Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Duke-Robert J. Margolis, MD, Center for Health Policy
1201 Pennsylvania Ave, NW, Suite 500, Washington, DC 20004
January 22, 2019

Welcome and Introductions

FDA Opening Remarks

Session I: Transforming Raw Data into Research-Ready Data

Digital Research Network

Patient-centered. Research ready.

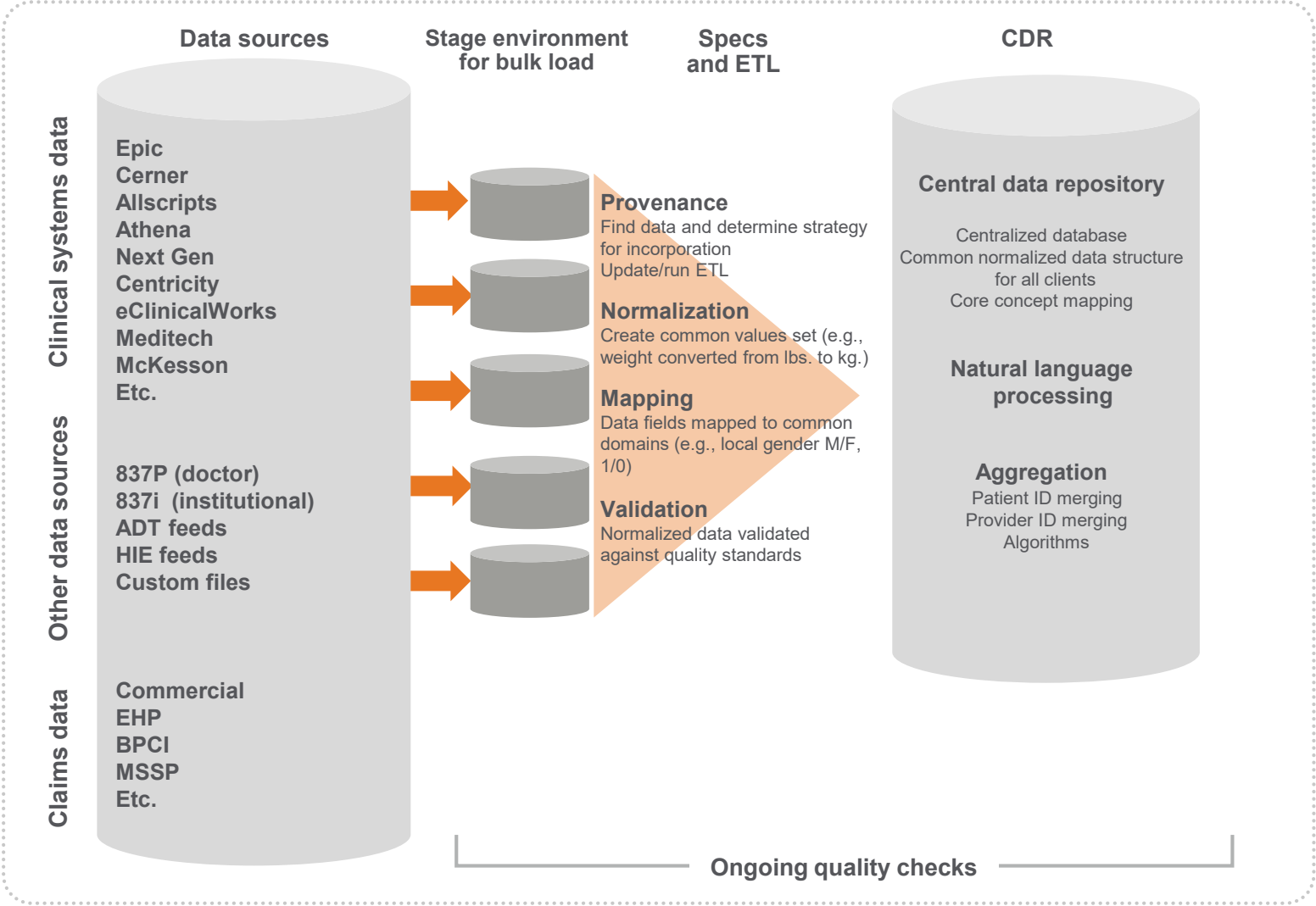
Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Session I: Transforming Raw Data into Research-Ready Data

January 22, 2019

OPTUM®

Optum 'Data Factory' High Level Overview



OPTUM®

High Level Overview of Optum Processes and Technologies for Data Extraction

Data Acquisition	Data load and normalization into proprietary data model	Quality Analytics
• Create secure data acquisition pipeline- through VPN or secured file transfer process (encrypted) • Ensure data flowing daily • Define expected standard file formats based on data type (HL7, Claims, etc.) • Reusable data extraction logic based on experience with multiple EMR/data warehouse structures		

OPTUM®

Optum Processes and Technologies for Data Extraction

Ensuring extraction of the most recent data from various data sources…

- Optum Analytics provides services under a Business Associate Agreement to our customers
- Our Customers provide access to their data to support certain Health Care Operations
  - Accurate and current data critical for Care Coordination activities
  - Work together to ensure access and accuracy

OPTUM®

Optum Processes and Technologies for Data Curation

Data Acquisition	Data load and normalization into proprietary data model	Quality Analytics
	• Leverage industry standards (Code sets) to normalize data as a part of Extraction Transformation Load process • Use Machine learning techniques to normalize free-text data sets from text fields or notes • Subject Matter Experts used for Labs and Medication Mapping • Internal Tools and Machine Learning processes developed to ensure consistency in data across all customers EMRs	

OPTUM®

Optum Processes and Technologies for Data Curation

Provenance Identification

- Analyze provider data stores (Multiple sources)
- Locate candidate sources in the raw data
- Characterize the data:
  - Variety of sources
  - Data type
  - Extent of population
  - Data quality

- If multiple data sources for one element, compare data and specify provenance cascade
- Document provenance for future reference and verification review

OPTUM®

Optum Processes and Technologies for Data Curation

Normalization – highly dependent on data type

- Structured Data

  Standard Terminology – use crosswalks

  Custom codes – use regular expressions, semantic logic, machine learning techniques
- Unstructured Data –

  requires extensive business requirement definition- NLP

Accuracy Verification during Mapping

- Structural testing concerns the format of data
- Semantic testing concerns the meaning of data
- Referential testing concerns the relationship between data

OPTUM®

Transforming Local Lab Result and Units to Normalized Values

Local Name	Local Result	Normal Range	Local Units	Mapped Name	Mapped Unit	Normalized Value
Prostate specific antigen	0.33	(null)	ng/ml	Prostate Specific Antigen	ng/ml	0.33
Albumin, serum	3630	3848-5304	mg/dl	Albumin	g/dl	3.63
Triglyceride	68	See lab report	(no units)	Triglycerides (TG)	mg/dl	68
C-reactive protein, serum	0.12	See lab report	mg/dl	C-reactive protein (CRP)	mg/L	1.2
Thyroid stimulating hormone	0.8	0.5-6.0	miu/l	Thyroid stimulating hormone (TSH)	uu/ml	0.8

OPTUM®

High Level Overview of Optum Processes and Technologies for Data Extraction

Data Acquisition	Data load and normalization into proprietary data model	Quality Analytics
		• Source to Target Mapping for new data sources • Analytical algorithms to validate normalized data sets using automated and semi-automated methods • Develop data integrity checking processes run during initiation and each monthly data refresh

OPTUM®

Data Quality Verification: Using Automated Analytics

Volumetric Analysis

- High Level Volumetric: examine trends over time for each table to identify any gaps in the data
- Mid Level Volumetric: examine trends over time of particular items of interest overall and by source of data
  - Volumes for specific lab tests, medication class

Linkage Reports: examine "joining" rates between the various tables to ensure consistency in patient IDs and encounter IDs (where available) across the various data sources.

OPTUM®

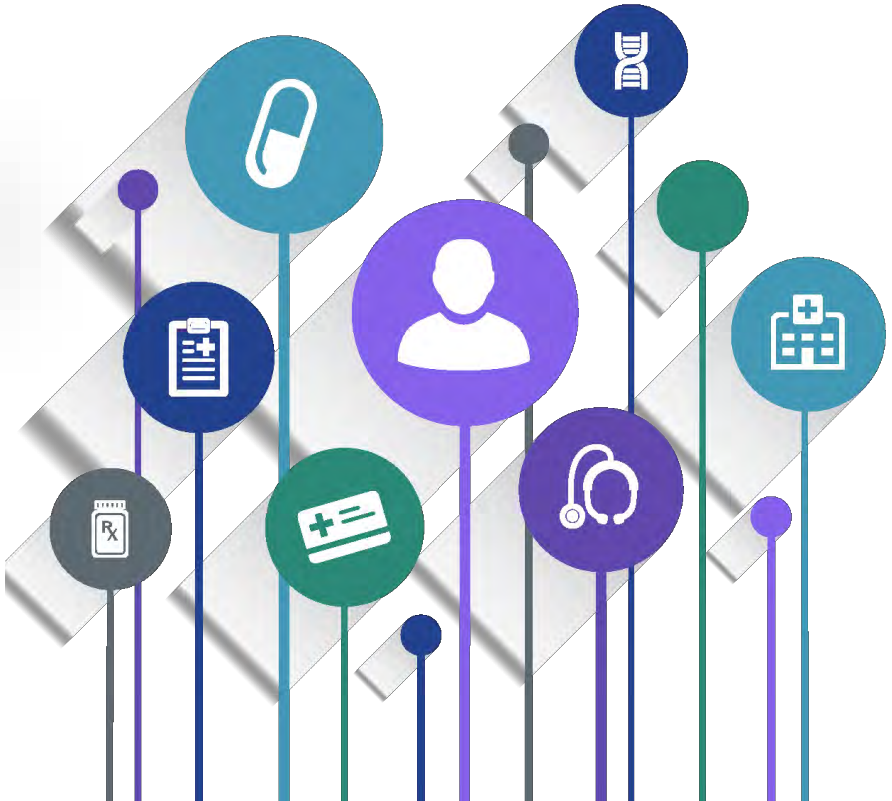
Thank you

Cynthia Senerchia
Vice President, Clinical Operations
Digital Research Network

OPTUM®

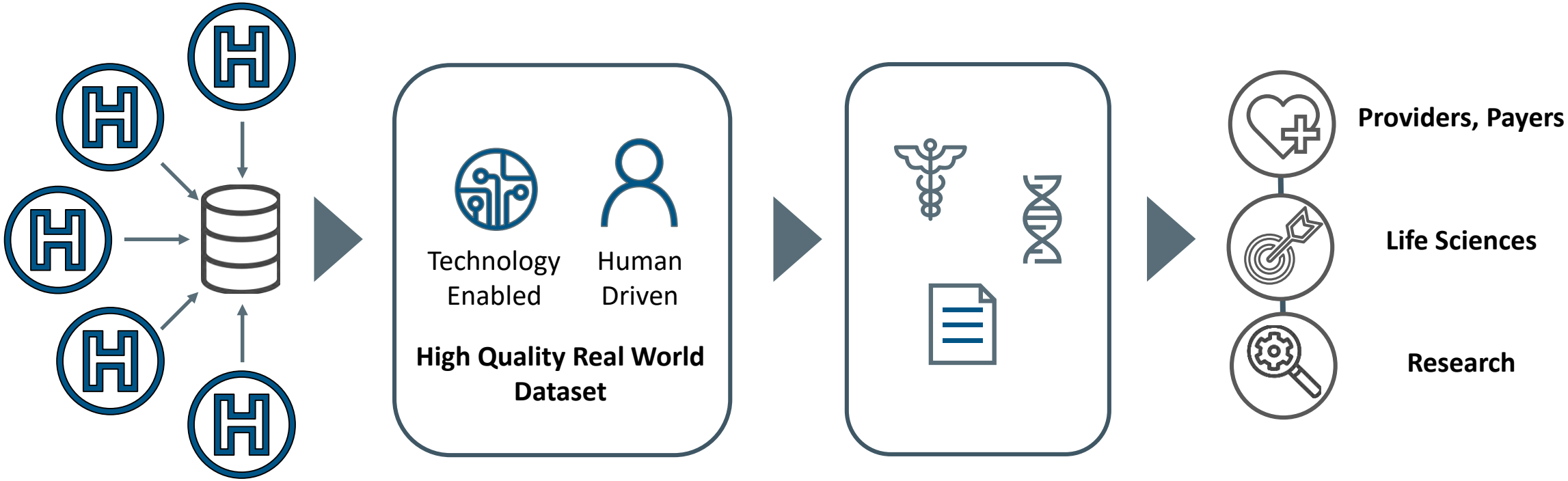
Session I: Transforming Raw Data into Research-Ready Data

COTA's Approach to Data Curation



COTAHEALTHCARE.COM

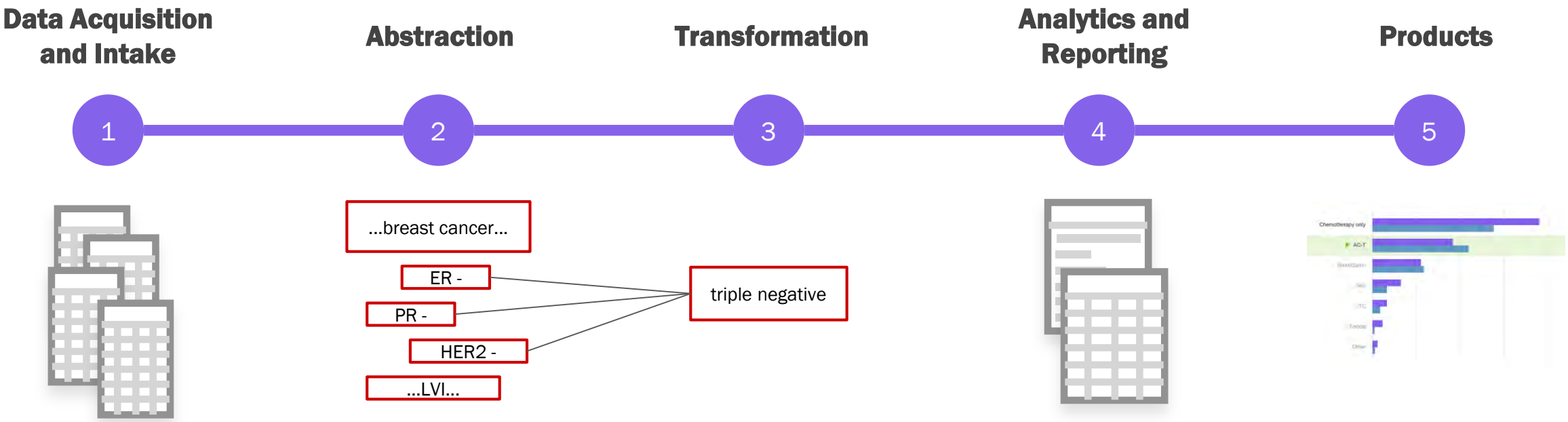
COTA transforms complex clinical data into Real World Data



1. Aggregate Clinical, Genomic and Financial Data Across Cota's Provider Network
2. Process Structured And Unstructured Data Via Data Enrichment Platform
3. Stratify RWD By CNA To Enable Comparison of "Clinically Identical" Patients
4. Analytics to Drive Insights Across The Healthcare Ecosystem

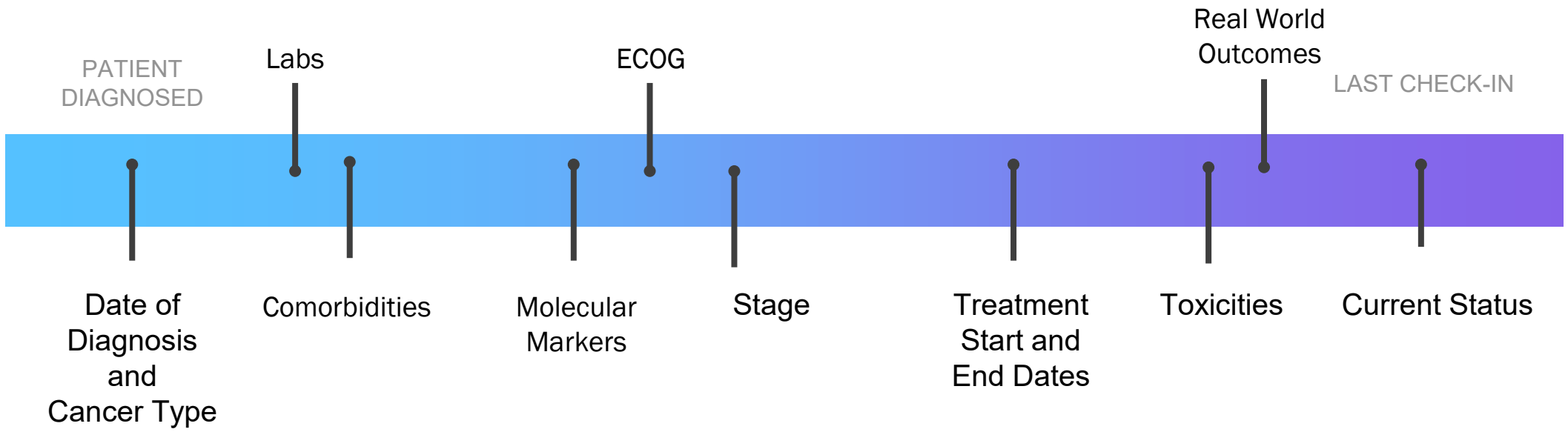
The Journey to Make COTA RWE

COTA RWE is derived via in-house technology that enables the collection and expression of comprehensive patient data supported by source attribution.



Event-Driven Patient Timeline

COTA's flexible model is designed to accommodate multiple similar facts over the entire patient timeline.



Data Acquisition and Intake

Abstraction begins when new documents and patient data are received.

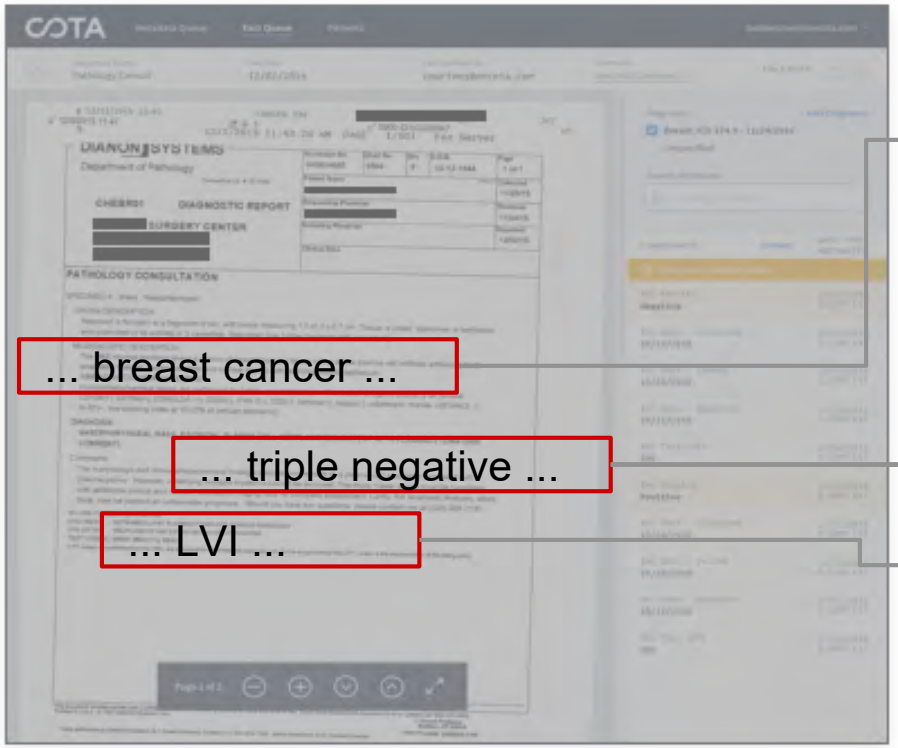
Data Source	Examples	File Type
Tabular data Data exported from one of the many sources in the provider's system or claims from Payer.	Tumor registry, utilization reports, BI reports, and claims	Character-delimited files (CSV)
EHR media All files are scanned or created by the provider's system.	Surgical Pathology Report, Visit notes	PDF, JPEG, TIFF
Programmatic EHR messages Data generated in digital text format from the provider's system.	ORU, ADTs, MDMs, RAS	HL7, CCD, FHIR

Abstraction

Clinical experts use standard and controlled terminology to turn unstructured information to structured data, which is then subject to robust review, rules, and quality assurance.

The Source
Pathology Report

Patient Facts
Interpreted Values



MRN 123, DISEASE

MRN 123, ER NEGATIVE

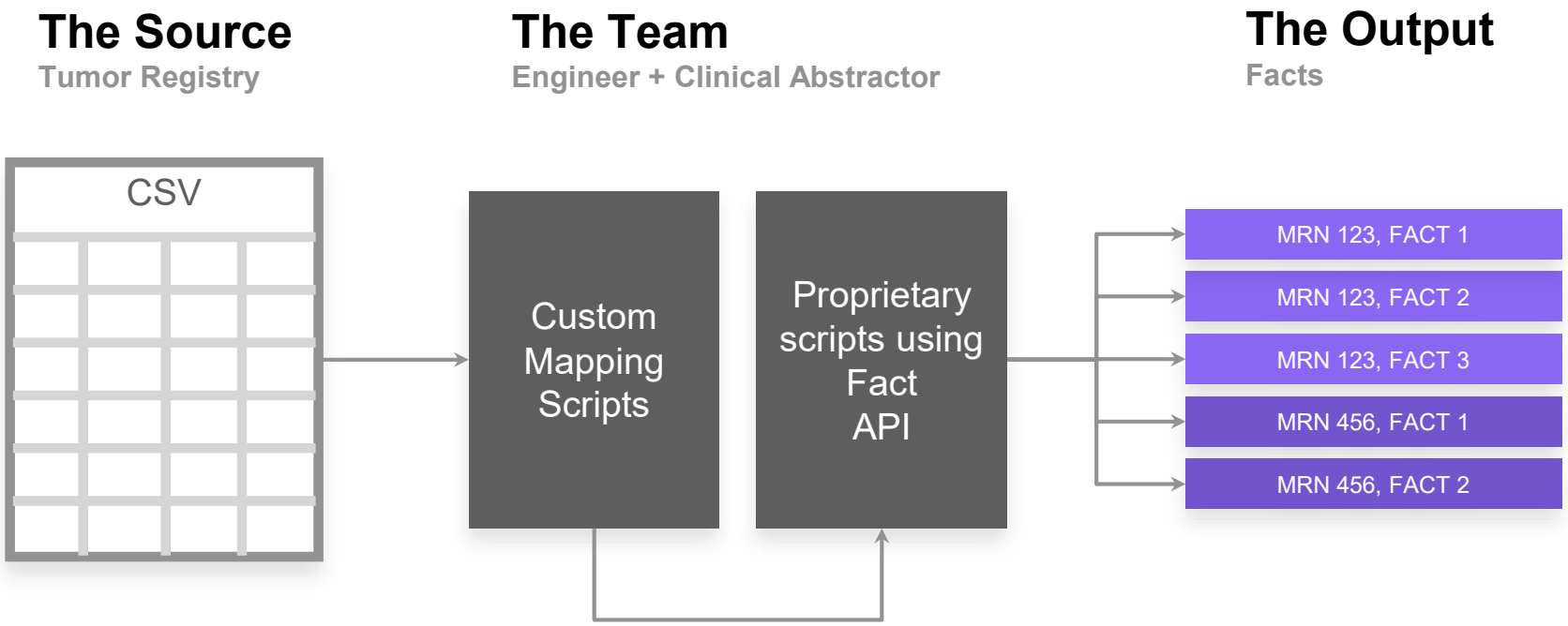
MRN 123, HER2 NEGATIVE

MRN 123, PR NEGATIVE

MRN 123, LYMPHOVASCULAR INVASION PRESENT

Abstraction

Structured and semi-structured sources are leveraged wherever possible, and augment manual abstraction, process optimization, and operational intelligence.



Transformation

The ETL layer handles all medical calculations, roll-ups, and normalizations, and generates data that powers COTA products and benchmarks.

Medical Calculations	Proprietary Calculations	Data Tables
• Staging • Time Deltas and events for Kaplan-Meier • Prognostic scoring systems	• CNA assignment • PHI scrubbing • Progression	• Staging • Molecular testing • Labs

Quality Assurance Overview

A multi-phase approach applying automated and human-driven activities is required to optimize and monitor data quality.

- Quality control at the point of data entry:
  - Data validation (restricted ranges, realistic dates, control lists, no free text)
  - Careful management of external data sources not entered by humans (SLAs, mapping, testing, data validation)
- Upfront abstractor testing against gold standard
- Ongoing abstractor monitoring using randomized double-blind abstraction and IRR measurement
- Programmatic checks for improbable scenarios

The Role of Technology

Natural Language Processing (NLP) has great potential to help, but we are concerned about accuracy.

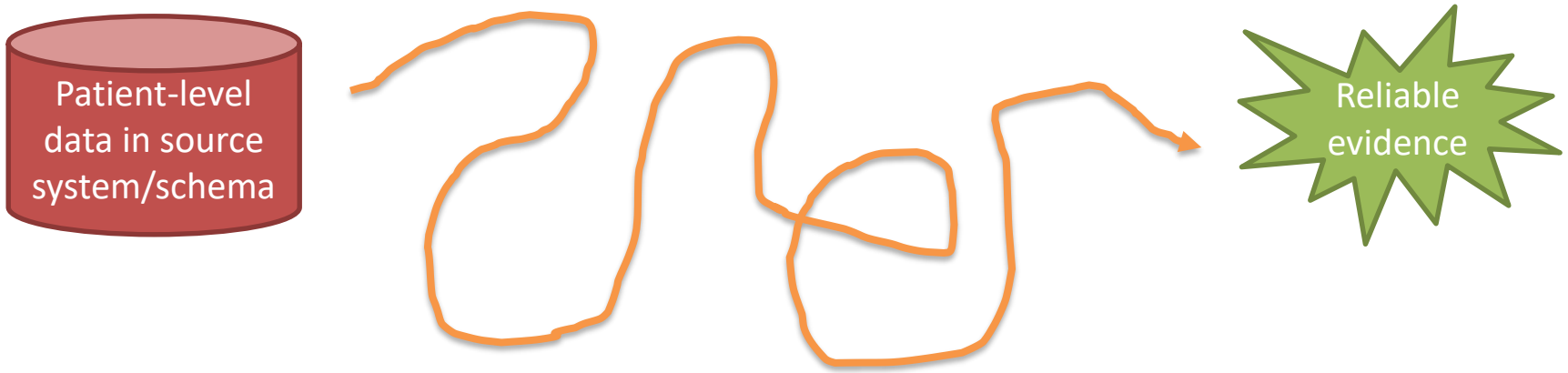
- Much of “what matters” in oncology is found only in complex physician narratives. NLP accuracy today is inadequate for these scenarios.
- Decisions regarding individual data elements are always made by humans with appropriate training.
- We rely on an increasingly sophisticated “suggestion engine” to improve human efficiency and accuracy.
- As accuracy improves, the suggestion engine will be compared against humans and IRR calculated.
- For individual data element/source combinations that prove superior to human abstractors, we can consider replacing human abstractors in the future.

Session I: Transforming Raw Data into Research-Ready Data

Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Patrick Ryan, PhD
Janssen Research and Development
Columbia University Medical Center

The journey to real-world evidence



One-time

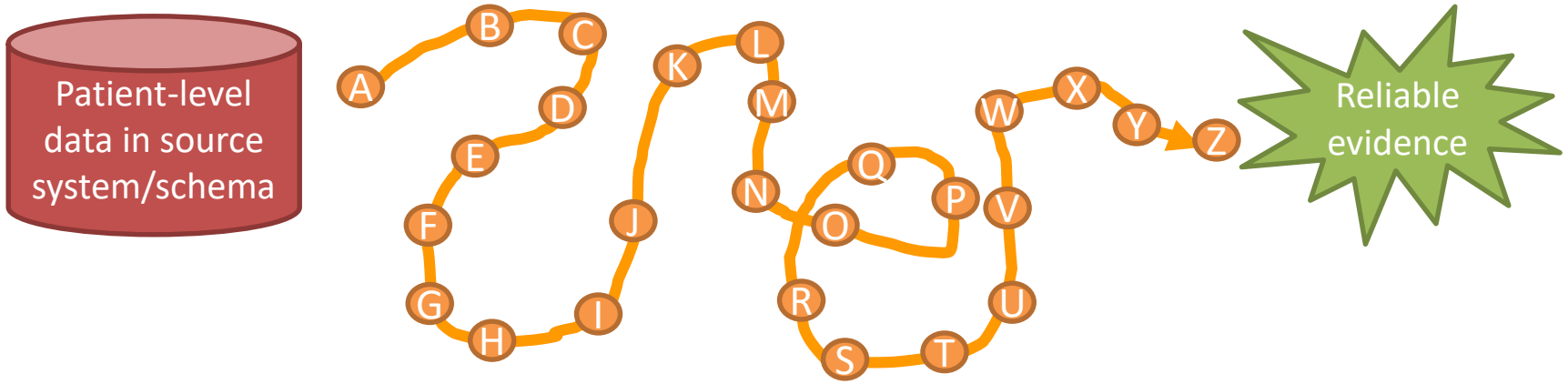
Repeated

Desired attributes for reliable evidence

Desired attribute	Question	Researcher	Data	Analysis		Result
Repeatable	Identical	Identical	Identical	Identical	=	Identical
Reproducible	Identical	Different	Identical	Identical	=	Identical
Replicable	Identical	Same or different	Similar	Identical	=	Similar
Generalizable	Identical	Same or different	Different	Identical	=	Similar
Robust	Identical	Same or different	Same or different	Different	=	Similar
Calibrated	Similar (controls)	Identical	Identical	Identical	=	Statistically consistent

Minimum requirements to achieve reproducibility

Desired attribute	Question	Researcher	Data	Analysis		Result
Reproducible	Identical	Different	Identical	Identical	=	Identical

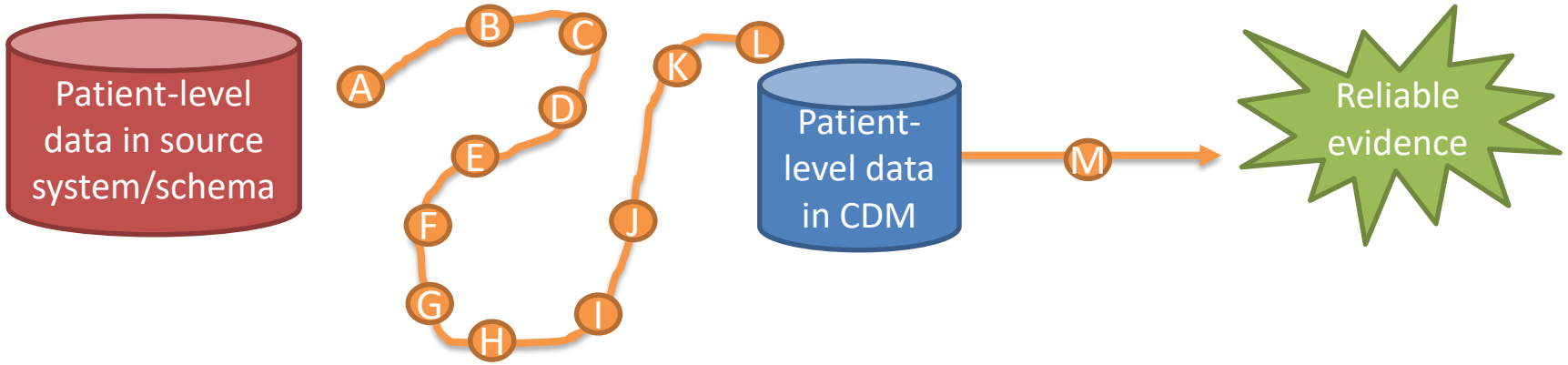


- Complete documented specification that fully describes all data manipulations and statistical procedures
- Original source data, no staged intermediaries
- Full analysis code that executes end-to-end (from source to results) without manual intervention

One-time | Repeated

How a common data model + common analytics can support reproducibility

Desired attribute	Question	Researcher	Data	Analysis		Result
Reproducible	Identical	Different	Identical	Identical	=	Identical



- Use of common data model splits the journey into two segments: 1) data standardization, 2) analysis execution
- ETL specification and source code can be developed and evaluated separately from analysis design
- CDM creates opportunity for re-use of data step and analysis step

One-time | Repeated

ETL: Real world scenario

PharMetrics Plus

CLAIMS

pat_id	claimno	from_dt	to_dt	diagprc_ind	Diag_admit	diag1
05917921689	IPA333393946	1/5/2006	1/5/2006	1	41071	41071

LRx/Dx

MEDICAL_CLAIMS

md_clm_id	ims_pat_nbr	dt_of_service	rxer_id	diag_cd
95963982102	80445908	8/1/2012 0:00	680488	41071

German DA

Problem Events

db_country	international_ practice_num	international_ doctor_num	international_ patient_num	age_at_event
GE	GE6326	GE8784	GE46478747	20

Diagnosis

db_country	international_dia gnosis_num	diagnosis_num	icd10_4_code	i
GE	GE2397573	2397573	I21.4	N (NS

Ambulatory EMR

Problem

Patient_id_synth	Diag_dt	Icd10_cd
271138	4/11/2013	I214

4 real observational databases, all containing an inpatient admission for a patient with a diagnosis of 'acute subendocardial infarction'

- Not a single table name the same…
- Not a single variable name the same….
- Different table structures (rows vs. columns)
- Different conventions (with and without decimal points)
- Different coding schemes (ICD9 vs. ICD10)

What does it mean to ETL to OMOP CDM? Standardize **structure** and **content**

PharMetrics Plus
Inpatient Claims

pat_id	claimno	from_dt	to_dt	diagprc_ind	Diag_admit
05917921689	IPA333393946	1/5/2006	1/5/2006	1	41071

Transform structure optimized for large-scale analysis for clinical characterization, population-level estimation, and patient-level prediction

PharMetrics Plus
CONDITION_OCCURRENCE

PERSON_ID	CONDITION_ START_DATE	CONDITION_ SOURCE_ VALUE	CONDITION_TYPE_CONCEPT_ID
05917921689	1/5/2006	41071	Inpatient claims - primary position
05917921689	1/5/2006	41071	Inpatient claims - 1st position

Maintain provenance by preserving source values and source location in standard structure

Augment content using international vocabulary standards that can be applied to any data source

PharMetrics Plus
CONDITION_OCCURRENCE

PERSON_ID	CONDITION_ START_DATE	CONDITION_ SOURCE_ VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
05917921689	1/5/2006	41071	Inpatient claims - primary position	44825429	444406

OMOP CDM = Standardized structure: same tables, same fields, same datatypes, same conventions across disparate sources

PharMetrics Plus

CLAIMS

pat_id	claimno	from_dt	to_dt	diagprc_ind	Diag_admit	diag1
05917921689	IPA333393946	1/5/2006	1/5/2006	1	41071	41071

LRx/Dx

MEDICAL_CLAIMS

md_clm_id	ims_pat_nbr	dt_of_service	rxer_id	diag_cd
95963982102	80445908	8/1/2012 0:00	680488	41071

German DA

Problem Events

db_country	international_practice_num	international_doctor_num	international_patient_num	age_at_event	date_of_event	international_diagnosis_num
GE	GE6326	GE8784	GE46478747	20	11/19/2014 0:00	GE2397573

Diagnosis

db_country	international_diagnosis_num	diagnosis_num	icd10_4_code	icd10_3_text	diagnosis_confidence
GE	GE2397573	2397573	I21.4	Non-ST elevation (NSTEMI) myocardial infarction	Confirmed

Ambulatory EMR

Problem

Patient_id_synth	Diag_dt	Icd10_cd
271138	4/11/2013	I214

PharMetrics Plus: CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
157033702	1/5/2006	41071	Inpatient claims - primary position
157033702	1/5/2006	41071	Inpatient claims - 1st position

LRX/DX: CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
80445908	8/1/2012	41071	Primary Condition

German DA : CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
46478747	11/19/2014	I21.4	EHR problem list entry

Ambulatory EMR : CONDITION_OCCURRENCE

PERSON_ID	CONDITION_START_DATE	CONDITION_SOURCE_VALUE	CONDITION_TYPE_CONCEPT_ID
271138	4/11/2013	I214	Primary Condition

- Consistent structure optimized for large-scale analysis
- Structure preserves all source content and provenance

OMOP CDM = Standardized content: common vocabularies across disparate sources

PharMetrics Plus: CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
05917921689	1/5/2006	41071	Inpatient claims - primary position	4482542	444406

LRx/Dx: CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
80445908	8/1/2012	41071	Primary Condition	4482542	444406

German DA : CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
6478747	11/19/2014	I21.4	EHR problem list entry	4557208	444406

Ambulatory EMR : CONDITION_OCCURRENCE

PERSON_ID	CONDITION _START _DATE	CONDITION _SOURCE _VALUE	CONDITION _TYPE _CONCEPT_ID	CONDITION _SOURCE _CONCEPT_ID	CONDITION _CONCEPT_ID
271138	4/11/2013	I214	Primary Condition	4557208	444406

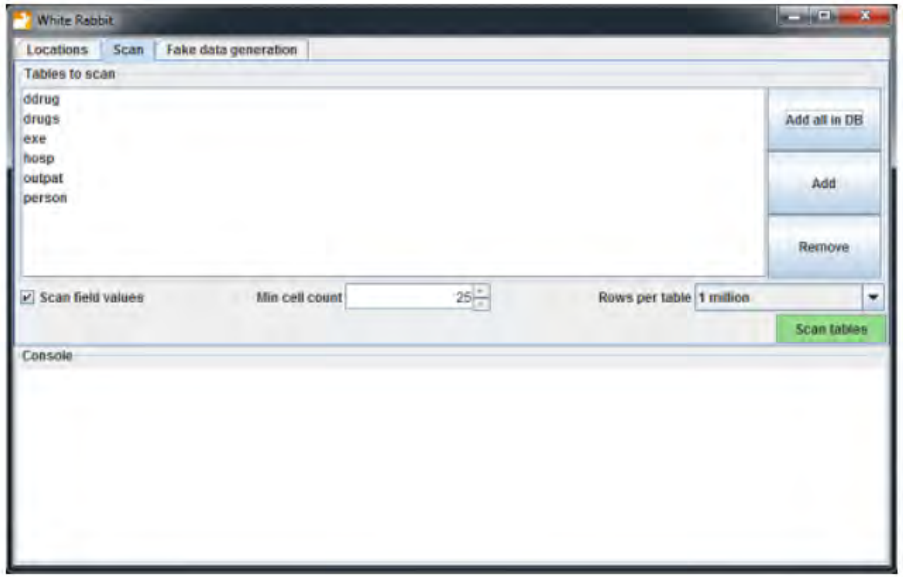
- Standardize across vocabularies to a common referent standard (ICD9/10→SNOMED)
- Source codes mapped into each domain standard so that now you can talk across different languages

- Standardize source codes to be uniquely defined across all vocabularies
- No more worries about formatting or code overlap

ETL best practices

- Create ETL specification design document to promote transparency
- Share ETL source code to enable reproducibility
- ETL unit testing to improve concordance between specification and implementation
- Enable data quality exploration at all stages of analysis lifecycle using standardized data characterization tools

Create ETL specification design document to promote transparency



White Rabbit

Rabbit in a Hat

https://github.com/OHDSI/WhiteRabbit

Share ETL source code to enable reproducibility

OHDSI / ETL-CDMBuilder

Unwatch 58 | Star 17 | Fork 17

Code | Issues 9 | Pull requests 1 | Projects 0 | Wiki | Insights | Settings

Branch: master | ETL-CDMBuilder / man / | Create new file | Upload files | Find file | History

ericaVoss Updating PREMIER documentation — Latest commit 06c367a on Aug 21, 2018

..		
CERNER	Adding Cerner Documentation and Updated CDM_BUILDER code.	10 months ago
CPRD	New CPRD test cases	6 months ago
HCUP	Loading CCAE/MDCR material	11 months ago
JMDC	Updating JMDC test cases and documentation	5 months ago
OPTUM_EXTENDED	Delete .RData	6 months ago
OPTUM_INTEGRATED	Optum Test Update	a year ago
OPTUM_ONCOLOGY	Optum - adding logic to handle multiple providers for an encounter/visit	a year ago
OPTUM_PANTHER	Optum Panther CDM v5.3.1 updates	6 months ago
PREMIER	Updating PREMIER documentation	5 months ago
SEER	Updated SEER document for CDM v5.2	a year ago
TRUVEN_CCAE_MDCR	Moving files around to preserve URLs already mentioned in publications.	6 months ago

https://github.com/OHDSI/ETL-CDMBuilder

ETL unit testing to improve concordance between specification and implementation

Observational Health Data Sciences and Informatics

Search

Recent Changes Media Manager Sitemap

Trace: • test_framework

Documentation

Getting Started with OHDSI

Common Data Model (CDM)

- CDM Specifications
- CDM Vocabulary

Convert Database to CDM (ETL)

- ETL creation best practices
- Example ETLs
- ETL Tools
- ETL Support

Tool Specific Documentation

- ATLAS
- ACHILLES
- White Rabbit
- Usagi
- Methods Library
- WebAPI
- Common Evidence Model

documentation:software:whiterabbit:test_framework

Rabbit-In-a-Hat testing framework

Rabbit-In-a-Hat can generate a framework for creating a set of unit tests. The framework consists of a set of R functions tailored to the source and target schema in your ETL. These functions can then be used to define the unit tests.

Unit testing assumes that you have your data in source format somewhere in a database. You should already have created an ETL process that will extract from the source database, transform it into CDM format, and load it into a CDM schema. The unit test framework can be used to make sure that your ETL process is doing what it is supposed to do. For this you will need to create a new, empty database with exactly the same structure as your source database, and a new empty database where a test CDM database will live. The framework can be used to insert test data into the empty source schema. You can then run your ETL process on the test data to populate the test CDM database. you can then use the framework to verify that the output of the ETL in the test CDM database is what you'd expect given the test source data.

Table of Contents

http://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:whiterabbit:test_framework

Enable data quality exploration at all stages of analysis lifecycle using standardized data characterization tools

Achilles

Data Sources Reports

Essential hypertension

OHDSI Achilles Heel Results Viewer: SynPUF 5 percent sample

Search:

Analysis Id	Rule Id	Warning Type	Message	Record Count
4	6	WARNING	4-Number of persons by race; data with unmapped concepts	
200	6	WARNING	200-Number of persons with at least one visit occurrence, by visit_concept_id; data with unmapped concepts	
301	6	WARNING	301-Number of providers by specialty concept_id; data with unmapped concepts	
400	6	WARNING	400-Number of persons with at least one condition occurrence, by condition_concept_id; data with unmapped concepts	
402	23	WARNING	402-Number of persons by condition occurrence start month, by condition_concept_id; 504 concepts have a 100% change in monthly count of events	504
420	22	WARNING	420-Number of condition occurrence records by condition occurrence start month; theres a 100% change in monthly count of events	
600	6	WARNING	600-Number of persons with at least one procedure occurrence, by procedure_concept_id; data with unmapped concepts	
602	23	WARNING	602-Number of persons by procedure occurrence start month, by procedure_concept_id; 267 concepts have a 100% change in monthly count of events	267
620	22	WARNING	620-Number of procedure occurrence records by procedure occurrence start month; theres a 100% change in monthly count of events	
700	6	WARNING	700-Number of persons with at least one drug exposure, by drug_concept_id; data with unmapped concepts	
702	23	WARNING	702-Number of persons by drug exposure start month, by drug_concept_id; 323 concepts have a 100% change in monthly count of events	323
720	22	WARNING	720-Number of drug exposure records by drug exposure start month; theres a 100% change in monthly count of events	
800	6	WARNING	800-Number of persons with at least one observation occurrence, by observation_concept_id; data with unmapped concepts	
802	23	WARNING	802-Number of persons by observation occurrence start month, by observation_concept_id; 60 concepts have a 100% change in monthly count of events	60
820	22	WARNING	820-Number of observation records by observation start month; theres a 100% change in monthly count of events	
902	23	WARNING	902-Number of persons by drug era start month, by drug_concept_id; 158 concepts have a 100% change in monthly count of events	158
920	22	WARNING	920-Number of drug era records by drug era start month; theres a 100% change in monthly count of events	
1000	6	WARNING	1000-Number of persons with at least one condition era, by condition_concept_id; data with unmapped concepts	
1002	23	WARNING	1002-Number of persons by condition era start month, by condition_concept_id; 480 concepts have a 100% change in monthly count of events	480
1020	22	WARNING	1020-Number of condition era records by condition era start month; theres a 100% change in monthly count of events	
	41	NOTIFICATION	No body weight data in MEASUREMENT table (under concept_id 3025315 (LOINC code 29463-7))	
	27	NOTIFICATION	Unmapped data over percentage threshold in:Procedure	
	27	NOTIFICATION	Unmapped data over percentage threshold in:Measurement	
	27	NOTIFICATION	Unmapped data over percentage threshold in:Condition	
	42	NOTIFICATION	[GeneralPopulationOnly] Percentage of outpatient visits is below threshold	

Showing 1 to 30 of 30 entries

Download Heel Results

Associated Analysis SQL

Associated Heel SQL

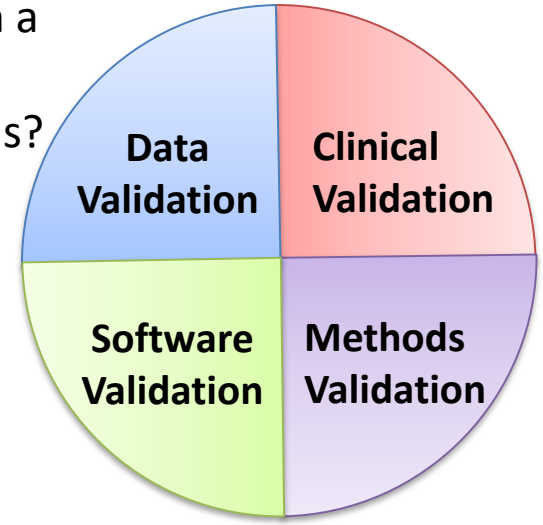
https://github.com/OHDSI/Achilles

The goal isn't "data quality", it's "evidence quality" so need to apply a more holistic approach to validation

Validation: "the action of checking or proving the accuracy of something"

Data : are the data completely captured with plausible values in a manner that is conformant to agreed structure and conventions?

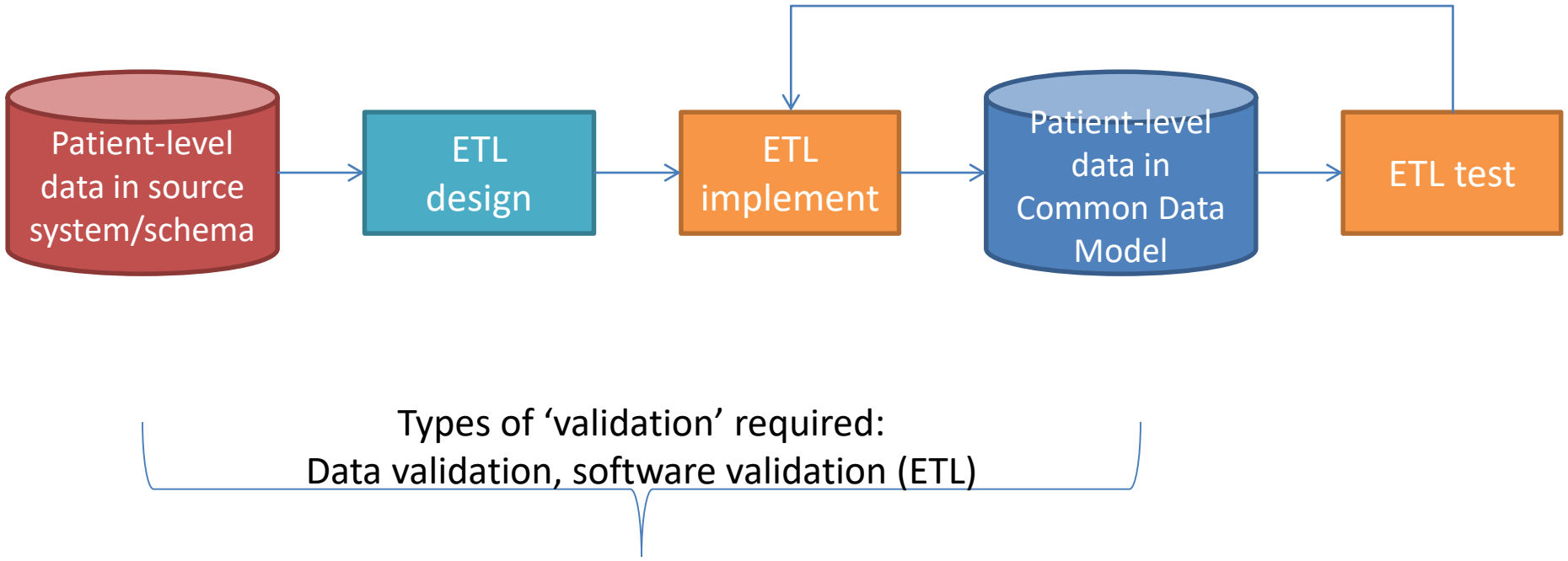
Clinical: to what extent does the analysis conducted match the clinical intention?



Software : does the software do what it is expected to do?

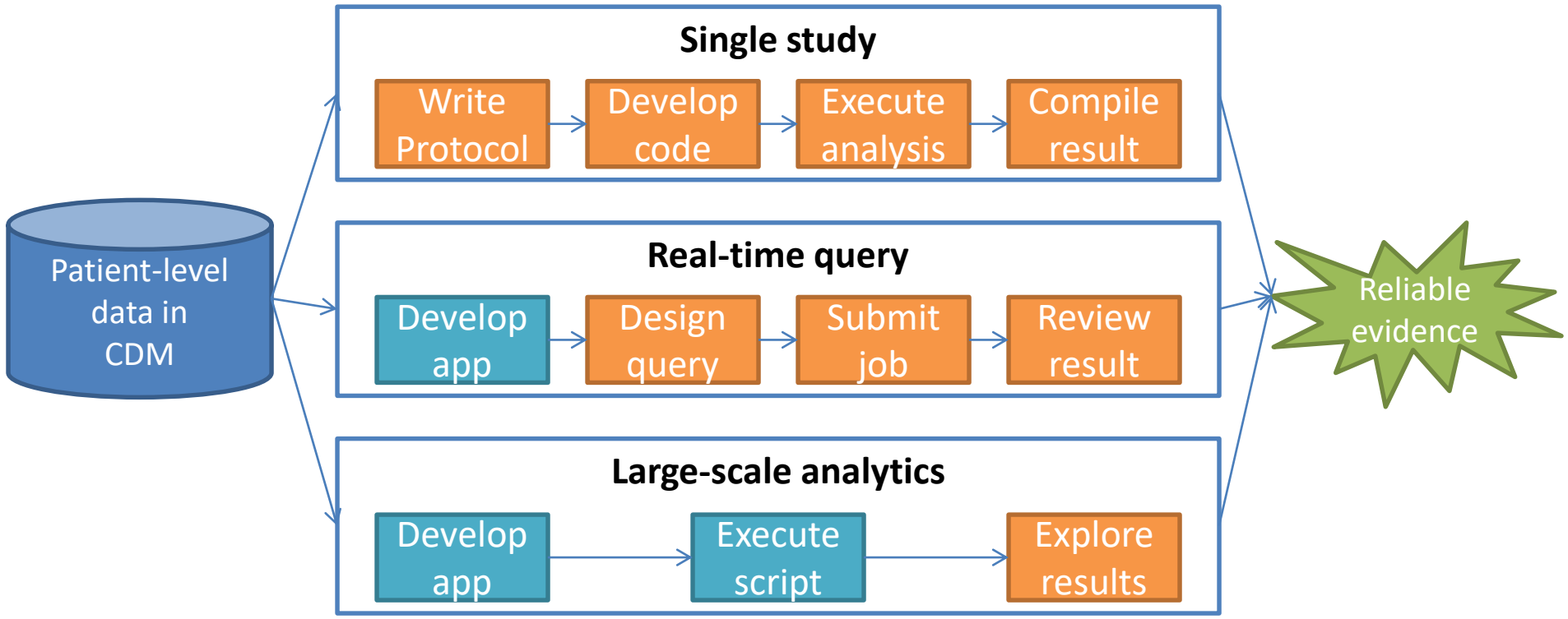
Statistical : do the estimates generated in an analysis measure what they purport to?

Structuring the journey from source to a common data model



One-time | Repeated

Structuring the journey from a common data model to evidence



Types of 'validation' required:
Software validation (analytics), Clinical validation, Statistical validation

One-time | Repeated

Session I: Transforming Raw Data into Research-Ready Data

Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Session I: Transforming Raw Data into Research-Ready Data

Jeffrey Brown, PhD
January 22, 2019

DEPARTMENT OF POPULATION MEDICINE
HARVARD MEDICAL SCHOOL
Harvard Pilgrim Health Care Institute

Data networks have different goals and needs

- Provide information about individuals, e.g., Health information exchanges
  - Exchange patient data for patient care at the point of care
  - *Need: real-time access, patient identity, minimal need for completeness or standardization (sending notes to read)*
- Provide information about groups, e.g., Sentinel
  - Public health surveillance
  - Health services research
  - Clinical trial planning and enrollment
  - Prediction modeling
  - **Regulatory decision-making and medical product efficacy**
  - *Need: size, fitness-for-use, methodology, data stability and standardization, transparency, reproducibility*

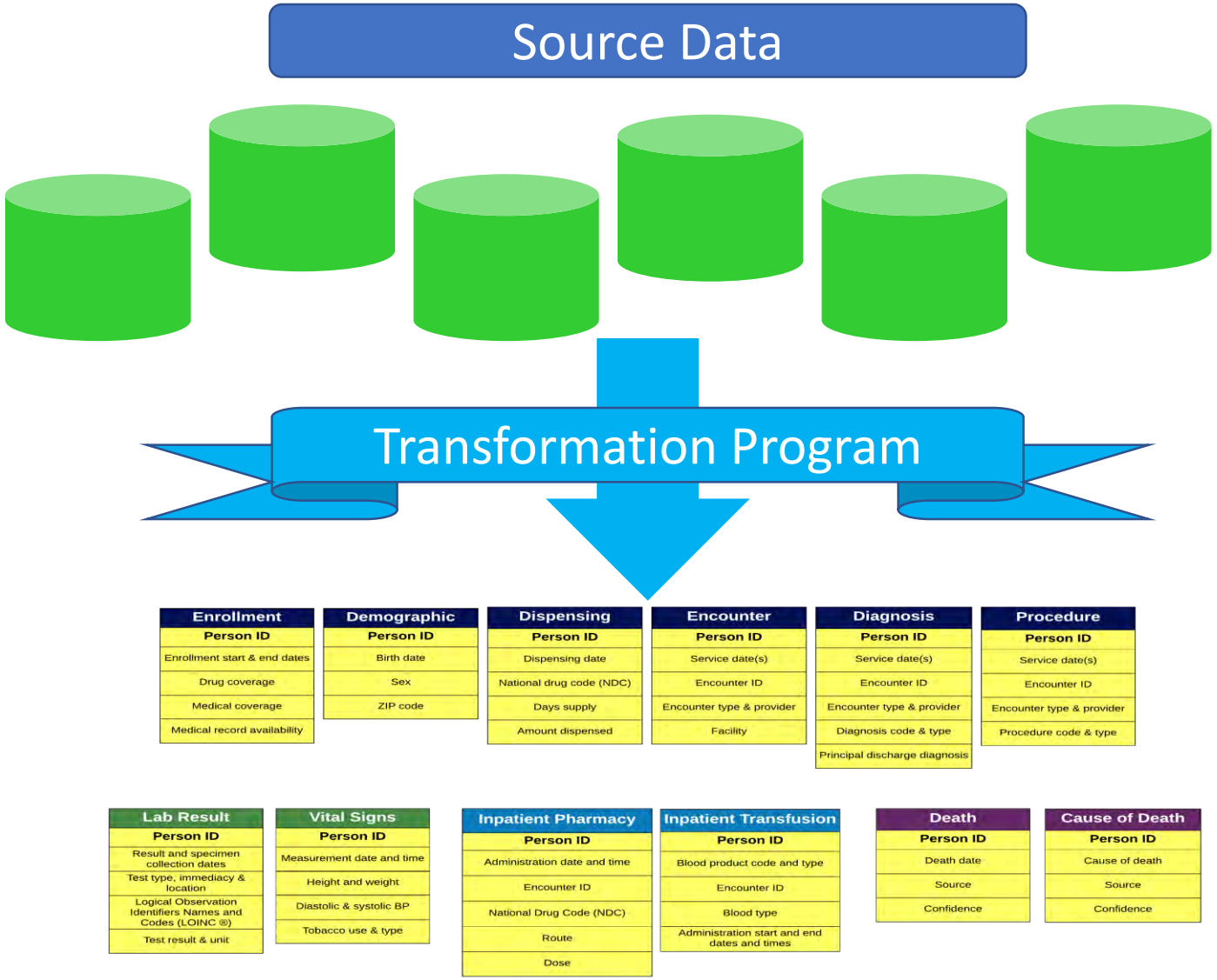
All data models have same basic concepts

- Information about people
  - Demographics (e.g., age, sex, race, ethnicity, residence)
  - Other characteristics (e.g., disease and family history)
- Information about care documented during medical encounters
  - Standard vocabularies document care during health care encounters
  - Vital signs, images, and other measurements
  - Notes
- Patient reported information
  - Within healthcare setting
  - In community (e.g., social media, fitness trackers, geolocation)

All data models have same basic approach to standardization

Unique Data Partner's source database structure

Data Partner's data transformed into Common Data Model format (every data refresh)



Sentinel principles for data curation

- Data model should maximize user control and transparency
  - Retain original data elements and values
  - Transform values only when necessary, e.g., sex, care setting
- Create phenotypes and derived variables as part of analysis – analytic code documents all transformations
- Quality assessment for entire data set for every refresh
- Data Partner participation is essential to assure that source data is appropriate for inclusion and use

Early binding versus late binding

- Sentinel data must be ready on demand - **early binding**
- Each data transformation is checked by operations team
  - 1,000s of checks and 50+ data refreshes a year
  - Checks for data model conformance, logic relationship, trends, outlier clinical validity
- Sentinel's early binding approach coupled with
  - Late-binding data quality review driven by the question and based on data and expert input
  - Validated analytic tools with embedded data quality output
  - Fitness-for-use is iterative process

Key questions

- Who is responsible for data curation?
- Who is responsible for assuring data fidelity between data source and data model?
- Who is responsible for determining whether a dataset is approved for use?
  - For every refresh at every Data Partner?
  - Is there a way to assure and document that the approved dataset is used for analysis?
- Do analytic tools use source data values or derived and mapped values?

Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Thank You

Session I: Transforming Raw Data into Research-Ready Data

Jeffrey Brown, PhD
January 22, 2019

DEPARTMENT OF POPULATION MEDICINE
HARVARD MEDICAL SCHOOL
Harvard Pilgrim Health Care Institute

Session I: Transforming Raw Data into Research-Ready Data

BREAK

Session II: Study Specific Data Curation to Establish a Fit-for-Purpose Dataset

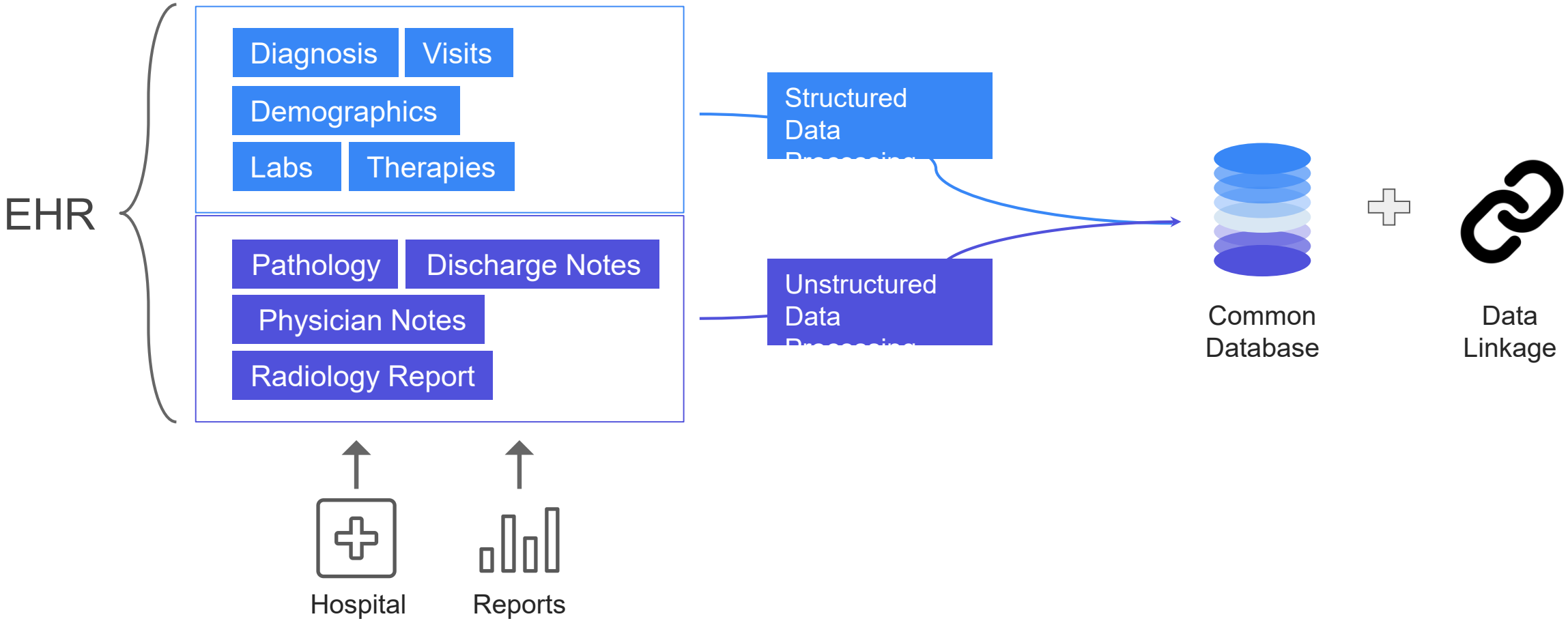
January 22, 2019

Session II: Study Specific Data Curation

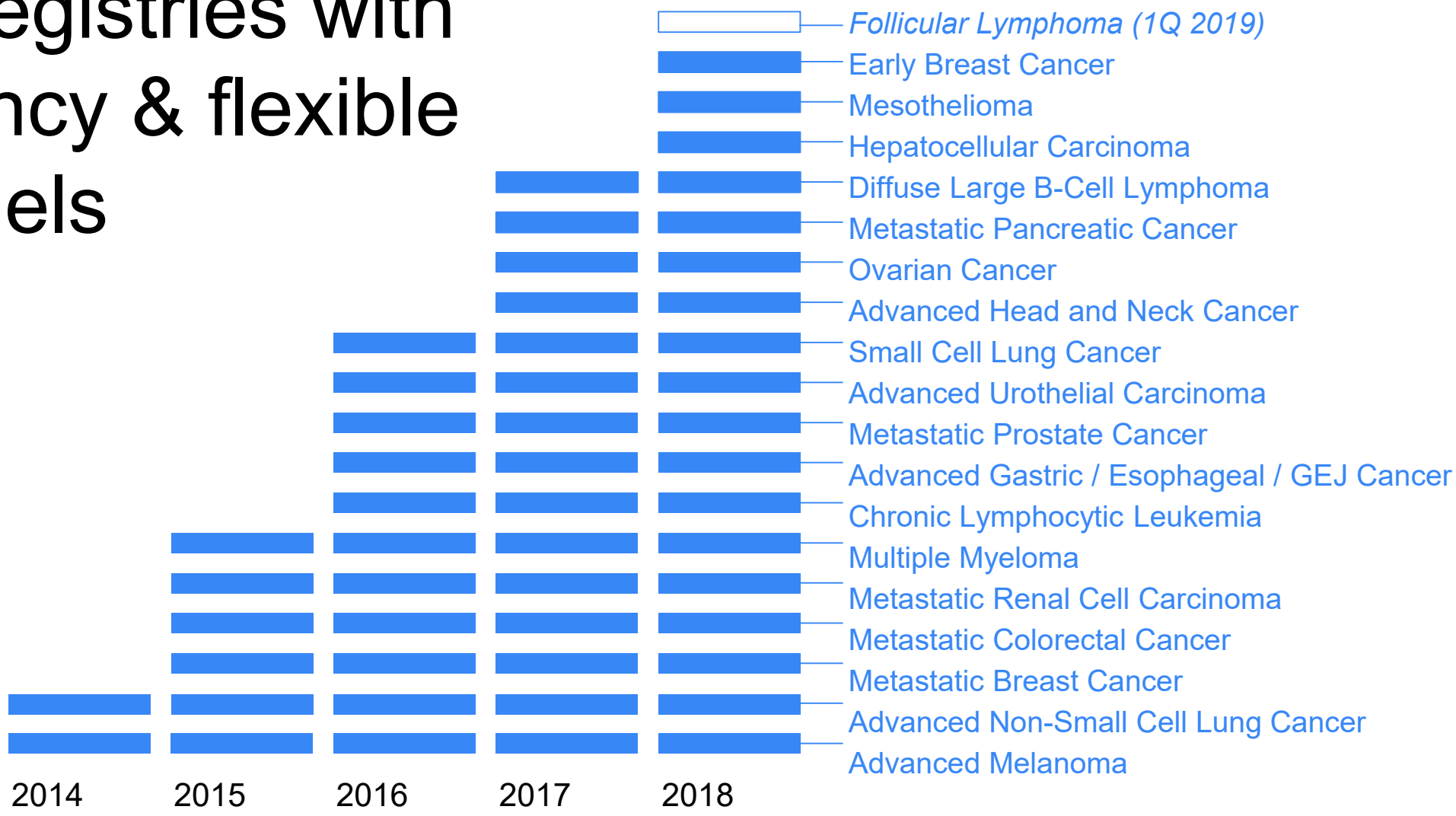
Amy Abernethy, MD, PhD
Chief Medical Officer / Chief Scientific Officer & SVP - Oncology, Flatiron Health *(a member of the Roche Group)*
Adjunct Professor of Medicine, Duke University School of Medicine
@dramyabernethy ♦♦ amy@flatiron.com

flatiron

Data source and curation



Longitudinal cancer-specific registries with 30d recency & flexible data models



flatiron

Configurable Abstraction

Tissue Collection Site

Section of PD-L1 Report

IHC Report

Lung, Right Upper Lobe Tissue

H&E

Review: Manual
Tumor Stained: 0
Intensity: 0

AssayType

Reference Range	
NEGATIVE	< 50 %
POSITIVE	>= 50 %

NEGATIVE

Result

0 50% 100%

PD-L1, 22C3

Review: Manual
Tumor Stained: 0
Intensity: 0

AssayType

Reference Range	
NEGATIVE	< 1 %
POSITIVE	>= 1 %

NEGATIVE

Result

Results: NEGATIVE, ELIGIBLE FOR OPDIVO®

0 50% 100%

PD-L1, 28-8

Comment:
All non-small cell lung cancer patients are eligible for OPDIVO® (nivolumab) regardless of their PD-L1 status.
The professional interpretation was performed at Clarient, Inc. 6455 Mission Court, West Bloomfield, MI, 48324. CLIA: 23D2013964

Lab Name

For every PD-1/PD-L1 test a patient receives, Flatiron biomarker Data Model captures:

- Test status
- Test result
- Date biopsy collected
- Date biopsy received by laboratory
- Date result received by provider
- Lab name
- Sample type
- Tissue collection site
- Type of test (e.g., FISH)
- Assay / kit (e.g., Dako 22C3)
- Percent staining & staining intensity

flatiron

Remaining study data is captured through trial-specific notes and documents in the EHR

Example: Domains in an oncology study with EHR data source

- Demographics (DM)
- Subject Visits (SV)
- Con Meds (CM)
- Exposure (EX)
- **Adverse Events (AE)**
- Disposition (DS)
- Med History (MH)
- Protocol Deviations (DV)
- I/E Criteria (IE)
- Lab Test Results (LB)
- Physical Exam (PE)
- Vital Signs (VS)
- Tumor ID (TU)
- Response (RS)
- Procedures (PR)
- Subject Elements (SE)
- Death (DD)
- Reproductive (RP)
- Healthcare Encounters (HO)

Example: Flatiron Note for Adverse Events

Adverse Event 01 Hide

Adverse Event: Adverse Event:

Description

Grade: Grade:

1 2 3 4 5

Start Date: Start Date:

DD-MMM-YYYY

Status: Status:

Active Resolved Progressed

End Date: End Date:

DD-MMM-YYYY

Cause: Cause:

Unknown Disease Treatment Treatment/disease Other

Certainty: Certainty:

Unknown Unrelated to Unlikely related to Possibly related to Probably related to Definitely related to

Study Treatment: Study Treatment:

Not changed Held Interrupted Dose reduced Withdrawn

Concomitant Medication Given: Concomitant Medication Given:

Yes No

Classification: Classification:

Adverse Event Serious Adverse Event Adverse Event of Special Interest

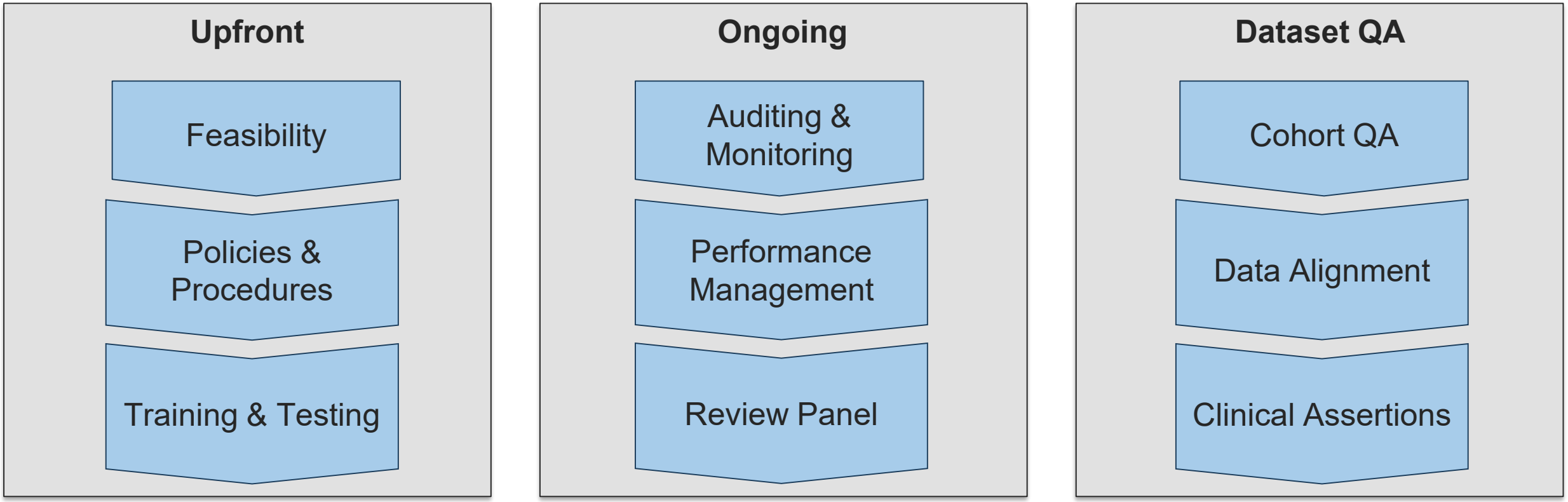
Comments: Edit Comments:

Clear

flatiron

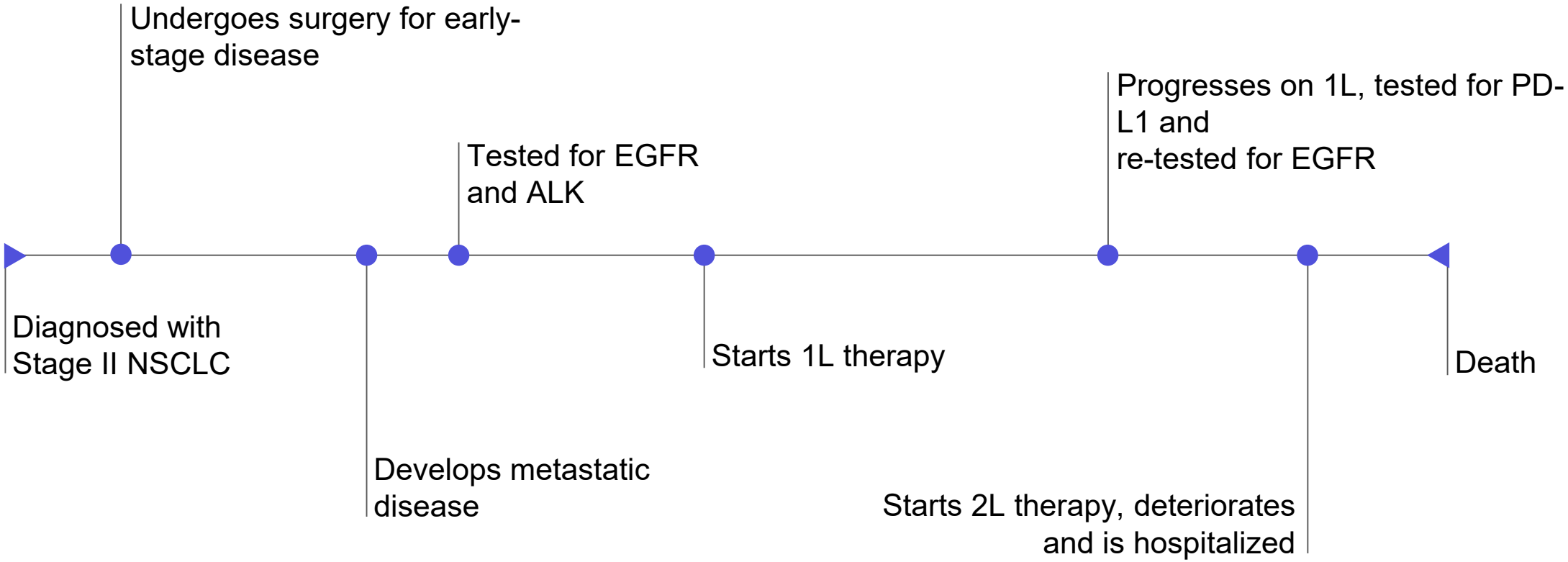
Configurable quality assurance & quality control

Centralized Controlled Environment



Asserting that this transformation is done properly

Data quality is in context



Diagnostic events are a combination of clinical, pathological, radiological, & biomarker data - *in context*

Undergoes surgery for early-stage disease

Tested for EGFR and ALK

Progresses on 1L, tested for PD-L1 and re-tested for EGFR

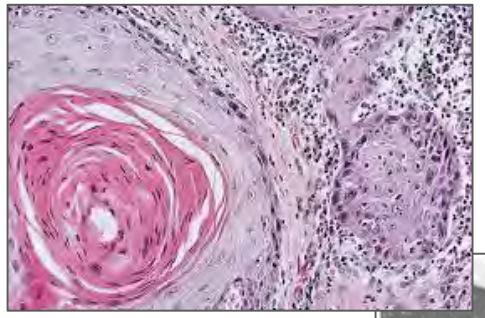
Diagnosed with Stage II NSCLC

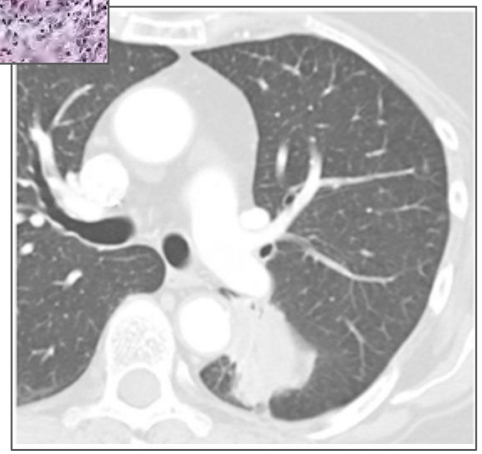
Starts 1L therapy

Death

Develops metastatic disease

Starts 2L therapy, deteriorates and is hospitalized





Gross Description

The specimen is received in formalin labeled with the patient's name. It consists of a 1 x 0.3 x 0.1 cm aggregate of pink-tan to red-pink soft tissue cores and fragments entirely submitted in one block.

Dictated by: GREGORY W SMITH PA
Entered: *02/06/12 - 1544 JAM*

Microscopic Description

The specimen consists of a well differentiated adenocarcinoma, favor lung primary. CK7 and TTF are positive. CK20 is negative. A colleague agrees with this malignant diagnosis.

Dictated by: THOMAS J GRIFONE MD
Entered: *02/07/12 - 1423 SML*

Diagnosis

SPECIMEN SUBMITTED AS TRUCUT BIOPSY LEFT LUNG NODULE:
- WELL-DIFFERENTIATED ADENOCARCINOMA, FAVOR LUNG PRIMARY.
- SEE ABOVE.

flatiron

Undergoes surgery for early-stage disease

Tested for EGFR and ALK

Progresses on 1L, tested for PD-L1 and re-tested for EGFR

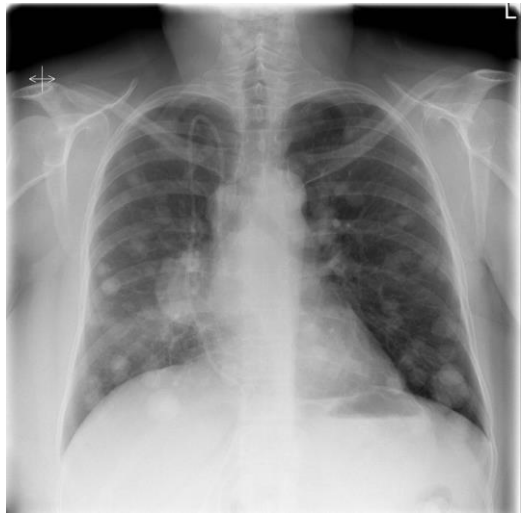
Diagnosed with Stage II NSCLC

Starts 1L therapy

Death

Develops metastatic disease

Starts 2L therapy, deteriorates and is hospitalized



Path?

Undergoes surgery for early-stage disease

Tested for EGFR and ALK

Progresses on 1L, tested for PD-L1 and re-tested for EGFR

Diagnosed with Stage II NSCLC

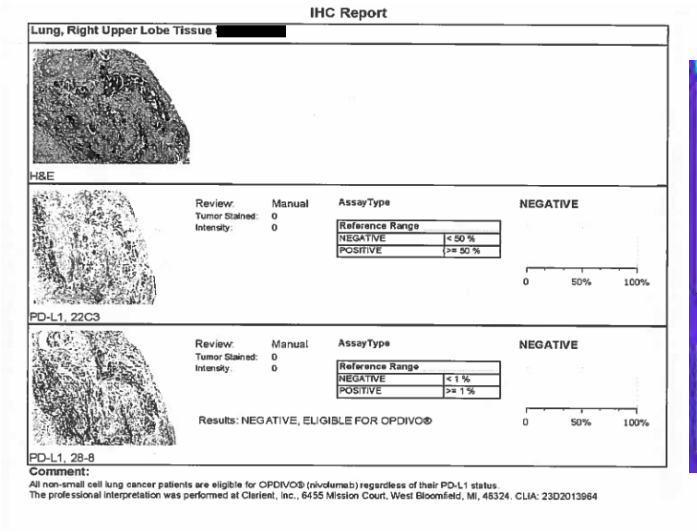
Starts 1L therapy

Death

Develops metastatic disease

Starts 2L therapy, deteriorates and is hospitalized

Time to progression is dependent on when patient is evaluated



IHC Report

Lung, Right Upper Lobe Tissue [redacted]

H&E

Review:	Manual	AssayType	NEGATIVE
Tumor Stained:	0		
Intensity:	0		

Reference Range	
NEGATIVE	< 50 %
POSITIVE	>= 50 %

0 50% 100%

PD-L1, 22C3

Review:	Manual	AssayType	NEGATIVE
Tumor Stained:	0		
Intensity:	0		

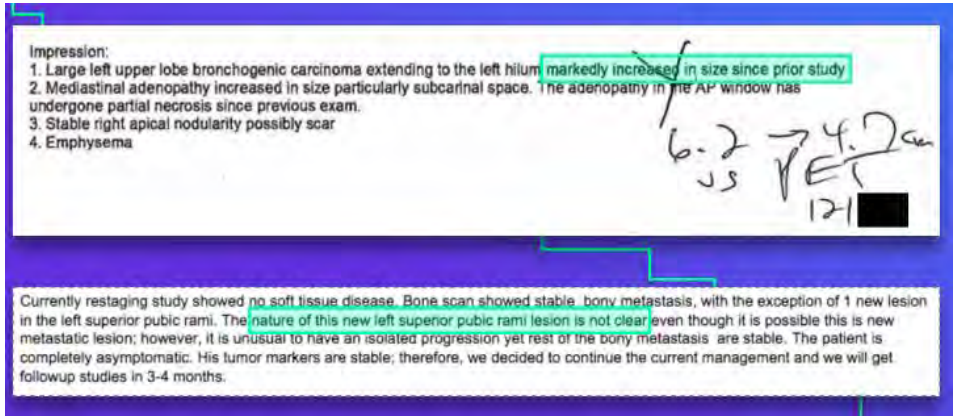
Reference Range	
NEGATIVE	< 1 %
POSITIVE	>= 1 %

Results: NEGATIVE, ELIGIBLE FOR OPDIVO®

0 50% 100%

PD-L1, 28-8

Comment:
All non-small cell lung cancer patients are eligible for OPDIVO® (nivolumab) regardless of their PD-L1 status.
The professional interpretation was performed at Clarient, Inc., 6455 Mission Court, West Bloomfield, MI, 48324. CLIA: 23D2013964



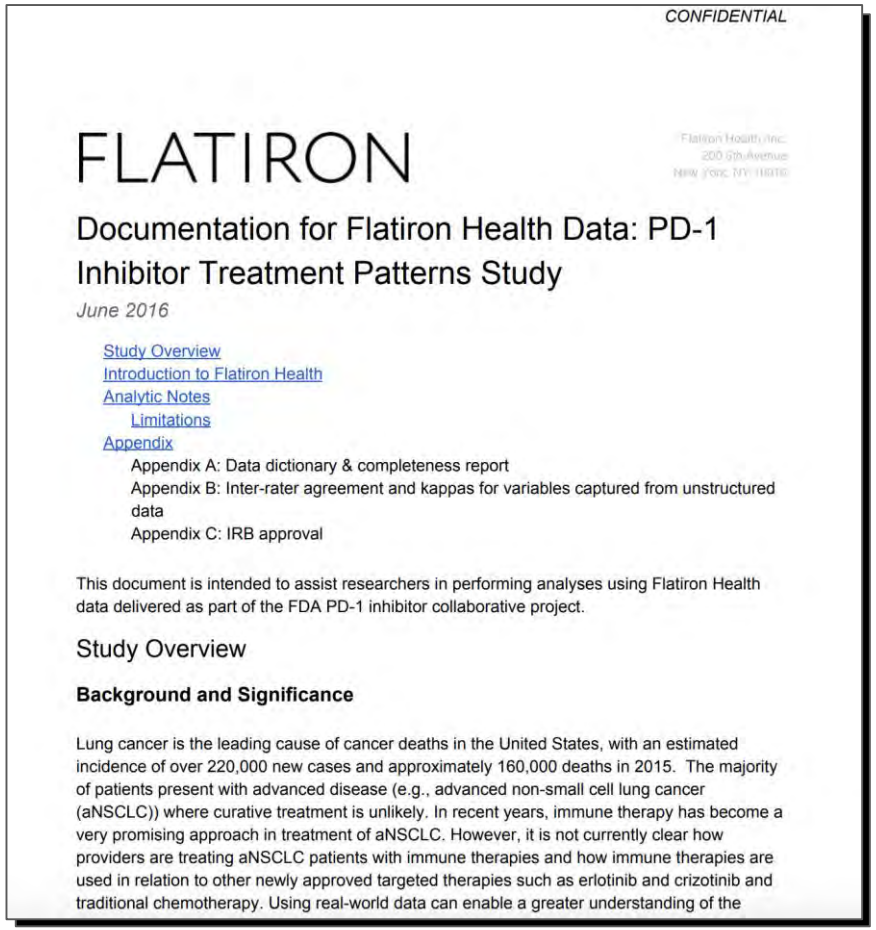
Impression:
1. Large left upper lobe bronchogenic carcinoma extending to the left hilum markedly increased in size since prior study
2. Mediastinal adenopathy increased in size particularly subcarinal space. The adenopathy in the AP window has undergone partial necrosis since previous exam.
3. Stable right apical nodularity possibly scar
4. Emphysema

6.2 → 4.7 cm
vs PET
121 [redacted]

Currently restaging study showed no soft tissue disease. Bone scan showed stable bony metastasis, with the exception of 1 new lesion in the left superior pubic rami. The nature of this new left superior pubic rami lesion is not clear even though it is possible this is new metastatic lesion; however, it is unusual to have an isolated progression yet rest of the bony metastasis are stable. The patient is completely asymptomatic. His tumor markers are stable; therefore, we decided to continue the current management and we will get followup studies in 3-4 months.

flatiron

Analytic guidance provided with data deliverables - e.g., sensitivity analysis, clinical verification



CONFIDENTIAL

FLATIRON

Flatiron Health Inc.
200 5th Avenue
New York, NY 10010

Documentation for Flatiron Health Data: PD-1 Inhibitor Treatment Patterns Study

June 2016

This document is intended to assist researchers in performing analyses using Flatiron Health data delivered as part of the FDA PD-1 inhibitor collaborative project.

Study Overview

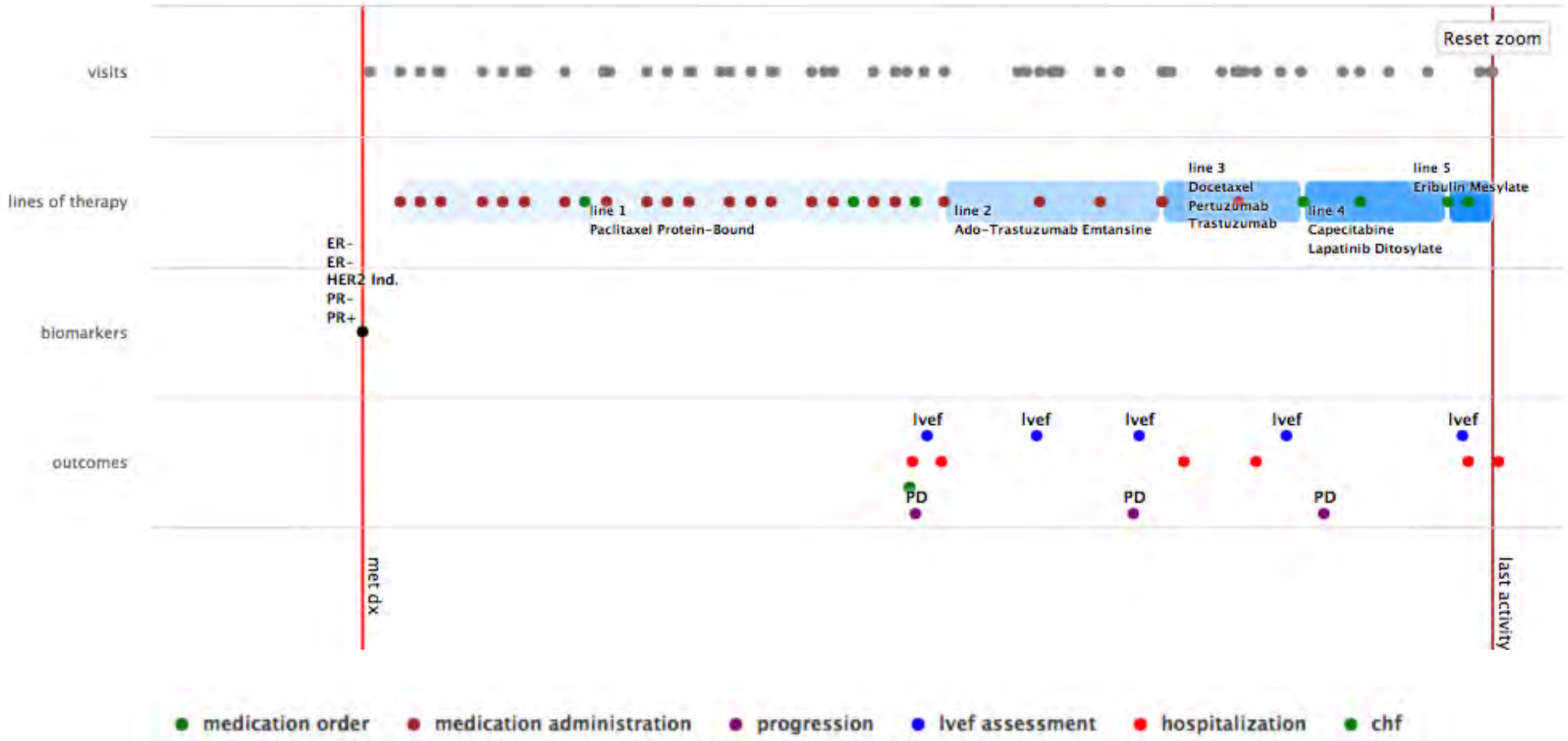
Background and Significance

Lung cancer is the leading cause of cancer deaths in the United States, with an estimated incidence of over 220,000 new cases and approximately 160,000 deaths in 2015. The majority of patients present with advanced disease (e.g., advanced non-small cell lung cancer (aNSCLC)) where curative treatment is unlikely. In recent years, immune therapy has become a very promising approach in treatment of aNSCLC. However, it is not currently clear how providers are treating aNSCLC patients with immune therapies and how immune therapies are used in relation to other newly approved targeted therapies such as erlotinib and crizotinib and traditional chemotherapy. Using real-world data can enable a greater understanding of the

- Deliver comprehensive analytic guide including:
  - Study Overview
  - Research Questions
  - Inclusion/Exclusion Criteria
  - Data Elements
  - Baseline Characteristics
  - Data Quality and Provenance
  - Data Freeze and Retention Process
  - Overview of Abstracted Variables Data Quality
  - Measure Inter-Rater Reliability
  - Interpreting Agreement
  - De-identification of Flatiron Data
  - Analytic Notes

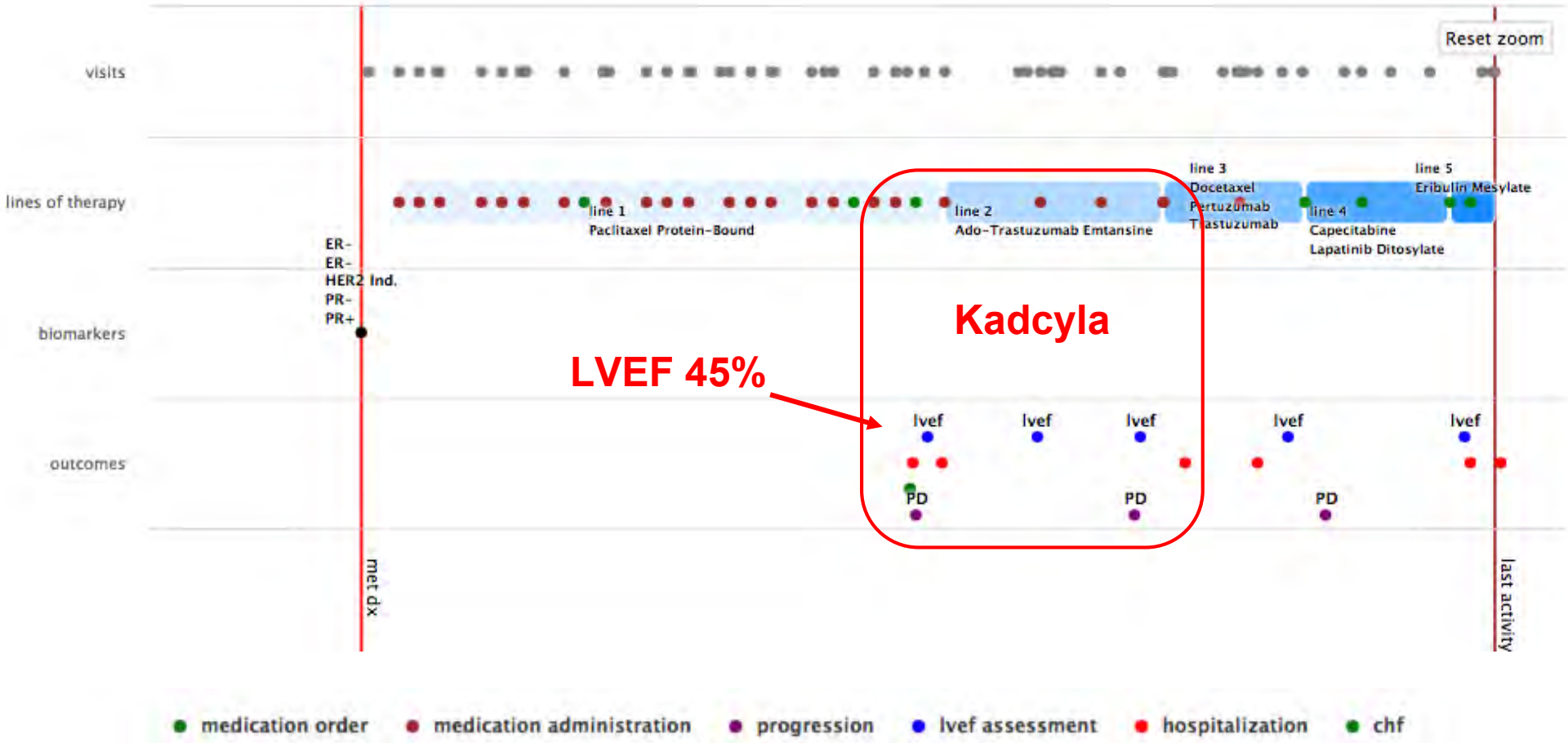
flatiron

Data Verification via Patient Journey Visualizer



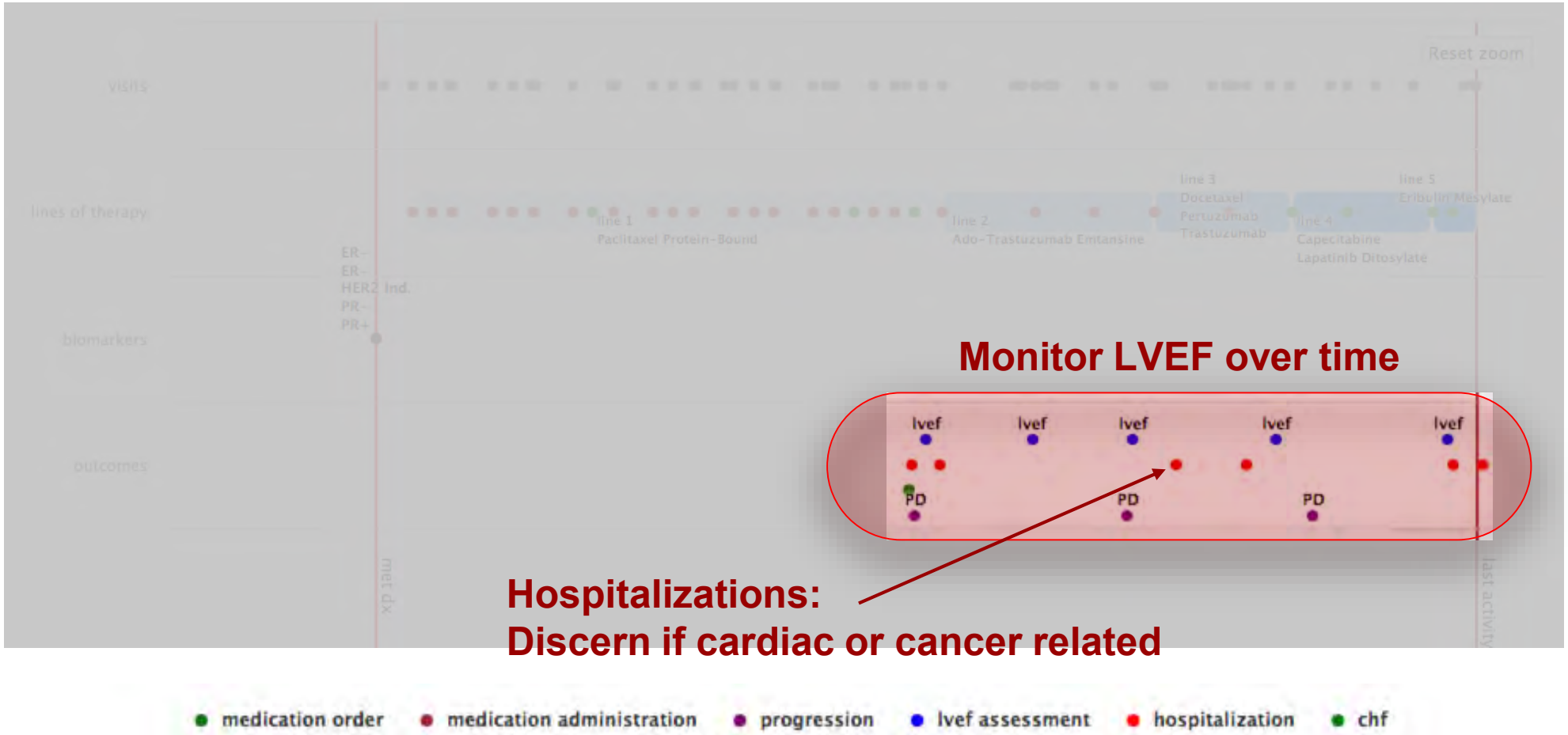
flatiron

Data Verification via Patient Journey Visualizer



flatiron

Data Verification via Patient Journey Visualizer



flatiron

Lingua Franca for Data Quality

Not all data elements are created equal

Document clinical data quality and completeness

Completeness of technology-enabled abstraction

Example: Advanced NSCLC

Variable	Structured data only	Flatiron data completeness
Metastatic diagnosis	26%	100%
Smoking status	0%[1]	94%
Histology	37%	99%[2]
Stage	61%	95%
ALK results (of those tested)	9%	100%[3]
EGFR results (of those tested)	11%	99%[3]

1 88% are free text in dedicated field in EHR (requiring hand abstraction)
2 Including 8% of patients with results pending or unsuccessful test
3 Including 6% of patients with results pending or unsuccessful test

Accuracy of technology-enabled abstraction

Example: Sites of metastases

Site of met	Inter-abstractor agreement	Kappa
Bone	97%	0.93
Brain	96%	0.91
Liver	92%	0.83
Lung	94%	0.87

flatiron

Example: Flatiron data completeness report

Name of table as it appears in the underlying data

Table:	Table Name

Summary of Table	
Rows	5264
Columns	9
Distinct patients	5264
% of broad cohort	100
Mean # of observations per patient	1

- Rows → Number of rows in the table. Some tables may contain multiple rows per patient
- Columns → Number of columns in the table, which is equivalent to the number of variables available in the table
- Distinct patients → Number of distinct patients represented in the table, which may not be equivalent to broad cohort
- % of broad cohort → % of distinct of the broad cohort represented in the table
- Mean # of observations per patient → Average number of rows per patient in the table

Summary of Table Variables				
Variable	Completeness % - row level	Completeness % - Patient level	Distribution (of non-null responses)	Methodology used to determine completeness
Variable 1	100	100	5264 distinct values	% of unique observations where PatientID is not null
Variable 2	100	98	173 distinct values	% of unique observations where PracticeID is not null
Variable 3	100	98	COMMUNITY: 100%	% of unique observations where PracticeType is not null
Variable 4	99.9	99.9	814 distinct values	% of unique observations where PrimaryPhysicianID is not null

Variable name as it appears in the underlying data

% of **patients in the table** with ***at least one non-null value*** (i.e., a field that is not empty) that is not a missingness string. **Note**: Patients with a non-null value similar to a missing value (e.g., "Group stage not known," "Other Payer - Type Unknown," etc.) are defined as ***missing*** (not complete) in this metric

% of **rows** with a non-null value (i.e., a field that is not empty) that is not a missingness string. **Note**: Rows with a non-null value similar to a missing value (e.g., "Group stage not known," "Other Payer - Type Unknown," etc.) are defined as ***missing*** (not complete) in this metric

Table Summary
Synthesis of key takeaways from the metrics shown above

Need a consistent approach to documenting quality of high risk or high value variables

Appendix B: Flatiron Health PD-L1: Inter-rater agreement and kappas on abstracted variable

Project: **FDA** **PD-1 inhibitors in aNSCLC**

Note: For questions where a high percentage of patients have a common answer (e.g., PD-L1 testing status), kappa may be significantly lower than inter-rater agreement. In these cases, it may be more accurate to use inter-rater agreement to measure reliability of the data.

Kappas scale	
Almost perfect	0.8 to 1.0
Substantial	0.6 to 0.8
Moderate	0.4 to 0.6
Fair	0.2 to 0.4
Slight	0 to 0.2

Table: Enhanced_AdvancedNSCLC

Summary of variable inter-rater agreement and kappas

Variable	Description of variable	Corresponding question(s) on abstraction form	Question type	Inter-rater agreement (exact day for dates)	Kappa (exact agreement)	Kappa (30-day window for dates)
DiagnosisDate	Date of initial diagnosis	Enter the date of initial diagnosis	date	0.795	0.794	0.902
AdvancedDiagnosisDate	Date of diagnosis of advanced disease: first recurrence or metastasis	Enter the date of the first diagnosis of metastatic or advanced NSCLC	date	0.695	0.695	0.796
MetastaticDiagnosisDate	Date of diagnosis of metastatic disease	Enter the date of initial diagnosis [for ~55% of patients in the cohort who are diagnosed metastatic]	date	0.795	0.794	0.902
		Enter the date of distant metastatic diagnosis [for ~45% of patients in the cohort who are diagnosed non-metastatic]	date	0.527	0.476	0.557
Histology	Histology	Select the histology	drop down	0.947	0.894	
GroupStage	Group stage at time of initial diagnosis	Select the group stage	drop down	0.848	0.768	
SmokingStatus	Documented history of smoking	Smoking status	drop down	0.934	0.695	
EgfrTested	Indicator of whether the tumor was tested for a EGFR mutation	Was the tumor tested for a EGFR mutation?	boolean	0.927	0.84	
AlkTested	Indicator of whether the tumor was tested for an ALK rearrangement	Was the tumor tested for an ALK rearrangement?	boolean	0.901	0.791	
PdL1Tested	Indicator of whether the tumor was tested for PD-L1 expression	Was the tumor tested for PD-L1 expression?	boolean	0.901	0.547	
KrasTested	Indicator of whether the tumor was tested for a KRAS mutation	Was the tumor tested for a KRAS mutation?	boolean	0.894	0.728	
Ros1Tested	Indicator of whether the tumor was tested for a ROS-1 rearrangement	Was the tumor tested for a ROS-1 rearrangement?	boolean	0.881	0.725	

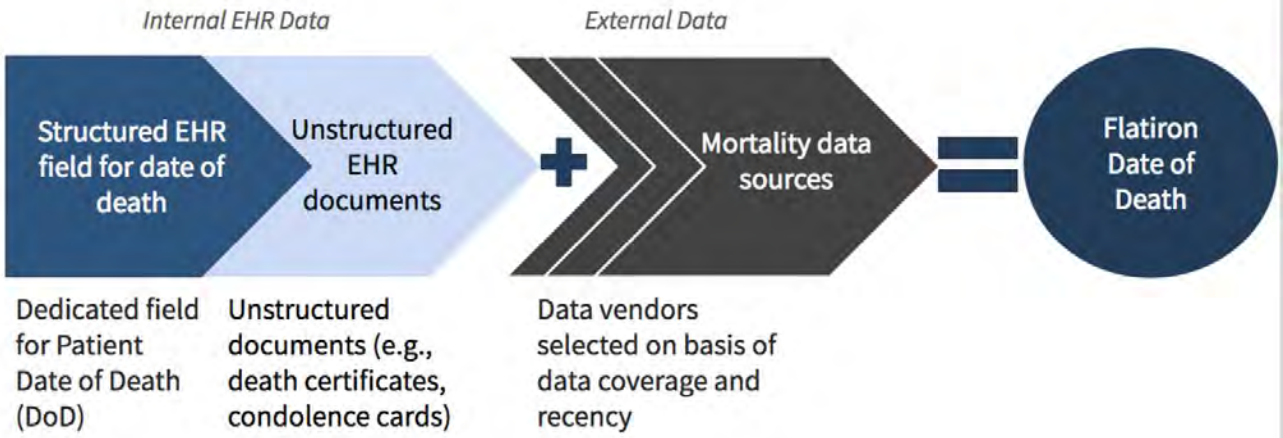
flatiron

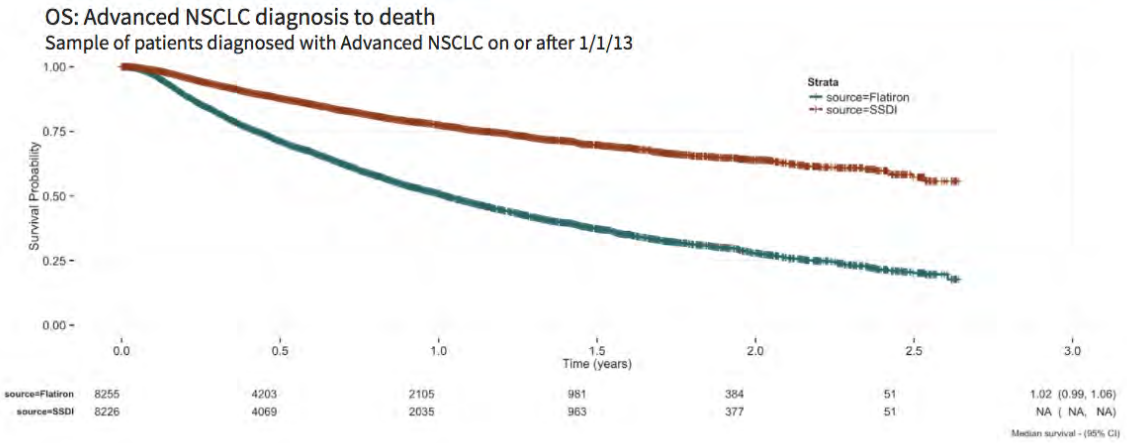
Validation of Oncology Endpoints

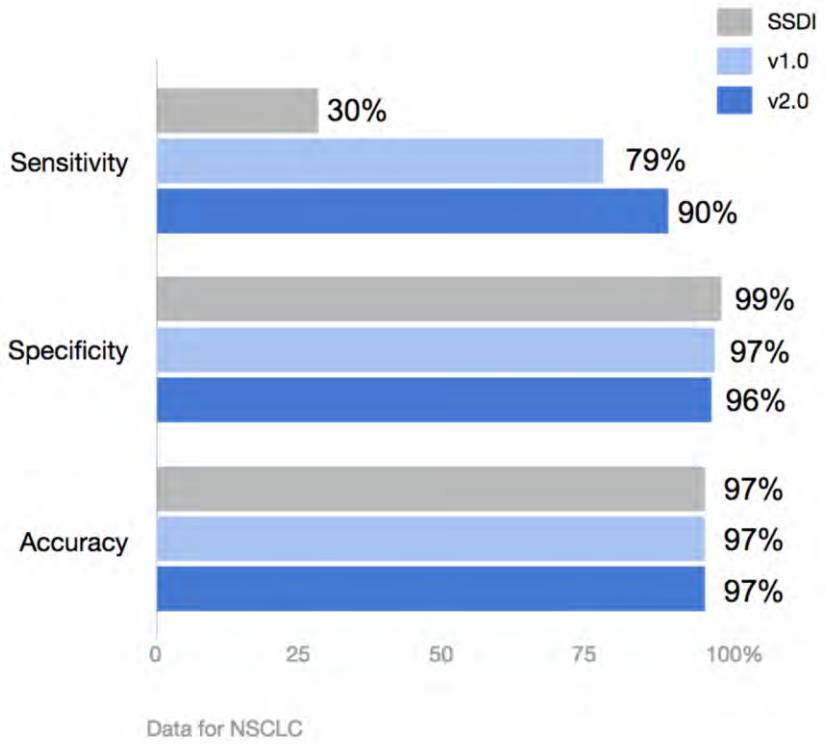
Data Quality & Validation Framework	
Face Validity	• Oncologist agreement with definition & approach
	• Regulator and other stakeholder agreement with definition & approach
Feasibility & Quality of Variables (structured & abstracted)	• Completeness of collected data
	• Inter-rater agreement on progression dates for duplicate abstracted patients
	• Qualitative feedback from abstractors reviewing the medical records
Validity of Outputs	• Likelihood of predicting a downstream event (e.g., overall survival)
	• Association between OS and PFS/TTP ○ Patient-level correlation ○ Responsiveness of endpoint to treatment effects

Evaluate data against a reference standard

E.g., gold standard = National Death Index

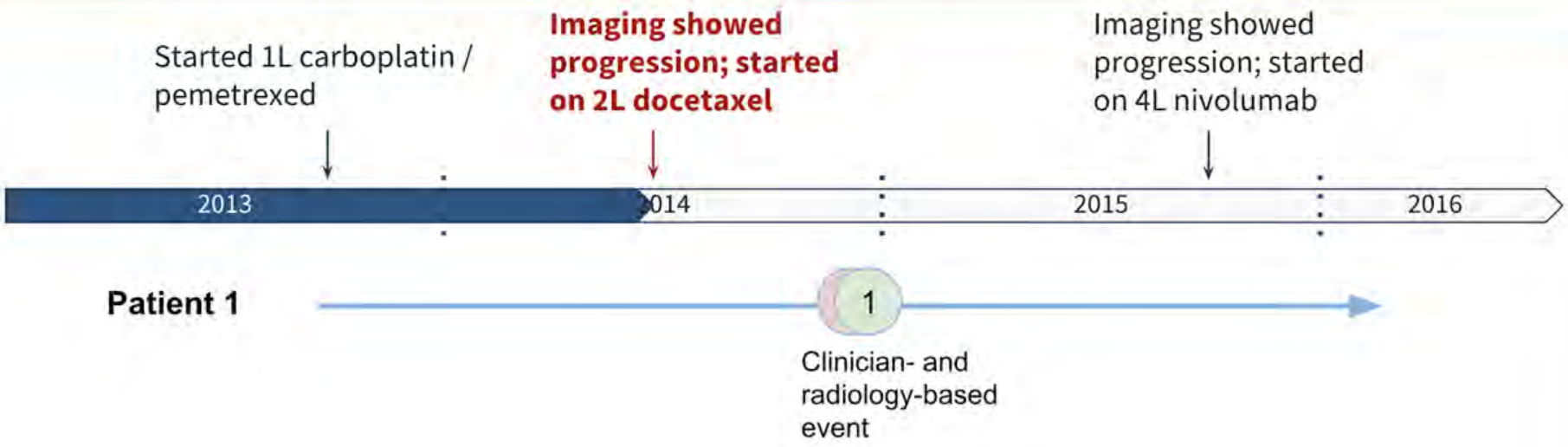






flatiron

rwP as a consensus endpoint



Maintain all underlying component information

	Event?	Date?
Clinician note	X	03/11/2015
Radiology report	X	03/05/2015
Pathology report		

Consensus:
Did a progression event occur? YES

Associated date:
Clinician-confirmed: 03/05/2015
Radiology-reported: 03/05/2015
Either: 03/05/2015

Confidence:
2 of 3 potential elements; consistent evidence of progression; source dates within a month; no pathology available (e.g., score = 7/10)

FLATIRON

Small cohorts

flatiron

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

Vemurafenib in Multiple Nonmelanoma Cancers with *BRAF* V600 Mutations

David M. Hyman, M.D., Igor Puzanov, M.D., Vivek Subbiah, M.D., Jason E. Faris, M.D., Ian Chau, M.D., Jean-Yves Blay, M.D., Ph.D., Jürgen Wolf, M.D., Ph.D., Noopur S. Raje, M.D., Eli L. Diamond, M.D., Antoine Hollebecque, M.D., Radj Gervais, M.D., Maria Elena Elez-Fernandez, M.D., Antoine Italiano, M.D., Ph.D., Ralf-Dieter Hofheinz, M.D., Manuel Hidalgo, M.D., Ph.D., Emily Chan, M.D., Ph.D., Martin Schuler, M.D., Susan Frances Lasserre, M.Sc., Martina Makrutzki, M.D., Florin Sirzen, M.D., Ph.D., Maria Luisa Veronese, M.D., Josep Tabernero, M.D., Ph.D., and José Baselga, M.D., Ph.D.

ABSTRACT

BACKGROUND

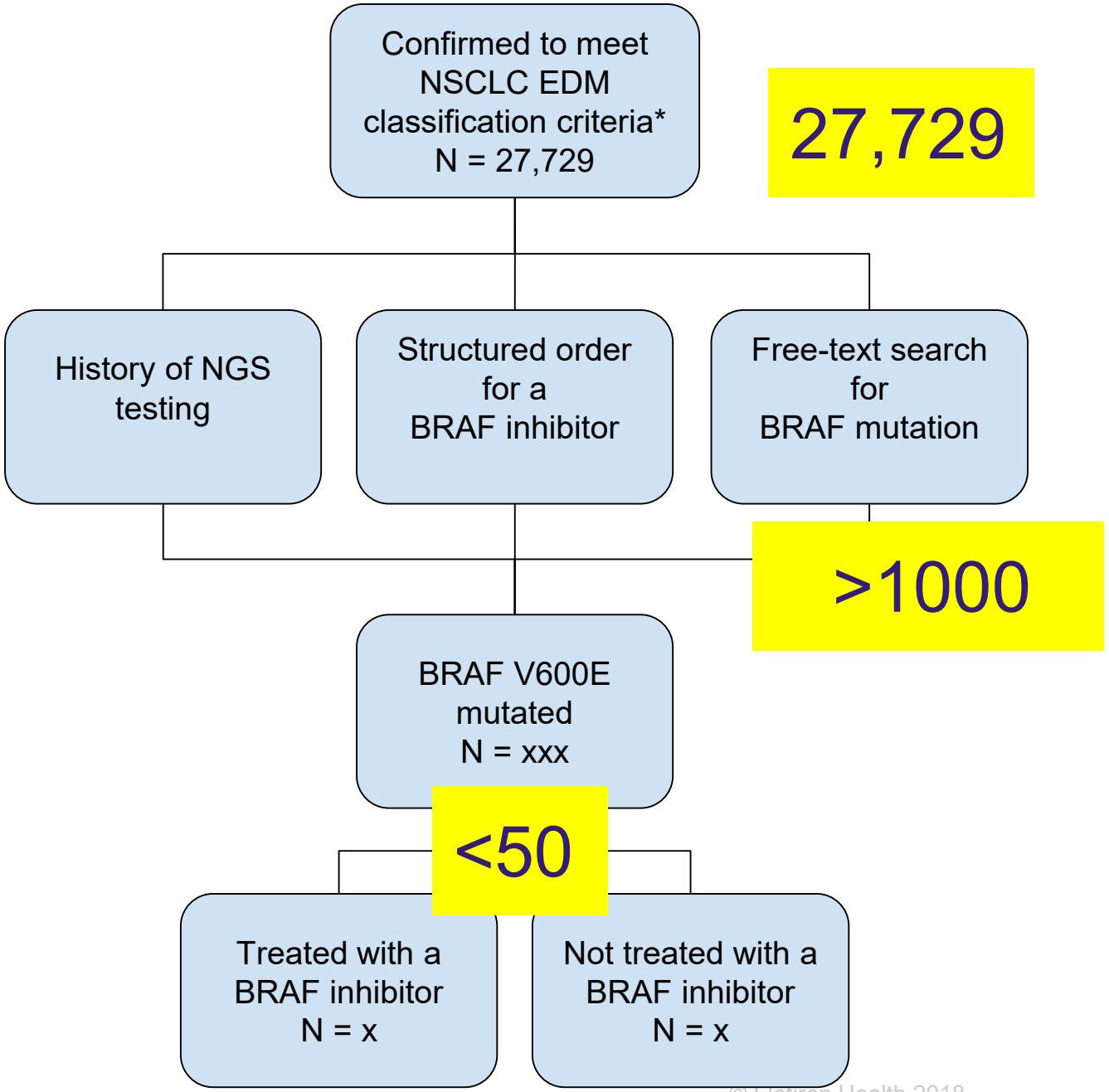
BRAF V600 mutations occur in various nonmelanoma cancers. We undertook a histology-independent phase 2 "basket" study of vemurafenib in *BRAF* V600 mutation–positive nonmelanoma cancers.

METHODS

We enrolled patients in six prespecified cancer cohorts; patients with all other tumor types were enrolled in a seventh cohort. A total of 122 patients with *BRAF* V600 mutation–positive cancer were treated, including 27 patients with colorectal cancer who received vemurafenib and cetuximab. The primary end point was the response rate; secondary end points included progression-free and overall survival.

RESULTS

In the cohort with non–small-cell lung cancer, the response rate was 42% (95% confidence interval [CI], 20 to 67) and median progression-free survival was 7.3 months (95% CI, 3.5 to 10.8). In the cohort with Erdheim–Chester disease or Langerhans'-cell histiocytosis, the response rate was 43% (95% CI, 18 to 71); the median treatment duration was 5.9 months (range, 0.6 to 18.6), and no patients



flatiron

18 months

flatiron

Source evidence: Radiology report

CT SCAN OF THE CHEST, ABDOMEN AND PELVIS WITH ORAL ANI
INTRAVENOUS CONTRAST:
History: Non-small-cell lung cancer.

IMPRESSION:
1.INCRE ASE IN SIZE AND NUMBER OF LEFT LOWER LOBE PULMONARY NODULES.
2.INCRE ASE IN MEDIASTINAL AND LEFT HILAR LYMPHADENOPATHY.
3.RESOLUTION OF A PRE VIOUS RIGHT LOWER LOBE PULMONARY NODULE.
4.NO SIGNIFI CANT CHANGE IN RENAL AND HEPATIC CYSTS.

Sincerely,

Clinician confirmation: Visit note one week later

Abd pain, nausea of unclear etiology.
Diarrhea.
Loss of appetite and weight loss.

MRI brain negative.
CT reviewed, progression of disease in the lungs, no abdominal pathology to explain nausea.

C diff (-).

Recommendation/Plan:

Discussed disease progression with and his family today.
He does not want to pursue chemotherapy.
With progressive wekaness and hemoptysis, will arrange for hospice at home.

Unstructured records contain crucial clinical context.

Session II: Study Specific Data Curation to Establish a Fit-for-Purpose Dataset

EHR-based studies and data validity

January 2019

Dan Riskin
Chief Executive Officer

Outline

In this brief talk, we will drill down into issues of data validity

- Introduction
- What is data validity?
- How is data accuracy assessed?
- Conclusion

The goal is a thoughtful discussion on data validity in EHR-based studies

VERANTOS

Outline

Who is speaking?

- Dan Riskin
  - Successful serial entrepreneur with products benefiting millions of patients
  - Adjunct Professor of Biomedical Informatics Research at Stanford
  - Testified on 21st Century Cures Initiative

- Verantos
  - Silicon Valley firm providing advanced EHR-based RWE studies
  - 3 of the top 10 biopharma firms are customers
  - Supported by NIH and NSF

The goal is a thoughtful discussion on data validity in EHR-based studies

VERANTOS


Emergency

What is data validity?

Study validity

What determines study validity?

A study is valid if the evidence is sufficient to make the clinical assertion

Validity is not a new expectation for physicians, researchers, or FDA

VERANTOS

The changing face of RWE

Product franchises are adding EHR-based studies to their RWE strategy

	Registry *(Traditional model)*	EHR *(New model)*
Benefits	Controlled data collection Tailored information	Scale and power Flexibility in subgroups
Challenges	Limited scale Limited flexibility	Data collected for clinical use Technically challenging

EHR-based studies represent the area of fastest growth in RWE

VERANTOS

Study validity

Study validity requires accuracy and generalizability

- Accuracy
  - Accuracy must be measured
  - Accuracy should be high enough to justify the clinical assertion
- Generalizability
  - The demographics and disease burden must be measured
  - These should adequately reflect characteristics of the target population
- Currently, regulators do not consistently require accuracy assessment in EHR-based studies, so this will be the focus of the talk

Data accuracy and generalizability are required if assertions are made

How is data accuracy assessed

Disruptive changes in EHR-based studies

Past EHR-based approaches do not translate to regulatory-grade studies

- Current use cases
  - Pharma uses purchased data sets for trial recruitment and marketing insight
  - Clinical assertions are not made in these uses, so accuracy is not measured
- Limitations in translating legacy data sets to regulatory-grade studies
  - Purchased EHR structured data sets have no underlying narrative or chart, so accuracy cannot be determined
  - When measured, these data sets have low cohort accuracy, with sensitivity < 50%
  - There is known bias, skewing toward higher sensitivity for sicker patients
- What is not good enough?
  - Not checking is not good enough
  - 50% accuracy is not sufficient to justify a 10% difference in study arms

The industry must move past legacy data and tech to meet requirements

The specificity fallacy

Some RWE firms report specificity but not sensitivity

- Why is specificity easier to measure than sensitivity?
  - Example: A pancreas cancer study uses 300 patients out of a 1 million patient EHR
  - The firm pulls the 300 charts from structured data and performs a chart abstraction to assess pancreas cancer false positives
  - The firm does not sample a portion of the million records to assess false negatives
  - Specificity is calculated, but sensitivityis ignored
- Why does ignoring sensitivity matter?
  - Sensitivity is where the error and bias resides
  - There is known skew in EHR accuracy… Sicker patients have more visits and are more likely to be added to the problem list
  - With a skew toward sicker patients, conclusions may be wrong or non-applicable

The industry cannot be allowed to test what's easy and ignore what's hard

Case study

How can a large biopharma firm run high quality RWE studies?

1. Firm X wanted to run a PCT and started by testing EHR cohort accuracy
    1. Requires underlying chart
    2. Requires willpower to actually check both specificity and sensitivity
2. Structured data accuracy was found to be insufficient for the assertion
    1. Structured data alone had cohort accuracy of 61.4% (F1-score, blended Sn and Sp)
    2. NLP alone brought cohort accuracy above 85%
        1. E.g. “Admitted for r/o **MI**.”
    3. NLP + additional AI brought accuracy to 95.3%
        1. E.g. “Admitted for r/o **MI**. C/o **chest pain**. EKG revealed **ST elev**. **Troponin** elevated.”
3. After enhancement, cohort accuracy met success criteria
    1. Support planned pragmatic clinical trial
    2. Will submit with a data validity report that measures accuracy for all key cohorts

Setting a high bar will keep healthcare safe and encourage innovation

VERANTOS

Looking at data accuracy

What happens when we look at cohort accuracy?

Feature	EHR structured	EHR unstructured
Hypercholesterolemia	Recall: 55.1% Precision 98.0%	Recall: 98.2% Precision 99.4%
Diabetes mellitus	Recall: 80.6% Precision 97.9%	Recall: 97.0% Precision 92.6%
Chronic kidney disease	Recall: 40.8% Precision 97.6%	Recall: 92.9% Precision 97.9%
Dementia	Recall: 62.1% Precision 100.0%	Recall: 93.1% Precision 90.0%

If the FDA says data accuracy matters, firms will measure accuracy

**This table is provided for demonstration purposes only and does not represent actual results*

VERANTOS

Conclusion

Conclusion

Advanced RWE requires advanced validity assessment

- When a clinical assertion is made, validity must be assessed
- Validity should include accuracy and generalizability
- Accuracy must include both sensitivity and specificity
- If underlying data are insufficiently valid for the assertion, the data must be demonstrably enhanced or the assertion limited
- Enhancement approaches include natural language processing, other AI-based approaches, and clinical documentation improvement

Regulators should require accuracy assessment (sensitivity and specificity) for all key cohort for all EHR-based studies

Thank You

dan.risk@verantos.com

www.verantos.com

Session II: Study Specific Data Curation to Establish a Fit-for-Purpose Dataset

Study-specific data curation in PCORnet®

Keith Marsolo, PhD

Department of Population Health Sciences, Duke University School of Medicine

Distributed Research Network Operations Center (DRN OC)

PCORnet Coordinating Center

pcornet® | The National Patient-Centered Clinical Research Network

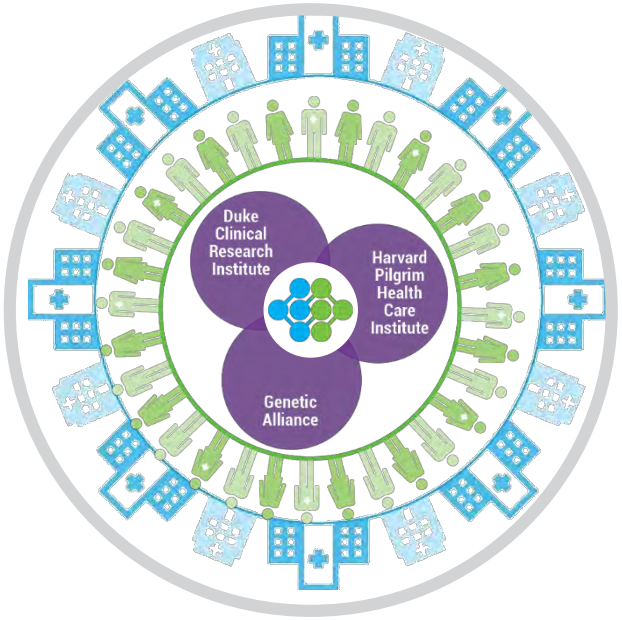
Disclosures

- Previously served as a consultant for Novartis

- *This work was supported through several Patient-Centered Outcomes Research Institute (PCORI) Program Awards (CC2-Duke-2016; ASP-1502-27079; OBS-1505-30699; OBS-1505-30683). All statements are solely those of the speaker and do not necessarily represent the views of PCORI, its Board of Governors or Methodology Committee.*

pcornet®

PCORnet® embodies a "network of networks" that harnesses the power of partnerships



9 Clinical Research Networks (**CRNs**) + 2 Health Plan Research Networks (**HPRNs**) + Patient Partners + 1 Coordinating Center = A national infrastructure for people-centered clinical research

pcornet®

PCORnet® Data Strategy

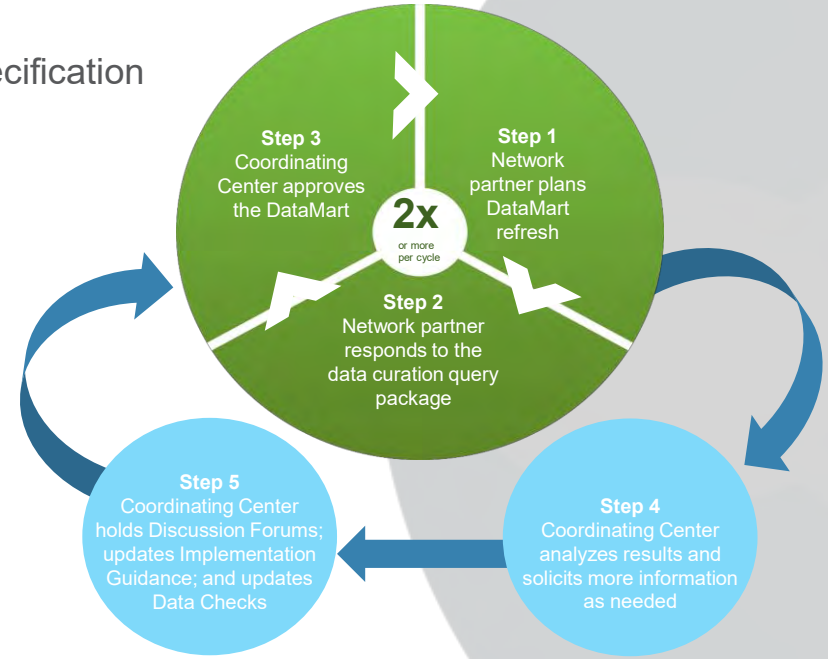
- Standardize data into a common data model
- Ensure that data support the question (data curation)
  - Foundational
  - Study-specific
- Operate a secure, distributed query infrastructure
  - Develop re-usable tools to query the data
  - Send questions to the data and only return required information
- Learn by doing and repeat

pcornet®

Assessing foundational data quality – Data Curation

- Purpose
  - Evaluate data quality and fitness-for-use across a broad research portfolio
  - Generate meaningful, actionable information for network partners, investigators and other stakeholders

- Resources
  - Implementation Guidance to accompany CDM specification
  - ETL Annotated Data Dictionary
  - Data quality checks
    - Conformance
    - Completeness
    - Plausibility
    - Persistence
  - Data curation query packages
  - Analyses and reports
  - Discussion Forums



pcornet®

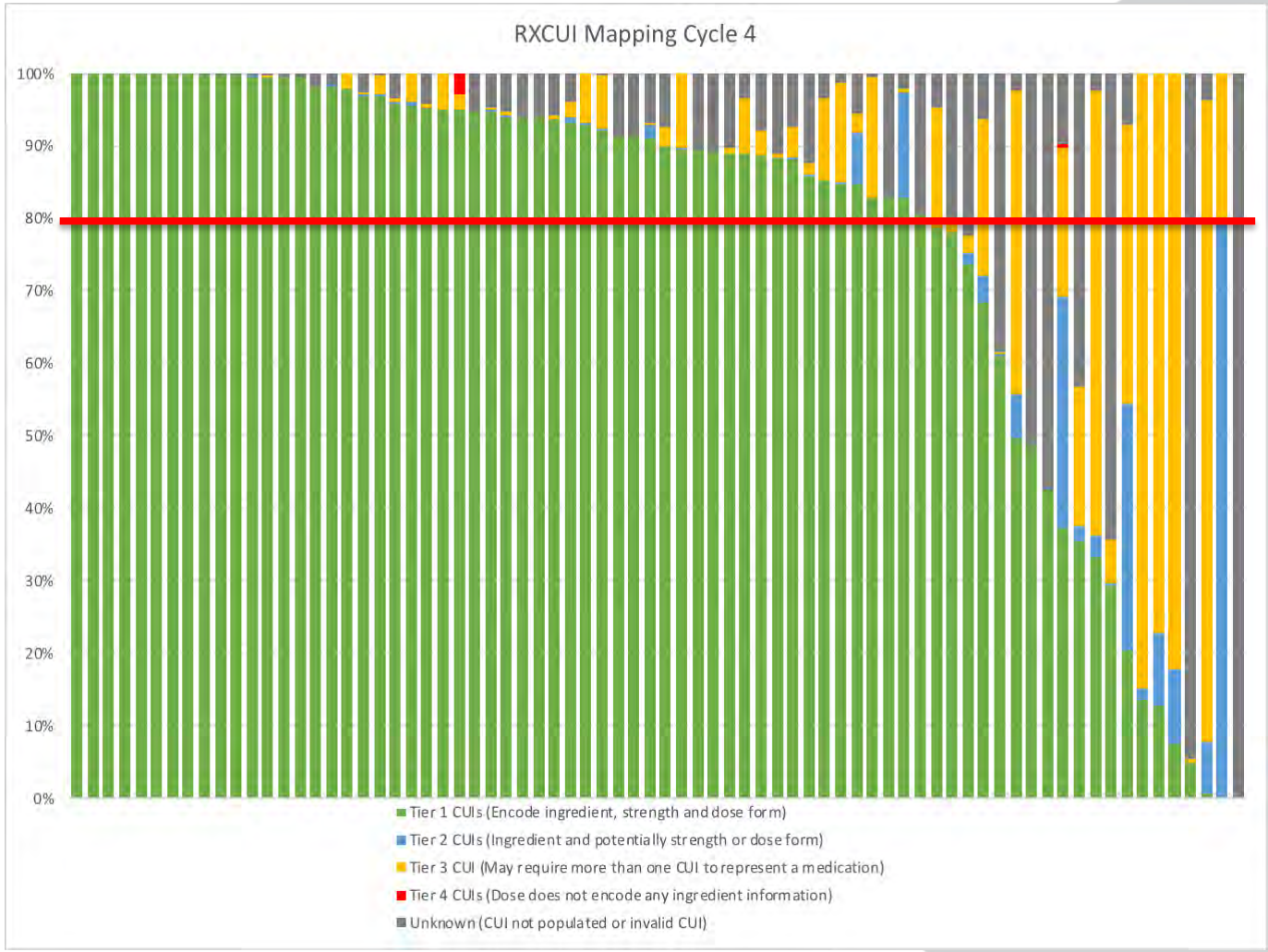
Study-specific data curation

- First challenge: convincing investigators that this step is even necessary (even more difficult if Coordinating Center is not the one running the study)
- Second challenge: what do to do with the results
  - Address the issue & incorporate into the foundational curation process (preferred)
    - Medication coding
    - Data latency
  - Consider proxy variables
    - Days supply
  - Leverage alternative data sources
    - Collect data on events directly from patients to supplement CDM (ADAPTABLE – out of scope for this talk)

pcornet®

Medication coding

- Information about the medication ingredient, strength, and dose form is needed for many studies
- Implementation Guidance developed to establish the preferred mapping strategy
- Data Curation added a data check to measure adherence to the guidance



pcornet®

Incorporating medication coding into data curation

Implementation Guidance Reference Table 4: Ordering of RxNorm Term Types

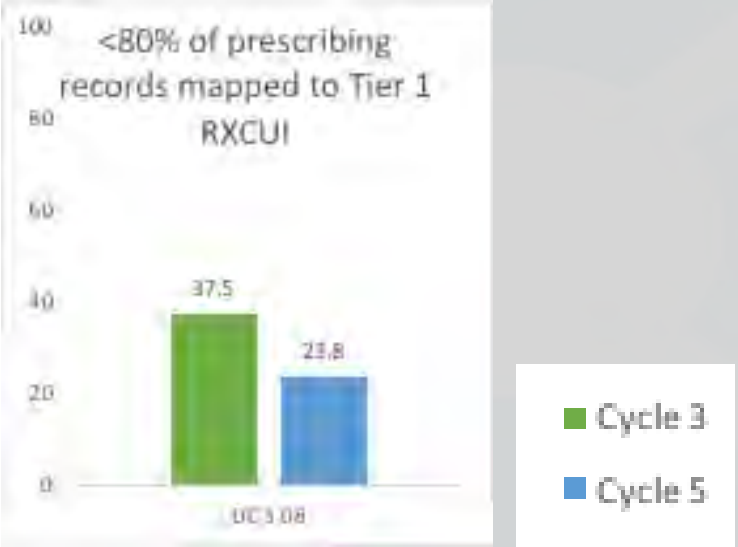
(Content from the UMLS [https://www.nlm.nih.gov/research/umls/rxnorm/docs/2015/appendix5.html] – Accesse

	RxNorm Term Type		Information incorporated			
	Code	Description	Ingredient(s)	Strength	Dose Form	Brand Name
Most Preferred	SBD	Semantic Branded Drug	X	X	X	X
	SCD	Semantic Clinical Drug	X	X	X	
	BPCK	Brand Name Pack	X	X	X	X
	GPCK	Generic Pack	X	X	X	
	SBDF	Semantic Branded Drug Form	X		X	X
	SCDF	Semantic Clinical Drug Form	X		X	
↓	SBDG	Semantic Branded Dose Form Group			X	X
	SCDG	Semantic Clinical Dose Form Group	X		X	
	SBDC	Semantic Branded Drug Component	X	X		X
	BN	Brand Name				X
	MIN	Multiple Ingredients	X			
	SCDC	Semantic Clinical Drug Component	X	X		
	PIN	Precise Ingredient	X			
Least Preferred	IN	Ingredient	X			
Do not use	DF	Dose Form			X	
Do not use	DFG	Dose Form Group			X	
Do not use	PSN	Prescribable Name				
Do not use	SY	Synonym				
Do not use	TMSY	Tall Man Lettering Synonym				

Table IVG. RXNORM Term Type Mapping

This table shows the number of records in the PRESCRIBING table by RXNORM Term Type tiers. Guidance on mapping prescribing orders to RXNORM is provided in the CDM. These data support Data Check 3.08 (less than 80% of prescribing orders are mapped to a RXNORM_CUI which fully specifies the ingredient, strength and dose form). Data check exceptions occur if the Tier 1 percentage is <80% or the numerator is 0. Exceptions are highlighted in blue and should be investigated and explained in the ETL ADD.

Term Type Tier	Term Type Tier Description	Term Types	Numerator	Percentage	Source table
Tier 1	RXNORM_CUI encodes ingredient(s), strength and dose form	SCD, SBD, BPCK, and GPCK	2,204	13.90	PRES_L3_RXCUI_TIER
Tier 2	RXNORM_CUI encodes ingredient(s) and potentially strength or dose form. Can still represent medications with multiple ingredients with a single RXCUI.	SBDF, SCDF, SBDG, SCDG, SBDC, BN,and MIN	7,118	44.90	PRES_L3_RXCUI_TIER
Tier 3	Requires more than one RXNORM CUI to represent medications with multiple ingredients.	SCDC, PIN, and IN	5,888	37.14	PRES_L3_RXCUI_TIER
Tier 4	RXNORM_CUI does not encode any ingredient information.	DF and DFG	0		PRES_L3_RXCUI_TIER
Unknown	RXNORM CUI was not populated or could not be matched to the reference table	n/a	642	4.05	PRES_L3_RXCUI_TIER

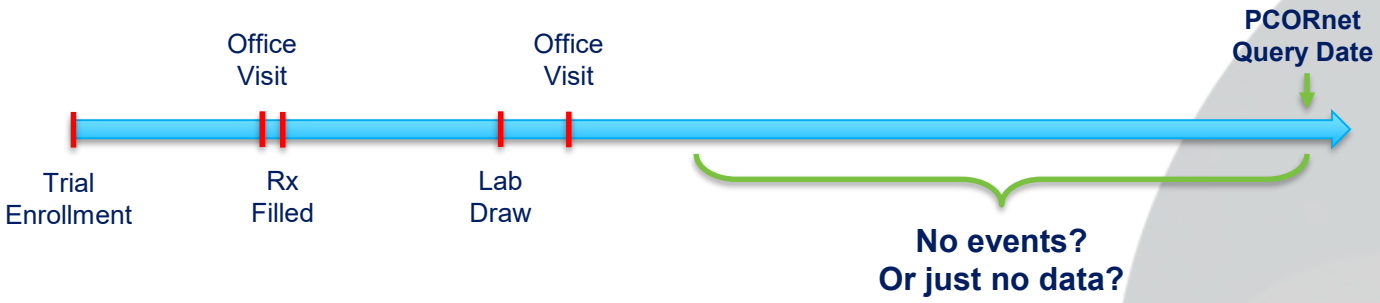


Note: all partners must pass this check starting July 2019

pcornet®

Data latency

- Latency / completeness of data



- Questions:
  - *"How complete & up-to-date are the data we're looking at?" (DSMB)*
  - *"What's the data censoring date for participants?" (Statistician)*
- Developed latency calculation & incorporated into data curation

pcornet®

Data latency as part of data curation

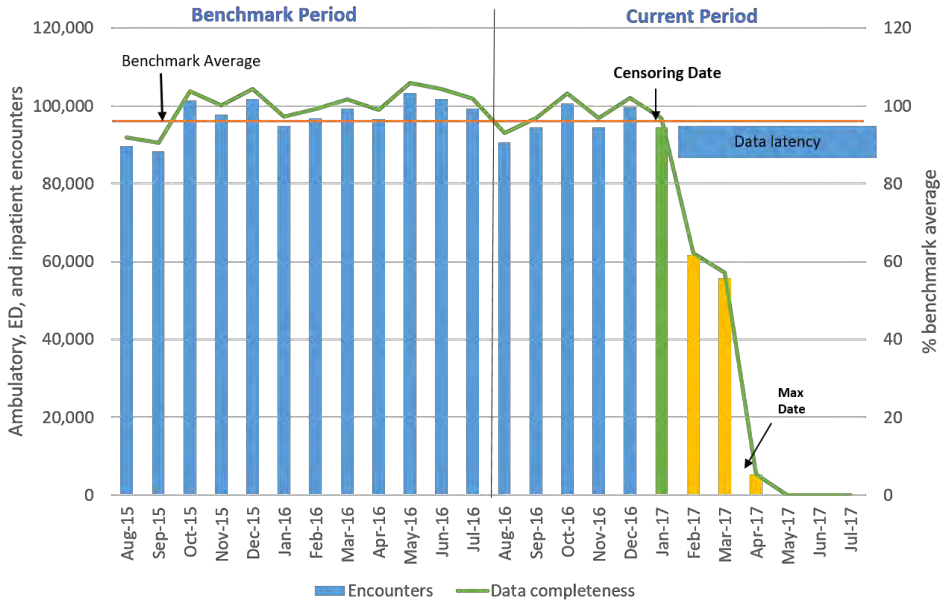


Table IVG. Data Latency and Completeness of Vital, Prescribing, and Lab Data, Past 2 Years

This table includes VITAL, PRESCRIBING, and LAB_RESULT_CM data from the most recent 24 month period; month -0 is the month the data curation query was run. Data completeness is determined by comparing the actual volume to the expected volume in each month. Expected volume is determined by taking the average volume during the benchmark period of months -12 to month -23. Data completeness is reported as a percentage of the benchmark average. Temporal differences may be affected by data availability, ETL processes, date shifting, secular trends, and/or changes in data provenance.

These data support Data Check 3.11 (vital, prescribing, or laboratory records are less than 75% complete three months prior to the current month). Data check exceptions occur if the month -3 result is <75% of the benchmark average or 0 records. Data check exceptions are highlighted in blue. Data check exceptions and unexpected results should be investigated and explained in the ETL ADD.

	Vitals		Prescriptions		Labs	
Month	Records	Percent of benchmark average	Records	Percent of benchmark average	Records	Percent of benchmark average
Month -0	60,980	9.8	16,015	13.0	82,977	13.0
Month -1	495,533	79.4	118,617	96.3	583,263	91.3
Month -2	560,362	89.7	121,318	98.5	604,813	94.7

pcornet

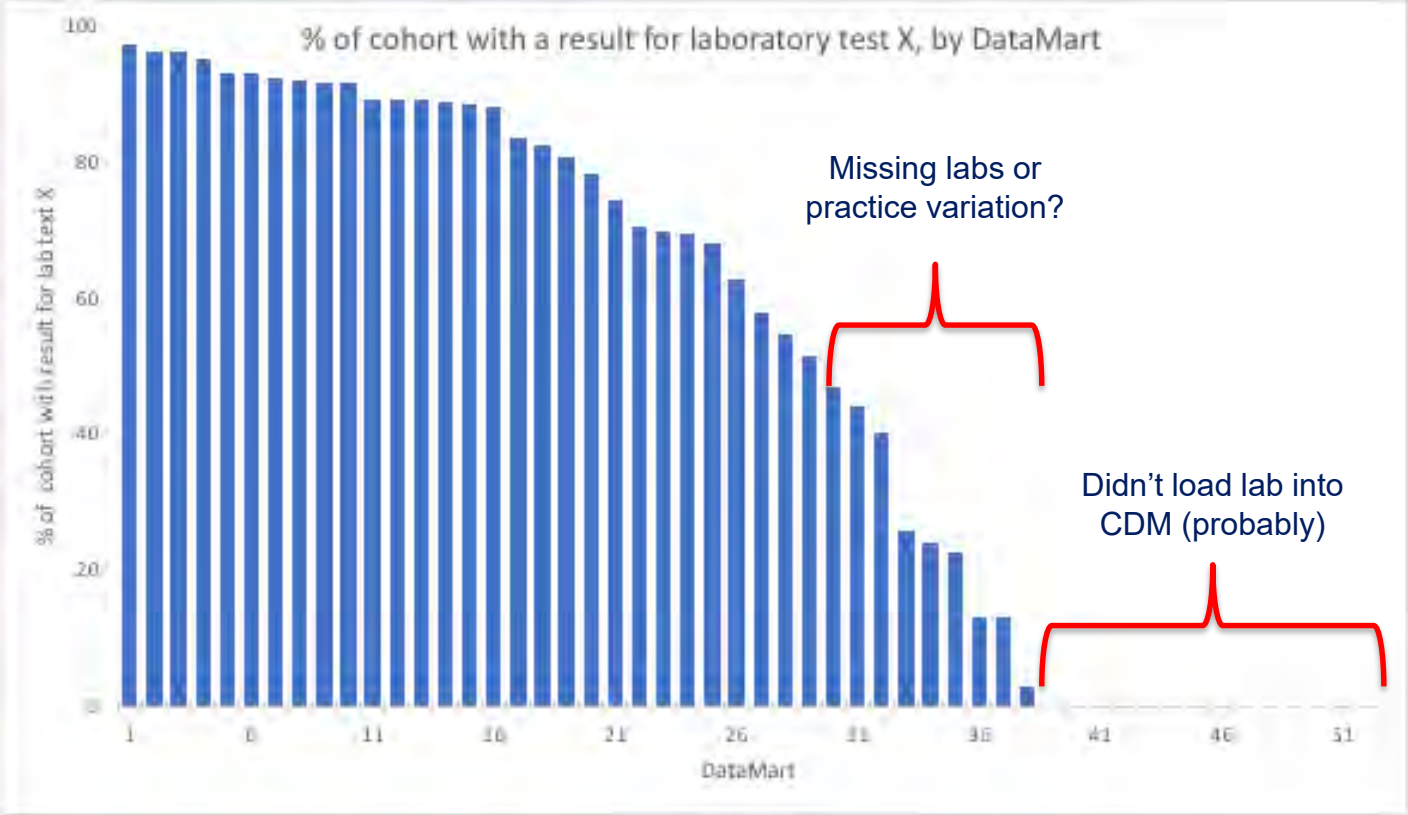
Proxy variables – days supply

- Study Aims: To evaluate the comparative effects of different types, timing, and amount of antibiotics prescribed during the first 2 years of life on:
  - Body mass index and risk of obesity at 5 and 10 years
  - Growth trajectories from infancy onwards
- Sample findings from study-specific characterization
  - Days supply – highly missing
  - Start date minus end date – low percent missing – **very different from the global measure**
- One key takeaway – a proxy variable for one study may not be suitable for another

pcornet®

Open issues (one example)

- Differentiating between data quality issues & normal practice variation



Next steps / recommendations

- Need to stress importance of fixing data issues that can be resolved
  - Datamart administrators are typically not the ones using the data, so they may not understand the impact of leaving things unaddressed
- Identify incentives that would improve data quality on the front end
  - Clinicians will support changes in workflow (within reason) if there's a benefit to them
  - Goes beyond research – precision medicine, analytics, etc. (better care?)
- Define guidance for what it means to be "regulatory grade"
  - Can we create a checklist as opposed to "we know it when we see it"?

pcornet®

Session II: Study Specific Data Curation to Establish a Fit-for-Purpose Dataset

LUNCH

Session III: Linking Multiple Data Sources

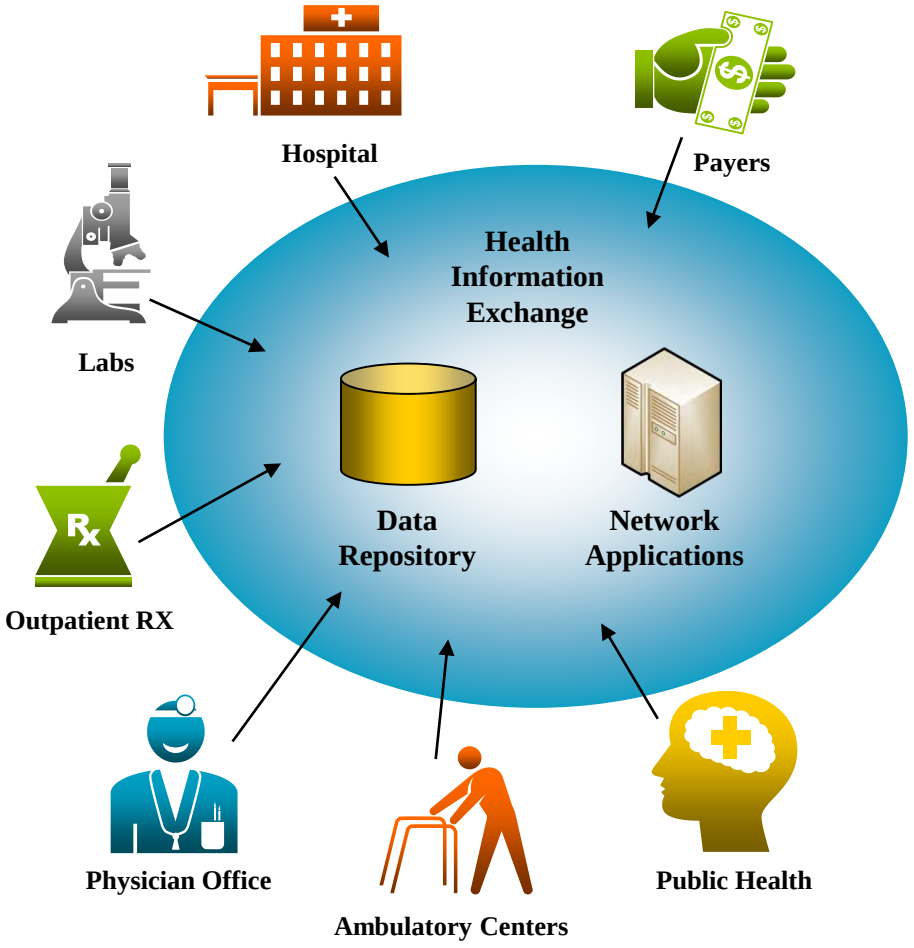
Regenstrief Institute
Better Care. Better Health.

Linking Multiple Data Sources: Considerations for Use Cases and Quality

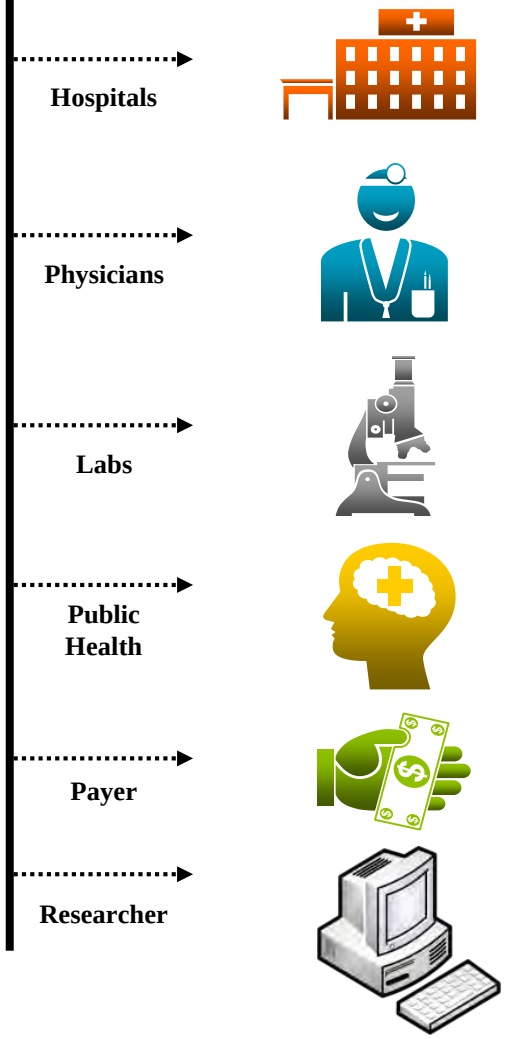
Shaun J. Grannis, MD, MS, FAAFP, FACMI
Director, Regenstrief Center for Biomedical Informatics
Regenstrief Clem McDonald Scholar for Biomedical Informatics
Associate Professor, Family Medicine, IU School of Medicine
Biomedical Research Scientist, Regenstrief Institute

Data Linkage: The Indiana Network for Patient Care (INPC)

Data Management



Data Access & Use



Hospitals
- Results delivery
- Secure document transfer
- Shared EMR
- Credentialing
- Eligibility checking

Physicians
- Results delivery
- Secure document transfer
- Shared EMR
- CPOE
- Credentialing
- Eligibility checking

Labs
- Results delivery

Public Health
- Surveillance
- Reportable conditions
- Results delivery
- De-identified, longitudinal clinical data

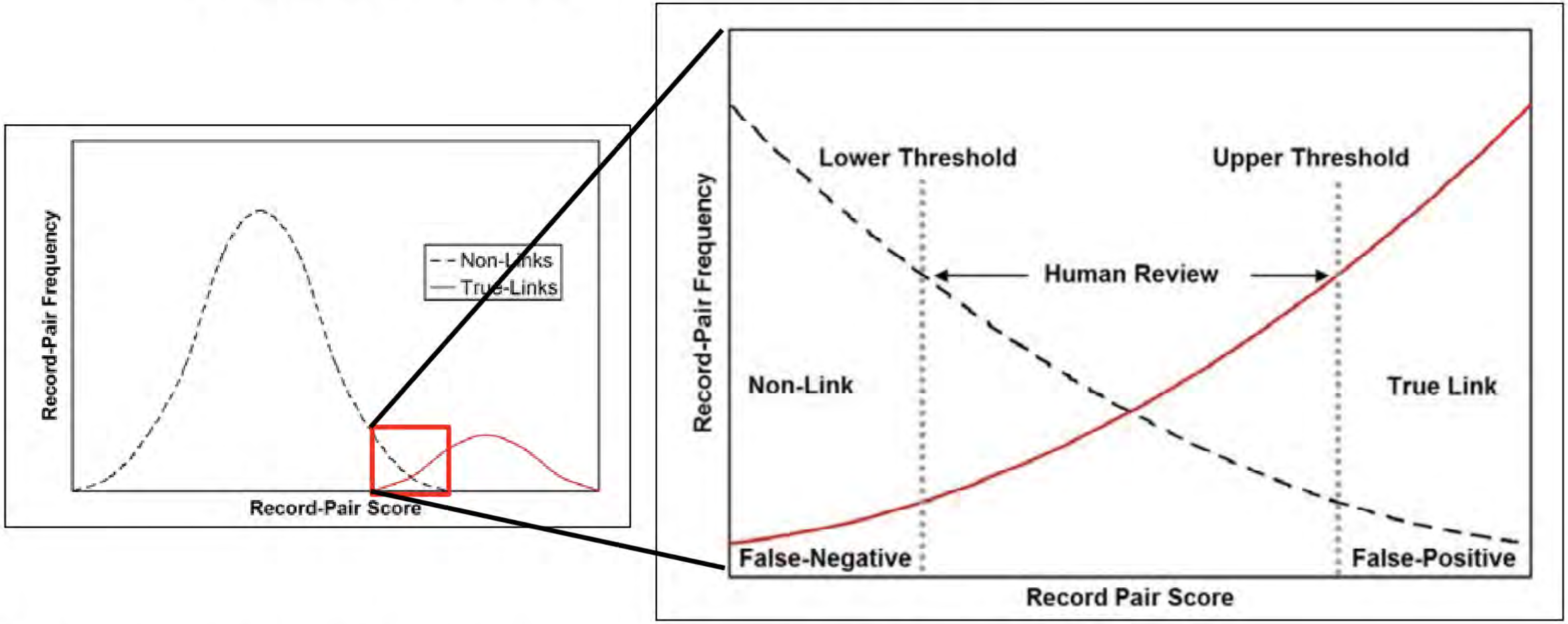
Payer
- Secure document transfer
- Quality Reporting

Researcher
- De-identified, longitudinal clinical data (OMOP CDM, i2b2)
- Subject Recruitment
- Clinical Trials

Figure 7-1. Examples of Matching Scenarios Broken Down By Dimensions of Workflow Timing and Human Supervision

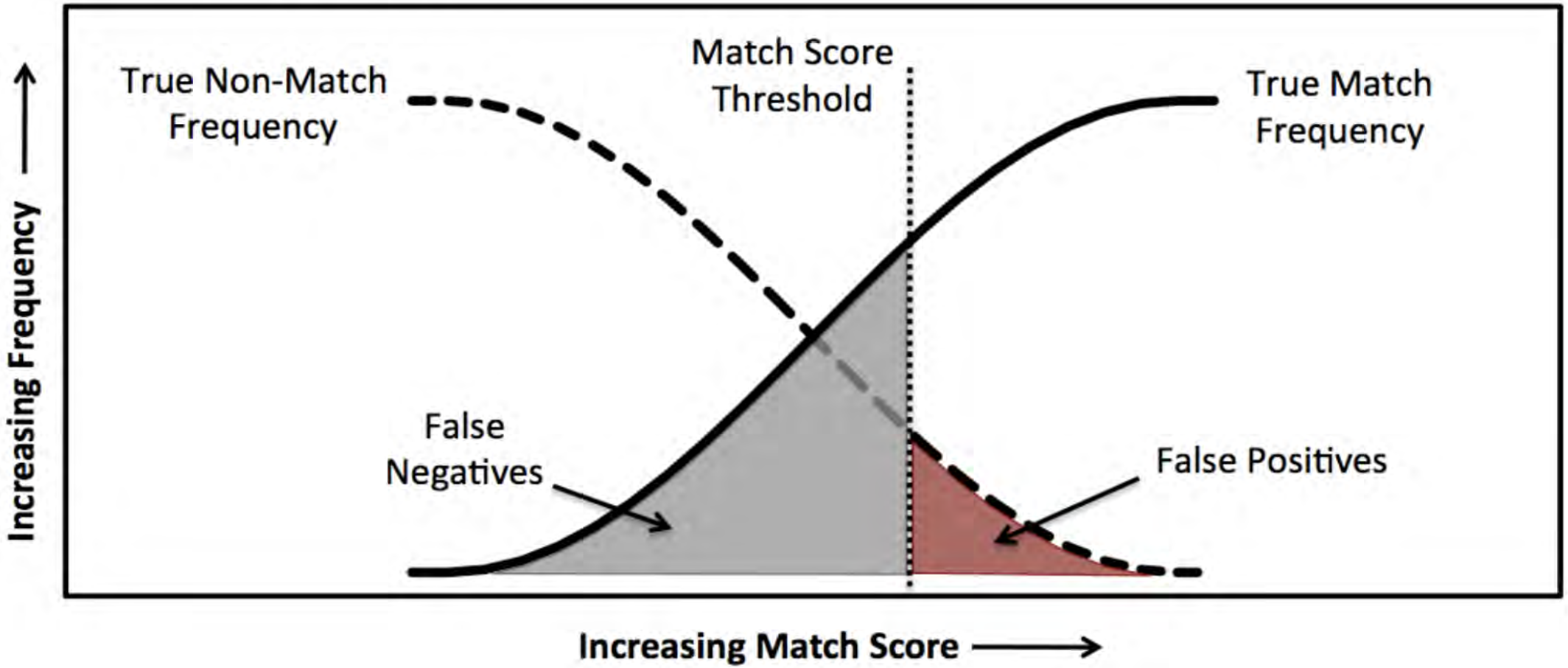
Human Supervision	Workflow Timing: Batch Mode	Workflow Timing: Real Time
Substantial Manual Supervision	**Reporting, Research**	**Health Care Enterprise (Hospitals)**
Little or No Manual Supervision	**De-identified Matching**	**Health Information Exchange**

Figure 5-1. Illustration of the Intermediate Score Range Where Both True Matches and Non-Matches Are Present



NOTE: To disambiguate these linkages, human review is often necessary.

Figure 7-2. Illustration of the Relationship Between False Positive and False Negative Matches



NOTE: As the match score threshold is increased, the number of false positives decreases, but false negatives increase. As the match score threshold is lowered, the number of false negatives decreases, but false positives increase.

Linkage Metrics

1. Algorithm metrics:
   - sensitivity (recall), PPV (precision), F-measure
2. Data Quality metrics:
   - completeness (missing rate)
   - accuracy/error rates (conformance to known data requirements/business rules)
   - discriminating power (various measures)
3. Business processes metrics
   - Data validation methods
   - Compliance with established process standards

How to compare across sites/regions?

Regenstrief Institute
Better Care. Better Health.

Linking Multiple Data Sources: Considerations for Use Cases and Quality

Shaun J. Grannis, MD, MS, FAAFP, FACMI
Director, Regenstrief Center for Biomedical Informatics
Regenstrief Clem McDonald Scholar for Biomedical Informatics
Associate Professor, Family Medicine, IU School of Medicine
Biomedical Research Scientist, Regenstrief Institute

Session III: Linking Multiple Data Sources

DATAVANT

Connecting the world's health data

What we do

1. Protect

De-identify datasets to protect patient privacy and reduce risk

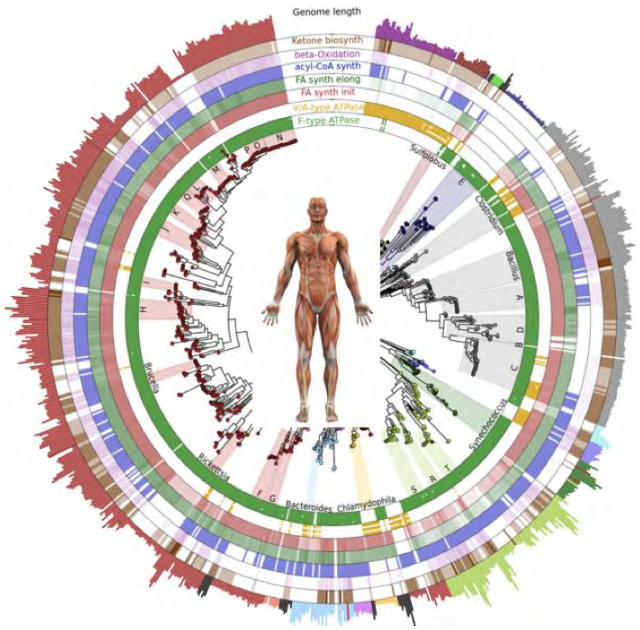
2. Link

Connect matching patient records across datasets to increase data completeness and dimensionality

3. Discover

Help institutions discover data sources that augment their knowledge of a population

Assembling a more holistic view of the patient…



…to expand the set of questions that can be answered in healthcare

Secure, HIPAA-Compliant De-identification

- Datavant's technology can be installed on-premise, meaning that **we don't need access to client's data or systems**
- We work with clients to **configure the de-identification rules required for a specific data layout and use case,** using Safe Harbor or the Expert Determination method to ensure compliance with HIPAA

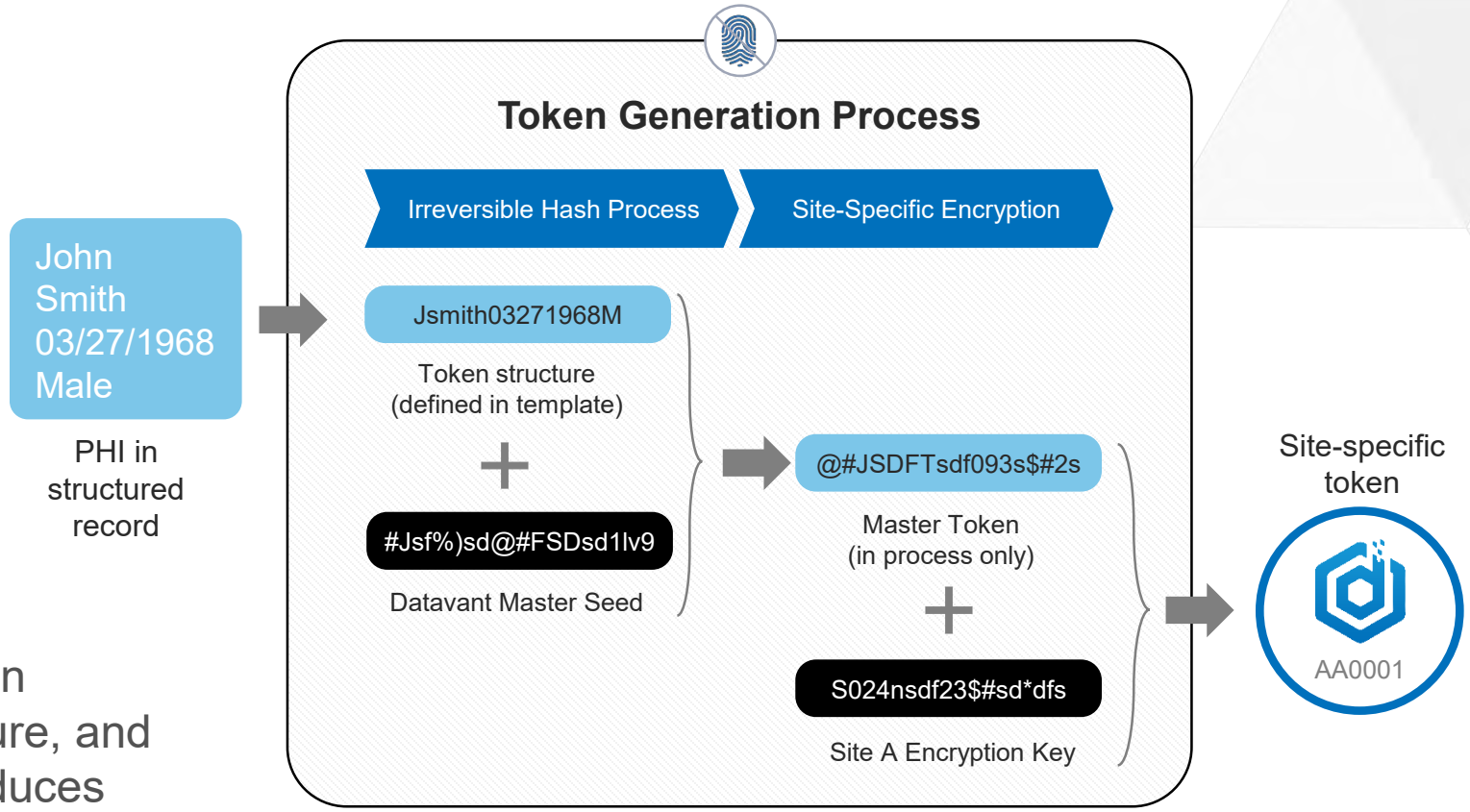
Input data		Configured template with rules	Output data
First Name	John	*Remove*	Null
Last Name	Smith	*Remove*	Null
Gender	Male	*Pass through*	Male
Date of Birth	March 27, 1968	*Convert to birth year*	1/1/1968
Address	5 Pine Street	*Remove*	Null
Zip Code	95401	*Convert to 3-digit zip area*	954
		Create token from name, DOB and gender	AA001

De-identification engine

Adding Anonymized Linking Tokens to Each Record

Token creation has two steps:

1. **Hashing**: Makes tokens irreversible, securing users from employee or Business Associate regulatory violations
2. **Encryption**: Makes tokens site-specific, protecting users from a partner's security breach



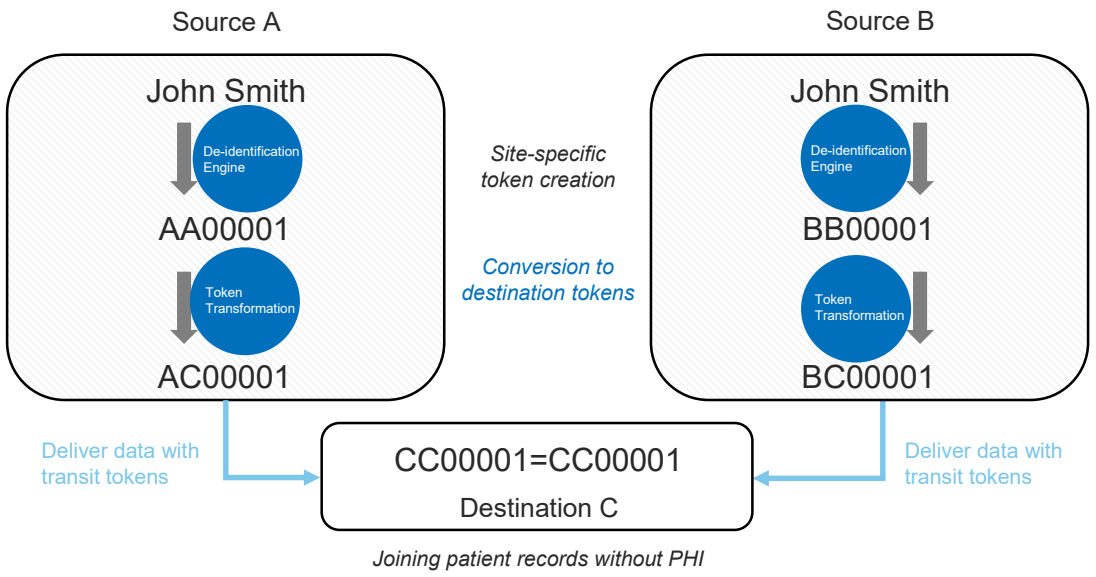
Our tokenization process has been cryptographically-certified as secure, and our de-identification software produces datasets that have been certified to be in **compliance with HIPAA**

Linking De-Identified Data With Tokens

Connect patient records across multiple datasets **without ever sharing PHI**

- Because tokens are site-specific, they cannot be matched across sites unless they are transformed.
- When both parties agree to exchange data, Datavant enables a second piece of software to convert tokens from one encryption key to another.
- In this way, tokens from different sources can be converted into a common encryption key to allow joining.
- Once in a common key, tokens from the different datasets are matched according to each user's needs.

Multiple sources sending data to recipient



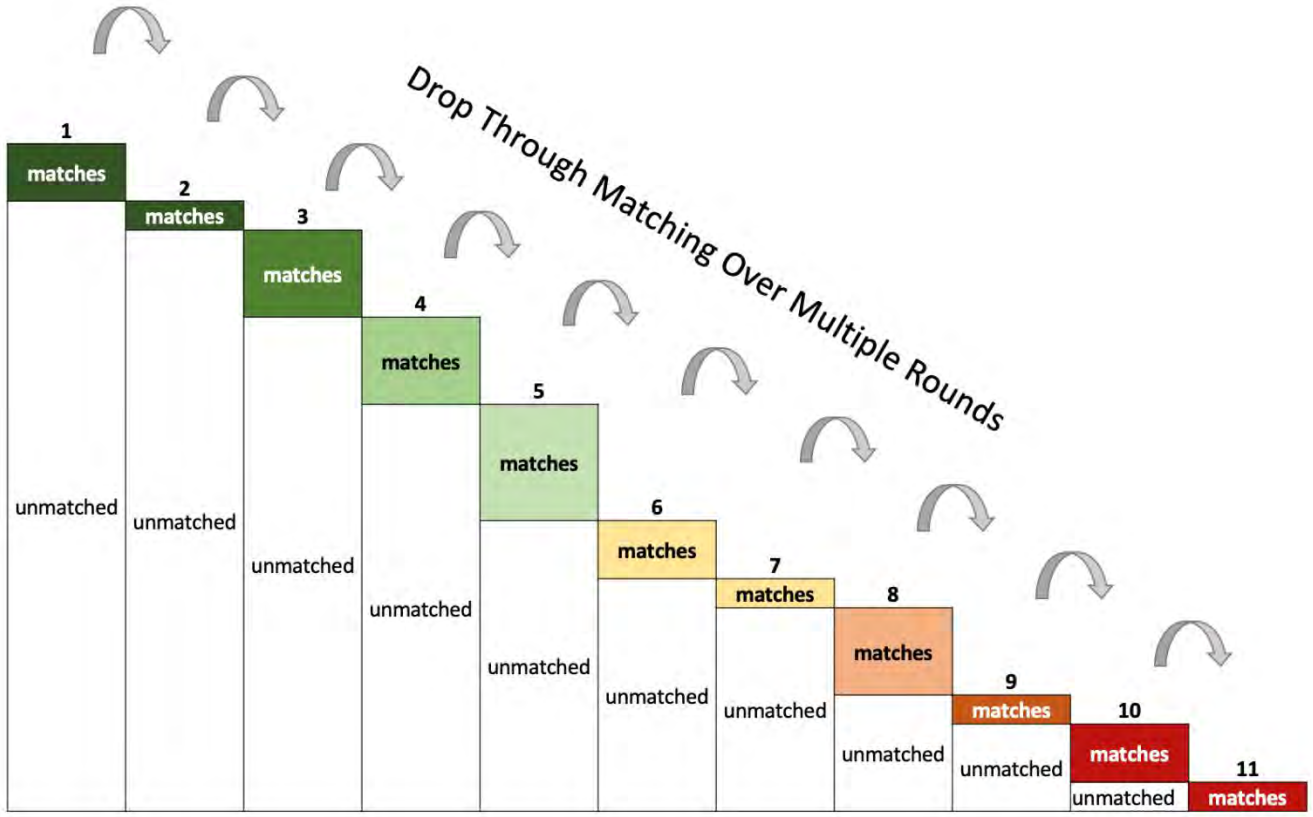
Logic to Support Stringent AND Broad Matching

We recommend not picking a single token or token combination for matching logic, but to instead take advantage of multiple matching options using a "drop through" or “waterfall” technique.

1. The most stringent set of tokens are used in the first round to define a match.
2. Any records matched in this round are put aside, and only unmatched records move to the next round.

This cycle is repeated using less and less stringent matching logic over multiple rounds.

Best matches are always made first, with only a few rounds used for stringent matching, and many rounds used for broad matching.



Appendix

Matching with Datavant Tokens

Using Datavant's software, companies can de-identify and tokenize patient records so that they can be linked across disparate datasets.

Patient records can be linked based on token matches (when tokens are in the same site key). The quality of a given match depends on the tokens used and on the specific matching logic.

Datavant has many different token types that are composed from different combinations of PII:

- Some designs are deterministic (using Social Security Number, for example)
- Most designs are probabilistic (based on a combination of non-unique fields such as: first name, last name, DOB and gender)

Datavant recommends adding <u>multiple</u> tokens to each data file to:

- Increase the chances that de-identified datasets will share common tokens and be join-able
- Increase accuracy of matching by having more tokens with which to confirm a match result
- Allows clients to select matching stringency – from strict to broad – depending on their specific use case and their sensitivity to either false positives or false negatives

DATAVANT

2 Embarcadero Center
9th Floor
San Francisco, CA 94111

415-520-1171
info@datavant.com

Sam Roosz
Head of Partnerships

+1.765.490.9385
sam@datavant.com

Session III: Linking Multiple Data Sources

January 22, 2019

TriNetX

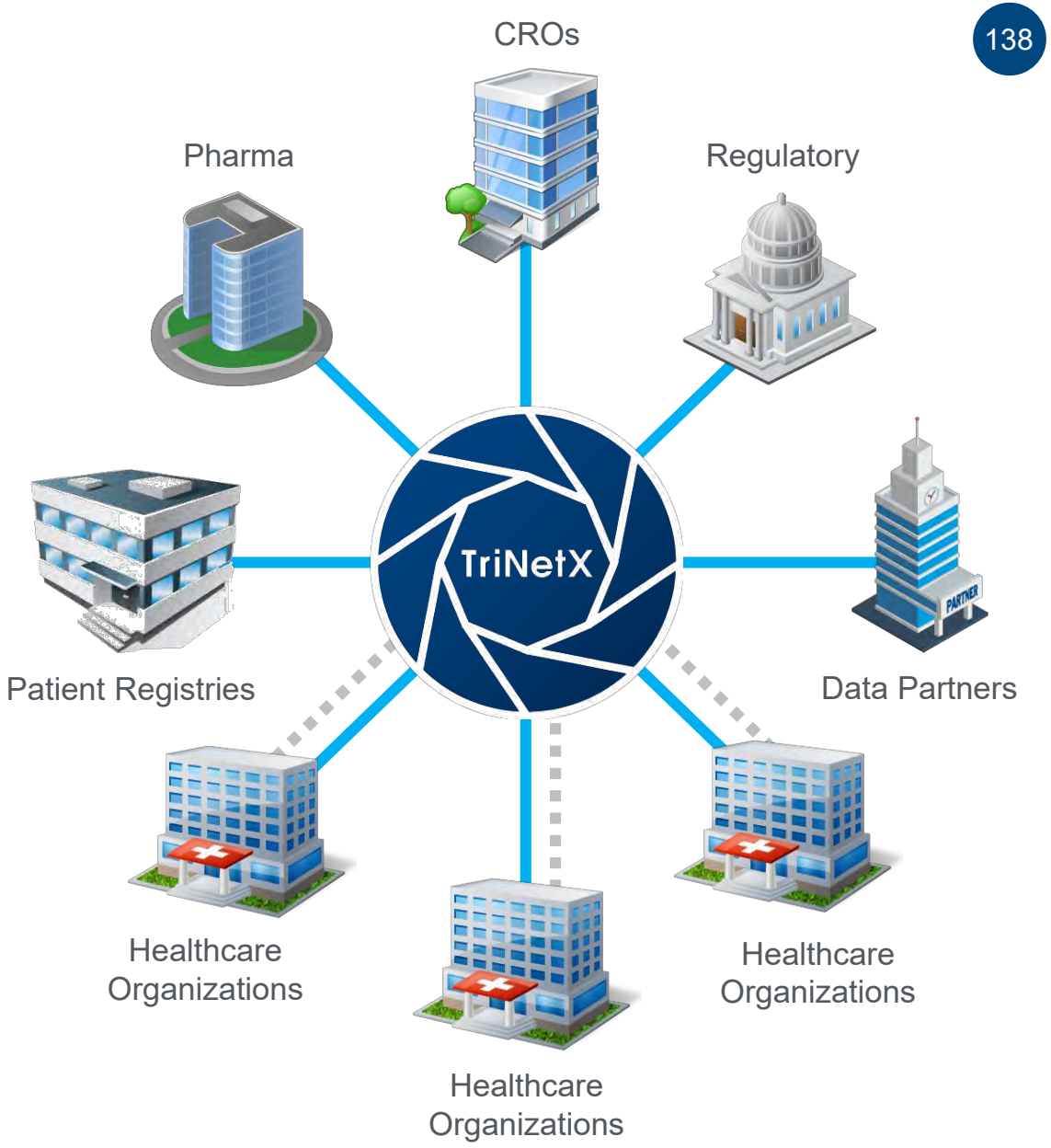
The Global Health Research Network

LINKING IN PRACTICE

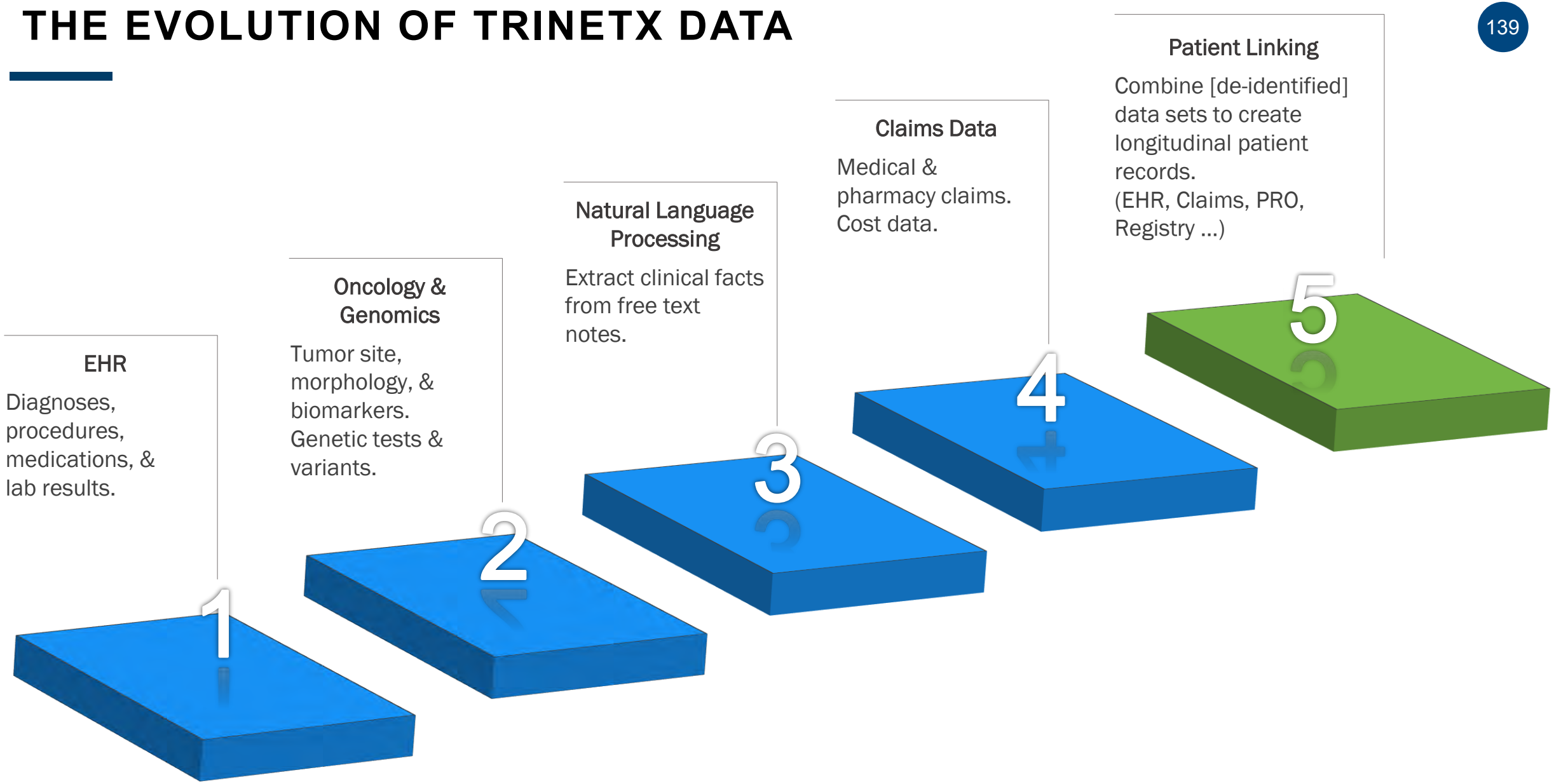
PRESENTED BY:
Steven Kundrot

TriNetX

- Global health research network
- Cloud-based platform enabling on-demand access to real-world data and analytic tools
- Data sourced and continuously refreshed from EMRs, Claims, PRO, registries and unstructured sources
- Path back to the patient via IRB and Honest Broker
- Data is downloadable
- Federated model & compliant with international privacy standards



THE EVOLUTION OF TRINETX DATA



TriNetX

LINKING: SOLUTION CONTEXT

Implementation within the context of a federated, global network …

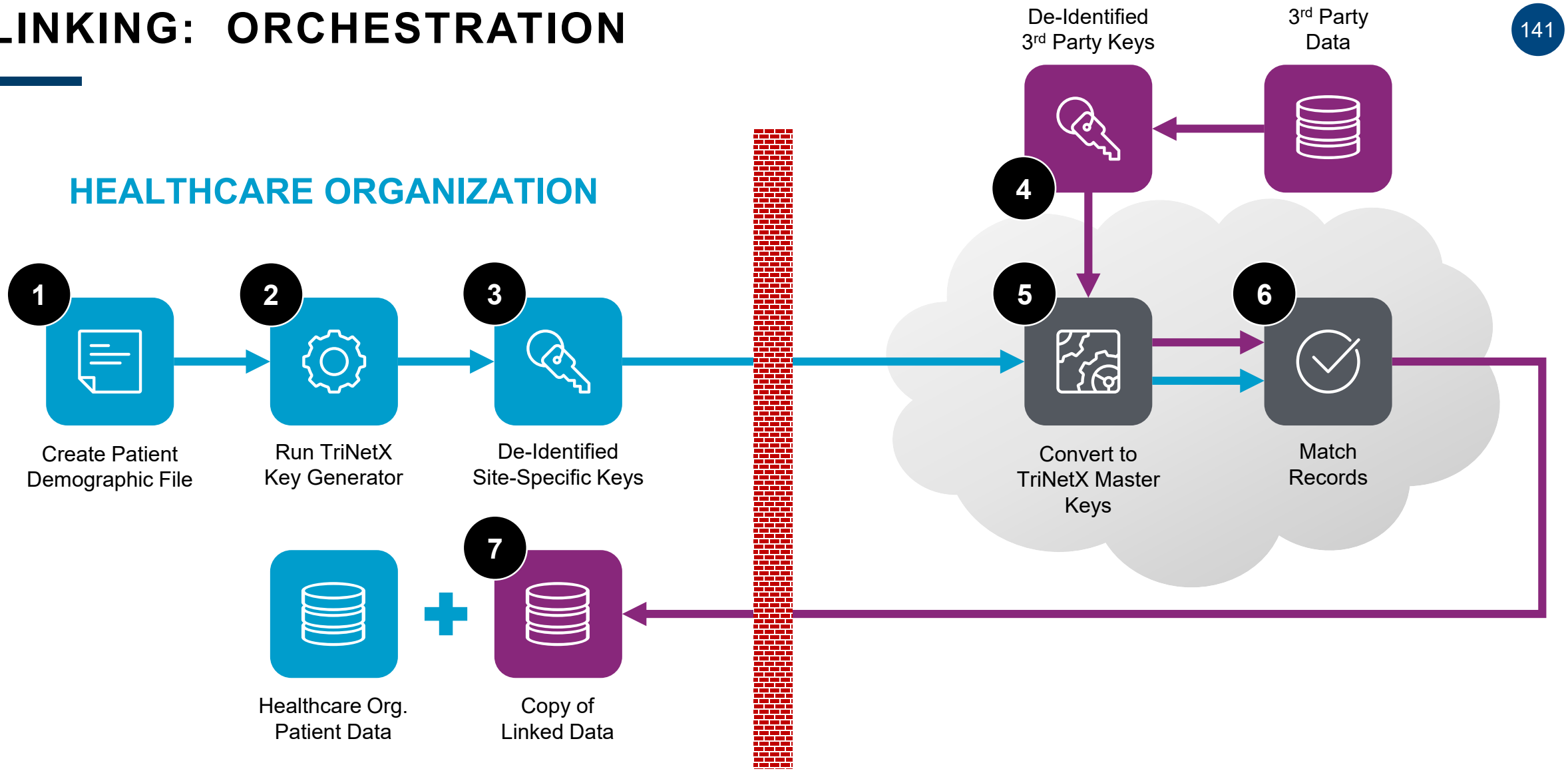
KEY ASSUMPTIONS

- Governance/privacy
- Broad applicability
- Matching validity
- Performance and scale
- Flexible implementation

VENDOR SNAPSHOT

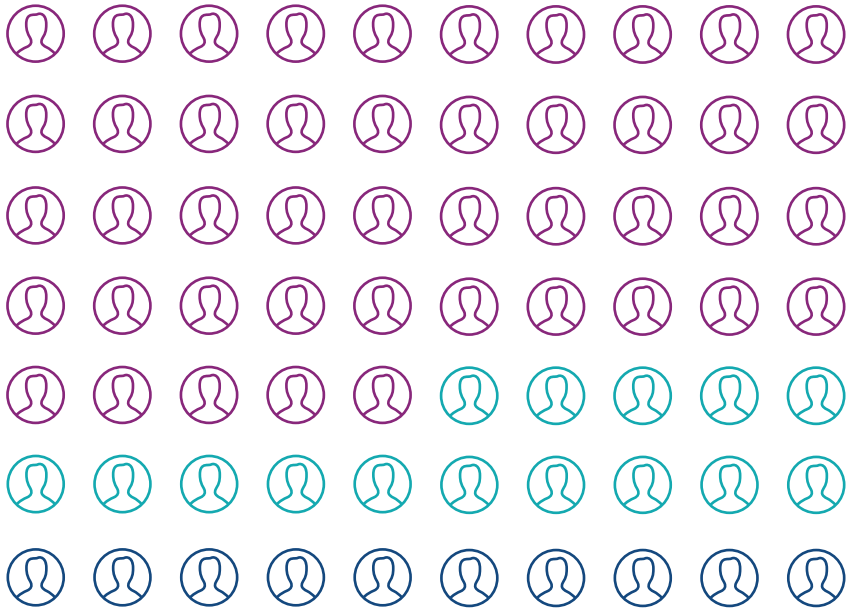
- Datavant / UPK
- Health Data Link
- Verato
- Experian
- Health Verity
- Symphony Health

LINKING: ORCHESTRATION



LINKING: GOVERNANCE

ELIGIBLE COHORT



AUTHORIZED FOR EXPORT ID

SOURCE OF ELIGIBILITY

- Healthcare Organization Data
- Linked Claims Data
- Linked Patient Reported Outcomes Data

LINKING: RESULTS

HCO	Total Patients	Linked Patients	Linking %
HCO 1	1,971,715	1,322,343	67%
HCO 2	918,569	693,199	75%

- Matches based on 99% probability
- Potential for pool and depth/breadth increase

HCO	Orphan Patients	Orphan Patients Recovered	% Recovered	% Patient Pool Increased
HCO 1	928,257	462,536	50%	44%
HCO 2	3,701	1,017	27%	0.1%

- Orphan patient: a patient w/o any facts before linking
- Patient pool increased

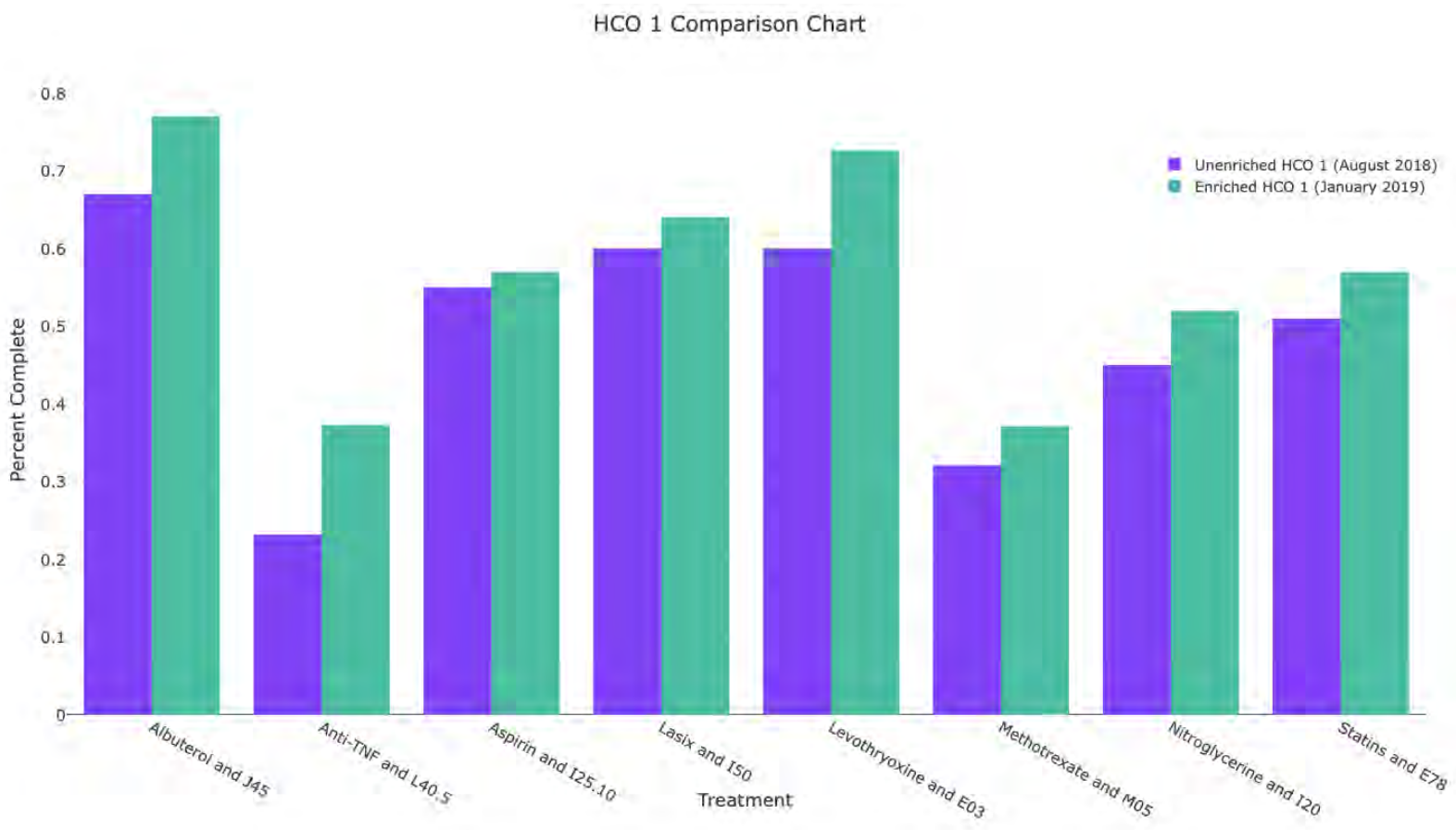
HCO	HCO Death Data	Linked Death Data	Death Addition	% Increase
HCO 1	144,264	309,462	271,686	188%
HCO 2	12,615	36,134	25,978	206%

- Depth of deceased knowledge increased
- Decease pool increased

HCO	HCO Facts	HCO Facts for Linked Patients	New Linked Facts	Depth, Breadth Potential Increase
HCO 1	1,046,431,944	845,492,762	326,163,485	39%
HCO 2	299,453,541	241,722,712	106,539,304	44%

- Potential for clinical depth/breadth increase
- Potential for longitudinal increase

LINKING: RESULTS



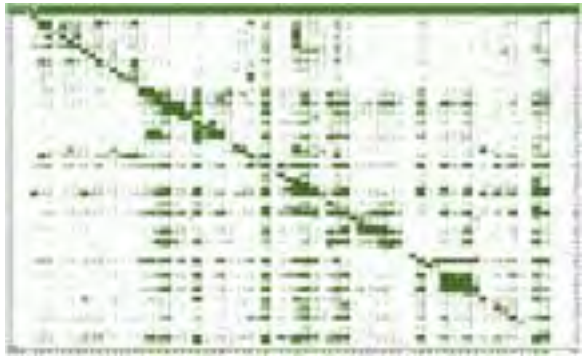
	RA Patients	# with 5yrs+ span	% with 5yrs+ span
Unenriched	3,540	1,650	47%
Enriched	4,160	3,220	77%

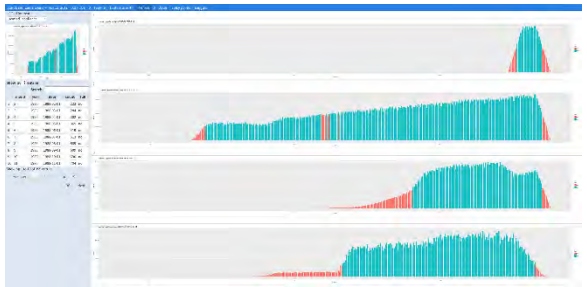
- Increase in completeness
- Increase in longitudinally

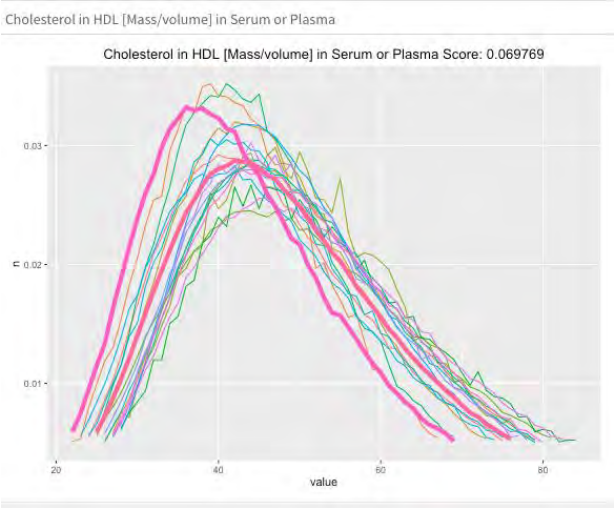
LINKING: WHAT'S NEXT

- Linking throughout our network
- On-going assessment of linking
  - Quality of matching
  - Depth/breadth significance
- Development of standard metrics
  - Transparent to community

HCOs	Linked Patients
HCO 1, HCO 2	2,035
HCO 1, HCO 2, HCO 3	34
…	
…	







THANK YOU!

125 Cambridgepark Drive, Suite 500
Cambridge, MA 02140 USA

857.285.6037

join@trinetx.com

trinetx.com

Session III: Linking Multiple Data Sources

BREAK

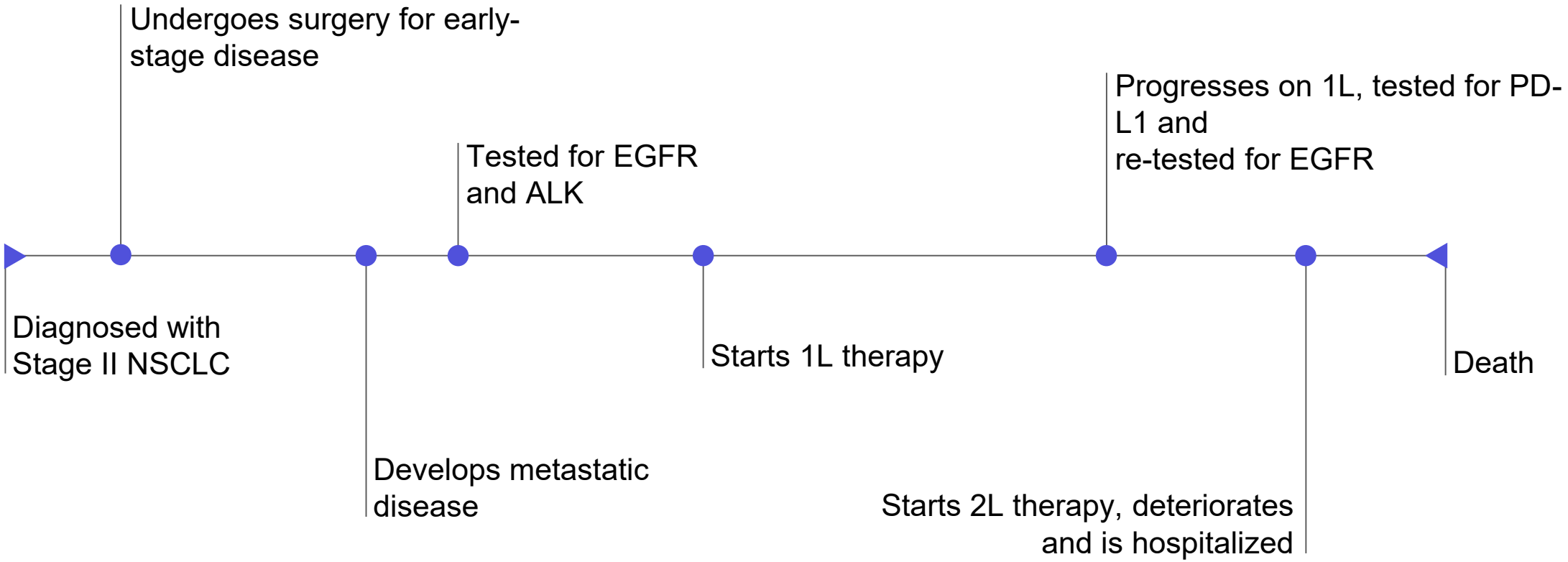
Session IV: Submitting Data Documentation for Traceability and Auditing

January 22, 2019

Session IV: Documentation for Traceability and Auditing

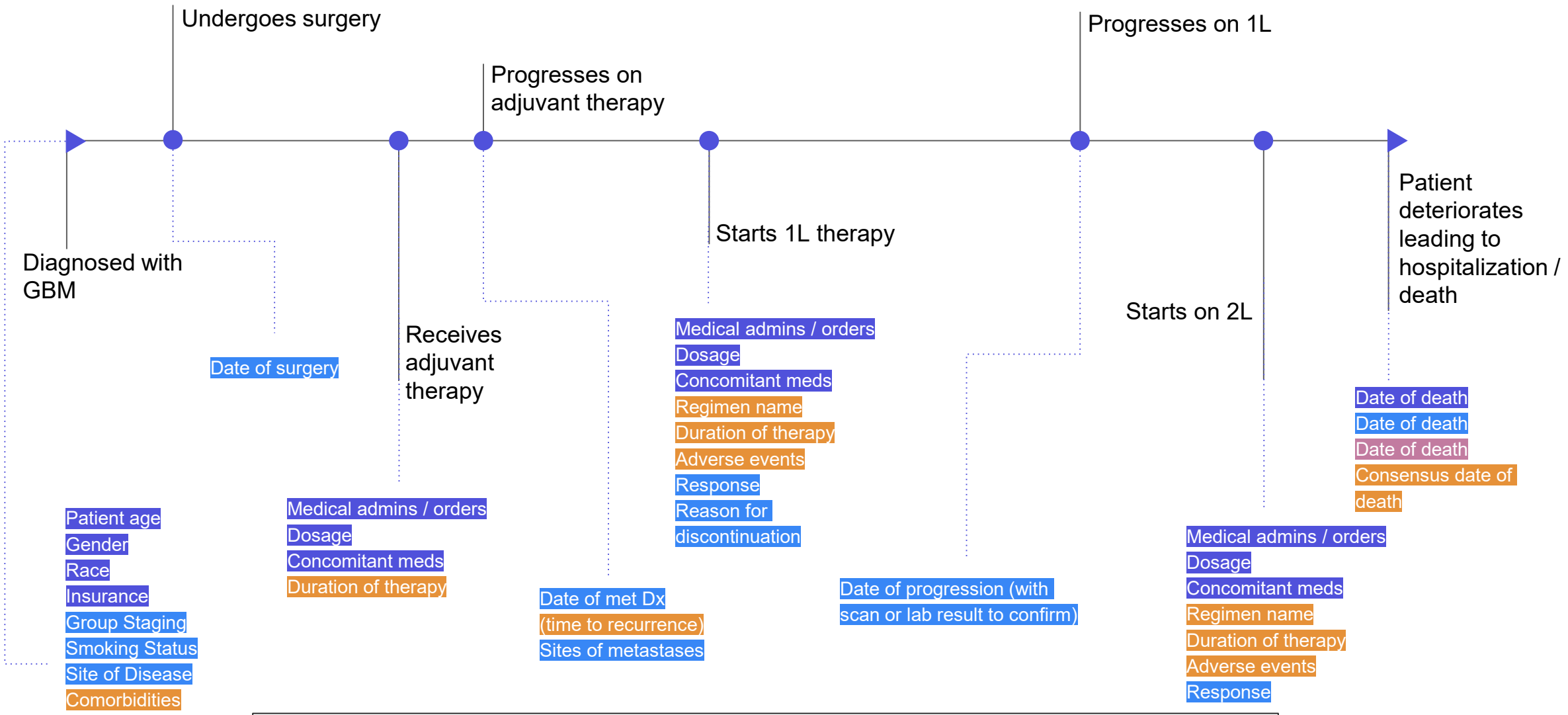
Amy Abernethy, MD, PhD
Chief Medical Officer / Chief Scientific Officer & SVP - Oncology, Flatiron Health *(a member of the Roche Group)*
Adjunct Professor of Medicine, Duke University School of Medicine
@dramyabernethy ♦♦ amy@flatiron.com

flatiron



Documentation of source, quality and provenance.

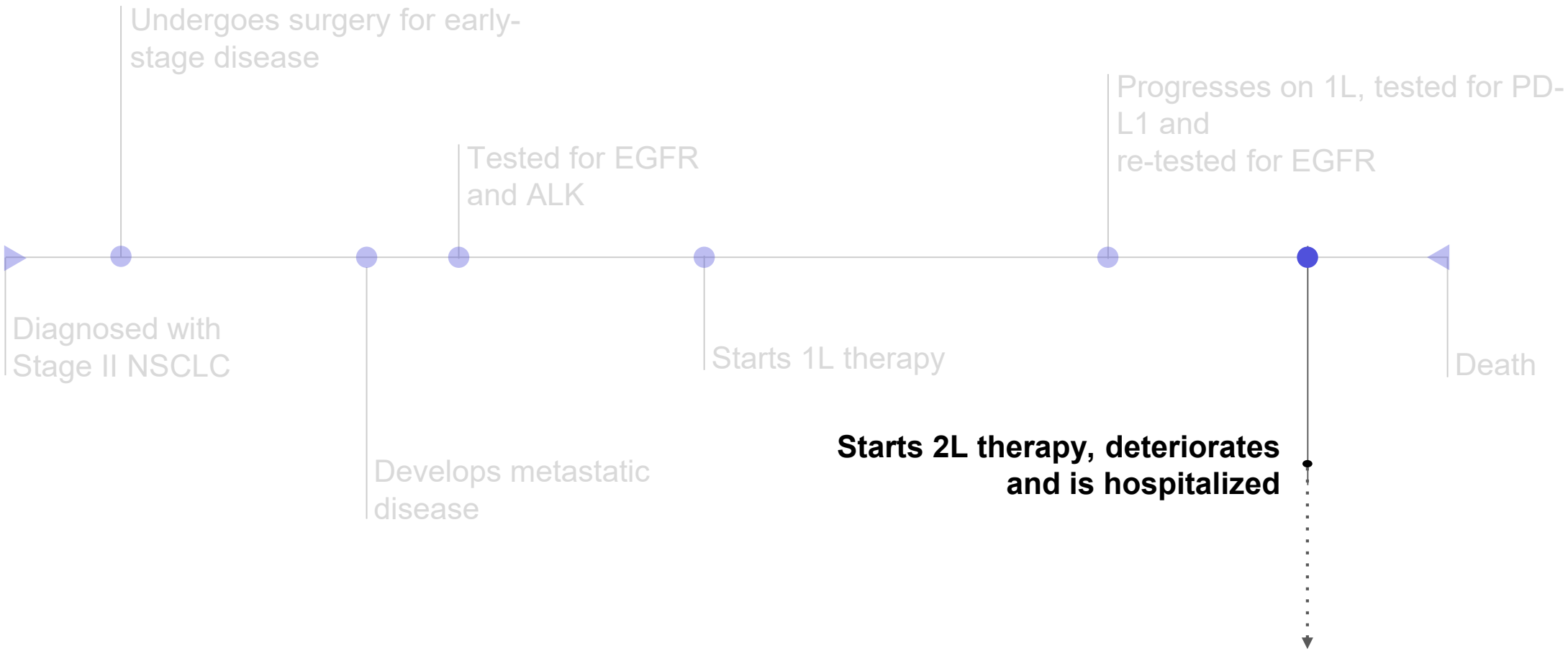
A comprehensive view of the patient journey



Structured EMR data | Unstructured EMR data | External mortality data | Combined / derived data

**Relative timing not exact*

flatiron



Patient	Stage at Dx	Biomarkers	2L Treatment	**Progression**	Date of Death
Jane Doe	II	EGFR-, ALK-, PD-L1-	nivolumab	**2017-03-08**	2017-04-12

Undergoes surgery for early-stage disease

Tested for EGFR and ALK

Progresses on 1L, tested for PD-L1 and re-tested for EGFR

Diagnosed with Stage II NSCLC

Develops metastatic disease

Death

Abstraction Details

> Abstracted by Sue Smith on 4/30/17 at 10:10am
> Physician notes and scan interpretation reviewed
> Medical record from West Florida Cancer Clinic

Quality of Progression abstraction
==========

> Completeness: 99%
> Sue Smith is 96% accurate at last testing
> Inter-abstractor agreement: 97%
> Kappa: 0.93

> Audit trail for any changes
> Dataset freeze and storage

Patient	Stage at Dx	Biomarkers	2L Treatment	Progression	Date of Death
Jane Doe	II	EGFR-, ALK-, PD-L1-	nivolumab	2017-03-08	2017-04-12

flatiron

Clinical Pharmacology & Therapeutics

Explore this journal >

Open Access Creative Commons

Development

Harnessing the Power of Real-World Evidence (RWE): A Checklist to Ensure Regulatory-Grade Data Quality

Rebecca A. Miksad, Amy P. Abernethy

First published: 6 December 2017 Full publication history

DOI: 10.1002/cpt.946 View/save citation

Cited by (CrossRef): 0 articles Check for updates

Citation tools

Early View

Browse Early View Articles

Online Version of Record published before inclusion in an issue

Abstract

The role of real-world evidence (RWE) in regulatory, drug development, and healthcare decision-making is rapidly expanding. Recent advances have increased the complexity of cancer care and widened the gap between randomized clinical trial (RCT) results and the evidence needed for real-world clinical decisions.[1] Instead of remaining invisible, data from the >95% of cancer patients treated outside of clinical trials can help fill this void.

DEFINING RWE

elements.[2]

RWE QUALITY

RWE is generated from high-quality data that are 1) d from relevant RWD sources, 2) cleaned, harmonized, ed to fill in gaps, and 3) include endpoints. Quality need to encompass the entire process to generate RWE, a sources and processing to defining appropriate use gure 1).

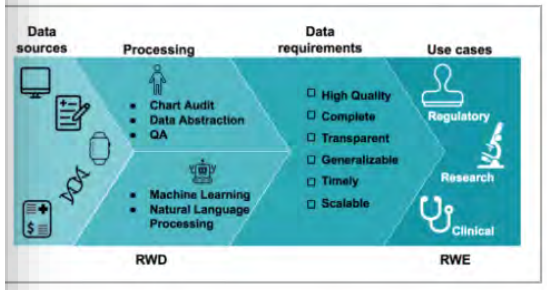


Figure 1. Open in figure viewer | Download Powerpoint slide

The journey from data to evidence. Real-world data (RWD) are data that are routinely collected in the form of electronic health records (EHRs), patient disease registries, wearables, genomic datasets, medical claims registries, and others. These data can be aggregated, linked, and processed to produce key conclusions in the form of real-world evidence (RWE). The proposed checklist can be used to assess if the quality of the RWD is regulatory-grade.

mal RWD source depends on the RWE hypothesis and .[3] As the EHR is a contemporaneous (prospective or ctive) account of the clinical narrative, it provides al details and longitudinal follow-up for outcomes. The

Meta-characteristics of RWD and RWE

Regulatory grade RWE, a potential checklist

- ☐ **Clinical Depth**
  Data granularity to enable appropriate interpretation and contextualization of patient information.
- ☐ **Completeness**
  Inclusion of both structured and unstructured information supports a thorough understanding of patient clinical experience.
- ☐ **Longitudinal Follow-up**
  Ability to review treatment history and track patient journey going forward over time.
- ☐ **Quality Monitoring**
  Systematic processes implemented to ensure data accuracy and quality.
- ☐ **Timeliness / Recency**
  Timely monitoring of treatment patterns and trends in the market to derive relevant insights.
- ☐ **Scalability**
  Efficient processing of information with data model that evolves with standard of care.
- ☐ **Generalizability**
  Representativeness of the data cohorts to the broader patient population.
- ☐ **Complete Provenance**
  Robust traceability throughout the chain of evidence.

Thank you

amy@flatiron.com

@dramyabernethy

flatiron

Appendix

Session IV: Submitting Data Documentation for Traceability and Auditing

Data documentation in the Aetion Evidence Platform

Jeremy Rassen, Sc.D.
President & Chief Science Officer
Aetion, Inc.

January 2019

The platform approach

At Aetion, we take a *platform* approach that combines:

- **Data ingestion**
- **Data storage**
- **Data measurement**
- **Analytic workflows**

This allows for testing, validation, and full **traceability and transparency.**

It also creates a **"closed system"** for documenting/archiving/auditing data transformations and provenance.

Stage 1 validation & reporting

Verify: do the loaded data match the provided data?

Part 1: rules-based “sanity checks”

- Do the imported datasets meet technical expectations?

Part 2: semi-automated validation

- Do the imported datasets meet scientific expectations?

Stage 2 reporting & versioning

As data are used, document each and every step.

<u>Part 1: archived, auditable reporting</u>

- Provide *natural language* reporting on how data are put to use in a study (e.g., data element -> measurement)

<u>Part 2: comprehensive versioning</u>

- Provide traceable versioning (provenance and history) of each measurement; taken together, becomes a full catalog of how a study came to be

“Stage 3” and beyond

Continue to document study beyond the data steps

- Epidemiological assumptions applied (eg, exposure grace period)
- Statistical methods used
- Relevant literature
- Results

Ae-ti-on

From aetiology (Greek):

The cause of diseases and disorders; the investigation or attribution of the cause or reason for something.

Session IV: Submitting Data Documentation for Traceability and Auditing

Data Documentation for Traceability and Auditing

J. Marc Overhage, MD, PhD

VP Intelligence Strategy and CMIO

January 22, 2019

Systematic Approach to Managing Big Data

Aggregate and normalize

Create and apply intelligence

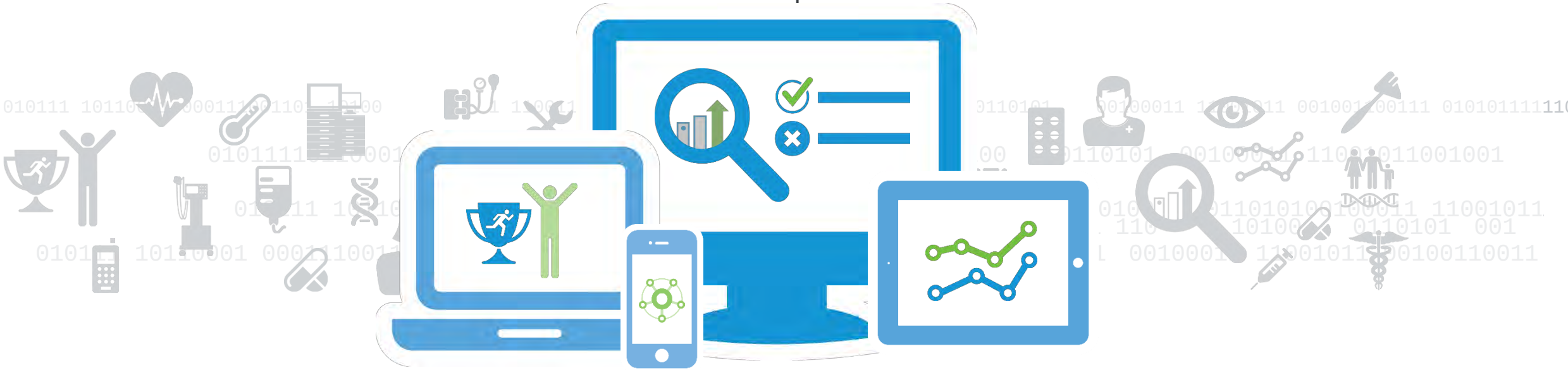
Act and measure

- Person
- Health coach
- Care manager
- Home health assistant
- Clinician
- Provider
- Data scientist
- Executive

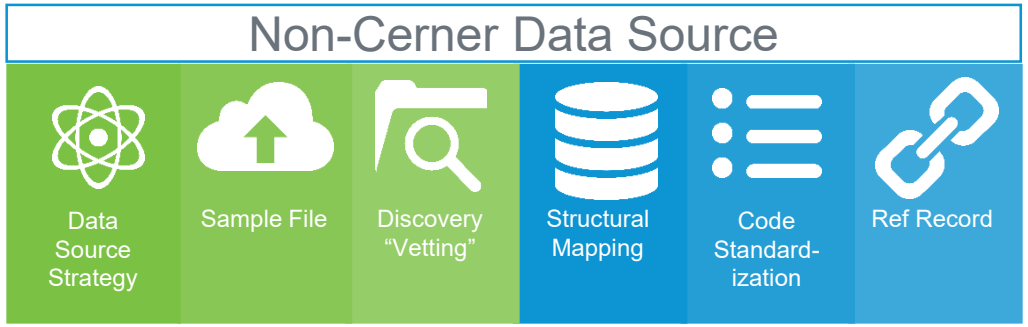
Data Integration

Data onboarding into *HealtheIntent*

- Data *sources*, data *sets*
  - Data source: A software system that sends data to HealtheIntent. This is typically a vendor (i.e. BCBS)
  - Data Set: Set of data file(s) from a Data Source that can be mapped to a data model in HealtheIntent (ie. medical claims, results, medications, demographics, allergies)
    - Many formats supported: HL7, X12, CCD, XML, CSV flat files
- File Frequency
  - how often will new data be received/extracted and uploaded to HealtheIntent

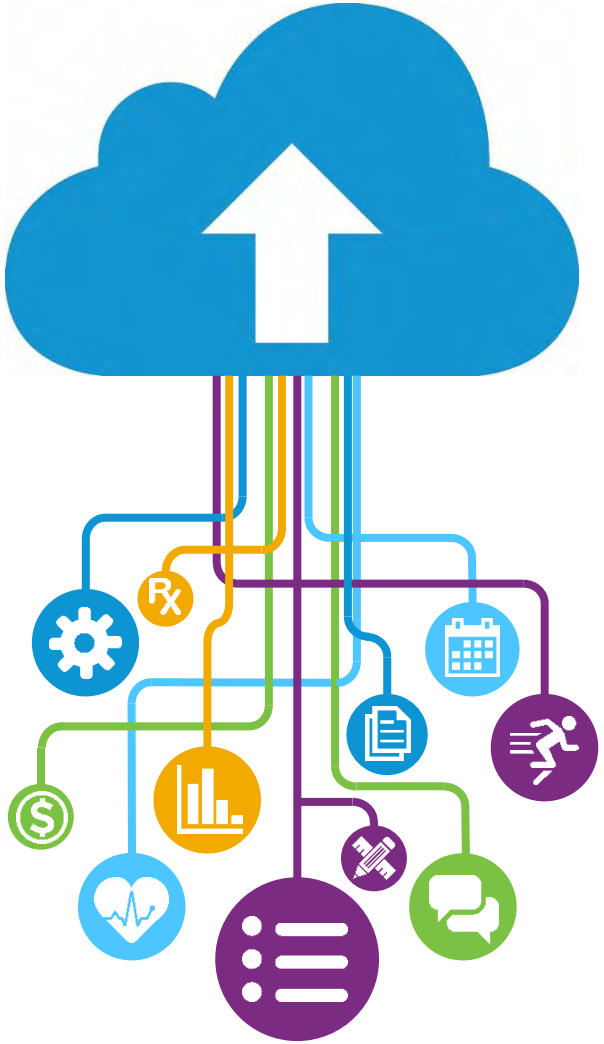


Loading Multiple Data Sources

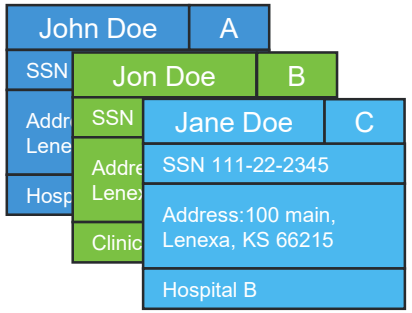


Data Vetting:
Data Vetting is the process of analyzing the raw data files for content, format, and consistency before we on-board into HealtheIntent

- This process requires collaborating sessions between Cerner, Client and Vendor and can take a few weeks to complete.



Reconcile records to a single source of truth



Identify like- reference records

SSN First name Phone
DOB Last name
Address Race Alias
Gender Ethnicity

No link | Manual | Auto link

Determine similarity score to confirm records match

EID 2468	
Record ID A	Record ID B
John Doe	Jon Doe
SN 111-22-1234	SN 111-22-1234
DOB 11/30/75	11/30/75
100 Main, Lenexa, KS	100 Main, Lenexa, KS

Assign unique EID number to linked records

Organize data into concepts

Aspirin (Multum d00170)

LOINC ICD-10
Medi-Span CPT
NDC ICD-9
MEDCIN

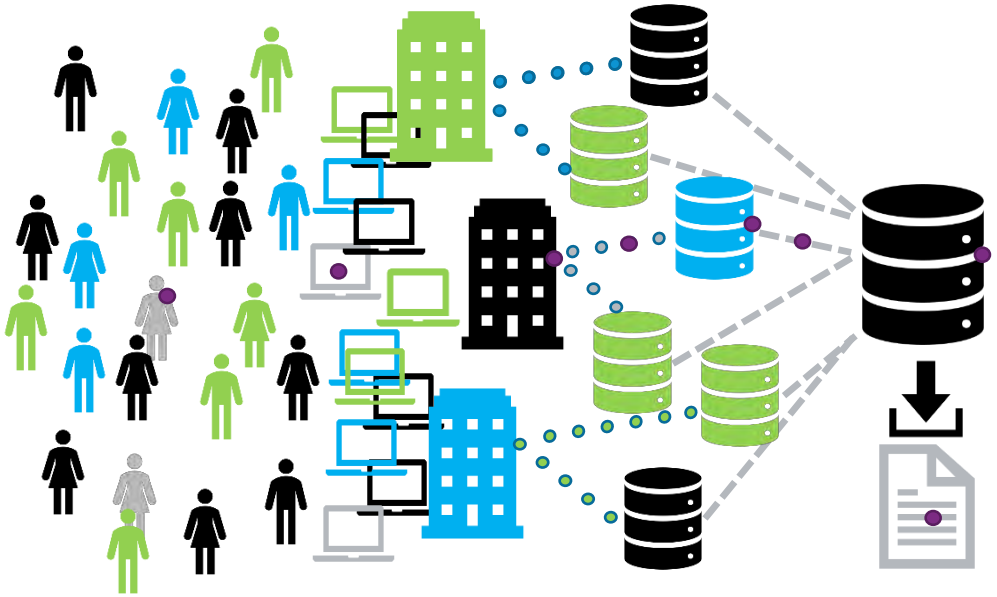
Allergies	Medications
Conditions	Procedures
Immunizations	Visits
Lab results	Vitals

Medications Most recent
Aspirin (Multum d00170) Mar 13, 2016

Medication	Date	Source
aspirin 300 mg oral delayed release tablet	3/24/2014	Westwatch Bay
aspirin 227.5 mg oral gum	10/17/2013	Baseline East
ASA 500 MG Oral Tablet [Bayer Aspirin]	9/23/2013	Westwatch Bay
Aspirin	4/23/2013	Get Well Now
aspirin	2/18/2013	Westwatch Bay
Aspirin	5/14/2012	Baseline East
aspirin 300 mg oral tablet	6/20/2011	Get Well Now

Provenance Tracking

- Provenance definition
  - According to HL7 FHIR specification, provenance is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust, and allowing reproducibility. Provenance assertions are a form of contextual metadata. Provenance indicates clinical significance in terms of confidence in authenticity, reliability, and trustworthiness, integrity, and stage in lifecycle, all of which may impact security, privacy, and trust policies.
  - Granularity of the entities – device, individual, institution
  - Documents versus data
- Provenance complexities
  - Individual
  - Institution/Organization
    - Multiple facilities
    - Multiple EHRs
    - Multiple EHR domains
    - Non-EHR systems
  - Multiple source inference
  - Aggregation entities – e.g. HIEs
  - Intermediaries and networks



Session IV: Submitting Data Documentation for Traceability and Auditing

Closing Remarks

Unpacking Real-World Data Curation: Principles and Best Practices to Support Transparency and Quality

Duke-Robert J. Margolis, MD, Center for Health Policy
1201 Pennsylvania Ave, NW, Suite 500, Washington, DC 20004
January 22, 2019