (e.g., suppression of sensitive information in data tables, disease surveillance) as opposed to broad operations with large, linked individual-level datasets aimed at facilitating descriptive epidemiological analyses. In addition, building a new package has enabled us to create a streamlined framework including only functions relevant to this research area. For instance, it integrates relevant functions originally developed by Statistics Norway to facilitate the creation of realistic simulated data.

Software design

The package `regkit` was built following the principles of modularity and flexibility, which increases its possible application in various research projects. For instance, although the package was originally designed for Norwegian data sources, its flexibility may allow for use with other national registries. Similarly, to improve interoperability across projects, `regkit` imports widely known packages (e.g., `ggplot2`, `dplyr`, `purrr`, etc.) that many users working

with register data are likely to have installed already. In addition, packages used in functions only relevant to a specific use case (e.g., `plot_map()`) are listed as *Suggest*, keeping the core installation as lean as possible. The current unit test suite provides good coverage of the main functions, ensuring the reliability of the package.

Considering that an average user of `regkit` (social scientists or epidemiologists) may have limited programming experience, we provide a user-friendly educational framework and accessible documentation of functions. For instance, one of the first challenges researchers working with population-based registers encounter is how to efficiently manipulate large datasets into smaller and tidier datasets with which they can work analytically. The `regkit` package includes reading and filtering functions (e.g., `read_diag_data()`, `filter_diag_data()`) that support files in parquet format (Apache Parquet, 2025), which enables users to work efficiently with larger-than-memory files in R without requiring deeper knowledge on the inner workings of parquet format objects. Furthermore, most functions automatically generate a log file that records and timestamps the function call, internal data transformations, warnings, errors, and general outputs. These logs 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 epidemiological analyses, such as the computation of prevalence and incidence rates, along with the function `plot_rates()` for visualizing the results. Moreover, there are some specific challenges related to Norwegian registry data that are addressed in the helper functions of `regkit`, 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, `regkit` 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 simulated individual-level datasets with the function `simulate_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.

Research impact statement

The most prominent use of the `regkit` package to date has been in providing analyses on time trends in autism diagnoses in Norway over recent years, published as part of a national public health report (Martinez Sanchez et al., 2025). The report received approximately 500,000 unique visits throughout 2025, and the analytic component based on work carried out using `regkit` was the most visited section. Core features of this report, which included regional breakdowns and visualization of autism incidence spanning 12 years and prevalence analyses with stratification by demographic characteristics retrieved from Statistics Norway, showcase the key advantages of the software. The package has additionally been used for analyses of time trends in other conditions (e.g., anxiety and depression), with publication of results forthcoming. Moreover, community-readiness has been demonstrated by the acceptance of talks and posters presenting `regkit` at conferences in the fields of epidemiology, autism research, and reproducibility in scientific publishing (e.g. Martinez Sanchez & Hannigan, 2025).

AI Usage Disclosure

No generative AI tools were used in the development of this software or the writing of this manuscript.

Acknowledgements

Due to the sensitive nature of the data used to develop and test `regkit`, the core development of the package 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). TSD is a secure research platform with limited internet access, therefore the commit history in the publicly available repository does not reflect contributions made within TSD. All the co-authors contributed conceptually to the iterative development of the package.

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 (#2022083). 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 program (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

Jentof, S., Iancu, D.-C., & Berntsen, Ø. (2026). *klassR: Classifications for Statistics Norway*. https://doi.org/10.32614/cran.package.klassr

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

Langsrud, Ø., & Bruusgaard, J. (2026). *PxWebApiData: PX-Web Data by API*. https://doi.org/10.32614/cran.package.pxwebapidata

Langsrud, Ø., & Lupp, D. (2025). *SSBtools: Algorithms and Tools for Tabular Statistics and Hierarchical Computations*. https://doi.org/10.32614/cran.package.ssbtools

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

Ludvigsson, J. F., Almqvist, C., Bonamy, A.-K. E., Ljung, R., Michaëlsson, K., Neovius, M., Stephansson, O., & Ye, W. (2016). Registers of the Swedish total population and their use in medical research. *European Journal of Epidemiology*, *31*(2), 125–136. https://doi.org/10.1007/s10654-016-0117-y

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., & Hannigan, L. (2025). 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. *Septentrio Conference Series*, *2*. https://doi.org/10.7557/5.8265

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; chapter 3: Time trends in the prevalence of autism diagnoses* (p. 85). Norwegian Institute of Public Health. ISBN: 978-82-8406-508-3

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

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

Viippola, E., Kuitunen, S., Rodosthenous, R. S., Vabalas, A., Hartonen, T., Vartiainen, P., Demmler, J., Vuorinen, A.-L., Liu, A., Havulinna, A. S., Llorens, V., Detrois, K. E., Wang, F., Ferro, M., Karvanen, A., German, J., Jukarainen, S., Gracia-Tabuenca, J., Hiekkalinna, T., … FinnGen. (2023). Data Resource Profile: Nationwide registry data for high-throughput epidemiology and machine learning (FinRegistry). *International Journal of Epidemiology*, *52*(4), e195–e200. https://doi.org/10.1093/ije/dyad091

White, R. (2026). *NIPH R Packages*. https://niphr.github.io/.