Claims Data Processing

Description

This document describes methods to incorporate insurance claims data in SEER*DMS. Insurance claims provide an extremely high volume of data. Therefore, the claims data processing in SEER*DMS must be automated as much as possible to avoid an over-burden on registry staff.

Claims data will initially represent:

- A large amount of raw data that must be stored outside of the SEER*DMS `RECORD` table.
- Data that should be matched against itself to remove functional duplicates.
- Data that should be matched against existing patient sets and tumors to determine linkage. Links to a Patient Set and CTC can be stored with the claim.
- Data that should be used to supplement treatment information for existing patient sets.

Claims data ***may*** represent in the future:

- A way to identify new patients or tumors.

This design focuses on claims, however the principles will eventually be applied to other data streams (i.e. diagnostic indices, pathology reports). Those types of data were considered in the design, but were not a driving factor.

Data Structure

The claims data will have such a large volume of records (including possible duplicates) that it will not be stored in the `RECORD` table. Doing so could potentially slow down the processing of abstracts, path reports, and other types of data that are stored in the RECORD table. In addition, the record table would have required significant changes to support the claims data structure.

Two options were considered for the claims data table structure:

1. A set of tables with fields that match all fields on a claim. This design would work for claims processing but new sets of tables would be required for each type of high volume data (diagnostic index, pharmacy, etc). In this design, the claims datastore would have included tables for services, drugs, treatments, diagnosis, etc.
2. A table which stores the claims data in a JSON data structure. JSON data structures are more flexible than database tables. The contents of the JSON field is known to the application, but not specified in the database. The advantage is we have a table that could be shared with other import types with no additional structural changes. In addition, the JSON data can store complex and deep structures. The disadvantage is that users of the data will need to access it in a different way than they are accustomed to. PostgreSQL fully supports this column type and has added operators and methods for processing it (see http://www.postgresql.org/docs/current/static/functions-json.html).

The two options were evaluated and Option #2 was chosen. The primary reason is that we plan to support many types of high volume data. Option #2 allows us to use a single data model for all high volume data types.

The table below shows how the data are stored. The `PRE_RECORD` table stores a JSON representation of the import in the `import_data` field. The import algorithm is responsible for converting the file to this structure.

PRE_RECORD

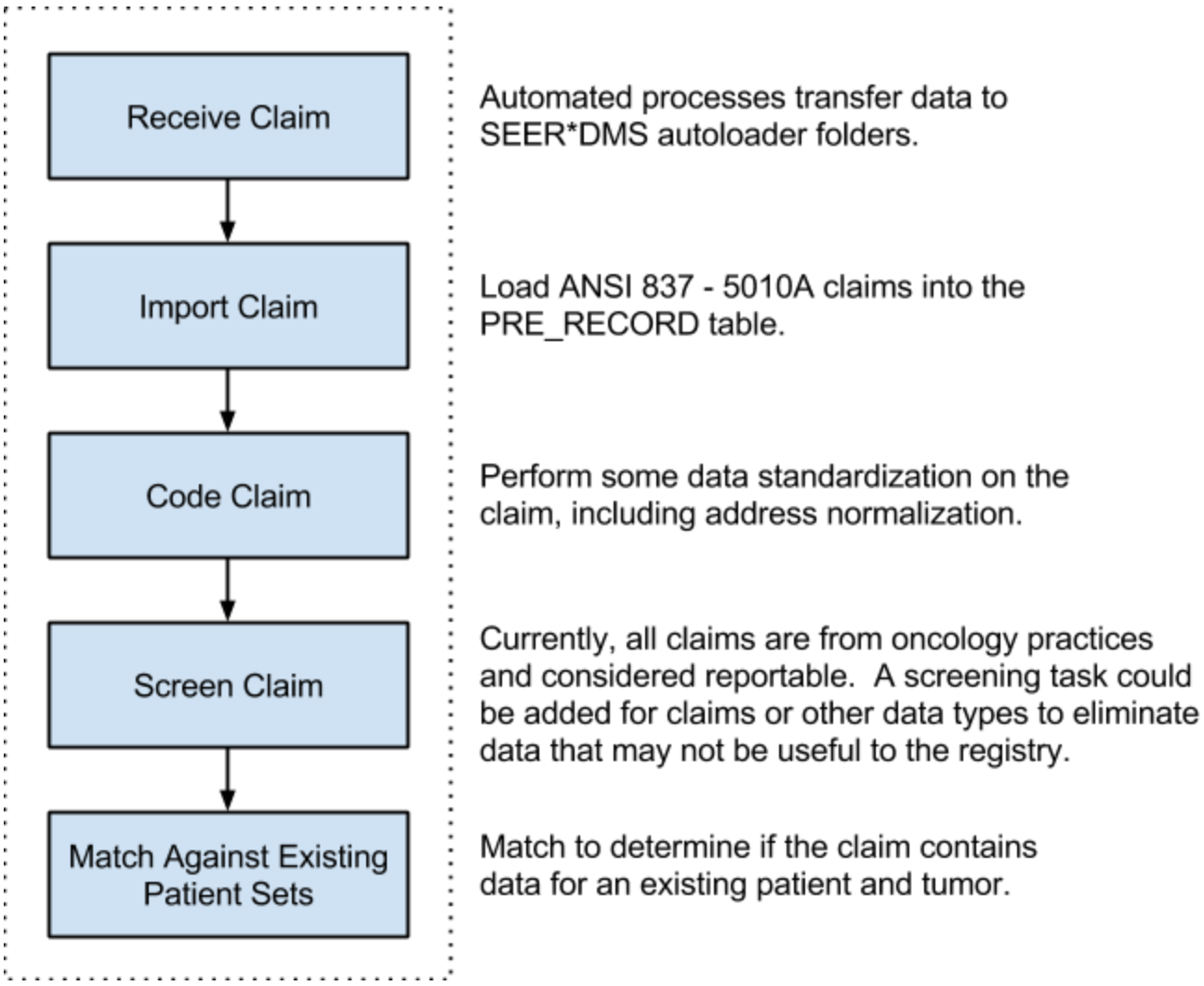
Name	Type	Nullable	Description
pre_record_id	numeric(22)	false	Unique identifier
type	numeric(2)	false	Type of data; initially only "CLAIM"
original_file	numeric(22)	false	The submission file this pre-record came from
rec_index	numeric(22)	false	Line number in the submission file
fac_id	numeric(22)	false	The facility this pre-record came from
import_data	jsonb	false	JSON data structure containing the relevant information from the submission file
matching_patient_id	numeric(22)	true	The matching SEER*DMS patient if one was found
matching_ctc_id	numeric(22)	true	The matching SEER*DMS tumor if one was found
date_loaded	timestamp	false	Date the item was added
date_last_modified	timestamp	false	Date the item was last modified

We need to also track hash codes. Each claim is hashed on the way in and only imported if the hash doesn't already exist. This is how we ensure the same claim is not imported more than once. The initial implementation only uses a full hash check, but will eventually use partial hashes as well.

Name	Type	Nullable	Description
pre_record_id	numeric(22)	false	ID of the pre-record for this hash
hash	varchar(40)	false	Full or partial hash

Once a record is persisted in the database it flows through the workflow described below. The workflow steps will be the same for all `PRE_RECORD` types, but the implementation of each activity may vary by type. Note: The workflow for `PRE_RECORD` data types is completely separate from the main workflow for `RECORD` data.

Processing New Claims in SEER*DMS



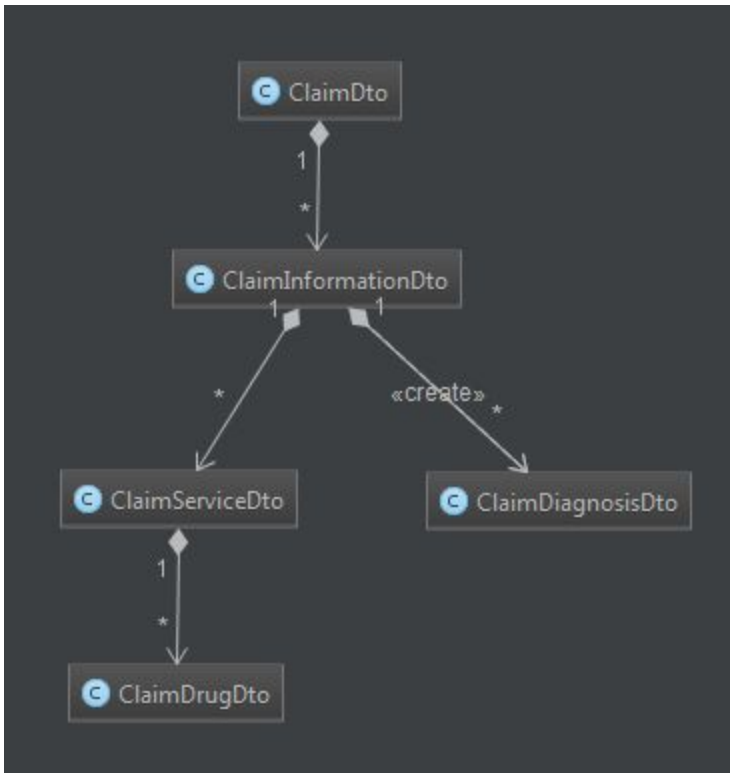
Receive Claim

Claims data are imported through standard SEER*DMS import mechanisms. It is expected that nearly all claims files will be imported via the autoloader.

Import Claim

We currently support the "ANSI 837 - 5010A" format for claims. We have written a parser which can handle this format as well as a few other X12 standard formats. In the future we plan to support additional X12 formats as they become available. We plan on open-sourcing the X12 parser in the future.

Each claim is mapped from the ANSI 837 format to a structure called `ClaimDto`. The `ClaimDto` entities have a similar structure as the raw claim, however we are only storing a subset of the data in order to keep the size of the database at a reasonable level. Each `ClaimDto` is stored in the `PRE_RECORD.IMPORT_DATA` field as the JSON representation of the object.



Claim entity diagram

ClaimDto

A `ClaimDto` represents the all the information relating to a claim. It is a subset of the complete data contained in the claim file. It contains a demographic information and a set of `ClaimInformationDto`.

The patient demographic information is read from the either the 2000C (patient hierarchical level) or 2000B loop (subscriber hierarchical level). The 2000C loop is only used when the patient is a dependant on the insurance. In that case, 2000B represents the insurance holder and not the patient.

Property	Loop	Segment	Element	Notes
provider_first_name	2010AA	NM1	NM104	If 2010AA NM102 = 1 (Person)
provider_middle_name	2010AA	NM1	NM105	If 2010AA NM102 = 1 (Person)
provider_last_name	2010AA	NM1	NM103	If 2010AA NM102 = 1 (Person)
provider_facility	2010AA	NM1	NM103	If 2010AA NM102 = 2 (Non-Person)
provider_qual	2010AA	NM1	NM102	
provider_npi	2010AA	NM1	NM109	
claim_type	2000B	SBR	SBR09	
claim_coverage_type	2000B	SBR	SBR05	Insurance Type Code

pat_gender	2010BA or 2010CA	DMG	DMG03	
pat_dob_year	2010BA or 2010CA	DMG	DMG02	
pat_dob_month	2010BA or 2010CA	DMG	DMG02	
pat_dob_day	2010BA or 2010CA	DMG	DMG02	
pat_last_name	2010BA or 2010CA	NM1	NM103	
pat_first_name	2010BA or 2010CA	NM1	NM104	
pat_middle_name	2010BA or 2010CA	NM1	NM105	
pat_name_type_id	2010BA or 2010CA	NM1	NM101	Entity Identifier Code
pat_name_type_qual	2010BA or 2010CA	NM1	NM102	
pat_name_id_qual	2010BA or 2010CA	NM1	NM108	
pat_address_street	2010BA or 2010CA	N3	N301	
pat_address_street_supp	2010BA or 2010CA	N3	N302	
pat_address_city	2010BA or 2010CA	N4	N401	
pat_address_state	2010BA or 2010CA	N4	N402	
pat_address_zip	2010BA or 2010CA	N4	N403	
pat_ssn	2010BA or 2010CA	REF	REF02	REF01 must be SY
patient_member_id	2010BA or 2010CA	NM1 or REF	NM109 or REF02	Stored in NM1 segment in 2010BA only REF01 must be 1W (2010CA only)
c_last _name	N/A	N/A	N/A	A list of names split on space and hyphen. Derived from pat_last_name, suffix is stripped off and placed in its own field.

c_name_suffix	N/A	N/A	N/A	Derived from pat_last_name, contains only the suffix which can be MD, JR, SR, MR, DR, II, III, IV, V, VI, VII, 2ND, 3RD, 4TH, ESQ, PHD, DDS, DVS, RN, LPN, DO. Coding algorithm accounts for suffixes that have periods with them too (JR. SR. M.D. etc.)
c_dolc_yyyy	N/A	N/A	N/A	The year of the latest date contained in all date fields of the claim
c_dolc_mm	N/A	N/A	N/A	The month of the latest date contained in all date fields of the claim
c_dolc_dd	N/A	N/A	N/A	The day of the latest date contained in all date fields of the claim
c_street_address	N/A	N/A	N/A	Street Address (number + street name) is standardized using an internal address parser
c_primary_site	N/A	N/A	N/A	ICD-O-3 primary site of the primary diagnosis
c_histology	N/A	N/A	N/A	ICD-O-3 histology of the primary diagnosis
c_behavior	N/A	N/A	N/A	ICD-O-3 behavior of the primary diagnosis
c_laterality	N/A	N/A	N/A	ICD-O-3 laterality of the primary diagnosis
c_grade	N/A	N/A	N/A	ICD-O-3 grade of the primary diagnosis.
claim_info				Collection of `ClaimInformationDto`

ClaimInformationDto

A `ClaimInformationDto` represents basic information about a claim including the facility and various dates. It also contains sets of `ClaimServiceDto` and `ClaimDiagnosisDto` entities.

Property	Loop	Segment	Element	Notes
fac_code_value	2300	CLM	CLM05 (0)	Composite Element
fac_code_qual	2300	CLM	CLM05 (1)	Composite Element
claim_freq_code	2300	CLM	CLM05 (2)	Composite Element
pat_control_num	2300	CLM	CLM01	
med_rec_num	2300	REF	REF02	REF01 must be EA

date_admission	2300	DTP	DTP03	DTP01 must be 435
date_discharge	2300	DTP	DTP03	DTP01 must be 096
date_first_contact	2300	DTP	DTP03	DTP01 must be 444
date_last_contact	2300	DTP	DTP03	DTP01 must be 304
date_initial_treatment	2300	DTP	DTP03	DTP01 must be 454
service_npi	2310C	NM1	NM109	Service Facility. Set only if different from provider facility
service_facility	2310C	NM1	NM103	Service Facility. Set only if Service NPI is different than provider NPI
services				Collection of `ClaimServiceDto`
diagnosis				Collection of `ClaimDiagnosisDto`

ClaimServiceDto

A `ClaimServiceDto` represents a single procedure. It also contains a set of drugs used in the service.

Property	Loop	Segment	Element	Notes
num	2400	LX	LX01	
quantity	2400	SV1	SV104	
unit	2400	SV1	SV103	
date_init_tx	2400	DTP	DTP03	DTP01 must be 454
date_last_contact	2400	DTP	DTP03	DTP01 must be 304
date_service	2400	DTP	DTP03	DTP01 must be 472
date_prescription	2400	DTP	DTP03	DTP01 must be 471
date_therapy_start	2400	DTP	DTP03	DTP01 must be 463
date_last_xray	2400	DTP	DTP03	DTP01 must be 455
date_test	2400	DTP	DTP03	DTP01 must be 738 or 739
procedure_qual	2400	SV1	SV101 (0)	
procedure	2400	SV1	SV101 (1)	
procedure_mod	2400	SV1	SV101 (2) SV101 (3) SV101 (4) SV101 (5)	Collection of procedure modifiers

diagnosis_pointer	2400	SV1	SV107(1-12)	Collection of diagnosis pointers. The diagnosis pointers point to the diagnoses this service is providing treatment for. The first pointer in the list is primary diagnosis for the service
drugs				Collection of `ClaimDrugDto`

ClaimDrugDto

A `ClaimDrugDto` represents a drug.

Property	Loop	Segment	Element	Notes
qual	2410	LIN	LIN02	
id	2410	LIN	LIN03	Only if LIN02 is N4
quantity	2410	CTP	CTP04	
unit	2410	CTP	CTP05	

ClaimDiagnosisDto

A `ClaimDiagnosisDto` represents a diagnosis.

Property	Loop	Segment	Element	Notes
type	2300	HI	HI(01-12) (0)	12 separate composite elements
diagnosis	2300	HI	HI(01-12) (1)	12 separate composite elements
c_primary_site	N/A	N/A	N/A	ICD-O-3 Primary site. The result of taking a diagnosis code (stored in the diagnosis property) in the claim in the ICD-9-CM or ICD-10-CM format and converting it using tables provided by NCI and incorporated into MorphologyUtils.
c_histology	N/A	N/A	N/A	ICD-O-3 Histology. The result of taking a diagnosis code (stored in the diagnosis property) in the claim in the ICD-9-CM or ICD-10-CM format and converting it using tables provided by NCI and incorporated into MorphologyUtils.
c_behavior	N/A	N/A	N/A	ICD-O-3 Behavior. The result of taking a diagnosis code (stored in the diagnosis property) in the claim in the ICD-9-CM or ICD-10-CM format and

				converting it using tables provided by NCI and incorporated into MorphologyUtils.
c_grade	N/A	N/A	N/A	ICD-O-3 Grade. The result of taking a diagnosis code (stored in the diagnosis property) in the claim in the ICD-9-CM or ICD-10-CM format and converting it using tables provided by NCI and incorporated into MorphologyUtils.
c_laterality	N/A	N/A	N/A	ICD-O-3 Laterality. The result of taking a diagnosis code (stored in the diagnosis property) in the claim in the ICD-9-CM or ICD-10-CM format and converting it using tables provided by NCI and incorporated into MorphologyUtils.
c_reportable	N/A	N/A	N/A	Using the case-finding lists from NAACCR or ICD-9-CM and ICD-10-CM, this flags whether the diagnosis code is reportable or not.
index	2300	HI	HI(01-12)	The position (1-12) of the diagnosis in the list of diagnosis codes.

JSON representation

The `ClaimDto` is converted to JSON for storage in the `PRE_RECORD` table. Here is an example:

```
{
    "claim_type": "CI",
    "pat_ssn": "123456789",
    "pat_last_name": "DOE-SMITH JR",
    "pat_first_name": "DAVID",
    "pat_middle_name": "J",
    "pat_address_street": "123 ELM ST",
    "pat_address_city": "SPRINGFIELD",
    "pat_address_state": "MI",
    "pat_address_zip": "12345",
    "pat_name_type_id": "IL",
    "pat_name_type_qual": "1",
    "pat_name_id_qual": "MI",
    "pat_gender": "M",
    "provider_facility": "MEDICAL GROUP",
    "provider_npi": "1234567890",
    "provider_qual": "2",
    "c_primary_site": "C779",
    "c_histology": "9860",
    "c_behavior": "3",
    "c_laterality": "0",
    "c_grade": "9",
```

```
"c_dolc_year": "2016",
"c_dolc_month": "02",
"c_prim_day": "02",
"c_last_name": [ "DOE", "SMITH", "DOESMITH"],
"c_name_suffix": "JR",
"claim_info": [
    {
        "fac_code_qual": "B",
        "claim_freq_code": "1",
        "date_first_contact": "20151223",
        "date_last_contact": "20160104",
        "date_initial_treatment": "20160202",
        "pat_control_num": "A37YH667",
        "fac_code_value": "11",
        "diagnosis": [
            {
              "type": "BK",
              "code": "2007",
              "index": 1,
              "c_primary_site": "C779",
              "c_histology": "9860",
              "c_behavior": "3",
              "c_laterality": "0",
              "c_grade": "9",
              "c_reportable": "1"
            },
            {
              "type": "BF",
              "code": "4134",
              "index": "2",
              "c_reportable": "0"
            },
            {
              "type": "BF",
              "code": "5559"
              "index": "2",
              "c_reportable": "0"
            }
        ],
        "services": [
            {
                "num": "2",
                "procedure_qual": "HC",
                "procedure_mod": [ "25" ],
                "date_service": "20050322-20050325",
                "quantity": "1",
                "unit": "UN",
                "procedure": "478331"
```

```
                "diagnosis_pointer" : [ 1 ]
            }
        ]
    }
  ]
}
```

Code Claim

The original fields from the claim are not changed. Calculated fields are added and can always be identified since they all start with "c_".

1. Last Name is split into last name and suffix. (`c_last_name` and `c_name_suffix`). The last name is split further on space and hyphen and the names are stored into a list. This was so we could account for the case that two claims matched on most demographic fields but a change in last name occurred (for example if a name changed from SMITH to SMITH-JONES or SMITH JONES). We do not split compound names (names such as DE LA CRUZ). We do this by checking if the uncoded last name contains any parts that are commonly part of compound last names. The list we currently use is [VERE, VON, VAN, DE, DEL, DELLA, DI, DA, PIETRO, VANDEN, DU, ST., ST, LA, TER]. Names with these parts will not be split but stored as a single entity in `c_last_name`. We also included a combined last name where all parts of a last name (regardless if they are a compound name or not) are combined together as names are sometimes stored this way in claims. For example, the coded last name for SMITH-JONES will contain SMITHJONES. These values are primarily used to help matching.
2. Date of Last Contact (this is taken to be the latest date out of all dates stored in the claim) is stored in `c_dolc_year`, `c_dolc_month`, `c_dolc_day`. It is set using dates from ClaimInformationDto and ClaimServiceDto.
3. Diagnosis code conversions. Currently, diagnosis codes in claims are coded using ICD-9-CM or ICD-10-CM. NCI provided two tables to convert ICD-9-CM and ICD-10-CM diagnosis codes to ICD-O-3 sites, histologies, behaviors, grades and lateralities. If a diagnosis code can be converted to ICD-O-3, then `c_primary_site`, `c_histology`, `c_behavior`, `c_grade`, and `c_laterality` will be set on that disease. In addition, the primary diagnosis values will also be stored on the root of the claim to make searching easier.
4. The street address (number and street name) is coded and stored in `c_street_address`.

The `PRE_RECORD` table will not support an audit log. The number of coded fields are small, and all coded values go into new fields so they do not overwrite existing values.

Screen Claim

All claims are screened to determine if they are cancer case. We are currently using the case-finding lists on the SEER Website (http://seer.cancer.gov/tools/casefinding) to determine reportability. The coded field, `c_reportable`, is set to 1 if the code is contained in the case-finding list; otherwise the field is set to 0.

Match against Existing Patient Sets and Tumors

The next step is to match the claim against all existing patient sets and determine if this is a new case.

The standard linkage algorithm that SEER*DMS already defines is used to find a demographic match. Low score matches are not used for claims data matching. Only matches with a score of x or above are used.

If a match is found, the link to the patient set is added as `PRE_RECORD.MATCHING_PATIENT_ID`. If there are no matches or if multiple multiple matches are found then `PRE_RECORD.MATCHING_PATIENT_ID` is left NULL. Unmatched claims will be periodically rematched to see if new patients have been added.

If a patient was found, `PRE_RECORD.MATCHING_CTC_ID` will be set if the primary diagnosis on the claim matches to an existing tumor within the patient set. The primary diagnosis was converted to ICD-O-3 (`c_primary_site`, `c_histology`, `c_behavior`, `c_grade`, `c_laterality`) in the coding step. Some claims will not have a diagnosis that converts to ICD-O-3 so they will not match any tumor. For the claims that do successfully convert, the ICD-O-3 primary diagnosis is compared with the corresponding values for each tumor linked to the matched patient. A tumor is considered a match if one of the following is true:

1. The site, histology, behavior, laterality in the ICD-O-3 primary diagnosis on the claim match their corresponding values in the tumor data exactly. The date of last contact in the claims data must also match the date of diagnosis on the tumor.
2. The MPH rules determine that the ICD-O-3 data from the primary diagnosis on the claim is at least a possible match to the tumor data.

If a match is found, the `CTC.CTC_ID` of the tumor is used to set `PRE_RECORD.MATCHING_CTC_ID`. If no match is found, then `PRE_RECORD.MATCHING_CTC_ID` is left `NULL`. It is possible that claims could have multiple primary diagnoses that correspond to distinct tumors. Right now, we are only using the first primary diagnosis found to match a tumor to set the PRE_RECORD.MATCHING_CTC_ID. A warning message is logged if another primary diagnosis corresponding to a distinct tumor is found to match a tumor linked to that patient. Multiple distinct primary tumors on a single claim have not been found during any test runs.