Received: 30 May 2023 | Revised: 27 September 2023 | Accepted: 2 October 2023
DOI: 10.1002/pds.5717

ORIGINAL ARTICLE

WILEY

IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model

Berta Raventós[1,2] | Martí Català[3] | Mike Du[3] | Yuchen Guo[3] | Adam Black[4] | Ger Inberg[5] | Xintong Li[3] | Kim López-Güell[3] | Danielle Newby[3] | Maria de Ridder[5] | Cesar Barboza[5] | Talita Duarte-Salles[1,5] | Katia Verhamme[5] | Peter Rijnbeek[5] | Daniel Prieto Alhambra[3,5] | Edward Burn[3]

[1]Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (IDIAPJGol), Barcelona, Spain

[2]Universitat Autònoma de Barcelona, Bellaterra (Cerdanyola del Vallès), Barcelona, Spain

[3]Centre for Statistics in Medicine (CSM), Nuffield Department of Orthopaedics, Rheumatology and Musculoskeletal Sciences (NDROMS), University of Oxford, Oxford, UK

[4]Odysseus Data Services, Cambridge, Massachusetts, USA

[5]Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Correspondence
Edward Burn, Botnar Research Centre,
Windmill Road, OX37LD, Oxford, UK.
Email: edward.burn@ndorms.ox.ac.uk

Funding information
European Medicines Agency

Abstract

Purpose: Real-world data (RWD) offers a valuable resource for generating population-level disease epidemiology metrics. We aimed to develop a well-tested and user-friendly R package to compute incidence rates and prevalence in data mapped to the observational medical outcomes partnership (OMOP) common data model (CDM).

Materials and Methods: We created IncidencePrevalence, an R package to support the analysis of population-level incidence rates and point- and period-prevalence in OMOP-formatted data. On top of unit testing, we assessed the face validity of the package. To do so, we calculated incidence rates of COVID-19 using RWD from Spain (SIDIAP) and the United Kingdom (CPRD Aurum), and replicated two previously published studies using data from the Netherlands (IPCI) and the United Kingdom (CPRD Gold). We compared the obtained results to those previously published, and measured execution times by running a benchmark analysis across databases.

Results: IncidencePrevalence achieved high agreement to previously published data in CPRD Gold and IPCI, and showed good performance across databases. For COVID-19, incidence calculated by the package was similar to public data after the first-wave of the pandemic.

Conclusion: For data mapped to the OMOP CDM, the IncidencePrevalence R package can support descriptive epidemiological research. It enables reliable estimation of incidence and prevalence from large real-world data sets. It represents a simple, but

Berta Raventós and Martí Català should be considered as joint first-authors

This work has been presented as a Software Demonstration in the Observational Health Data Sciences and Informatics (OHDSI) Symposium held in Bethesda, USA, on 14–16 October 2022.

extendable, analytical framework to generate estimates in a reproducible and timely manner.

KEYWORDS

common data model, incidence, OMOP, prevalence, R package

Key Points

- The 'IncidencePrevalence' R package enables reproducible and rapid analyses of incidence rates and prevalence of health events in data sets mapped to the OMOP CDM.
- It is flexible enough to accommodate specific research needs, such as different age and sex stratifications and prior observation time requirements, among others.
- We demonstrated its validity by replicating two earlier studies, and we illustrated its rapid execution time by conducting benchmarking analyses in four large-scale databases from The Netherlands, Spain and the United Kingdom.
- The package is suitable for large-scale network studies and it can be used for any health event registered in the CDM, providing a useful tool for those generating evidence from data mapped to the OMOP CDM.

Plain Language Summary

Routinely-collected health care data are a valuable resource to measure the occurrence of health events in a population. They can provide insights on the rate of new cases over a period of time (incidence), and the proportion of new and pre-existing cases occurring at a particular time point (point prevalence) or over a period of time (period prevalence). However, the statistical analysis required to compute these metrics is not straightforward and requires programming knowledge. In order to reduce its complexity, we developed an R package named 'IncidencePrevalence' to compute incidence and prevalence rates in data mapped to the observational medical outcomes partnership (OMOP) common data model (CDM). To validate the package, we calculated incidence rates of COVID-19 using data from Spain and the United Kingdom, and replicated two previously published studies using data from the Netherlands and the United Kingdom. We then compared the obtained results to those previously published and we assessed the time required to do a given analysis across different databases. The results obtained were very similar to those previously published and required short execution times. The package developed will be useful for researchers generating evidence on the incidence and prevalence of health-related events using OMOP-formatted data.

1 | BACKGROUND

Population-level estimates of incidence rates and prevalence provide important information on the occurrence of disease and other health-related events, such as medication use or measurements. Routinely-collected health care data offers a valuable resource for informing such estimates, with capture of both existing or new occurrences of the outcomes of interest and the number of individuals at risk of the outcome at a given time point. Calculating incidence rates and prevalence does involve several design choices, such as decisions about observable person-time (e.g., the amount of lookback time used to search for a prevalent disease-related medical encounter). These decisions can have a sizeable impact on results, and careful consideration is required as to what is most appropriate for the research question at hand.[1,2]

The observational medical outcomes partnership (OMOP) Common Data Model (CDM) provides a standardised structure and vocabulary for health data.[3] The adoption of the OMOP CDM allows for the same analytic code to be run against multiple data sources. Network studies are run in a distributed manner with common analytic code run by each site and aggregated results returned, without the need to share patient-level data. The development of robust, and open-source R packages can help facilitate such research, reduce the time required to perform it, and increase the trust in the results generated.

We aimed to develop an R package, named IncidencePrevalence, to support the analysis of incidence and prevalence in data sets mapped to the OMOP CDM in a standardised, reproducible and timely manner. This package can be used to calculate rates of any event registered in the CDM, such as medical conditions, medication use or clinical measurements. Furthermore, it is flexible enough to accommodate specific research needs, such as different age and sex stratifications and prior observation time requirements, among others. In this paper, we demonstrate the functionality of this R package by comparing its results to previously published data using four real world healthcare databases from the Netherlands, Spain and the United Kingdom. In addition, we demonstrate its performance by measuring the time taken to compute a benchmark analysis using the same databases.

2 | OBJECTIVES

The goal of the IncidencePrevalence package is to provide a flexible approach to calculate incidence rates and prevalence (point and period) in databases mapped to the OMOP CDM. In this way, we aim to complement the growing body of resources and tools provided by the Observational Health Data Sciences and Informatics group (OHDSI; https://www.ohdsi.org/) and to facilitate the development of studies exploring the incidence and prevalence of health-related events across domains. In addition, we have assessed the face validity of the package and performed a benchmark analysis to measure its execution time on four real world healthcare databases.

3 | METHODS

3.1 | Package

3.1.1 | Package details and software dependencies

Version 0.5.1 of IncidencePrevalence is written in R (version 4.2.1), organised using roxygen2,[4] and depends on the following existing packages: checkmate,[5] cli,[6] dbplyr,[7] dplyr,[8] lubridate,[9] glue,[10] magrittr,[11] rlang,[12] purrr,[13] tidyr,[14] tidyselect,[15] stringr[16] and zip.[17] IncidencePrevalence is designed to be used against data mapped to the OMOP CDM, and therefore, data sets must be first converted to the CDM prior to using the package presented in this paper.[18–20] IncidencePrevalence can connect to several database management systems through the DataBase Interface (DBI) R package.[21] To allow for a pipe friendly syntax, CDM table references are stored in a single object (referred as cdm object) through the CDMConnector R package.[22] This object is created from the DBI database connection and includes a list of references to the tables in the OMOP CDM. Full details about the CDM object and how it interacts with the CDM can be found in the CDMConnector package documentation.[22]

This R package was built using a test-driven development approach. Unit tests were developed on mock OMOP CDM data populated with test-specific synthetic data so as to test that the R package had properly encoded the requirements of the software.[23] These unit tests included checks on the output format of results, and numerous logical checks of results. Tests were also implemented to make sure that edge cases were well-handled. For example, we developed tests to confirm multiple outcomes occurring in the outcome washout periods (see 'Main functions' below) were handled accordingly. Expected errors were also checked, and in such circumstances informative error messages are returned to the user. Once tests were passing, code was then refactored to optimise speed and memory usage.

TABLE 1 Main functions of the IncidencePrevalence R package.

Functions	Description
'generateDenominatorCohortSet()'	Identify a set of denominator populations based on specific criteria.
'estimateIncidence()'	Estimation of incidence rates in a given set of denominator populations.
'estimatePointPrevalence()'	Estimation of point prevalence in a given set of denominator populations.
'estimatePeriodPrevalence()'	Estimation of period prevalence in a given set of denominator populations.

3.1.2 | Main functions

IncidencePrevalence includes different functions to identify study populations of interest (referred as denominator populations in this paper) and to estimate incidence, point- and period-prevalence of an outcome of interest in these populations. The user interface was collaboratively designed with epidemiologists to ensure that function and parameter names were intuitive. For instance, 'outcomeWashout' was the term used to label the parameter that determined the amount of time required to exclude participants with a previous history of the health event under investigation, as established by prior epidemiological studies.[2] Here, we provide a brief explanation of its main functions (Table 1) and design choices available to the user (Figure 1).

The *'generateDenominatorCohortSet()'* function can be used to identify one or more denominator populations. Each denominator population is defined according to a study start and end date (in their absence the earliest observation period start date and latest observation period end date in the database will be used). Individuals contribute time to denominator populations only if they are observed in the database during the study period and meet the stratification criteria. These criteria can be defined using different input parameters, including age, sex and the amount of prior observation required to enter the study (Figure 1A). The *'generateDenominatorCohortSet()'* function identifies the denominator population for each of the combinations of the above-mentioned parameters, and assigns a unique cohort identifier to each population. The contribution of individuals to the denominator population is limited to the specific study period during which they meet all the specified stratification criteria (such as age and prior history requirements). Consequently, the entry date for individuals is defined as the date when the final inclusion criterion is met, and their contribution of time ends as soon as any of the requirements are no longer fulfilled. Individuals can appear multiple times in the

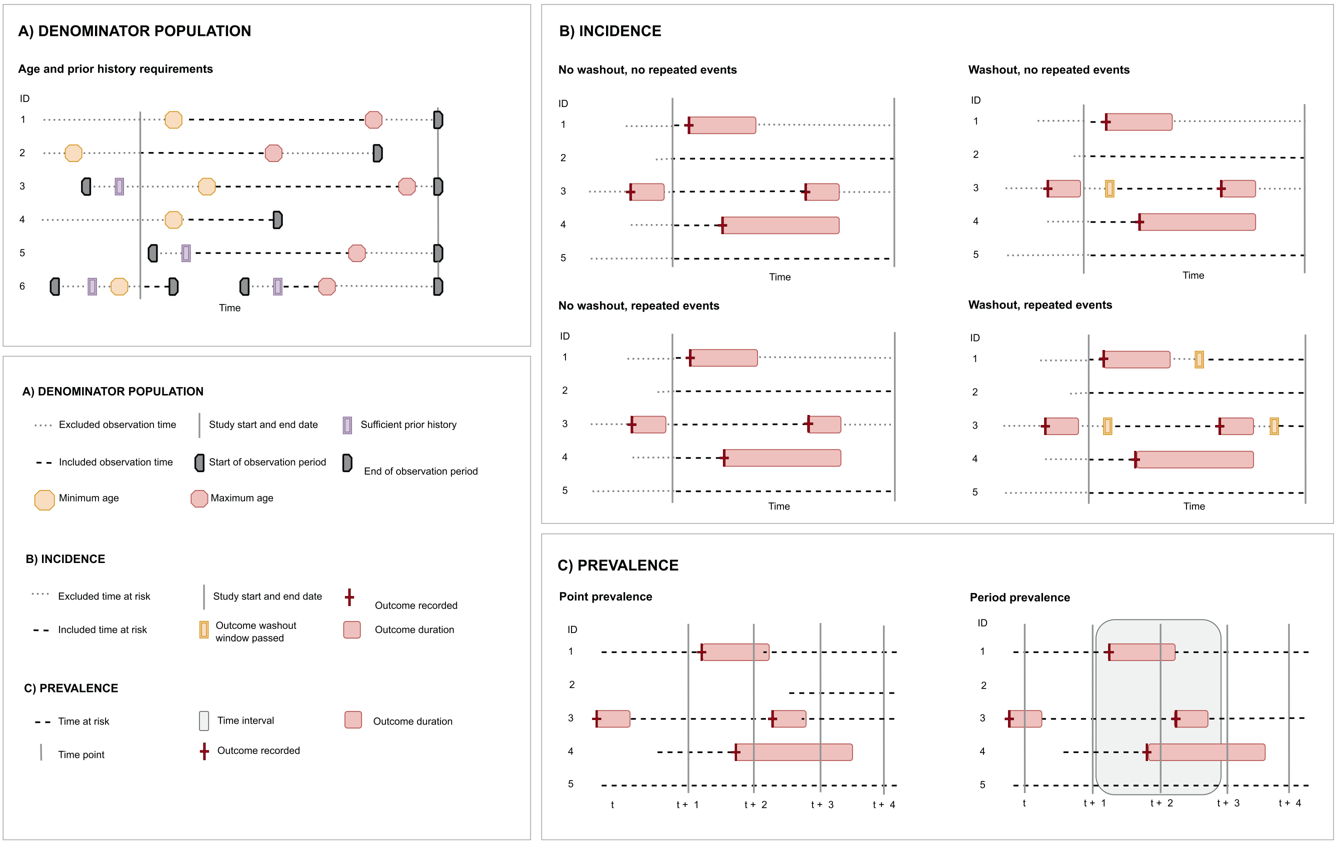

FIGURE 1 Summary of the main parameters of the IncidencePrevalence R package.

denominator cohort as part of different populations, such as when multiple age criteria are applied, or as part of the same population if the database includes multiple observation periods per person. In addition, the denominator population can also be restricted to individuals within a particular subgroup of the total database population (i.e., an existing OMOP CDM cohort that defines individuals with either time-invariant or time-varying characteristics of interest). The output generated by this function includes a table with the cohort of identified individuals structured as an OMOP CDM cohort table, as well as an attribute containing the patient attrition (i.e., the number of individuals who are excluded and the reason behind its exclusion).

Further functions are available to calculate incidence and prevalence once one or multiple denominator populations and outcome cohorts of interest (see 'Outcome definition' below) have been identified.

Incidence rates can be calculated using the 'estimateIncidence()' function. Users can decide whether to restrict their analysis to the first event per person or to include all events captured during their time contributed to the denominator population, allowing for more than one event per person (repeated events). In addition, an outcome washout period can be specified so as to define a minimum amount of time following an event during which individuals do not contribute time to the study. Incidence rates are calculated by dividing the number of events by the person-time. Person-time is defined as the sum of the participants' time contributed to the study. In cases where repeated events are not permitted, individuals are censored at the first occurrence of the outcome. If repeated events are allowed, patients' time contribution is not censored at the occurrence of the outcome but can be temporarily paused according to the defined outcome washout period.

The 'estimatePointPrevalence()' and 'estimatePeriodPrevalence()' functions can be used to calculate the proportion of people included in the denominator population who are in the outcome cohort at a specified time point (point prevalence) or within a given time interval (period prevalence), respectively. Point prevalence can be calculated at a specific time point (start, middle or end) within each calendar time interval. In addition, when computing period prevalence, the user can choose whether participants are required to be observed throughout the entire calendar time interval being under study or if they are only required to present for 1 day of the interval in order to contribute to the prevalence calculation of that period. Whether to restrict participants to those fully observed throughout the calendar time interval in the period prevalence calculation can substantially affect the results obtained, particularly when the outcome might affect the follow-up period or when populations are at risk of reaching the end of their follow-up.

The estimates obtained in the 'estimateIncidence()', 'estimatePointPrevalence()' and 'estimatePeriodPrevalence()' functions are computed for each time interval and each unique combination of input parameters. The calendar time intervals allowed in this package

TABLE 2 Example usage of the IncidencePrevalence R package.

Example	R code
Identify a set of denominator populations with individuals aged between 0 and 64 and 65 to 100 from 2008 to 2018. For each age group, identify three different cohorts with males, females, and both sexes. Require 365 days of prior observation to be included.	cdm < − generateDenominatorCohortSet(cdm = cdm, name=‘denominator’, cohortDateRange = c(as.Date(‘2008-01-01’), as.Date(‘2018-01-01’))), ageGroup = list(c(0, 64), c(65, 100)), sex = c(‘Male’, ‘Female’, ‘Both’), daysPriorObservation = 180)

Estimate yearly incidence rates allowing repeated events during the study period. Restrict the analysis to individuals who had not experienced the outcome in the previous 180 days, and only compute rates in complete years observed in the database.	inc < − estimateIncidence(
	cdm = cdm,
	denominatorTable = ‘denominator’,
	outcomeTable = ‘outcome’,
	interval = ‘years’,
	repeatedEvents = TRUE,
	outcomeWashout = 180,
	completeDatabaseIntervals = TRUE
	)

Estimate point prevalence at the beginning of each calendar year.	prev_point <−
	estimatePointPrevalence(
	cdm = cdm,
	denominatorTable = ‘denominator’,
	outcomeTable = ‘outcome’,
	interval = ‘years’,
	timePoint = ‘start’
	)

Estimate yearly period prevalence. Restrict the analysis of each time interval to individuals who had been observed throughout the interval, and only compute rates in complete years observed in the database.	prev_period <−
	estimatePeriodPrevalence(
	cdm = cdm,
	denominatorTable = ‘denominator’,
	outcomeTable = ‘outcome’,
	interval = ‘years’,
	completeDatabaseIntervals = TRUE,
	fullContribution = TRUE
	)

include weeks, months, quarters, years and, for incidence, overall (i.e. the complete study period specified in *generateDenominatorCohortSet)*. To satisfy data protection regulations, a minimum number of events for time intervals can be required in order to report results. In addition, the user can decide whether to include calendar time intervals that are not fully captured in the database. By default, confidence intervals are computed using the exact and Wilson Score methods for incidence and prevalence rates, respectively.[24,25] An example usage of the R package with the annotated code can be found in Table 2. The output of these three functions includes a main table with results. Incidence rates are reported as the number of events per 100 000 person-years and prevalence as expressed as a proportion of population. Given that the population included in the analysis of incidence rates and prevalence might differ from the one obtained using ‘*generateDenominatorCohortSet()*’, we have added an additional attrition attribute to report the number of participants who were included in the analysis. Individuals that contributed to an analysis can also be identified to further characterise the study population before sharing aggregated results. Lastly, we have incorporated specific functions to easily plot incidence and prevalence results.

3.1.3 | Outcome definition

Outcome cohorts need to be defined and instantiated prior to using this package. The table instantiated to the CDM object must include information on the subject and a cohort identifier as well as the start and end date in which individuals qualify for the cohort. For instance, in the case of a cohort comprising individuals with a chronic condition, the outcome table would consist of a row per diagnosed patient, where the start date corresponds to the date of diagnosis and the end date represents the end of follow up for each patient. However, for cohorts involving recurrent events such as non-chronic conditions or medication use, individuals can belong to the same cohort for multiple time periods (one row per observed episode and patient). How to define cohorts according to the OMOP format is elaborated elsewhere and it is outside the scope of this package.[26] Note that cohort definitions can have a considerable impact on the subsequent results. For instance, cohort start and end date determine when the incident event takes place and for how long it is considered prevalent.

3.2 | Package validation and benchmarking

We performed three validation studies by replicating figures from publicly available sources on COVID-19 incidence,[27,28] and two multinational network studies, one on incidence of Adverse events of special interest (AESI) of COVID-19 vaccines,[29] and the other on prevalence of use of ranitidine and other Histamine$_2$-Receptor Antagonists.[30]

We used data sources that were previously used in those studies, including the Clinical Practice Datalink (CPRD) Aurum and Gold from the UK,[31,32] the Integrated Primary Care Information (IPCI) from the Netherlands,[33] and the Information System for Research in Primary Care (SIDIAP) from Catalonia, Spain.[34] All of these databases contain routinely collected data from primary care, and were previously mapped to the OMOP CDM. Of those, SIDIAP is the only database that is linked at an individual level to hospital data (Conjunt Mínim Bàsic de Dades d'Alta Hospitalària, CMBD-AH).[35]

We computed the overall monthly incidence rates of COVID-19 (defined as COVID-19 diagnoses or positive test results for SARS-CoV-2) in CPRD Aurum and SIDIAP. As a comparison, we used incidence rates based on aggregated counts of COVID-19 cases (defined as positive test results for SARS-CoV-2) from official UK and Catalan government websites for data on COVID-19.[27,28] The denominators to compute incidence rates were obtained from the UK Office for National Statistics and the Statistical Institute of Catalonia (Institut d'Estadística de Catalunya).[36,37] Data from the UK was restricted to England.

We computed incidence rates of AESIs in CPRD GOLD, and we restricted the analysis to events having more than five occurrences. Lastly, we computed yearly period prevalence rates of use of ranitidine in IPCI. For both analyses, we used the same age stratifications and prior observation requirements as the ones reported in the original studies.[29,30] We used the same study periods than those of the original study to compute incidence rates of AESIs in CPRD GOLD. For the prevalence of use of ranitidine in IPCI, we calculated yearly period prevalence rates using data from 2006 to 2018.[30,33] We

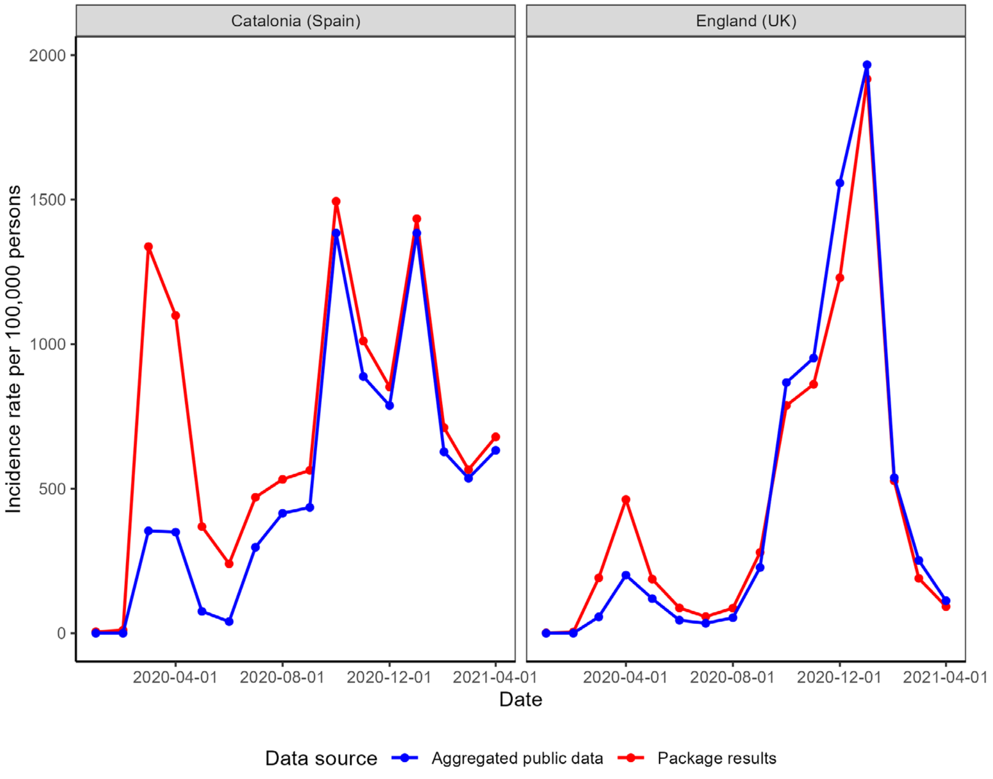

FIGURE 2 Comparison of incidence rates based on aggregated public data and incidence rates computed with the IncidencePrevalence R package in Spain's Information System for Research in Primary Care database (SIDIAP) and UK's Clinical Practice Research Datalink (CPRD) Aurum database. For the package results, we used data from SIDIAP and CPRD Aurum. Data from CPRD Aurum was restricted to England. To execute the package, the denominator population spanned from January 2020 to April 2021. Individuals were required to have 180 days of prior observation, and we only allowed the first event per person.

visually inspected differences between package results and those reported in the original published data. We measured running times of the package by performing a benchmark analysis in CPRD Aurum, CPRD GOLD, SIDIAP and ICPI. To do so, we created the *benchmarkIncidencePrevalence()* function to measure the running times for a given set of design choices. In this case, we measured the execution time required to analyze incidence rates and prevalence of a previously simulated outcome cohort with 10% prevalence using a denominator population with four age groups and two sex stratifications from 1 January 2014 to 31 December 2019.

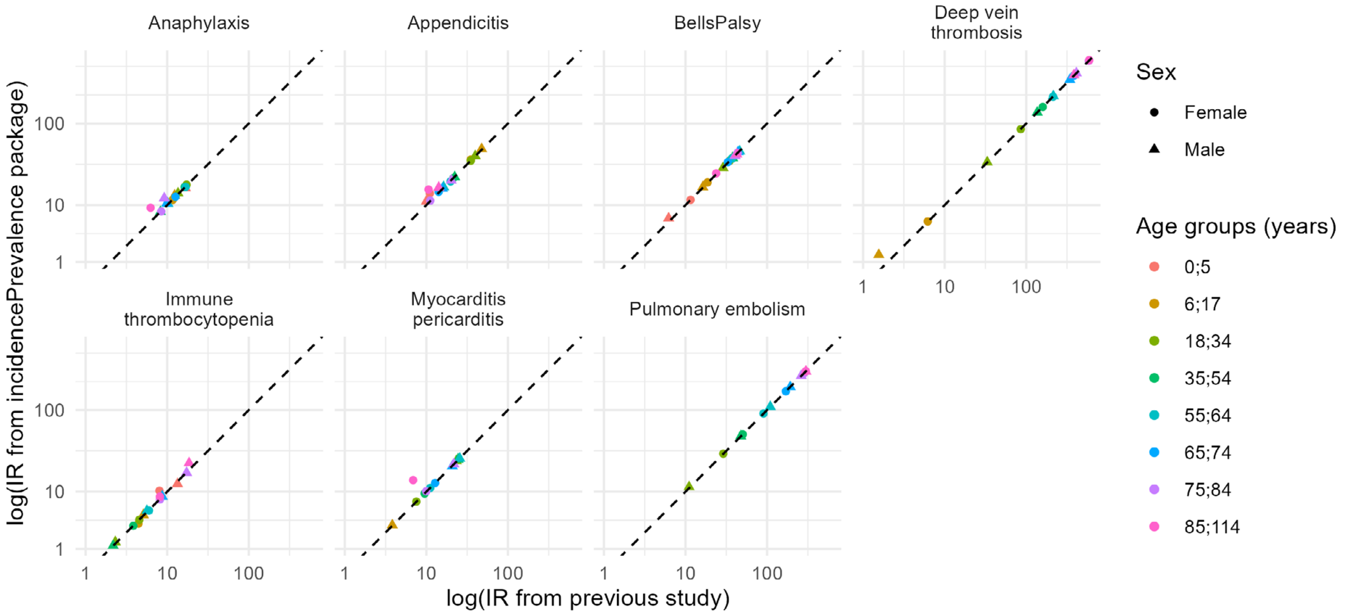

FIGURE 3 Age and sex-stratified incidence rates for adverse events of special interest in UK's Clinical Practice Research Datalink GOLD database. The age-sex incidence rates were estimated among people observed for at least 365 days before 1 January 2017, and the study period was 2017 to 2019. Events were identified using the same cohort definitions as the previous published study, with event-specific washout periods were applied accordingly. IR: incidence rates.

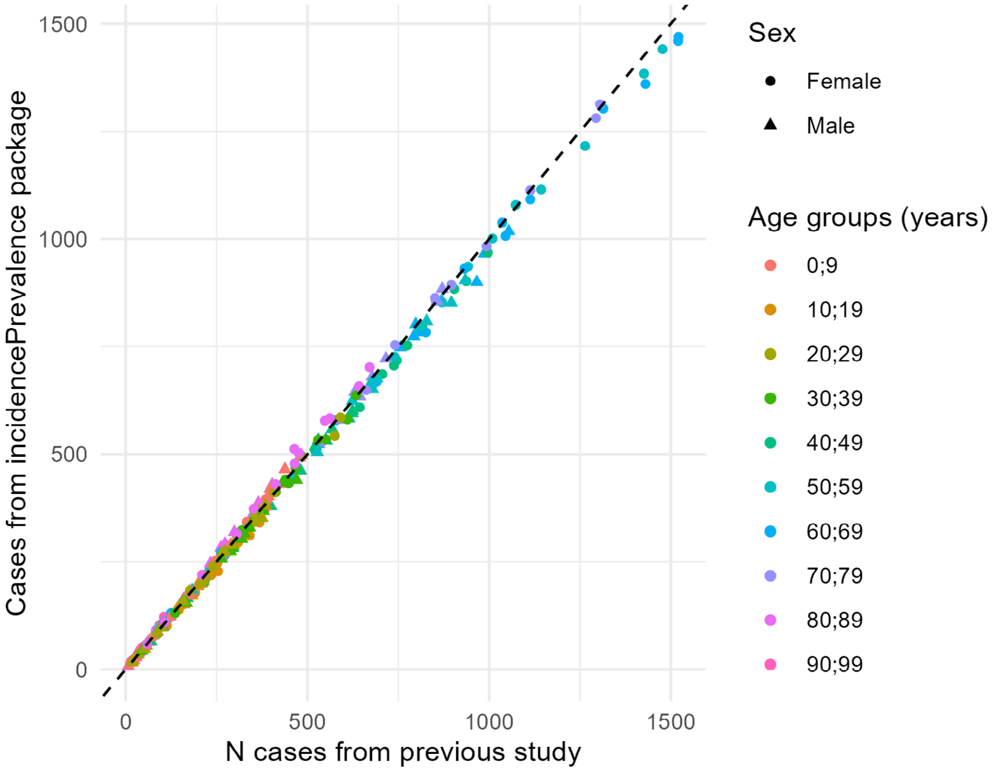

FIGURE 4 Age and sex stratified prevalence rates of use of ranitidine in The Netherlands' Integrated Primary Care Information (IPCI) database. The denominator population spanned from 2006 to 2019, and was stratified by 10-year age groups and sex. Individuals were not required to have prior data visibility. Yearly period prevalence was computed each calendar year. PR: prevalence rates.

TABLE 3 Benchmarking results.

	CPRD AURUM (*n* = 39 999 011)	CPRD GOLD (*n* = 15 662 217)	SIDIAP (*n* = 8 265 343)	IPCI (*n* = 2 612 850)
Task (minutes taken, %)[a]				
Generating denominator (8 cohorts)	19.66 (20.5)	8.37 (21.2)	3.2 (8.0)	0.77 (8.6)
Yearly point prevalence, 1 outcome(s)	11.7 (12.2)	5.23 (13.2)	4.89 (12.3)	0.9 (10.1)
Monthly point prevalence, 1 outcome(s)	16.08 (16.7)	6.46 (16.4)	7.64 (19.2)	1.74 (19.5)
Yearly period prevalence, 1 outcome(s)	11.65 (12.1)	4.98 (12.6)	5.21 (13.1)	0.94 (10.5)
Monthly period prevalence, 1 outcome (s)	16.58 (17.3)	6.44 (16.3)	8.34 (21)	1.99 (22.3)
Yearly incidence, 1 outcome (s)	7.51 (7.8)	3.09 (7.8)	3.69 (9.3)	0.75 (8.4)
Monthly incidence, 1 outcome (s)	12.83 (13.4)	4.94 (12.5)	6.85 (17.2)	1.84 (20.6)
Total	96.01 (100)	39.51 (100)	39.82 (100)	8.93 (100)

Abbreviations: CPRD, clinical practice research datalink; IPCI: integrated primary care information; SIDIAP, information system for research in primary care.

[a]Note that execution times not only rely on package performance, but also on the server and the environment in which analyses are performed.

4 | RESULTS

4.1 | Package

IncidencePrevalence is freely available under the Apache License (Version 2.0) and can be obtained from CRAN (https://cran.r-project.org/web/packages/IncidencePrevalence/index.html), where full details and instructions for setup and use are provided.

4.2 | Package validation and benchmarking

The source code underwent rigorous quality checks through the execution of more than 900 tests, obtaining over 99% test coverage.

We plotted incidence rates of COVID-19 over time using aggregated public data and data from CPRD Aurum and SIDIAP using the IncidencePrevalence package (Figure 2). We observed substantial discrepancies between rates during the early months of the pandemic in both CPRD Aurum and SIDIAP, obtaining higher rates when using the IncidencePrevalence package. These discrepancies were more pronounced in SIDIAP and differed from public data up to September 2020. In CPRD Aurum, rates largely coincided after June 2020 (Figure 2).

We visually inspected the correlation between package results and prior studies for incidence rates of AESIs (Figure 3) and prevalence rates of ranitidine use (Figure 4). Both figures showed strong, positive correlation, and were indicative of high agreement.

Running times of the benchmark analysis are reported in Table 3, overall and by task (e.g., generating denominator, yearly point prevalence). In total, running times were 96.01 min in CPRD Aurum (*n* = 39 999 011), 39.51 min in CPRD Gold (*n* = 15 662 217), 39.42 min in SIDIAP (*n* = 8 265 343), and 8.93 min in IPCI (n = 2 612 850). The median running time was 39.67 min.

5 | DISCUSSION

Reliable information about disease frequency is crucial for researchers, clinicians and policymakers. Here, we introduced IncidencePrevalence, an open-source R package to calculate incidence and prevalence rates of health events recorded in data sets mapped to the OMOP CDM. Moreover, we illustrated its use and we compared the results from this package to previously published data on incidence of COVID-19 and potential AESIs for COVID-19 vaccines, and prevalence of use of ranitidine. Lastly, we performed a benchmarking analysis which demonstrated its high performance across databases.

IncidencePrevalence enables standardised and reproducible analyses of large-scale data sets, while providing flexibility to support different design choices. Compared to writing custom code on a study-by-study basis, our package makes the analysis of incidence and prevalence rates more accessible for epidemiologists with basic programming experience in R. This R package builds on existing tools such as those for working with databases from R (e.g., DBI, dbplyr, CDMConnector), and can be used alongside other OHDSI resources for defining outcome and strata cohorts.[23] Our software includes a comprehensive approach that allows the analysis from identifying a denominator population to computing both incidence and prevalence estimates. In addition, it provides functionality to allow for various design choices (e.g., age and sex stratifications, calendar time intervals), to track relevant information (e.g., analysis settings, patient's attrition), and to easily export the obtained results. Future releases of the package will incorporate additional functionality to calculate incidence rate ratios and standardised incidence rates.

Here, we used this R package to replicate three studies using four healthcare data sets mapped to the OMOP CDM from the Netherlands, Spain and the United Kingdom, obtaining similar results to those previously published. Discrepancies in incidence rates of COVID-19 can be attributed to the COVID-19 definition used in each analysis. For the IncidencePrevalence results, the definition used included both clinical diagnoses and positive test results for SARS-CoV-2. However, official data only considered positive test results as COVID-19 cases. While the most notable differences occurred during the first months of the pandemic (when testing was not widely available), incidence rates largely coincide afterwards. To replicate the overall and age-sex-specific incidence and prevalence estimates for AESI and ranitidine use, we used the same study periods (only for the AESI comparison) and outcome definitions than those utilised in the previously published studies. We also utilised the same stratifications and prior observation requirements. While the original studies were informed by multiple databases, we chose to replicate each study using a single, distinct database. The rationale behind this choice was to demonstrate the package's performance computing different estimates across different data sources. The obtained estimates were consistent with previous results, and were indicative of high agreement. In terms of performance, IncidencePrevalence showed fast execution times considering the sample sizes of the data sets used, which ranged from 2.6 million individuals in IPCI to 39 million in CPRD Aurum. Note that execution times not only rely on package performance, but also on the server and the environment in which analyses are performed.

The IncidencePrevalence package also has limitations. First, it is designed to be used against routinely collected data and so, data quality issues inherent to observational studies must be considered. In this line, the quality of the data and the phenotypes used to define event cohorts, including the duration of entry event, will impact the obtained results. Therefore, databases converted to the OMOP CDM should be ideally assessed for quality prior to using this package or generating evidence.[26,38] This package can only be used in data sets mapped to the OMOP CDM. While this might limit its applicability, the description of the design choices and the approach employed to evaluate the face validity of the package, can hold value for other researchers who are planning to conduct similar analyses or develop similar tools using data sources not mapped to the OMOP CDM. Lastly, the face validity assessment performed in this work is solely based on showcasing the similarity between the results obtained with the IncidencePrevalence package compared to previously published studies using study-specific custom code rather than relying on quantitative statistical measures (as there is no 'gold standard'). Therefore, we cannot definitively confirm the accuracy of the results based on face validity assessment, underlining the importance of the numerous unit tests in the package to check code correctness.

In-depth analysis of health event frequencies using large real-world data sets requires well-tested, and flexible software solutions. The standardisation of healthcare data to a common format, such as the OMOP CDM, makes the development of standardised analytic tools possible. The IncidencePrevalence R package enables reproducible and rapid analyses of incidence and prevalence rates in data sets mapped to the OMOP CDM. We have demonstrated the use of the package by replicating three earlier studies in four different databases. Moreover, we have demonstrated that it has relatively low execution times when analyzing large data sets. The package can be used for any health event registered in the CDM and will provide a useful tool for those generating evidence from data mapped to the OMOP CDM.

AUTHOR CONTRIBUTIONS

Edward Burn, Berta Raventós, Martí Català, Mike Du, Yuchen Guo, Adam Black, Ger Inberg and Kim López-Güell authored the software package. Danielle Newby, Daniel Prieto Alhambra, Maria de Ridder, Talita Duarte-Salles and Peter Rijnbeek provided epidemiological expertise and user feedback during the development of the package. Mike Du, Yuchen Guo and Kim López-Güell were responsible for testing the package and benchmarking its performance. Xintong Li, Maria de Ridder, Berta Raventós and Martí Català were responsible for performing the validation studies. Berta Raventós and Martí Català wrote the initial version of the manuscript. All authors interpreted the results, critically reviewed the manuscript and approved the final version for submission.

ACKNOWLEDGEMENTS

We acknowledge the contributions by Luis Pinheiro, who critically reviewed the content of this manuscript, and provided useful feedback.

FUNDING INFORMATION

Development of the IncidencePrevalence R package was funded by the European Medicines Agency as part of the Data Analysis and Real World Interrogation Network (DARWIN EU®). This manuscript represents the views of the DARWIN EU® Coordination Centre only and cannot be interpreted as reflecting the views of the European Medicines Agency or the European Medicines Regulatory Network.

CONFLICT OF INTEREST STATEMENT

DP-A's research group has received research grants from the European Medicines Agency, from the Innovative Medicines Initiative, from Amgen, Chiesi, and from UCB Biopharma; and consultancy or speaker fees from Astellas, Amgen and UCB Biopharma. KV and PR work for a research group that—in the past 5 years—received unconditional research grants from, GSK, Amgen, Chiesi, Astra-Zeneca, UCB, J&J, the European Medicines Agency and the Innovative Medicines Initiative.

DATA AVAILABILITY STATEMENT

In accordance with current European and national law, the data used in this study is only available for the researchers participating in this study. Thus, we are not allowed to distribute or make publicly available the data to other parties. The source code of the R package is publicly available at: https://github.com/darwin-eu/IncidencePrevalence

ETHICS STATEMENT

The protocol for this research was approved by the independent scientific advisory committee for Medicine and Healthcare products Regulatory Agency database research (protocol number 20_000211 and 22_002279), the clinical research ethics committee of Fundació Institut Universitari per a la recerca a l'Atenció Primària de Salut Jordi Gol i Gurina (project code: 21/052-PCV), and the Integrated Primary Care Information review board (protocol number 2019-13).

ORCID

Berta Raventós https://orcid.org/0000-0002-4668-2970
Martí Català https://orcid.org/0000-0003-3308-9905
Mike Du https://orcid.org/0000-0002-9517-8834
Yuchen Guo https://orcid.org/0000-0002-0847-4855
Adam Black https://orcid.org/0000-0001-5576-8701
Ger Inberg https://orcid.org/0000-0001-8993-8748
Xintong Li https://orcid.org/0000-0002-6872-5804
Kim López-Güell https://orcid.org/0000-0002-8462-8668
Danielle Newby https://orcid.org/0000-0002-3001-1478
Maria de Ridder https://orcid.org/0000-0002-6555-2741
Cesar Barboza https://orcid.org/0009-0002-4453-3071
Talita Duarte-Salles https://orcid.org/0000-0002-8274-0357
Katia Verhamme https://orcid.org/0000-0001-8162-4904
Peter Rijnbeek https://orcid.org/0000-0003-0621-1979
Daniel Prieto Alhambra https://orcid.org/0000-0002-3950-6346
Edward Burn https://orcid.org/0000-0002-9286-1128

REFERENCES

1. Rassen JA, Bartels DB, Schneeweiss S, Patrick AR, Murk W. Measuring prevalence and incidence of chronic conditions in claims and electronic health record databases. *Clin Epidemiol*. 2018;11:1-15.
2. Roberts AW, Dusetzina SB, Farley JF. Revisiting the washout period in the incident user study design: why 6-12 months may not be sufficient. *J Comp Eff Res*. 2015;4(1):27-35.
3. 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;19(1):54-60.
4. Wickham H, Danenberg P, Csárdi G, et al. roxygen2: In-Line Documentation for R (version 7.2.3). 2022 https://CRAN.R-project.org/package=roxygen2
5. Lang M, Bischl B, Tóth D. checkmate: Fast and Versatile Argument Check; (version 2.1.0). 2022 https://CRAN.R-project.org/package=checkmate
6. Csárdi G, Wickham H, Müller K. cli: Helpers for Developing Command Line Interfaces; (version 3.6.0). 2023 https://cran.r-project.org/web/packages/cli/index.html
7. Wickham H, Girlich M, Ruiz E. dbplyr: A 'dplyr' Back End for Databases; (version 2.1.1). 2022 https://CRAN.R-project.org/package=dbplyr
8. Wickham H, François R, Henry L, Müller K. dplyr: A Grammar of Data Manipulation; (version 1.0.10). 2022 https://CRAN.R-project.org/package=dplyr
9. Spinu V, Grolemund G, Wickham H, et al. lubridate: Make Dealing with Dates a Little Easier; (version 1.9.0). 2022 https://CRAN.R-project.org/package=lubridate
10. Hester J, Bryan J. glue: Interpreted String Literals; (version 1.6.2). 2022 https://CRAN.R-project.org/package=glue
11. Milton Bache S, Wickham H, Henry L. magrittr: A Forward-Pipe Operator for R; (version 2.0.3). 2022 https://CRAN.R-project.org/package=magrittr
12. Henry L, Wickham H. rlang: Functions for Base Types and Core R and 'Tidyverse' Features; (version 1.0.6). 2022 https://CRAN.R-project.org/package=rlang
13. Wickham H, Henry L. purrr: Functional Programming Tools; (version 1.0.1). 2023 https://CRAN.R-project.org/package=purrr
14. Wickham H, Girlich M. tidy r: Tidy Messy Data; (version 1.2.1). 2022 https://CRAN.R-project.org/package=tidyr
15. Henry L, Wickham H. tidyselect: Select from a Set of Strings; (version 1.2.0). 2022 https://CRAN.R-project.org/package=tidyselect
16. Wickham H. stringr: Simple, Consistent Wrappers for Common String Operations; (version 1.5.0). 2022 https://CRAN.R-project.org/package=stringr
17. Csárdi G, Podgórski K, Geldreich R. zip: Cross-Platform 'zip' Compression; (version 2.2.2). 2022 https://CRAN.R-project.org/package=zip
18. Raventós B, Fernández-Bertolín S, Aragón M, et al. Transforming the information system for research in primary care (SIDIAP) in Catalonia to the OMOP common data model and its use for COVID-19 research. *Clin Epidemiol*. 2023;15:969-986.
19. Junior EPP, Normando P, Flores-Ortiz R, et al. Integrating real-world data from Brazil and Pakistan into the OMOP common data model and standardized health analytics framework to characterize COVID-19 in the global south. *J Am Med Inform Assoc*. 2023;30(4):643-655.
20. Papez V, Moinat M, Voss EA, et al. Transforming and evaluating the UK Biobank to the OMOP common data model for COVID-19 research and beyond. *J Am Med Inform Assoc*. (published correction appears in J Am Med Inform Assoc. 2023 Apr 19;30(5):1006). 2022;30(1):103-111.
21. R Special Interest Group on Databases (R-SIG-DB), Wickham H, Müller K. DBI: R Database Interface; (version 1.1.3). 2022 https://CRAN.R-project.org/package=DBI
22. Black A. CDMConnector: Connect to an OMOP Common Data Model; (version 1.0.0). 2022 https://cran.r-project.org/web/packages/CDMConnector/index.html
23. Wickham H. Testthat: get started with testing. *R J*. 2011;3(1):5-10.
24. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934;26:404-413.
25. Wilson EB. Probable inference, the law of succession, and statistical inference. *J Am Stat Assoc*. 1927;22(158):209-212.
26. Observational Health Data Sciences and Informatics. The Book of OHDSI. 2019 https://ohdsi.github.io/TheBookOfOhdsi/
27. UK Coronavirus Dashboard. UK Health Security Agency. Accessed January 2023 https://api.coronavirus.data.gov.uk/v2/data?areaType=region&metric=newCasesBySpecimenDate&format=csv
28. Dades COVID. Generalitat de Catalunya. Accessed January 2023 https://www.dadescovid.cat
29. Li X, Ostropolets A, Makadia R, et al. Characterising the background incidence rates of adverse events of special interest for COVID-19 vaccines in eight countries: multinational network cohort study. *BMJ*. 2021;373:n1435.
30. Verhamme K, Rijnbeek P, de Ridder M. Ranitidine and Other Histamine-H_2-Receptor antagonist – A Drug Utilisation Study. 2023. Accessed April 3, 2023 https://www.encepp.eu/encepp/openAttachment/studyResult/39070;jsessionid=zE6hWB4IKJqtXhO0lOvz6lyTpFXdYTQeZ3jgv_OV8U5oJuxZ0FjF!-2033451844
31. Herrett E, Gallagher AM, Bhaskaran K, et al. Data resource profile: clinical practice research datalink (CPRD). *Int J Epidemiol*. 2015;44(3):827-836.
32. Wolf A, Dedman D, Campbell J, et al. Data resource profile: clinical practice research datalink (CPRD) Aurum. *Int J Epidemiol*. 2019;48(6):1740-1740g.

33. de Ridder MAJ, de Wilde M, de Ben C, et al. Data resource profile: the integrated primary care information (IPCI) database, The Netherlands. *Int J Epidemiol*. 2022;51(6):e314-e323.
34. Recalde M, Rodríguez C, Burn E, et al. Data resource profile: the information system for research in primary care (SIDIAP). *Int J Epidemiol*. 2022;51(6):e324-e336.
35. Burn E, Fernández-Bertolín S, Voss EA, et al. Establishing and characterising large COVID-19 cohorts after mapping the information system for research in primary care in Catalonia to the OMOP common data model. medRxiv. 2021:21266734.
36. Office for National Statistics (ONS). Statistical bulletin, Population estimates for the UK, England, Wales, Scotland and Northern Ireland: mid-2021. 2022. Accessed April 2023 https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2021
37. Statistical Institute of Catalonia. Population on 1 January. 2022. Accessed April 2023 https://www.idescat.cat/indicadors/?id=aec&n=15223&lang=en
38. Blacketer C, Defalco FJ, Ryan PB, Rijnbeek PR. Increasing trust in real-world evidence through evaluation of observational data quality. *J Am Med Inform Assoc*. 2021;28(10):2251-2257.

How to cite this article: Raventós B, Català M, Du M, et al. IncidencePrevalence: An R package to calculate population-level incidence rates and prevalence using the OMOP common data model. *Pharmacoepidemiol Drug Saf*. 2023;1-11. doi:10.1002/pds.5717