Characterization of Confounding Covariate Data Capture in the U.S. Food and Drug Administration's Sentinel Distributed Database

Presented at the 2025 ISPE Annual Meeting

Jummai Apata[1], Ashish Rai[2], Diane Emerton[3], Youjin Wang[1], Wei Liu[1]

[1]Center for Drug Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA;
[2]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA, USA;
[3]Department of Population Medicine, Harvard Pilgrim Health Care Institute, Boston, MA, USA

BACKGROUND

- The Sentinel System is FDA's active post-market medical product safety surveillance system for assessing the use and safety of regulated medical products.
- The Sentinel Distributed Database (SDD) comprises mainly United States (U.S.) administrative claims data provided by Data Partners (health insurance providers, healthcare systems, academic medical centers).
- A known limitation of claims data is its inadequate capture of key common confounding covariates, which could lead to biased study estimates potentially negatively impacting the validity of study findings.

OBJECTIVE

To evaluate the capture of selected confounding covariates, including vitals data, in the SDD over time, by assessing their incidence and prevalence.

METHODS

- **Study design:** Descriptive study characterizing prevalence and incidence of conditions often considered to be confounders in pharmacoepidemiology studies.
- **Study period:** January 1, 2010 to December 31, 2023.
- **Data source:** 14 Sentinel Data Partners (DPs) contributed.
- **Cohort eligibility:** Individuals of all ages (except for calculatable Body Mass Index [BMI] cohorts which only included 18+ years) who had at least 365 days of continuous prior enrollment in both medical and drug plans.
- **Index dates:** Dates of all qualifying occurrences of cohort-defining codes in claims or measurements in vitals data.
- **Incident confounding covariates:** Episodes were excluded if the observed confounding condition was present in the prior 365 days.
- **Confounding covariate measures:** Cohorts' incident and prevalent confounding conditions included **obesity, overweight, smoking, alcohol and drug abuse/dependence** (described using diagnosis codes, procedure codes and National Drug Codes [NDC]), **height and weight for BMI, blood pressure [BP]** (measurements from vitals data), and **race**.
- The number of qualifying occurrences of the cohort-defining variables were stratified by year and sex.

RESULTS

- About 260 million members from 14 DPs overall, including 8.3 million members from seven DPs contributing vitals data, were eligible for cohort entry (**Table 1**).

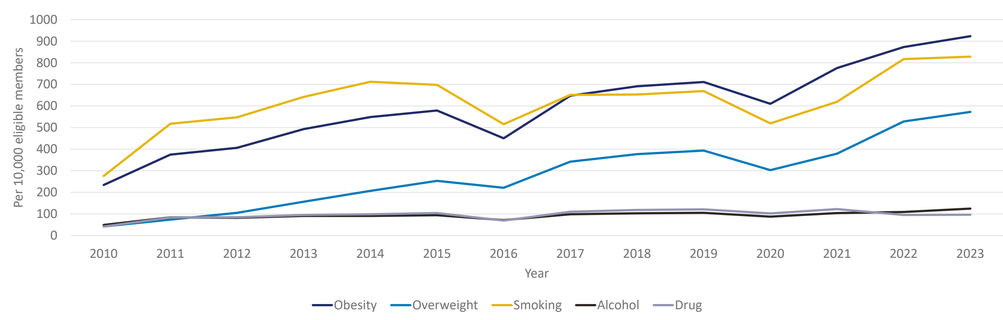
Table 1. Summary of Eligible Member Cohorts in the SDD From January 1, 2010 to December 31, 2023

	Members with an Index Date	Eligible Members[1]
Incidence of Confounding Conditions Observed in the SDD (diagnosis, procedure, NDC codes)		
Incident obesity	41,808,406	256,076,279
Incident overweight	24,525,780	261,059,251
Incident smoking	40,664,965	256,469,018
Incident alcohol abuse or dependence	8,161,856	261,760,913
Incident drug abuse or dependence	8,826,792	262,067,942
Prevalence of Confounding Conditions Observed in the SDD (diagnosis, procedure, NDC codes)		
Prevalent obesity	48,379,256	262,961,230
Prevalent overweight	26,351,868	262,961,230
Prevalent smoking	46,526,773	262,961,230
Prevalent alcohol abuse or dependence	9,343,054	262,961,230
Prevalent drug abuse or dependence	9,721,582	262,961,230
Prevalence of Confounding Conditions Observed in the SDD (demographic, vitals data)		
Has non-missing value for reported race	118,798,367	136,735,309
Has diagnosis code that indicates a range of BMI	64,912,835	262,961,230
Has non-missing values for both height and weight within +/- 6 months[2]	3,548,406	6,079,044
Has non-missing value for diastolic blood pressure[2]	5,091,900	8,275,067
Has non-missing value for systolic blood pressure[2]	5,091,957	8,275,067

[1]Eligible Members and Member-Years are reflective of the number of patients that met all cohort entry criteria on at least one day during the query period.
[2]Only from 7 Data Partners contributing vitals data
N/A, Not Applicable

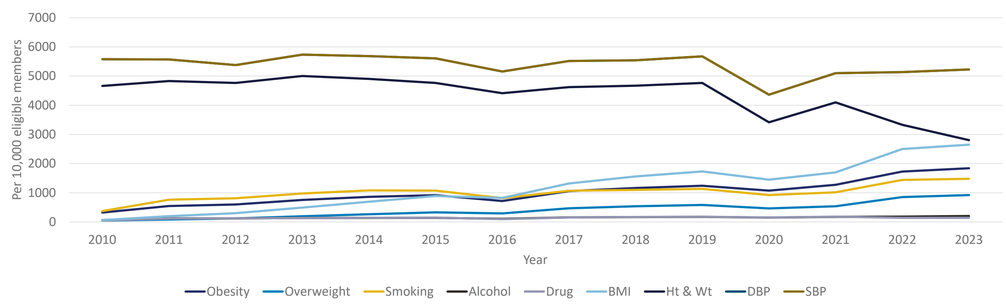
- The incidence and prevalence of obesity, overweight, smoking, and the prevalence of BMI from diagnosis/procedure/NDC codes steadily increased over time with slight dips occurring in 2015-2016 and 2020-2021 (**Figure 1, Figure 2**).
- The capture of confounding covariates obtained from diagnosis/procedure/NDC codes peaked in 2023 with the highest observed values being incidence of 923 per 10,000 eligible members for obesity (**Figure 1**), and prevalence of 2,649 per 10,000 eligible members for BMI (**Figure 2**).
- Among the confounding covariates recorded in this study, drug and alcohol abuse/dependence had the lowest rates. Both the incidence and prevalence of these conditions remained below 204 per 10,000 eligible members throughout the study period (**Figure 1, Figure 2**).

Figure 1. Incidence of Confounding Covariates in the SDD From Diagnosis/Procedure/NDC Codes



The prevalence of diastolic/systolic BP and BMI obtained from vitals data peaked in 2013 at 5,735 and 5,003 per 10,000 eligible members, respectively, declining afterward (**Figure 2**).

Figure 2. Prevalence of Confounding Covariates in the SDD From Diagnosis/Procedure/NDC Codes and Vitals Data

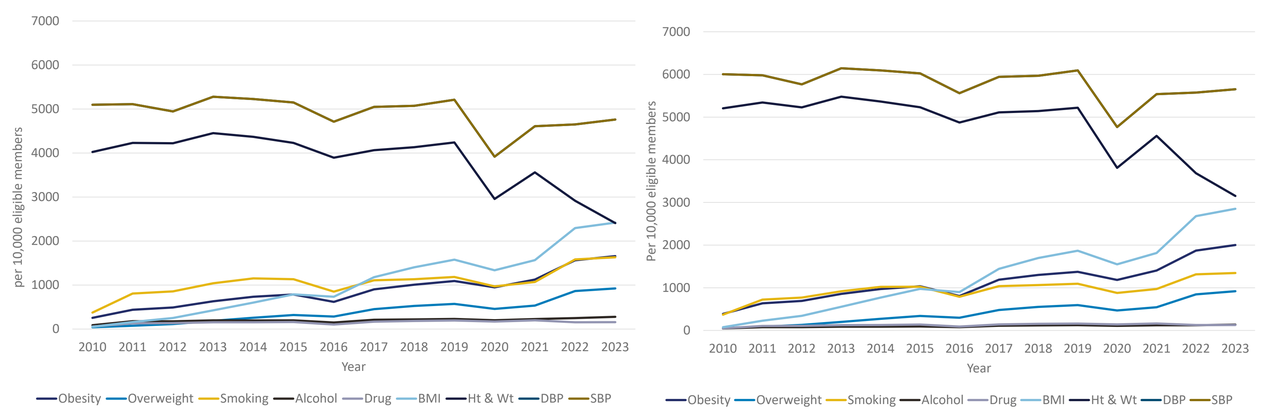


Ht & Wt (Height and Weight), DBP (Diastolic Blood Pressure), and SBP (Systolic Blood Pressure), were obtained from vitals data while all other covariates were obtained from diagnosis, procedure or NDC codes.

A slightly higher prevalence of alcohol and drug abuse/dependence was observed in males compared to females (**Figure 3, Figure 4**).

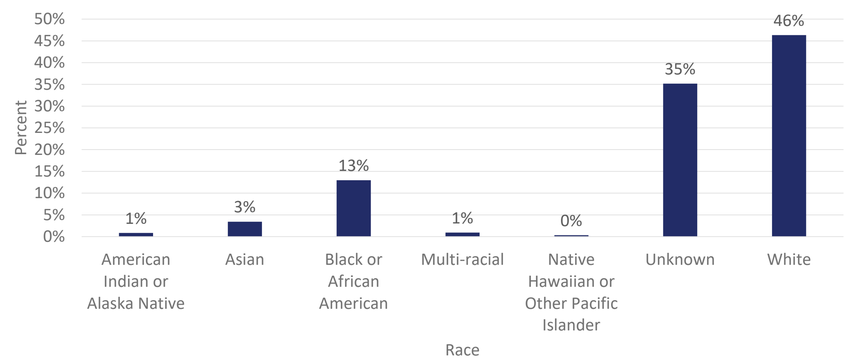
Figure 3. Prevalence of Confounding Covariates in Males in the SDD

Figure 4. Prevalence of Confounding Covariates in Females in the SDD



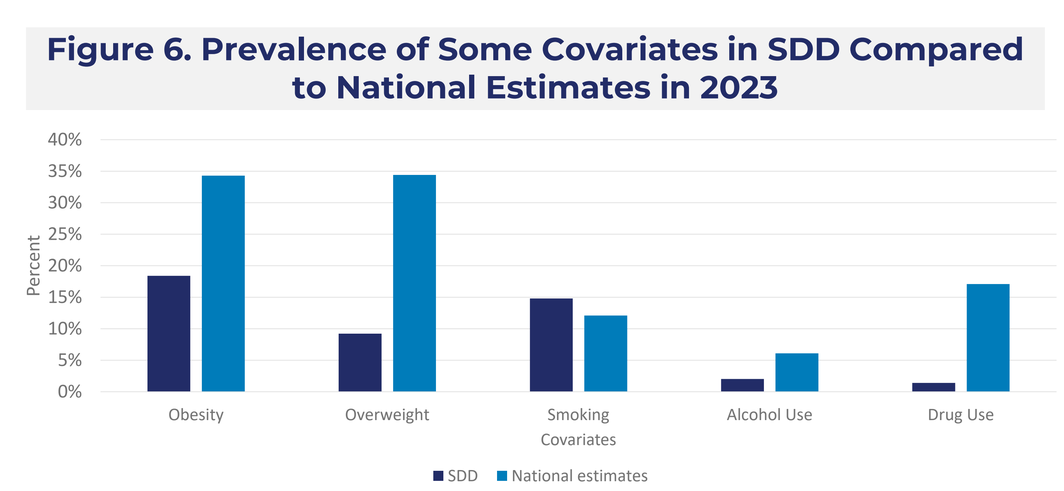
Ht & Wt (Height and Weight), DBP (Diastolic Blood Pressure), and SBP (Systolic Blood Pressure), were obtained from vitals data while all other covariates were obtained from diagnosis, procedure or NDC codes.

Figure 5. Members With Reported Race on Their Index Date



Race was unknown in 35% of members with claims data from January 2020-December 2021 (**Figure 5**).

Figure 6. Prevalence of Some Covariates in SDD Compared to National Estimates in 2023



National estimates for obesity, overweight, smoking (current smokers), and alcohol use (heavy drinking) were obtained from CDC Behavioral Risk Factor Surveillance System (BRFSS).
National estimate for drug use disorder (substance use disorder) data was obtained from Substance Abuse and Mental Health Services Administration (SAMHSA) National Survey on Drug Use and Health.
Definitions used in national estimates may not be directly comparable to definitions in the SDD.

CONCLUSION

- The capture of key confounding covariates in the SDD has improved over time; however, for most covariates, prevalence remains lower than expected given the national estimates (**Figure 6**), indicating they may still be insufficiently captured.
- In planning drug safety studies, each data source should be carefully evaluated for fitness for use, relevance, reliability, etc. to ensure it can adequately address the specific study question and account for these key confounding covariates.
- Linking claims data with electronic health records (EHR) is an increasingly common practice aimed at improving the capture of confounding variables and strengthening the validity of observational research when combined with innovative analytic methods.

ACKNOWLEDGEMENTS/DISCLOSURES

This project was supported by Task Order 75F40124F19014 under Master Agreement 75F40119D10037 from the U.S. Food and Drug Administration (FDA).

The contents are those of the authors and do not necessarily represent the official views of, nor an endorsement by, FDA/HHS or the U.S. Government.

A.R. and D.E. are employees of HPHCI, an organization which conducts work for government and private organizations, including pharmaceutical companies.

Many thanks are due to the Sentinel Data Partners who provided data used in the analysis.