RWD Lineage Initiative

A traceability standard to trust SDTM generated from RWD for regulatory decision-making

Meet the Speakers

Tasha Nagamine

Title: CTO

Organization: Droice Labs

Tasha is an entrepreneur and seasoned technologist with 10+ years of experience in AI, research, and tech strategy. She brings deep expertise in RWD to build data-driven products that have processed over 100 million patient lives. Tasha received her BS in physics from Brown University and left a PhD in AI/deep learning at Columbia University to start Droice.

Anita Umesh, Ph.D.

Title: Biomedical Data Standards Specialist

Organization: Roche/Genentech

Anita is a member of Roche/Genentech's Data Standards and Governance Group. Originally trained as a molecular biologist/biochemist with research experience in cardiovascular science and clinical informatics experience in oncology, she contributes to the Data Tabulation efforts to develop modeling strategies of complex oncology and molecular data into SDTM. She has volunteered on a number of CDISC groups since 2016.

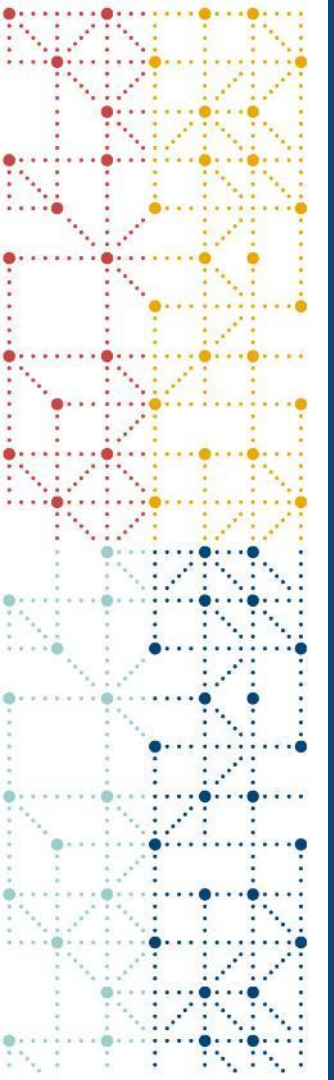
Disclaimer and Disclosures

- *The views and opinions expressed in this presentation are those of the author(s) and do not necessarily reflect the official policy or position of CDISC.*

- *The author(s) have no real or apparent conflicts of interest to report.*

Agenda

1. Introduction
2. Project overview
3. Progress & Updates



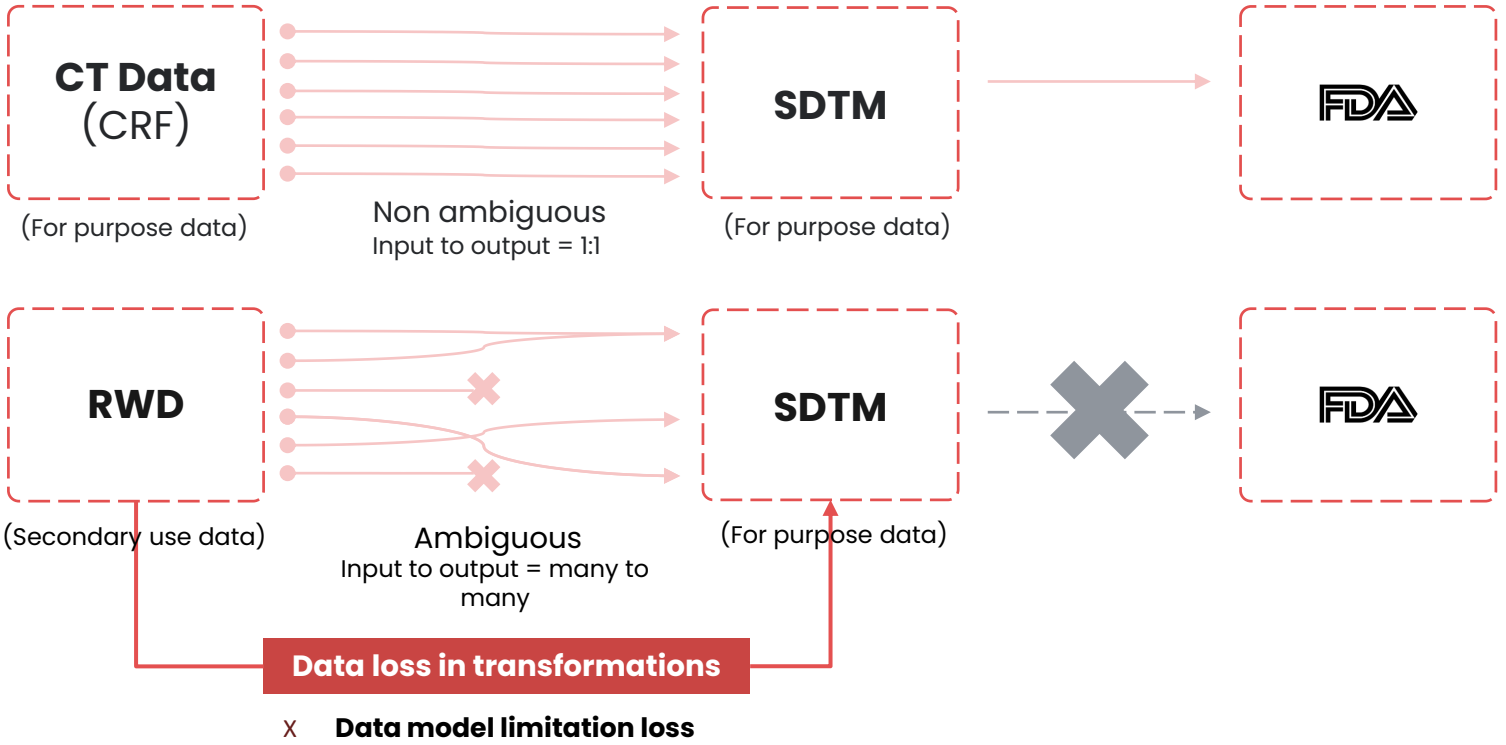
Introduction

Introducing RWD Lineage

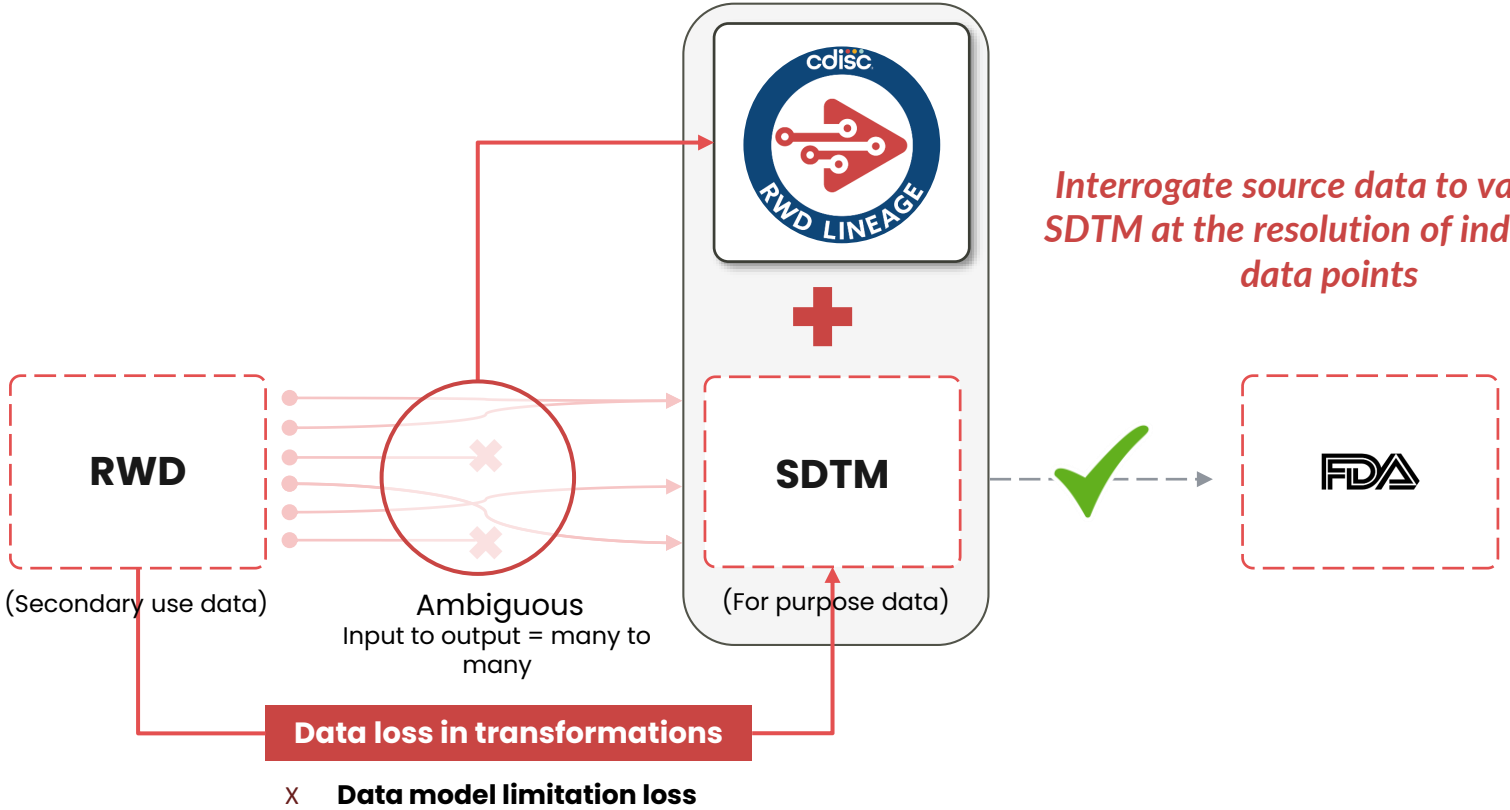
Project Goal

Create a CDISC data exchange standard for lineage metadata that is supplied along with RWD-derived SDTM, which provides the data reliability required by FDA to use RWE as primary evidence.

Reliability in RWD: Challenges

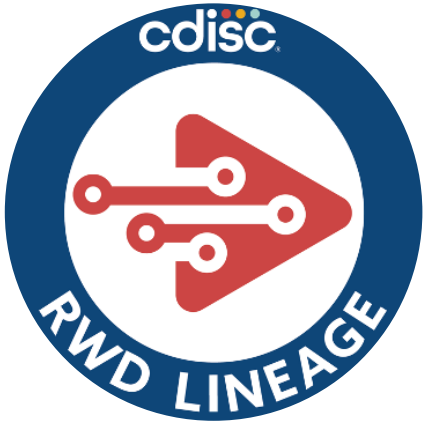


RWD LINEAGE for Reliable RWE



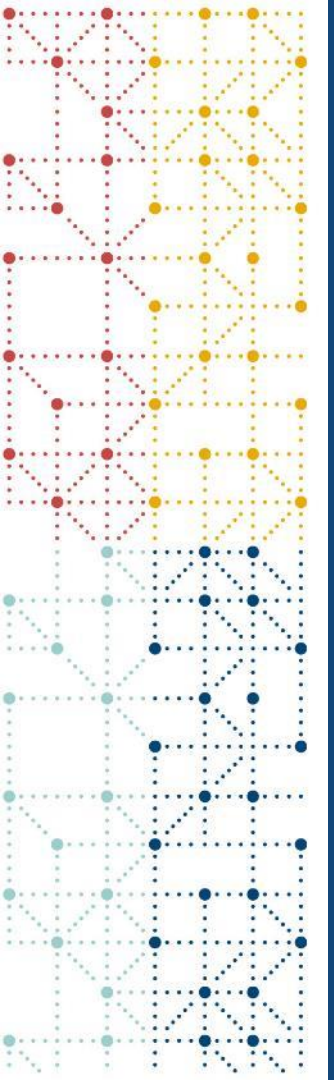
Interrogate source data to validate SDTM at the resolution of individual data points

RWD LINEAGE for Reliable RWE



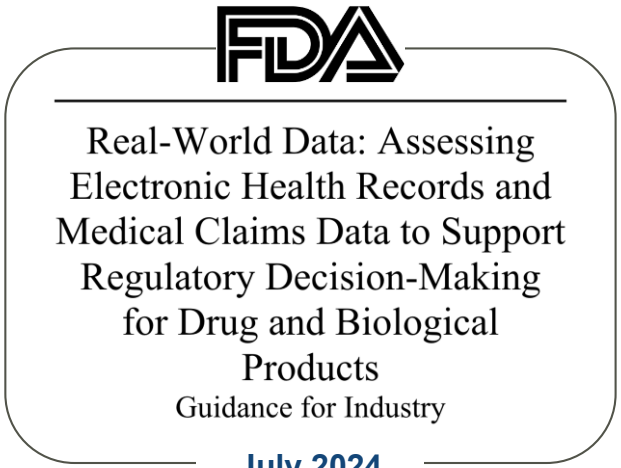
Traceability & Reliability

- ✓ FDA Review
- ✓ FDA Inspection



Project Overview

RWD Lineage: Scope



"For all study designs, it is important to ensure the **reliability**… of the data used to help support a regulatory decision. For the purposes of this guidance, the term **reliability includes accuracy, completeness, and traceability**."

RWD Lineage: Scope

- Creation of a machine-readable, CDISC data exchange standard that is a metadata model for storing element-level data lineage (RWD Lineage).
- RWD Lineage will be a standardized and comprehensive representation of data lineage for each source patient data element that specifies either 1) the location of the element in the output SDTM dataset (Positive Lineage), or 2) that the element was not used in the output analysis dataset (Negative Lineage).
- RWD Lineage should provide the data reliability (data element traceability and ability to quantify errors) needed for FDA to accept RWD-derived SDTM as primary evidence.

Use Cases

Utilizing RWE as **primary evidence**, e.g.:

1. External controls
2. Pragmatic elements

Use Cases: FDA Needs to Interrogate Source RWD

SDTM

MH						
…	USUBJID	MHSEQ	MHTERM	MHPRESP	MHOCCUR	…
…	**001**	**1**	**HISTORY OF MYOCARDIAL INFARCTION**	**Y**	**Y**	…
…	**001**	**2**	**TYPE 2 DIABETES**	**Y**	**N**	…
…	**001**	**3**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**002**	**15**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**003**	**27**	**HYPERTENSION**	**Y**	**N**	…
…	…	…	…	…	…	…

FDA

Real-World Data: Assessing Electronic Health Records and Medical Claims Data to Support Regulatory Decision-Making for Drug and Biological Products

Guidance for Industry

July 2024

"...sponsors should evaluate the completeness, accuracy, and plausibility of the data, including **verifying data against its original source** (e.g., discharge notes, pathology reports, registry records)…"

"...subject matter experts' **review of medical records** (including structured and unstructured data) may be a preferred reference standard for validation of clinical events identified by diagnosis codes or automated algorithms…"

Project Scope: Requirements

- Lineage must support FDA's definition of reliability (accuracy, completeness, traceability)
  - Data point-level traceability
- Lineage must support quantitative validation of RWD against source data
- Supports multi-entity problems
  - Needs to deal with all stages of RWD processing
  - Data: sources > intermediate parties > sponsors > FDA
- Must address stakeholder gaps
  - e.g., FDA reviewer needs to know how changing inclusion/exclusion affects the cohort
  - Lineage should take stakeholder use cases (e.g., audit, validation, review, QA) into account (e.g., using software)

Reliability: Atomic Lineage for Individual Data Points

SDTM

MH						
…	USUBJID	MHSEQ	MHTERM	MHPRESP	MHOCCUR	…
…	**001**	**1**	**HISTORY OF MYOCARDIAL INFARCTION**	**Y**	**Y**	…
…	**001**	**2**	**TYPE 2 DIABETES**	**Y**	**N**	…
…	**001**	**3**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**002**	**15**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**003**	**27**	**HYPERTENSION**	**Y**	**N**	…
…	…	…	…	…	…	…

Reliability: Atomic Lineage for Individual Data Points

SDTM

MH						
…	USUBJID	MHSEQ	MHTERM	MHPRESP	MHOCCUR	…
…	**001**	**1**	**HISTORY OF MYOCARDIAL INFARCTION**	**Y**	**Y**	…
…	**001**	**2**	**TYPE 2 DIABETES**	**Y**	**N**	…
…	**001**	**3**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**002**	**15**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**003**	**27**	**HYPERTENSION**	**Y**	**N**	…
…	…	…	…	…	…	…

Source RWD

PT_DX		
PT_ID	ICD10	TERM
21962	I10	Essential hypertension
21962	N18.3	Chronic kidney disease, stage 3
…	…	…

Reliability: Atomic Lineage for Individual Data Points

SDTM

MH						
…	USUBJID	MHSEQ	MHTERM	MHPRESP	MHOCCUR	…
…	**001**	**1**	**HISTORY OF MYOCARDIAL INFARCTION**	**Y**	**Y**	…
…	**001**	**2**	**TYPE 2 DIABETES**	**Y**	**N**	…
…	**001**	**3**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**002**	**15**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**003**	**27**	**HYPERTENSION**	**Y**	**N**	…
…	…	…	…	…	…	…

Source RWD

PT_DX		
PT_ID	ICD10	TERM
21962	I10	Essential hypertension
21962	N18.3	Chronic kidney disease, stage 3
…	…	…

VITALS		
Patno	Vital	Value
19251	BP	150/110
19251	BMI	28.3
…	…	…

NOTES	
PT_ID	TEXT
19251	Medical history: Type 2 DM on insulin, CHF, HTN, CKD3, …
19251	Discharge summary: …
…	…

Reliability: Atomic Lineage for Individual Data Points

SDTM

MH						
…	USUBJID	MHSEQ	MHTERM	MHPRESP	MHOCCUR	…
…	**001**	**1**	**HISTORY OF MYOCARDIAL INFARCTION**	**Y**	**Y**	…
…	**001**	**2**	**TYPE 2 DIABETES**	**Y**	**N**	…
…	**001**	**3**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**002**	**15**	**HYPERTENSION**	**Y**	**Y**	…
…	…	…	…	…	…	…
…	**003**	**27**	**HYPERTENSION**	**Y**	**N**	…
…	…	…	…	…	…	…

True positive

True positive

False negative

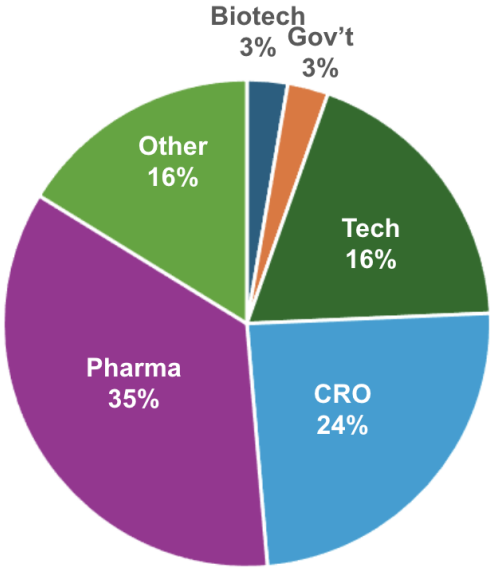
Interrogate source data to validate the accuracy of individual data points

Atomic lineage provides traceability of individual data points, including coordinates and values, from target SDTM to original source data

Updates

Introducing RWD Lineage Team

- Team kickoff meeting: July 30, 2024
- 64 team members registered
- Participation from 37 unique organizations
- Meets every other Tuesday, 11am Eastern
- 15-20 attendees per meeting
- https://wiki.cdisc.org/display/RWDLIN/RWD+Lineage



Introducing RWD Lineage

Motivation

- To generate reliable RWD in SDTM for regulatory use, additional information is needed to audit source data and quantify the information loss and performance of data transformations.
    - Traceability + Quantitative Quality

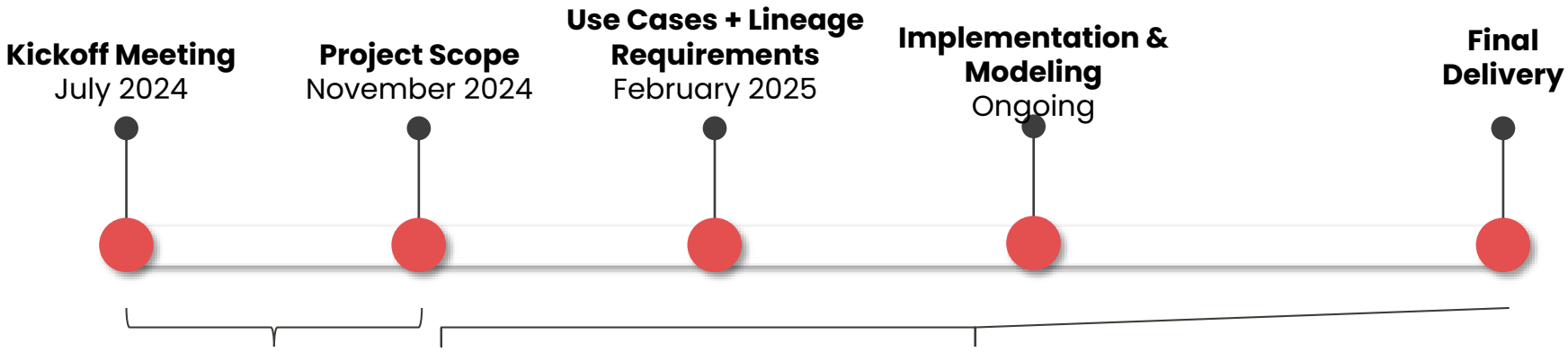
Initial definition:

- RWD Lineage will be a standardized and comprehensive representation of data lineage for each source patient data element that specifies either 1) the location of the element in the output SDTM dataset (Positive Lineage), or 2) that the element was not used in the output analysis dataset (Negative Lineage).

Standards development:

- RWD Lineage and quality will be represented in a CDISC standard metadata model.
- This model will be a machine-readable data exchange standard

Project Timeline



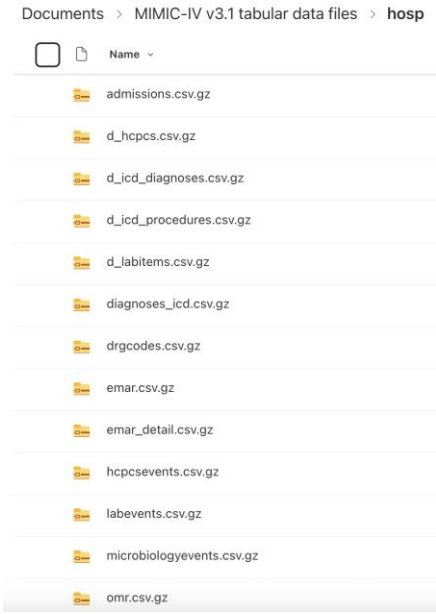
- Decide use cases
- Define timelines

- Implement use cases to inform the lineage
- Package lineage into a standard

https://wiki.cdisc.org/display/RWDLIN/RWD+Lineage+Project+Charter

RWD Lineage Implementation Updates

- Source data selected for use: Medical Information Mart for Intensive Care IV (MIMIC-IV) from Physionet
- Indications and Use Case
  - Parkinson's Disease, ECA
  - Breast Cancer*
- Source (22 tables) -> SDTM -> Lineage to source
- Expressed in Define.xml



*in alignment with CDISC 360i

RWD Lineage Implementation Updates - Breast Cancer

- Cohort creation specifications on wiki page

Study indication	Variable	Variable type	Criteria	MIMIC source	Table name	MIMIC-IV v3.1 specification
Breast cancer	Breast cancer	Screening cohort	Patients with Invasive breast cancer	MIMIC-IV v3.1	diagnosis_icd	subject_id where for • icd_version = '9', icd_code is '174*' (where * is any number) • icd_version = '10', icd_code is 'C50*' (could have 0-3 digits after 'C50') Note that these should filter out patients with 'Malignant neoplasm' of 'breast', and is expected to include both female and male patients
Breast cancer	Sex	Inclusion criteria	Female	MIMIC-IV v3.1	patients	subject_id where gender = 'F'
Breast cancer	Age	Inclusion criteria	>18 and <60 at diagnosis	MIMIC-IV v3.1	diagnoses_icd,d_icd_diagnoses,admissions,patients	1. 'diagnosis_icd' table where seq_num = '1' for subject_id containing • icd_version = '9', icd_code is '174*' (where * is any number) • icd_version = '10', icd_code is 'C50*' (could have 0-3 digits after 'C50') 2. for subject_id identified in (1), use the hamd_id as a key to identify the earliest 'admittime' for subjects from the 'admissions' file 3. for subject_id identified in (2), calculate 'Age at diagnosis' by taking the year from earliest 'admittime' in 'admissions' file which is the year of diagnosis, and calculate a. (year of diagnosis) - ('anchor_year' from 'patients' table) b. Age at diagnosis = ('anchor_age' from 'patients' table) + (difference determined in 3a) 4. 'patients' table for subjects_id the calculated 'Age at diagnosis' in step 3b = >18 or <60

Breast Cancer Cohort Generation

patients

- gender = 'F'
- calculate age at diagnosis >18 - <60

(admissions.year_of_earliest_diagnosis) - (patients.anchor_year)

diagnosis_icd

- icd_version = '9', icd_code = '174*'
- icd_version = '10', icd_code = 'C50*'
- use key = hamd_id for cases where seq_num = '1'

admissions

- key = hamd_id
- identify earliest 'admittime'

⇒ year_of_earliest_diagnosis

join on hamd_id

calculate age at diagnosis to identify patients with age at diagnosis >18 <60

emar

medications like

'%Doxorubicin%', '%Adriamycin%', '%Epirubicin%', '%Ellence%',

'%Paclitaxel%', '%Taxol%', '%Docetaxel%', '%Taxotere%',

'%5-fluorouracil%', '%Capecitabine%', '%Xeloda%',

'%Cyclophosphamide%', '%Carboplatin%', '%Paraplatin%'

procedures_icd

procedures

WHERE (icd_code LIKE '0HB%' AND icd_version = 10)

OR (icd_code LIKE '852%' AND icd_version = 9)

OR (icd_code LIKE '854%' AND icd_version = 9)

Source-to-Target Output

subject_id	gender	anchor_age	anchor_year	anchor_year_group	dod
1	F	43	2143	2011 - 2013	5/23/2150
2	F	60	2149	2017 - 2019	5/11/2150
3	F	59	2189	2017 - 2019	3/26/2189
4	F	57	2115	2017 - 2019	12/11/2115

STUDYID	DOMAIN	USUBJID	MHSEQ	MHTERM	MHDECOD	MHSTDTC	MHDTC	VISITNUM	VISIT	MHCAT	MHSTAT	MHOCCUR	MHEVDTYP
MIMICIVBC	MH	MIMICIVBC-12	1	C50919	Malignant neop	2149-01-29 22	2149-01-29 22	4	ICU ADMISSIC	MALIGNANCY HISTORY		Y	INITIAL DIAGNOSIS
MIMICIVBC	MH	MIMICIVBC-12	2	C50912	Malignant neop	2149-03-02 23	2149-03-02 23	7	ICU ADMISSIC	MALIGNANCY HISTORY		Y	
MIMICIVBC	MH	MIMICIVBC-12	3	C50912	Malignant neop	2149-11-05 18	2149-11-05 18	10	ICU ADMISSIC	MALIGNANCY HISTORY		Y	
MIMICIVBC	MH	MIMICIVBC-17	1	C50912	Malignant neop	2149-03-03 0:0	2149-03-03 0:0	1	ICU ADMISSIC	MALIGNANCY HISTORY		Y	INITIAL DIAGNOSIS
MIMICIVBC	MH	MIMICIVBC-17	2	C50512	Malignant neop	2149-09-01 17	2149-09-01 17	4	ICU ADMISSIC	MALIGNANCY HISTORY		Y	
MIMICIVBC	MH	MIMICIVBC-18	1	C50911	Malignant neop	2115-11-02 18	2115-11-02 18	1	ICU ADMISSIC	MALIGNANCY HISTORY		Y	INITIAL DIAGNOSIS
MIMICIVBC	MH	MIMICIVBC-18	2	C50919	Malignant neop	2115-11-29 22	2115-11-29 22	2	ICU ADMISSIC	MALIGNANCY HISTORY		Y	
MIMICIVBC	MH	MIMICIVBC-19	1	C50012	Malignant neop	2189-03-07 1:1	2189-03-07 1:1	1	ICU ADMISSIC	MALIGNANCY HISTORY		Y	INITIAL DIAGNOSIS
MIMICIVBC	MH	MIMICIVBC-19	2	C50012	Malignant neop	2189-03-25 20	2189-03-25 20	2	ICU ADMISSIC	MALIGNANCY HISTORY		Y	

Read More…

- FDA RWD/RWE Guidance for Industry
  - Real-World Data: Assessing Electronic Health Records and Medical Claims Data To Support Regulatory Decision-Making for Drug and Biological Products
  - Real-World Data: Assessing Registries To Support Regulatory Decision-Making for Drug and Biological Products
  - Use of Electronic Health Record Data in Clinical Investigations
  - Considerations for the Use of Real-World Data and Real-World Evidence To Support Regulatory Decision-Making for Drug and Biological Products
  - Data Standards for Drug and Biological Product Submissions Containing Real-World Data
- PhUSE-US 2024
  - Transforming RWD for Regulatory Submissions: How to Use SDTM for RWD

Thank You!

Join us as a volunteer!

Thank You!