Comparative Evaluation of Methods for Defining Observation Periods in Healthcare Databases and Their Impact on Incidence Rate Estimates

Clair Blacketer[1,2], Patrick Ryan[1,3], Peter Rijnbeek[2], Martijn Schuemie[1,4]

[1]Janssen Research & Development, Raritan, NJ, [2]Department of Medical Informatics, Eramus MC, Rotterdam, NL, [3]Department of Biomedical Informatics, Columbia University, New York, NY, USA, [4]Department of Biostatistics, University of California, Los Angeles, CA, USA

Background

Defining the Observation Period table when standardizing data to the OMOP Common Data Model (CDM) is inherently challenging. Observation periods determine the span of time during which patient data is considered complete and reliable for analysis, impacting the validity of any findings drawn from such data. Traditionally, prospective studies track patients from enrollment through follow-up, carefully recording all events to precisely calculate this time. However, secondary use of real-world data (RWD), such as electronic health records (EHR) and administrative claims, often involves inferring these periods from existing records, which can be challenging and inconsistent. Studies have shown significant discrepancies between EHR data (encounter-based) and data that rely on enrollment, like US claims, highlighting the need for robust methods to define observation periods accurately[1–3].

The primary objective of this study is to test different definitions of observation periods to understand their impact on the evidence generated from real-world health data. By comparing multiple approaches across various databases, we aim to identify best practices for defining observable time, particularly in encounter-based datasets where this is most challenging in order to offer methodological guidance for researchers and data analysts to ensure more accurate and reliable outcomes in epidemiological research.

Methods

This study includes 11 databases standardized to the OMOP Common Data Model (CDM) version 5.3 or higher, including six enrollment-based databases (Merative™ Marketscan® Commercial Database, Merative™ Marketscan® Multi-State Medicaid Database, Merative™ Marketscan® Medicare Database, Optum's Clinformatics® De-Identified Data Mart, JMDC, and IQVIA® Adjudicated Health Plan Claims Data) and five encounter-based databases (Optum® De-Identified Electronic Health Records, IQVIA® Longitudinal Patient Database in Australia, IQVIA® Disease Analyzer France, IQVIA® Disease Analyzer Germany, and Premier).

We conduct a methodological experiment using these databases, replicating the study "Characterising the background incidence rates of adverse events of special interest for covid-19 vaccines in eight countries: multinational network cohort study" by Li et al [4].

This replication focuses on five adverse events (acute myocardial infarction, deep vein thrombosis, pulmonary embolism, anaphylaxis, and narcolepsy) to represent a range of conditions from rare to prevalent and acute to less acute.

The study employs several observation period definitions:

1. **Persistence + Surveillance Method**: A persistence window is the maximum allowed number of days between event records to create an era of persistent observation. The surveillance window is the number of days added to the end of the persistent observation era as a period of surveillance prior to the end of the observation period. This study applies 8 persistence windows (180, 365, 548, 730, 1095, 1460, 1825, 2190 days) combined with 5 surveillance windows (0, 30, 90, 180, 270 days) across each database for two event types (all events and medical events). Medical events refer to any condition, drug, procedure, observation, device, etc. except for pharmacy drug dispensings. All events refer to medical events + pharmacy drug dispensings.
2. **Age + Gender Method**: We calcuate average time between health events by person, age, and gender, generating dynamic persistence and surveillance windows (99th percentile for persistence; 10th, 25th, median, 75th, and 90th percentiles for surveillance).
3. **Min/Max Method**: Defined observation periods as the time from the earliest to the latest observed event, applied in encounter-based databases only.

Enrollment time is used as the gold standard in the 6 enrollment-based databases. Two additional patient populations were created for comparison: one included enrollment time for patients with any health event, and the other for those with any medical event only. To compare the incidence rates (IRs) generated using these methods, we calculated the mean squared error (MSE) between the IRs from the different observation period definitions and those generated using enrollment time. This approach allowed us to identify the observation period definitions that most closely approximated the gold standard IRs.

Results

The results demonstrated that the persistence + surveillance method showed a decreasing trend in IRs across both enrollment- and encounter-based databases (Figure 1). There is a clear inverse relationship between persistence and surveillance observed in Figure 2. A 548-day persistence window combined with 270-day surveillance windows produces IRs close to those generated using enrollment time when only medical events are used. As the persistence window increases the surveillance window then must to decrease to keep the IR close to the enrollment IR. This suggests a balance is necessary between the two values to accurately estimate the observation period. When all health events are used this trend begins earlier, with the 365-day persistence combined with the 270-day surveillance. This indicates that higher persistence and surveillance values should be used if only medical events are available. In MDCR the IRs that come close to the enrollment IR begin at a

persistence of 365 days and surveillance of 90 days when only medical events are used and 180-day persistence combined with 30-day surveillance when all events are used.

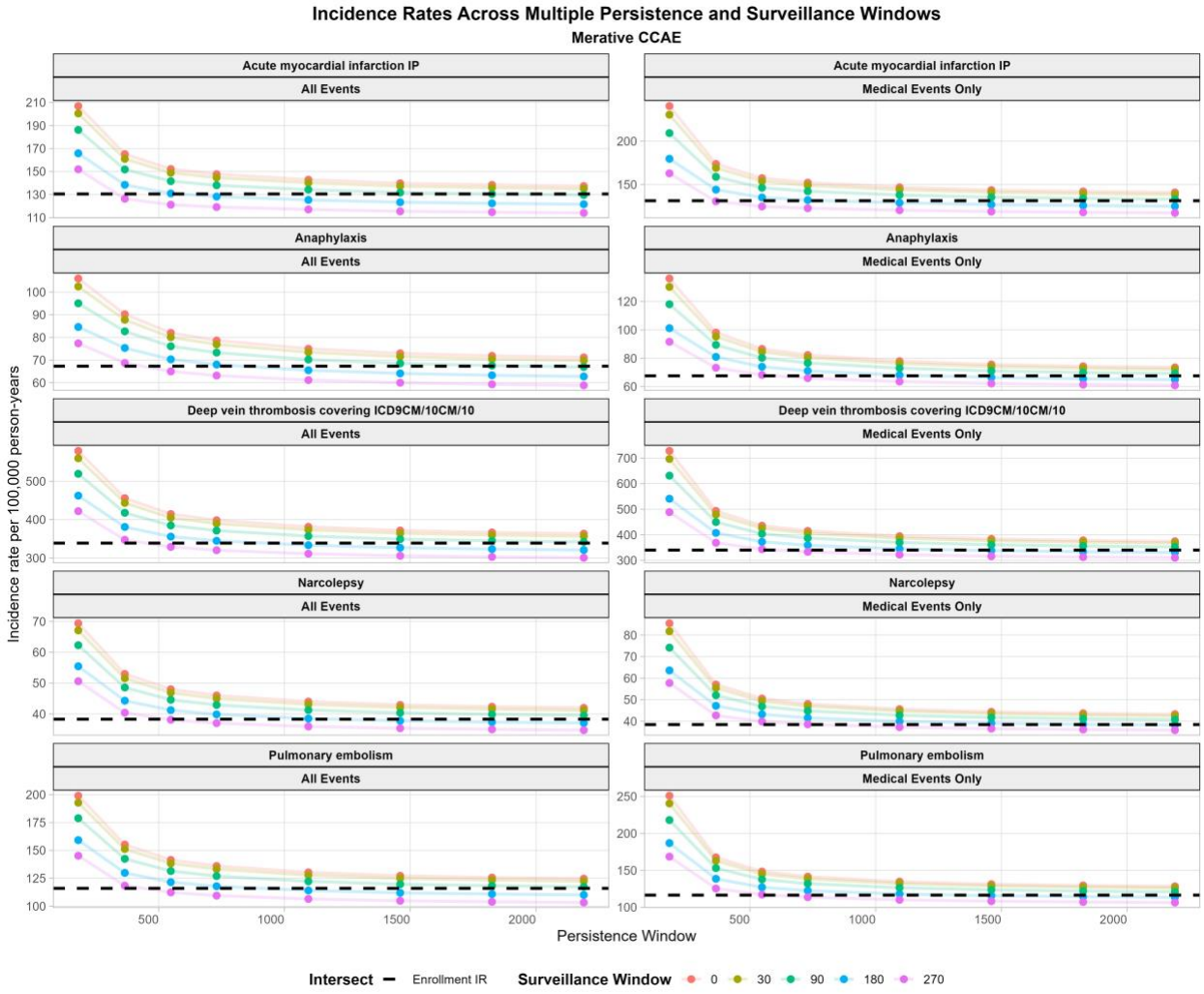

Figure 1: Incidence rates in Merative CCAE across multiple persistence and surveillance windows, five outcomes, and two event types

The age + gender approach, while showing some alignment with the gold standard, did not perform as well, particularly for rare outcomes like narcolepsy and anaphylaxis. The min/max method, shown in Figure 2 as persistence 9999 and surveillance 0, was found to be less reliable. In some claims databases, like IQVIA Pharmetrics, it came close to the enrollment IR while in others it did not, like MDCR and Optum Extended.

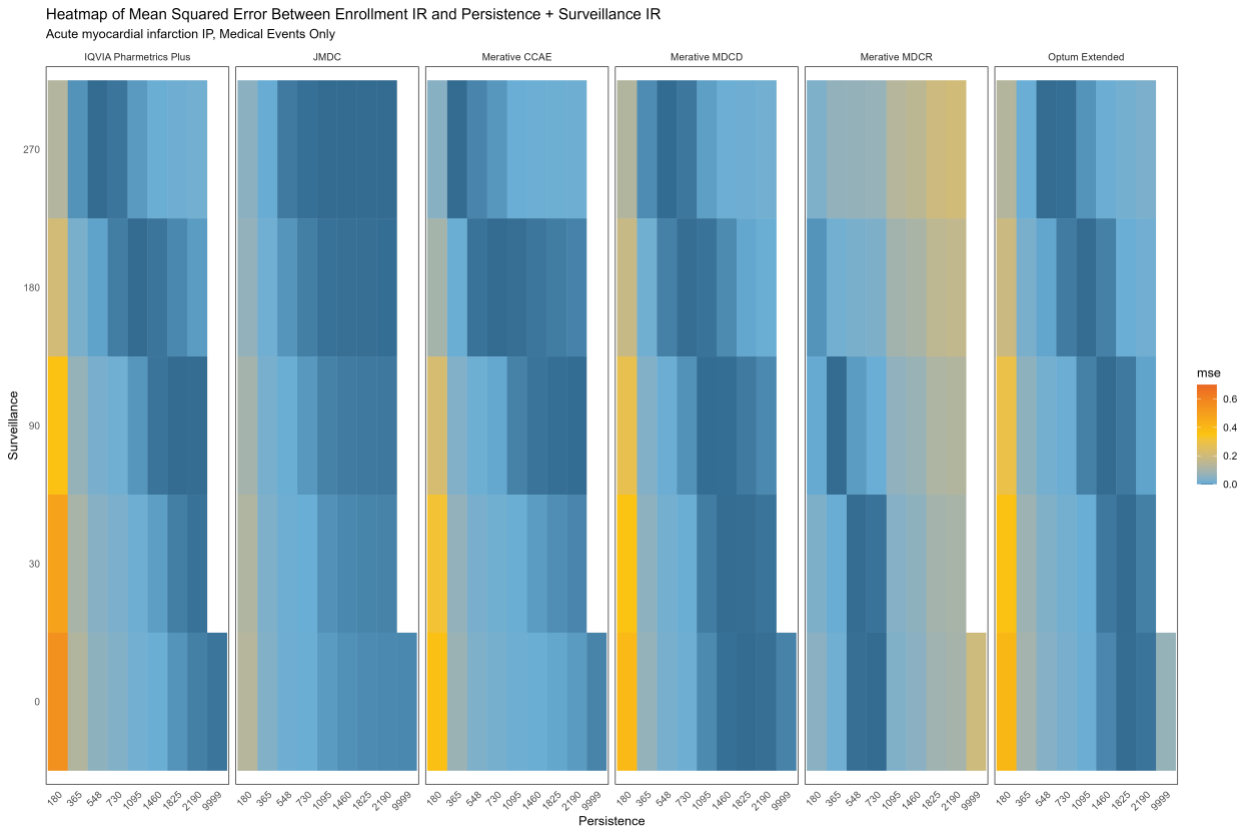

Figure 2: Heatmap of mean squared error between enrollment incidence rate and persistence + surveillance IR using medical events only for the outcome acute myocardial infarction with the min/max approach expressed as persistence 9999 and surveillance 0.

Discussion

This study underscores the importance of accurately defining observation periods in healthcare databases. The persistence + surveillance method is recommended for encounter-based databases, with specific persistence and surveillance windows based on the events available and median age, as detailed in Table 1.

Table 1: Persistence and surveillance recommendations for defining the observation period in EHR data (*P = persistence, S = surveillance)*.

	Events Available	
Median Age	Medical Only	Medical + Pharmacy
< 70	P1460 + S180	P1095 + S180
>= 70	P548 + S30	P365 + S30

Future research should focus on refining the age + gender model to better capture healthcare utilization patterns. Additionally, explicit documentation of observation period

definitions in study protocols is crucial for transparency and reproducibility in epidemiologic research.

Conclusion

This research underscores the necessity of methodological rigor in defining observation periods within healthcare databases. By demonstrating the impact of different definitions on IR estimates, the study provides a foundation for improving evidence generation from observational health data. Future work should focus on refining age and gender models and further exploration of the implications of observation period definitions across diverse healthcare settings.

References

1 Lin KJ, Singer DE, Glynn RJ, *et al*. Identifying Patients With High Data Completeness to Improve Validity of Comparative Effectiveness Research in Electronic Health Records Data. *Clin Pharmacol Ther*. 2018;103:899–905.

2 Hsu C-N, Huang K, Lin F-J, *et al*. Continuity and Completeness of Electronic Health Record Data for Patients Treated With Oral Hypoglycemic Agents: Findings From Healthcare Delivery Systems in Taiwan. *Frontiers in Pharmacology*. 2022;13.

3 Lin KJ, Glynn RJ, Singer DE, *et al*. Out-of-system Care and Recording of Patient Characteristics Critical for Comparative Effectiveness Research. *Epidemiology*. 2018;29:356.

4 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.