Benchmarking Disease Prevalence in a Large Scale Electronic Health Record Data Network: An Assessment of Chronic and Rare Disease in the United States in 2023

RWD105

Amanda M. Moore, PharmD, PhD[1] , Matt Scranton, MSc[1] , E. Susan Amirian, PhD[1] , Jeffrey Brown, PhD[1]
[1]TriNetX, LLC, Cambridge, MA, USA

BACKGROUND

The generalizability and transportability of studies conducted using real-world data depend on the data representativeness, or how well the sampled population resembles the broader target population.

OBJECTIVE

To characterize the representativeness of the TriNetX Dataworks-USA electronic health records (EHR) network to the general United States (US) healthcare-seeking population by benchmarking the estimated 2023 prevalences of common chronic conditions and selected rare diseases against published literature sources.

METHODS

- TriNetX Dataworks-USA is a federated research network comprised of de-identified EHR data sourced directly from 71 healthcare organizations for over 120 million patients and mapped to standardized terminologies. Data in the network is refreshed on average every 2 to 4 weeks. The results of these analyses reflect the network data as of April 2025.
- The study cohort included all adult patients (aged ≥18 years) with a recorded encounter in the TriNetX Dataworks-USA network from January 1, 2022 to December 31, 2023.
- The prevalences of the 10 most common and costly chronic diseases, as reported by the CDC Chronic Disease Indicators (CDI) and National Health Interview Survey (NHIS)[1,2], and selected rare diseases were estimated by dividing the number of patients with ≥1 or ≥2 relevant diagnosis codes before the end of 2023 by the number of patients with recorded activity from 2022-2023 using the TriNetX LIVE™ platform.
- TriNetX-estimated prevalences were presented against prevalences reported in other published population-based studies.

RESULTS

- As of April 2025, over 36 million patients had a recorded encounter in the TriNetX Dataworks-USA EHR network from 2022-2023 (Table 1). Patients had an average of 1,199 clinical facts in their medical record (Fig. 1) and over half of the cohort (56.2%) had more than 5 years of medical history. 95% of clinical facts spanned the time window between 2010-2024.

RESULTS (continued)

- Prevalence of rare diseases generally overestimated when compared to literature estimates (Fig. 2).
- Prevalence of common conditions compares to CDC estimates, including asthma (TriNetX 5.3-8.5%, CDC 8.7-9.9%), coronary heart disease (TriNetX 5.2-7.7%, CDC 4.9%), and diabetes (TriNetX 9.1-11.8%, CDC 9.6-12.1%) (Fig. 3). Prevalence in asthma stratified by age, sex, and race/ethnicity follow similar patterns to CDC reported estimates (Fig. 4).

Table 1. Cohort baseline patient characteristics

Encounter in 2022-2023 (N=36,576,341)			
CHARACTERISTIC[1]	n (%)	CHARACTERISTIC[1]	n (%)
Age at index (years), n (%)		Race, n (%)	
Mean/SD	50 (19)	White	22,416,711 (61.3)
18-44	15,260,002 (41.7)	Black or AA	5,520,013 (15.1)
45-64	12,165,777 (33.3)	Asian	1,704,607 (4.7)
≥65	10,180,032 (27.8)	NH/PI	242,323 (0.7)
Sex, n (%)		AI/AN	168,155 (0.5)
Male	15,093,358 (41.3)	Other	1,774,733 (4.9)
Female	20,357,064 (55.7)	Unknown	4,479,800 (12.2)
Unknown	1,125,789 (3.1)	Ethnicity, n (%)	
US geographic distribution, n (%)		Hispanic or Latino	3,075,850 (8.4)
Northeast	10,319,706 (28.2)	Not Hispanic or Latino	22,287,608 (60.9)
Midwest	7,368,389 (20.1)	Unknown	11,212,884 (30.7)
South	14,111,199 (38.6)		
West	3,612,303 (9.9)		
Unknown	1,164,744 (3.2)		

[1]Percentages may not reach or exceed 100% due to patient obfuscation and rounding on the TriNetX LIVE™ platform

Figure 1. Average number of clinical facts, per patient

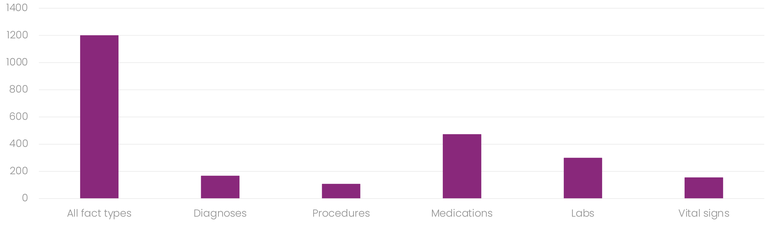


Figure 2. Prevalences of selected rare diseases, by source

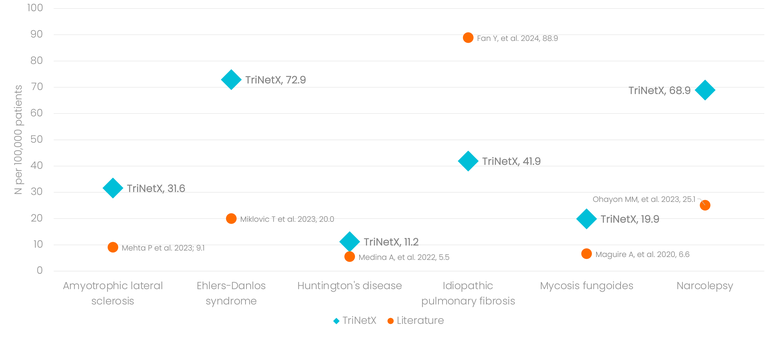


Figure 3. Prevalences of common chronic conditions, by source

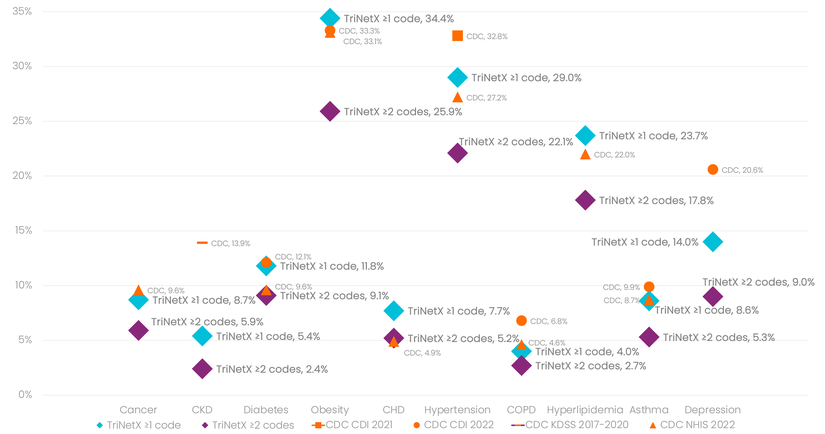
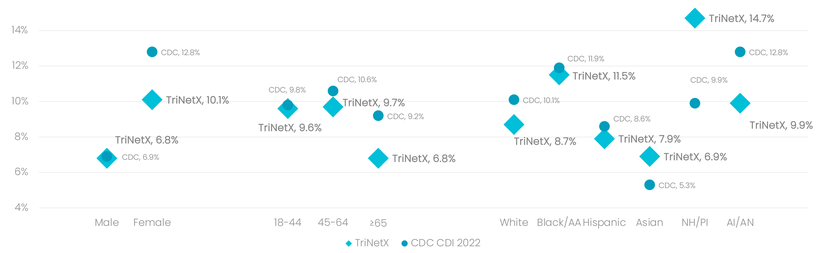


Figure 4. Asthma prevalence, by sex, age, race/ethnicity



CONCLUSION

- The prevalences of common chronic diseases observed in the TriNetX Dataworks-USA network are consistent with CDC estimates in the healthcare-seeking US population.
- Given that the network includes many academic medical centers, the prevalences of rare diseases may be overrepresented in the network.
- Limitations of this research include differential definitions/coding schema used to identify conditions. Diagnoses were identified primarily by ICD-10-CM diagnosis codes and did not include lab values (e.g., eGFR) or medications. The CDC CDI and NHIS are survey-based for self-reported responses, which may limit their generalizability.

REFERENCES
1. Centers for Disease Control and Prevention (CDC). Chronic Disease Indicators (CDI). https://www.cdc.gov/cdi/index.html. Accessed 9 Apr 2025. 2. CDC. National Health Interview Survey (NHIS). https://www.cdc.gov/nchs/nhis/index.html. Accessed 9 Apr 2025. 3. CDC. Kidney Disease Surveillance System (KDSS). https://nccd.cdc.gov/ckd. Access 9 Apr 2025. 4. Mehta P et al. *Amyotroph Lateral Scler Frontotemporal Degener.* 2023: 1-7. 5. 5. Miklovic T et al. *StatPearls.* 2023. 6. Medina A et al. *Mov Disord.* 2022; 37: 2327-2335. 7. Fan Y et al. *Chest.* 2022; 166: A3282. 8. Maguire A et al. *Acta Derm Venerol.* 2020; 100: 5631. 9. Ohayon MM et al. *Sleep Med X.* 2023; 6: 100095.

Download Poster PDF

ISPOR 2025 | Montréal, QC, Canada
Design services were provided by Kelly Marsocci and were funded by TriNetX, LLC.