Sentinel

ICPE 2022 Symposium

Data Harmonization, Standardization, and Quality Assessment in Distributed Health Data Networks: Lessons from Around the World

Presented at the 38th International Conference on Pharmacoepidemiology & Therapeutic Risk Management

Sentinel

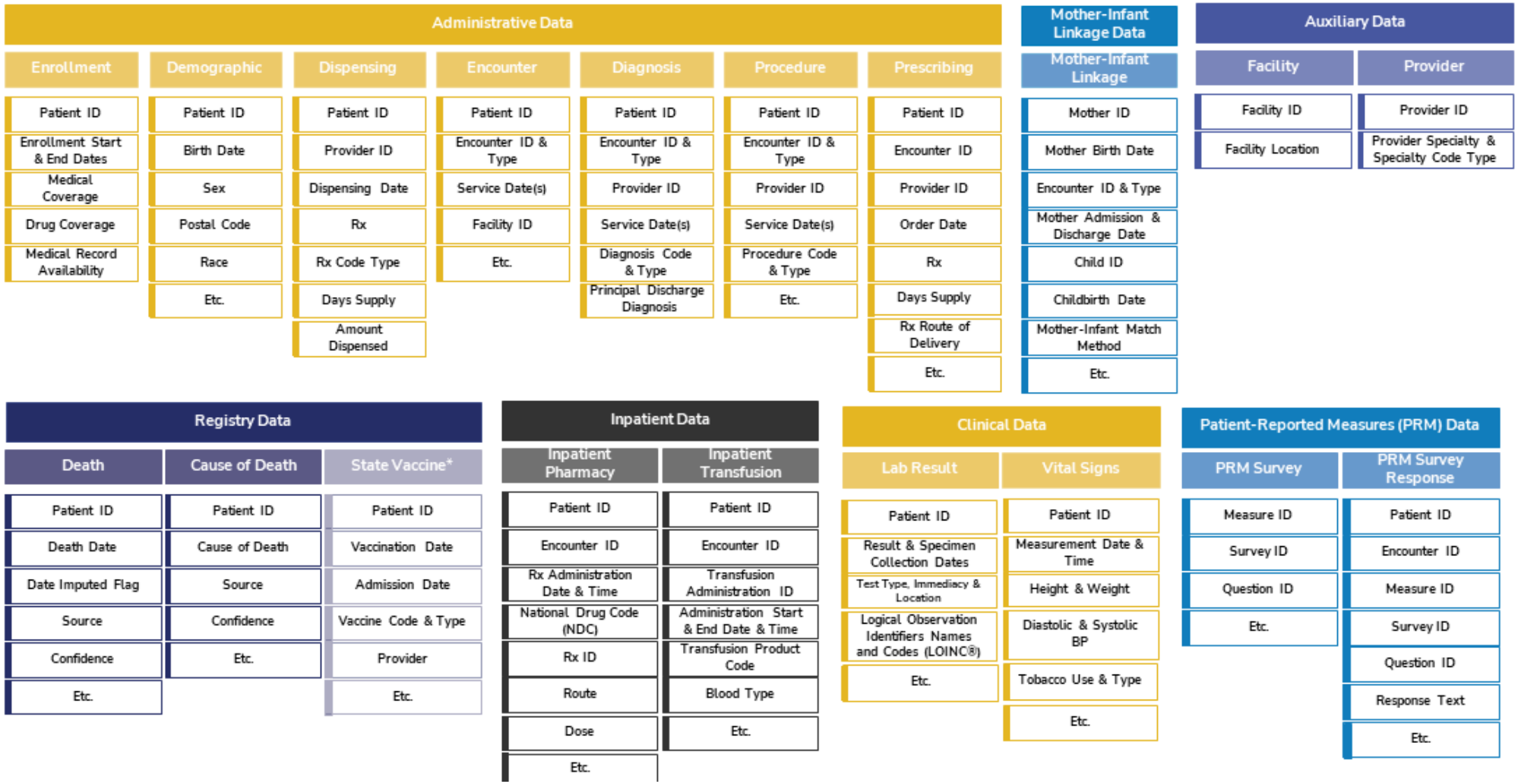
Sentinel's Data Curation Experience

38th International Conference on Pharmacoepidemiology & Therapeutic Risk Management

Judith C. Maro, PhD

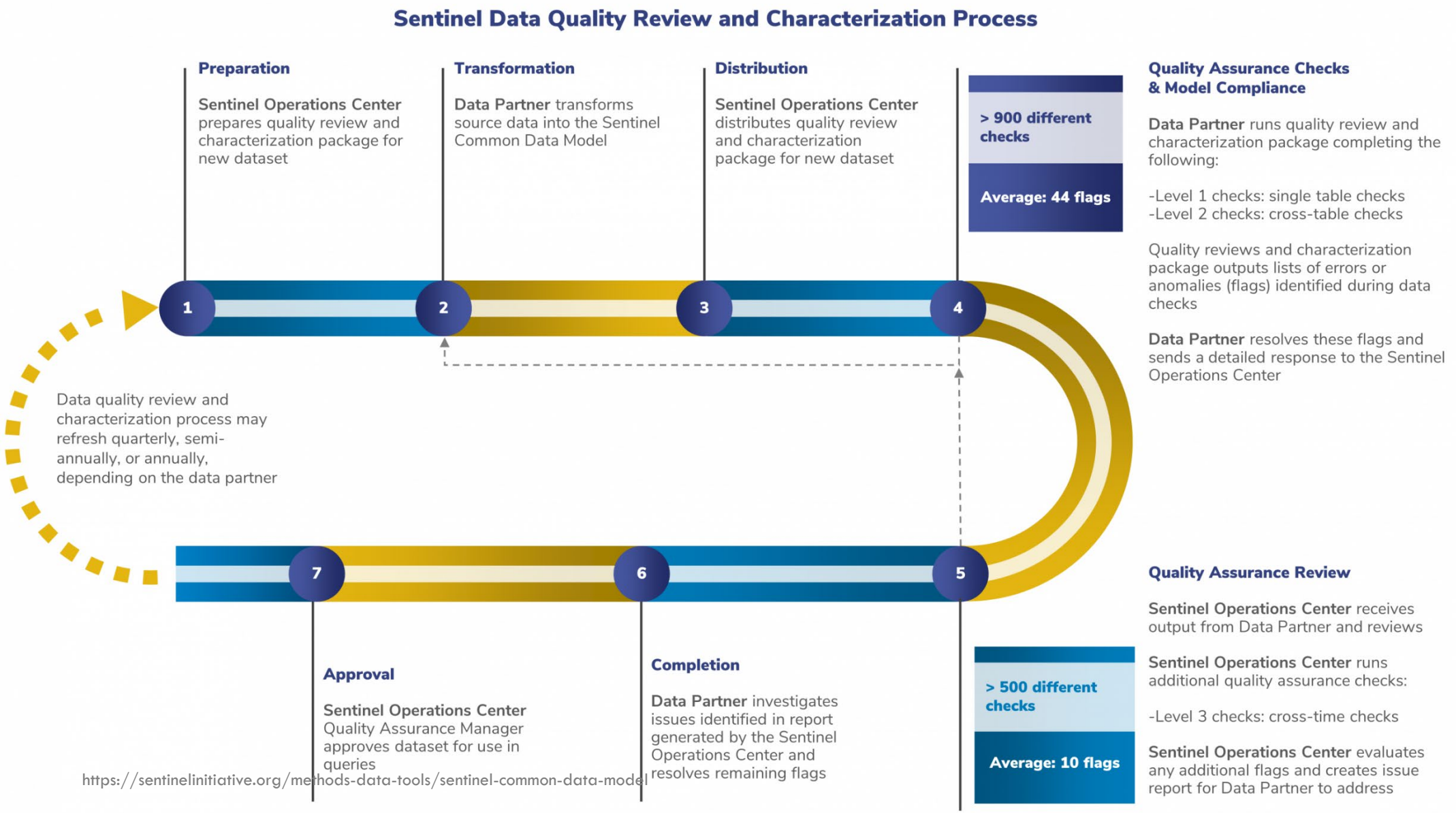
August 26, 2022

Sentinel Common Data Model – 12 Years of Curation



https://sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

Checking every Data Delivery from each Data Partner



https://sentinelinitiative.org/methods-data-tools/sentinel-common-data-model

Types of Data Quality Checks and Examples

Level 1 Checks:
Single table checks

- ✓ **Completeness**
  Admission date is not missing value
- ✓ **Validity**
  Admission date is in date format

Level 2 Checks:
Cross-table checks

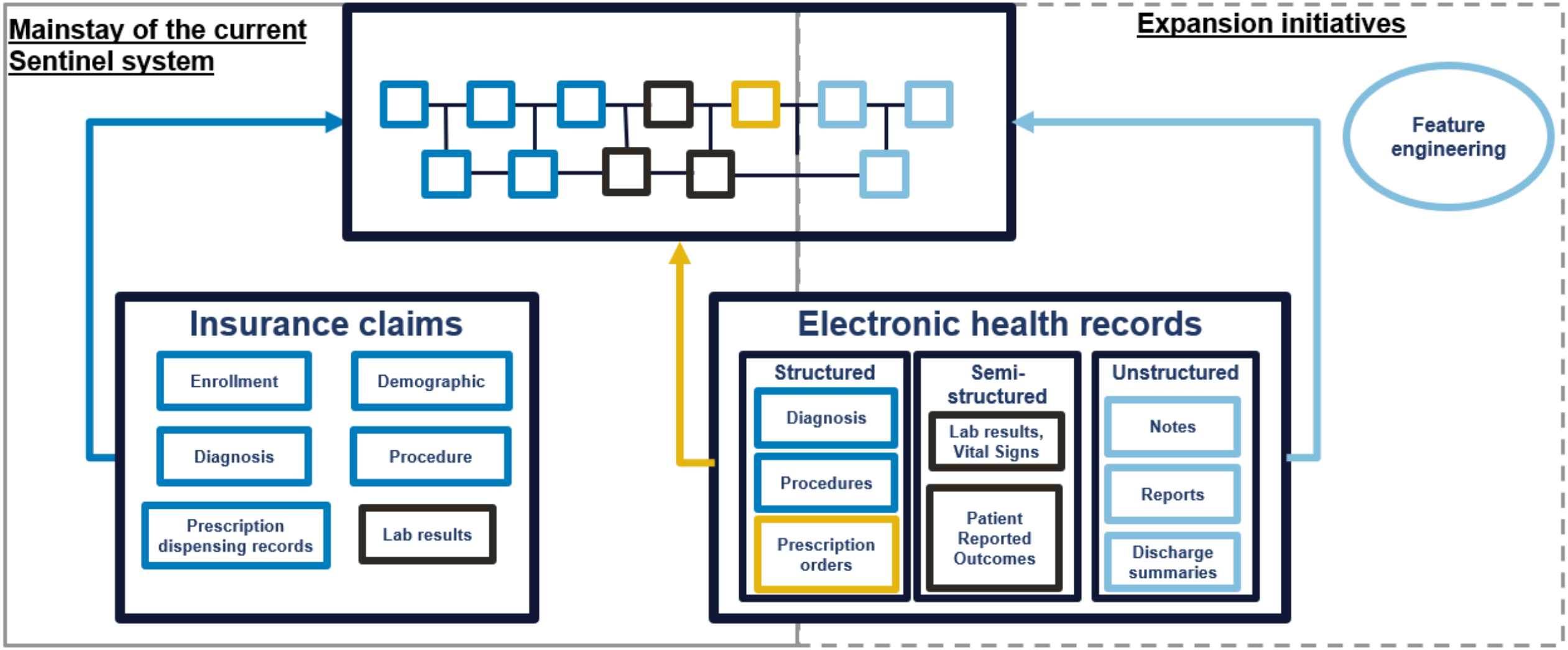
- ✓ **Accuracy**
  Admission date occurs before the patient's discharge
- ✓ **Integrity**
  Admission date occurs within the patient's active enrollment period

Level 3 Checks:
Cross-time checks

- ✓ **Consistency of Trends**
  There is no sizable percent change in admission date record counts by month-year

https://sentinelinitiative.org/about/how-sentinel-gets-its-data#data-quality-review-and-characterization-in-the-sentinel-operations-center

EHR Data in the Sentinel Common Data Model



Adapting the Data Quality Metrics for International Work

Easiest Changes:

- Adding Standard Coding Libraries (Drug Information Number for Canada, Dictionary of Medicines and Devices for the United Kingdom, Anatomic Therapeutic Classification, etc.) used in ex-US settings

Harder Changes:

- Modifying Postal Code/Regional Demographic checks to be accurate
- Imputing Days Supply of Pharmacy Dispensings

Hardest Changes:

- Modifying Race, Ethnicity Categories to be geographically appropriate

FDA to develop a $100M Real World Evidence Medical Data Enterprise with 10M additional EHR lives

- The Sentinel Innovation Center was tasked with finding the best way to use data from the 10M additional EHR lives along with the best source of the additional 10M EHR lives

- We did some preliminary data quality assessment in Standalone EHR data (i.e., not linked to claims information) that was not necessarily structured into a common data model.

1. Gottlieb, S. (2018, June 10). FDA budget matters: Notes on data and Real World Evidence. U.S. Food and Drug Administration. https://www.fda.gov/news-events/fda-voices/fda-budget-matters-cross-cutting-data-enterprise-real-world-evidence

Data Quality Metrics for Standalone EHR data

- Protocol is publicly available at:

  https://www.sentinelinitiative.org/sites/default/files/documents/Sentinel_Data_Quality_Metrics_for_Electronic_Health_Records_v1.0_public_final.pdf

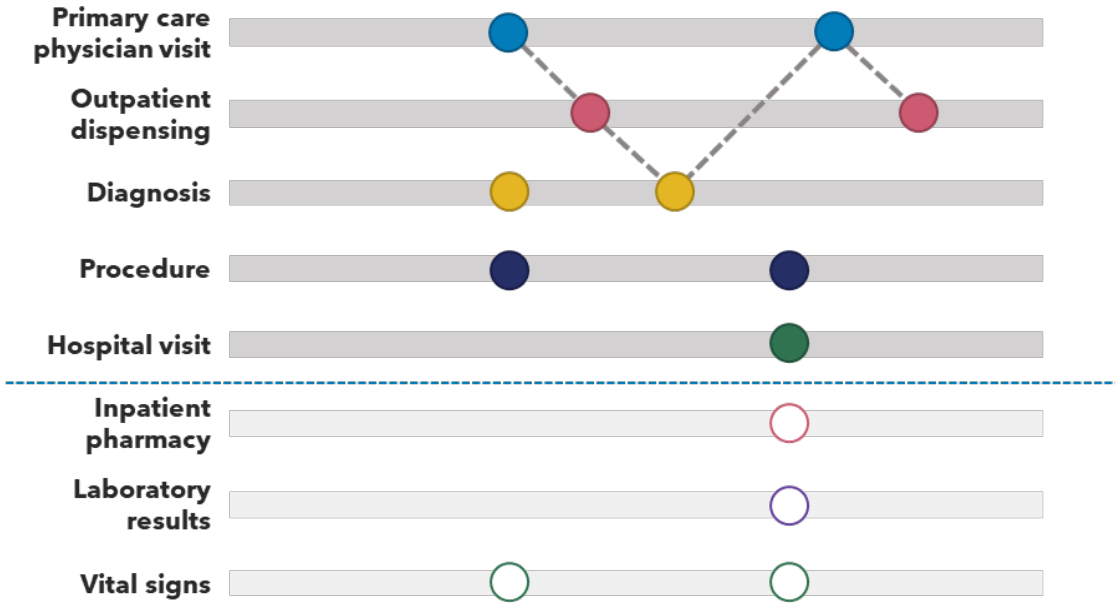
New Challenges with EHR Data for Sentinel

1. Missingness from EHR "leakage" when care is sought outside of the system
2. Important clinical information still kept in unstructured or semi-structured free-text fields
3. Inconsistent data standardization and integration
    - Data harmonization across different EHRs or even same EHR at different sites may be resource-intensive

Capturing Patient Experience

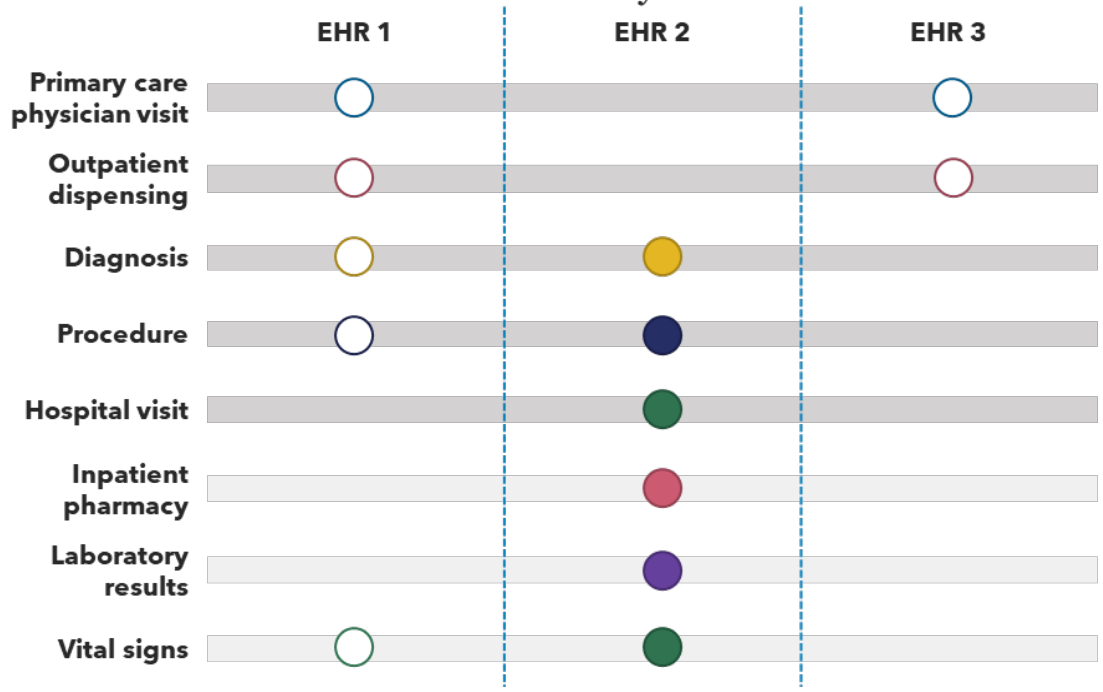
Claims Data

- Comprehensive data across all encounters & billable settings
- Misses some clinical detail



Electronic Healthcare Data

- Detailed data within an encounter
- Misses encounters outside health system



Filled circles indicate captured data; Open circles indicate missing data; Figures adapted from Sentinel Operations Center and FDA collaborations

EHR Data Quality Metrics for Standalone EHR Data

Measure	Description
Missing demographic elements	Missingness of select demographic variables at single point in time
Utilization in 65+	Facts by year interval in 65-year + population
Infant utilization	Facts in newborns measured in first 6, 12, and 24 months of life
Counts by diagnosis code	Fact counts by first-three characters of diagnosis count and year
Invalid age	Fact counts per year and invalid category (<0 years, >120 years)
Post-death events	Facts and birth counts following death, by year of death
Fact types by year	Fact counts by year and type
Missing encounter data	Missingness of admission and discharge dates by encounter setting
Post-discharge facts	Fact counts in intervals (3-, 7-, 30-, 90-days) following inpatient discharge
Encounter attributes by year	Basic stats by encounter month-year and setting, e.g., patient counts, distinct encounters, fact type counts

Note: Yellow-boxed metrics are distributional in nature whereas non-boxed metrics are more about identifying data errors and measuring gaps in data longitudinality.

EHR Data Source - TriNetX

Quality metrics were tested using TriNetX data to help inform protocol refinement.

- 71 U.S.-based Health Care Organizations (HCOs)*
- 5 calendar years (2015 to 2019)
- 69.2M distinct patients with demographic data

- Medical Facts:
  - 8.6B laboratory records
  - 6.6B medication orders
  - 4.6B vital measurements
  - 3.6B diagnosis records
  - 2.1B procedure records
- Encounters:
  - 2.1B Total Encounters
  - 1.4B Ambulatory
  - 57.5M Emergency Department
  - 50.8M Inpatient

* All 71 HCOs responded with data for at least two metrics. If an HCO did not populate relevant data for a metric, they are excluded from reporting. These 71 US-based HCOs are a subset of the overall TriNetX data. For additional information on this source, see https://trinetx.com/

Medical Facts Following Inpatient Discharge (Metric 4.2)

Assumption: Following an inpatient stay, a patient should have follow-up care in the next 90 days.

Percent of Patients with Procedures or Diagnoses Following Inpatient Discharge								
	1-3 Days Following		1-7 Days Following		1-30 Days Following		1-90 Days Following	
	Outpatient	Inpatient/ Outpatient	Outpatient	Inpatient/ Outpatient	Outpatient	Inpatient/ Outpatient	Outpatient	Inpatient/ Outpatient
Average	2.6%	8.0%	2.8%	10.8%	3.1%	13.2%	3.2%	13.9%
Median	1.1%	0.8%	1.3%	0.9%	1.5%	1.0%	1.6%	1.1%
Min	0%	0%	0%	0%	0%	0%	0%	0%
Max	8.1%	91.6%	8.7%	91.6%	9.3%	95.1%	9.5%	97.6%

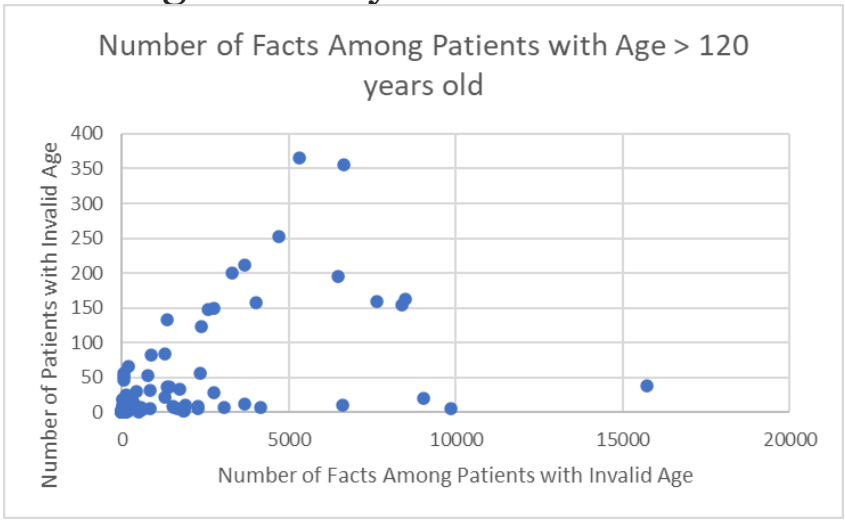
- Percent of patients with procedures or diagnoses following inpatient discharge increases with evaluation window
- Comparison between HCOs by characteristic is limited by distribution and outliers
  - Outpatient (N=4) vs. Inpatient/Outpatient (N=58)

Patients with "Invalid" Age >120 years (Metric 3.5)

Assumption: No one lives to be over 120 years old.

- 46/70 HCOs have at least 1 year in which they have a patient aged >120 years

Count of Patients with Invalid Age in 2019	
	>= 120 years
Average	17.3
Median	1
Min	0
Max	366



- Minor number of data anomalies for total volume; HCO with 366 Patients >120 y.o. had 82,910 patients >65 in the same year.
- Across all Year Groups, the average number of facts per 120+ y.o. person was 90 facts, but median was 11, suggesting average is driven by outlier HCOs.
- In one medium-sized, urban general HCO, the number of medical facts among patients 120+ y.o. ballooned (x20-40 more) with the addition of NLP/linkage.

Final Reflections

- Adapting the Sentinel Common Data Model to incorporate longitudinal EHR data, even international EHR, not especially difficult

- Assessing data quality in standalone EHR data differed from administrative claims data in different ways
  - More attention must be paid to documenting EHR "leakage," the biggest risk for longitudinal work in US-based databases
  - Units of analysis (i.e., medical facts vs. encounters)
  - Various standards for aggregation of the information

- To operationalize to high-quality surveillance systems like Sentinel, additional structure would likely need to be created to understand what is "normal" for EHRs. With so much missingness, defining a benchmark becomes very difficult.

- Although NLP techniques may improve capture, it's important to ascertain where the additional capture is manifesting.

Discussion

This slide presentation represents the work of many, many contributors in the Sentinel System.

Backup