
AMPLITY INSIGHTS

# Creating Value with Real World Data and Natural Language Processing

An Amplity Health White Paper

SEPTEMBER | 2020

Amplity HEALTH

# Table of Contents



Note: This report contains information from numerous sources that Amplity Health believes to be reliable but for which accuracy cannot be guaranteed. The reader assumes all responsibility for how they use this information.

Amplity Insights

# Creating Value with Real World Data and Natural Language Processing

## Introduction

The era of big data in health care is upon us, powered by the digitization of all healthcare information. Fundamental to this trend has been the widespread adoption of the electronic health record (EHR). (1) However, enthusiasm regarding the use of EHR data in research has been tempered by concerns over data quality – inconsistent formatting, missingness, variability in recording and the fact that many important variables are not routinely coded but contained in free text. (2) Indeed, most data in EHRs are recorded in free text, including physician notes, radiology reports, pathology reports, etc. Mining such free text presents a challenge. Natural language processing (NLP) provides promising tools to make sense of these unstructured data (3); however, NLP is not systematically applied to code all concepts of interest, and the state of the art does not yet give human-level accuracy.

Clinicians instinctively know this, and it underlies their ongoing motivation to enter much of the data into EHRs as free text. (3) Unlike the entry of coded information, narrative text captures a more nuanced patient narrative, can be told from different perspectives, and allows expressions of feelings. In addition, entering coded data places additional burdens on clinicians; it may take longer to find and enter codes than required to summarize a consultation in text. Free text permits clinicians to precisely describe clinical findings, describe supportive evidence for a diagnosis, and provide context for relevant psychosocial problems. Text can allow clinicians to summarize their thought processes in diagnosis or in coming to a specific treatment recommendation. These issues are particularly important for oncology patients and those with rare diseases. In oncology, NLP has not proven sufficiently reliable, such that a leading aggregator of oncology EHR data relies on technology-directed chart abstraction by clinicians. In rare diseases, the diagnosis may not be recognized but may be apparent from reviewing clinical notes.

### Examples of insights that can be derived from free text that may not be obtainable from coded sources:

- Full range of laboratory and physical examination
- Past medical history
- Diagnoses that were not coded
- Specific information about devices
- Reasons for alteration or discontinuation of a therapeutic regimen as well as alternatives considered
- Patient-reported symptomatology, frequently in their own words
- Referral patterns

We describe here a novel source of real-world data that provides researchers with direct access to transcriptions of doctors' notes. To enable such access, the text narratives have undergone redactions and dates have been shifted to provide appropriate protection of patient confidentiality according to HIPAA; linking these data to other RWD sources is prohibited for the same reason. Through agreements with physicians and transcription services providing the data, follow-up contact with particular physicians can be initiated regarding potential enrollment of patients into clinical trials. Patients in the database have insurance benefits that span the full range of coverage and payment designs.

## Methods

The Amplity Insights database is composed of detailed narrative records of physicians' encounters with patients, in the form of clinical notes dictated by physician. To de-identify these notes, medical transcription companies license a software tool used for deidentification called doc2deid from its developer, PH18, Inc. (Santa Cruz, California). The transcription companies provide these de-identified data to PH18, which exclusively licenses them to Amplity Insights. The data are uploaded on a rolling basis to the Amplity Insights database.

The deidentification process used to generate the Amplity Insights data set complies with the Safe Harbor standards of the Privacy Rule of the Health Insurance Portability and Accountability Act (HIPAA) of 1996. The Safe Harbor approach mandates the removal or redaction of 18 specific fields that could be used to identify a patient. In the unstructured clinical notes used as a source for the Amplity Insights database, these fields could occur in many different formats. Therefore, natural language processing (NLP) algorithms were incorporated into the doc2deid software and certified by an independent third-party advisor, Bradley Malin, Ph.D., of Vanderbilt University Medical Center of Nashville, Tennessee with expertise in HIPAA deidentification. The NLP algorithms were designed to find and redact, change, or remove any identifying fields, regardless of their format. For example, dates mentioned in unstructured text might take many forms as full dates or consist only of a month in which a medically relevant event occurred. Names must be recognized either in full or in part, such as a family name with an honorific, and a middle initial correctly classified as a component of a name. The system replaced certain types of identifiers, such as identification numbers, with randomly generated equivalents. The algorithm also was able to access dictionaries of names and of places, and randomly substitute the names of cities with de-identification tags, for example. Such identifiers were noted anywhere they occurred in a document following their first use.

Dates were changed according to a specific rule designed to shift them backward by up to 365 days, creating one year of ambiguity regarding any date. For each patient, the system randomly generated a number between 1 and 365 as the number of days by which to shift all dates for that patient. This approach preserved the relative amount of time between all dated events in a patient's record (for example, the amount of time that passed between administration of a drug and a measure of its effectiveness). Maintaining the year of ambiguity around dated events makes it unlikely that the Amplity Insights database would be linkable to any other patient database, as linkages could re-introduce specificity to dates and risk identifying patients.

In an iterative process, the advisor reviewed successive randomly-chosen batches of records to find any remaining identifiers, as well as over-redactions (erroneous removal, redaction, or change of non-identifiers). Seventy-five percent of remaining identifiers consisted of portions of provider names, such as a middle initial or provider name fragment in a file name. Partial dates were also observed as remaining identifiers, but no explicit patient information. After each round of review, the algorithm was subsequently refined, and another batch of records generated for review. The iterative process continued over four rounds, with fewer instances of remaining identifiers/over-redactions observed with each round, until the records in the final (fourth) batch did not show any such instances. At this point, the advisor was able to certify compliance with the Safe Harbor deidentification standards as specified in the HIPAA Privacy Rule. This process is repeated twice annually to ensure continued compliance.

The raw de-identified data from the partners are stored on a server and parsed using a variety of scripts that perform additional quality checks on the data, preventing duplicates and inconsistencies. These scripts also create certain fixed regions of metadata for each record, such as year of birth, gender, and de-identified dates of service. The resulting files are then indexed using both public (e.g., National Cancer Institute (NCI), Medical Subject Headings (MeSH), Medical Dictionary for Regulatory Activities (MedDRA), etc.) and private ontologies, and added to the Amplity Insights database. The fixed regions of data, and use of established indexing systems, allow users to build efficient queries. Queries of the database employ NLP and text-mining methods to parse the fields created during indexing with combinations of words, phrases, logical expressions, ontological elements and proximity searches. The selected set of records identified by a query can be expressed in a number of ways. Most commonly, they are depicted in a table showing frequency counts. Columns display queried variables, and rows represent patient records. Within a table, direct links back to the raw de-identified record allow immediate verification. Researchers using the Amplity Insights database therefore use either the raw records, a data table, or a combination of the two.

To project data to the US population, data were weighted to align Amplity Insights demographic characteristics with national estimates from the National Health Interview Survey (NHIS) for each year from 2014 to 2017. Amplity Insights estimates are presented alongside NHIS and NHANES equivalents in Table 1 on the next page. US population estimates from the Census Bureau's Population Estimates Program (PEP) are also included for comparison. Standard weighting methodology was applied (Appendix 1). As some racial/ethnic categories (ex. Asian, Middle Eastern, Pacific Islander, and multiracial) showed low representation within the Amplity Insights database, they were aggregated into a single category, denoted as "Other." To facilitate weighting across a wide range of age groups, five age categories were defined to accurately describe children and adolescents (0 to 17 years of age), young adults (18 to 24 years), two groups of mature adults (aged 25 to 44 years and 45 to 64 years), and older adults (65+ years) in the weighted results.

**Table 1:** Demographic comparisons among data from Amplity Insights (raw and weighted), NHIS, NHANES, and the US Census

| | 2017 | | | | | 2016 | | | | | 2015 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Raw RHD | Weighted RHD | NHIS | NHANES | Population Estimate (Census) | Raw RHD | Weighted RHD | NHIS | NHANES | Population Estimate (Census) | Raw RHD | Weighted RHD | NHIS | NHANES | Population Estimate (Census) |
| **Gender** | | | | | | | | | | | | | | | |
| Male | 44.6% | 48.9% | 48.9% | N/A | 49.2% | 45.3% | 48.9% | 48.9% | 48.8% | 49.2% | 43.8% | 48.8% | 48.9% | 48.8% | 49.2% |
| Female | 55.4% | 51.1% | 51.1% | N/A | 50.8% | 54.7% | 51.1% | 51.1% | 51.2% | 50.8% | 56.2% | 51.2% | 51.1% | 51.2% | 50.8% |
| **Race/ Ethnicity** | | | | | | | | | | | | | | | |
| Non-Hispanic Caucasian | 79.7% | 61.8% | 61.9% | N/A | 60.7% | 85.5% | 61.1% | 62.3% | 60.6% | 61.2% | 87.6% | 64.4% | 62.7% | 60.6% | 61.6% |
| Non-Hispanic African American | 16.1% | 12.7% | 12.9% | N/A | 12.5% | 10.0% | 13.4% | 12.9% | 11.9% | 12.5% | 8.2% | 12.9% | 12.8% | 11.9% | 12.4% |
| Hispanic | 3.1% | 18.4% | 18.1% | N/A | 18.1% | 3.8% | 18.6% | 17.9% | 17.6% | 17.9% | 2.9% | 16.9% | 17.7% | 17.6% | 17.6% |
| Other | 0.4% | 7.1% | 7.1% | N/A | 8.7% | 0.5% | 6.8% | 6.9% | 9.8% | 8.5% | 1.2% | 5.8% | 6.8% | 9.8% | 8.3% |
| **Age** | | | | | | | | | | | | | | | |
| 0-17 | 7.2% | 23.0% | 23.0% | N/A | 22.6% | 8.6% | 23.1% | 23.1% | 24.0% | 22.8% | 4.6% | 23.1% | 23.3% | 24.0% | 22.9% |
| 18-24 | 4.1% | 9.3% | 9.3% | N/A | 9.4% | 4.5% | 9.4% | 4.5% | 9.4% | 8.2% | 9.5% | 4.8% | 9.6% | 8.2% | 9.7% |
| 25-44 | 18.5% | 26.2% | 26.2% | N/A | 26.5% | 17.2% | 26.2% | 26.2% | 26.5% | 26.4% | 19.4% | 26.1% | 26.1% | 26.5% | 26.4% |
| 45-64 | 32.1% | 26.1% | 26.1% | N/A | 25.9% | 31.7% | 26.3% | 26.3% | 26.1% | 26.1% | 35.3% | 26.5% | 26.4% | 26.1% | 26.1% |
| 65+ | 38.1% | 15.4% | 15.5% | N/A | 15.6% | 38.0% | 15.1% | 15.1% | 15.1% | 15.2% | 35.9% | 14.8% | 14.7% | 15.1% | 14.9% |
| **Census Region** | | | | | | | | | | | | | | | |
| Northeast | 22.3% | 18.3% | 18.3% | N/A | 17.3% | 22.0% | 18.2% | 18.2% | N/A | 17.4% | 25.1% | 17.2% | 17.2% | N/A | 17.5% |
| Midwest | 29.9% | 21.9% | 21.9% | N/A | 20.9% | 32.8% | 21.8% | 21.8% | N/A | 21.0% | 11.6% | 22.1% | 22.1% | N/A | 21.1% |
| South | 33.6% | 36.4% | 36.4% | N/A | 38% | 35.2% | 36.2% | 36.2% | N/A | 37.9% | 19.9% | 37.5% | 37.5% | N/A | 37.7% |
| West | 14.2% | 23.4% | 23.4% | N/A | 23.8% | 9.9% | 23.8% | 23.8% | N/A | 23.7% | 43.5% | 23.1% | 23.1% | N/A | 23.6% |

**Estimates weighted using NHIS**

**NHANES dataset covers two years — 2017-2018 is not yet available; 2015-2016 estimates are reported for both 2015 and 2016**

## Results

As of August 2018, the Amplity Insights database includes 30,128,000 records of physician-patient encounters, with approximately 2,150,000 records added to this total per month. These encompass records from more than 205,000 providers at more than 40,000 clinics and hospitals, and 14 million patients.

### Longitudinality

**Patients with multiple records over time.** Approximately one third of patients in the Amplity Insights database have multiple records of clinical encounters (Table 2). For most patients, these records were recorded within an 18-month timespan. More than 10% of patients have at least five records in the database, and these may encompass multiple providers and substantial amounts of patient history. Patients with major illnesses are more likely to have multiple records in the database, reflecting a greater intensity of care. This is illustrated by the pattern for patients with any type of lung cancer (Table 2). Approximately one quarter have at least five records over an 18-month period during 2016-2018.

**Table 2:** Number of records available for all patients in the Amplity Insights database

| Number of records available for each patient | All Patients | % | | Lung Cancer Patients | % |
|---|---|---|---|---|---|
| **1** | 6,773,645 | 64.4% | | 16,568 | 40.2% |
| **2** | 1,575,144 | 15.0% | | 6,723 | 16.3% |
| **3** | 706,027 | 6.7% | | 4,388 | 10.7% |
| **4** | 398,147 | 3.8% | | 3,010 | 7.3% |
| **5+** | 1,067,498 | 10.2% | | 10,485 | 25.5% |
| **Total** | 10,520,461 | 100.0% | | 41,174 | 100.0% |

*Table presents counts as of August 1, 2018.

**Individual records capture longitudinal patient history.** Clinicians may also report past care from any provider in the records available for each patient. This allows an assessment of longitudinal history for any patient. Even among patients with only one record, nearly half have three or more years of history captured within that record. The historical narrative referenced for a particular patient in the database has a mean duration of almost 14 years (5095 days) and a median duration of approximately 4 years (1460 days, SD: 6538 days).

## Demographic profile

As shown in Table 1, raw data related to demographic characteristics in the Amplity Insights database for the period 2014-2017 show a different demographic pattern than those from federal sources, such as NHIS. With the application of an appropriate weighting schema (see Appendix), the Amplity Insights demographic profile is much closer to NHIS estimates and to those of NHANES and the Census bureau.

**Gender.** The raw Amplity Insights data slightly under-report males each year (43.8%-45.3%) when compared to NHIS estimates (approximately 49%).

**Race/Ethnicity.** Across the years 2014-7, the proportion of African- Americans in the Amplity Insights database (8.2%-16.1%) was fairly similar to the proportion in NHIS (12.8%-12.9% across years). However, Caucasians were over-represented in the Amplity Insights database (79.7%-87.6%) compared to NHIS estimates (61.9%-63.4%). Hispanics were under-represented (2.9%-5.0% Amplity Insights, 17.4%-18.1% NHIS) as were people in the "Other" category (0.4%-1.2% Amplity Insights, 6.4%-7.1% NHIS).

**Age.** The three youngest age groups were under-represented in the Amplity Insights database vs. NHIS estimates (0-17 years: 4.6%-8.6% Amplity Insights, 23.0%-23.5% NHIS; 18-24 years: 4.1%-5.1% Amplity Insights, 9.3%-9.7% NHIS; and 25-44 years: 17.2%-19.4% Amplity Insights, 26.1%-26.2% NHIS), while the oldest two age groups were over-represented (45-64 years: 31.7%-35.3% Amplity Insights, 26.1%-26.4% NHIS; and 65+ years: 35.5%-38.1% Amplity Insights, 14.4%-15.5% NHIS).

**Geography.** The Census region distribution in the Amplity Insights database varied substantially among different years (Northeast: 22.0%-31.5%, Midwest: 7.6%-32.8%, South: 19.9%-33.6%, West: 9.9%-43.5%), while the same distribution in NHIS remained fairly consistent (Northeast: 16.9%-18.3%, Midwest: 21.8%-22.6%, South: 36.2%-37.6%, West: 22.9%-23.8%).

## Research samples

**Oncology.** As of August 2018, more than 500,000 records for more than 200,000 patients and approximately 30,000 providers document care for breast cancer. Other cancers with more than 200,000 records from more than 100,000 patients and 20,000 providers include colorectal, lung, and prostate cancer. Less common than these, pancreatic cancer is still represented by more than 50,000 records, 20,000 patients, and 10,000 providers.

**Rare diseases.** Because the Amplity Insights database shows narrative descriptions of patients' presentation and symptoms, it is possible to not only identify patients with a diagnosis but also potential patients with suggestive signs and symptoms. Table 3 compares the numbers of patients diagnosed with three orphan diseases to undiagnosed patients whose records show multiple signs and symptoms associated with those diseases. The latter group may present a rich source of potential enrollees for randomized controlled trials (RCTs); finding patients with orphan diseases remains a major challenge in recruiting RCTs. The signs and symptoms included were chosen by clinical experts and have been used by AMPLITY INSIGHTS clients to identify patients who may have the orphan diseases of interest.

**Table 3:** Comparison of numbers of patients and records with diagnosed orphan diseases versus recorded signs and symptoms of diseases

| Query | Actual disease diagnosis counts | | Signs and symptoms used to suggest potential diagnosis* in undiagnosed patients | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Patients | Records | 2+ | | 3+ | | 4+ | | 5+ | | 6+ | |
| | | | Patients | Records | Patients | Records | Patients | Records | Patients | Records | Patients | Records |
| **Paroxysmal Nocturnal Hemoglobinuria (PNH)** | 1,016 | 1,698 | 1,303 | 1,998 | 55 | 74 | 1 | 1 | 0 | 0 | 0 | 0 |
| **Atypical hemolytic uremic syndrome (aHUS)** | 162 | 470 | 936 | 1,334 | 34 | 38 | 2 | 2 | 0 | 0 | 0 | 0 |
| **Lysosomal Acid Lipase Deficiency (LAL-D)** | 28 | 29 | 67,459 | 97,864 | 8,393 | 10.477 | 1,095 | 1,214 | 100 | 103 | 6 | 6 |

- PNH: Coombs negative hemolytic anemia; hemoglobinuria; renal dysfunction with hemolysis; aplastic anemia; myelodysplastic syndrome; unexplained thrombosis; pancytopenia/cytopenia with hemolysis, refractory iron deficiency anemia, non-response to therapy, or thrombosis
- aHUS: Thrombotic microangiopathy; hemolytic uremic syndrome; ADAMTS13>5%; malignant hypertension; consult with a maternal fetal medicine specialist; hemolysis, elevated liver enzymes, and low platelet count (HELLP) syndrome; eclampsia; pre-eclampsia
- LAL-D: Elevated low-density lipoprotein (LDL) or alanine aminotransferase (ALT); liver cirrhosis; splenomegaly; hepatomegaly; hepatosplenomegaly; non-alcoholic fatty liver disease; microvesicular steatosis; bilateral adrenal calcification

## Annotated Extracts from Sample Records

Below are excerpts from two sample patient records along with some commentary. Full transcripts are included in Appendix 2.

**Patient 1** is a 14-year-old boy from Pennsylvania who has been followed by his doctor for the past two years for nonalcoholic fatty liver disease (NAFLD) with elevated liver enzymes and obesity. A liver biopsy during that time also showed evidence of progressive fibrosis and steatosis, although no organomegaly.

> As you know, **NAME is now a 14-year-old male who we initially saw back in **DATE[10/06/2014] with elevated liver enzymes in the setting of moderate obesity. Subsequent investigations were revealing for evidence of nonalcoholic fatty liver disease and steatohepatitis from a liver biopsy in **DATE[01/05/2015] that showed portal and periportal fibrosis with early bridging stage II/IV. **NAME demonstrated excellent weight management with stabilization of weight over several months' time period. His liver enzymes, which had peaked in the 250 range improved. Last year they were 64 and 113, this year stable at 68 and 129.
>
> LABORATORY DATA: AST and ALT of 68 and 129, total bilirubin 0.5 with direct 0.3, total protein and albumin of 7.7 and 4.8.
>
> IMAGING: Abdominal ultrasound. Impression: Fatty liver.
>
> ABDOMEN: Soft, nontender, nondistended. No organomegaly.
>
> [Comment: Notably, these symptoms are hallmarks of the rare condition called lysosomal acid lipase deficiency, or LAL-D, an autosomal recessive disorder with a late-onset form diagnosed in children and adolescents. It is often missed due to the nonspecific constellation of symptoms, with patients identified as having NAFLD but not given genetic tests that can identify the disorder, which arises from a single mutation in the LIPA gene. (4,5) This patient's record does not reflect genetic testing despite a family history of related symptoms.]
>
> FAMILY HISTORY: As noted before, dad with elevated liver enzymes and suspicion for nonalcoholic fatty liver disease. Interestingly dad himself is now significantly obese.
>
> [Comment: The record also does not indicate the results of a lipid profile, with elevated low-density lipoprotein-cholesterol (LDL-C) and low high-density lipoprotein-cholesterol (HDL-C) also common in LAL-D, (Burton et al, 2015; Himes et al, 2016) although further blood work may have been planned after the appointment]

We will obtain **NAME today in clinic, which will help us non-invasively follow disease progression.

Persistent liver enzyme elevation despite decent weight management strategies; however, stressed some improvements that could be made over the coming year. Otherwise, no additional changes in medical management. We …stressed the importance of compliance with vitamin E therapy. In the absence of any significant complications, we will plan to see him back in approximately 1 year's time.

**Patient 2** is a 72-year-old Caucasian woman from California with a 20-year history of breast cancer, who seeks a second opinion from an oncologist. The oncologist records her initial therapy in 1996, the subsequent detection in 2012 on MRI of bone metastases, and history of hormonal therapies since that time. Since 2012, the patient has been seen at the oncologist's clinic and also at another hospital, whose care is also recorded here.

DOB: **DATE[02/26/1944]
**DATE[12/30/2016]

The patient has not been seen since **DATE[03/05/2015] and she comes in now to re-establish care, having had treatment both here and at **NAME Hospital.

HISTORY The patient is currently a 72-year-old woman who in 1996 while in the **ADDR system, noticed a left-sided breast upper inner quadrant nodule. She states she had a lumpectomy and lymph node dissection for what was a grade 3, hormone receptor positive invasive ductal carcinoma. She received two months of CMF chemotherapy, radiation, and then four months of CMF chemotherapy, but states she was told she did not need hormonal therapy. She has had progressive back pain, particularly over the last couple of months and had an MRI of the lumbar spine on **DATE[02/18/2012] at **NAME Imaging with the findings of extensive bone marrow signal abnormality in the L5 and right S1 vertebrae, highly suspicious for metastatic breast cancer as it appeared osteoblastic with the recommendation for further investigation.

I last saw her on **DATE[03/26/2012] when she was in to discuss the biopsy results from the L5 transverse process which showed adenocarcinoma consistent with breast cancer, but without enough material for ER, PR, and HER-2 staining.

She went to **NAME in the interval and had a repeat biopsy showing her to have hormone receptor positive, HER-2 negative, invasive ductal carcinoma of the breast and she received radiation therapy to the T9 lesion and the left hip with mostly resolution of her pain.

She was started on hormonal therapy with Femara, but had intolerable arthralgias and was switched to anastrozole, but had bilateral carpal tunnel syndrome and then was switched to tamoxifen since November, 2012 and continued tamoxifen through October 2016 when she had CT scan of the chest, abdomen, and pelvis and a bone scan (for some reason **NAME does not appear to believe in PET/CT scans) with increased L5 activity and she was given radiation therapy and changed to Aromasin for which she tolerated this only two weeks and felt like she was going to die and after completion of radiation therapy was started on Arimidex as of **DATE[12/19/2015] (of note, she had already been on this previously and stopped it because of bilateral carpal tunnel syndrome) and was told to add Afinitor one month later.
She comes in now for another opinion.

[Comment: The reasons for switching hormonal therapies are recorded here as related to the patient's reported side effects. This oncologist furnishing a second opinion appears to question the other hospital's selection of imaging modalities.]

Her tumor markers have previously been normal and we have been using PET/CT scan locally although **NAME seems like they like CT scan and separate bone scan. I told the patient it seems to be more efficient to use a PET/CT scan at signs of symptomatic progression or on a three to four month basis, but changing modalities mid treatment between centers would only cause potential confusion as to disease progression or not and, at this point, unfortunately she needs to stay with a more laborious CT scan and bone scan separately.

[Comment: The treatment decision is also questioned.]

Treatment strategy. She … was on tamoxifen from November 2012 until October 2016 when she had a single site of bony progression. If this is the only site of progression and she received radiation therapy, there seemed to be, in my opinion, no reason to switch from tamoxifen.

[Comment: The record includes imaging and pathology results and many lab values dictated into the record between 2012 and 2016, both normal and abnormal, with some redacted to avoid identifiable specificity.]

Labs **DATE[03/02/2012] shows a white count of 2.9, hemoglobin 10.9, platelets 208. Normal creatinine. Normal CEA. Normal CA15-3. Normal LDH. Normal AST, ALT, and alkaline phosphatase.

[Comment: The record also captures the strategy for switching treatment and coordinating care with the other hospital, the rationale for why the oncologist chose particular therapies, and the patient's satisfaction with the plan.]

At this point, given the superiority of head-to-head comparisons of Faslodex and aromatase inhibitor therapy, the recommendation is for Faslodex alone or Faslodex with Ibrance. Since she has bone-only disease and minimal progression and due to her multiple side effects previously, I think it is reasonable to start with Faslodex as a single agent and add Ibrance at progression. At progression thereafter, I would recommend switching her to tamoxifen and Afinitor as resensitization or even tamoxifen alone if she truly only one site of disease progression that was treated with radiation therapy while on tamoxifen. Thereafter, the consideration of chemotherapy could be considered but with bone-only disease I think she will have months to years of time to review this.

## Cohort analysis

Clinical notes capture nuances of provider observations about important events that may not be coded in administrative claims. An example is suicidal ideation, which may be recorded by providers in varied ways: "…admits to continued suicidal thoughts…" "…plan to hang herself…" "she thought about harming herself for the last 3 days…" NLP can identify such phrases and create a more accurate picture of patients with suicidal ideation than could be gleaned from claims data or reliance on codes.

A cohort analysis of suicidal ideation in the Amplity Insights database shows treatment patterns for 191 patients before and after physicians noted their suicidal tendencies. Selected patients were required to have at least three visits, with at least one year elapsing between them, and at least one visit before and after the date when suicidal ideation (SI) was first recorded (SI date). All patients had been diagnosed with depression and had a history of using a named selective serotonin reuptake inhibitor (SSRI) or serotonin and norepinephrine reuptake inhibitor (SNRI). Patients were categorized by which therapy they had been taking prior to their SI date, with patients on multiple therapies counted under each. The most common medications taken by patients prior to SI were sertraline (28%), fluoxetine (26%), venlafaxine (17%), citalopram (15%), duloxetine (15%), and escitalopram (13%). The most frequent pattern after the SI date was for the clinical notes to mention that the patient continued to take the same drug they had been taking before the SI date. However, most (74%) changed dosage on or after their SI date, compared to 56% of non-SI patients. The non-SI patients appear to add therapies in subsequent visits after their first. Discontinuation was rare, with 4-5% of patients in each cohort discontinuing their initial therapy. The clinical records also show patterns of hospitalization, with nearly all (96%) of SI patients hospitalized on or after the visits after the SI date, compared to 59% of the non-SI cohort.

## Discussion

Currently, there is intense interest in leveraging real world data for purposes ranging from population health management to quality improvement efforts to drug development. The Amplity Insights database is an alternative source of RWD to administrative claims and electronic health records.

While it is not possible to link Amplity Insights to other RWD datasets, separately querying it can provide important insights that may not be evident elsewhere. In particular, the ability to review the narrative as written by clinicians may provide insights into their thought processes in diagnosis or in coming to a specific treatment recommendation. It may also provide a rich source of rare disease patients for recruitment into clinical trials, as it may help identify patients who have not as yet been given a diagnosis. The nuances in the narratives can also provide greater depth to understanding a patient's journey through the healthcare system. As such, we believe it is a valuable addition to the RWD sources available to researchers.

### References


1. M.L. Berger, V. Doban. Big Data, Advanced Analytics and the Future of Comparative Effectiveness Research. Journal of Comparative Effectiveness Research 2014; 3(2): 167-176.

2. G. Daniel, C. Silcox, J Bryan, M. McLellan, M. Romaine, K. Frank. Characterizing RWD Quality and Relevancy for Regulatory Purposes. Duke-Margolis Center for Health Policy, October 1, 2018.

3. E. Ford, J.A. Carroll, H.E. Smith, D. Scott, J.A. Cassell. Extracting Information from the Text of Electronic Medical Records to Improve Case Detection. Journal of the American Medical Informatics Association 2016;23:1007–1015.

4. B.K. Burton, P.B. Deegan, G.M. Enns, O. Guardamagna, et al, Clinical Features of Lysosomal Acid Lipase Deficiency. Journal of Pediatric Gastroenterology and Nutrition 2015; 61:619-625.

5. R.W. Himes, S.E. Barlow, K. Bove, N.M. Quintanilla, et al, Lysosomal Acid Lipase Deficiency Unmasked in Two Children with Nonalcoholic Fatty Liver Disease Pediatrics 2016; 138(4): e20160214.

## Want to Learn More?

Please reach out to schedule a call with one of our experts to learn more about Amplity Insights, or dataset, or one of our other commercialization capabilities.

phone **800.672.0676, Press 1**

email **contact@amplity.com**

web **Amplity.com**

### About Amplity Health

A true partner to global healthcare companies, Amplity builds transformational solutions by challenging boundaries and the status quo. Amplity has the expertise, infrastructure, and insights to help clients overcome any healthcare commercialization hurdle. With wide-ranging capabilities such as patient identification solutions, multichannel messaging, inside sales and service resourcing, field sales, campaign planning and execution, and more, Amplity has what it takes to maximize marketing efforts and set strategies up for success.

## Appendix 1

**Weighting Procedures**

Interim weights specific to each of the four US Census regions were created for each of 90 separate demographic groups defined by different combinations of age, gender, and race, including missing values for each variable. An interim weight of 1 was assigned to the group defined by missing data for all three demographic variables (age, gender, and race). Final weights were computed by multiplying the interim weight by an adjustment factor for each Census region. A weight larger than 1 was assigned to under-represented groups, and a weight less than one to over-represented groups. For 2014-6, weights ranged from 0.014 to 303.534, with an average of 1. Applied weights fell within the range of similar weighting systems applied in similar (medical visit/ emergency room visit) federal datasets. For example, in the dataset from the Drug Abuse Warning Network, Emergency Department (DAWN-ED), an average weight of 22.108 (maximum of 862.824) was used in the final year of data collection (2012). The Medical Expenditure Panel Survey (MEPS) routinely has weights in excess of 50,000. Therefore, no trimming was applied.

## Appendix 2

**Transcripts of full records**

Patient ID 3925411620159117

Dictation ID: **ID Patient Name: **NAME Speaker: **ID Procedure Date: **DATE[08/02/2016] Account Number: **ID Business Entity: **BUS Medical Record Number **ID Signer Name:**NAME (**PHONE) Work Type: Evaluation (03) Referring Physician: **NAME-**NAME PCP: **NAME **NAME-**NAME CC List: **NAME I had the pleasure of seeing **NAME in the **NAME Clinic at the **ID of **ADDR for follow-up of his nonalcoholic fatty liver disease.

As you know, **NAME is now a 14-year-old male who we initially saw back in **DATE[10/06/2014] with elevated liver enzymes in the setting of moderate obesity. Subsequent investigations were revealing for evidence of nonalcoholic fatty liver disease and steatohepatitis from a liver biopsy in **DATE[01/05/2015] that showed portal and periportal fibrosis with early bridging stage II/IV. **NAME demonstrated excellent weight management with stabilization of weight over several months' time period. His liver enzymes, which had peaked in the 250 range improved. Last year they were 64 and 113, this year stable at 68 and 129. He has demonstrated some weight gain, although proportional to his age and trajectory, he has not been overly impressive. He was at the 79th percentile last year and is at the 83rd percentile this year. He has not had any recent hospitalizations or emergency room visits. He has prescribed vitamin E, but is noncompliant with this medication. He has done some strategies to improve weight management, although mom notes that there has been an increase in his sedentary practices, particularly as this relates to gaming and playing on his phone.

PAST MEDICAL HISTORY: Reviewed and otherwise unchanged.

FAMILY HISTORY: As noted before, dad with elevated liver enzymes and suspicion for nonalcoholic fatty liver disease. Interestingly dad himself is now significantly obese.

SOCIAL HISTORY: Finished 7th grade going into 8th grade, doing well.

REVIEW OF SYSTEMS: As noted above and otherwise unremarkable.

LABORATORY DATA: AST and ALT of 68 and 129, total bilirubin 0.5 with direct 0.3, total protein and albumin of 7.7 and 4.8. IMAGING: Abdominal ultrasound. Impression: Fatty liver.

MEDICATIONS: He was prescribed vitamin E, but has been noncompliant.

PHYSICAL EXAMINATION:

VITAL SIGNS: Weight of 63.2 kg, which is at the 83rd percentile, this is up from 55.1 kg a year ago at the 80th percentile. Blood pressure 121/77, heart rate 76, respiratory rate 16.

GENERAL: Well developed, well nourished, no acute distress.

HEAD: Normocephalic, atraumatic.

EYES: Extraocular muscles intact. Conjunctivae clear. No scleral icterus.

EARS: Normal external ears. No evidence of excoriation.

NECK: Supple, no thyromegaly. Trachea midline.

CHEST: Clear to auscultation bilaterally. No acute distress.

CARDIOVASCULAR: Regular rate and rhythm. No murmur. Good distal pulses.

ABDOMEN: Soft, nontender, nondistended. No organomegaly.

SKIN: Clear of significant visible or palpable lesions.

LYMPHATIC: No appreciated lymphadenopathy.

NEUROLOGIC: Normal neurologic exam. No deficits. ASSESSMENT AND PLAN: **NAME is a 14-year-old male with elevated liver enzymes and obesity concerning for nonalcoholic fatty liver disease. Nonalcoholic fatty liver disease. Persistent liver enzyme elevation despite decent weight management strategies; however, stressed some improvements that could be made over the coming year. We will obtain **NAME today in clinic, which will help us non-invasively follow disease progression. Otherwise, no additional changes in medical management. We will reinstitute therapy and stressed the importance of compliance with vitamin E therapy. In the absence of any significant complications, we will plan to see him back in approximately 1 year's time. It was a pleasure seeing him today in the **NAME Clinic.

DATE
08/02/2016
DOB
2002
AGE
14
SPECIALTY
Gastroenterology
STATE
Pennsylvania
PROVIDERID
a0eea7b61b15cac5cddfdacf52ed076e
GENDER
Male
PATIENTID
3925411620159117
ETHNICITY
WORDCOUNT
695
DOCUMENTID
P107_008_391515.txt

Patient ID 6819752620659309
FOLLOW UP VISIT RE: **NAME DOB: **DATE[02/26/1944] **DATE[12/30/2016]

The patient has not been seen since **DATE[03/05/2015] and she comes in now to re-establish care, having had treatment both here and at **NAME Hospital.

HISTORY The patient is currently a 72-year-old woman who in 1996 while in the **ADDR system, noticed a left-sided breast upper inner quadrant nodule. She states she had a lumpectomy and lymph node dissection for what was a grade 3, hormone receptor positive invasive ductal carcinoma. She received two months of CMF chemotherapy, radiation, and then four months of CMF chemotherapy, but states she was told she did not need hormonal therapy. She has had progressive back pain, particularly over the last couple of months and had an MRI of the lumbar spine on **DATE[02/18/2012] at **NAME Imaging with the findings of extensive bone marrow signal abnormality in the L5 and right S1 vertebrae, highly suspicious for metastatic breast cancer as it appeared osteoblastic with the recommendation for further investigation.

I last saw her on **DATE[03/26/2012] when she was in to discuss the biopsy results from the L5 transverse process which showed adenocarcinoma consistent with breast cancer, but without enough material for ER, PR, and HER-2 staining.

She went to **NAME in the interval and had a repeat biopsy showing her to have hormone receptor positive, HER-2 negative, invasive ductal carcinoma of the breast and she received radiation therapy to the T9 lesion and the left hip with mostly resolution of her pain.

She was started on hormonal therapy with Femara, but had intolerable arthralgias and was switched to anastrozole, but had bilateral carpal tunnel syndrome and then was switched to tamoxifen since November, 2012 and continued tamoxifen through October 2016 when she had CT scan of the chest, abdomen, and pelvis and a bone scan (for some reason **NAME does not appear to believe in PET/CT scans) with increased L5 activity and she was given radiation therapy and changed to Aromasin for which she tolerated this only two weeks and felt like she was going to die and after completion of radiation therapy was started on Arimidex as of **DATE[12/19/2015] (of note, she had already been on this previously and stopped it because of bilateral carpal tunnel syndrome) and was told to add Afinitor one month later.

She comes in now for another opinion.

REVIEW OF SYSTEMS She is having to use OxyContin q.12 hours and oxycodone as needed for breakthrough, which she does not like, and instead is using CBD to put off the side effects of her aromatase inhibitor therapy as well as from the radiation therapy for which she developed some esophagitis and lost weight and is only now starting to gain weight and eat better. No hemoptysis, hematemesis, no melena, bright red blood per rectum. Arthralgias comes and go, under control at this time with CBD, OxyContin, and oxycodone.

PAST MEDICAL HISTORY Malaria, 1993. Osteopenia. History of breast cancer, as noted above. Osteonecrosis of the jaw, diagnosed 2016, with partial plate replacement. **NAME 2016, coronary artery bypass graft x2.

PAST SURGICAL HISTORY Tonsillectomy; tubal ligation, 1991; left breast lumpectomy, 1996; breast reconstruction on the left side, 2007, for repair of a defect.

MEDICATIONS See accompanying electronic medical record.

ALLERGIES Quinine causes GI upset. Bilateral carpal tunnel syndrome from Arimidex. Aromasin causes her to feel like she is going to die. Osteonecrosis of the jaw from Xgeva.

SOCIAL HISTORY She lives in **ADDR and is going back in two weeks to continue her oncology care at **NAME. She is retired from **NAME and has been on disability since 1997. Lifelong nonsmoker, rare alcohol. She and her husband are looking forward to his retirement in the summer of 2012.

REVIEW OF SYSTEMS She is having some right lower back pain in the sacral area. She states she had a mechanical fall but that this pain has been there for more than a year, has waxed and waned and not changing in nature although she did specifically point to the right sacrum and states that while she has no weakness, it causes pain and can come and go and flare at times. No nausea or vomiting, no change in bowel or bladder habits, no melena or bright red blood per rectum, rashes or arthralgias.

PHYSICAL EXAMINATION (Done in the presence of **NAME) Vital Signs: Weight 126.9 pounds, down from 136 pounds in March 2015. BP 149/81. Pulse 94.

Pain score 5/10 in right shoulder and right hip.

HEENT: Normocephalic, atraumatic. Oropharynx is clear.

Neck: Supple.

Lymph Nodes: No cervical, clavicular or axillary adenopathy.

Lungs: Clear bilaterally.

Heart: Regular rate and rhythm.

Abdomen: Soft and nontender.

Extremities: No clubbing, cyanosis or edema.

Back: No tenderness to palpation to the back.

Laboratory Data: Labs **DATE[01/28/2012] shows a normal renal panel, normal electrolytes, normal vitamin D, normal AST, ALT, albumin, globulin and alkaline phosphatase, normal B12.

Labs **DATE[03/02/2012] shows a white count of 2.9, hemoglobin 10.9, platelets 208. Normal creatinine. Normal CEA. Normal CA15-3. Normal LDH. normal AST, ALT, and alkaline phosphatase.

**DATE[02/19/2015], white count 3.1, hemoglobin 11.8, platelet count 171. normal AST, ALT, albumin, globulin, alkaline phosphatase, CEA, and CA15-3.

**DATE[11/18/2016], white count 4.9, hemoglobin 11.0, platelet count 201, normal alkaline phosphatase, normal AST, ALT at 45.

Imagine Data: MRI of the lumbar spine on **DATE[02/18/2012] at **NAME Imaging with the findings of extensive bone marrow signal abnormality in the L5 and right S1 vertebrae, highly suspicious for metastatic breast cancer as it appeared osteoblastic with the recommendation for further investigation.

PET/CT scan done at **NAME Imaging on **DATE[03/09/2012] shows excessive metabolic metastases present in bony L5 and S1 with a single small metastasis involving the right side of T9 without any other bony disease or visceral disease. Of note, the L5 and S1 lesions are quite pronounced and are involving the vertebral bodies and surrounding elements.

**DATE[03/03/2012] negative CT scan of the brain.

Lumbar spine, three views, shows the number alignment of the lumbar spine without evidence of compression fracture, moderate disk and facet degenerative changes at L3, L4, L5, and S1, increased bony trabeculation, possible cortical thickening at the L5 vertebral body concerning for Pagets. Recommend correlative MRI.

Carotid duplex ultrasound **DATE[03/01/2012] shows no hemodynamically significant stenosis.

**DATE[11/26/2012] at **NAME Memorial with a CT scan of the chest, abdomen and pelvis shows no significant change in blastic bone metastases when compared to **DATE[07/28/2012]. There is increased size of the uterus and endometrial complex with recommendation for ultrasound of the pelvis if warranted.

Bone scan **DATE[11/26/2012] at **NAME Memorial Hospital shows stable osseous metastatic disease compared to **DATE[07/28/2012].

PET/CT scan at **NAME Imaging and H3LCRV on **DATE[02/22/2013] showed improved skeletal metastases with now the lesions being isometabolic and no activity with signs of sclerosis without any new disease noted in the viscera or bones.

Bone scan at **NAME **DATE[09/18/2014] shows evidence of active osseous metastatic disease with the main tumor burden in the sacrum, but compared to prior bone scan of **DATE[06/01/2014] there are mixed changes for the sacral lesions with the overall findings stable.

CT scan of the chest, abdomen and pelvis with contrast **DATE[09/18/2014] shows stable osseous metastases when compared to **DATE[06/01/2014]. No new measurable disease in the chest, abdomen and pelvis.

MRI of the lumbar spine **DATE[01/16/2016] at **ADDR Hospital shows progression of presumed osseous metastasis at the S1 level, stable appearance of presumed metastasis at L5, spinal stenosis L4-L5 and L5-S1.

CT chest, abdomen, and pelvis **DATE[09/22/2016] at **NAME shows no visceral disease.

Bone scan October 2016 at **NAME, by report of the patient, said to have increased only in L5.

PATHOLOGY Biopsy of the L5 vertebral body shows HER-2 negative, hormone receptor positive adenocarcinoma consistent with breast.

IMPRESSION/PLAN The patient is a 72-year-old woman with a history of breast cancer in 1996, treated in the **ADDR system with lumpectomy and lymph node dissection, two months of chemotherapy with CMF, followed by radiation therapy, and then four months of additional CMF chemotherapy without hormonal therapy who had metastatic lytic and blastic lesions in L5 and S1 as well as a T9 lesion without any visceral disease who was biopsied locally and then in **ADDR and shown to be hormone receptor positive and HER-2 negative metastatic breast cancer who was treated with Femara, stopped due to arthralgias, and then anastrozole, stopped due to bilateral carpal tunnel syndrome, and has been on tamoxifen from November, 2012 until October 2016 when she was said to have increased activity in L5/S1 as her only site of progression (I did not have the bone scan report to confirm this) who was switched off tamoxifen for some reason and given Aromasin which caused her to have severe side effects and then given radiation to this area and was switched to Arimidex on **DATE[12/19/2015] although she had had bilateral carpal tunnel syndrome from this previously.

Diagnosis. Metastatic adenocarcinoma of the breast that appears to be bone only.

Biopsy. She had a biopsy showing this to be hormone receptor positive, HER-2 negative invasive ductal carcinoma of the breast.

Bone disease. She had been on Xgeva, but has developed osteonecrosis of the jaw and has been taken off of this treatment moving forward.

Treatment strategy. She had been given aromatase inhibitor therapy upfront for metastatic bone-only hormone receptor positive, HER-2 negative breast cancer first with Femara, stopped due to arthralgias, and then anastrozole, stopped due to bilateral carpal tunnel syndrome, and was on tamoxifen from November 2012 until October 2016 when she had a single site of bony progression. If this is the only site of progression and she received radiation therapy, there seemed to be, in my opinion, no reason to switch from tamoxifen. However, she has been withheld from tamoxifen for two months now and has been recommended to be on Arimidex, which she has already seen before and has bilateral carpal tunnel syndrome and was switched to tamoxifen for that reason. Further treatment. At this point, given the superiority of head-to-head comparisons of Faslodex and aromatase inhibitor therapy, the recommendation is for Faslodex alone or Faslodex with Ibrance. Since she has bone-only disease and minimal progression and due to her multiple side effects previously, I think it is reasonable to start with Faslodex as a single agent and add Ibrance at progression. At progression thereafter, I would recommend switching her to tamoxifen and Afinitor as resensitization or even tamoxifen alone if she truly only one site of disease progression that was treated with radiation therapy while on tamoxifen. Thereafter, the consideration of chemotherapy could be considered but with bone-only disease I think she will have months to years of time to review this.

Disease assessment. Her tumor markers have previously been normal and we have been using PET/CT scan locally although **NAME seems like they like CT scan and separate bone scan. I told the patient it seems to be more efficient to use a PET/CT scan at signs of symptomatic progression or on a three to four month basis, but changing modalities mid treatment between centers would only cause potential confusion as to disease progression or not and, at this point, unfortunately she needs to stay with a more laborious CT scan and bone scan separately. Arthralgias and pain from aromatase inhibitor therapy. It should be noted the patient is using oxycodone, OxyContin, and CBD for control of symptoms and I told her that another reason to stop her current hormonal therapy and switch to Faslodex is that she may not need all these other pain medications. She is gratified for this information.

Follow up. She is agreeable to starting with Faslodex now and holding the Afinitor and holding the Ibrance, in particular in combination with the Faslodex, and using single agent Faslodex and I will communicate this to her oncologist at **NAME and she can be treated there moving forward after she returns in two months.

I thank you all.
DATE
12/30/2016
DOB
1944
SPECIALTY
Oncology
STATE
California
PROVIDERID
44622
ZIP
945XX
GENDER
Female
PATIENTID
6819752620659309
ETHNICITY
Caucasian
WORDCOUNT
2545
DOCUMENTID
P102_5871581.txt
FULLPATH
2965\22\P102_5871581.txt