Chapter 2: Records and Patient Sets

In SEER*DMS, patient data are stored in "records" and "patient sets." Records contain the original data submitted to the registry by hospitals and other facilities. Record data include medical data submitted by reporting facilities and follow-up information received from non-medical organizations. Each record contains a single report of data -- a NAACCR record contains data from a single abstract, an HL7 E-path record contains data from a single pathology report, etc.

A "patient set" contains the data consolidated from all of the patient's records and provides links to those records. The patient set includes summarized values such as the summarized treatment fields and derived values such as the derived collaborative stage fields.

A registrar edits the patient set fields and, typically, does not change values in the original records. Therefore, the records serve as a permanent reference of the original values sent to the registry; while patient set data are submitted by the registry to SEER, NAACCR, and other organizations.

In this chapter, you'll learn about

- Records
    - Linked and Unlinked Records
- Patient Sets
    - Cancer/Tumor/Case (CTC)
    - Staging Data
    - Summary Treatment (Summary TX)
    - Treatment Procedures (TX and TXr)
    - Facility Admission
    - SEER Course
    - Incoming and Linked Records
    - Auto-building Patient Sets and CTCs
- Audit Logs
- SEER*DMS Databases: Production, Warehouse, and Test

Records

In the SEER*DMS database, the record table contains the original data submitted to the registry. Records are created in the database when data are loaded from files or entered manually (as described in *Chapter 5: Importing Data Files* and *Chapter 6: Data Entry*).

Each record travels through the workflow, triggering the automated and manual "tasks" that must be performed to process the data. The path of a record through the workflow is determined by the record's type. There is a diagram in the SEER*DMS Documentation that shows the workflow path of each record (see *Workflow Diagrams by Record Type* in the documentation section of the SEER*DMS Portal at https://seer.cancer.gov/seerdms/portal). The following are the record types supported by SEER*DMS; some record types may not be available in your registry's configuration.

Abstract Records:

- **Health Record** – An initial abstract record submitted by a medical facility in a registry-defined file format.
- **NAACCR Abstract** – An initial abstract record submitted by a medical facility in the NAACCR record layout or extended file layout (a multi-line format with one line in the NAACCR format followed by optional lines for expanded text fields).

- **NAACCR Modified Record** – An abstract record that was modified and re-submitted to the registry. This record is identical in format to the NAACCR record (the record type field is coded with an 'M').
- **Short Health Record** – A SEER*DMS short health record is a partial or incomplete abstract. A partial abstract may include some patient demographic information, complete or partial histopathology information, treatment information, or other information not sufficient to complete a full case abstract.   Typically, these records provide additional information regarding follow-up, treatment, and admissions related to a previously abstracted case. In rare situations, a short health record may contain the only data related to a case and may be used to build a new CTC.

Pathology Reports:

- **HL7 E-Path** – Electronically transmitted pathology reports transmitted in an HL7 standard file format. These records typically trigger the creation of Abstract Facility Leads. Registry-defined procedures determine whether a CTC or patient set is built when the HL7 record is first processed or if an attempt is made to obtain an abstract first.
- **Pathology Report** – Pathology reports imported into SEER*DMS using a registry-defined file layout. These records typically trigger the creation of Abstract Facility Leads. Registry-defined procedures determine whether a new CTC or patient set is built when the pathology report record is first processed or if an attempt is made to obtain an abstract first.

Casefinding Records:

- **Casefinding** – Data obtained from medical reports which identify new potential cases. These records typically trigger the creation of Abstract Facility Leads. Registry-defined procedures determine whether a new CTC or patient set is built when the casefinding record is first processed or if an attempt is made to obtain an abstract first.
- **NAACCR Casefinding** – A casefinding record transmitted in the NAACCR file format. In some registry algorithms, the record is considered a casefinding record if Class of Case indicates that it is a report of pathology or other lab specimens.

Update Records:

- **NAACCR Update** – The NAACCR short format record used to submit field-specific corrections to data previously submitted to the registry. SEER*DMS matches the record to existing patient set data but does not specifically attempt to match the update record to the original NAACCR record. If SEER*DMS identifies possible matches but cannot identify a definite match, a manual matching task will be created. If SEER*DMS identifies a patient set that is an exact match, SEER*DMS attempts to auto-consolidate the data. A manual Consolidate FUP task will be created if the system cannot complete the auto-consolidation.
- **Registry Update** – A registry-defined record used to submit field-specific corrections to data previously submitted to the registry. SEER*DMS matches the record to existing patient set data but does not specifically attempt to match the update record to the original abstract record. If SEER*DMS identifies possible matches but cannot identify a definite match, a manual matching task will be created. If SEER*DMS identifies a patient set that is an exact match, SEER*DMS attempts to auto-consolidate the data. A manual Consolidate FUP task will be created if the system cannot complete the auto-consolidation.

Other Types of Records:

- **Death Certificate** – A record created from death certificate data files provided by the state. Death certificate data are used to update vital status and cause of death information in patient set data, and to identify potential new cases. If a death certificate indicates a new reportable cancer, an Abstract Facility Lead is created. If an abstract can not be

obtained, death clearance follow-back procedures are performed and a Death Certificate Only case may be created.

- **Death Notice** – Contains information published in public death notices. These records are used to update follow-up variables for existing patients. In the workflow, Death Notice records move through the edits, screening, and matching task. If no match is found, they are held to be re-matched at a later time.
- **Follow-up Only** – Data reported by a facility indicating a history of cancer. These data may provide follow-up information for existing patient sets. In the workflow, Follow-up Only records move through the automatic edit checks and the auto-matching task. If no match is found, they are held to be re-matched at a later time.
- **Follow-up Transmit** – A short record containing follow-up information (date, contacts, current address, etc) for patients previously reported by the facility.
- **Health Index Record** - Data from hospital listings such as diagnostic index, discharge logs, or surgery logs. These data are used for quality assurance related to casefinding and may provide follow-up and treatment data that are consolidated into existing patient sets.
- **ME Cohort** – Data from the Hawai'i component of the US Public Health Service's Multi-ethnic Cohort (MEC) Study which are linked annually to the Hawai'i Tumor Registry data.
- **Supplemental** – Follow-up information from non-medical organizations. These data are matched to the database and used in passive follow-up procedures. Typically, supplemental data are not stored in the record table of the database. The data are matched against the database and used to update patient set data via automated tasks. Updates made to patient set data are recorded in the patient set audit logs. In general, supplemental data are only stored if a manual matching or consolidation task is required to process the data. However, some registries choose to store certain types of supplemental data.

Record type is a general classification that can be assigned to various kinds of data provided in multiple file formats. For example, the Supplemental record type is used for department of motor vehicle, CMS/HCFA, NDI, and voter registration data. Health Records may come from multiple sources (including your registry's abstracting tool) and in different file formats. Specific file layouts are defined in registry-specific configuration files. For information about your registry's configuration, please refer to the Imports section of the SEER*DMS Help menu and the System Files section of System Administration.

In SEER*DMS, records provide a mechanism for storing and accessing the original data values submitted to the registry. Record data do not have to be error-free; records may have null data fields or fields with errors. Problems with the data are generally resolved in the patient set data fields at the time of consolidation. This is a new concept for most cancer registry data editors. In general, it is recommended that registry staff do not change record data fields except to correct errors that prevent the record from being processed in SEER*DMS (see *Chapter 7: Resolving Record Errors* for more information).   All changes are documented in the record's audit log which includes the original data value for each field that was modified. In addition, the original data files are archived and can be accessed by your registry's system administrators.

Linked and Unlinked Records

A "link" is an association between a record and the patient set. The link is established after the record has been matched against the database. A record that has moved through the workflow may be linked to an existing patient set that it matched, linked to a newly created patient set, or held in the database as an "unlinked" record.

- If an incoming record matches an existing patient set, the record is linked to the patient set and the data from the record are consolidated with the data in the patient set. This involves a combination of manual and automated processes in the Main or Passive FUP branches of the workflow.

- o Main Workflow.   For example, an abstract record for a reportable cancer and its matching patient set are forwarded to a manual Consolidate task. In this task, a registrar uses the tools in the Patient Set Editor to complete the consolidation by linking the incoming record and incorporating its data into the consolidated patient set data fields.
- o Passive FUP Workflow. Records are automatically linked and consolidated to a matching patient set in the Passive FUP workflow. A manual task to complete the consolidation is only created if discrepancies between the record and patient set prevent auto-consolidation. Records that follow a passive follow-up route through the workflow include Supplemental, Death Notice, non-reportable Death Certificates, and Follow-up Only records.

- If a record contains reportable data for a new patient, a new patient set can be auto-built using the record's data. The record is linked to the newly created patient set. Reportable abstract records are used to build patient sets in all registries. Death Certificate records, Pathology Report, HL7, Casefinding, NAACCR Casefinding, and Short Health records can also be used to build patient sets.

"Unlinked" records are records that have not been incorporated into a patient set, including:

- Records that are in the workflow and have yet to be matched.
- Records for a new patient that do not contain enough data to create a patient set.   These would include records that only contain follow-up data.
- Records for new patients that do have sufficient data to create a patient set, but the registry prefers to wait for an abstract or other record type. For example, some registries prefer to build a patient set from an abstract rather than a pathology report. If a pathology report is received before the abstract, the pathology report would exit the workflow and remain in the database as an unlinked record. If an abstract is not received, the pathology report can be used to create a patient set at a later time (see *Chapter 28: System Administration*).

The Patient Set Editor provides access to all linked records, enabling users to review the raw data that contributed to the summarized and consolidated Patient Set.   However, some supplemental data are consolidated into the patient but not saved in records. Supplemental data are only stored in a record if a manual task is required to process the data. If a record is created, the record is linked and consolidated to the matching patient. If there is no need to create a record, the changes are made to the patient set in an automatic consolidation task. The following information is recorded in an audit log entry: ID of the import in which the data were loaded, name of the data file, line number in the file, original and new value of each field that was changed.

Patient Sets

A "patient set" is the packet of data in the SEER*DMS database which contains all data related to a particular patient. It includes demographic data, information on all reportable cancers, admissions information, diagnostic procedures, treatment information and text documentation. A "patient set" also includes direct links to the original records that contributed to the patient set data.

The diagram on the following page illustrates the basic structure of a patient set. There is one set of patient-level data, but there can be multiple CTCs – one to represent each of the patient's cancer diagnoses.   In the same regard, each CTC has one set of CTC-level data but may have multiple treatment procedures and admissions. The original records are linked to the patient set.

Note: This diagram is intended to provide an introduction to the patient set and is not comprehensive. There are other data entities associated with a patient set that are not noted in the diagram (images, special studies, AFLs, etc). More complete data diagrams are provided in the Database Documentation section of the SEER*DMS Web Portal (seer.cancer.gov/seerdms/portal).

Patient Set

Patient-level or "Demographic" Data

Patient-level data fields describe the person. For example, the patient's date of birth is stored here and is **not** stored separately with each CTC. Patient-level data are shown on the Demographics page of the editor and include:

-- Demographics (name, sex, race, SSN, marital status, aliases, etc.)
-- Follow-up information (date of last contact, vital status, follow-up physician, etc)
-- Primary and secondary causes of death
-- Patient contact information and current address; and informants

CTC

CTC Level Data

Data Shown on:
-- CTC Page
-- Staging Page
-- Summary TX Page

One set of these data per CTC

Facility and Admission Data

-- Shown on Facility/Adm Page
-- Facility Identifiers
-- Admissions 1 or many

A CTC can have 1 or more Fac/Adm

Course Data

Course Level Fields
e.g., course start date

Treatment Procedure Data

A Course can have 1 or more TX

A CTC can have 1 or more courses

A Patient Set can have 1 or more CTCs

Linked Records

The original or "pristine" records that provided the data for the patient set are stored separately, but they are "linked". A link is an association so that you can open the records while you are viewing the patient set.

There can be any number of linked records.

Cancer/Tumor/Case (CTC)

Cancer/Tumor/Case (CTC) refers to an instance of the disease of interest. Generally speaking, a Cancer/Tumor/Case is a neoplasm with topography, histology, and behavior codes that meet the SEER, local, or special study case reporting guidelines. While this is usually a cancer or a tumor, some code combinations are not truly tumors and others are arguably not cancer. Case is sometimes used interchangeably with "tumor" or "cancer" but case can also be used to refer to a person in the database.

In the documentation, we sometimes use "tumor" instead of CTC, usually within a standard data item name. We may also use "cancer" or "tumor" or "case" to refer to an instance of a disease that has not yet been registered in the database.

In the SEER*DMS database, CTC is the name of a data entity within a patient set. All diagnostic, staging, and treatment data for a reportable neoplasm are stored in a CTC. The CTC data include summarized treatment information, as well as data regarding each admission and treatment for that particular CTC. A patient set may include multiple CTCs, one for each primary cancer diagnosed for the patient.

Staging

Data fields describing the extent or severity of the cancer are displayed on the Staging page of the patient set editor. These include fields from standard staging systems such as Collaborative Stage (CS), the TNM system developed by the American Joint Committee on Cancer (AJCC), and SEER summary staging. Staging data fields are stored in the CTC table of the database.

Software provided by the Collaborative Stage Task Force is integrated into SEER*DMS. The CSv2 software provides coding information for the field lookups and calculates the derived CS fields.

Summary Treatment (Summary TX)

Summary treatment fields are stored in the CTC table of the database and displayed on the Summary TX page of the patient set editor. These fields contain the consolidated treatment values and associated dates for the treatments related to the CTC.

A SEER*DMS polisher sets the value of each Summary TX field to a likely or suggested value. A summarized date of therapy is also calculated by the polisher. The suggested summary value may or may not be the best value when combinations of treatment are being consolidated. For example; the correct consolidation for two single agent chemotherapy entries may be a single agent chemotherapy or multi-agent chemotherapy. If the value set by SEER*DMS is not the best, you may over-ride the value. Instructions for visually editing and consolidating treatment data are provided in *Chapter 11: The Patient Set Editor*.

SEER Course

SEER requires that all treatment administered during the first course of therapy be included for each reported case. See the *SEER Program Coding and Staging Manual* for definitions of first course of therapy. Some registries also collect information on subsequent therapy. This could be treatment for recurrence of cancer or for disease progression. SEER*DMS allows the registry to identify the treatment course and will distinguish between treatment courses when deriving suggested best values for treatment. The default value for treatment course is first course of therapy.

Treatment Procedures (TX and TXr)

CTC treatment data are stored in the *treatment_procedure* table of the database and displayed on TX and TXr pages of the patient set editor. The Facility ID number associated with the place of treatment is listed next to the page link in the patient navigation box. For example, TX (2344) indicates that this treatment was reported by and administered at facility FAC-2344.

TXr is typically used to identify a treatment that was reported by one facility, but given at a different facility. TXr may also be used for registry-specific designations, for example, it is also used to identify treatment procedures with Previous Flag = Yes in SEER*DMS MDCSS.

If the treatment facility is known, the facility number of both the reporting facility and the treatment facility will appear following the TXr. For example, TXr (6431~2344) indicates that according to a cancer report submitted by facility FAC-2344, this treatment was given at facility FAC-6431. A code of 9999 indicates that the facility is unknown. For example, TXr (9999~2344) indicates that according to a record submitted by facility FAC-2344, this treatment was given at an unknown facility. MDCSS Only -- if the reporting facility field is blank for TXr data and the Previous Flag is set to 1, '????' will be displayed in the navigation for reporting facility.

When treatment data are incorporated into a CTC from a NAACCR abstract record, both TX and TXr pages are created if there are treatment data in the NAACCR summary fields that are not included in the NAACCR hospital fields. A TX data page is created from the NAACR RX hospital fields. A TXr data page would be created for the data stored in the NAACCR summary fields that is additional to the hospital information. The performing hospital for the TXr data is set to the "unknown" facility code which is usually FAC-9999.

Facility Admission

The Facility Admission data catalogues information gathered from each patient encounter at a given facility. If a patient has multiple admissions or encounters at the same facility, the facility may provide admission data for each encounter. These data are stored in the *facility_admission* tables of the database.

Incoming and Linked Records

When a patient set is viewed during a consolidation task, records are listed as either "incoming" or "linked". An incoming record is a record which is a tentative match to the patient set. The record is currently being processed in the workflow. It was selected during a Match-Consolidate task for consolidation with the patient set data. If the match is confirmed during the consolidation task, the record will be linked to the patient set and the data will be consolidated. For more information please refer to *Chapter 12: Consolidating Data*.

Auto-Building Patient Sets and CTCs

An entire patient set or components of an existing patient set must be created when new reportable data are incorporated into the database. For example, a new patient set must be created when it is determined that data for a reportable cancer are related to a new patient. A new CTC must be created within an existing patient set when it is determined that data matching an existing patient represents a new, reportable tumor.

To reduce the work involved in creating a new patient set, SEER*DMS auto-builds patient set components based on record data. The transfer of record data to the patient set is automated, when possible. The values of the appropriate data fields are auto-filled, as determined by registry-specific algorithms which are specified for each record type. For some fields, the value is copied directly from a record field to a corresponding patient set field. The values of other patient set data fields are imputed based on multiple fields in the record, or by recoding the value of a single

field in the record. The field mappings and recoding algorithms defined for your registry are documented on the Auto-Build page of the Help menu.

These are the circumstances under which auto-building occurs:

- A reportable abstract record does not match an existing patient set. The system auto-builds a new patient set based on that record.   A single CTC is created within the patient set. Within that CTC, the number of auto-built admissions and treatments will vary based on the record type and the amount of data available on the record.
- A reportable abstract record matches a patient already in the database, but describes a cancer for which that patient has no CTC. During the Consolidate task, a CTC is auto-built when the record is linked to a "New CTC" in the patient set. Within that CTC, the number of auto-built admissions and treatments will vary based on the record type and the amount of data available on the record.
- A reportable record matches an existing patient and CTC, but describes a new facility admission or treatment relevant to this CTC. In this case, the system attempts to auto-build new facility and treatment data fields for that CTC.

SEER*DMS auto-builds patient set data fields using data from a single record. If multiple records for a new patient are consolidated, the patient set data fields are auto-built from one of the records when the Consolidate task is started. SEER*DMS selects an abstract record (NAACCR Abstract, NAACCR Modified, or Health Record) to use for auto-building. If there are two or more abstract records, the abstract record that was first loaded into SEER*DMS is used. The other incoming records are then consolidated with the new patient set.

Audit Logs

Whenever a record or patient set data field is changed by a user or the system, an entry is made in an audit log. A separate audit log is maintained for each record and patient set in the system. The information in the audit log includes the user or process which modified the data, comments related to the change, the date and time of the modification, and the original and modified value of each data field that was changed. A user may provide separate comments for the individual data items or one comment for all data items changed within a single session.

When the audit log is displayed in the record or patient set editor, the changes are grouped by events. Events include workflow tasks, system tasks, and other processes that modify patient data. The audit log events in SEER*DMS are:

- **Ad-hoc Editing** – The changes were not made during a worklist task; a user opened the record or patient set via the patient lookup and modified one or more data items.
- **AFL Update** – Changes made to an AFL were incorporated into the record that triggered the AFL. At this time, reportability is the only field that is changed via this process. The record's reportability is changed to auditable when the AFL status is set to "Not a Reportable Cancer".
- **AFUP Update** – Patient set data were updated when a response to Active Follow-up (AFUP) was entered.
- **Auto-Build** – The patient set or a CTC was auto-built from a record.
- **Auto-Coding** - A change was made to a record as it entered the workflow. Registry-specific coding algorithms are implemented in this automated workflow task.
- **Auto-Consolidation** - The change was made during an automated process to consolidate incoming data.   Typically, this would involve the consolidation of follow-up data. In most registry configurations, the consolidation of other data is performed in manual tasks.
- **Auto-Screening** – The change was made during the automated screening task. This task sets the reportability status of an incoming record.

- **Consolidate** - The change was made by a user during a Consolidate worklist task.
- **Consolidate FUP**- The change was made by a user during a Consolidate FUP worklist task.
- **Data Entry** – One audit log entry is made when a record is saved in a data entry session.
- **Direct Database Update** – Record or patient set data were modified by an external script (audit log entries are not created by all external scripts).
- **Edits Sys Task** – The change was made by a SEER*DMS "polisher" which executed after the Patient Set Edits system task completed. A polisher is a background utility that, according to SEER or registry-defined algorithms, imputes data values based on values in other fields.
- **Image Data Entry** – Record data were updated during a manual Image Data Entry task. An Image Data Entry task is created for each image record imported via the autoloader.
- **Initial Conversion** – The change was made during an automated process to load data from the legacy system.
- **Legacy Audit Log** – Changes documented in this event were made on the legacy system. The audit log information was transferred with the data.
- **Link AFL Source** – This event indicates that a record was auto-linked to a patient set. The record that initiated an AFL is auto-linked to a patient set when its AFL is closed by an abstract linked to the patient set.
- **Link Modification** – This event is only used in the record audit log. It indicates that a user opened a patient set and changed the way a record was linked (e.g., changed a record from linked at patient level to linked to an existing CTC). The record itself was not opened directly, but was modified via a change made in the patient set editor.
- **Mass Change** – Changes were made as specified in a mass change import. The Import ID is listed with the event information.
- **Pathology Screening** - The change was made by a user during a Pathology Screening worklist task.
- **QC Patient Set** - The change was made by a user during a QC Patient Set worklist task.
- **QC Record** - The change was made by a user during a QC Record worklist task.
- **Rapid Follow-up** – Patient set data were updated during an automated Rapid Follow-up worklist task.
- **Resolve Pat Set Errors** – The change was made by a user during a Resolve Patient Set Errors worklist task.
- **Resolve Record Errors** - The change was made by a user during a Resolve Record Errors worklist task.
- **Review Data** - The change was made by a user during a Review Data worklist task.
- **Review Special Studies** - The change was made by a user during a Review Special Studies worklist task.
- **Screening** - The change was made by a user during a Screening worklist task.
- **System** – Changes were made by a SEER*DMS system task.
- **Visual Edit Pat Set** – The change was made by a user during a Visual Edit Patient Set worklist task.

SEER*DMS Databases: Production, Warehouse, and Test

The PostgreSQL database that is accessed via the production version of the SEER*DMS application is referred to as the "live database" or "production database". This database contains up-to-the-minute changes made by the registry's editing or coding staff.

The "data warehouse" is a read-only version of the live database which is created nightly. To reduce load on the production database, it is recommended that external applications query the data warehouse unless the query requires updates made since the previous night. The data warehouse contains all tables in the live database, the NAACCR table, and specialized views which are optimized for extracts, reporting, and other analysis. Views are aggregated, sometimes summarized, copies of data that are specifically structured for dynamic queries and analytics.

The "test server" database contains the data accessed by the test server's version of SEER*DMS. This database is updated on an ad hoc basis by registry IT staff. A registry system administrator executes a reload script that loads a snapshot of the production database from the previous night (the data are loaded from the same snapshot as the data warehouse). If you log into the test server and save a change, that change will immediately be reflected in reports and SQL queries that use the Test Server Database. There is no data warehouse associated with the test server.

The SEER*DMS Web Portal (seer.cancer.gov/seerdms/portal) is the most comprehensive resource of documentation related to the SEER*DMS databases. The following materials are available online:

- Description of each table in the database
- Mapping of database fields to the labels used in SEER*DMS
- Mapping of database fields to NAACCR items
- Diagrams showing table relationships
- NAACCR Mart and NAACCR Extract Documentation