

From register data to useful information: Framework for automating real-world evidence reporting

Jani Miettinen¹, Reijo Sund^{1,2}

¹ Institute of Clinical Medicine, University of Eastern Finland, Kuopio, Finland; ² Clinical Research Centre, Kuopio University Hospital, Kuopio, Finland

Jani Miettinen, MSc, Institute of Clinical Medicine, University of Eastern Finland, PO Box 1627, FI-70211 Kuopio, FINLAND. Email: jani.miettinen@uef.fi

Abstract

Real-world data (RWD) are increasingly used in scientific research. However, transforming such data into structured and reproducible scientific evidence remains challenging. We present an automated analytical framework for transforming real-world data into standardized and reproducible real-world evidence (RWE) research reports. The novelty of this work lies in the integration of data harmonization, automated statistical modeling, and report generation within a lightweight graphical user interface supported by a simple common data model.

The framework standardizes key analytical components, including cohort construction, time-to-event modelling, and descriptive and summary reporting. As a proof of concept, the system is demonstrated using longitudinal data from OSTPRE cohort (Kuopio Osteoporosis Risk Factor and Prevention Study). The implementation illustrates how an automated workflow can efficiently generate transparent and consistent RWE outputs suitable both for research and healthcare system evaluation.

The framework is scalable to broader data ecosystems, such as regional hospital data repositories, enabling more detailed analyses and the production of robust real-world evidence. This approach supports more efficient utilization of real-world data in scientific and clinical decision-making.

Keywords: data science, big data, medical informatics, open source software, automation, health care research

Introduction

Healthcare systems routinely generate large volumes of patient information during clinical care. These data, stored in registries as by-products of treatment processes, represent real-world data (RWD) and capture diagnoses, treatments, and outcomes without experimental intervention. RWD has become increasingly valuable for research, especially for studying disease patterns and treatment effects [1,2]. However, converting raw RWD into reliable real-world evidence (RWE) requires structured data modelling, appropriate statistical methods, and adherence to reporting standards [3,4].

Data modelling provides a conceptual representation of healthcare data that can be translated into a logical database structure [5]. A core objective in health research is to harmonize heterogeneous datasets through common data models (CDMs), which standardize variables to enable reproducible analyses such as regression or survival modelling. CDMs vary in scope; the appropriate model depends on research aims and the level of detail required, and broadly any generalized structure supporting multi-source analysis may be considered a CDM [6].

The Observational Medical Outcomes Partnership (OMOP CDM) is widely used and supported by an active open data science community offering extensive tools [7–9]. OMOP excels in categorical harmonization and federated analyses but is resource-intensive to implement. For many local research needs – especially within regulated environments such as those governed by Finnish secondary-use legislation or the emerging European Health Data Space (EHDS) – lighter models may offer sufficient structure while enabling more agile workflows.

Producing scientific analyses from RWD is often time-consuming. Although commercial tools like Tableau provide strong visualization features, they are not designed for reproducible scientific workflows and may incur substantial licensing costs. These limitations highlight the need for open-source, transparent, and scalable analytical solutions [10].

R Shiny enables interactive, web-based applications, and many existing Shiny tools support tasks such as exploring patient trajectories, survival analysis, or pattern visualization [9,11–14]. However, these tools typically address isolated analytical components rather than providing an end-to-end research workflow. In contrast, our framework integrates data harmonization, reproducible model execution, and structured reporting within a unified automated pipeline, supporting full exposure–response analysis rather than single-purpose modules.

RWE plays an increasingly important role in healthcare decision-making [2,15,16]. This work presents a framework for transforming RWD into RWE, using the association between diabetes and ischemic heart disease as a proof of concept. Beyond this case study, the framework contributes methodologically by operationalizing automated workflows, defining a lightweight CDM, and enabling extensible analytical pipelines. Central to the framework is a Shiny-based application that automates scientific analyses and supports systematic evaluation of exposure–response relationships. The primary contribution of this study is therefore methodological: it provides a generalizable framework for assessing associations between pairs of conditions without redesigning analyses for each new case, with the epidemiological example included solely to demonstrate the framework’s applicability.

Material and methods

A framework for transforming RWD into RWE can be understood by organizing the analytical workflow into sequential layers. These layers form a structured pipeline that progresses from raw registry data to interpretable scientific results. Figure 1 summarizes the overall process, beginning with raw data ingestion, followed by cleaning and preprocessing, harmonization into a common data model (CDM), statistical analysis, and automated reporting. Each layer builds on the previous one, ensuring methodological transparency and reproducibility. In this study, we focus on the CDM, Analytics, and Reporting layers, as these constitute the core methodological components of the application developed for the OSTPRE (Kuopio Osteoporosis Risk Factor and Prevention Study) dataset. The key steps

required to achieve an automated reporting system are data modelling (Step 1) and analysis automation (Step 2).

Raw data layer

Health and social welfare databases provide RWD that is typically messy, unstructured, and unsuitable for direct analysis [17]. Raw entries may include duplicates, inconsistent timestamps, and heterogeneous coding formats. Therefore, preprocessing steps are required to remove errors, consolidate repeated events, and organize information into clinically meaningful representations such as longitudinal patient histories and diagnostic timelines [18]. These preparations form the foundation for creating a harmonized structure in the subsequent CDM layer.

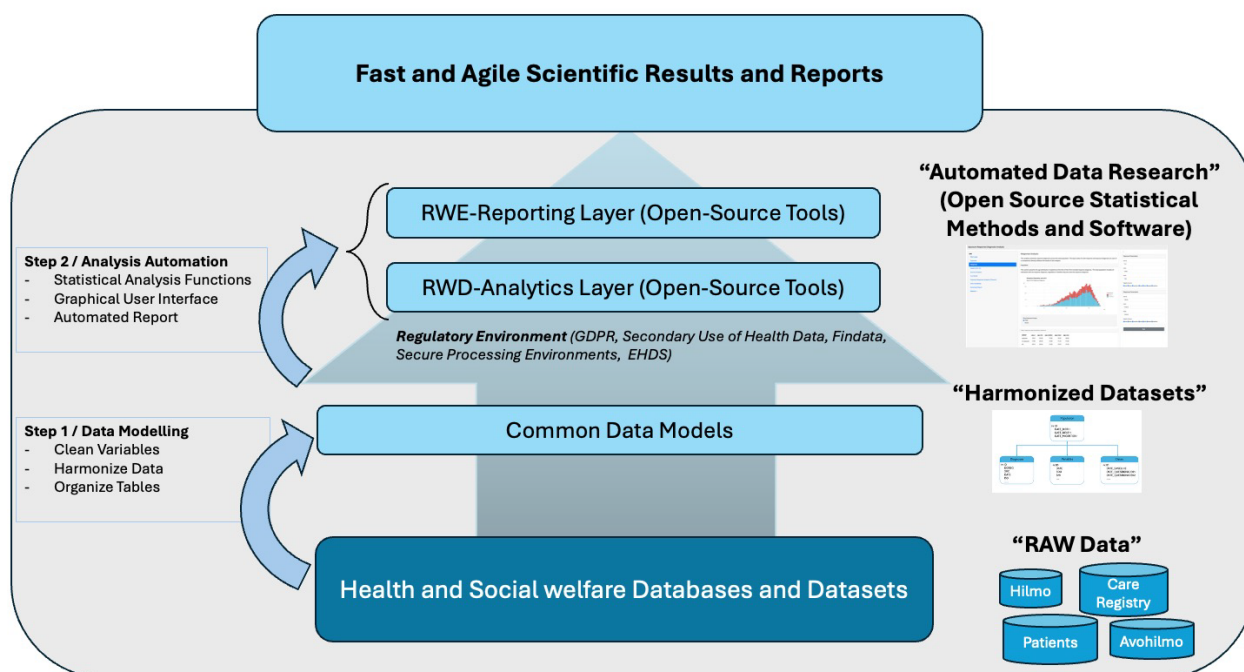


Figure 1. Data pipeline from raw registries through CDM harmonization to automated analytics and reporting. Step 1 highlights the importance of data harmonization, while Step 2 illustrates the automation of statistical analysis and reporting.

Common data model layer

The CDM establishes a harmonized, analysis-ready structure in which variable names, identifiers, and row formats are standardized. This aligns with tidy-data principles: each variable in a column, each observation in a row, and each value in a single cell [19]. By transforming heterogeneous source files into a uniform format, the CDM enables seamless integration of datasets from different registry systems.

CDM Construction using the OSTPRE cohort

The OSTPRE cohort includes approximately 14,200 women aged 47–56 at baseline in the late 1980s, later expanding to around 16,000 participants. The cohort has been followed for multiple decades, producing extensive questionnaire data, lifestyle information, and hospital visit records, making it ideal for studying long-term comorbidity patterns beyond osteoporosis [20–23].

The dataset comprises several subcomponents: demographic variables, year-specific questionnaires,

and hospital visits categorized using ICD-8, ICD-9, and ICD-10 codes. Preparation for analysis required systematic cleaning to remove errors, correct entries, and resolve missing values. Afterwards, related data sources were merged into structured groups, such as unified diagnosis events, demographic profiles, and time-dependent factors including death or emigration.

The CDM created for this research (Figure 2) is intentionally lightweight, designed to support the analytical requirements of the application rather than replicate comprehensive models such as OMOP. Several refinement cycles were needed to ensure compatibility with downstream statistical methods. Once finalized, transforming updated OSTPRE datasets into the CDM became a reproducible and efficient process.

Analytics layer

Once harmonized into the CDM, the data enter the Analytics Layer, where statistical methods suitable for longitudinal RWD are applied. We developed a

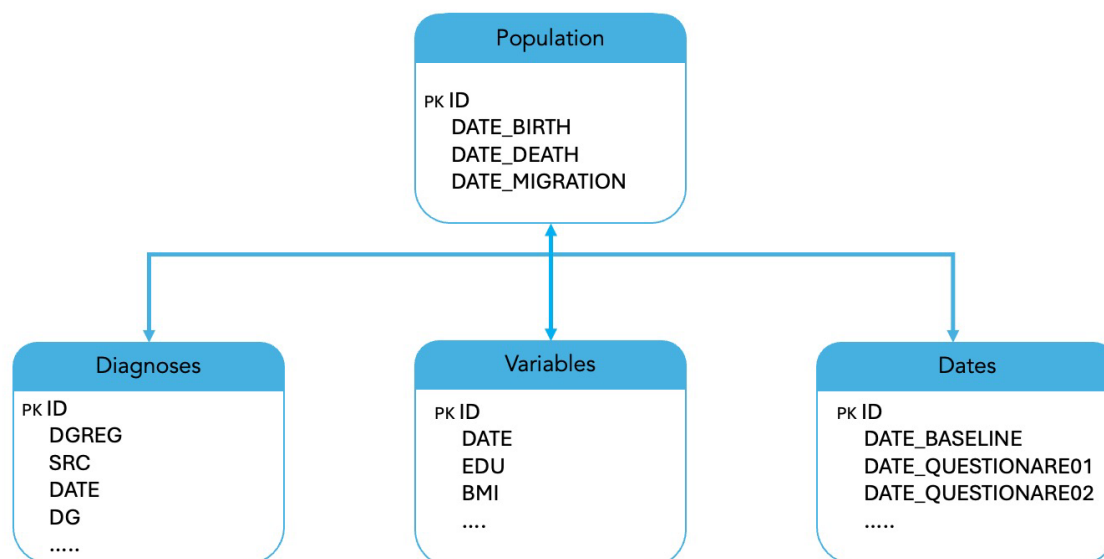


Figure 2. CDM structure integrating identifiers, ICD-8/9/10 codes, questionnaire cycles, and key variables for longitudinal analysis.

tailored research protocol to assess associations between two medical conditions, with methods carefully selected based on prior case studies where they proved most effective for drawing conclusions within the given research setting. While the statistical techniques themselves are standard, their correct and consistent application in a longitudinal, time-dependent context is nontrivial, even when performed manually. This protocol ensures that analyses are reproducible, transparent, and aligned with the structure of the underlying registry data.

To quantify how exposure conditions influence subsequent outcomes while accounting for follow-up time, censoring, and relevant covariates, we employed several complementary statistical methods:

Cross-tabulation provides a contingency table showing the frequency of co-occurrence between an exposure and a response condition. A Chi-square test evaluates whether the observed distribution differs from what would be expected under independence, indicating whether any association is unlikely to be due to chance. This method is useful for initial exploration before applying more complex time-dependent analyses.

Survival analysis examines the time until an event—such as onset of a medical condition or death—occurs. The analysis requires a defined starting point (e.g., baseline questionnaire, first occurrence of a diagnosis, or a specific age) and can use different time scales such as age or time since exposure. A key feature is the ability to handle censored observations, where follow-up ends before the event is observed. Survival models can also incorporate competing risks when events such as death preclude the outcome of interest. These techniques allow estimation of cumulative incidence and comparison of event probabilities between exposed and unexposed groups.

Standardized Incidence Ratio (SIR) quantifies whether the incidence of a condition in the study population differs from what would be expected based on age-specific rates in a reference population. In this study, SIRs were estimated using Poisson regression, stratifying follow-up into one-year age groups after exposure. This approach reveals whether the response condition occurs more or less frequently in the exposed group than expected, and how relative risk evolves over time.

Cox Proportional Hazards Model estimates hazard ratios to assess how covariates – such as BMI, age, or education – affect the instantaneous risk of the response condition over time. It accounts for varying follow-up durations and censoring, making it well suited to longitudinal registry data. The model also supports time-dependent covariates, enabling individuals who develop the exposure condition during follow-up to transition from unexposed to exposed groups. This allows examination of both the direct exposure–response relationship and the modifying effects of additional risk factors.

Implementation

All analytical steps were implemented as modular R functions that process CDM data, prepare model inputs, and generate standardized outputs. These functions have been compiled and documented into the *healthpopR* R-package [24], which supports reproducible computation and reduces analytic flexibility bias [25].

RWE-reporting layer

The Reporting Layer organizes the outputs produced by the Analytics Layer, focusing on results that are most relevant for interpretation and decision-making. Because statistical analyses generate large amounts of intermediate information, the

Reporting Layer highlights key evidence rather than presenting all outputs.

These analytical functions were integrated into a Shiny application that operates directly on the CDM. The application provides a browser-based interface where users can define exposure and response conditions using regular expressions for ICD-8/9/10 codes. Once defined, the application processes the data, fits the appropriate statistical models, and presents results through interactive visualizations and aggregated tables. All outputs preserve privacy by reporting only aggregated information, and results can be exported for documentation or further research use.

Results

To demonstrate the automated functionality of the framework and the Shiny application as a proof of concept, we present a real-world study example. The relationship between diabetes and ischemic heart disease (IHD) represents a well-suited scenario for illustrating the capabilities of the application in generating real-world evidence reports. Diabetes is a highly prevalent chronic condition worldwide, and cardiovascular complications remain among its most serious long-term outcomes [26]. A large body of epidemiological research has established that individuals with diabetes are at a markedly increased risk of developing IHD compared with the general population, with risk estimates ranging from a 2- to 4-fold increase [27,28]. Moreover, both conditions are widely coded and tracked in national health registries, making them particularly appropriate for demonstrating how registry-based follow-up studies can be standardized and automated. Within the application, the researcher can define exposure (diabetes) and

response (IHD) diagnoses consistently across decades using ICD coding systems. The full report is available on the GitHub page [29]. To provide step-by-step evidence-based conclusion, we introduce most important results.

Cross-tabulation approach allows researchers to quickly assess whether the prevalence of the response condition differs between exposed and unexposed groups, offering a baseline understanding before analyzing more complex, time-sensitive results. In the diabetes–IHD example, individuals with diabetes had a substantially higher proportion of IHD (60.5%) compared with those without diabetes (43.1%). This difference was highly significant ($\chi^2 = 409.5$, $df = 1$, $p < 0.001$), with a small-to-moderate effect size ($\phi = 0.16$). While informative, cross-tabulation does not capture how the incidence of IHD evolves over time following a diabetes diagnosis.

Competing-risk analysis is used to estimate the cumulative incidence of a response condition and death following an exposure condition (Figure 3). At the start of follow-up, the cumulative incidence may appear relatively high due to an optional data-handling rule in the framework: response events that occurred prior to the exposure can be reassigned to the exposure date to maintain temporal consistency. This approach visualizes the portion of patients who already had an IHD diagnosis at or before the time of diabetes onset. In the application, researchers can modify this reassignment depending on the study design and research question. As an illustrative example, the cumulative incidence of IHD increased steadily throughout follow-up, accounting for the competing risk of death. This analysis provides a clear view of how the automated workflow captures exposure–response dynamics in a reproducible and interpretable manner.

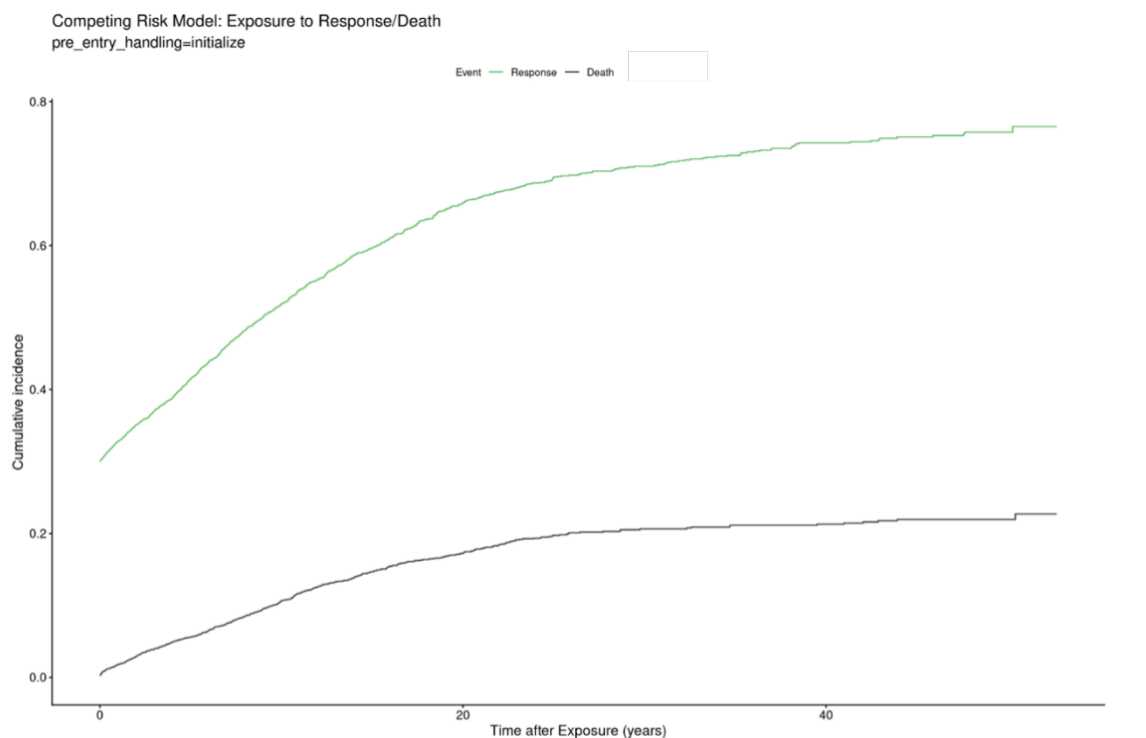


Figure 3. Competing-risk cumulative incidence curves for ischemic heart disease (IHD) and all-cause mortality following diabetes diagnosis. To maintain temporal consistency, IHD events recorded before the diabetes diagnosis were reassigned to the diagnosis date.

Standardized incidence ratio (SIR) analysis was conducted to compare the observed incidence of ischemic heart disease (IHD) in exposed individuals with the incidence expected based on reference population rates. This approach allows researchers to quantify how exposure alters risk relative to a standardized baseline and to examine temporal patterns of risk over follow-up. In the diabetes–IHD example, the SIR was highest during the first year after exposure (SIR 2.82, 95% CI 2.45–3.24), indicating a pronounced short-term elevation in risk. Although the relative risk decreased in subsequent exposure periods, it remained significantly elevated throughout follow-up, ranging from 1.51 (95% CI 1.35–1.68) during 1–4 years after exposure to approximately 1.7 during 5–14 years, and increasing again to 2.04 (95% CI 1.79–2.32) after 15 years. Overall, individuals with any exposure had a 79%

higher risk of IHD compared with unexposed individuals (SIR 1.79, 95% CI 1.66–1.94). These findings provide a clear overview of both the magnitude and temporal dynamics of exposure-related risk, complementing the results from cross-tabulation and competing-risk analyses.

A Cox proportional hazards model was used to evaluate the association between diabetes and the risk of ischemic heart disease (IHD). Two models were fitted. Model 1 included baseline age and diabetes status as a time-dependent variable. Model 2 was additionally adjusted for body mass index (BMI) and educational level, both measured at study baseline. Baseline age was included in both models to account for small variations in participants’ age at cohort entry.

In Model 1, individuals with diabetes had a 57% higher risk of IHD compared with those without diabetes (HR 1.57, 95% CI 1.47–1.69, $p < 0.001$). After adjustment for BMI and education in Model 2, the association was attenuated but remained statistically significant, with diabetes associated with a 37% increased risk of IHD (HR 1.37, 95% CI 1.27–1.47, $p < 0.001$).

Educational level showed a strong inverse association with IHD risk. Compared with individuals with low education, those with medium education had a 12% lower risk (HR 0.88, 95% CI 0.84–0.93, $p < 0.001$), while those with high education had a 40% lower risk (HR 0.60, 95% CI 0.53–0.68, $p < 0.001$).

BMI was independently associated with IHD risk. Compared with participants with healthy weight, overweight was associated with a 24% higher risk (HR 1.24, 95% CI 1.17–1.31, $p < 0.001$) and obesity with a 55% higher risk (HR 1.55, 95% CI 1.45–1.67, $p < 0.001$), whereas underweight was not significantly associated with IHD risk (HR 0.93, 95% CI 0.48–1.79, $p = 0.83$).

Importantly, the full report also provides diagnostic plots, such as tests of proportional hazards and covariate contributions [29]. For Model 2, the diagnostics indicate that exposure, BMI, and other covariates have notable contributions to model fit, allowing researchers to evaluate model assumptions and performance directly within the automated workflow without requiring additional programming.

These results are consistent with the findings from cross-tabulation, competing-risk, and SIR analyses, further confirming that diabetes is associated with an elevated risk of IHD. By providing time-to-event estimates and adjustment for key covariates, the Cox models complement earlier analyses and illustrate how the automated workflow captures both

the magnitude and temporal dynamics of exposure–response associations.

Discussion

We developed an automated analytical pipeline designed to generate reproducible results for investigating relationships between exposure and outcome conditions. In this use case, the OMOP CDM would also have been a suitable option; however, its implementation would have required substantially greater resources and time for data transformation and infrastructure development. If an OMOP-based data structure had already been available, developing applications on top of it would have been a reasonable approach. In future work, we aim to facilitate interoperability by enabling streamlined data transfer from OMOP-based structures to the simplified model supported by our application.

While comprehensive CDM frameworks allow highly detailed and standardized representations of healthcare data, simpler data models can adequately address most research-oriented analytical needs, particularly in well-defined registry settings. Maintaining a clear understanding of data provenance, analytical assumptions, and workflow structure remains essential regardless of the underlying data model.

Generalizing a statistical model to new datasets often introduces errors, as these datasets typically contain variations in variable values or other characteristics [26]. Such discrepancies are an inherent challenge in modeling but can be managed through iterative development: when issues arise, the model can be refined and its functions adjusted to improve performance. This iterative process reflects the dynamic nature of data modeling, particularly when new analytical requirements emerge [30]. Revisiting the data model is necessary, while

maintaining simplicity as a core principle ensures that the application remains adaptable. Additional robust analyses can also be integrated as researchers use the system, and advanced functionalities are planned for future updates.

The dataset used in this project was transformed into a generalized format, enabling potential extension to nationwide registries. The application itself is scalable to broader infrastructure because it operates on data structures which are generated by Finnish healthcare registries. A current limitation is that the application was developed and tested with a dataset of 16,000 individuals, which may affect computational performance when applied to larger populations depending on available computing resources. However, moving analytical functions to database servers could substantially improve processing times [31,32]. Performance would also likely improve with more powerful computational infrastructure, as the application was evaluated on a relatively small server with four CPU cores and 8 GB of RAM.

The application's foundation on open-source software fosters continuous development and encourages the integration of new statistical methods and analytical procedures [33]. This approach aligns with the core principles of Open Science, which advocate for transparency and collaboration by sharing code and methodologies through platforms such as GitHub [34]. Looking ahead, the application will undergo iterative enhancements to incorporate additional functionalities and advanced analyses, and it will be validated across diverse datasets to ensure scalability and robustness.

The application has been evaluated in research settings, where medical students, many without prior programming experience, were able to generate statistical analyses by specifying exposure and outcome conditions through the graphical interface.

This demonstrates the accessibility of the framework and its potential to lower the technical barrier for conducting structured observational analyses. By requiring only substantive clinical interest rather than coding expertise, the system supports broader engagement in data-driven research. Although the underlying registry data cannot be made publicly available due to data protection regulations, the application is currently being used in ongoing research projects at the University of Eastern Finland, supporting its practical feasibility and operational stability.

This type of framework exemplifies a low-cost solution built on real-world data with open data science principles [35]. Instead of relying on external companies to develop complex platforms, meaningful results can be obtained directly from registry data in a straightforward manner. Additionally, the application saves time for statisticians by automating the initial steps of research, allowing them to focus on more advanced analyses.

Conclusion

Overall, the framework provides a rapid, low-cost, and reproducible approach for generating real-world evidence (RWE) to support research and healthcare decision-making. By automating key analytical steps, it enables researchers to identify associations efficiently and with minimal manual effort, thereby facilitating both scientific discovery and practical evaluation of healthcare outcomes. Because the application outputs only aggregated results, researchers can conduct analyses without accessing person-level data, supporting secure and privacy-preserving use of sensitive health information. The framework can also be scaled to larger data environments, such as regional hospital data pools, to generate more detailed insights and expand analytical capacity. As a proof-of-concept, this

work demonstrates a cost-efficient pathway for RWE generation and lays the groundwork for developing more advanced analytical systems in the future.

The authors used ChatGPT (GPT-5, OpenAI, 2025) to assist in improving English grammar and

sentence clarity. The authors reviewed and edited all content generated by the model and take full responsibility for the final text.

Conflict of interest statement

Authors declare no conflicts of interest.

References

[1] Gregg EW, Patorno E, Karter AJ, Mehta R, Huang ES, White M, Patel CJ, McElvaine AT, Cefalu WT, Selby J, Riddle MC, Khunti K. Use of Real-World Data in Population Science to Improve the Prevention and Care of Diabetes-Related Outcomes. *Diabetes Care*. 2023 Jul 1;46(7):1316-1326. <https://doi.org/10.2337/dc22-1438>

[2] Dang A. Real-World Evidence: A Primer. *Pharmaceut Med*. 2023 Jan;37(1):25-36. <https://doi.org/10.1007/s40290-022-00456-6>

[3] Zou KH, Li JZ, Imperato J, Potkar CN, Sethi N, Edwards J, Ray A. Harnessing Real-World Data for Regulatory Use and Applying Innovative Applications. *J Multidiscip Healthc*. 2020 Jul 22;13:671-679. <https://doi.org/10.2147/JMDH.S262776>

[4] Zou KH, Berger ML. Real-World Data and Real-World Evidence in Healthcare in the United States and Europe Union. *Bioengineering*. 2024 Aug 2;11(8):784. <https://doi.org/10.3390/bioengineering11080784>

[5] Simsion G, Milton SK, Shanks G. Data modeling: Description or design? *Inf Manage*. 2012 May 1;49(3):151–63. <https://doi.org/10.1016/j.im.2012.01.003>

[6] Finster M, Wenzel M, Taghizadeh E. Common data models and data standards for tabular health data: a systematic review. *BMC Med Inform Decis*

Mak. 2025 Nov 13;25(1):422. <https://doi.org/10.1186/s12911-025-03267-2>

[7] Overhage JM, Ryan PB, Reich CG, Hartzema AG, Stang PE. Validation of a common data model for active safety surveillance research. *J Am Med Inform Assoc*. 2012 Jan-Feb;19(1):54-60. <https://doi.org/10.1136/amiajnl-2011-000376>

[8] Raventós B, Català M, Du M, Guo Y, Black A, Inberg G, Li X, López-Güell K, Newby D, de Ridder M, Barboza C, Duarte-Salles T, Verhamme K, Rijnbeek P, Prieto Alhambra D, Burn E. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. *Pharmacoepidemiol Drug Saf*. 2024 Jan;33(1):e5717. <https://doi.org/10.1002/pds.5717>

[9] Glicksberg BS, Oskotsky B, Thangaraj PM, Giangreco N, Badgeley MA, Johnson KW, Datta D, Rudrapatna VA, Rappoport N, Shervey MM, Miotto R, Goldstein TC, Rutenberg E, Frazier R, Lee N, Israni S, Larsen R, Percha B, Li L, Dudley JT, Tatonetti NP, Butte AJ. PatientExploreR: an extensible application for dynamic visualization of patient clinical history from electronic health records in the OMOP common data model. *Bioinformatics*. 2019 Nov 1;35(21):4515-4518. <https://doi.org/10.1093/bioinformatics/btz409>

[10] Zhou S, Brunke L, Tao A, Hall AW, Bejarano FP, Panerati J, Schoellig AP. What Is the Impact of Releasing Code With Publications? *Statistics from the*

- Machine Learning, Robotics, and Control Communities. *IEEE Control Systems*. 2024 Aug;44(4):38-46. <https://doi.org/10.1109/MCS.2024.3402888>
- [11] Chang W, Cheng J, Allaire JJ, Sievert C, Schloerke B, Aden-Buie G, et al. shiny: web application framework for R [Internet]. 2025 [cited 2025 Dec 18]. Available from: <https://CRAN.R-project.org/package=shiny>. <https://doi.org/10.32614/CRAN.package.shiny>
- [12] Cruz T, Jimenez FG, Bravo ARQ, Ander E. DataXploreFines: generalized data for informed decision, making, an interactive Shiny Application for data analysis and visualization [preprint]. arXiv:2307.11056. doi:10.48550/arXiv.2307.11056
- [13] Sessler T, Quinn GP, Wappett M, Rogan E, Sharkey D, Ahmaderaghi B, Lawler M, Longley DB, McDade SS. surviveR: a flexible shiny application for patient survival analysis. *Sci Rep*. 2023 Dec 13;13(1):22093. <https://doi.org/10.1038/s41598-023-48894-9>
- [14] Miller DM, Shalhout SZ. StoryboardR: an R package and Shiny application designed to visualize real-world data from clinical patient registries. *JAMIA Open*. 2023 Jan 6;6(1):ooac109. <https://doi.org/10.1093/jamiaopen/ooac109>
- [15] Justo N, Espinoza MA, Ratto B, Nicholson M, Rosselli D, Ovcinnikova O, García Martí S, Ferraz MB, Langsam M, Drummond MF. Real-World Evidence in Healthcare Decision Making: Global Trends and Case Studies From Latin America. *Value Health*. 2019 Jun;22(6):739-749. <https://doi.org/10.1016/j.jval.2019.01.014>
- [16] Jansen MS, Dekkers OM, le Cessie S, Hooft L, Gardarsdottir H, de Boer A, Groenwold RHH. Real-World Evidence to Inform Regulatory Decision Making: A Scoping Review. *Clin Pharmacol Ther*. 2024 Jun;115(6):1269-1276. <https://doi.org/10.1002/cpt.3218>
- [17] Liu F, Panagiotakos D. Real-world data: a brief review of the methods, applications, challenges and opportunities. *BMC Med Res Methodol*. 2022 Nov 5;22(1):287. <https://doi.org/10.1186/s12874-022-01768-6>
- [18] Sund R. Utilisation of administrative registers using scientific knowledge discovery. *Intell Data Anal*. 2003 Nov 1;7(6):501-19. <https://doi.org/10.3233/IDA-2003-7602>
- [19] Wickham H, Cetinkaya-Rundel M, Grolemund G. R for data science. 2nd edition. [Internet]. Authors; 2023 [cited 2025 Nov 19]. Available from: <https://r4ds.hadley.nz/>
- [20] Moilanen A, Kopra J, Kröger H, Sund R, Rikonen T, Sirola J. Characteristics of Long-Term Femoral Neck Bone Loss in Postmenopausal Women: A 25-Year Follow-Up. *J Bone Miner Res*. 2022 Feb;37(2):173-178. <https://doi.org/10.1002/jbmr.4444>
- [21] Tuppurainen M, Honkanen R, Kröger H, Saarikoski S, Alhava E. Osteoporosis risk factors, gynaecological history and fractures in perimenopausal women--the results of the baseline postal enquiry of the Kuopio Osteoporosis Risk Factor and Prevention Study. *Maturitas*. 1993 Sep;17(2):89-100. [https://doi.org/10.1016/0378-5122\(93\)90004-2](https://doi.org/10.1016/0378-5122(93)90004-2)
- [22] Heikkinen J, Honkanen RJ, Quirk SE, Williams LJ, Koivumaa-Honkanen H. Long-term life satisfaction in ageing women with work disability due to mental and musculoskeletal disorders. *Maturitas*. 2023 Dec;178:107849. <https://doi.org/10.1016/j.maturitas.2023.107849>
- [23] Nissinen T, Sund R, Suoranta S, Kröger H, Väänänen SP. Combining Register and Radiological Visits Data Allows to Reliably Identify Incident Wrist Fractures. *Clin Epidemiol*. 2023 Sep 20;15:1001-1008. <https://doi.org/10.2147/CLEP.S421013>

- [24] Miettinen J, Sund R. HealthpopR R package [Internet]. 2025 [cited 2025 Nov 19]. Available from: <https://janikmiet.github.io/healthpopR/>
- [25] Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020 Jun;582(7810):84-88. <https://doi.org/10.1038/s41586-020-2314-9>
- [26] Nahhas RW. Introduction to regression methods for public health using R. 1st ed. Chapman & Hall; 2024. 456 p. <https://doi.org/10.1201/9781003263197>
- [27] Kannel WB, McGee DL. Diabetes and cardiovascular disease. The Framingham study. *JAMA*. 1979 May 11;241(19):2035-8. <https://doi.org/10.1001/jama.241.19.2035>
- [28] Haffner SM, Lehto S, Rönkä T, Pyörälä K, Laakso M. Mortality from coronary heart disease in subjects with type 2 diabetes and in nondiabetic subjects with and without prior myocardial infarction. *N Engl J Med*. 1998 Jul 23;339(4):229-34. <https://doi.org/10.1056/NEJM199807233390404>
- [29] Miettinen J. RWE example report [Internet]. 2025 Dec [cited 2025 Dec 19]. Available from: <https://janikmiet.github.io/shinyERA-example-report/>
- [30] Pai AU. Agile Data Science: How scrum masters can drive data-driven projects. *Eur J Comput Sci Inf Technol*. 2025;13(44):58-67. <https://doi.org/10.37745/ejcsit.2013/vol13n445867>
- [31] Dijkman R, Gao J, Syamsiyah A, van Dongen B, Grefen P, ter Hofstede A. Enabling efficient process mining on large data sets: realizing an in-database process mining operator. *Distrib Parallel Databases*. 2020 Mar 1;38(1):227-53. <https://doi.org/10.1007/s10619-019-07270-1>
- [32] Yu X, Youill M, Woicik M, Ghanem A, Serafini M, Aboulnaga A, Stonebraker M. PushdownDB: accelerating a DBMS using S3 computation. In: *IEEE 36th International Conference on Data Engineering (ICDE)*. 2020 Apr; Dallas, TX, USA. p. 1802-1805. <https://doi.org/10.1109/ICDE48307.2020.00174>
- [33] Daswito R, Besral B, Ilmaskal R. Analysis using R software: a big opportunity for epidemiology and public health data analysis. *J Health Sci Epidemiol*. 2023 Apr 29;1(1):1-5. <https://doi.org/10.62404/jhse.v1i1.9>
- [34] Fortunato L, Galassi M. The case for free and open source software in research and scholarship. *Philos Trans A Math Phys Eng Sci*. 2021 May 17;379(2197):20200079. <https://doi.org/10.1098/rsta.2020.0079>
- [35] Bowrin K, Briere JB, Levy P, Millier A, Clay E, Toumi M. Cost-effectiveness analyses using real-world data: an overview of the literature. *J Med Econ*. 2019 Jun;22(6):545-553. <https://doi.org/10.1080/13696998.2019.1588737>