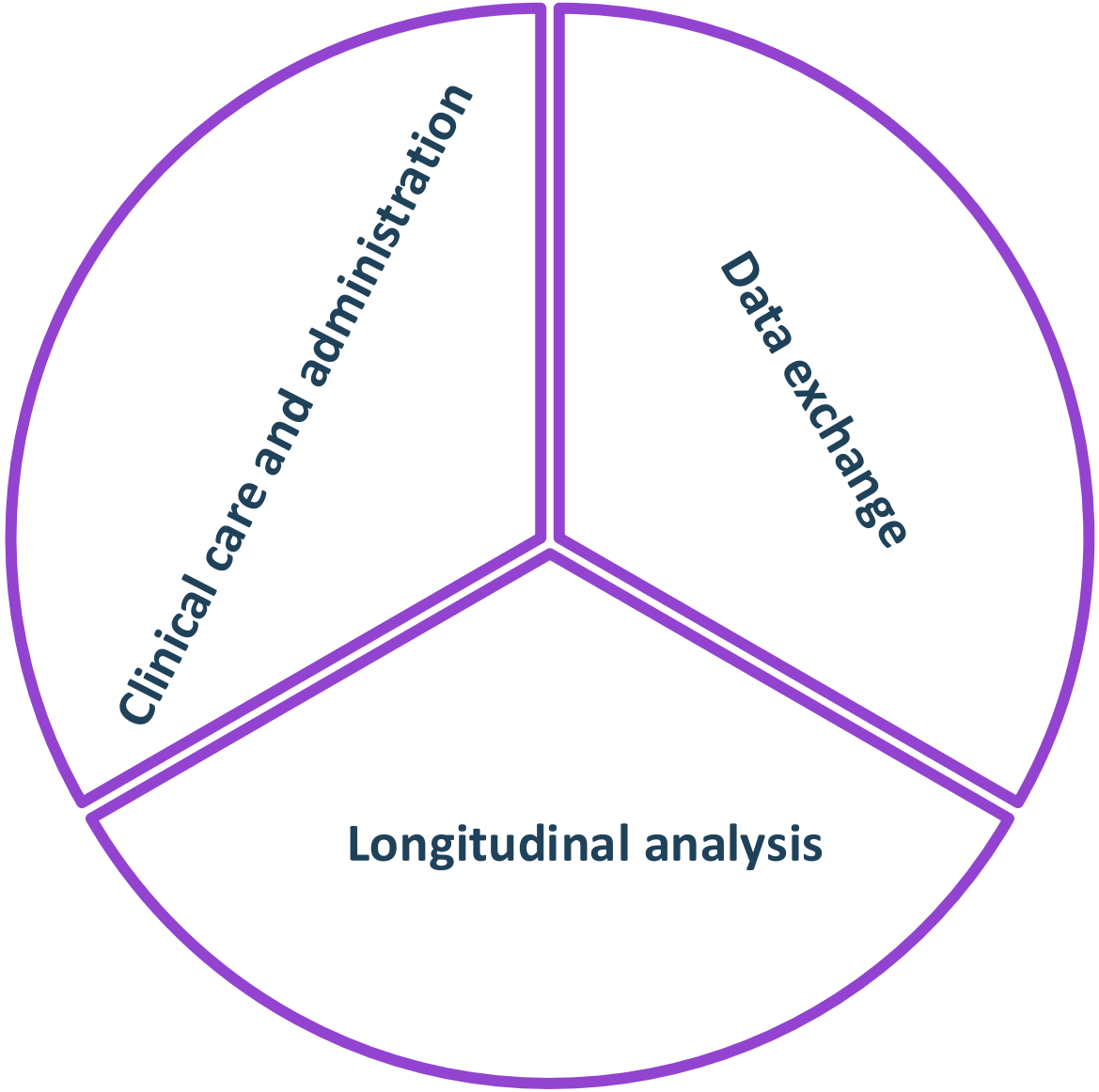
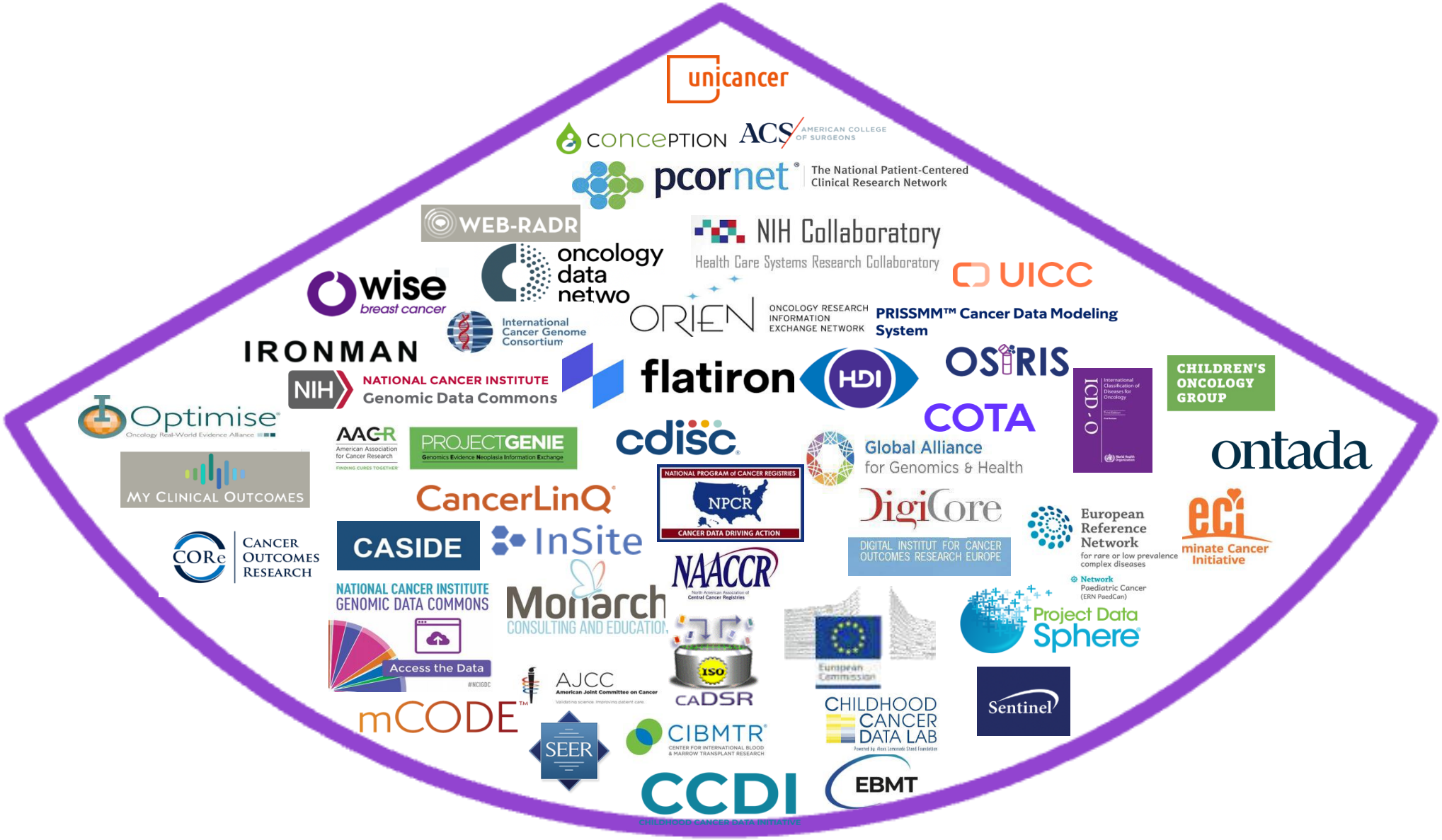
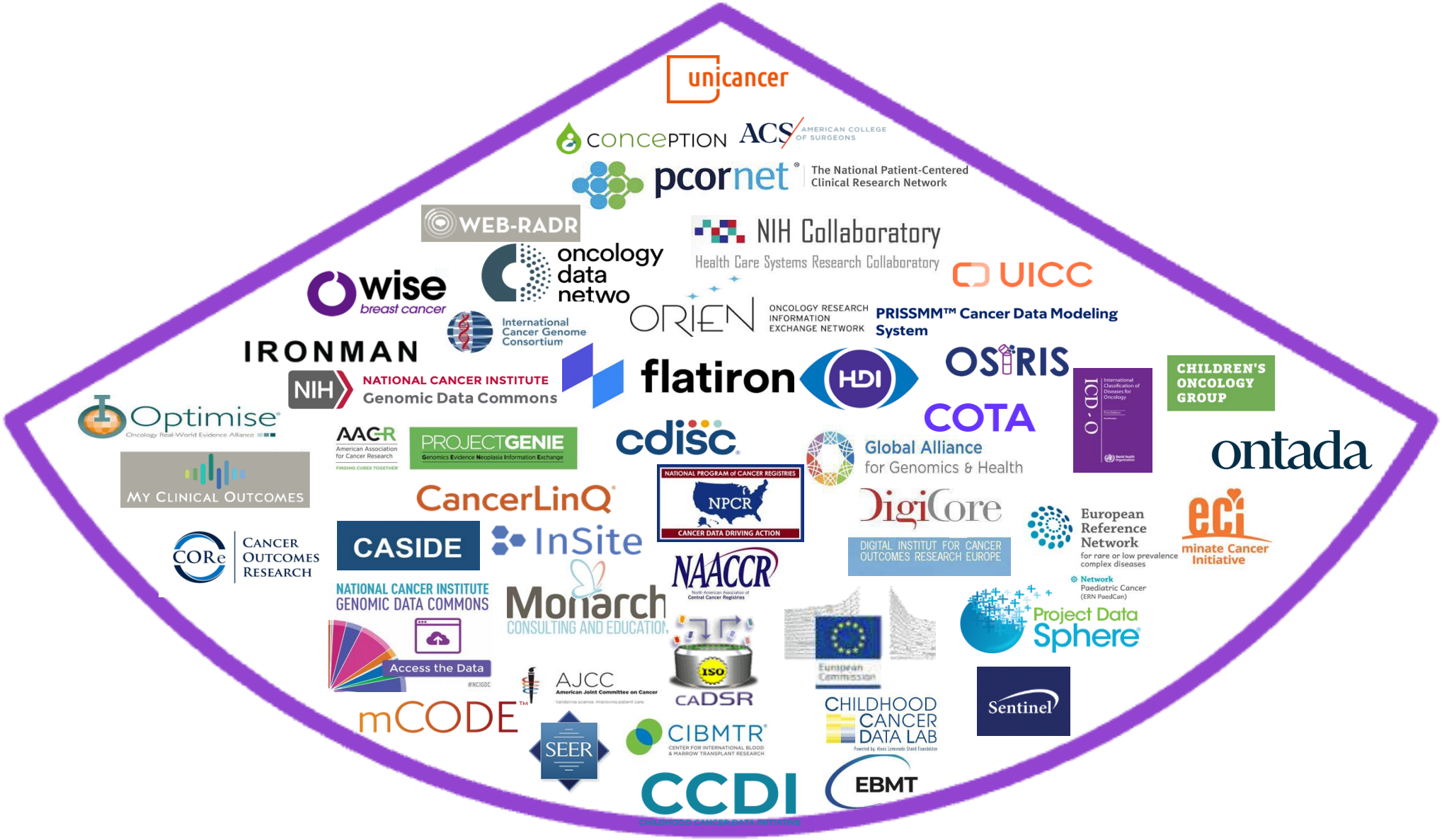
Genomic Data in a Closed World Environment

Asieh Golozar, MD PhD MPH MHS


Clinical care and administration
Data exchange
Longitudinal analysis


unicancer
conception
ACS AMERICAN COLLEGE OF SURGEONS
pcornet
The National Patient-Centered Clinical Research Network
WEB-RADR
NIH Collaboratory
Health Care Systems Research Collaboratory
UICC
oncology data netwo
wise breast cancer
ORIEN
ONCOLOGY RESEARCH INFORMATION EXCHANGE NETWORK
PRISSMM™ Cancer Data Modeling System
International Cancer Genome Consortium
IRONMAN
NIH NATIONAL CANCER INSTITUTE Genomic Data Commons
flatiron
HDI
OSIRIS
CHILDREN'S ONCOLOGY GROUP
Optimise
Oncology Real-World Evidence Alliance
COTA
ICD-O
AACR American Association for Cancer Research
PROJECTGENIE Genomics Evidence Neoplasia Information Exchange
cdisc
Global Alliance for Genomics & Health
ontada
MY CLINICAL OUTCOMES
NATIONAL PROGRAM of CANCER REGISTRIES NPCR CANCER DATA DRIVING ACTION
CancerLinQ
DigiCore
DIGITAL INSTITUT FOR CANCER OUTCOMES RESEARCH EUROPE
European Reference Network for rare or low prevalence complex diseases
eCi minate Cancer Initiative
CORe CANCER OUTCOMES RESEARCH
CASIDE
InSite
NAACCR
Network Paediatric Cancer (ERN PaedCan)
NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS
Monarch CONSULTING AND EDUCATION
Project Data Sphere
Access the Data
ISO
caDSR
European Commission
AJCC American Joint Committee on Cancer
mCODE™
SEER
CIBMTR CENTER FOR INTERNATIONAL BLOOD & MARROW TRANSPLANT RESEARCH
CHILDHOOD CANCER DATA LAB
Sentinel
CCDI CHILDHOOD CANCER DATA INITIATIVE
EBMT

https://xkcd.com/927/


unicancer
conception
ACS AMERICAN COLLEGE OF SURGEONS
pcornet
The National Patient-Centered Clinical Research Network
WEB-RADR
NIH Collaboratory
Health Care Systems Research Collaboratory
oncology data netwo
wise breast cancer
UICC
ORIEN
ONCOLOGY RESEARCH INFORMATION EXCHANGE NETWORK
PRISSMM™ Cancer Data Modeling System
International Cancer Genome Consortium
IRONMAN
NIH NATIONAL CANCER INSTITUTE Genomic Data Commons
flatiron
HDI
OSIRIS
CHILDREN'S ONCOLOGY GROUP
Optimise
Oncology Real-World Evidence Alliance
COTA
ICD-O
AACR American Association for Cancer Research
PROJECTGENIE Genomics Evidence Neoplasia Information Exchange
cdisc
Global Alliance for Genomics & Health
ontada
MY CLINICAL OUTCOMES
NATIONAL PROGRAM of CANCER REGISTRIES NPCR CANCER DATA DRIVING ACTION
CancerLinQ
DigiCore
DIGITAL INSTITUT FOR CANCER OUTCOMES RESEARCH EUROPE
European Reference Network for rare or low prevalence complex diseases
eCi minate Cancer Initiative
CORe CANCER OUTCOMES RESEARCH
CASIDE
InSite
NAACCR
Network Paediatric Cancer (ERN PaedCan)
NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS
Monarch CONSULTING AND EDUCATION
Project Data Sphere
Access the Data
ISO
caDSR
European Commission
AJCC American Joint Committee on Cancer
mCODE™
SEER
CIBMTR CENTER FOR INTERNATIONAL BLOOD & MARROW TRANSPLANT RESEARCH
CHILDHOOD CANCER DATA LAB
Sentinel
CCDI CHILDHOOD CANCER DATA INITIATIVE
EBMT

Formal logic in knowledge representation

Open-world assumption

1. What we know – is true
2. What we don‘t know – we don‘t know

Closed-world assumption

1. What we know – is true
2. What we don‘t know – is false

→ RWD is Closed World **implicitly**

Example:

- ICD-10-CM has all diseases
- Our analyses, like rates, incidence, causal effect expect, we expect:
  - The code if the patient suffers from it
  - The code to be absent if not
- There is no such thing as “No MI today”.

Closed and Open World Domains

Closed

- Conditions, including
  - Histology
  - Topology
  - Metastases
  - Lymph nodes
  - Stages
  - Grades
- Drugs
- Procedures
- Devices

Open

- Images
- Wave forms
- Omic variants
- Locations
- Survey answers
- Free text

So, if it is not a new standard, what do we need?

Genomic Data in a Closed World Environment

Precision Oncology

If we ever want to apply epidemiological methods used in clinical research to the study genomic variation, it

… requires genomic variants that are:

- unique,
- comprehensive,
- searchable,
- which means, standardized

However, genomic variant data…

… are not standardized for interoperability

… are too narrow when measured through pre-defined panels and overwhelmed with meaningless information when obtained through Next Generation Sequencing

… are obtained through a myriad of different detection methods

… occur in different modalities, which are somewhat interconnected

They have their own world of representation

They need to become features

Without features, epidemiological methods cannot work

- Genomic markers need to be turned into **features** of an **analytical dataset**:

Met exon 14 skipping mutation

- Splice region variant
- Chromosome 7q31.2
- Exon 14 missing
- Location 116411884-116411895

	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5	Feature 6	Feature 7	Feature 8	Feature 9	Feature 10	Feature 11	Feature 12
Person 1												
Person 2												
Person 3												
Person 4												
Person 5												
Person 6												
Person 7												
Person 8					1							
Person 9												
Person 10												
Person 11												
Person 12												
Person 13											1	
Person 14												
Person 15												
Person 16												
Person 17		1										
Person 18												

→ ETLing is easier said than done

We get:

- Different modalities, which are often mixed
- Different sources
- Different detection methods
- Flat files with complex syntax requirements but lax compliance

We need:

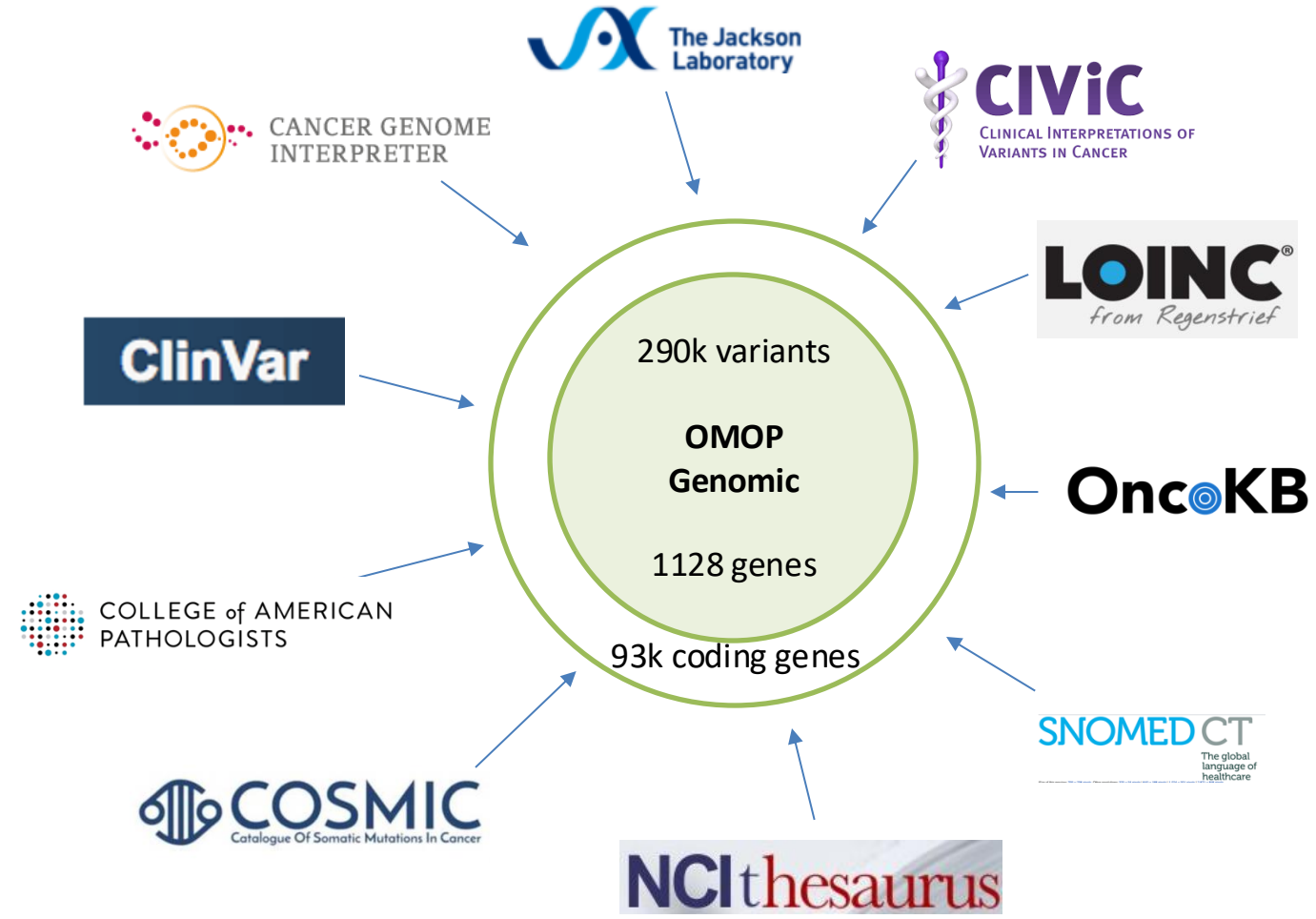
- Precise variants
- Standardized variants
- Meaningful hierarchy
- Closed World assumption

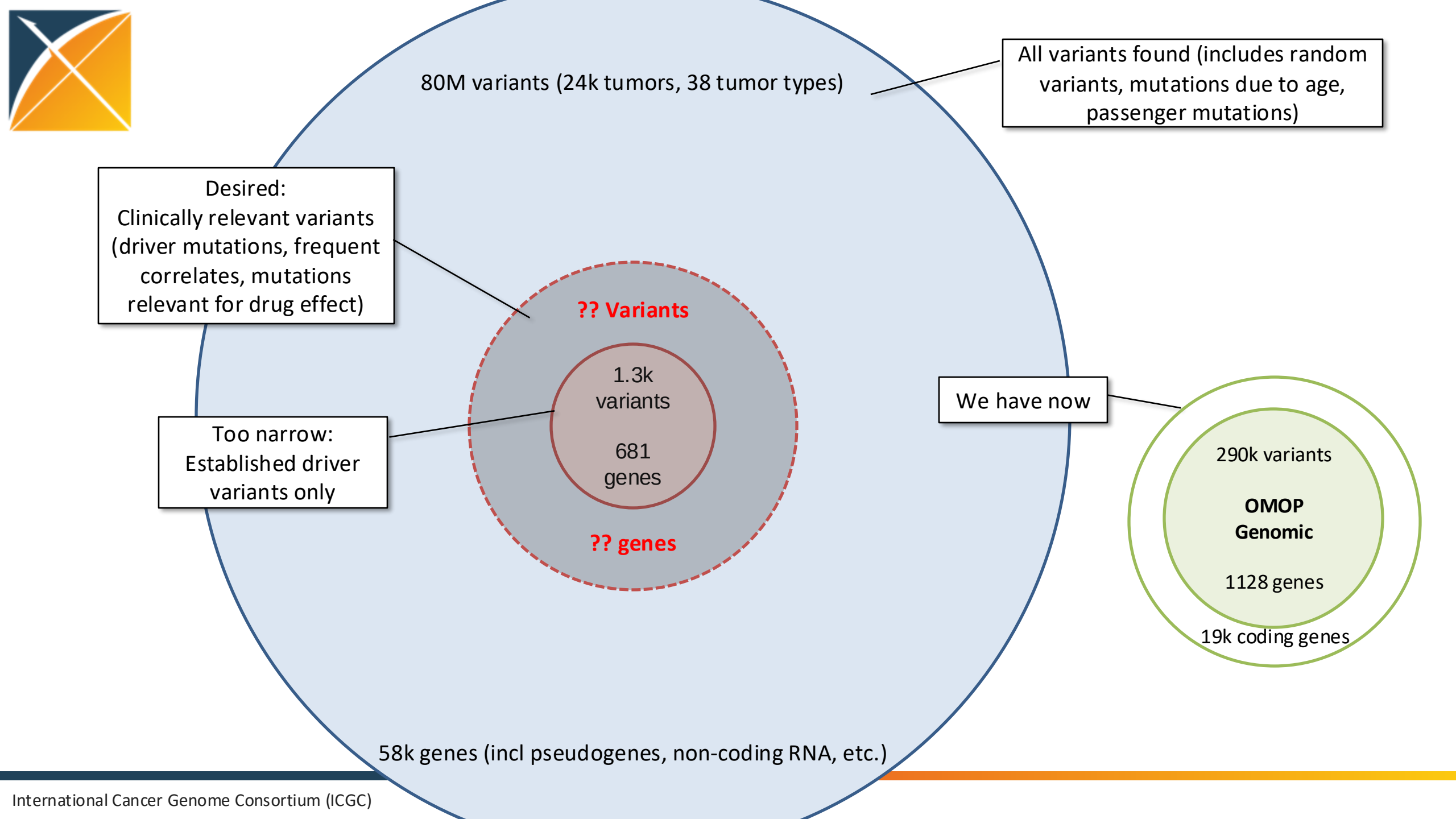
Recording Variants in the OMOP CDM

OMOP Genomic is built from relevant sources

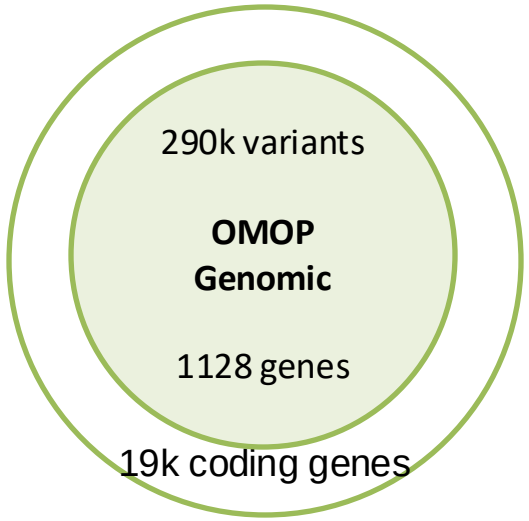
… by

- Combining public repositories
- Deduping them




80M variants (24k tumors, 38 tumor types)
All variants found (includes random variants, mutations due to age, passenger mutations)
Desired:
Clinically relevant variants (driver mutations, frequent correlates, mutations relevant for drug effect)
?? Variants
1.3k variants
681 genes
?? genes
Too narrow:
Established driver variants only
We have now
290k variants
OMOP Genomic
1128 genes
19k coding genes
58k genes (incl pseudogenes, non-coding RNA, etc.)

OMOP Genomic contains



Genetic Variation: 19,297

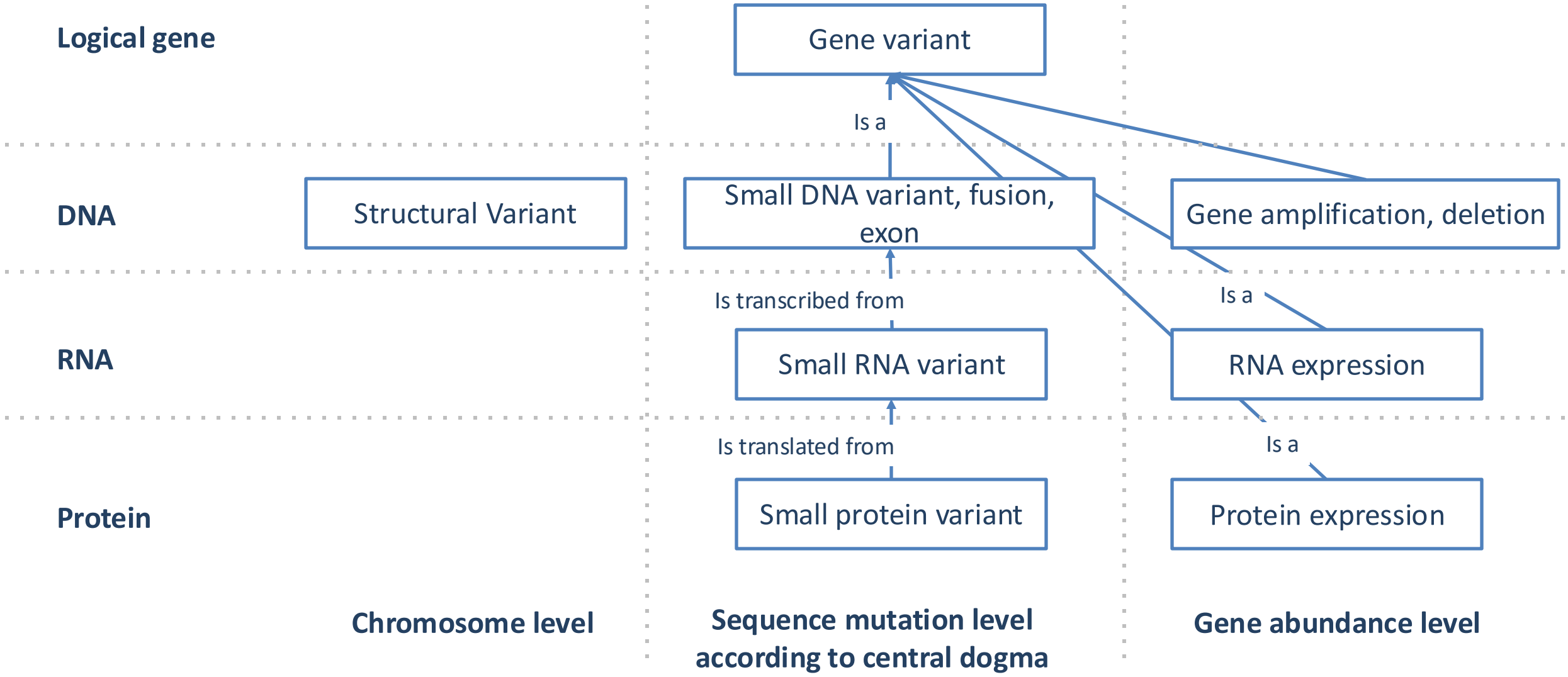
Structural Variant: 3,226

Gene DNA Variant: 83,460

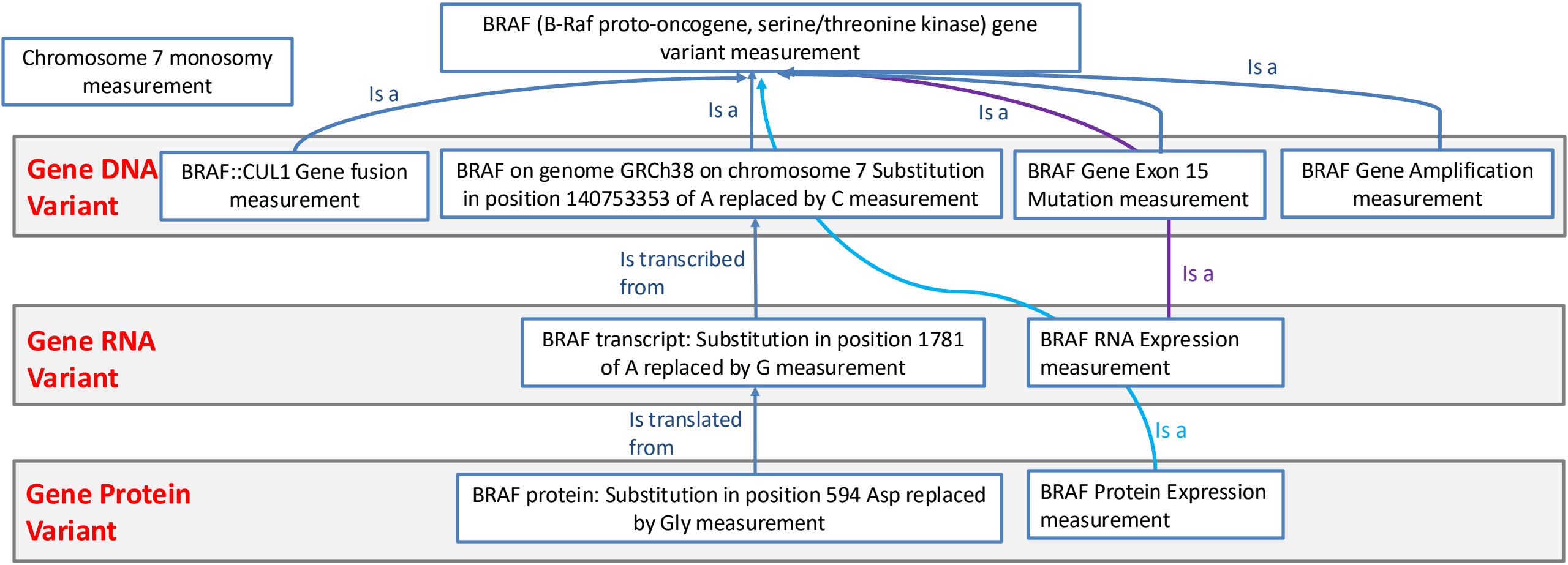
Gene RNA Variant: 92,988

Gene Protein Variant: 89,795

Hierarchical relationships inside OMOP Genomic



Hierarchical relationships inside OMOP Genomic
EXAMPLE



Summary

1. Genomic variants must enter epidemiological research.
2. For that, they need to become uniquely identifiable features.
3. For that, we need them to be
   - Well, unique
   - Well, identifiable
   - Comprehensive (to cancer)
   - Finite

   } = Closed World
4. OMOP gives you all that.
5. **Join the Journey**