Representation of Unstructured Data Across Common Data Models (DI2)

Final Report – Availability of Priority Data Elements from Chart Annotation

Prepared by: Keith Marsolo, PhD,[1,2] Ruth Reeves, PhD;[3] Li Zhou, MD, PhD;[4] Lesley Curtis, PhD;[1,2] Tyler Erikson, MS;[2] Judy Maro, PhD;[5] Kathleen Shattuck, MPH;[5] Jill Whitaker, MSN, RN-BC;[3] Tina French, RN, CPHQ;[3] Liz Hanchow, RN, MSN;[3] Suzanne Blackley, MA;[4] John Laurentiev, BS;[4] Sarah Dutcher, PhD, MS;[6] Efe Eworuke, PhD;[6] Aida Kuzucan, PharmD, PhD;[6] Joseph Plasek, PhD;[4]

Author affiliations: [1]Department of Population Health Sciences, Duke University School of Medicine, Durham, NC; [2]Duke Clinical Research Institute, Duke University School of Medicine, Durham, NC; [3]Vanderbilt University Medical Center Department of Biomedical Informatics, Nashville, TN; [4]Harvard Medical School and Brigham and Women's Hospital, Boston MA; [5]Department of Population Medicine, Harvard Medical School and Harvard Pilgrim Health Care Institute, Boston, MA; [6]US Food and Drug Administration, Silver Spring, MD

Version 1.0
January 31, 2023

The Sentinel System is sponsored by the U.S. Food and Drug Administration (FDA) to proactively monitor the safety of FDA-regulated medical products and complements other existing FDA safety surveillance capabilities. The Sentinel System is one piece of FDA's Sentinel Initiative, a long-term, multi-faceted effort to develop a national electronic system. Sentinel Collaborators include Data and Academic Partners that provide access to healthcare data and ongoing scientific, technical, methodological, and organizational expertise. The Sentinel Coordinating Center is funded by the FDA through the Department of Health and Human Services (HHS) Contract number HHSF223201400030I.

Representation of Unstructured Data Across Common Data Models

Final Report - Availability of Priority Data Elements from Chart Annotation

Table of Contents

History of Modifications

Version	Date	Modification	Author
0.1	09/30/2022	Original Version	Keith Marsolo & Project WG
0.2	12/05/2022	Incorporation of additional results and feedback from FDA and WG members	Keith Marsolo & Project WG
1.0	1/31/2023	Final version based on feedback from FDA and WG members	Keith Marsolo & Project WG

Introduction and Motivation

The overarching goal of the “Representation of unstructured data across Common Data Models” project is to provide guidance to the Sentinel Network on how best to incorporate information derived from unstructured data into a Common Data Model (CDM) framework. There are three main project objectives, which are to: 1) identify the priority data elements or concepts that are important for pharmacoepidemiological safety studies that FDA could potentially ask data partners to extract from unstructured data; 2a) survey the natural language processing (NLP) solutions that are in use across the Sentinel ecosystem; 2b) assess the overall availability of priority concepts (e.g., medication exposure, smoking status) within unstructured data at two different Data Partners; and 3) develop recommendations on how to best represent natural language processing (NLP)-derived data elements within the Sentinel CDM (SCDM).

This document describes activities related to Objective 2b, assessing the availability of priority concepts at two different Data Partners. To conduct this assessment, we annotated the charts of selected patients to summarize the availability of specific concepts, which had previously been identified as a priority by FDA in Objective 1 (see Table 1 below). Part of the motivation behind this work is a potential future state where Sentinel Data Partners with access to electronic health record (EHR) data have processed some or all of their clinical notes through one or more natural language processing (NLP) pipelines. As NLP pipelines are increasingly offered as commodity software-as-a-service by cloud computing providers like Amazon and Microsoft, we believe this future state is increasingly likely. Understanding the available content that exists within unstructured text will provide context on the type of information that may be available for analysis in Sentinel. While some projects or studies may require the development of new pipelines or processes to extract new or emerging concepts (as seen with the COVID-19 pandemic), we believe that many analyses will be able to take advantage of the “stock” outputs of existing NLP pipelines.

Table 1: Concepts labeled as medium or high priority by FDA.

Domain	Additional detail
Anatomy	
Cancer Pathology	Site, Procedure, Histology
Condition	Diagnoses, Signs / Symptoms, associated metadata (e.g., time period, nature, severity, scale), type (family history, medical history, primary diagnosis)
Genomics	Variant, mutation, expression level
Smoking status	
Admission-Discharge-Transfer (ADT)-type event	
Care Setting	
Diagnostic procedures / tests	
Healthcare profession / specialty	
Medication	Name / Ingredient, associated metadata (e.g., dose, form, route), medication timing, indication
Treatment / procedures	
Concepts from existing Sentinel projects	Anaphylaxis, acute pancreatitis, COVID-19 (+/-), suicidality
Oxygen support	Use of supplemental oxygen, volume, status (change in volume, method or single use)

Physical exam findings	
Death	Date and cause
Hospice care	
General metadata associated with multiple domains	Date, frequency, time, negation, uncertainty

We focused our annotation efforts on two use cases – hospitalized patients with COVID-19, and cancer. Within each use case, we looked at a subset of the clinical narrative that is associated with the patient's electronic health record, as opposed to annotating the entire patient chart or even all notes associated with a selected encounter. We made this decision because we could not necessarily assume that future Sentinel Data Partners will have processed all possible note types (e.g., all hospital discharge summaries have been processed, but not respiratory therapy notes). Characterizing the information available within a portion of the patient's chart can inform the planning of future analyses that seek to leverage information derived from unstructured text.

Methods

The use cases for the annotation task were selected based on input from the FDA and SOC. The broad focus was identified as part of the prioritization process (e.g., COVID-19, cancer), with additional refinement based on available data, project timeline and budget. For each use case, a general population definition was drafted, as well as a strategy for selecting patients and notes within each cohort.

Annotation occurred at two Data Partners – Brigham and Women's Hospital (BWH) and Vanderbilt University Medical Center (VUMC). Epic was the underlying EHR at BWH, while records at VUMC were sourced from Epic for the COVID-19 cohort and from both Epic and StarPanel for the cancer cohort (due to the date of Vanderbilt's transition to Epic). Data analysis was conducted at Duke University. With guidance from FDA, all institutions received determinations from their local Institutional Review Boards (IRB) that this work did not constitute research and fell under the FDA's public health surveillance exemption.

Population definition and sampling strategy

COVID-19: The first use case targeted hospitalized patients with COVID-19, with a specific focus on documenting oxygen use, as the use of supplemental oxygen is a factor in determining COVID-19 and outcomes yet is under-coded in administrative claims data. The qualifying index event was an inpatient encounter with an admitting diagnosis of COVID-19 (U07.1) between April 1, 2020, and December 31, 2021. Patients must have been at least 18 years old at the time of the admission. A decision was made to limit the cohort size to a total of 70 patients. Two cohorts of equal size were defined – patients with a billing code for supplemental oxygen, and those without (relevant codes for supplemental oxygen can be found in **Appendix E – Billing codes for supplemental oxygen**). Patients were randomly selected from the underlying population for inclusion in the annotation cohort.

Cancer: The second use case was focused on cancer, specifically looking at patients who received an order for darzalex (daratumumab), a medication used to treat forms of multiple myeloma. The rationale behind this was to determine if the unstructured text provided enough detail to ascertain the specific indication behind the prescription, as this can be lacking in administrative claims. The index event was a new prescription/order for darzalex between January 1, 2016, and November 30, 2021, with no prescription/order in the prior 3 years. A total of 30 patients were selected at random from this underlying population.

Annotation task

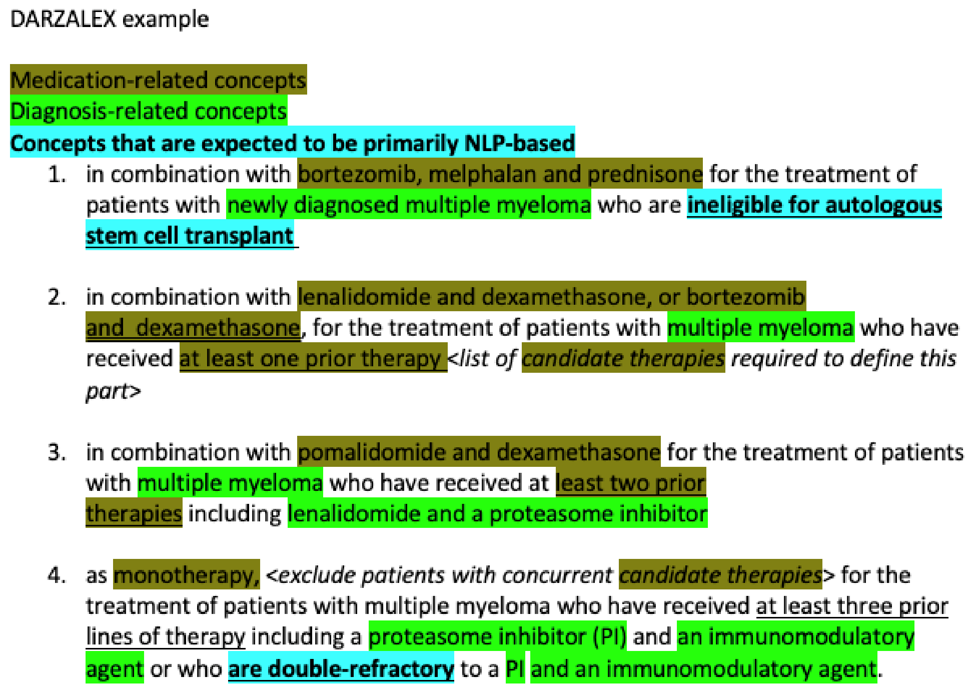
COVID-19: Within the COVID-19 cohort, the primary annotation task examined the discharge summary associated with the COVID-19 hospitalization, annotating the existence of priority concepts that were likely to present in that note type. The rationale for choosing the discharge summary was that if Sentinel were going to complete an analysis of inpatient encounters that relied on NLP concepts that had already been extracted from the note, discharge summaries would be more likely to have been processed than other specialty note types (e.g., respiratory therapist notes), even if those specialty notes would be more likely to include documentation related oxygen use. Stratifying by billing codes for supplemental oxygen would ensure there is a mix of patients who did and did not receive oxygen compared with a purely random sample of hospitalized patients.

As a secondary activity, we also had each site run a query to quantify the notes that include keywords related to oxygen use (number of notes by type) for the patients in the COVID-19 cohort. This secondary analysis will allow a characterization of the degree of "missingness," as discharge summaries are not expected to contain all details related to oxygen use.

Cancer: The primary annotation task for this use case was to annotate selected concepts of interest found within the physician / clinic note that is associated with the visit where the patient was prescribed darzalex. The concepts of interest included those priority concepts that were likely to be present in the note, as well as those that were associated with the label for darzalex, which would allow us to determine whether the physician note contained sufficient information to determine the specific indication. The label for darzalex is shown in Figure 1 below, with the different concepts highlighted. The medication and diagnosis-related concepts could potentially be found in both structured and unstructured text, while those in blue would be expected to be primarily found within unstructured text.

SQL code for cohort selection

Draft SQL code was developed to assist in the process of identifying patients, but because the clinical notes were obtained from local systems and not any kind of CDM-based research repository, each site was able to tailor the process to fit their local environment and select the relevant notes. The draft SQL code provided basic logic and code sets that could be replicated in a local query or identification strategy. The example SQL code for the cohorts can be found in **Appendix A – SQL Code to select COVID-19 patients** and **Appendix C – SQL code to select darzalex patients**.



DARZALEX example

Medication-related concepts
Diagnosis-related concepts
Concepts that are expected to be primarily NLP-based

1. in combination with bortezomib, melphalan and prednisone for the treatment of patients with newly diagnosed multiple myeloma who are **ineligible for autologous stem cell transplant**

2. in combination with lenalidomide and dexamethasone, or bortezomib and dexamethasone, for the treatment of patients with multiple myeloma who have received at least one prior therapy *<list of candidate therapies required to define this part>*

3. in combination with pomalidomide and dexamethasone for the treatment of patients with multiple myeloma who have received at least two prior therapies including lenalidomide and a proteasome inhibitor

4. as monotherapy, *<exclude patients with concurrent candidate therapies>* for the treatment of patients with multiple myeloma who have received at least three prior lines of therapy including a proteasome inhibitor (PI) and an immunomodulatory agent or who **are double-refractory** to a PI and an immunomodulatory agent.

Figure 1: Labeled indications for darzalex. Medication-related concepts are highlighted in brown, diagnosis-related concepts in green. The concepts that are primarily expected to be extracted via NLP are highlighted in blue.

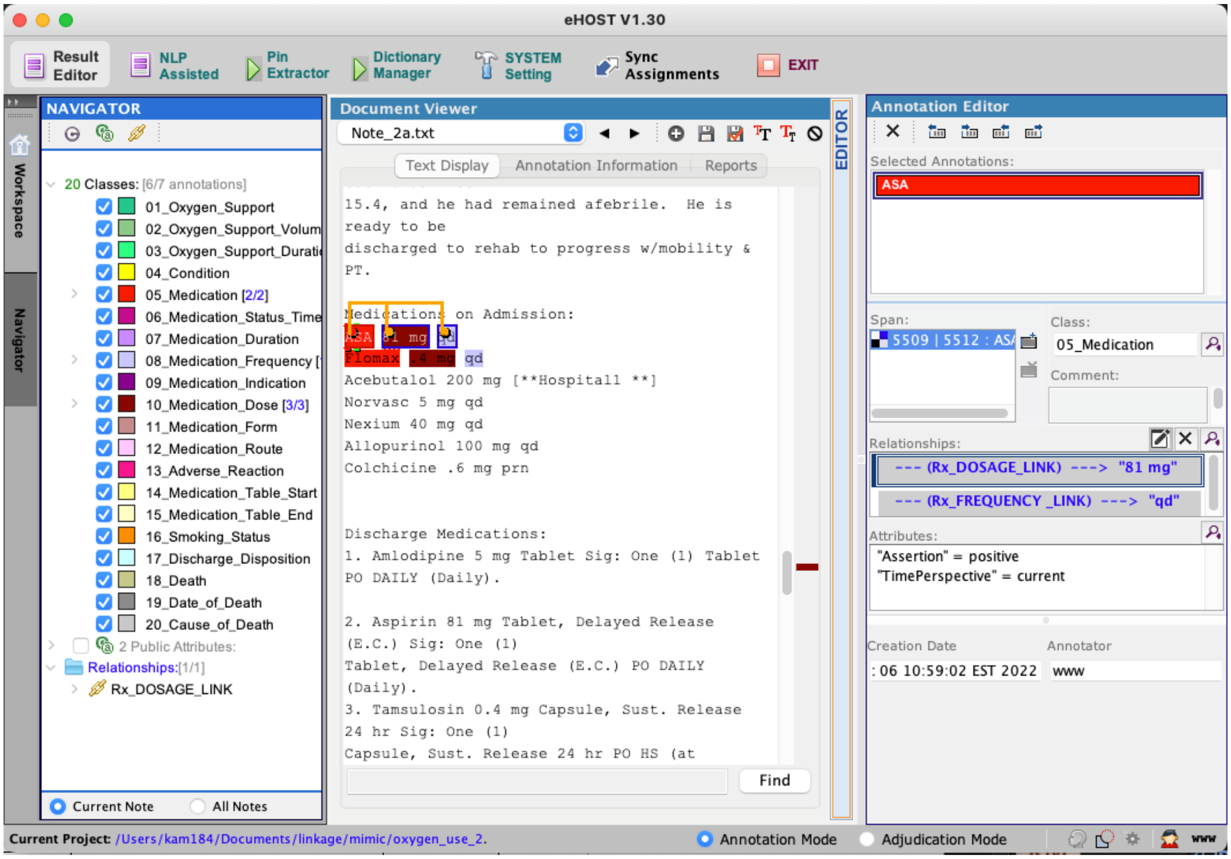
Annotation guide

Annotations were completed using the extensible Human Oracle Suite of Tools (eHOST) software package.[1] An annotation guide was drafted to assist in this task. Each of the main priority concepts (e.g., medications) were defined as a "primary class," with additional "attributes" to indicate other metadata (e.g., positive mention, historical / resolved status). Some primary classes also had associated "secondary" sub-classes (e.g., medication dose). When present in the same portion of the text, these primary and secondary classes could be linked with a relationship (see example in Figure 2 below). The annotation guide included examples of the text that would be included for each class as well as guidance on how to proceed in certain instances. It was then used to create a schema file that would configure eHOST so that the classes, attributes and relationships would appear in the user interface. The annotation guides can be found in **Appendix B – Annotation Guidelines for the COVID-19 (Oxygen Use) cohort** and **Appendix D – Annotation Guidelines for the Cancer (Darzalex)**

[1] Brett South, Shuying Shen, Jianwei Leng, Tyler Forbush, Scott DuVall, and Wendy Chapman. 2012. A Prototype Tool Set to Support Machine-Assisted Annotation. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*, pages 130–139, Montréal, Canada. Association for Computational Linguistics.

eHOST supports annotation of concepts, their attributes, and relations between concepts. The tool has the ability to 1) generate and pre-annotate texts using dictionaries, 2) provide machine-assisted annotation using real-time string machine to represent concepts, 3) support encoding of standard clinical vocabularies such as the UMLS and SNOMED-CT with fast API calls and query, and 4) compute inter-annotator agreement using flexible match criteria. (Description from https://github.com/chrisleng/ehost)

cohort, and example eHOST schemas in **Appendix F – XML schema for eHOST (COVID-19 / Oxygen Use annotation guidelines)** and **Appendix G – XML schema for eHOST (Darzalex annotation guidelines)**.



20 Classes: [6/7 annotations]

- ☑ 🟩 01_Oxygen_Support
- ☑ 🟩 02_Oxygen_Support_Volum
- ☑ 🟩 03_Oxygen_Support_Durati
- ☑ 🟨 04_Condition
- ☑ 🟥 05_Medication [2/2]
- ☑ 🟪 06_Medication_Status_Time
- ☑ 🟪 07_Medication_Duration
- ☑ ⬜ 08_Medication_Frequency [
- ☑ 🟪 09_Medication_Indication
- ☑ 🟥 10_Medication_Dose [3/3]
- ☑ 🟫 11_Medication_Form
- ☑ ⬜ 12_Medication_Route
- ☑ 🟥 13_Adverse_Reaction
- ☑ 🟨 14_Medication_Table_Start
- ☑ ⬜ 15_Medication_Table_End
- ☑ 🟧 16_Smoking_Status
- ☑ ⬜ 17_Discharge_Disposition
- ☑ 🟫 18_Death
- ☑ ⬛ 19_Date_of_Death
- ☑ ⬜ 20_Cause_of_Death

□ 2 Public Attributes:

Relationships:[1/1]

Rx_DOSAGE_LINK

Note_2a.txt

```
15.4, and he had remained afebrile.  He is
ready to be
discharged to rehab to progress w/mobility &
PT.

Medications on Admission:
ASA 81 mg qd
Flomax .4 mg qd
Acebutalol 200 mg [**Hospital1 **]
Norvasc 5 mg qd
Nexium 40 mg qd
Allopurinol 100 mg qd
Colchicine .6 mg prn


Discharge Medications:
1. Amlodipine 5 mg Tablet Sig: One (1) Tablet
PO DAILY (Daily).

2. Aspirin 81 mg Tablet, Delayed Release
(E.C.) Sig: One (1)
Tablet, Delayed Release (E.C.) PO DAILY
(Daily).
3. Tamsulosin 0.4 mg Capsule, Sust. Release
24 hr Sig: One (1)
Capsule, Sust. Release 24 hr PO HS (at
```

Selected Annotations: ASA

Span: 5509 | 5512 : AS/

Class: 05_Medication

Comment:

Relationships:

--- (Rx_DOSAGE_LINK) ---> "81 mg"

--- (Rx_FREQUENCY _LINK) ---> "qd"

Attributes:

"Assertion" = positive

"TimePerspective" = current

Creation Date: :06 10:59:02 EST 2022

Annotator: www

Current Project: /Users/kam184/Documents/linkage/mimic/oxygen_use_2.

● Annotation Mode ○ Adjudication Mode

Figure 2: Screenshot of the eHOST software with an annotation for a medication (ASA), with secondary classes for dose (81 mg) and frequency (qd). Relationships between the classes have also been defined (RX_DOSAGE_LINK and RX_FREQUENCY_LINK). Note that for this example, medications present on admission are defined as "current."

The COVID-19 (oxygen use) annotation guide was developed first. An initial draft was created and each team of annotators from the two Data Partner sites was asked to annotate several notes to determine if they had questions or if they found example text that was not well-described within the document. The guide went through several iterations, first using de-identified discharge summaries from the MIMIC-III (Medical Information Mart for Intensive Care) database while regulatory approval was pending,[2] and then with local examples from the COVID-19 cohort. Once the annotators were satisfied that the guide provided enough

[2] MIMIC (Medical Information Mart for Intensive Care) is a large, freely available database comprising deidentified health-related data from patients who were admitted to the critical care units of the Beth Israel Deaconess Medical Center (from https://mimic.mit.edu/docs/about/).

Johnson, A., Pollard, T., & Mark, R. (2016). MIMIC-III Clinical Database (version 1.4). *PhysioNet*. https://doi.org/10.13026/C2XW26.
Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. Scientific Data, 3, 160035.

instructions on how to proceed, it was considered ready for testing whether the annotators were sufficiently trained, as measured by the agreement rate between them. The cancer annotation guide was developed after the COVID-19 guide. The definitions for overlapping concepts were reused, with the focus on defining those concepts specific to the cancer use case. These cancer concepts went through the same iterative process to refine the annotation guide. The notes used in the development of the annotation guide were not utilized in the actual annotation task.

Annotation strategy

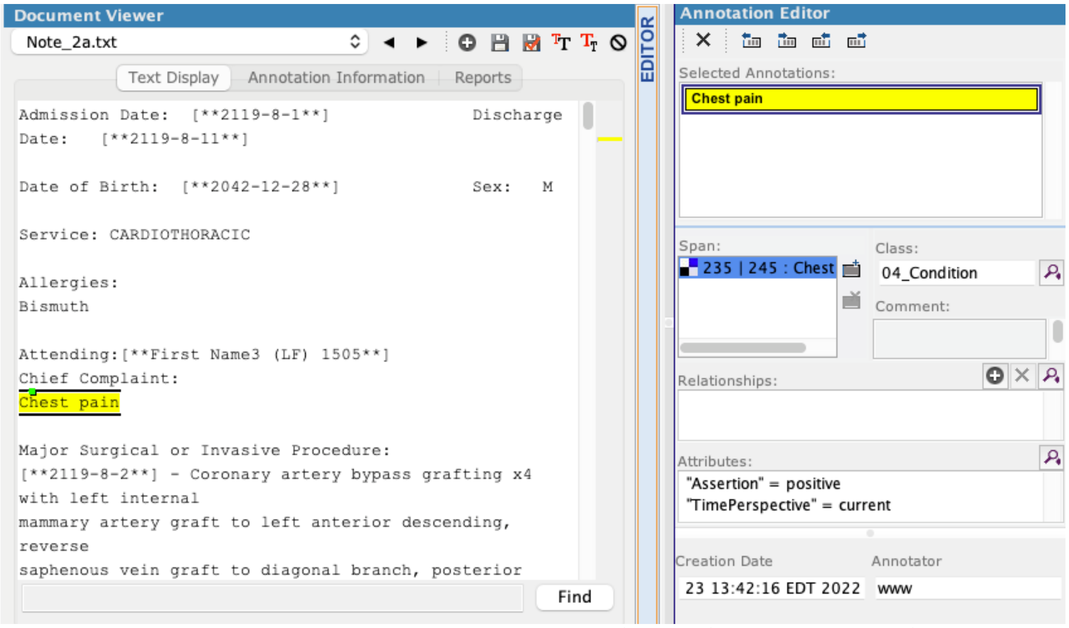
Within each cohort, each team of two annotators was asked to double-annotate a set of notes (5-6 in total) and compute the inter-annotator agreement at the class / attribute level. If the overall percentage was above 80%, that site could proceed and single-annotate the remainder of the notes within that cohort. If the percentage was less than ~80%, the annotation guide was clarified to help resolve any discrepancies and the notes annotated again, allowing inter-annotator agreement to be recomputed. If the percentage was still less than 80%, additional notes were selected, and the process repeated.

Once the annotation was complete, each site ran a procedure that would remove all of the potentially identifying personal health information from the XML annotation file that was generated by eHOST (e.g., text snippets highlighted during the annotation process – beyond elements like date of birth/death or dates of service, there is a small theoretical risk that an annotation could inadvertently contain personal health information, so institutions are leery about sharing the raw files). An example can be found in Figure 3 and Figure 4 below. These redacted files were shared with the team at Duke for final analysis. A manually curated list of text snippets was later generated by each site for their COVID-19 cohort.

Analysis

Descriptive statistics were computed on the single-annotation notes, summarizing the characteristics of the “primary” class concepts and the associated attributes across Data Partner sites, along with the secondary class concepts. For the primary medication class, we also computed statistics on the presence of associated secondary classes (e.g., how often was dose recorded alongside medication, alone and in combination with route, frequency, etc.).

For the COVID-19 cohort, we also report on the other non-discharge summary notes that contain oxygen keywords (number by note type). We also provide an aggregated table of text snippets for each of the classes.



```
Admission Date:  [**2119-8-1**]              Discharge
Date:   [**2119-8-11**]

Date of Birth:  [**2042-12-28**]             Sex:   M

Service: CARDIOTHORACIC

Allergies:
Bismuth

Attending:[**First Name3 (LF) 1505**]
Chief Complaint:
Chest pain

Major Surgical or Invasive Procedure:
[**2119-8-2**] - Coronary artery bypass grafting x4
with left internal
mammary artery graft to left anterior descending,
reverse
saphenous vein graft to diagonal branch, posterior
```

Chest pain

Span: 235 | 245 : Chest

Class: 04_Condition

Comment:

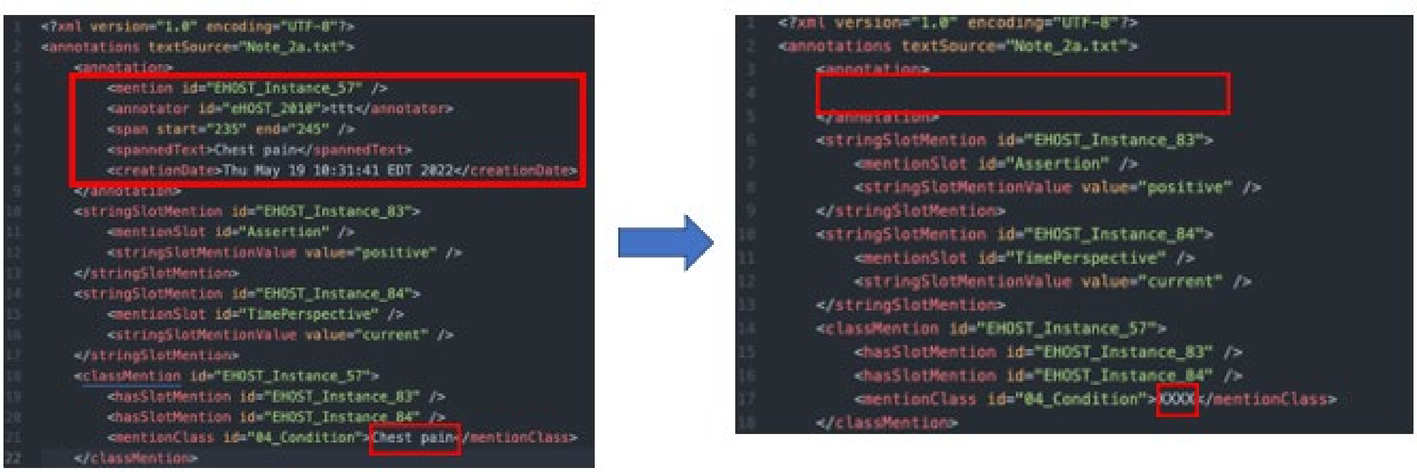
Relationships:

Attributes:
"Assertion" = positive
"TimePerspective" = current

Creation Date: 23 13:42:16 EDT 2022

Annotator: www

Figure 3: Example annotation within the eHOST software for a Condition of "chest pain", and an assertion of "positive" and a time perspective of "current."



```
<?xml version="1.0" encoding="UTF-8"?>
<annotations textSource="Note_2a.txt">
    <annotation>
        <mention id="EHOST_Instance_57" />
        <annotator id="eHOST_2010">ttt</annotator>
        <span start="235" end="245" />
        <spannedText>Chest pain</spannedText>
        <creationDate>Thu May 19 10:31:41 EDT 2022</creationDate>
    </annotation>
    <stringSlotMention id="EHOST_Instance_83">
        <mentionSlot id="Assertion" />
        <stringSlotMentionValue value="positive" />
    </stringSlotMention>
    <stringSlotMention id="EHOST_Instance_84">
        <mentionSlot id="TimePerspective" />
        <stringSlotMentionValue value="current" />
    </stringSlotMention>
    <classMention id="EHOST_Instance_57">
        <hasSlotMention id="EHOST_Instance_83" />
        <hasSlotMention id="EHOST_Instance_84" />
        <mentionClass id="04_Condition">Chest pain</mentionClass>
    </classMention>
```

```
<?xml version="1.0" encoding="UTF-8"?>
<annotations textSource="Note_2a.txt">
    <annotation>

    </annotation>
    <stringSlotMention id="EHOST_Instance_83">
        <mentionSlot id="Assertion" />
        <stringSlotMentionValue value="positive" />
    </stringSlotMention>
    <stringSlotMention id="EHOST_Instance_84">
        <mentionSlot id="TimePerspective" />
        <stringSlotMentionValue value="current" />
    </stringSlotMention>
    <classMention id="EHOST_Instance_57">
        <hasSlotMention id="EHOST_Instance_83" />
        <hasSlotMention id="EHOST_Instance_84" />
        <mentionClass id="04_Condition">XXXX</mentionClass>
    </classMention>
```

Figure 4: Annotation represented in XML output by eHOST (left). Potentially sensitive details are highlighted in red. Information related to the text snippet was removed before analysis (right).

Results

Cohort selection

Generic SQL code was provided to each site to aid in the cohort selection process, but each site had to tailor the approach to their local environment. As an example, the BWH SQL code for the cancer cohort can be found in **Appendix H – BWH Code for Cancer Cohort**.

Within the COVID-19 cohort, each site annotated hospital discharge summaries, which are a standard note type across institutions. The definition of the cancer cohort allowed for slight nuance in the notes that were selected. BWH chose Progress Notes that mentioned darzalex but varied by clinic specialty. The breakdown can be found in Table 2 below. VUMC primarily selected notes authored by Nurse Practitioners, Physician Assistants, or Physicians in an oncology clinic. This selection was done through the OHDSI instance of the VUMC EHR data, using the visit metadata and associating note author IDs with healthcare roles from the National Uniform Claims Committee vocabulary and Care Site IDs (clinic type).

Table 2: Breakdown of specialty and note type for the BWH cancer cohort (n=30).

Specialty	Note Type	Count
Hematology Oncology	Progress Note	2
Infusion Therapy	Progress Note	15
Medical Oncology	Progress Note	4
Medicine	Progress Note	1
Myeloma	Progress Note	2
Nursing	Plan of Care	2
Nursing	Plan of Care	1
Oncology	Plan of Care	1
Oncology	Progress Note	1
Pharmacy	Progress Note	1

Inter-annotator agreement

The results for the inter-annotator agreement at each site are provided for the COVID-19 cohort in Table 2 below. For the cancer cohort, both Data Partner sites ended up below the 80% threshold on their first pass through the notes and needed to re-annotate after revising the guidelines, but both ended up with >80% agreement at the classes + attribute level.

Table 3: Inter-annotator agreement for BWH and VUMC on the COVID-19 cohort at the class, class + attribute and class + attribute + relationship level (reported as %)

	BWH	VUMC
Classes	89.0	87.5
Classes + attributes	80.1	79.7
Classes + attributes + relationship	76.6	71.9

Statistics on annotation concepts

Various statistics from the single-annotated notes are provided in the tables below. Table 4 provides a summary of the non-medication concepts for the COVID-19 cohort at BWH and VUMC. The cohort numbers are broken out into two categories: patients WITH a billing code for supplemental oxygen (oxygen cohort) during the hospitalization, and patients WITHOUT a billing code (non-oxygen cohort). For each of these sub-cohorts, we report the number of notes that have an instance of a given concept (or concept / attribute combination) and well as the total number of annotations (to improve readability, the counts for attribute values of "unknown" and "hypothetical" were combined into a single option in the table due to their relatively low numbers. From an annotation perspective, we acknowledge that they have different meanings). We report these numbers as percentages as well. The concepts "template start" and "template end" were defined to denote the start and end of any structured or semi-structured data that would have been pulled into the discharge summary from elsewhere in the EHR (e.g., laboratory results or administered medications). These structured tables add significant burden to the annotation task, and since they exist elsewhere in the EHR as queryable data, we chose not to spend resources on their annotation. Table 5 provides similar statistics about the number of notes and overall number of annotations for the medication-related concepts. These include the overall mentions of a medication and potential "secondary" classes that describe other attributes about the medication (e.g., dose, route, form, frequency, etc.). The "adverse reaction" class was defined to note instances where a patient reports an issue with a medication or a reason for not taking it (e.g., allergic, causes hives).

To quantify the degree that oxygen use might be mentioned in other parts of the clinical narrative, we asked Data Partner sites to count the number of times that the oxygen keywords (identified through the annotation exercise) were present in the rest of the unstructured text associated with the patient's hospitalization. These mentions are summarized in Table 6. To quantify the number of different *note types* that might contain a mention of oxygen use, we also counted the number of times each oxygen keyword was present in a note, by note type. These numbers are shown in Table 7. Both Tables 6 and 7 report numbers from BWH.

In order to illustrate the different phrases that are present in the discharge summaries, we also report the top 5 text snippets for each annotation concept (by number of mentions). These terms are reported in Table 8. While we ignore case when grouping terms, for the purpose of this exercise, different spellings of the same underlying phrase (e.g., COVID and COVID-19) are treated as different terms.

The annotations of the cancer cohort are summarized in Tables 9 and 10 (non-medication concepts and medication-related concepts, respectively). To help illustrate the degree that medication metadata (e.g., dose and route) in a medication annotation, we have summarized the medication annotations and noted the number of times that a given attribute is present or absent as part of that annotation (e.g., none present, all present, dose and route only, etc.). The top combinations are reported for the VUMC cancer annotations in Table 11. Note that because we did not annotate structured templates within the clinical note (e.g., medications pulled from the patient med list), these numbers may be somewhat of an undercount.

Table 4: Characteristics of non-medication concepts/attributes across patients in the COVID-19 cohort for BWH and VUMC.

	Patients **WITH** billing codes for supplemental oxygen				Patients **WITHOUT** billing codes for supplemental oxygen			
	BWH		**VUMC**		**BWH**		**VUMC**	
Concept	**Notes with at least 1 annotation of the concept (N=27) total notes**	**Total annotations (N=2,719)**	**Notes with at least 1 annotation of the concept (N=35) total notes**	**Total annotations (N=3,120)**	**Notes with at least 1 annotation of the concept (N=33) total notes**	**Total annotations (N=3,022)**	**Notes with at least 1 annotation of the concept (N=35) total notes**	**Total annotations (N=2,243)**
Oxygen Support	27 (100.0%)	102 (3.8%)	31 (88.6%)	210 (6.7%)	28 (84.8%)	90 (3.0%)	16 (45.7%)	54 (2.4%)
Assertion								
Negative	16 (59.3%)	27 (26.5%)	N/A	N/A	18 (64.3%)	25 (27.8%)	7 (43.8%)	7 (13.0%)
Positive	23 (85.2%)	73 (71.6%)	31 (100.0%)	209 (99.5%)	15 (53.6%)	63 (70.0%)	11 (68.8%)	42 (77.8%)
Unknown/Hypothetical	2 (7.4%)	2 (2.0%)	1 (3.2%)	1 (0.5%)	2 (7.1%)	2 (2.2%)	2 (12.5%)	5 (3.7%)
Time perspective								
Current	27 (100.0%)	91 (89.2%)	30 (96.8%)	190 (90.5%)	27 (96.4%)	82 (91.1%)	15 (93.8%)	46 (85.2%)
History	7 (25.9%)	9 (8.8%)	9 (29.0%)	16 (7.6%)	4 (14.3%)	6 (6.7%)	4 (25.0%)	6 (11.1%)
Predicted	2 (7.4%)	2 (2.0%)	4 (12.9%)	4 (1.9%)	2 (7.1%)	2 (2.2%)	2 (12.5%)	2 (3.7%)
Change Status								
Change	6 (22.2%)	8 (7.8%)	17 (54.8%)	45 (21.4%)	3 (10.7%)	3 (3.3%)	5 (31.3%)	10 (18.5%)
Singular	27 (100.0%)	94 (92.2%)	31 (100.0%)	165 (78.6%)	27 (96.4%)	87 (96.7%)	16 (100.0%)	44 (81.5%)
Other O2 indicators								
Volume	25 (92.6%)	43 (42.2%)	22 (71.0%)	75 (35.7%)	14 (50.0%)	38 (42.2%)	9 (56.3%)	26 (48.1%)
Duration	N/A	N/A	11 (35.5%)	14 (6.7%)	1 (3.6%)	1 (1.1%)	5 (31.3%)	8 (14.8%)
Condition	27 (100.0%)	1,667 (61.3%)	35 (100.0%)	2,152 (69.0%)	33 (100.0%)	1,911 (63.2%)	35 (100.0%)	1,394 (62.1%)
Assertion								
Hypothetical	15 (55.6%)	194 (11.6%)	21 (60.0%)	169 (7.9%)	17 (51.5%)	145 (7.6%)	15 (42.9%)	107 (7.7%)
Negative	24 (88.9%)	312 (18.7%)	32 (91.4%)	364 (16.9%)	30 (90.9%)	299 (15.6%)	33 (94.3%)	297 (21.3%)
Positive	27 (100.0%)	1,129 (67.7%)	35 (100.0%)	1,580 (73.4%)	33 (100.0%)	1,420 (74.3%)	35 (100.0%)	964 (69.2%)
Unknown	13 (48.1%)	32 (1.9%)	16 (45.7%)	39 (1.8%)	14 (42.4%)	47 (2.5%)	11 (31.4%)	26 (1.9%)
Time perspective								
Current	27 (100.0%)	1,459 (87.5%)	35 (100.0%)	2,002 (93.0%)	33 (100.0%)	1,702 (89.1%)	35 (100.0%)	1,308 (93.8%)
History	14 (51.9%)	50 (3.0%)	25 (71.4%)	150 (7.0%)	19 (57.6%)	73 (3.8%)	32 (91.4%)	85 (6.1%)
Predicted	14 (51.9%)	158 (9.5%)	N/A	N/A	14 (42.4%)	136 (7.1%)	1 (2.9%)	1 (0.1%)
Smoking Status	2 (7.4%)	3 (0.1%)	1 (2.9%)	1 (0.0%)	1 (3.0%)	1 (0.0%)	3 (8.6%)	3 (0.1%)
Assertion								
Negative	1 (50.0%)	1 (33.3%)	N/A	N/A	N/A	N/A	N/A	N/A
Positive	1 (50.0%)	2 (66.7%)	1 (100.0%)	1 (100.0%)	1 (100.0%)	1 (100.0%)	3 (100.0%)	3 (100.0%)
Time perspective								
Current	2 (100.0%)	3 (100.0%)	N/A	N/A	1 (100.0%)	1 (100.0%)	3 (100.0%)	3 (100.0%)
History	N/A	N/A	1 (100.0%)	1 (100.0%)	N/A	N/A	N/A	N/A
Death	0 (0.0%)	0 (0.0%)	1 (2.9%)	2 (0.1%)	2 (6.1%)	3 (0.1%)	0 (0.0%)	0 (0.0%)
Date of Death	N/A	N/A	N/A	N/A	2 (100.0%)	3 (100.0%)	N/A	N/A
Discharge Disposition	18 (66.7%)	36 (1.3%)	32 (91.4%)	40 (1.3%)	26 (78.8%)	55 (1.8%)	34 (97.1%)	62 (2.9%)
Template Start	27 (100.0%)	64 (2.4%)	35 (100.0%)	57 (1.8%)	30 (90.9%)	65 (2.2%)	35 (100.0%)	138 (6.5%)
Template End	27 (100.0%)	64 (2.4%)	34 (97.1%)	57 (1.8%)	30 (90.9%)	64 (2.1%)	35 (100.0%)	138 (6.5%)

Table 5: Characteristics of medication-related concepts / attributes across patients in the COVID-19 cohort for BWH and VUMC.

	Patients **WITH** a billing code for supplemental oxygen				Patients **WITHOUT** a billing code for supplemental oxygen			
	BWH		**VUMC**		**BWH**		**VUMC**	
Secondary	**Notes with at least 1 annotation of the concept (N=27) total notes**	**Total annotations (N=2,719)**	**Notes with at least 1 annotation of the concept (N=35) total notes**	**Total annotations (N=3,120)**	**Notes with at least 1 annotation of the concept (N=33) total notes**	**Total annotations (N=3,022)**	**Notes with at least 1 annotation of the concept (N=35) total notes**	**Total annotations (N=2,243)**
Medication	27 (100.0%)	541 (19.9%)	35 (100.0%)	447 (14.3%)	33 (100.0%)	571 (18.9%)	35 (100.0%)	304 (14.3%)
Assertion								
Negative	16 (59.3%)	50 (9.2%)	8 (22.9%)	12 (2.7%)	23 (69.7%)	65 (11.4%)	13 (37.1%)	20 (6.6%)
Positive	25 (92.6%)	198 (36.6%)	31 (88.6%)	197 (44.1%)	29 (87.9%)	205 (35.9%)	27 (77.1%)	104 (34.2%)
Unknown/Hypothetical	10 (37.0%)	26 (4.8%)	2 (5.7%)	18 (4.0%)	11 (33.3%)	30 (5.3%)	1 (2.9%)	15 (0.3%)
Time perspective								
Current	25 (92.6%)	228 (42.1%)	29 (82.9%)	161 (36.0%)	31 (93.9%)	258 (45.2%)	29 (82.9%)	107 (35.2%)
History	7 (25.9%)	21 (3.9%)	22 (62.9%)	61 (13.6%)	5 (15.2%)	13 (2.3%)	13 (37.1%)	26 (8.6%)
Predicted	9 (33.3%)	25 (4.6%)	5 (14.3%)	5 (1.1%)	7 (21.2%)	29 (5.1%)	3 (8.6%)	6 (2.0%)
Medication attributes								
Medication timing	19 (70.4%)	46 (8.5%)	7 (20.0%)	14 (3.1%)	16 (48.5%)	41 (7.2%)	8 (22.9%)	14 (4.6%)
Medication duration	18 (66.7%)	46 (8.5%)	13 (37.1%)	32 (7.2%)	16 (48.5%)	34 (6.0%)	12 (34.3%)	24 (7.9%)
Medication frequency	16 (59.3%)	78 (14.4%)	11 (31.4%)	35 (7.8%)	17 (51.5%)	70 (12.3%)	14 (40.0%)	25 (8.2%)
Medication indication	16 (59.3%)	49 (9.1%)	27 (77.1%)	72 (16.1%)	20 (60.6%)	65 (11.4%)	21 (60.0%)	65 (21.4%)
Medication dose	7 (25.9%)	16 (3.0%)	4 (11.4%)	4 (0.9%)	17 (51.5%)	29 (5.1%)	5 (14.3%)	5 (1.6%)
Medication form	6 (22.2%)	18 (3.3%)	4 (11.4%)	4 (0.9%)	7 (21.2%)	12 (2.1%)	4 (11.4%)	4 (1.3%)
Medication route	5 (18.5%)	10 (1.8%)	12 (34.3%)	20 (4.5%)	9 (27.3%)	17 (3.0%)	2 (5.7%)	2 (0.7%)
Adverse Reaction	2 (7.4%)	4 (0.1%)	17 (48.6%)	43 (1.4%)	4 (12.1%)	4 (0.1%)	34 (97.1%)	68 (3.0%)
Assertion								
Negative	N/A	N/A	N/A	N/A	N/A	N/A	16 (47.1%)	18 (26.5%)
Positive	2/2 (100.0%)	4 (100.0%)	17 (100.0%)	41 (95.3%)	N/A	N/A	19 (55.9%)	50 (73.5%)
Unknown	N/A	N/A	2 (11.8%)	2 (4.7%)	4 (100.0%)	4 (100.0%)	N/A	N/A
Time perspective								
Current	2/2 (100.0%)	4 (100.0%)	17 (100.0%)	42 (97.7%)	4 (100.0%)	4 (100.0%)	34 (100.0%)	39 (57.4%)
History	N/A	N/A	1 (5.9%)	1 (2.3%)	N/A	N/A	12 (35.3%)	29 (42.3%)

Table 6: Number of mentions of terms used to denote supplemental oxygen across all notes associated with the COVID-19 cohort at BWH. Results broken out by oxygen and non-oxygen cohorts. Terms with fewer than 50 mentions across either cohort are grouped into the "Other" category.

Terms used to denote use of supplemental oxygen	Oxygen Cohort	Non-oxygen Cohort
o2	1564	1242
nc	621	398
oxygen	594	429
o2 requirement	166	125
nasal cannula	155	105
hfnc	139	11
o2 device	130	134
oxygen requirement	109	78
supplemental oxygen	98	65
intubation	95	147
supplemental o2	90	87
t-piece	86	0
oxymizer	68	17
o2 device: none	49	61
Other mention of supplemental oxygen (15 distinct terms)	189	210

Table 7: Mentions of oxygen terms by note type across all notes associated with the hospitalization of patients in the COVID-19 cohort at BWH. Results broken out by Oxygen and Non-oxygen cohorts. Note types with fewer than 10 mentions in either cohort are grouped into the "Other" category

Note Type	Oxygen Cohort	Non-oxygen Cohort
Progress Notes	861	954
Assessment & Plan Note	114	110
Consults	67	63
Plan of Care	103	61
ED Provider Notes	65	60
ED Notes	47	44
H&P	43	42
Discharge Summary	36	36
ED Triage Notes	33	28
Nursing Summary	18	16
Research Notes	10	14
Discharge Instructions	7	12
ACP (Advance Care Planning)	0	10
ED Progress/Update Note	2	10
CDI Query	11	9
Other Note Types (26 distinct values)	63	58

Table 8: Top terms (by number of mentions) in each of the Concept classes for the COVID-19 cohort at BWH and VUMC.

Concept Class	**BWH**	**VUMC**
Oxygen support	O2, nc, oxygen, o2 device, supplemental oxygen	BiPAP, nasal canula, O2, oxygen, extubated
Oxygen support volume	2l, (l/min): 1, 4l, 1l, 2 l	15L, 2L, 3L, with ambulation (# of L/min): 4, 6L
Condition	COVID-19, COVID, cough, fever, shortness of breath	COVID, COVID-19, cough, HTN, hypoxia
Medication	Dexamethasone, remdesivir, metformin, Tylenol, antibiotics	Remdeivir, dexamtethasone, lovenox, decadron, lasix
Medication duration	5 days, 1 week, 4 days, 10 days, 10-day course	3 days, 7 days, 5 days, 10 days, x3 days
Medication frequency	Daily, bid, prn, qd, nightly	BID, daily, TID, PRN, Q12h
Medication indication	COVID, cough, COVID-19, fevers, CAP	COVID, pain, Afib, anticoagulation, DVT
Medication dose	6mg, 5mg, 40mg, 20mg, 10mg	10mg, 80mg, 50mg, 120mg, 5mg
Medication form	Inhaler, gtt, patch, injections, mdi	Gtt, infusion, eyedrop, injection, patch
Medication route	IV, PO, Oral, IM, SQ	IV, PO, Oral, intravenous, topical
Adverse reaction	Allergies, gi side effects, drug allergies, flushing	Allergies, rash, hives, itching, mouth ulcers
Smoking status	Smoking, smoker, tobacco, extensive smoking history	Smoker, smoking
Discharge disposition	Home, home or self-care, home with services, self-care, skilled nursing facility	Home or self-care, home, skilled nursing facility, SNF, Hospice Scatter Bed

Table 9: Characteristics of non-medication concepts for the cancer cohort between BWH and VUMC.

Concept	BWH		VUMC	
	Notes with at least 1 annotation of the concept (N=24) total notes	Total annotations (N=1,354)	Notes with at least 1 annotation of the class characteristic (N=28) total notes	Total annotations (N=7,379)
Cancer/Tumor	8 (33.3%)	51 (3.8%)	28 (100.0%)	188 (2.6%)
Assertion				
Negative	1 (12.5%)	1 (2.0%)	2 (7.1%)	2 (1.1%)
Positive	8 (100.0%)	46 (90.2%)	28 (100.0%)	179 (95.2%)
Unknown/Hypothetical	1 (12.5%)	4 (3.9%)	5 (17.9%)	7 (3.7%)
Time perspective				
Current	8 (100.0%)	47 (92.2%)	28 (100.0%)	173 (92.0%)
History	1 (12.5%)	2 (3.9%)	6 (21.4%)	15 (8.0%)
Predicted	1 (12.5%)	2 (3.9%)	N/A	N/A
Other cancer indicator				
Cancer Stage[1]	0 (0.0%)	0 (0.0%)	21 (75.0%)	10 (5.3%)
Stem Cell Transplant	2 (8.3%)	4 (0.3%)	22 (78.6%)	87 (1.2%)
Assertion				
Negative	N/A	N/A	22 (9.1%)	3 (3.4%)
Positive	2 (100.0%)	4 (100.0%)	192 (86.4%)	67 (77.0%)
Unknown/Hypothetical	N/A	N/A	72 (31.8%)	17 (19.5%)
Time perspective				
Current	2 (100.0%)	2 (50.0%)	192 (86.4%)	55 (63.2%)
History	1 (50.0%)	2 (50.0%)	152 (68.2%)	26 (29.9%)
Predicted	N/A	N/A	32 (13.6%)	6 (6.9%)
Refractory	2 (8.3%)	2 (0.1%)	8 (28.6%)	15 (0.2%)
Assertion				
Negative	N/A	N/A	5 (62.5%)	5 (33.3%)
Positive	2 (100.0%)	2 (100.0%)	7 (87.5%)	9 (60.0%)
Unknown/Hypothetical	N/A	N/A	1 (12.5%)	1 (6.7%)
Time perspective				
Current	2 (100.0%)	2 (100.0%)	5 (62.5%)	6 (40.0%)
History	N/A	N/A	6 (75.0%)	9 (60.0%)
Gene/Protein	6 (25.0%)	251 (18.5%)	27 (96.4%)	114 (1.5%)
Assertion				
Negative	2 (33.3%)	18 (7.2%)	2 (7.4%)	2 (1.8%)
Positive	6 (100.0%)	233 (92.8%)	27 (100.0%)	112 (98.2%)
Time perspective				
Current	6 (100.0%)	249 (99.2%)	27 (100.0%)	103 (90.4%)
History	2 (33.3%)	2 (0.8%)	5 (18.5%)	12 (10.5%)
Condition	18 (75.0%)	350 (25.8%)	28 (100.0%)	1,795 (24.3%)
Negative	8 (44.4%)	114 (32.6%)	27 (96.4%)	739 (41.2%)
Positive	18 (100.0%)	204 (58.3%)	28 (100.0%)	1,038 (57.8%)
Unknown/Hypothetical	1 (5.6%)	33 (9.4%)	9 (32.1%)	31 (17.2%)
Time perspective				
Current	18 (100.0%)	292 (83.4%)	28 (100.0%)	1,427 (79.5%)
History	5 (27.8%)	39 (11.1%)	24 (85.7%)	372 (20.7%)
Predicted	3 (16.7%)	20 (5.7%)	1 (3.6%)	1 (0.1%)
Test/Procedure	16 (66.7%)	132 (9.7%)	28 (100.0%)	2,820 (38.2%)
Assertion				
Negative	2 (12.5%)	2 (1.5%)	18 (64.3%)	57 (2.0%)
Positive	16 (100.0%)	126 (95.5%)	28 (100.0%)	2,718 (96.4%)
Unknown/Hypothetical	3 (18.8%)	4 (3.0%)	9 (32.1%)	45 (15.9%)
Time perspective				
Current	16 (100.0%)	98 (74.2%)	28 (100.0%)	768 (27.2%)
History	3 (18.8%)	8 (6.1%)	26 (92.9%)	1,968 (69.8%)
Predicted	6 (37.5%)	26 (19.7%)	13 (46.4%)	84 (3.0%)
Smoking Status	2 (8.3%)	3 (0.2%)	6 (21.4%)	6 (0.1%)
Assertion				
Negative	1 (50.0%)	1 (33.3%)	6 (100.0%)	6 (100.0%)
Positive	1 (50.0%)	2 (66.7%)	N/A	N/A
Time perspective				
Current	1 (50.0%)	1 (33.3%)	N/A	N/A
History	1 (50.0%)	2 (66.7%)	6 (100.0%)	6 (100.0%)
Template Start	5 (20.8%)	7 (0.5%)	27 (96.4%)	67 (0.9%)
Template End	5 (20.8%)	7 (0.5%)	27 (96.4%)	66 (0.9%)

[1] Cancer Stage was only found in the training data at BWH.

Table 10: Characteristics of medication-related concepts for the cancer cohort between BWH and VUMC.

	BWH		VUMC	
Medication-related concepts	Notes with at least 1 annotation of the class characteristic (N=24) total notes	Total annotations (N=1,354)	Notes with at least 1 annotation of the class characteristic (N=28) total notes	Total annotations (N=7,379)
Medication	24 (100.0%)	548 (40.5%)	28 (100.0%)	2,364 (32.0%)
Assertion				
Negative	10 (41.7%)	4 (8.0%)	23 (82.1%)	94 (4.0%)
Positive	24 (100.0%)	473 (86.3%)	28 (100.0%)	2,153 (91.1%)
Unknown/Hypothetical	3 (12.5%)	31 (5.7%)	10 (35.7%)	117 (4.9%)
Time perspective				
Current	24 (100.0%)	333 (60.8%)	28 (100.0%)	828 (35.0%)
History	8 (33.3%)	152 (27.7%)	26 (92.9%)	1,360 (57.5%)
Predicted	10 (41.7%)	63 (11.5%)	17 (60.7%)	176 (7.4%)
Medication Attributes				
Timing	20 (83.3%)	141 (25.7%)	27 (96.4%)	802 (33.9%)
Duration	12 (50.0%)	26 (4.7%)	23 (82.1%)	183 (7.7%)
Frequency	12 (50.0%)	42 (7.7%)	25 (89.3%)	405 (17.1%)
Indication	7 (29.2%)	11 (2.0%)	25 (89.3%)	137 (5.8%)
Dose	19 (79.2%)	89 (16.2%)	23 (82.1%)	498 (21.1%)
Form	14 (58.3%)	23 (4.2%)	7 (25.0%)	13 (0.5%)
Route	16 (66.7%)	42 (7.7%)	19 (67.9%)	189 (8.0%)
Adverse Reaction	9 (37.5%)	17 (1.3%)	25 (89.3%)	87 (1.2%)
Assertion				
Negative	2 (22.2%)	2 (11.8%)	3 (12.0%)	3 (3.4%)
Positive	4 (44.4%)	5 (29.4%)	24 (96.0%)	83 (95.4%)
Unknown/Hypothetical	4 (44.4%)	10 (58.8%)	1 (4.0%)	1 (1.1%)
Time perspective				
Current	5 (55.6%)	7 (41.2%)	25 (100.0%)	68 (78.2%)
History	N/A	N/A	14 (56.0%)	31 (35.6%)
Predicted	4 (44.4%)	10 (58.8%)	N/A	N/A

Table 11: Co-occurrence of medication-related metadata (e.g., dose, timing) within an annotation for a medication in the VUMC cancer cohort (n=2,364). '1' indicates presence, '0' denotes absence. Combinations with a percentage below 1% of the total are not shown (n=35).

Timing	Duration	Frequency	Indication	Dose	Form	Route	FREQUENCY	Percent
0	0	0	0	0	0	0	1056	44.67
1	0	0	0	0	0	0	526	22.25
0	0	0	1	0	0	0	77	3.26
1	0	0	0	1	0	0	76	3.21
0	0	1	0	1	0	0	69	2.92
0	0	0	0	1	0	0	64	2.71
0	0	1	0	0	0	0	62	2.62
0	0	1	0	1	0	1	52	2.20
1	0	1	0	1	0	1	48	2.03
1	0	1	0	1	0	0	45	1.90
0	1	0	0	0	0	0	44	1.86
1	0	1	0	0	0	0	24	1.02

Discussion

When looking at the annotations for the COVID-19 cohort (Tables 4 and 5), most of the primary concepts (e.g., medications, conditions) are present in almost every note across both Data Partner sites, and the percentage of overall annotations that correspond to those concepts are also similar (e.g., conditions ~60%, medications ~15-20%). When there are differences across Data Partner sites, such as with discharge disposition, the percentages within site between the oxygen and non-oxygen groups are more similar, which may be an indicator of the difference in documentation patterns. This can be seen in the results of the medication-related concepts, where BWH has consistently higher documentation of the medication-related metadata (e.g., dose, route), which holds across both groups. On the other hand, VUMC is more likely to document adverse reactions or reasons for not taking a medication (e.g., allergies). Some of the biggest differences across Data Partner sites are seen in the overall level of documentation of oxygen support, though there are again similarities when looking just at positive mentions of oxygen support. For instance, 100% of the patients in the BWH oxygen cohort have an indication of oxygen use, with 85% having a positive mention. Within the VUMC oxygen cohort, 88.6% have a mention of oxygen, with all having a positive mention. In the non-oxygen cohort, 84.8% of BWH patients have a mention of oxygen use, compared with 45.7% among VUMC patients. This is a fairly large difference, but again looking at positive mentions, there is a smaller spread. Approximately 45% of BWH patients have a positive mention of oxygen support (n=15/33), and ~31% of VUMC patients (n=11/35). This illustrates both the potential drawbacks of relying on billing codes to indicate oxygen use and the fact that documentation of positive/current items may be more consistent across Data Partner sites than negative mentions.

Looking at the mentions of oxygen use across note type (Table 7), we can see the wide variety of notes within the EHR where oxygen can be documented. One key takeaway from this analysis is that it is important to understand the underlying documentation practices in order to select the correct notes. Manual review / annotation is typically not feasible for such a large number of notes, particularly over a large population, and even automated processing can be a challenge depending on the complexity of the extraction task. This is especially true for more novel concepts that are not part of existing pipelines. Therefore, engagement with potential data partners and clinical experts is key in order to make sure that the relevant content is targeted for any downstream analysis.

For the Cancer cohort, there were some differences in the notes that were selected, particularly the underlying specialty, which led to different levels of information density (Tables 9 and 10). This is in contrast to the COVID-19 cohort, where both Data Partner sites annotated hospital discharge summaries, which tend to contain the same types of information across institutions (as evident by looking at the top terms by institution by concept [Table 8]). VUMC notes have more than 5 times that number of annotations within their cancer cohort compared with BWH (7379; 1354). Both Data Partner sites chose notes that met the inclusion criteria, though the VUMC note selection was less random, specifically targeting those with the most information. For conditions, the overall percentage of the total annotations was roughly the same across Data Partner sites (~25%), though they were found in more VUMC notes compared with BWH (100% to 75%). There were far more mentions of tests and procedures within the VUMC notes, both in number of notes containing a mention and the percentage of the total. Smoking status was not well documented across either Data Partner site. All notes mentioned medications, as was expected given the cohort definition, and the total percentages were similar across Data Partner sites. For the gene/protein class, there were more notes in the VUMC cohort that contained a mention compared to BWH (96.4% compared to 25%), though BWH had more annotations overall and they represented a higher percentage of the total (18.5% BWH; 1.5% VUMC). There were a limited number of overall mentions of stem cell transplants or patients being refractory (non-responsive) to a treatment, but they were found more often in VUMC notes compared with BWH. The last two concepts were of particular interest because of their inclusion in the label indication. It is likely that this information is somewhere within the patient's chart, but the particular note (or visit within their overall care journey) may be different by institution. The variation across classes by Data Partner site is an informative finding, illustrating the differences that can occur when selecting note types, particularly as they vary by specialty.

When looking at the presence of medication metadata within a medication annotation within the VUMC cancer cohort (Table 11), we see that most annotations do not contain any other metadata attributes (~44%). Timing is the most mentioned attribute, but only appearing in ~22% of annotations. One important note about this result is that the annotators skipped any templated text that might have been pulled in from structured portions of the EHR (e.g., medication list). These results were captured in the "Template Start" concept and there is at least one in almost every VUMC cancer note, so it is possible that pulling in the text from the medication list would increase those numbers. However, since those data are already available in a structured field, it would be more straightforward to just use the data in that format than to parse it out via NLP. As a result, one of the main takeaways from this particular sub-analysis is that medication mentions in the free-text portion of the note do not include much additional metadata.

Finally, many of the concept classes used in the annotation exercise are broad (e.g., conditions, treatments and procedures), but these groupings were chosen because existing NLP pipelines are generally robust at extracting data of these types. For more novel concepts where there has been less NLP development (e.g., oxygen support, refractory for stem cell transplant), having a more defined concept definition can make the annotation task more straightforward, while also generating a training corpus for the development of a pipeline that can automatically extract these terms.

Appendix A – SQL Code to select COVID-19 patients

```
/* identify all patients with an encounter diagnosis of COVID-19 (U07.1) who were admitted
between 4/1/2020 and 12/31/2021 and who were >= 18 at the time of admission.  Pseudo-code
written based on structure of the PCORnet CDM.  Ignoring CODE_TYPE variables for diagnosis
and procedures (e.g., ICD10, CPT, HCPCS) */
overall_cohort as (
select
    DIAGNOSIS.PATID,
    DIAGNOSIS.ENCOUNTERID,
    DIAGNOSIS.DX,
    DIAGNOSIS.ADMIT_DATE,
    DEMOGRAPHIC.BIRTH_DATE
    from
        DIAGNOSIS
        join DEMOGRAPHIC
            on DEMOGRAPHIC.PATID = DIAGNOSIS.PATID
        where DX = ‘U07.1’
        and ADMIT_DATE between ‘2020-04-01’ and ‘2021-12-31’
        and datediff (year, DIAGNOSIS.ADMIT_DATE, DEMOGRAPHIC.BIRTH_DATE)
= >= 18
),

/* from the initial cohort of patients, identify those with a procedure code for supplemental
oxygen assumes that the ENCOUNTERID for the procedure record is the same the
ENCOUNTERID associated with the diagnosis record.   Ignoring any supplemental oxygen codes
that might be coded as diagnoses (ICD-10 Z codes) */

oxygen as (
select
    PROCEDURES.PATID
    from
        PROCEDURES
        where PROCEDURES.ENCOUNTERID = overall_cohort. ENCOUNTERID
        and PROCEDURES.PX in
(94660,5A09357,5A09457,5A09557,31500,43753,94002,94003,94657,94656,94004,09HN8BZ,
0BH18EZ,0DH57BZ,0WHQ73Z,0WHQ7YZ,5A09457,09HN7BZ,5A09557,0CHY7BZ,5A09357,5A
1945Z,5A1955Z,0BH17EZ,0BH13EZ,0CHY8BZ,0DH58BZ,5A1935Z,33989,33988,36822,33957,3
3958,33959,33962,33963,33964,33965,33966,33969,33984,33985,33986,33951,33952,33953,3
3954,33955,33956,33946,33947,33948,33949,33987,5A15223,5A1522F,5A1522G,5A1522H,5A15
A2F,5A15A2G,5A15A2H)
),
/* non-oxygen patients are those not in the oxygen cohort */

no_oxygen as (
select
    overall_cohort. PATID
    from
    overall_cohort
    where overall_cohort. PATID (not in oxygen. PATID).
);
```

Appendix B – Annotation Guidelines for the COVID-19 (Oxygen Use) cohort

Oxygen Administration Annotation Guidelines

Guideline Terms: Class, Attribute, Value, Instance, Relation, Link

***Class*:** Select words or phrase representing the specified concept (numbered below)

- ⊗ **Please note:** Select most comprehensive representation of the concept in question (e.g., "soft systolic murmur", as opposed to "systolic murmur") without including any extraneous words irrelevant to the concept (keep it concise)

Attribute: Select a value from the attribute list of the annotated class.

- ⊗ **Please note:** Do not select text to indicate these values of the associated class, use the drop-down menu

Relation: Link an instance of the annotated class to at least 1 annotated instance of a designated instance of another class

Instance: Refers to a particular span of text annotated in some way (e.g., as a class, or assigned an attribute)

Link: Select the instance you want to relate to another, & right click with cursor within the span of the instance to link to.

Prior to linking, be sure to click on the "**+**" icon within the Relationships pane to change it to a pencil icon

The Attribute *Assertion* applies to **PRIMARY CLASSES**. The assertion values are the following:

- positive Ex: "I believe that x" "Consistent with" "Most likely" "Compatible with" "Probable"
- negative Ex: "no evidence of x" "x has not been administered" "condition x is absent"
- uncertain Ex: "Cannot rule out x" "Suggestive of x" "May be x ""Presumably x"

Please Note: **THIS VALUE DOES NOT PERTAIN TO FUTURE STATES, PLANNED CARE OR PREDICTED CONDITIONS.**

USE TIMEPERSPECTIVE (**predicted**) FOR SUCH CASES

- hypothetical Ex: "if x occurs, contact so & so" "Monitor Pt for x" "In all patients with condition x"

The Attribute *TimePerspective* applies to **PRIMARY CLASSES**. The TimePerspective values are the following:

- current Ex: "Pt is taking x" "He complains of x" (include preexisting conditions, e.g., PMH of COPD)
- history Ex: "history of pneumonia" "historically X" "Past use of x" (and not ongoing or present)
- predicted Ex: "Pt is scheduled for transfusion tomorrow AM" "Begin taking x at midnight"

Interaction between these Attributes can occur. Ex: "Pt has no history of ***X***."
X would be assigned Assertion [negative] and *TimePerspective* [history]

Class Assignment Guidelines

1) **Oxygen Support** PRIMARY CLASS: Mark references to supplemental oxygen device or oxygen delivery method, or refence to oxygen administration
   - ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
   - ✓ *attribute* (unique to class) ***Change Status:*** [singular (default) if no other O2 device mentioned, change]
   - ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Nasal Cannula	Oxygenation	ECMO	on oxygen	Non-rebreather mask	Oxygen conserving device
High-flow O2		BIPAP	oxygen delivery	Invasive mechanical ventilation	

⊗ **Please note** change status value ‘change’ should only be used if there are more than one instance of the secondary class Oxygen Support Volume linked to the instance of Oxygen Support

1) **Oxygen Support Volume** Secondary Class: Mark mentions of oxygen volume

Relation*: 02* ***Volume Link*** → Link the annotated instance of “Oxygen Support Volume” to the annotated instance of “Oxygen Support” with which it is associated in the document

Example Expressions

3L		2 Liters	

⊗ **Please note**: If oxygen device or oxygen delivery method is not mentioned, annotate any word that let you know the topic was related to oxygen as the oxygen support.

2) **Oxygen Support Duration** Secondary Class: Mark mentions of a time interval that a supplemental oxygen device or oxygen delivery method is stated to be in use

Relation*: 02* ***Interval Link*** → Link the annotated instance of “Oxygen Support Duration” to the annotated instance of “Oxygen Support” with which it is associated in the document

Example Expressions

		For 2 days	6 hours

3) **Condition** PRIMARY CLASS: Mark references to diagnoses, signs or symptoms. Mark findings where clinical evidence of abnormalities are mentioned.
   - ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
   - ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Nausea		PNA	LE edema	Orthopnea
Chest pain	Soft systolic murmur		Hypertension	

⊗ **Please note**: Do not mark any evidence from procedures (e.g., “Splenectomy”) or test results (e.g., “CXR shows airspace opacities”) Test results, or procedures should only be marked as indicators of conditions if they are mentioned in summary, but not from imported lab or procedure tables

4) **Medication** PRIMARY CLASS: Mark references to medications and drugs
   - ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
   - ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Hydrocortisone Active **TimePerspective** (current)		**Etomidate** discontinued **TimePerspective** (history)	**Steroids**
Pressures discontinued **TimePerspective** (history)		**Vancomycin** prescription ordered **TimePerspective** (predicted)	**Nebulizers**

- ⊗ **Please note**: Do not mark any medications that are part of a medication list. Please see 14 and 15 below for instructions on annotating medication lists.
- ⊗ **Please note** indicators of liquid nutrition, saline dispensed via intravenous drip should be considered a medication

5) **Medication Timing** Secondary Class: Mark references to time or date-time or event referencing a *singular point* in time (e.g., transfusion) for medication administration

Relation*: **Rx Time-Stamp Link*** → Link the annotated instance of "Medication Timing" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

X drug at **6:00 AM**		Start lisinopril **tomorrow morning**	
		Give Benadryl **1 hour prior to transfusion**	

- ⊗ **Please note**: Medication Duration and Medication Frequency are their own classes.

Instances referring to medication discontinuation can be captured by assigning the drug in question the TimePerspective value [*history*]. Examples such as "taper steroids", similarly, can be captured as TimePerspective value [*current*].

6) Medication Duration Secondary Class: Mark references to duration phrases that are associated with a medication administration

Relation*: **Rx Duration Link*** → Link the annotated instance of "Medication Duration" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

For 1 week	Vancomycin x **5 days**	over **30 days**	Patient on **day 5/5** of Zosyn

7) Medication Frequency Secondary Class: Mark references to a series of times or the frequency of a medication administration.

Relation*: **Rx FREQUENCY LINK*** → Link the annotated instance of "Medication Frequency" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Every 4 hours		**b.i. d**	**Twice daily**	**Q3h**
PRN	**Q6h**	Start Seroquel **hs**	**Every other day**	Insulin **before meals**

- ⊗ **Please note**: References to regularly occurring events in relation to medication administration such as meals, or nights (hs), successive number of days, etc. are indicators of frequency and should be marked

8) Medication Indication Secondary Class: Mark references to phrases indicating the reason a medication is being taken

Relation: ***Rx Reason Link*** → Link the annotated instance of "Medication Indication" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

wheezes			For pain	If swelling occurs
			Persistent cough	

⊗ **Please note**: Do not double-annotate medications indications as [Condition]

9) Medication Dose Secondary Class: Mark references to phrases indicating the medication dosage amount or measurement

Relation: ***Rx Dosage Link*** → Link the annotated instance of "Medication Dose" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

6 units		**1L**	250 mg	10 mL

10) Medication Form Secondary Class: Mark references to medication form associated with the administration of a medication

Relation: ***Rx Form Link*** → Link the annotated instance of "Medication Form" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Inhaler			Otic Suspension (Otic Drops; Ear Drops)	

11) Medication Route Secondary Class: Mark references to the route of medication

Relation: ***Rx Route Link*** → Link the annotated instance of "Medication Route" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Sublingual		Inhalation	Transendocardial	
PO	**IV**	**IM**		

12) **Adverse Reaction** Secondary Class (with assertion value): Mark references to adverse reactions to medications, including allergies, sensitivities and contraindications

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Relation: ***Rx Adversity Link*** → Link the annotated instance of "Adverse Reaction" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Hives		Anaphylaxis	Allergies	Rash

⊗ **Please note**: Generally, the medication mentioned (the text with the primary classification **Medication**) as causing the adverse reaction should be assigned the assertion value [negative]. If it is indicated that the med is being taken in spite of the adverse reaction, the medication should be left with the default assertion value [positive]

13) **Template Start:** Mark the first word of a table or other tabular artifact representing data imported from medications, laboratory results, or procedures tables from the EHR

14) Template End: Mark the last word of a table or other tabular artifact representing data imported from medications, laboratory results, or procedures tables from the EHR

15) Smoking Status PRIMARY CLASS: Mark indications of tobacco exposure

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Never **smoker** **TimePerspective** (current) **Assertion** (negative)	**Tobacco user** - E-Cigarettes or another Vaporizer	Work **secondhand tobacco exposure**
Former **smoker** **TimePerspective** (history) **Assertion** (positive)	Current every day **smoker**	**Tobacco user** - Pipe tobacco

16) Discharge Disposition Mark setting of discharge disposition

Example Expressions

assisted living facility	home/self-care	skilled nursing	hospice
Intermediate Care Facility	Nursing Home	Against Medical Advice	**Transfer to OSH**

⊗ **Please note**: Discharge status due to death should be annotated under the Class "**Death**"

17) Death Mark mentions of mortality

Example Expressions

		Expired (Deceased)	

18) Date of Death Secondary Class: Mark references to the date or date-time of death

Relation: ***Mortality Date Link*** → Link the annotated instance of "Date of Death" to the annotated instance of "Death" with which it is associated in the document

Example Expressions

Expired at **6:08pm**			

19) Cause of Death Secondary Class: Mark references to the cause of death

Relation: ***Mortality Causal Link*** → Link the annotated instance of "Cause of Death" to the annotated instance of "Death" with which it is associated in the document

Example Expressions

				Asystole
	PEA arrest			

Appendix C – SQL code to select darzalex patients

```
/*
Select patients with a prescription for darzalex (daratumumab) with an ORDER_DATE
between 1/1/2016 and 11/30/2021
and 0 occurrences of a prescription for darzalex (daratumumab) in the prior 1 to 1095
days

From this cohort, select 30 patients at random
*/

overall_cohort as (
select
        PRESCRIBING.PATID,
        PRESCRIBING.ENCOUNTERID,
        PRESCRIBING.RX_ORDER_DATE,
        PRESCRIBING.RXNORM_CUI,
        PRESCRIBING.RAW_RX_MED_NAME,

        from
        PRESCRIBING

        where (RXNORM_CUI in (1721952, 1721947, 1721956, 1726440, 1721953,
        1721954, 1721955, 1721951,1726439, 1721948, 1721950, 1721949)
                or RAW_RX_MED_NAME like %darzalex%
                or RAW_RX_MED_NAME like %daratumumab%)
                and RX_ORDER_DATE between ‘2016-01-01’ and ‘2021-11-30’
),

select overall_cohort. PATID
from overall_cohort

LEFT JOIN PRESCRIBING on overall_cohort. PATID = PRESCRIBING.PATID

where (PRESCRIBING.RXNORM_CUI in (1721952, 1721947, 1721956, 1726440,
1721953, 1721954, 1721955, 1721951,1726439, 1721948, 1721950, 1721949)
        or PRESCRIBING.RAW_RX_MED_NAME like %darzalex%
        or PRESCRIBING.RAW_RX_MED_NAME like %daratumumab%)
        AND PRESCRIBING.RX_ORDER_DATE >= DATEADD (day, -1095,
        overall_cohort. RX_ORDER_DATE)
        AND PRESCRIBING.RX_ORDER_DATE <= DATEADD (day, -1, overall_cohort.
        RX_ORDER_DATE);
```

Appendix D – Annotation Guidelines for the Cancer (Darzalex) cohort

Darzalex Annotation Guidelines

Guideline Terms: Class, Attribute, Value, Instance, Relation, Link

Class: Select words or phrase representing the specified concept (numbered below)

- ⊗ **Please note:** Select most comprehensive representation of the concept in question (e.g., “soft systolic murmur”, as opposed to “systolic murmur”) without including any extraneous words irrelevant to the concept (keep it concise)

Attribute: Select a value from the attribute list of the annotated class.

- ⊗ **Please note**: Do not select text to indicate these values of the associated class, use the drop-down menu

Relation: Link an instance of the annotated class to at least 1 annotated instance of a designated instance of another class

Instance: Refers to a particular span of text annotated in some way (e.g., as a class, or assigned an attribute)

Link: Select the instance you want to relate to another, & right click with cursor within the span of the instance to link to.

Prior to linking, be sure to click on the “**+**” icon within the Relationships pane to change it to a pencil icon

Class Assignment Guidelines

The Attribute *Assertion* applies to **PRIMARY CLASSES**. The assertion values are the following:

- positive Ex: “I believe that x” “Consistent with” “Most likely” “Compatible with” “Probable”
- negative Ex: “no evidence of x” “x has not been administered” “condition x is absent”
- uncertain Ex: “Cannot rule out x” “Suggestive of x” “May be x ““Presumably x”

Please Note: THIS VALUE DOES NOT PERTAIN TO FUTURE STATES, PLANNED CARE OR PREDICTED CONDITIONS.

USE TIMEPERSPECTIVE (**predicted**) FOR SUCH CASES

- hypothetical Ex: “if x occurs, contact so & so” “Monitor Pt for x” “In all patients with condition x”

The Attribute *TimePerspective* applies to **PRIMARY CLASSES**. The TimePerspective values are the following:

- current Ex: “Pt is taking x” “He complains of x” (include preexisting conditions, e.g., PMH of COPD)
- history Ex: “history of pneumonia” “historically X” “Past use of x” (and not ongoing or present)
- predicted Ex: “Pt is scheduled for transfusion tomorrow AM” “Begin taking x at midnight”

Use the tense of the sentence, wherever possible to decide which time frame a given instance falls within; in short, no-tense instances without a tense indicator, assign the default value “current”

Interaction between these Attributes can occur. Ex: “Pt has no history of ***X***.”
X would be assigned Assertion [negative] and *TimePerspective* [history]

20) **Condition** PRIMARY CLASS: Mark references to diagnoses, signs or symptoms. Mark findings where clinical evidence of abnormalities are mentioned.

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Nausea		PNA	LE edema	Orthopnea
Chest pain	Soft systolic murmur		Hypertension	Elevated LFT's
Chest CT [Test/Procedures/Treatment]	with small focal area of consolidation in the L lung base			

⊗ **Please note**: Test results or findings (e.g., "CXR shows airspace opacities") should be annotated as Condition. These should only be marked if they are mentioned in summary, not those imported lab or procedure tables.

↕ Relation: ***Findings Link*** → Link instances of test results or findings to the annotated instance of the class "Tests / Procedures / Treatment" with which it is associated in the document.

֍ Conditions that are not stated as findings or results do not need to be linked to any other class.

⊗ **Please note**: Do not label cancer indications as Condition. Such indications should be assigned to the Cancer/Tumor class below.

21) **Tests / Procedures / Treatments** PRIMARY CLASS: Mark references to tests and/or procedures used for diagnostic purposes, as well as any procedures used for treatment.

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Echocardiogram	Chest x-ray	Gallbladder surgery	Splenectomy
Chest CT			

⊗ **Please note**: Do not mark any medications as a treatment. These should be included in the Medication class. Do not mark stem cell transplant as treatment. This should be included in the stem cell class.

22) **Cancer/Tumor** PRIMARY CLASS: Mark references to cancer or tumor.

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Cancer	tumor	(malignant) neoplasm	
Multiple myeloma		Kahler's disease	

⊗ **Please note**: Do not mark instances of 'mass' unless note is explicit that it is a cancer or a tumor

23) **Cancer Stage** Secondary Class: Mark references to cancer or tumor stage.

Relation: ***Cancer Stage Link*** → Link the annotated instance of "Stage" to the annotated instance of "Cancer" with which it is associated in the document

Example Expressions

Stage I	Stage II	III	IV

24) **Gene / Protein** PRIMARY CLASS: Mark references to genetic markers or proteins (e.g., gene, variant, expression level).

- ✓ *attribute:* ***Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute:* ***TimePerspective*** [current (default), history, predicted]

Example Expressions

BIRC5		BRAC2 positive	gain(1q)
LTBP1	MCM2	t (4;14), t (14;16), del(17p)	TOP2A

25) **Stem Cell Transplant** PRIMARY CLASS: Mark references to stem cell transplant.

- ✓ *attribute:* ***Assertion*** [positive (or eligible), negative (or ineligible), uncertain, hypothetical]
- ✓ *attribute:* ***TimePerspective*** [current (default), history, predicted]

Example Expressions

⊗ **Please note**: Mark indications that the patient is eligible for a stem cell transplant as a positive mention. Mark ineligibility as a negative mention.

26) **Refractory** PRIMARY CLASS: Mark that a cancer is refractory to a medication (fails to improve or stops responding).

- ✓ *attribute:* ***Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute:* ***TimePerspective*** [current (default), history, predicted]

Example Expressions

27) **Medication** PRIMARY CLASS: Mark references to medications and drugs

- ✓ *attribute:* ***Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute:* ***TimePerspective*** [current (default), history, predicted]

Example Expressions

Hydrocortisone Active **TimePerspective** (current)		**Etomidate** discontinued **TimePerspective** (history)	**Steroids**
Pressures discontinued **TimePerspective** (history)		**Vancomycin** prescription ordered **TimePerspective** (predicted)	**Nebulizers**

⊗ **Please note**: Do not mark any medications that are part of a medication list. Please see 14 and 15 below for instructions on annotating medication lists.

⊗ **Please note** indicators of liquid nutrition, saline dispensed via intravenous drip should be considered a medication

28) **Medication Timing** Secondary Class: Mark references to time or date-time or event referencing a *singular point* in time (e.g., transfusion) for medication administration

Relation: ***Rx Time-Stamp Link*** → Link the annotated instance of "Medication Timing" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

X drug at **6:00 AM**	C1 D8	Start lisinopril **tomorrow morning**	
cycle 1 day 1		Give Benadryl **1 hour prior to transfusion**	

- ⊗ **Please note**: Medication Duration and Medication Frequency are their own classes.
- ⊗ **Please note**: Instances referring to medication discontinuation can be captured by assigning the drug in question the TimePerspective value [*history*]. Examples such as "taper steroids", similarly, can be captured as TimePerspective value [*current*].
- ⊗ **Please note**: Instances referring to periodicity of a drug regimen, such as "**cycle 1 day 1"** should be assigned to this class. Use the sentence tense for deciding what TimePerspective value should be assigned. If the phrase occurs without any tense, use the default "current" as the TimePerspective value.

29) Medication Duration Secondary Class: Mark references to duration phrases that are associated with a medication administration

Relation: ***Rx Duration Link*** → Link the annotated instance of "Medication Duration" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

For 1 week	Vancomycin x **5 days**	over **30 days**	Patient on **day 5/5** of Zosyn

30) Medication Frequency Secondary Class: Mark references to a series of times or the frequency of a medication administration.

Relation: ***Rx FREQUENCY LINK*** → Link the annotated instance of "Medication Frequency" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Every 4 hours		b.i. d	Twice daily	Q3h
PRN	Q6h	Start Seroquel **hs**	Every other day	Insulin before meals

- ⊗ **Please note**: References to regularly occurring events in relation to medication administration such as meals, or nights (hs), successive number of days, etc. are indicators of frequency and should be marked

31) Medication Indication Secondary Class: Mark references to phrases indicating the reason a medication is being taken

Relation: ***Rx Reason Link*** → Link the annotated instance of "Medication Indication" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

wheezes			For pain	If swelling occurs
			Persistent cough	

- ⊗ **Please note**: Do not double-annotate medications indications as [Condition]

32) Medication Dose Secondary Class: Mark references to phrases indicating the medication dosage amount or measurement

Relation*: **Rx Dosage Link*** → Link the annotated instance of "Medication Dose" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

6 units		1L	250 mg	10 mL

33) **Medication Form** Secondary Class: Mark references to medication form associated with the administration of a medication

Relation*: **Rx Form Link*** → Link the annotated instance of "Medication Form" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Inhaler			Otic Suspension (Otic Drops; Ear Drops)	

34) **Medication Route** Secondary Class: Mark references to the route of medication

Relation*: **Rx Route Link*** → Link the annotated instance of "Medication Route" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Sublingual		Inhalation	Transendocardial	
PO	IV	IM		

35) **Adverse Reaction** Secondary Class (with assertion value): Mark references to adverse reactions to medications, including allergies, sensitivities and contraindications

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Relation*: **Rx Adversity Link*** → Link the annotated instance of "Adverse Reaction" to the annotated instance of "Medication" with which it is associated in the document

Example Expressions

Hives		Anaphylaxis	Allergies	Rash

⊗ **Please note**: Generally, the medication mentioned (the text with the primary classification **Medication**) as causing the adverse reaction should be assigned the assertion value [negative]. If it is indicated that the med is being taken in spite of the adverse reaction, the medication should be left with the default assertion value [positive]

36) **Template Start:** Mark the first word of a table or other tabular artifact representing data imported from medications, laboratory results, or procedures tables from the EHR

37) **Template End:** Mark the last word of a table or other tabular artifact representing data imported from medications, laboratory results, or procedures tables from the EHR

38) **Smoking Status** PRIMARY CLASS: Mark indications of tobacco exposure

- ✓ *attribute: **Assertion*** [positive, negative, uncertain, hypothetical]
- ✓ *attribute: **TimePerspective*** [current (default), history, predicted]

Example Expressions

Never **smoker** **TimePerspective** (current) **Assertion** (negative)	**Tobacco user** - E-Cigarettes or other Vaporizer	Work **secondhand tobacco exposure**
Former **smoker** **TimePerspective** (history) **Assertion** (positive)	Current every day **smoker**	**Tobacco user** - Pipe tobacco

39) Death Mark mentions of mortality

Example Expressions

		Expired (Deceased)	

⊗ **Please note**: We do not expect to see many mentions of mortality as part of this cohort.

40) Date of Death Secondary Class: Mark references to the date or date-time of death

Relation: ***MORTALITY DATE LINK*** → Link the annotated instance of “Date of Death” to the annotated instance of “Death” with which it is associated in the document

Example Expressions

Expired at **6:08pm**			

41) Cause of Death Secondary Class: Mark references to the cause of death

Relation: ***MORTALITY CAUSAL LINK*** → Link the annotated instance of “Cause of Death” to the annotated instance of “Death” with which it is associated in the document

Example Expressions

				Asystole
	PEA arrest			

Appendix E – Billing codes for supplemental oxygen

Code	Code_Type	Description
Z99.81	ICD10-CM	Dependence on supplemental oxygen
94660	CPT/HCPCS	Continuous positive airway pressure ventilation (CPAP), initiation and management
5A09357	ICD10-PCS	Assistance with Respiratory Ventilation, less than 24 Consecutive Hours, Continuous Positive Airway Pressure
5A09457	ICD10-PCS	Assistance with Respiratory Ventilation, 24-96 Consecutive Hours, Continuous Positive Airway Pressure
5A09557	ICD10-PCS	Assistance with Respiratory Ventilation, Greater than 96 Consecutive Hours, Continuous Positive Airway Pressure
Z99.11	ICD10-CM	Dependence on respirator [ventilator] status
Z99.1	ICD10-CM	Dependence on respirator
31500	CPT/HCPCS	emergency endotracheal intubation procedure
43753	CPT/HCPCS	gastric intubation with aspiration and lavage
94002	CPT/HCPCS	Ventilation assist and management, initiation of pressure or volume preset ventilators for assisted or controlled breathing
94003	CPT/HCPCS	Ventilation assist and management, initiation of pressure or volume preset ventilators for assisted or controlled breathing
94657	CPT/HCPCS	ventilator management
94656	CPT/HCPCS	ventilator management
94004	CPT/HCPCS	Ventilation assist and management, initiation of pressure or volume preset ventilators for assisted or controlled breathing
09HN8BZ	ICD10-PCS	Insertion of Airway into Nasopharynx, Via Natural or Artificial Opening Endoscopic
0BH18EZ	ICD10-PCS	Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening Endoscopic
0DH57BZ	ICD10-PCS	Insertion of Airway into Esophagus, Via Natural or Artificial Opening
0WHQ73Z	ICD10-PCS	Insertion of Infusion Device into Respiratory Tract, Via Natural or Artificial Opening
0WHQ7YZ	ICD10-PCS	Insertion of Other Device into Respiratory Tract, Via Natural or Artificial Opening
5A09457	ICD10-PCS	Assistance with Respiratory Ventilation, 24-96 Consecutive Hours, Continuous Positive Airway Pressure
09HN7BZ	ICD10-PCS	Insertion of Airway into Nasopharynx, Via Natural or Artificial Opening
5A09557	ICD10-PCS	Assistance with Respiratory Ventilation, Greater than 96 Consecutive Hours, Continuous Positive Airway Pressure
0CHY7BZ	ICD10-PCS	Insertion of Airway into Mouth and Throat, Via Natural or Artificial Opening
5A09357	ICD10-PCS	Assistance with Respiratory Ventilation, less than 24 Consecutive Hours, Continuous Positive Airway Pressure
5A1945Z	ICD10-PCS	Respiratory Ventilation, 24-96 Consecutive Hours
5A1955Z	ICD10-PCS	Respiratory Ventilation, Greater than 96 Consecutive Hours
0BH17EZ	ICD10-PCS	Insertion of Endotracheal Airway into Trachea, Via Natural or Artificial Opening
0BH13EZ	ICD10-PCS	Insertion of Endotracheal Airway into Trachea, Percutaneous Approach
0CHY8BZ	ICD10-PCS	Insertion of Airway into Mouth and Throat, Via Natural or Artificial Opening Endoscopic
0DH58BZ	ICD10-PCS	Insertion of Airway into Esophagus, Via Natural or Artificial Opening Endoscopic
5A1935Z	ICD10-PCS	Insertion of Airway into Esophagus, Via Natural or Artificial Opening Endoscopic

33989	CPT/HCPCS	Removal of left heart vent by thoracic incision (eg, sternotomy, thoracotomy) for ECMO/ECLS
33988	CPT/HCPCS	Insertion of left heart vent by thoracic incision (eg, sternotomy, thoracotomy) for ECMO/ECLS
36822	CPT/HCPCS	Insertion of cannula(s) for prolonged extracorporeal circulation for cardiopulmonary insufficiency (ECMO) (separate procedure)
33957	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; reposition peripheral (arterial and/or venous) cannula(e), percutaneous, birth through 5 years of age (includes fluoroscopic guidance, when performed)
33958	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; reposition peripheral (arterial and/or venous) cannula(e), percutaneous, 6 years and older (includes fluoroscopic guidance, when performed)
33959	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; reposition peripheral (arterial and/or venous) cannula(e), open, birth through 5 years of age (includes fluoroscopic guidance, when performed)
33962	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; reposition peripheral (arterial and/or venous) cannula(e), open, 6 years and older (includes fluoroscopic guidance, when performed)
33963	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; reposition of central cannula(e) by sternotomy or thoracotomy, birth through 5 years of age (includes fluoroscopic guidance, when performed)
33964	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; reposition central cannula(e) by sternotomy or thoracotomy, 6 years and older (includes fluoroscopic guidance, when performed)
33965	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; removal of peripheral (arterial and/or venous) cannula(e), percutaneous, birth through 5 years of age
33966	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; removal of peripheral (arterial and/or venous) cannula(e), percutaneous, 6 years and older
33969	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; removal of peripheral (arterial and/or venous) cannula(e), open, birth through 5 years of age
33984	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; removal of peripheral (arterial and/or venous) cannula(e), open, 6 years and older
33985	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; removal of central cannula(e) by sternotomy or thoracotomy, birth through 5 years of age
33986	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; removal of central cannula(e) by sternotomy or thoracotomy, 6 years and older
33951	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; insertion of peripheral (arterial and/or venous) cannula(e), percutaneous, birth through 5 years of age (includes fluoroscopic guidance, when performed)
33952	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; insertion of peripheral (arterial and/or venous) cannula(e), percutaneous, 6 years and older (includes fluoroscopic guidance, when performed)
33953	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; insertion of peripheral (arterial and/or venous) cannula(e), open, birth through 5 years of age

33954	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; insertion of peripheral (arterial and/or venous) cannula(e), open, 6 years and older
33955	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; insertion of central cannula(e) by sternotomy or thoracotomy, birth through 5 years of age
33956	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; insertion of central cannula(e) by sternotomy or thoracotomy, 6 years and older
33946	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; initiation, veno-venous
33947	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; initiation, veno-arterial
33948	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; daily management, each day, veno-venous
33949	CPT/HCPCS	Extracorporeal membrane oxygenation (ECMO)/extracorporeal life support (ECLS) provided by physician; daily management, each day, veno-arterial
33987	CPT/HCPCS	Arterial exposure with creation of graft conduit (eg, chimney graft) to facilitate arterial perfusion for ECMO/ECLS (List separately in addition to code for primary procedure)
Z92.81	ICD10-CM	Personal history of extracorporeal membrane oxygenation (ECMO)
5A15223	ICD10-PCS	Extracorporeal Membrane Oxygenation, Continuous
5A1522F	ICD10-PCS	Extracorporeal Oxygenation, Membrane, Central
5A1522G	ICD10-PCS	Extracorporeal Oxygenation, Membrane, Peripheral Veno-arterial
5A1522H	ICD10-PCS	Extracorporeal Oxygenation, Membrane, Peripheral Veno-venous
5A15A2F	ICD10-PCS	Extracorporeal Oxygenation, Membrane, Central, Intraoperative
5A15A2G	ICD10-PCS	Extracorporeal Oxygenation, Membrane, Peripheral Veno-arterial, Intraoperative
5A15A2H	ICD10-PCS	Extracorporeal Oxygenation, Membrane, Peripheral Veno-venous, Intraoperative

Appendix F – XML schema for eHOST (COVID-19 / Oxygen Use annotation guidelines)

```
<?xml version="1.0" encoding="UTF-8"?>
<eHOST_Project_Configure Version="1.0">
  <Handling_Text_Database>false</Handling_Text_Database>
  <OracleFunction_Enabled>false</OracleFunction_Enabled>
  <AttributeEditor_PopUp_Enabled>false</AttributeEditor_PopUp_Enabled>
  <OracleFunction>true</OracleFunction>
  <AnnotationBuilder_Using_ExactSpan>false</AnnotationBuilder_Using_ExactSpan>
  <OracleFunction_Using_WholeWord>true</OracleFunction_Using_WholeWord>
  <GraphicAnnotationPath_Enabled>true</GraphicAnnotationPath_Enabled>
  <Diff_Indicator_Enabled>true</Diff_Indicator_Enabled>
  <Diff_Indicator_Check_CrossSpan>true</Diff_Indicator_Check_CrossSpan>
  <Diff_Indicator_Check_Overlaps>false</Diff_Indicator_Check_Overlaps>
  <StopWords_Enabled>false</StopWords_Enabled>
  <Output_VerifySuggestions>false</Output_VerifySuggestions>
  <Pre_Defined_Dictionary_DifferentWeight>false</Pre_Defined_Dictionary_DifferentWeight>
  <PreAnnotated_Dictionaries Owner="NLP_Assistant" />
  <attributeDefs>
    <attributeDef>
      <Name>Assertion</Name>
      <is_Linked_to_UMLS_CUICode_and_CUILabel>false</is_Linked_to_UMLS_CUICode_and_CUILabel>
      <is_Linked_to_UMLS_CUICode>false</is_Linked_to_UMLS_CUICode>
      <is_Linked_to_UMLS_CUILabel>false</is_Linked_to_UMLS_CUILabel>
      <defaultValue>positive</defaultValue>
      <attributeDefOptionDef>hypothetical</attributeDefOptionDef>
      <attributeDefOptionDef>uncertain</attributeDefOptionDef>
      <attributeDefOptionDef>negative</attributeDefOptionDef>
      <attributeDefOptionDef>positive</attributeDefOptionDef>
    </attributeDef>
    <attributeDef>
      <Name>TimePerspective</Name>
      <is_Linked_to_UMLS_CUICode_and_CUILabel>false</is_Linked_to_UMLS_CUICode_and_CUILabel>
      <is_Linked_to_UMLS_CUICode>false</is_Linked_to_UMLS_CUICode>
      <is_Linked_to_UMLS_CUILabel>false</is_Linked_to_UMLS_CUILabel>
      <defaultValue>current</defaultValue>
      <attributeDefOptionDef>predicted</attributeDefOptionDef>
      <attributeDefOptionDef>history</attributeDefOptionDef>
      <attributeDefOptionDef>current</attributeDefOptionDef>
    </attributeDef>
  </attributeDefs>
  <Relationship_Rules>
    <Relationship_Rule>
      <Name>02_INTERVAL_LINK</Name>
      <Definition>((01_Oxygen_Support)) ((03_Oxygen_Support_Duration)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>02_VOLUME_LINK</Name>
      <Definition>((01_Oxygen_Support)) ((02_Oxygen_Support_Volume)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>MORTALITY_DATE_LINK</Name>
      <Definition>((18_Death)) ((19_Date_of_Death)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RX_REASON_LINK</Name>
      <Definition>((05_Medication)) ((09_Medication_Indication)) </Definition>
      <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_ADVERSITY_ LINK</Name>
    <Definition>((05_Medication)) ((13_Adverse_Reaction)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_DOSAGE_LINK</Name>
    <Definition>((05_Medication)) ((10_Medication_Dose)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_DURATION_LINK</Name>
    <Definition>((05_Medication)) ((07_Medication_Duration)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_FORM_LINK</Name>
    <Definition>((05_Medication)) ((11_Medication_Form)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_FREQUENCY _LINK</Name>
    <Definition>((05_Medication)) ((08_Medication_Frequency)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_ROUTE_LINK</Name>
    <Definition>((05_Medication)) ((12_Medication_Route)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
  <Relationship_Rule>
    <Name>Rx_TIME-STAMP_LINK</Name>
    <Definition>((05_Medication)) ((06_Medication_Status_Time) |(06_Medication_Timing)) </Definition>
    <attributeDefs />
  </Relationship_Rule>
</Relationship_Rules>
<classDefs>
  <classDef>
    <Name>01_Oxygen_Support</Name>
    <RGB_R>0</RGB_R>
    <RGB_G>192</RGB_G>
    <RGB_B>128</RGB_B>
    <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
    <attributeDef>
      <Name>Change Status</Name>
      <is_Linked_to_UMLS_CUICode_and_CUILabel>false</is_Linked_to_UMLS_CUICode_and_CUILabel>
      <is_Linked_to_UMLS_CUICode>false</is_Linked_to_UMLS_CUICode>
      <is_Linked_to_UMLS_CUILabel>false</is_Linked_to_UMLS_CUILabel>
      <defaultValue>singular</defaultValue>
      <attributeDefOptionDef>change</attributeDefOptionDef>
      <attributeDefOptionDef>singular</attributeDefOptionDef>
    </attributeDef>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>02_Oxygen_Support_Volume</Name>
    <RGB_R>128</RGB_R>
    <RGB_G>192</RGB_G>
    <RGB_B>128</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
```

```
<classDef>
  <Name>03_Oxygen_Support_Duration</Name>
  <RGB_R>0</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>04_Condition</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>0</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>05_Medication</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>0</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>06_Medication_Timing</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>07_Medication_Duration</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>128</RGB_G>
  <RGB_B>255</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>08_Medication_Frequency</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>192</RGB_G>
  <RGB_B>255</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>09_Medication_Indication</Name>
  <RGB_R>128</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>10_Medication_Dose</Name>
  <RGB_R>128</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>0</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
```

```
</classDef>
<classDef>
  <Name>11_Medication_Form</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>128</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>12_Medication_Route</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>192</RGB_G>
  <RGB_B>255</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>13_Adverse_Reaction</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>14_Template_Start</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>15_Template_End</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>192</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>16_Smoking_Status</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>128</RGB_G>
  <RGB_B>0</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>17_Discharge_Disposition</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>255</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>18_Death</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>192</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
```

```
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>19_Date_of_Death</Name>
      <RGB_R>128</RGB_R>
      <RGB_G>128</RGB_G>
      <RGB_B>128</RGB_B>
      <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>20_Cause_of_Death</Name>
      <RGB_R>192</RGB_R>
      <RGB_G>192</RGB_G>
      <RGB_B>192</RGB_B>
      <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
  </classDefs>
</eHOST_Project_Configure>
```

Appendix G – XML schema for eHOST (Darzalex annotation guidelines)

```xml
<?xml version="1.0" encoding="UTF-8"?>
<eHOST_Project_Configure Version="1.0">
  <Handling_Text_Database>false</Handling_Text_Database>
  <OracleFunction_Enabled>false</OracleFunction_Enabled>
  <AttributeEditor_PopUp_Enabled>false</AttributeEditor_PopUp_Enabled>
  <OracleFunction>true</OracleFunction>
  <AnnotationBuilder_Using_ExactSpan>false</AnnotationBuilder_Using_ExactSpan>
  <OracleFunction_Using_WholeWord>true</OracleFunction_Using_WholeWord>
  <GraphicAnnotationPath_Enabled>true</GraphicAnnotationPath_Enabled>
  <Diff_Indicator_Enabled>true</Diff_Indicator_Enabled>
  <Diff_Indicator_Check_CrossSpan>true</Diff_Indicator_Check_CrossSpan>
  <Diff_Indicator_Check_Overlaps>false</Diff_Indicator_Check_Overlaps>
  <StopWords_Enabled>false</StopWords_Enabled>
  <Output_VerifySuggestions>false</Output_VerifySuggestions>
  <Pre_Defined_Dictionary_DifferentWeight>false</Pre_Defined_Dictionary_DifferentWeight>
  <PreAnnotated_Dictionaries Owner="NLP_Assistant" />
  <attributeDefs>
    <attributeDef>
      <Name>Assertion</Name>
      <is_Linked_to_UMLS_CUICode_and_CUILabel>false</is_Linked_to_UMLS_CUICode_and_CUILabel>
      <is_Linked_to_UMLS_CUICode>false</is_Linked_to_UMLS_CUICode>
      <is_Linked_to_UMLS_CUILabel>false</is_Linked_to_UMLS_CUILabel>
      <defaultValue>positive</defaultValue>
      <attributeDefOptionDef>hypothetical</attributeDefOptionDef>
      <attributeDefOptionDef>uncertain</attributeDefOptionDef>
      <attributeDefOptionDef>negative</attributeDefOptionDef>
      <attributeDefOptionDef>positive</attributeDefOptionDef>
    </attributeDef>
    <attributeDef>
      <Name>TimePerspective</Name>
      <is_Linked_to_UMLS_CUICode_and_CUILabel>false</is_Linked_to_UMLS_CUICode_and_CUILabel>
      <is_Linked_to_UMLS_CUICode>false</is_Linked_to_UMLS_CUICode>
      <is_Linked_to_UMLS_CUILabel>false</is_Linked_to_UMLS_CUILabel>
      <defaultValue>current</defaultValue>
      <attributeDefOptionDef>predicted</attributeDefOptionDef>
      <attributeDefOptionDef>history</attributeDefOptionDef>
      <attributeDefOptionDef>current</attributeDefOptionDef>
    </attributeDef>
  </attributeDefs>
  <Relationship_Rules>
    <Relationship_Rule>
      <Name>CancerStageLink</Name>
      <Definition>((04_Cancer_Stage)) ((03_Cancer_Tumor)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>FindingsLink</Name>
      <Definition>((01_Condition)) ((02_Test_Procedure_Treatment)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>MortalityCausalLink</Name>
      <Definition>((22_Cause_of_Death)) ((20_Death)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>MortaltyDateLink</Name>
      <Definition>((21_Date_of_Death)) ((20_Death)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxAdversityLink</Name>
      <Definition>((16_Adverse_Reaction)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxDosageLink</Name>
      <Definition>((13_Medication_Dose)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxDurationLink</Name>
      <Definition>((10_Medication_Duration)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxFormLink</Name>
      <Definition>((14_Medication_Form)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxFrequencyLink</Name>
      <Definition>((11_Medication_Frequency)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxReasonLink</Name>
      <Definition>((12_Medication_Indication)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxRouteLink</Name>
      <Definition>((15_Medication_Route)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
    <Relationship_Rule>
      <Name>RxTimeStampLink</Name>
      <Definition>((09_Medication_Timing)) ((08_Medication)) </Definition>
      <attributeDefs />
    </Relationship_Rule>
  </Relationship_Rules>
  <classDefs>
    <classDef>
      <Name>01_Condition</Name>
      <RGB_R>255</RGB_R>
      <RGB_G>255</RGB_G>
      <RGB_B>0</RGB_B>
      <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>02_Test_Procedure_Treatment</Name>
      <RGB_R>0</RGB_R>
      <RGB_G>192</RGB_G>
      <RGB_B>128</RGB_B>
      <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>03_Cancer_Tumor</Name>
      <RGB_R>0</RGB_R>
      <RGB_G>192</RGB_G>
      <RGB_B>192</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>04_Cancer_Stage</Name>
  <RGB_R>0</RGB_R>
  <RGB_G>192</RGB_G>
  <RGB_B>255</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>05_Gene_Protein</Name>
  <RGB_R>0</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>0</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>06_Stem_Cell_Transplant</Name>
  <RGB_R>0</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>07_Refractory</Name>
  <RGB_R>0</RGB_R>
  <RGB_G>255</RGB_G>
  <RGB_B>192</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>08_Medication</Name>
  <RGB_R>255</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>0</RGB_B>
  <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>09_Medication_Timing</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>0</RGB_G>
  <RGB_B>128</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>10_Medication_Duration</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>128</RGB_G>
  <RGB_B>255</RGB_B>
  <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
  <Source>eHOST</Source>
</classDef>
<classDef>
  <Name>11_Medication_Frequency</Name>
  <RGB_R>192</RGB_R>
  <RGB_G>192</RGB_G>
```

```
    <RGB_B>255</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>12_Medication_Indication</Name>
    <RGB_R>128</RGB_R>
    <RGB_G>0</RGB_G>
    <RGB_B>128</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>13_Medication_Dose</Name>
    <RGB_R>128</RGB_R>
    <RGB_G>0</RGB_G>
    <RGB_B>0</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>14_Medication_Form</Name>
    <RGB_R>192</RGB_R>
    <RGB_G>128</RGB_G>
    <RGB_B>128</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>15_Medication_Route</Name>
    <RGB_R>255</RGB_R>
    <RGB_G>192</RGB_G>
    <RGB_B>255</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>16_Adverse_Reaction</Name>
    <RGB_R>255</RGB_R>
    <RGB_G>0</RGB_G>
    <RGB_B>128</RGB_B>
    <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>17_Template_Start</Name>
    <RGB_R>255</RGB_R>
    <RGB_G>255</RGB_G>
    <RGB_B>128</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>18_Template_End</Name>
    <RGB_R>255</RGB_R>
    <RGB_G>255</RGB_G>
    <RGB_B>192</RGB_B>
    <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
    <Source>eHOST</Source>
  </classDef>
  <classDef>
    <Name>19_Smoking_Status</Name>
    <RGB_R>255</RGB_R>
```

```
      <RGB_G>128</RGB_G>
      <RGB_B>0</RGB_B>
      <InHerit_Public_Attributes>true</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>20_Death</Name>
      <RGB_R>192</RGB_R>
      <RGB_G>192</RGB_G>
      <RGB_B>128</RGB_B>
      <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>21_Date_of_Death</Name>
      <RGB_R>128</RGB_R>
      <RGB_G>128</RGB_G>
      <RGB_B>128</RGB_B>
      <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
    <classDef>
      <Name>22_Cause_of_Death</Name>
      <RGB_R>192</RGB_R>
      <RGB_G>192</RGB_G>
      <RGB_B>192</RGB_B>
      <InHerit_Public_Attributes>false</InHerit_Public_Attributes>
      <Source>eHOST</Source>
    </classDef>
  </classDefs>
</eHOST_Project_Configure>
```

Appendix H – BWH Code for Cancer Cohort

```
---1. Identify MedicationID of Darzalex(daratumumab)

SELECT MedicationID, MedicationDescription
into darzalexID
from MedicationReferenceTable
where MedicationDescription like '%Darzalex%' or MedicationDescription like '%daratumumab%'
;

-- 2. Identify all the patient with prescription of darzalex

select distinct PatientID, PatientEncounterID, cast (OrderDateTime as date) as 'OrderDateTime' --
, MedicationID, MedicationDescription
into darzalex_order_all
from MedicationOrders om
where exists (select MedicationID from darzalexID mid were om. MedicationID = mid.
MedicationID)
        and om. OrderDateTime between '2016-01-01' and '2021-11-30'
;

-- 3. Select the order without Darzalex order in the prior 1-1095 days & random sample 100
patients first

select top 100 *
into darzalex_order_all_clean
from
        (select *, datediff (day, prior_OrderDateTime, OrderDateTime) as 'day_interval'
        from
                (select *, lag (OrderDateTime, 1) over (partition by PatientID order by
OrderDateTime asc) as 'prior_OrderDateTime'
                from darzalex_order_all) t1
        ) t2
where day_interval is null or day_interval>1095
order by newid ()
;
```