Estimating Observable Time in the Absence of Defined Enrollment

Clair Blacketer[1,2], Patrick Ryan[1,3], Frank DeFalco[1], Martijn Schuemie[1,4], Peter Rijnbeek[2]
[1]Janssen Research & Development, Raritan, NJ, [2]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

Traditional, prospective epidemiologic studies enroll a specific cohort of people and interacts with them individually, following them for a period of time(1). Any outcome of interest, loss to follow-up, or deaths are diligently recorded so that the amount of time each person was observed can be calculated. This is particularly important when calculating rates of disease. In contrast, retrospective studies that make use of existing observational data must define observable time for each patient. This is relatively simple in databases that have the notion defined enrollment, for example in US insurance claims and countries that require registration with a General Practitioner (GP). Patients are enrolled either in the health plan or the GP practice and all encounters with the health care system or practice are observed and recorded during the period of enrollment(2). However, it is much more difficult in databases that are primarily encounter-based, such as electronic health records (EHRs). Often the recorded date and time from first observed encounter to the last observed encounter is used, but a guideline from the OHDSI community suggests concatenating encounters that occur within 18 months of each other into periods of observable time(3,4). This pilot study explores multiple approaches for the creation of the observation period, the implications of these definitions, and recommendations for choosing a definition when a database lacks defined enrollment.

Methods

For this pilot we use the US administrative claims database Merative Commercial Claims and Encounters database (described in Appendix 1) that has been mapped to the OMOP CDM v5.4(5). We chose a US claims database because it has clearly defined enrollment per person, which is the time they are enrolled in a health insurance plan. Given that a certain percentage of patients in a health plan do not use the healthcare system at all, we used the enrollment for patients with at least one health event as the gold standard observation period. As an alternate approach, we then used observed health care encounters to create alternative definitions of the observation period without utilization of the defined enrollment.

All health events were de-duplicated on event date such that each date the person interacted with the health system was only represented one time. Then, we applied combinations of persistence and surveillance windows to those events to create eras of observable time. The persistence window allows for a maximum of some number of days between event records. The surveillance window then added some number of days to the end of the era of persistent observation as an additional period of surveillance prior to the observation period end date.

Figure 1 demonstrates how persistence and surveillance work together to create observation periods.

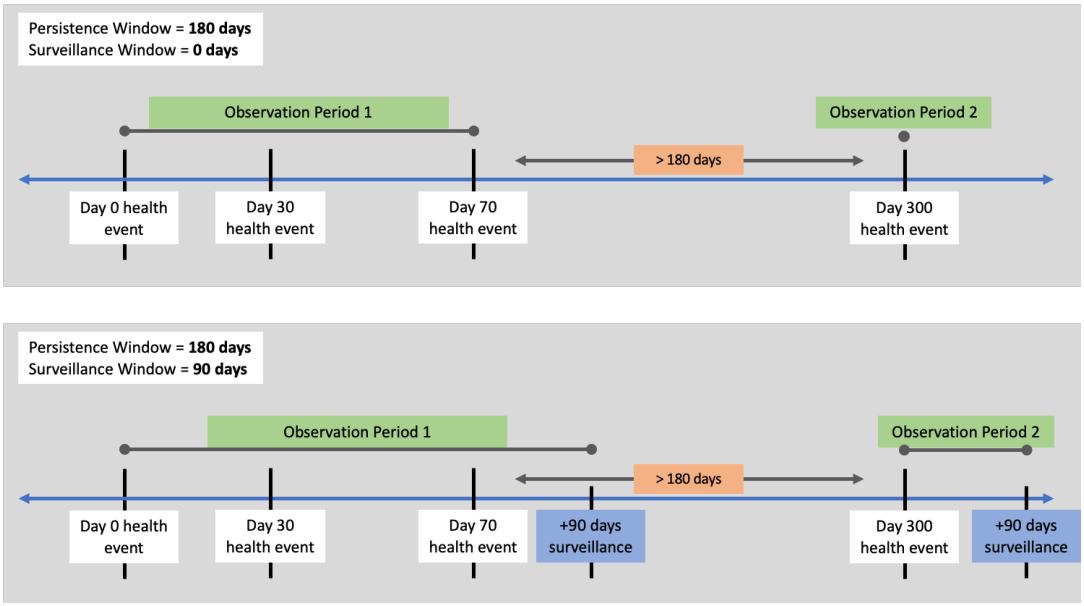

Figure 1. Graphical representation of the persistence and surveillance windows to create periods of observable time

We used 8 different persistence windows (in days): 180, 365, 548, 730, 1095, 1460, 1825, 2190 to combine events into eras of observable time. We also used 3 surveillance windows (in days) of 0, 90, and 180 days. We created two event definitions with which to build the eras: 1) all events 2) all events excluding drug dispensings. This choice was made because many databases without defined enrollment have records of prescriptions written but not prescriptions filled. Using 8 persistence windows, 3 surveillance windows, and2 event definitions, a total of 48 (8*3*2) observation period permutations were defined.

To understand the impact of the choice of persistence and surveillance windows and to determine which combination comes closest to the gold standard of defined enrollment for patients with at least one event, we replicated 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(6). This study was chosen because incidence rates are highly responsive to changes in observable time and the 15 adverse events of special interest included both rare and common conditions. The target cohort included anyone observed between 2017 and 2019 with at least 365 days of continuous prior observation. The outcomes chosen were non-haemorrhagic and haemorrhagic stroke, acute myocardial infarction, deep vein thrombosis, pulmonary embolism, anaphylaxis, Bell’s palsy, myocarditis or pericarditis, narcolepsy, appendicitis, immune thrombocytopenia, disseminated intravascular coagulation, encephalomyelitis (including acute disseminated encephalomyelitis), Guillain-Barré syndrome, and transverse myelitis. The code used to generate the incidence rates can be accessed at https://github.com/ohdsi-studies/Covid19VaccineAesiIncidenceCharacterization.

Results

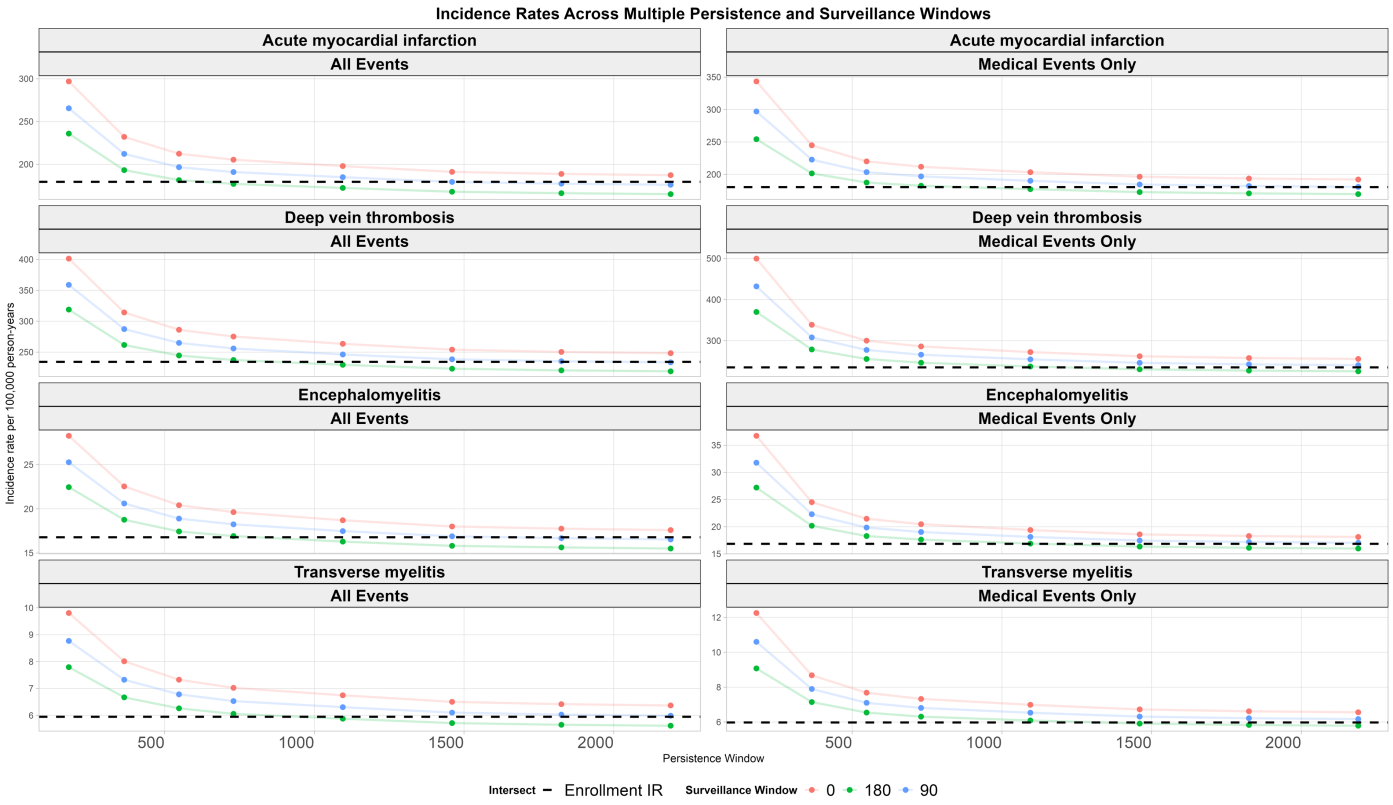

Figure 2. Incidence Rates per 100K person-years by outcome, event type, persistence, and surveillance windows

Figure 2 shows the incidence rates per 100,000 person-years for three of the fifteen outcomes: acute myocardial infarction (AMI), encephalomyelitis, and transverse myelitis. The y-axis is the incidence rate, and the x-axis is the persistence window used. Each color represents the surveillance window. Red is 0 days, blue is 90 days, and green is 180 days. The dotted black line is the incidence rate as calculated using the enrollment period. Two plots are shown for each outcome, one where all events were used to determine the incidence rate and the other where only medical events (excluding drug dispensings) were used.

For all three outcomes, event types, and surveillance windows, the lowest persistence window of 180 days overestimated the incidence rate between 1.5 to 3 times that using enrollment time. When only medical events were used, the incidence rates were consistently higher than when all events were used. The surveillance window of 0 never reached the enrollment line while the surveillance window of 180 reached the enrollment line for all outcomes between persistence windows of 600 – 900 days when all events were used and between persistence windows of 900 – 1250 days when only medical events were used. At the higher persistence

windows of 2000 days and above, the surveillance window of 180 regularly underestimated the incidence rate using enrollment time.

Limitations

This pilot study is limited in that these analyses were only conducted in one database covering one country and one type of observational health data. Additionally, we are using insurance claims to create the experimental observation periods which may not be generalizable to EHRs. Future work includes creating eras of observable time based on patterns of utilization by age and sex and conducting the experiments across multiple databases with defined enrollment to determine if the same patterns are seen as found in this pilot.

Conclusion

Eras of observable time in encounter-based databases without defined enrollment can be estimated by applying persistence and surveillance windows to observed medical events. The windows used should consider the data available as longer windows should be used when drug dispensings are not included. The results also suggest that a balance of persistence and surveillance should be used when employing this method.

References

1. Mahmood SS, Levy D, Vasan RS, Wang TJ. The Framingham Heart Study and the Epidemiology of Cardiovascular Diseases: A Historical Perspective. Lancet. 2014 Mar 15;383(9921):999–1008.

2. Voss EA, Makadia R, Matcho A, Ma Q, Knoll C, Schuemie M, et al. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. J Am Med Inform Assoc. 2015 May;22(3):553–64.

3. Observation Period Considerations for EHR Data [Internet]. [cited 2023 Jun 6]. Available from: http://ohdsi.github.io/CommonDataModel/ehrObsPeriods.html

4. Janssen ETL Documentation [Internet]. Janssen ETL Documentation. [cited 2023 Jun 6]. Available from: https://ohdsi.github.io/ETL-LambdaBuilder/

5. OMOP Common Data Model [Internet]. [cited 2021 Jan 14]. Available from: http://ohdsi.github.io/CommonDataModel/

6. Li X, Ostropolets A, Makadia R, Shoaibi A, Rao G, Sena AG, 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 Jun 14;373:n1435.

Appendix 1 – Data Source

The data source used in the analysis is the Merative® MarketScan® Commercial Database (CCAE) which includes health insurance claims across the continuum of care (i.e., inpatient, outpatient, outpatient pharmacy, carve-out behavioral healthcare) as well as enrollment data from large employers and health plans across the United States who provide private healthcare coverage for more than 155 million employees, their spouses, and dependents. This administrative claims database includes a variety of fee-for-service, preferred provider organizations, and capitated health plans.