Clarivate Consulting Services

Measurement of prevalence and incidence in open claims data: assessment of different methodologies

RWD63

Authors: Shirley Xiaolei Li[1], James Young[1], Adam Idica[2], Aishwarya Uday[1], Ekta Bhatnager[3], Hemanth Nair[2]

[1]Clarivate Analytics, Toronto, Canada; ; [2]Clarivate Analytics, USA; [3]Clarivate Analytics, Bangalore, India ;

Background

- **Prevalence and incidence are fundamental epidemiologic measures** for public health planning, care management, and drug development. Healthcare databases provide cost-effective, large-scale data but were not designed for research
- **Real-World Data (RWD) are widely used to estimate disease prevalence and incidence**; however, differences in data sources, analytic choices, and assumptions introduce substantial variation and limit comparability
- **Open claims data offer** the advantage of real-time availability and **massive scale (>200 million patients/year) but come with limitations** such as: 1) **No enrollment data** — denominators must be inferred, 2) **duplicate claims** across multiple clearinghouses, 3) **fragmented longitudinal patient records**, high attrition, and left/right truncation (patients enter study after start or exit before study end), and 4) **Inability to distinguish 'no diagnosis' from 'no data'**, introducing risk of misclassification and false positives
- Prior work shows prevalence and incidence estimates are highly sensitive to design choices: Rassen et al. (2019) observed 1.8–8.3× variation in prevalence in EHR/closed claims, Breskin et al. (2024) demonstrated similar sensitivity across open and closed claims, and Baser et al. (2023) found comparable utilization estimates after deduplication, though prevalence in open claims remains less well characterized.

Objective

- **This study evaluated how methodological specifications affect prevalence and incidence estimates** derived from an open claims database.

Methods

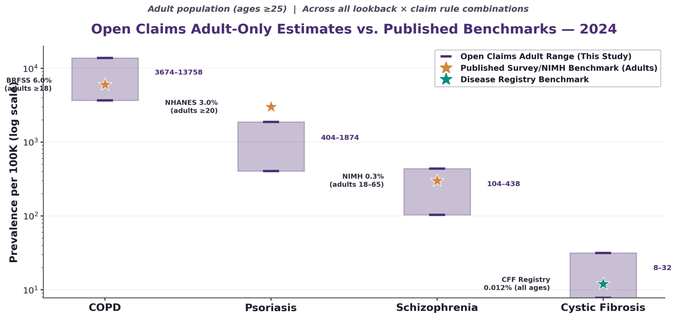
- **Annual prevalence and incidence were estimated** using Clarivate's Real-World Data repository, a database comprising **open claims** obtained from clearinghouses (digital hubs that transmit electronic claims from healthcare providers to government and commercial payers) and electronic health record (EHR)
- Four disease areas were studied:
  1) **Cystic fibrosis** (CF) — very rare (prevalence ~12/100K)
  2) **Schizophrenia** — rare (~300/100K)
  3) **Psoriasis** — moderate (~3,000/100K)
  4) **COPD** — common (~5,000–6,000/100K)
- Three specifications varied (Figure 1):
  Look-back period: 1 year, 2 years, or all available history beginning in 2016
  Disease identification: ≥ 1 claim, or ≥ 2 claims 30-365-days apart
  Continuous enrollment (CE) proxy: no CE, ≥1 claim every 12 months, or ≥1 claim every 6 months
- Outcomes included annual point prevalence and cumulative incidence per 100,000, estimated for years 2018 - 2024

Results

COMPARISON WITH PUBLISHED LITERATURE

- Most open claims estimates fall within published ranges, except psoriasis which falls below benchmark values (Figure 2)

Figure 2: Comparison of results for 2024 with published literature



EFFECT OF LOOKBACK PERIOD (Figure 3)

- **All-time** lookback yielded **1.5–2.1× higher prevalence vs. 1-year**, consistent with Rassen et al.'s 1.8–2.4× in MarketScan
- Lookback has an inverse effect on incidence**: longer lookback classifies more cases as prevalent**, reducing incidence by 1.2–1.5× (all-time vs. 1-year)

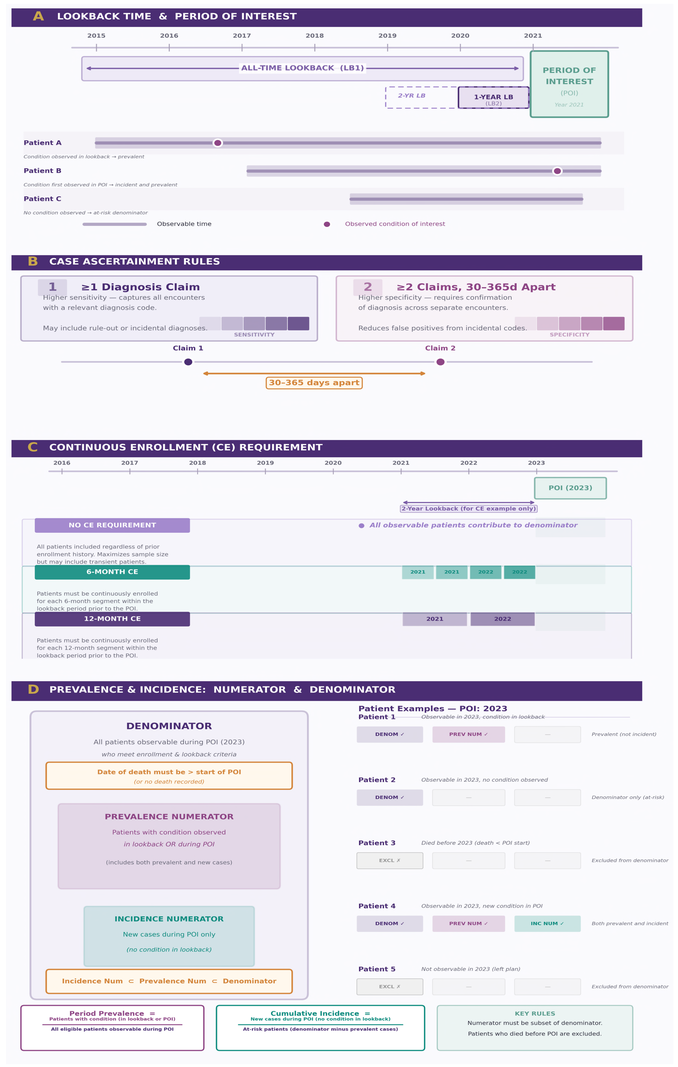
EFFECT OF CASE ASCERTAINMENT

- **≥2 claims reduced prevalence by 2.0–2.3× vs. ≥1 claim**, aligning with Cohen et al. (2020) in MarketScan.
- **≥2 claims reduces incidence 3.3–6.6× vs. ≥1 claim**, reflecting removal of false-positive incident (and prevalent) cases.

EFFECT OF CONTINUOUS ENROLLMENT

- Continuous enrollment**: 6-month CE increased prevalence rates by 1.25–1.67× and 12-month CE by 1.05–1.27× vs. no CE**, driven by a 26% smaller but sicker denominator population
- **Continuous enrollment had mixed effects on incidence**: 6-month CE increased incidence rates for COPD, psoriasis, and schizophrenia (1.20–1.27×), but halved CF incidence (0.52×), likely reflecting differential enrichment of sicker populations across disease areas

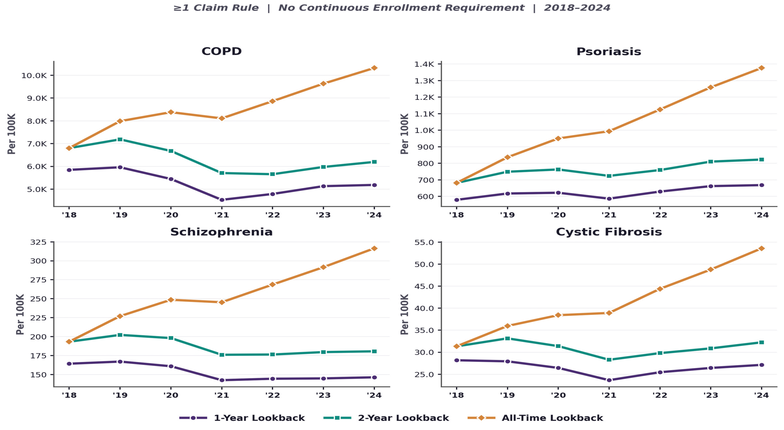
Figure 1: Study Design Framework: 2023 Example



TRENDING

- All trends show a decrease in 2021 driven by a large increase in the denominator, as a backlog of patients requiring non-urgent care sought treatment post-COVID. If this utilization shift differs between the numerator and denominator, resulting trends may be misleading.
- Consistent with Rassen et al., 2019, use of an all-time lookback window created a spurious temporal trend in prevalence and incidence: incidence declined 2018–2024 for COPD (−25%) and schizophrenia (−20%), likely reflecting accumulation of prevalent cases with increasing database history.

Figure 3: Effect of lookback period on prevalence estimates



COMBINED VARIATION

- Combined variation spans 2.9–4.2× across all methods for prevalence.
- Combined incidence variation is 5.7–12.2× — substantially wider than prevalence — with CF showing the widest range.

Figure 4: Comparison of 2024 prevalence across all methods

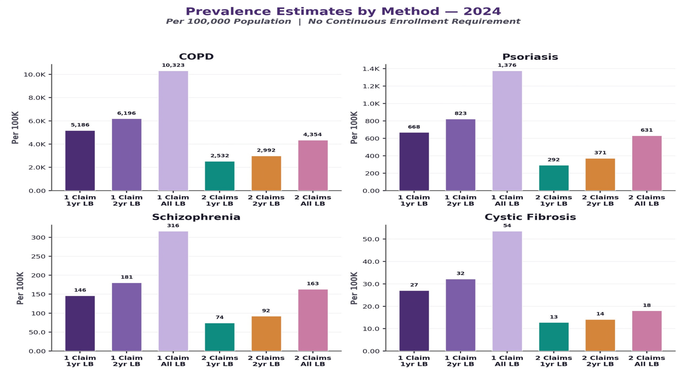
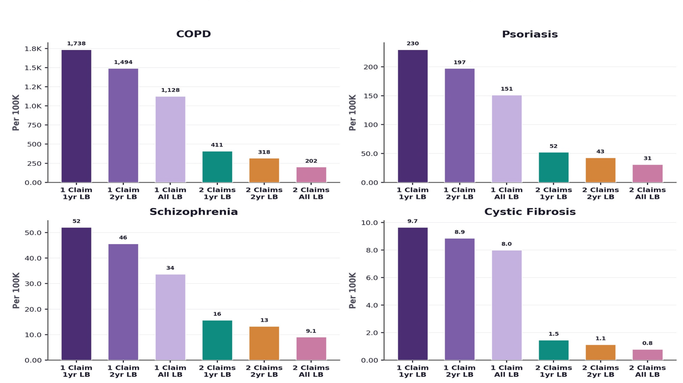


Figure 4: Comparison of 2024 incidence across all methods



Limitations

- Open claims data lack formal enrollment files, so denominators are inferred from claims activity; the true at-risk population may differ from the observed population
- Case definitions relied on ICD diagnosis codes without chart-review validation; sensitivity and specificity may vary across diseases and over time
- Published benchmarks (BRFSS, NHANES, NIMH, CFF Registry) use different populations, age ranges, and case definitions, limiting direct comparability
- Results may not generalize beyond the commercially insured US population captured in open claims. The study did not assess the impact of patient deduplication algorithms, which may differentially affect prevalence estimates in open claims.

Conclusions

- Open claims data yield prevalence and incidence estimates closely aligned with registry and survey benchmarks, reinforcing prior evidence that open claims can complement closed claims and EHR data for disease burden estimation (Secora et al., 2022)
- Methodological sensitivity in Clarivate open claims (2.9–12.2× variation) is comparable to the 1.8–8.3× reported by Rassen et al. (2019) in closed claims/EHR, confirming that design choices — not data source — are the primary driver of estimate variability
- Case ascertainment rule has the largest impact on estimates, yet it is not a core dimension in the Rassen et al. (2019) framework. Given its outsized effect — particularly on incidence — case ascertainment should be formally incorporated as a key design parameter alongside lookback time and continuous enrollment when estimating disease burden in open claims data
- Incidence is particularly sensitive to design choices (5.7–12.2× vs. 2.9–4.2× for prevalence) because lookback determines which cases are "new" while case ascertainment determines which are real. CE requirements introduce selection bias by reducing the denominator ~26% while enriching for sicker patients
- Open claims estimates are influenced by external factors such as COVID-era utilization shifts, requiring additional adjustments when assessing temporal trends
- Standardized reporting per the Rassen et al. framework and ISPOR-ISPE good practices (Berger et al., 2017; Wang et al., 2023), with triangulation across data sources and transparent parameter selection, should be the norm for producing interpretable, comparable real-world evidence for healthcare planning and drug development

Full reference list available upon request

Presented at ISPOR 2026, Philadelphia, PA | May 18, 2026