Data lag in a large open and closed claims dataset: Navigating the completeness-timeliness tradeoff

Andrew R. Weckstein, BA[1]; Elizabeth M. Garry, PhD, MPH[1]; Karthik Uppaluri, BS[1]; Ulka Campbell, PhD, MPH[1]; Monica Gierada, MPH[1]; Emily Rubinstein, MPH[1]; Reyna Klesh, MS[2]; Nicolle M. Gatto, PhD, MPH[1]

[1]Aetion, New York, NY; [2]HealthVerity, Philadelphia, PA

Background & Objective

Unlike traditional adjudicated (closed) insurance claims, open claims data sourced from non-insurer intermediaries may be available for research purposes within weeks or even days. However the tradeoff between the timeliness and completeness of these data remains to be further evaluated. An improved understanding of latency in open and closed claims is needed to support researchers in assessing data fitness, determining if and how to truncate data to ensure valid capture of study events, and to otherwise inform appropriate use of these data types for public health surveillance and research purposes.

➔ **Objective: Describe latency and completeness trends of open and closed medical claims in a large real-world dataset**

Methods

Data: This study compared claims data received in Apr 2021 (*initial datacut*) to a subsequently updated cut of data received 15 months later (*updated datacut*), encompassing two different HealthVerity medical claims sources:

- **Closed claims** - Adjudicated claims sourced from insurers. Captures all encounters submitted for reimbursement for patients with active enrollment/eligibility.
- **Open claims** - Adjudicated and pre-adjudicated claims sourced predominantly from providers and clearinghouses. Not associated with any insurance enrollment/eligibility files, therefore only encounters at providers within the HealthVerity open claims network are observable.

Population: Cohort of patients observable in *initial* and *updated* datacuts for both open and closed claims sources, defined as having:

- ≥ 1 closed medical claim and ≥ 1 open medical claim Apr 2020 to Apr 2021, and
- Continuous enrollment in medical benefit during study period Apr 2020 - Apr 2021 (based on closed source only)

Data Analysis: We describe latency over calendar time with the following metrics:

- Patient claim event counts (by week) defined within each datacut as the number of patients with ≥1 claim event during a given calendar week.
- Percent completeness (by week) defined by the ratio of the number of patients with claim events in the *initial datacut* relative to the number of patients with claim events in the *updated datacut*, for the same calendar week. Claim capture in the *updated datacut* is considered the benchmark for 100% completeness during the study period.

We report metrics overall, by open and closed claims and separately for the following event types:

- All-cause medical claims, by care setting (all medical settings; outpatient; inpatient)
- COVID-19 (U07.1) claims, by care setting (claim with ICD-10 CM of U07.1 from any medical setting; U07.1 outpatient; U07.1 inpatient)

Results

Figure 1. Data latency trends in closed and open claims sources, all medical claim events

Trends plotted Apr 2020 to Jun 2021 for cohort with N=10.7 million patients

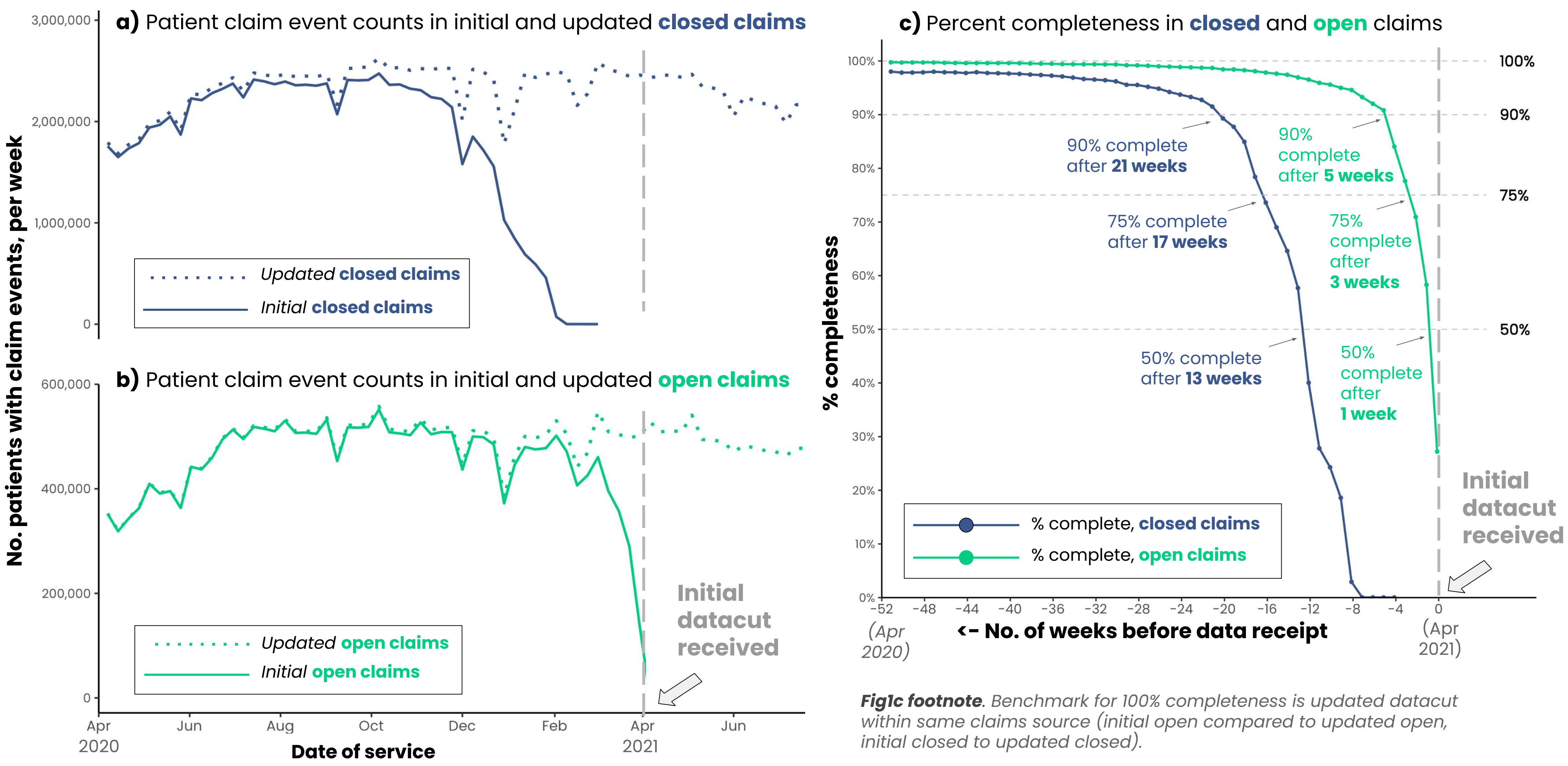

Fig1c footnote. Benchmark for 100% completeness is updated datacut within same claims source (initial open compared to updated open, initial closed to updated closed).

Table 1. Percent completeness by time since data receipt, all-cause and COVID-19 (U07.1) claim events

Assessed Apr 2020 to Apr 2021 among cohort with N=10.7 million patients

	% Completeness, Closed vs Open Claims		
No. weeks before initial data receipt	All medical settings	Outpatient	Inpatient
All-cause medical claim events			
1 week	0% vs 58%	0% vs 59%	0% vs 38%
2 weeks	0% vs 71%	0% vs 72%	0% vs 51%
4 weeks	<1% vs 84%	<1% vs 85%	0% vs 62%
8 weeks	3% vs 95%	3% vs 95%	2% vs 90%
12 weeks	40% vs 97%	40% vs 97%	26% vs 95%
24 weeks	94% vs 99%	94% vs 99%	92% vs 98%
1 year (52 weeks)	98% vs 100%	98% vs 100%	97% vs 100%
COVID-19 (U07.1) claim events			
1 week	0% vs 45%	0% vs 51%	0% vs 26%
2 weeks	0% vs 61%	0% vs 68%	0% vs 37%
4 weeks	0% vs 65%	0% vs 77%	0% vs 38%
8 weeks	1% vs 92%	1% vs 92%	2% vs 90%
12 weeks	30% vs 94%	30% vs 94%	32% vs 94%
24 weeks	92% vs 98%	92% vs 98%	92% vs 98%
1 year (52 weeks)	97% vs 99%	97% vs 100%	97% vs 98%

Figure 2. Data latency trends in closed and open claims sources, COVID-19 (U07.1) claim events

Trends plotted Apr 2020 to Jun 2021 for cohort with N=10.7 million patients

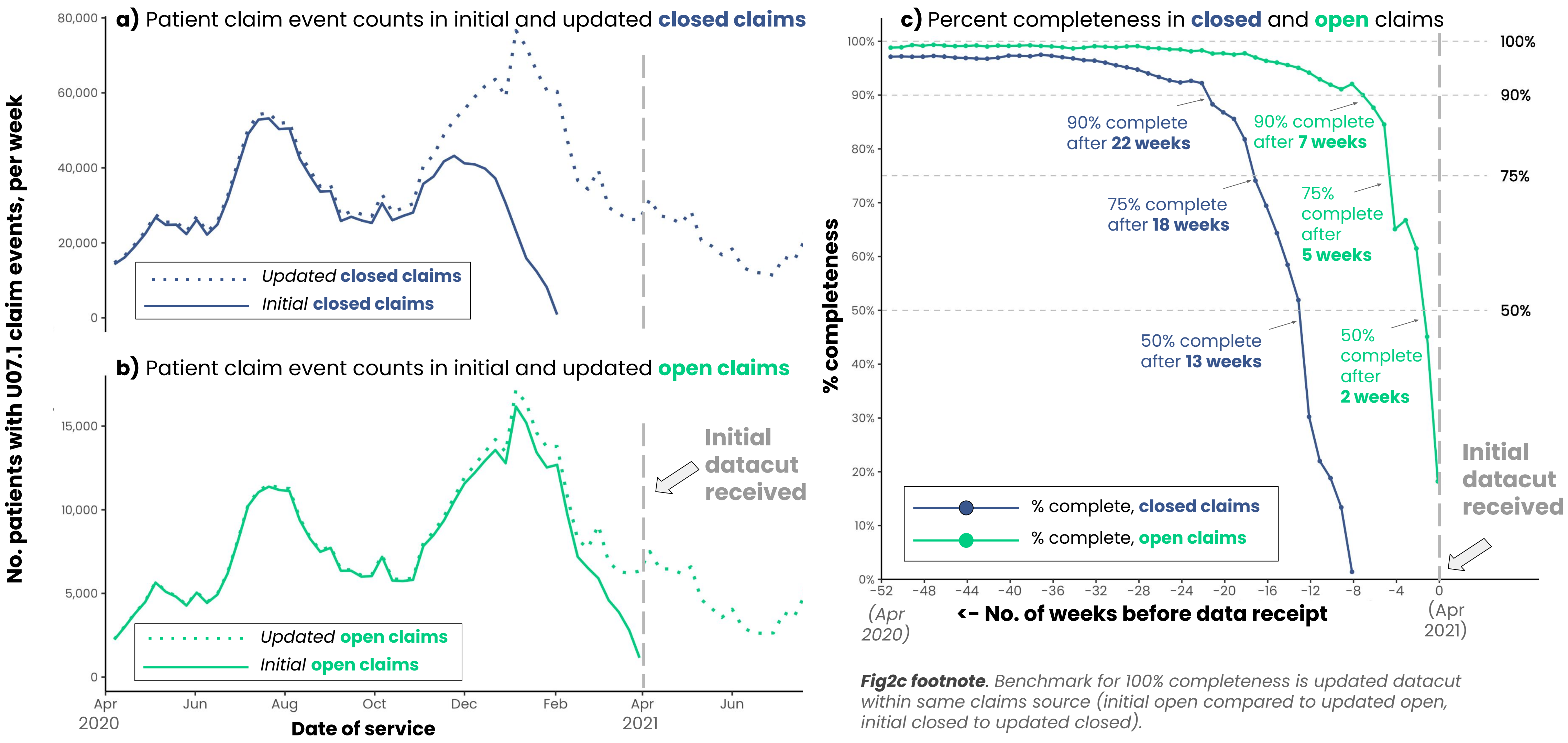

Fig2c footnote. Benchmark for 100% completeness is updated datacut within same claims source (initial open compared to updated open, initial closed to updated closed).

Table 2. Cumulative weekly claim event counts, all-cause and COVID-19 (U07.1) claim events

Assessed Apr 2020 to Apr 2021 among cohort with N=10.7 million patients

	Closed Claims	Open Claims	Open / Closed (%)
All-cause medical claim events			
All medical settings	121,344,554	24,964,180	20.6%
Outpatient	120,433,871	23,768,678	19.7%
Inpatient	4,872,772	1,981,605	40.7%
COVID-19 (U07.1) claim events			
U07.1 all settings	2,060,723	441,154	21.4%
U07.1 outpatient	1,900,810	361,474	19.0%
U07.1 inpatient	290,073	102,262	35.3%

***Table 2 footnote.** For each claim event type, maximum of 1 event per patient per week. Cumulative event counts are from the updated datacut.*

Key takeaway from Table 2: Using a rough approximation for open claims observability (requirement for ≥ 1 open claim during study period), open claims captured only ~20% of total closed claim events. Open claims capture within this study cohort was higher for inpatient (~41%) relative to outpatient events (~20%).

Conclusions

Reduced latency for open relative to closed claims; outpatient relative to inpatient claims.
Open claims in this real-world dataset were available more quickly than closed claims, with >75% of claim events available within 3 weeks for open claims versus 17 weeks for closed claims. For both open and closed sources, time lag for outpatient claims was shorter than inpatient claims. Lag for all-cause claim events was slightly shorter than COVID-19-specific claim events.

Open claims provided reliable estimates of how measures were changing over time, but underestimated the magnitude of overall population measures.
During the 4-6 month lag period required for closed claim processing, week-by-week trends in open claim measures tracked closely to (updated) closed claims trends. However the absolute magnitude of open and closed-based measures differed, with open claims offering only a partial view of closed claim events.

Refined definitions for open claims observability (more specific denominator populations) may yield improved population-level estimates.
This study evaluated completeness by comparing claims within the same data sources at different time points, which does not account for gaps in observability (open claims lack enrollment files), nor for agreement in pre-adjudicated versus adjudicated claims content. Further research is needed to describe the validity of open claims on these dimensions, and to demonstrate when and how such sources can be used for public health surveillance and evidence generation.

AETION

Disclosures

This project was unfunded. ARW, EMG, KU, UC, MG, ER, & NMG hold stock options or equity in Aetion Inc and were full-time employees of Aetion when this research was conducted. RK is an employee of HealthVerity, Inc.