RWD31

Impact of Method Chosen for Defining Observable Time in Linked Open Data Sources

Anna Swenson[1], Gursimran Basra[1], Paul Buzinec*, Kathryn Starzyk* | [1]OM1, Inc, Boston, MA, USA;
*Formerly OM1, Inc.

Background

- Electronic medical record (EMR) and open claims data sources are increasingly used for real-world evidence generation, however, defining observable patient time (the time during which healthcare events are expected to be recorded if they occur) in the absence of enrollment information is challenging.
- Researchers must make assumptions when defining observable time (also referred to as observation periods), and these assumptions are difficult to validate.
- With the use of common data models such as OMOP, definitions of observation periods are increasingly pre-specified at the database level without regards for the characteristics of the underlying data[1].
- Defining observable time in open data real-world data sources is an important study design consideration. When multiple data sources are linked together, the definitions of observation periods become even more complex and may lead to selection bias.

Objective

To explore how varying methods for defining observation periods in linked EMR and open medical claims data impact sample size, comorbidity prevalence, medication usage, and HCRU across three conditions.

Methods

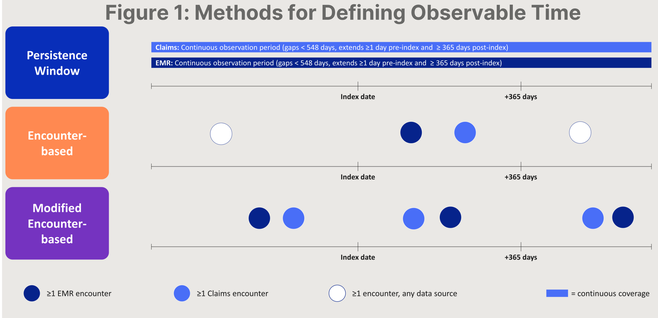
Eligible patients were from Atopic Dermatitis (AD), Rheumatoid Arthritis (RA), and Major Depressive Disorder (MDD) cohorts curated from the OM1 RWDC, a deterministically linked, de-identified, individual-level dataset containing EMR with open medical and pharmacy claims for patients in the United States. The study period was from 2013 to 2024; the index date is the patient's first diagnosis date for the condition in the OM1 data.

We compared three methods for defining observable time in linked datasets:
Method 1: Persistence window: Period of continuous data source type-specific observed encounters with gaps < 548 days rolled into a continuous period. Requires complete EMR/claims overlap for the period from at least one day before to >365 days after the index.
Method 2: Encounter-based: Patients were required to have:
- At least one encounter of any type on or before index
- At least one EMR and one medical claims encounter between index and 365 days
- At least one encounter of any type > 365 days after index

Method 3: Modified Encounter-based: Patients must have both an EMR and medical claims encounter during three time periods: prior to/on index date; on or within the year following the index date; and any time after index+365.

Demographics, comorbidities, medications, and HCRU were compared in the 12-month period post-index for three cohorts across the three observable time definitions. Pairwise absolute standardized mean differences (|SMD|) were used to compare methods; |SMD|≥0.10 was considered a meaningful imbalance[2].

Figure 1: Methods for Defining Observable Time



Results

Initial patient counts were 94,368 (AD), 282,850 (RA), and 1,063,161 (MDD).

After applying observation period criteria, the Persistence Window method resulted in the largest sample size reductions (-59.6%[RA] to -74.7%[MDD]), Encounter-based method the smallest (-36.2%[RA] to -46.8%[MDD]), and Modified Encounter-Based had intermediate reductions (-56.5%[RA] to -66.7%[MDD]). **(Figure 2)**

Mean age and Charlson Comorbidity Index ≥2 were highest for the Persistence Window method and lowest in the Encounter-based method. **(Tables 1-3)**

Comorbidity and medication prevalence varied only slightly across methods (most |SMD| < 0.10), while age, BMI, race distribution, CCI, and HCRU showed larger imbalances — particularly for the Encounter-based method when compared to the Persistence Window method. (**Figures 3-5)**. Persistence Window and Modified Encounter-based methods yield highly similar populations (|SMD| <0.10 for nearly all variables); Encounter-based produces a younger, healthier cohort with fewer outpatient visits.

Outpatient visit counts showed larger differences, with the persistence method having the highest mean visit counts and the encounter-based method the lowest (**Tables 1-3**).

Although results were largely consistent across conditions, the magnitude of differences varied. The RA cohort showed consistency across methods for most variables, with only outpatient visits showing a meaningful difference when comparing the Persistence Window and Encounter-based method. The AD and MDD cohorts showed greater variability between methods, likely representing differences in care patterns and documentation practices (**Figures 3-5**).

Conclusions

- Specifying observable time in open data sources requires trade-offs among data completeness, sample size and representativeness.
- Stricter definitions of linked data availability yielded smaller, older, and sicker populations with more complete data.
- Count data, specifically HCRU, were more sensitive to the method of defining observation periods than prevalence data.
- Method choice must align with study goals to ensure fit-for-purpose data.
- Potential for selection bias resulting from observation period definitions should be carefully assessed and mitigated with attention to condition-specific cohort characteristics.

Figure 2: Percent Reduction in Sample Size Across Three Condition Cohorts According to Method for Defining Observable Time

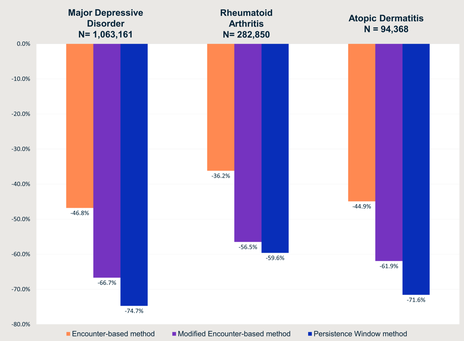


Table 1: Demographic Characteristics of Rheumatoid Arthritis Cohort by Observation Period Method

	Persistence Method N=114,278	Encounter-based Method N=180,494	Modified Encounter-based Method N=123,088
Age, mean (SD)	62.2 (13.1)	61.2 (13.5)	61.7 (13.2)
Female	88,789 (77.7%)	139,554 (77.3%)	95,492 (77.6%)
Race			
Asian	1,654 (1.4%)	2,928 (1.6%)	1,771 (1.4%)
Black or African American	11,141 (9.7%)	17,417 (9.6%)	11,930 (9.7%)
Other/Multiple	5,886 (5.2%)	10,399 (5.8%)	6,526 (5.3%)
Unknown	7,638 (6.7%)	14,466 (8.0%)	7,676 (6.2%)
White	87,959 (77.0%)	135,284 (75.0%)	95,185 (77.3%)
Insurance Type			
Commercial	36,711 (32.1%)	62,546 (34.7%)	40,269 (32.7%)
Medicaid	728 (0.6%)	1,434 (0.8%)	890 (0.7%)
Medicare	22,843 (20.0%)	35,408 (19.6%)	24,584 (20.0%)
Multiple/Other	42,609 (37.3%)	61,695 (34.2%)	44,713 (36.3%)
None/Missing	11,387 (10.0%)	19,411 (10.8%)	12,632 (10.3%)
CCI*			
0	5 (0.0%)	10 (0.0%)	8 (0.0%)
1	53,268 (46.6%)	92,759 (51.4%)	58,425 (47.5%)
≥2	61,004 (53.4%)	87,724 (48.6%)	64,654 (52.5%)
BMI available	108,451 (94.9%)	166,941 (92.5%)	116,180 (94.4%)
Outpatient Visits 12 months post-index, mean (SD)	11.8 (9.0)	10.7 (8.7)	11.3 (8.9)

*RA contributes to CCI; patients with RA diagnosis necessarily have CCI ≥1. The <1% of patients with a CCI = 0 is due to discrepancies in coding between CCI and the RA database.

Table 2: Demographic Characteristics of Major Depressive Disorder Cohort by Observation Period Method

	Persistence Method N=268,908	Encounter-based Method N=565,755	Modified Encounter-based Method N=354,331
Age, mean (SD)	42.4 (17.3)	40.3 (16.8)	41.5 (17.0)
Female	193,258 (71.9%)	395,240 (69.9%)	252,070 (71.1%)
Race			
Asian	2,889 (1.1%)	5,948 (1.1%)	3,879 (1.1%)
Black or African American	13,140 (4.9%)	25,374 (4.5%)	17,337 (4.9%)
Other/Multiple	7,294 (2.7%)	14,380 (2.5%)	9,854 (2.8%)
Unknown	100,917 (37.5%)	248,268 (43.9%)	134,297 (37.9%)
White	144,668 (53.8%)	271,785 (48.0%)	188,964 (53.3%)
Insurance Type			
Commercial	142,438 (53.0%)	312,186 (55.2%)	191,385 (54.0%)
Medicaid	4,809 (1.8%)	12,335 (2.2%)	7,021 (2.0%)
Medicare	15,799 (5.9%)	27,727 (4.9%)	19,542 (5.5%)
Multiple/Other	66,051 (24.6%)	123,652 (21.9%)	81,976 (23.1%)
None/Missing	39,811 (14.8%)	89,855 (15.9%)	54,407 (15.4%)
CCI			
0	164,045 (61.0%)	375,868 (66.4%)	224,238 (63.3%)
1	51,633 (19.2%)	97,938 (17.3%)	65,662 (18.5%)
≥2	53,229 (19.8%)	91,945 (16.3%)	64,429 (18.2%)
BMI available	127,307 (47.3%)	230,399 (40.7%)	161,889 (45.7%)
Outpatient Visits 12 months post-index, mean (SD)	17.4 (15.1)	15.0 (13.9)	16.3 (14.6)

Table 3: Demographic Characteristics of Atopic Dermatitis Cohort by Observation Period Method

	Persistence Method N=26,835	Encounter-based Method N=51,994	Modified Encounter-based Method N=35,925
Age, mean (SD)	52.9 (21.9)	47.9 (23.4)	50.8 (22.3)
Female	17,137 (63.9%)	32,379 (62.3%)	22,686 (63.1%)
Race			
Asian	741 (2.8%)	1,879 (3.6%)	1,155 (3.2%)
Black or African American	3,218 (12.0%)	6,523 (12.5%)	4,403 (12.3%)
Other/Multiple	905 (3.4%)	1,857 (3.6%)	1,265 (3.5%)
Unknown	2,898 (10.8%)	7,619 (14.7%)	4,124 (11.5%)
White	19,073 (71.1%)	34,116 (65.6%)	24,978 (69.5%)
Insurance Type			
Commercial	10,248 (38.2%)	21,754 (41.8%)	14,259 (39.7%)
Medicaid	689 (2.6%)	1,581 (3.0%)	948 (2.6%)
Medicare	6,561 (24.4%)	10,328 (19.9%)	7,987 (22.2%)
Multiple/Other	1,307 (4.9%)	2,212 (4.3%)	1,598 (4.4%)
None/Missing	8,030 (29.9%)	16,119 (31.0%)	11,133 (31.0%)
CCI			
0	11,444 (42.6%)	26,522 (51.0%)	17,032 (47.4%)
1	6,655 (24.8%)	12,330 (23.7%)	8,683 (24.2%)
≥2	8,736 (32.5%)	13,141 (25.3%)	10,210 (28.4%)
BMI available	11,459 (42.7%)	18,849 (36.3%)	14,204 (39.5%)
Outpatient Visits 12 months post-index, mean (SD)	11.1 (10.8)	9.4 (9.6)	10.0 (10.1)

Figure 3: Pairwise |SMD| Across Method Comparisons, Rheumatoid Arthritis

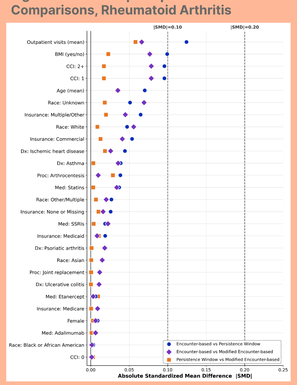


Figure 4: Pairwise |SMD| Across Method Comparisons, Major Depressive Disorder

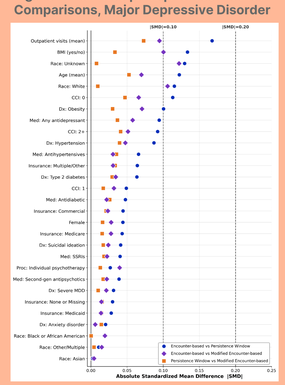
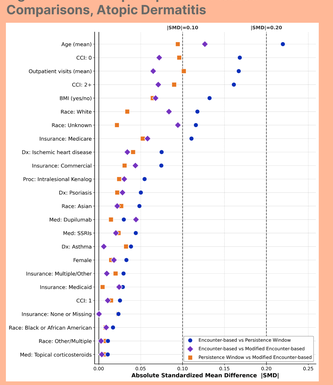


Figure 5: Pairwise |SMD| Across Method Comparisons, Atopic Dermatitis



BMI= Body Mass Index; CCI=Charlson Comorbidity Index; Dx=Diagnosis; Proc= Procedure, Med=Medication; SSRI= Selective Serotonin Reuptake Inhibitors

Conference

Presented at ISPOR 2026 · Philadelphia, PA · May 17–20, 2026

References

1. Blacketer C, DeFalco F, Conover M, Ryan P, Schuemie M, Rijnbeek P. J. Evaluation of the impact of defining observable time in real-world data on outcome incidence. J Am Med Inform Assoc. 2025;32:ocaf119. doi:10.1093/jamia/ocaf119.
2. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. Stat Med. 2009;28(25):3083-3107. doi:10.1002/sim.3697