White Paper

Amplity Insights

HEOR/RWE

2023

Amplity HEALTH

Table of Contents

Amplity Insights 03

Overview 03

Natural Language Processing 06

Representativeness 08

Longitudinal Perspective 12

Compliance 13

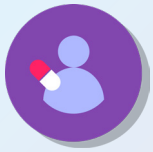
References 14

Amplity Insight's Data Attributes

Amplity Insight's database provides advantages relative to other commercially available data sources.

Large pool of unique patients allows for research on common and rare disease populations

The Amplity Insight's database is an open system of unstructured electronic medical transcription records for inpatient and outpatient care provided in the U.S. as part of routine clinical care. Each health care encounter transcription record could include initial office consultations, follow-up visits, urgent care visits, emergency department and hospital admissions and discharges, postoperative consultations, office notes, and referral letters. This database includes approximately 55 million records for 25 million patients who received care from 150,000 multi-specialty physicians across 50 states and 2 U.S. territories (Guam, Puerto Rico) between 2010 and 2022 and continues to grow with new patients and medical records added monthly.



25 million unique patients

55 million medical transcription records

150,000 multi-specialty physicians

* Totals correct at time of analysis, July of 2023. New data added monthly.

The Amplity Insight's database not only captures common specialties (e.g., cardiology, pulmonology/pulmonary) but also captures those specialties not commonly captured from other databases (e.g., dentistry, chiropractic, pharmacy service providers, rehabilitation services, etc.).

Examples of physician specialties are provided on Figure 1 and Table 1.

Figure 1. Amplity Database Percent Physician Specialty Distribution

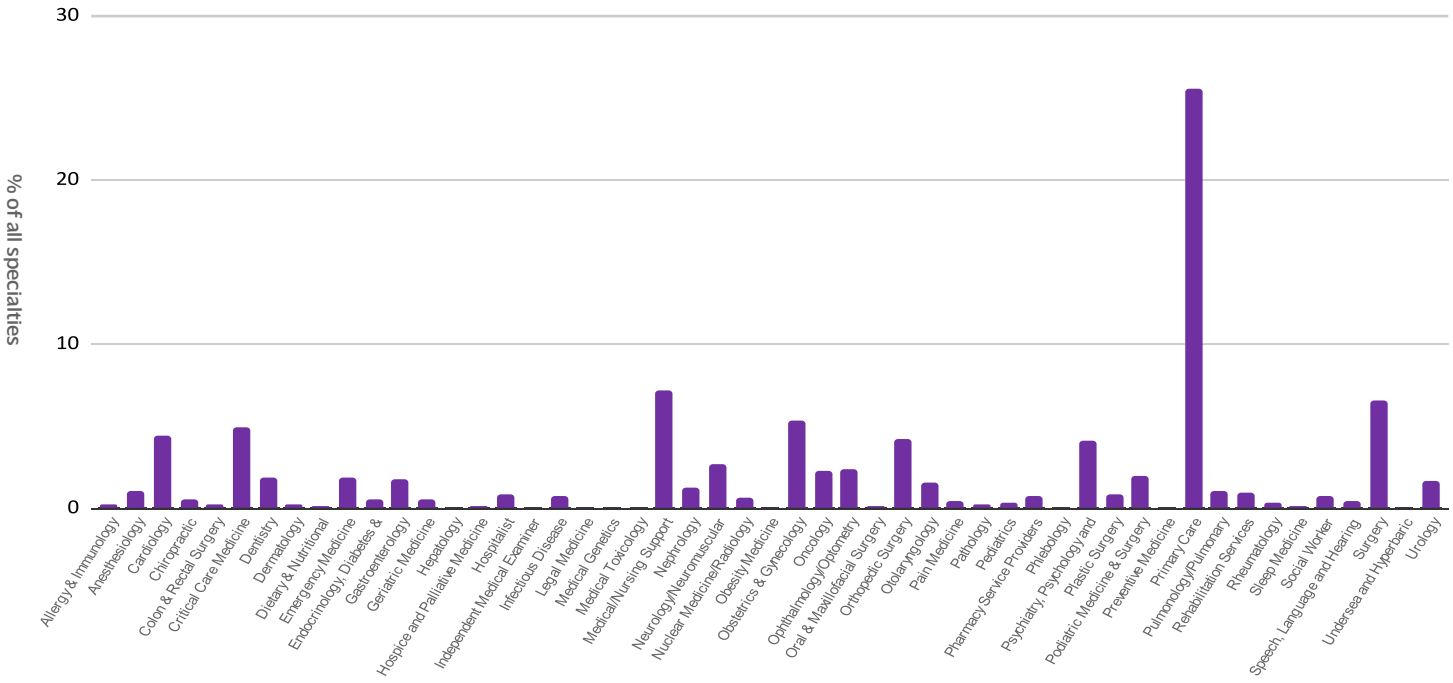


Table 1.

Physician Specialties	Amplity n	% of All Specialties
Allergy & Immunology	202	0.2
Anesthesiology	1,243	1
Cardiology	5,461	4.4
Chiropractic	644	0.5
Colon & Rectal Surgery	290	0.2
Critical Care Medicine	5,984	4.9
Dentistry	2,348	1.9
Dermatology	277	0.2
Dietary & Nutritional	120	0.1
Emergency Medicine	2,283	1.9
Endocrinology, Diabetes & Metabolism	648	0.5
Gastroenterology	2,174	1.8
Geriatric Medicine	645	0.5
Hepatology	23	0
Hospice and Palliative Medicine	118	0.1
Hospitalist	1,030	0.8
Independent Medical Examiner	13	0
Infectious Disease	851	0.7
Legal Medicine	33	0.03
Medical Genetics	34	0.03
Medical Toxicology	4	0.003
Medical/Nursing Support	8,919	7.2

Nephrology	1,595	1.3
Neurology/Neuromuscular Medicine	3,271	2.7
Nuclear Medicine/Radiology	774	0.6
Obesity Medicine	12	0.01
Obstetrics & Gynecology	6,545	5.3
Oncology	2,809	2.3
Ophthalmology/Optometry	2,946	2.4
Oral & Maxillofacial Surgery	115	0.1
Orthopedic Surgery	5,211	4.2
Otolaryngology	2,018	1.6
Pain Medicine	444	0.4
Pathology	258	0.2
Pediatrics	414	0.3
Pharmacy Service Providers	889	0.7
Phlebology	5	0.004
Psychiatry, Psychology and Behavioral health	5,045	4.1
Plastic Surgery	968	0.8
Podiatric Medicine & Surgery	2,489	2
Preventive Medicine	59	0.05
Primary Care	31,527	25.6
Pulmonology/Pulmonary	1,192	1
Rehabilitation Services	1,128	0.9
Rheumatology	333	0.3
Sleep Medicine	106	0.1
Social Worker	814	0.7
Speech, Language and Hearing Service Providers	439	0.4
Surgery	8,160	6.6
Undersea and Hyperbaric Medicine	14	0.01
Urology	2,105	1.7

Care Settings

5.7% of physician specialties are **inpatient providers**

90.8% of physician specialties are **outpatient providers**

2.0% of outpatient providers are **Emergency service providers**

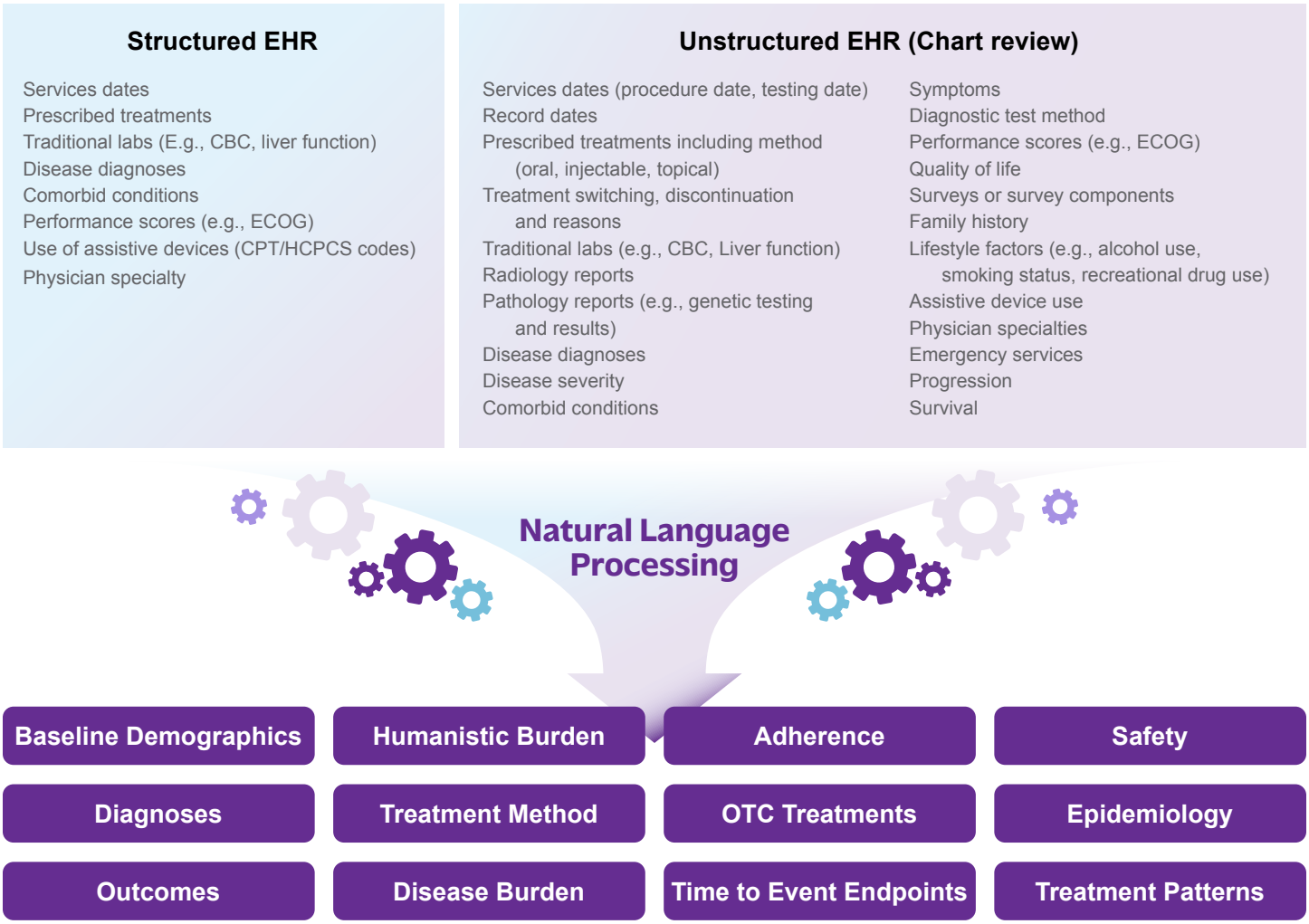
25.2% of physicians are affiliated with **academic centers**

74.8% of physicians are **community providers**

Natural Language Processing

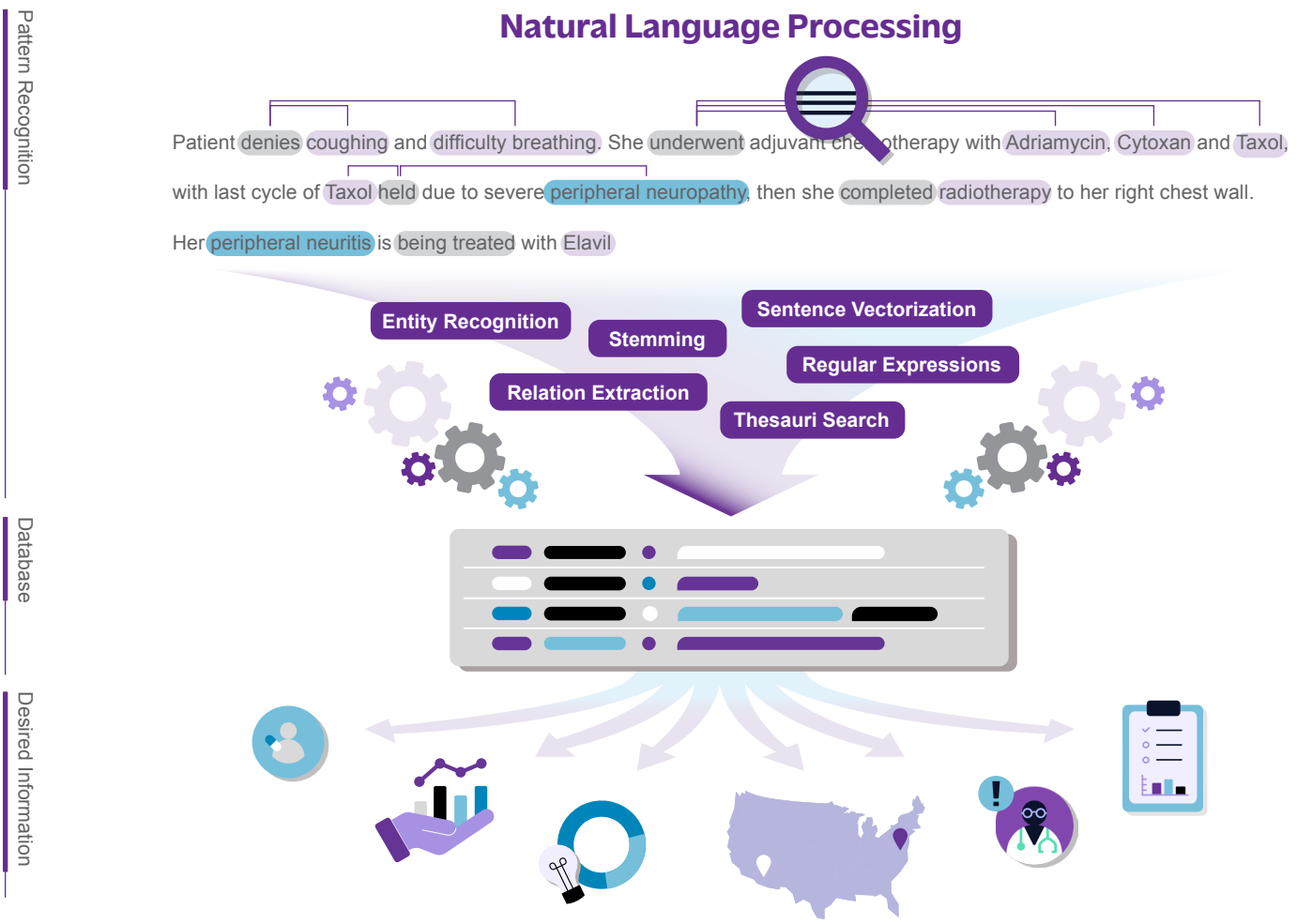
Medical transcription notes offer the granularity of information typically only available from a chart review study without the time requirement necessary to create an electronic chart review form (eCRF) and without the limitation to sample size due to time and cost restrictions. We use NLP to create variables that are less likely to be found in a structured data field (e.g., symptoms, reasons for treatment discontinuation, over the counter medication usage, diagnostic method used, mutations, race/ethnicity, health-related quality of life, death, vaccine recommendations, etc., Figure 2) while also capturing those more common variables (e.g., age, gender, prescription medication use, comorbidities, laboratory values, etc., Figure 2).

Figure 2



Natural Language Processing (NLP) methodology utilizes pattern recognition to find predefined sequences of text. It allows lists of terms to be searched simultaneously as well as the ability to represent linguistic constructions such as sentences and phrases (Figure 3).

Figure 3



Clients specify the specific variables of interest and for each of the variables, every distinct value (whether categorical, binary, or numeric) are coded as machine-understandable NLP patterns by the Insights' team of NLP analysts to extract the desired information and generate a structured table that can be used for further analyses. Examples of text, the associated value, and variable of interest are provided on Table 2.

Table 2

Example Text	Value	Variable
LABORATORY DATA: Today white blood cell count 8.65, hemoglobin 13.9, hematocrit 41.3, and platelet count 221.	Hemoglobin - 13.9	Lab Measure/Lab Value
The patient had a normal hemoglobin of 14 last month and on admission to the emergency room at Frick, his hemoglobin was […]	Hemoglobin - 14	Lab Measure/Lab Value
Asthma: We will continue with Symbicort inhaler and albuterol inhalers.	Budesonide - Inhalation	Admin Method
Upon coming to the Hospice Inn, the patient was started on a morphine drip with breakthrough pain as well.	Morphine - IV	Admin Method
Patient was started on clonazepam 0.25 mg 2 times a day, which patient reported great improvement from.	Clonazepam - Start	Tx Patterns

Patient's carbamazepine was decreased to 200 mg daily.	Carbamazepine - Dose Decrease	Tx Patterns
She has a history of MRSA and an underlying diabetes type 2.	Type 2 Diabetes	Comorbidities
She also has a history of hypothyroidism, coronary artery disease, congestive heart failure, chronic renal insufficiency, depression, hyperlipidemia, chronic back [...]	Coronary Artery Disease	Comorbidities
The patient also has significant tobacco abuse.	Tobacco	Addition
HISTORY OF PRESENT ILLNESS: The patient is a 26-year-old woman with a history of severe alcohol use disorder as well as anxiety and depression.	Alcohol	Addition

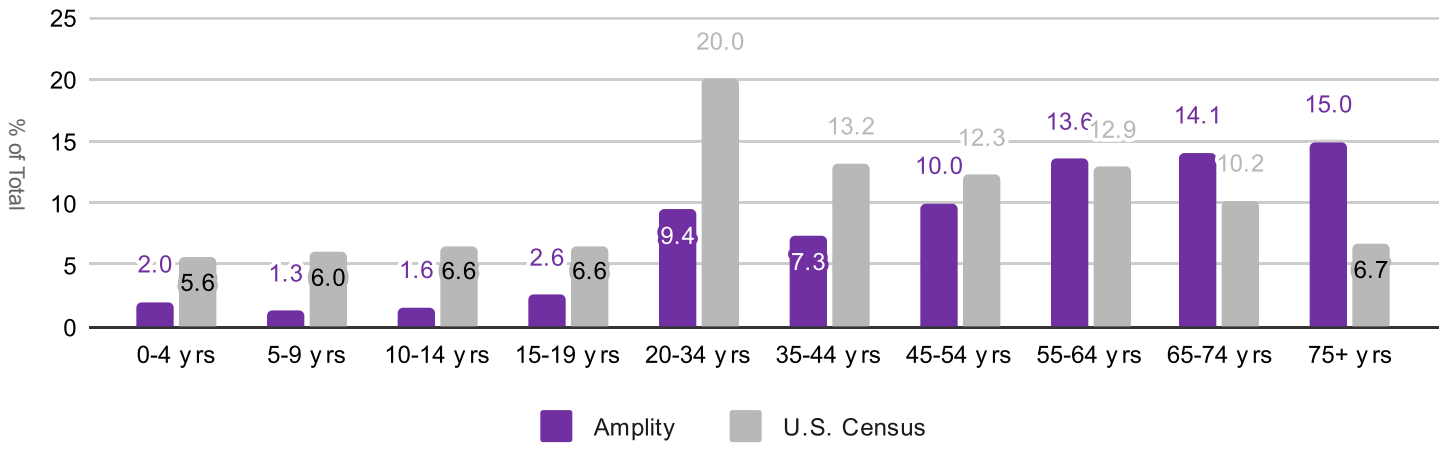
Amplity Insight's Representativeness

Amplity Insight's data analysts leverage best practices in data science to ensure data quality. All variables are assessed for precision (variable results tagged as true positive or false positive) and ensuring full transparency, the results are reported and discussed with the client and updates to logic conducted where applicable to increase precision. The goal is to achieve a positive predictive value of 95% for common variables (age, gender, race, etc.) and 90% on more complex customized variables (e.g., severity using specific definitions, treatment response, disease specific pain, etc.). Similar precision thresholds have been reported in a prior systematic literature review of various NLP methodology including methods for validity assessment.[1]

Age

Older age groups are overrepresented in the Amplity Insight's database as compared to the U.S. Census[2] estimates (Figure 4) however, this is similar to the May 2021 NCHS Data Brief using data from the National Ambulatory Medical Care (NAMC) survey that reported characteristics of office-based physician visits. The rate for adults aged 65 and over (550 per 100 adults) were higher than the rates for children aged 1-17 (153 per 100 children), adults aged 18-44 (173 per 100 adults) and adults aged 45-64 years (302 per 100 adults).[3]
The Amplity Insight's database captures large number of patients across all age categories to allow meaningful research insights within age groups of interest.

Figure 4. Amplity Database Percent Age Distribution Compared to U.S. Census[2]



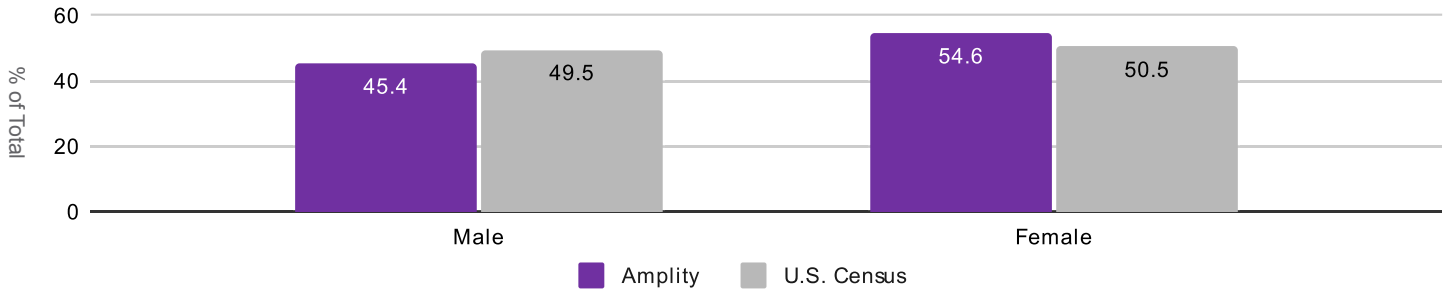
Age Categories	Amplity n
0-4 yrs	417,673
5-9 yrs	265,464
10-14 yrs	322,334
15-19 yrs	527,157
20-24 yrs	685,444
25-29 yrs	851,098
30-34 yrs	958,477
35-39 yrs	991,551
40-44 yrs	1,018,393
45-49 yrs	1,219,857
50-54 yrs	1,515,254
55-59 yrs	1,781,771
60-64 yrs	1,919,976
65-69 yrs	1,997,787
70-74 yrs	1,853,465
75-79 yrs	1,542,374
80-84 yrs	1,189,267
85+ yrs	1,420,547

Gender

Amplity Insight's data analysts leverage best practices in data science to ensure data quality. All variables are assessed for precision (variable results tagged as true positive or false positive) and ensuring full transparency, the results are reported and discussed with the client and updates to logic conducted where applicable to increase precision. The goal is to achieve a positive predictive value of 95% for common variables (age, gender, race, etc.) and 90% on more complex customized variables (e.g., severity using specific definitions, treatment response, disease specific pain, etc.). Similar precision thresholds have been reported in a prior systematic literature review of various NLP methodology including methods for validity assessment.[1]

Males are slightly underrepresented compared to the U.S. Census national estimates[2] (45.4% Amplity Insights vs 49.5% national estimates, Figure 5). This is consistent with the May 2021 NCHS Data Brief using data from the National Ambulatory Medical Care (NAMC) survey that reported characteristics of office-based physician visits and found that the visit rate among females were higher than the visit rate for males (females: 308 visits per 100 females, males: 224 visits per 100 males).[3]

Figure 5. Amplity Insights Gender Distribution Compared to U.S. Census[2]

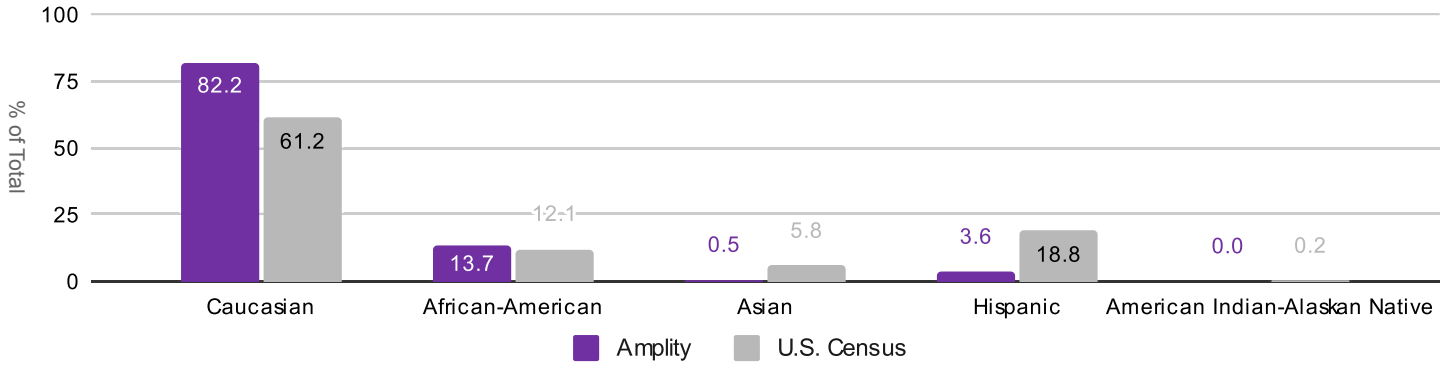


Gender	Amplity n
Female	14,560,131
Male	12,094,022

Race/Ethnicity

Compared to the U.S. Census estimates[2], the proportion of African Americans in the Amplity Insight's database is similar to the proportion reported nationally (U.S.: 12.1%, Amplity Insights: 13.7%, Figure 6). Caucasians are overrepresented compared with national estimates (U.S.: 61.2%, Amplity Insights: 80.9%); Asians and Hispanics are underrepresented relative to the U.S. Census estimates however, sufficient sample size to conduct research and gain meaningful insights.

Figure 6. Amplity Insights Race/Ethnicity Distribution Compared to U.S. Census[2]

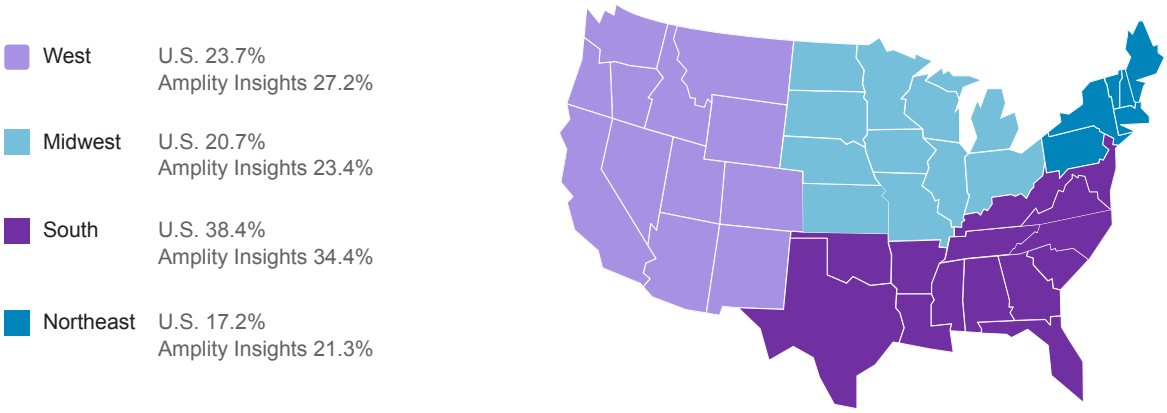


Race/Ethnicity	Amplity n
Caucasian	6,736,193
African-American	1,140,074
Asian	60,916
Hispanic	383,485
Pacific Islander	333
Middle-Eastern	3,604
Multiracial	628

U.S. Census Region

When compared with the U.S. Census[2], the Amplity Insights database is close to nationally representative estimates with a small overrepresentation in the West (U.S.: 23.7%, Amplity Insights: 27.2%, Figure 7), Midwest (U.S.: 20.7%, Amplity Insights: 23.4%), and Northeast (U.S.: 17.2%, Amplity Insights: 21.3%) and small underrepresentation in the South (U.S.: 38.4%, Amplity Insights: 34.4%).

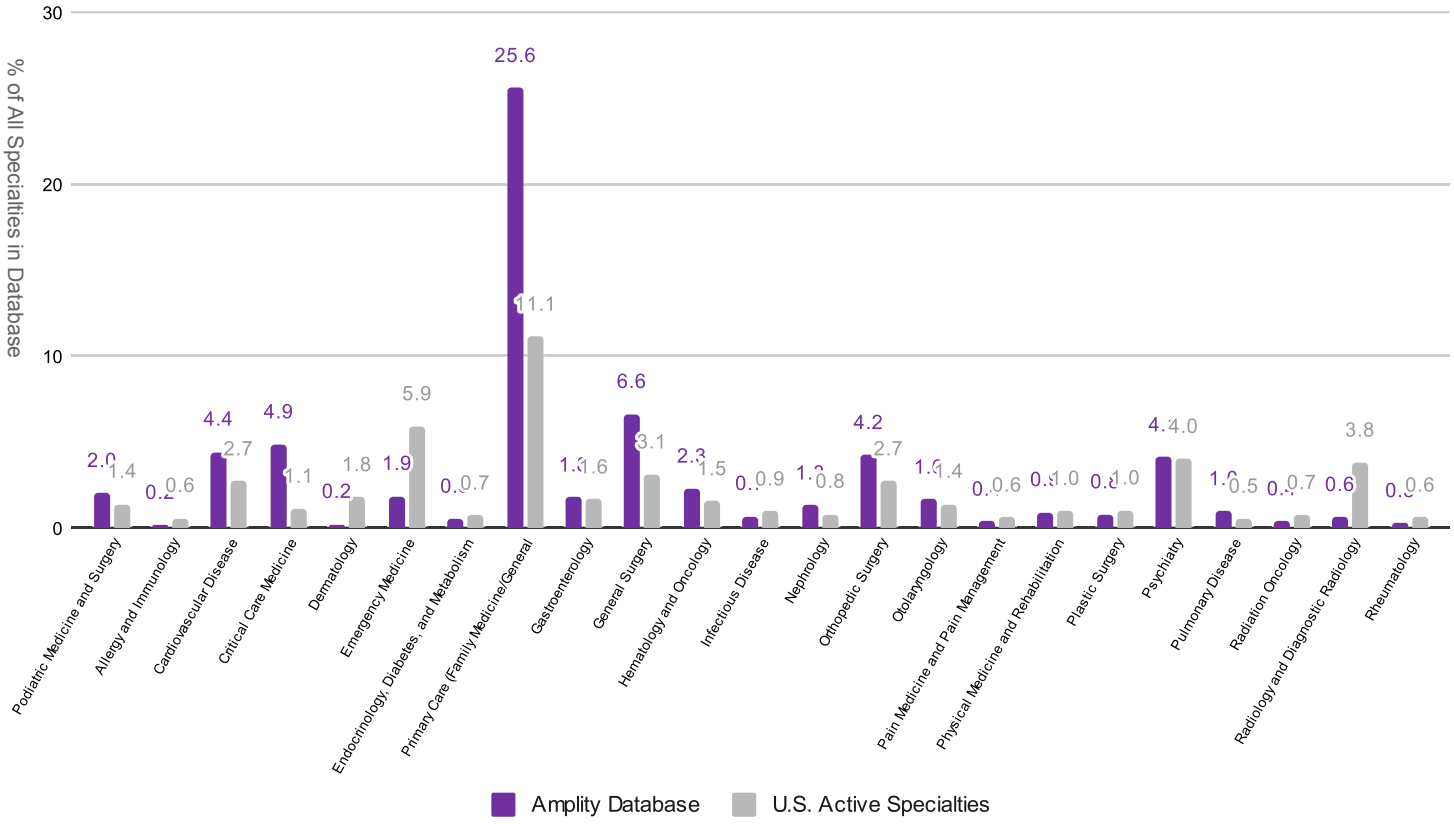
Figure 7. Amplity Insights Geographic Census Region Distribution Compared to U.S. Census Regions[2]



Physician Specialty

The distribution of physician specialties in the Amplity Insight's database overall is representative to the U.S. Active specialties[4] with primary care overrepresented while the distribution of other specialties represented in the Amplity Insight's database like that of the distribution of U.S. active physician specialties (Figure 8).

Figure 8. Amplity Database Percent Physician Specialty Distribution Compared to Active U.S. Physician Specialty[3]



Longitudinal Perspective

In addition to the capability to capture more granular data, choosing a real-world database that gives you the ability to understand and examine the full patient journey from before and following the diagnosis is valuable. An additional outcome reported in the Amplity Insight's database not typically captured in other databases without the linkage of multiple data sources, is mortality (e.g., obituary data or social security death index). The Amplity Insight's database includes approximately 75,730 mortality records sourced from death summary records and/or death notes. Additional detail that can be captured from the death records includes cause of death where available. Amplity leverages expertise from clinicians, epidemiologists/HEOR scientists, and senior data scientists who have spent years working with health information in various forms (claims, structured and unstructured EHRs, medical transcription) and bring that knowledge to tailor an approach that meets your strategic objectives.

On average, patients in the Amplity Insight's database have 2 medical transcription records although patients can have anywhere between 1 and 100 medical transcription records (Table 3). This translates to between 0* to 91 months (7.6 years) of follow-up time per patient within the database.

* Patients with a minimum follow-up time of 0 months occurs when those patients have multiple records occurring in the same month.

Table 3

Number of records per patient	
Mean	2.48
Standard Deviation	3.99
Median	1
Min	1
Max	100

Word counts per record	
Mean	561.66
Standard Deviation	346.77
Median	495
Min	17
Max	3,478

The uniqueness of medical transcription records are that one record could have years' worth of history documented within a single medical transcription encounter record (Table 3). The average number of words within Amplity Insight's medical transcription records is 561 words (median 495 words) with a range between 17 words to 3,478 words. Rule A et al, conducted a study evaluating length of outpatient progress notes for nearly 3 million outpatient encounter notes written across 46 specialties, reported median word counts ranging between 197 words and 1,604 words with word counts varying across specialties.[5] The Amplity Insight's data base records have a larger word count range compared to what has been published previously but it is important to note that the Insight's database is comprehensive and includes medical transcription records from physician and non-physician practitioners and is inclusive of medical transcription records originating from both inpatient and outpatient settings. Number of transcription records and word counts can differ by disease. Examples of medical record transcription counts and related word counts from common and rare diseases are provided on Table 3.

Table 4.

Disease Specific Records Stratified by Disease

	COPD	AATD	SMA	Oncology
Number of records per patient				
Mean	2.56	1.94	1.69	1.99
Standard Deviation	3.67	2.34	1.98	2.81

Median	1	1	1	1
Min	1	1	1	1
Max	99	47	35	98
Word counts per record				
Mean	720.48	923.44	741.82	605.8
Standard Deviation	381.27	529.12	452.85	354.06
Median	654	806	630	545
Min	23	51	70	7
Max	3,500	3,492	3,245	3,499

# Patients	1,805,338	6,292	1,916	935,701
# Disease Specific Records	4,643,466	12,199	3,246	1,864,882

Omni Records*/Non-Disease Specific Records Stratified by Disease

	COPD	AATD	SMA	Oncology
Number of records per patient				
Mean	5.15	4.63	3.96	4.47
Standard Deviation	9.15	9.11	7.77	8.67
Median	2	2	2	2
Min	1	1	1	1
Max	200	186	172	200
Word counts per record				
Mean	627.8	738.15	605.87	605.19
Standard Deviation	368.42	470.75	415.23	352.95
Median	561	630	511	543
Min	12	36	33	7
Max	3,500	3,492	3,245	3,500

# Patients	1,805,338	6,292	1,916	935,701
# Omni Records	9,492,257	30,881	8,401	4,314,366

* Omni represents longitudinal records for a patient that may not mention the disease or medication directly.
Disease specific records and Omni records* (non-disease specific) for each patient linked by a unique patient identifier
Abbreviations COPD: Chronic Obstructive Pulmonary Disease; AATD: Alpha 1 Antitrypsin Deficiency; SMA: Spinal Muscular Atrophy

HIPPA Compliant Data

Protecting the privacy of patient data is an integral component of research. All medical transcription records run through a proprietary software that de-identifies protected health information (PHI) in accordance with the Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy Rule.[6] The de-identification methodology have been certified by a third party to verify that they meet HIPAA requirements for fully de-identified data under the HIPAA Safe Harbor standard.[7]

References

1. Koleck TA, Dreisbach C, Bourne PE, et al. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. J Am Med Inform Assoc. 2019;26(4):364–379.

2. 2021 American Community Survey 1-Year Estimates as captured from the United States Census Bureau: https://data.census.gov/profile?q=United+States&g=010XX00US. Data Accessed April 2023.

3. Ashman JJ, Santo L, Okeyode T. Characteristics of office-based physician visits, 2018. NCHS Data Brief, no 408. Hyattsville, MD: National Center for Health Statistics. 2021. DOI: https://dx.doi.org/10.15620/cdc:105509external icon.

4. U.S. Physician specialty distribution is American Medical Association. AMA Physician Masterfile (December 2019). 2 Podiatrist counts captured from the U.S. Bureau of Labor and Statistics, https://www.bls.gov/oes/current/oes291081.htm#(1).Acessed April 2023. Results for May 2021.

5. Rule A, Bedrick S, Chiang MF, Hribar MR. Length and Redundancy of Outpatient Progress Notes Across a Decade at an Academic Medical Center. JAMA Netw Open. 2021;4(7):e2115334. doi:10.1001/jamanetworkopen.2021.15334.

6. HIPAA Privacy Rule, U.S. Department of Health and Human Services, National Institutes of Health, https://privacyruleandresearch.nih.gov/pr_08.asp.

7. De-identification assessment and HIPPA Certification of Amplity Insight's database conducted by Bradley Malin, Ph.D., Consultant, Privasense, LLC., 4/4/2022.