Common Data Models Video 2 – Final Script

In an ideal world, healthcare data would flow seamlessly.

Data would be captured accurately and in standardized formats at every point of care—whether in a clinic, hospital, pharmacy or even at home.

These data would move effortlessly between systems.

In the first video, we saw how efforts like USCDI, FHIR, and TEFCA are improving interoperability—making healthcare data more accessible and connected. But even if data flow seamlessly, challenges remain when using them for research, population health, surveillance, and quality improvement. --- Unlike clinical care or billing, these use cases require structured data that can be meaningfully compared across sources. In this video, we'll explore these challenges and the solutions that make healthcare data more useful beyond patient care.

Imagine this:

Clinicians and patients make decisions with complete, reliable information.

Billing happens automatically and without errors.

Researchers quickly uncover answers to critical questions about treatments and outcomes.

This is the ideal state. But achieving it isn't easy.

Two major barriers stand in the way of this vision.

The first challenge is the sheer volume and complexity of healthcare data. Healthcare data come from many sources:

- Primary care visits,
- Laboratories,
- Hospital stays,
- Pharmacy records,
- And even at-home monitoring devices.

Each of these systems collects similar types of information—like test results, medications, and clinical notes—but they record them differently. For example, one hospital might call a test result "cholesterol," while another labels it "lipid levels." Without a common structure,

these data are hard to share or compare across systems. ---Even more challenging, these systems don't always communicate with each other, leaving researchers and clinicians to deal with fragmented or incompatible data.

The second challenge is misaligned incentives, which limit the adoption of standardized, high-quality data practices.

Clinicians face overwhelming workloads, with more patients to care for and less time to document. Under these time constraints, they focus on recording data that are tied to billing, like procedures and services, while leaving out other important details—such as social or environmental factors—that aren't reimbursed.

This creates several issues:

Data needed for research or quality improvement are often incomplete.

- The data that are recorded may reflect financial priorities rather than clinical needs.
- And there's little incentive for healthcare providers to adopt better data standards.

For instance, a hospital may meticulously document procedures to maximize reimbursement but fail to capture a patient's housing or food insecurity—factors that could significantly impact health outcomes.

Together, these two barriers—data complexity and misaligned incentives— result in significant variability in how healthcare data are captured and stored, making it challenging to use them effectively for patient care, population health, or research.

To address the challenges of data variation, Common Data Models, or CDMs, provide a solution.

A CDM is like a universal blueprint for healthcare data. Converting data into a CDM organizes and standardizes the data, ensuring it looks and behaves the same way across different systems.

Think of it as a bridge—helping us move from the messy, fragmented reality of healthcare data to something structured and research-ready.

As healthcare research has evolved, multiple Common Data Models have been developed, each designed for specific purposes. Some focus on drug safety, while others are optimized for comparative effectiveness research or observational studies.

Three of the more widely used CDMs are: Sentinel, OMOP, and PCORnet.

Sentinel focuses on monitoring the safety of drugs and medical devices after they've reached the market. It relies heavily on insurance claims data, which are generated for billing purposes. These data include diagnoses, procedures, and the services provided to justify payments.

Because Sentinel retains data in their original structure, it offers flexibility. Researchers can adapt their analyses to different needs, but this flexibility comes at a cost.

With data so close to their original form, running analyses often requires more sophisticated queries and advanced data processing.

OMOP supports large observational studies and evaluates drug safety. It's designed to handle data from many sources, including EHRs, insurance claims, and clinical registries.

OMOP transforms raw data into highly standardized formats, making them easier to compare and analyze. For example, diagnoses are converted into SNOMED codes, which ensures consistency across studies.

This strict standardization can sometimes oversimplify the data, resulting in false precision or loss of nuanced information.

PCORnet is optimized for studies comparing different treatments or approaches to patient care. It relies on data from EHRs but can also incorporate insurance claims data.

PCORnet keeps its data close to the source, which allows researchers to link back to patient records. This is particularly useful for clinical trial recruitment.

Because PCORnet prioritizes flexibility over standardization, running queries can sometimes be more challenging.

But why not use a single CDM for everything? The answer lies in the unique needs of these different applications.

For example, Sentinel focuses on insurance claims data that are useful for post-market drug surveillance, while PCORnet prioritizes patient-centered research and maintains links to clinical trial recruitment and OMOP prioritizes computational efficiency. Each CDM was

built for a specific purpose, and while there's overlap, no single model meets all needs equally well.

Each of these CDMs was built for a specific purpose. While they have different strengths and challenges, together, they help researchers make sense of healthcare data and answer critical questions.

Here's how a Common Data Model works in practice. Imagine we're comparing two hypertension treatments using data from three hospitals with different EHR systems.

The first step in this process is data transformation. Raw data—like labs, medications, and imaging results—must be standardized into the CDM format. For example, one hospital might label a cholesterol test as "total cholesterol," while another calls it "lipid levels." To ensure compatibility across systems, these are mapped to a common standard, such as a LOINC code.

Next comes data curation. This step ensures the data are accurate and complete. Researchers review the data for inconsistencies, validate codes, and flag missing information. For instance, they might confirm that medication codes are correct or that lab results fall within realistic ranges. Curation is critical because research findings depend on high-quality data.

Finally, the transformed and curated data are grouped into computable phenotypes. These are standardized definitions that represent complex clinical conditions.

These phenotypes ensure consistency across datasets and help identify eligible patients for the study.

For hypertension, a phenotype might include ICD-10 codes for the diagnosis, prescriptions for antihypertensive medications, and blood pressure readings consistently above 140/90.

Through this process of transformation, curation, and phenotyping, CDMs organize fragmented healthcare data into a usable format for research and analysis.

CDMs aren't one-size-fits-all. Researchers often need to add custom elements—called sidecars—to capture unique data for a study.

For example:

A study on asthma might use sidecars to include environmental data, like air pollution levels.

A clinical trial could collect patient-reported outcomes using a phone app, storing this data separately but linking it back to the CDM.

This flexibility allows CDMs to adapt to specialized needs without requiring major changes to the core structure.

Instead of standardizing to one CDM, the focus is on ensuring consistency in base data. When foundational data are well-organized, transitioning between models is straightforward.

Mapping data to a CDM often requires hiring specialized engineers and analysts to handle the complexity of data transformation.

Smaller clinics may struggle with the cost of implementation as they don't participate in enough research to justify the expense.

Training IT staff and clinicians to standardize and maintain data correctly can take significant time and resources.

Once data are standardized, they can be reused across multiple studies, saving time and effort.

CDMs provide tools and analyses that can be reused across studies.

Ultimately, CDMs enable faster, more accurate research that directly benefits patient care.

CDMs are essential for turning fragmented healthcare data into something we can use.

While implementing and maintaining a CDM requires resources, the payoff is worth it for large organizations.

Emerging standards for the exchange of data between health systems like FHIR and USCDI provide a foundational set of data elements and formats, potentially lowering the barrier of entry for smaller institutions. These tools could standardize data earlier in the process, making it easier to work with.

But for now, CDMs remain one of the best tools for unlocking the full potential of healthcare data.