

Assessing the Concordance Between Self-Reported Nicotine Product Use and EMR Diagnosis Codes in the All of Us Research Program

David Goldfarb¹, Mary Catherine Minnig¹, Shivani Aggarwal^{1,2}
¹Landmark Science, Inc., New York, NY, USA ²Landmark Science, Inc., Los Angeles, CA, USA.



Background & Rationale

- Accurate ascertainment of nicotine and tobacco use (e.g., cigarettes, e-cigarettes, smokeless tobacco) is critical for respiratory and cancer outcomes research.
- However, capturing nicotine exposure using secondary data sources is challenging because standardized ontologies (e.g., ICD-9/10, SNOMED) capturing nicotine/tobacco use are routinely used but have limited granularity in reflecting true exposure patterns.
- Misclassification of nicotine exposure can bias effect estimates and attenuate observed associations in epidemiologic analyses.
- The All of Us (AoU) Research Program provides an opportunity to compare nicotine use captured via self-report and linked EMRs

Objective: To evaluate concordance between diagnostic codes for nicotine/tobacco use and self-reported survey data in the All of Us database.

Methods

Study Design

- Cross-sectional analysis of AoU data among participants with completed self-report questionnaires and linked EMR records.

Study Population

- Adult participants (≥18 years) who completed lifestyle surveys and had linked EMR data.
- Diagnosis codes from the EMR were required to be within 1 year of the self-report lifestyle questionnaires for the primary analytic cohort (n=33,665).

Data Extraction & Exposure Ascertainment

Data Source and Extraction

- Data were extracted from the AoU Registered Tier Dataset, v8 using BigQuery within the All of Us Researcher Workbench.
- Analyses were conducted in RStudio/Jupyter using R.

Tobacco Use Ascertainment

- Nicotine/tobacco use in EMRs was identified using standardized ontology codes from:
 - ICD-9-CM
 - ICD-10-CM
 - SNOMED-CT
- Self-report nicotine and tobacco use was evaluated from lifestyle questionnaires.

Analysis

- We evaluated concordance between self-reported nicotine/tobacco use and EMR diagnosis codes among the primary analytic cohort overall (Figure 2) and by product type (Table 2) including cigarettes, e-cigarettes, cigars, hookah, and smokeless tobacco.
- We assessed proportion of participants with self-reported use but no EMR diagnosis codes (n=90,637; Table 3).
- Descriptive analyses were stratified by demographic characteristics (Table 1 & Figure 2).
- Sensitivity analyses included participants with tobacco/nicotine-related EMR diagnosis codes recorded (n=283,773; Table 4).

Key Findings

- Among 571,624 participants who completed lifestyle surveys, 62,466 (10.9%) had at least one nicotine/tobacco-related diagnosis code recorded in the EMR and a total of 33,665 (5.9%) participants had ≥1 diagnosis code within a year of the survey date (Figure 1).
- Many unique nicotine/tobacco-related diagnosis codes were found in participants' EMRs. The most common were F17.210 (Nicotine dependence, cigarettes, uncomplicated) and F17.200 (Nicotine dependence, unspecified, uncomplicated). These codes accounted for 81% of all nicotine/tobacco-related entries (Figure 3) and were recorded for 20,676 (61.4%) and 17,817 (52.9%) of participants in the primary analytic cohort, respectively (data not shown).
- Self-reported use patterns and patterns of survey underreporting differed by age group, sex/gender, and race/ethnicity (See call-outs in Figure 5).
- Overall concordance between EMR nicotine/tobacco-related diagnosis codes and any survey-reported nicotine use exceeded 90% for most diagnosis codes (Figure 4). Concordance was lowest for codes related to tobacco use during pregnancy and highest for F17.2 (Nicotine Dependence; data not shown).
- Concordance between EMR diagnosis codes and survey responses differed by product type. Cigarette use had high concordance with cigarette-related codes. Smokeless product use had moderate concordance with chewing tobacco codes. E-cigarette use had poor concordance with non-specific nicotine codes, since no codes explicitly captured e-cigarette use. (Table 2)
- A substantial proportion of participants reported nicotine/tobacco use but had no corresponding nicotine/tobacco diagnosis codes recorded in their EMRs either within a year of their survey (Table 3) or any time before or after their survey (Table 4).

Limitations

- Efforts have been made to triangulate use patterns. However, due to social desirability bias, nicotine/tobacco use may be underreported in both the EMR and the self-reported surveys, and EMR documentation practices may vary.
- Limiting the analytic cohort to participants with both survey and EMR data may overestimate concordance since patients with both data types may be more engaged in healthcare or have the need to visit a doctor more frequently.
- Currently, nicotine and tobacco use questions are not administered longitudinally, limiting the ability to capture product switching or multi-product use over time.

Why is this Research Important?

- Accurate capture of nicotine/tobacco exposure is essential for studies evaluating respiratory disease and cancer risk.
- These findings demonstrate that the reliance on standardized nicotine/tobacco dependence diagnosis codes alone may: 1) underestimate nicotine exposure, particularly for emerging nicotine products such as e-cigarettes and nicotine pouches, and 2) fail to ascertain the specifics of multi-product use.
- Researchers using real-world data sources to measure nicotine exposure should consider triangulating with multiple data sources, including self-reported survey data (where available) and unstructured EMR data, to maximize accurate exposure assessment.

Results

Figure 1. Attrition Diagram for Participants Enrolled in the All of Us Research Program

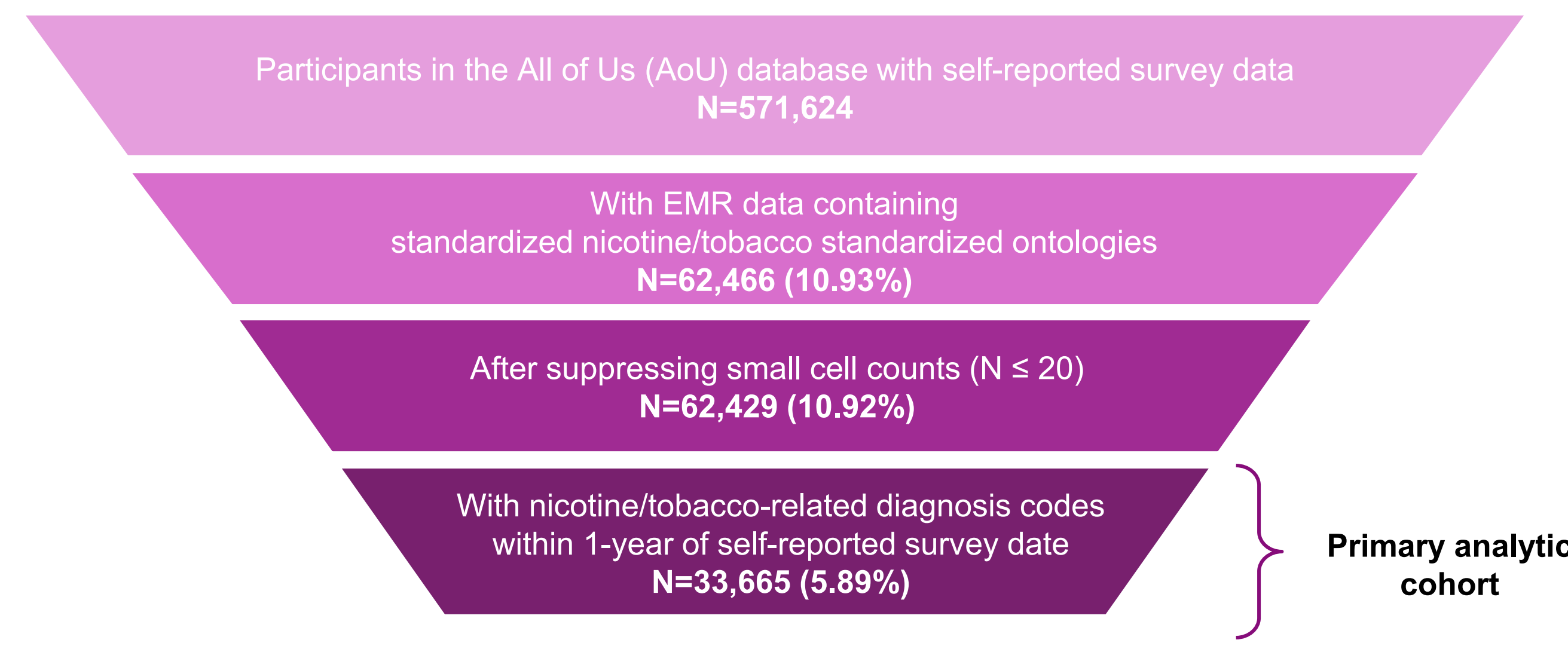


Table 1. Baseline Demographic Characteristics of the Primary Analytic Cohort

	Self-reported use among participants with diagnosis codes within 1-year of survey date						
	Overall N=33,665 n (%)	Multi-Product Use N=20,066 n (%)	Multi-Use/ No Cigarettes N=1,109 n (%)	Cigarettes Only N=8,855 n (%)	E-Cigarettes Only N=228 n (%)	Smokeless Only N=122 n (%)	No Tobacco/ Nicotine Use N=2504 n (%)
Age (years)							
<30	870 (2.6%)	452 (2.3%)	150 (13.5%)	61 (0.7%)	58 (25.4%)	2 (1.6%)	110 (4.4%)
30-44	7,252 (21.5%)	4,922 (24.5%)	436 (39.3%)	1021 (11.5%)	77 (33.8%)	25 (20.5%)	594 (23.7%)
45-59	10,049 (29.8%)	6,322 (31.5%)	300 (27.1%)	2381 (26.9%)	40 (17.5%)	46 (37.7%)	739 (29.5%)
≥60	15,494 (46%)	8,370 (41.7%)	223 (20.1%)	5392 (60.9%)	53 (23.2%)	49 (40.2%)	1061 (42.4%)
Sex							
Male	15,921 (47.3%)	10161 (50.6%)	623 (38.6%)	3477 (39.3%)	71 (31.1%)	99 (81.1%)	1062 (42.4%)
Female	16,880 (50.1%)	9446 (47.1%)	453 (25.5%)	5179 (58.5%)	137 (60.1%)	23 (18.9%)	1312 (52.4%)
Other/Missing	864 (2.6%)	459 (2.3%)	33 (3%)	199 (2.2%)	20 (8.8%)	0 (0%)	130 (5.2%)
Race/Ethnicity							
White	14,906 (44.3%)	10621 (52.9%)	428 (38.6%)	2973 (33.6%)	80 (35.1%)	61 (50.0%)	521 (20.8%)
Black	9,712 (28.8%)	4435 (22.1%)	283 (25.5%)	3523 (39.8%)	65 (28.5%)	21 (17.2%)	1074 (42.9%)
Asian	228 (0.7%)	136 (0.7%)	14 (1.3%)	50 (0.6%)	6 (2.6%)	0 (0%)	18 (0.7%)
Other/Unknown	8,819 (26.2%)	4874 (24.3%)	384 (34.6%)	2309 (26.1%)	77 (33.8%)	40 (32.8%)	891 (35.6%)

Counts and percentages may not sum to totals because certain product categories including cigar only (n=624) and hookah only (n=157) are not presented.

Figure 5. Self-Reported Use Patterns by Demographic Characteristics of the Primary Analytic Cohort

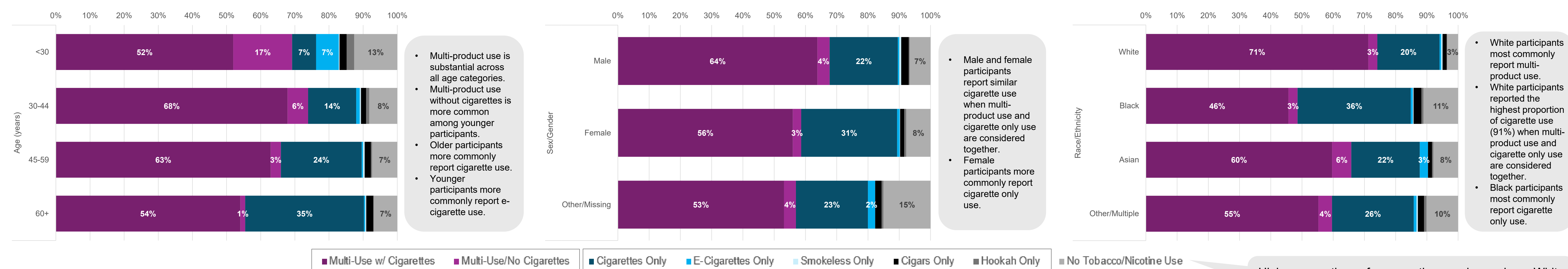


Table 2. Concordance of Diagnosis Codes with Self-Reported Use By Product

Code: Concept Name	Total N	% Agreement with Self-Reported Product Use				
		Cigarettes	E-Cigarettes	Smokeless	Cigars	Hookah
No product specified or implied						
F17.200: Nicotine dependence, unspecified, uncomplicated	17,817	87%	39%	18%	48%	21%
56294008: Nicotine dependence	1,396	91%	31%	15%	59%	19%
F17.201: Nicotine dependence, unspecified, in remission	432	88%	27%	16%	48%	13%
F17.209: Nicotine dependence, unspecified, with unspecified nicotine-induced disorders	122	91%	48%	29%	55%	26%
133042: Nicotine Dependence	81	89%	27%	0%	38%	0%
30310000: Nicotine-induced organic mental disorder	37	95%	0%	0%	57%	0%
F17: Nicotine dependence	27	82%	0%	0%	0%	0%
724697004: Harmful pattern of use of nicotine	27	93%	0%	0%	0%	0%
F17.2: Nicotine dependence	25	92%	0%	0%	0%	0%
F17.208: Nicotine dependence, unspecified, with other nicotine-induced disorders	25	88%	0%	0%	0%	0%
Tobacco-specific/No product specified or implied						
89765005: Tobacco dependence syndrome	191	86%	48%	21%	58%	20%
191889006: Tobacco dependence in remission	70	91%	0%	0%	51%	0%
305.1: Tobacco use disorder	2,905	90%	39%	18%	49%	22%
990: Toxic effect of tobacco	63	48%	0%	0%	38%	0%
Product-Specific						
Cigarettes						
F17.210: Nicotine dependence, cigarettes, uncomplicated	20,676	88%	39%	18%	48%	23%
F17.211: Nicotine dependence, cigarettes, in remission	525	89%	25%	16%	51%	14%
F17.219: Nicotine dependence, cigarettes, with unspecified nicotine-induced disorders	192	93%	35%	18%	50%	14%
F17.213: Nicotine dependence, cigarettes, with withdrawal	121	94%	43%	22%	48%	0%
F17.218: Nicotine dependence, cigarettes, with other nicotine-induced disorders	118	93%	39%	20%	61%	0%
16077091000119100: Tobacco dependence caused by cigarettes	43	98%	61%	0%	63%	0%
Smokeless Tobacco (Chewing Tobacco)						
F17.220: Nicotine dependence, chewing tobacco, uncomplicated	319	56%	30%	71%	60%	19%
F17.221: Nicotine dependence, chewing tobacco, in remission	29	79%	0%	0%	86%	0%
F17.229: Nicotine dependence, chewing tobacco, w/ unspecified nicotine-induced disorders	29	0%	0%	72%	0%	0%
Other tobacco product						
F17.290: Nicotine dependence, other tobacco product, uncomplicated	1,651	70%	51%	24%	61%	30%
F17.299: Nicotine dependence, other tobacco product, w/ unspecified nicotine-induced disorders	58	79%	36%	0%	64%	0%
F17.291: Nicotine dependence, other tobacco product, in remission	42	88%	0%	0%	62%	0%

Note: Darker colors indicate higher % agreement. Numbers will not sum to analytic cohort totals because diagnostic codes are not mutually exclusive and more than one code can be used per patient. Data on codes related to tobacco use in pregnancy not shown.

Figure 2. Types of Nicotine & Tobacco Products (Adapted from World Health Organization resources)

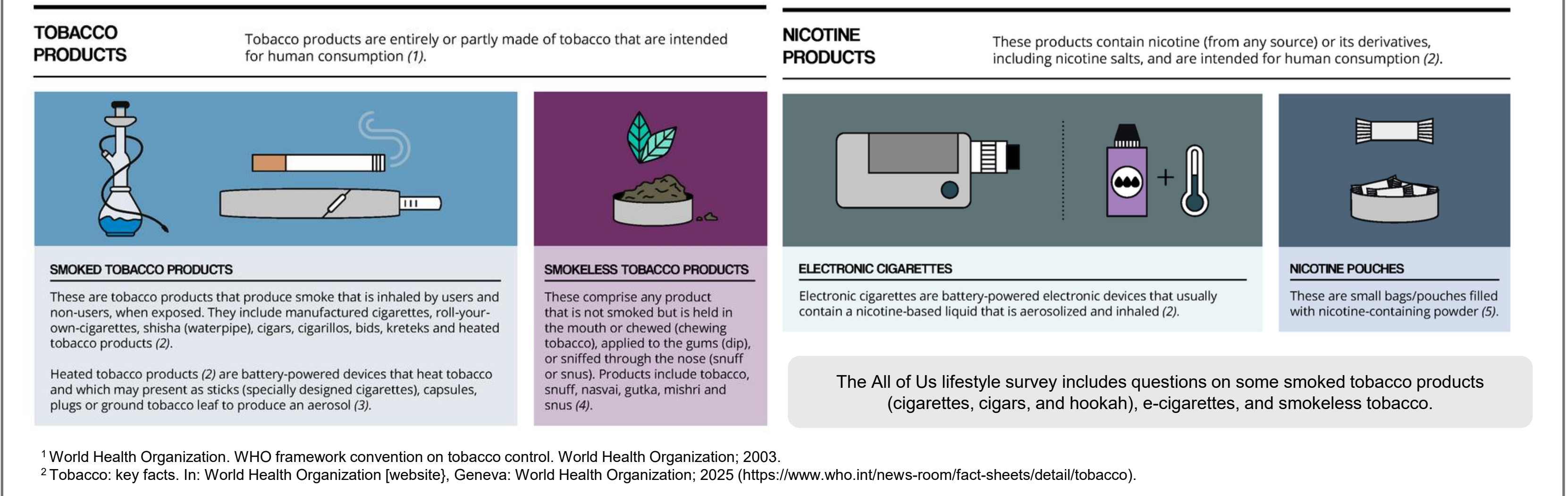


Figure 3. Top 10 Nicotine/Tobacco-Related Diagnosis Codes Present in the EMRs of the Primary Analytic Cohort

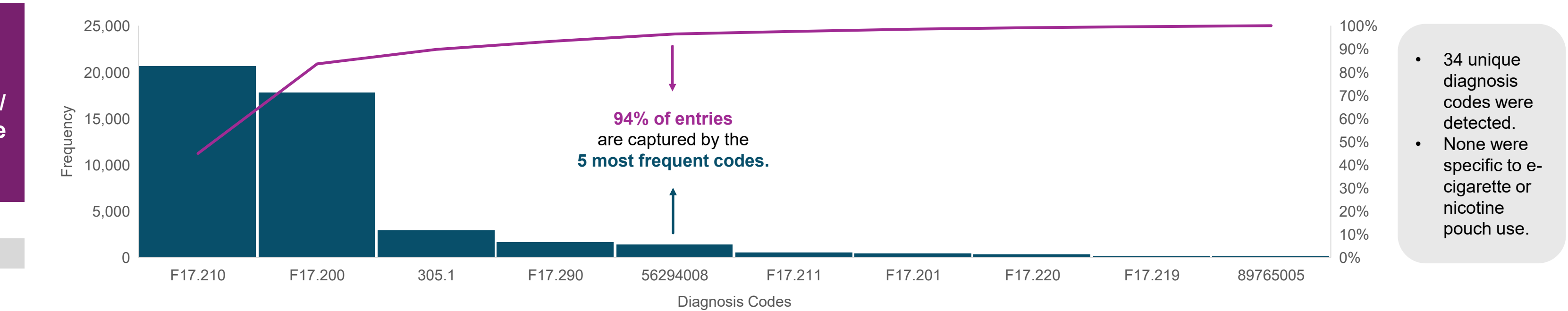


Figure 4. Concordance of Select Diagnosis Codes with Any Self-Reported Use

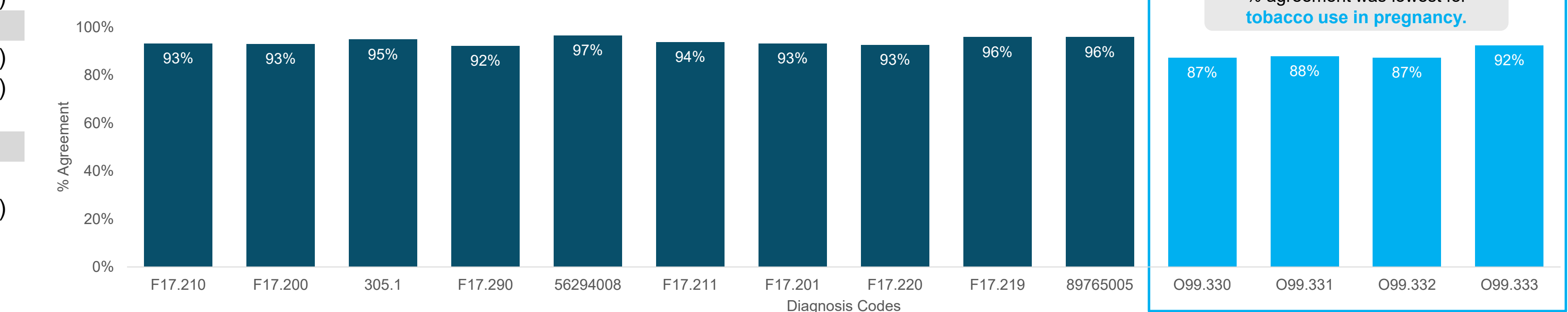


Table 3. Percent of Self-Reported Survey Responses by Product, Among Patients with No Nicotine/Tobacco Diagnosis Codes within a Year of Survey (N=90,637)

Product in Survey	Self-Reported Use		Prefer Not to Answer/ Don't Know/Skip
	Yes	No	
Cigarettes	31%	67%	2%
E-cigarettes	16%	83%	2%
Smokeless	9%	90%	1%
Cigars	35%	63%	2%
Hookah	18%	80%	2%

A substantial proportion of participants report nicotine/tobacco use but have no diagnosis codes.

The same patterns were seen when the 1-year window was relaxed.

"True Negatives"
Concordance between self-report of no use and no diagnosis codes within a year of the survey.

Table 4. Percent of Self-Reported Survey Responses by Product, Among Patients with No Nicotine/Tobacco Diagnosis Codes (N=283,773)

Product in Survey	Self-Reported Use		Prefer Not to Answer/ Don't Know/Skip
	Yes	No	
Cigarettes	30%	68%	2%
E-cigarettes	12%	86%	2%
Smokeless	7%	91%	2%
Cigars	34%	65%	2%
Hookah	16%	82%	1%