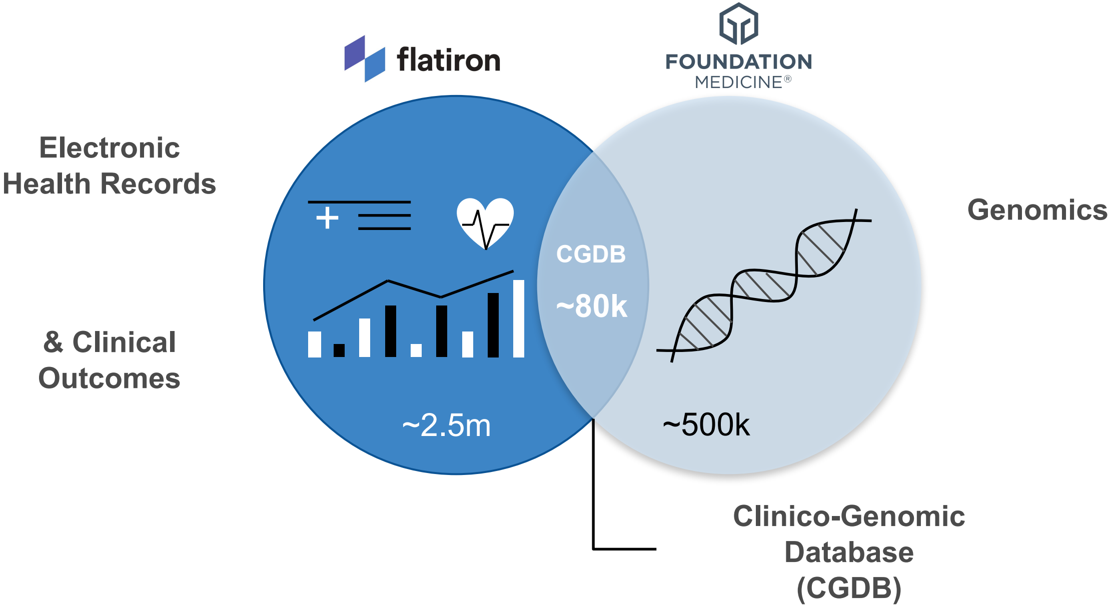
Introduction to the Clinico-Genomic Database

Sarah McGough, PhD
Senior Data Scientist, Real-World Data | Roche

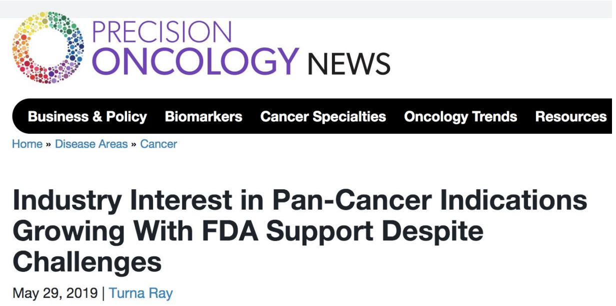
The Clinico-Genomic Database links Flatiron **electronic health records** with Foundation Medicine (FMI) **comprehensive genomic profiling** for tens of thousands of cancer patients in the U.S.



How might the CGDB used for health research?

Cancer biology is enormously **complex** and tied to the **human genome**. With the CGDB, we can:

- Understand **prevalence of key cancer biomarkers** in the real-world patient population
- Inform **biomarker-targeted therapy** options & outcomes for patients
- Identify **genomic drivers** of outcomes (e.g. prognositic biomarkers)
- Contribute to **tumor-agnostic** “pan-tumor” research



PRECISION ONCOLOGY NEWS

Industry Interest in Pan-Cancer Indications Growing With FDA Support Despite Challenges

May 29, 2019 | Turna Ray

FDA NEWS RELEASE

FDA approves third oncology drug that targets a key genetic driver of cancer, rather than a specific type of tumor

The linkage of information from multiple diverse sources in the CGDB gives rise to missing data challenges.

Generally missing data

Observations are missing without a clear structure or group dependency

(not to be confused with MCAR- missing completely at random!)

Structured or “block” missing data

Observations are missing across “blocks” of disparate information sets → Represents a special type of heterogeneity in the data

	Block 1	Block 2	Block 3
X1	🟥	🟥	🟥
X2	NA	🟨	🟨
X3	🟩	🟩	NA
X4	NA	NA	🟦

Data might be purely **unmeasured** by block,

or

not possible to measure by block

The CGDB contains comprehensive genomic profiling from multiple genomic tests, each of which targets a different set of genes.

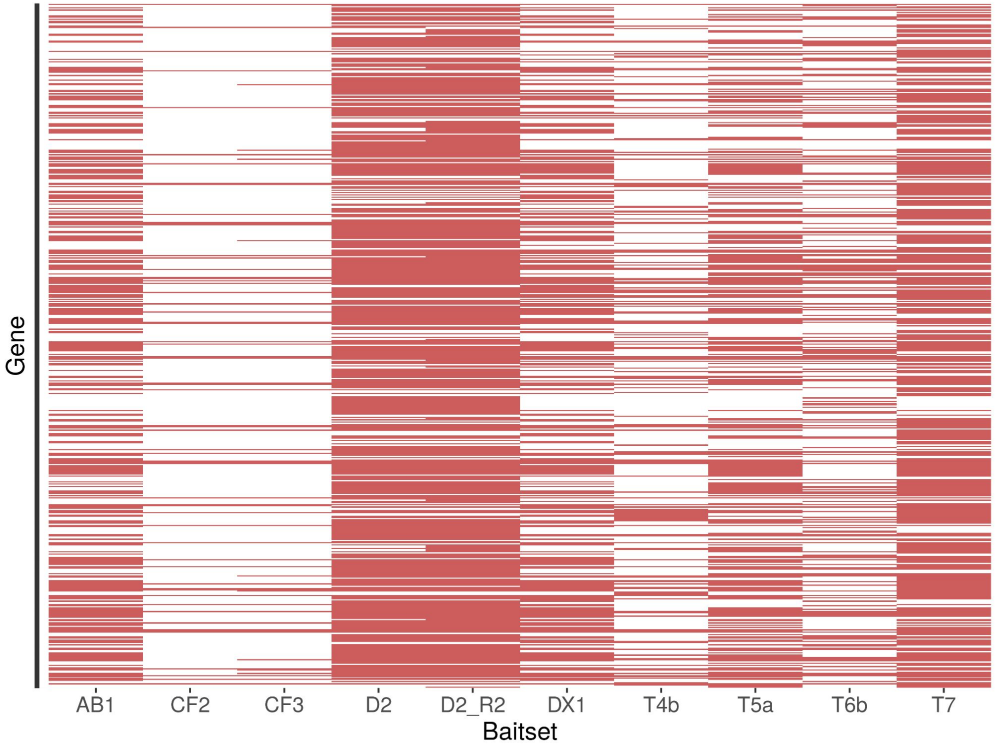
Each patient usually receives **1 test**.

Of **596 unique genes** measured in the CGDB, only **30** are measured commonly across all tests.

- Tests are ordered to *target* specific treatment, prognosis, and disease progression goals
- Tests *evolve* over time
- *Solid* tissue vs. *liquid* biopsy

Here, genes are **"block missing"** by test type.

Measured genes by test (baitset).

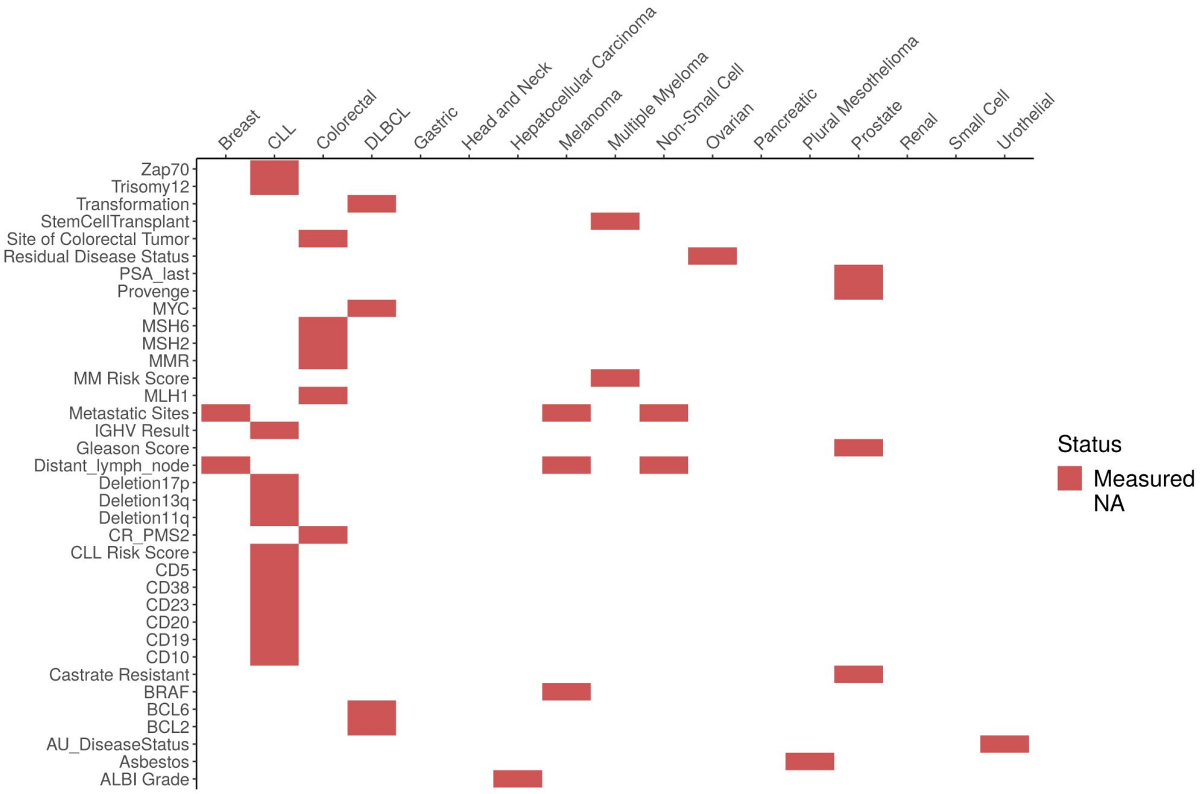


The CGDB combines information across dozens of cancer types, but some information is cancer-specific.

Each cancer type collects cancer-specific information, such as the **Gleason Score** for Prostate Cancer patients or **Stem Cell Transplant** for DLBCL patients

Here, variables are **“block missing”** by cancer type.

But, imputation is not appropriate.



Amassing data in large volumes from multiple sources presents an amazing precision medicine opportunity- but also a structured missing data problem.

We must address structured missing data anytime that we hope to use the CGDB in its totality, including:

- Making insights across **cancer types** → tumor-agnostic “pan-tumor” research
- Studying **genomic features** across patients → oncology biomarker research

More generally: this challenge will rise with the increasing development of multi-modal data sources.

What sorts of solutions are available to us? Can we influence how researchers and industry work with these kinds of data?

Doing now what patients need next