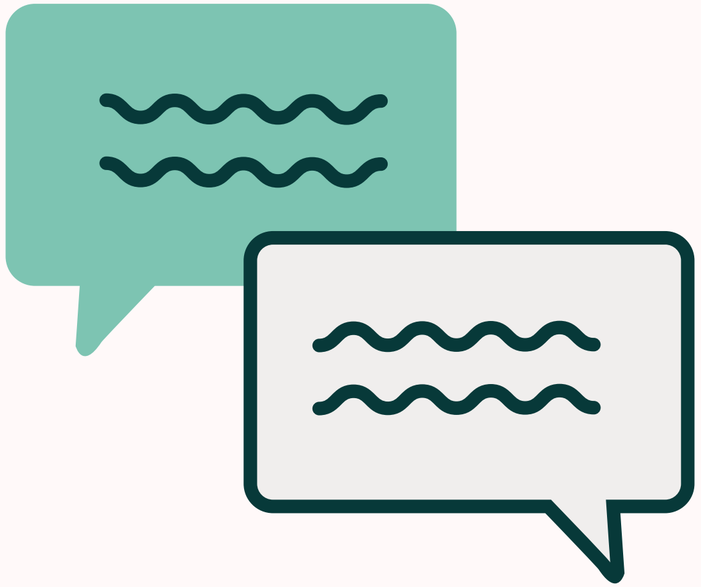
DE-MYSTIFYING BIOMEDICAL [META]DATA

Guest Lecture for LIS 545 (Spring 2021)

Kaitlin Throgmorton, MLIS

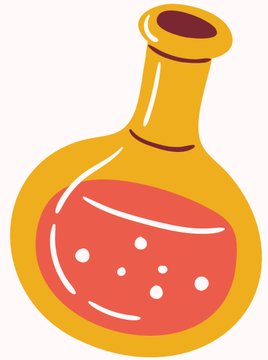
Environment

- Interrupt any time!
- I'm not great at chat while talking, so either jump in, or wait for a break, at which point I'll check in on chat.
- I love seeing videos on, but also totally get that we are all screentimed out.
- I've also got staggered Q&A breaks throughout, and I'll leave you with my contact info.



Outcomes

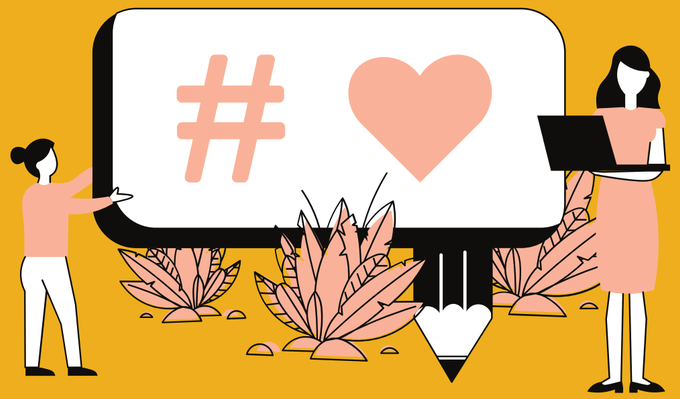
- Instill appreciation for the complexity and heterogeneity of biomedical data
- Demystify bioinformatics, and maybe even inspire you to consider it
- Provide a behind-the-scenes look at how data standards proliferate



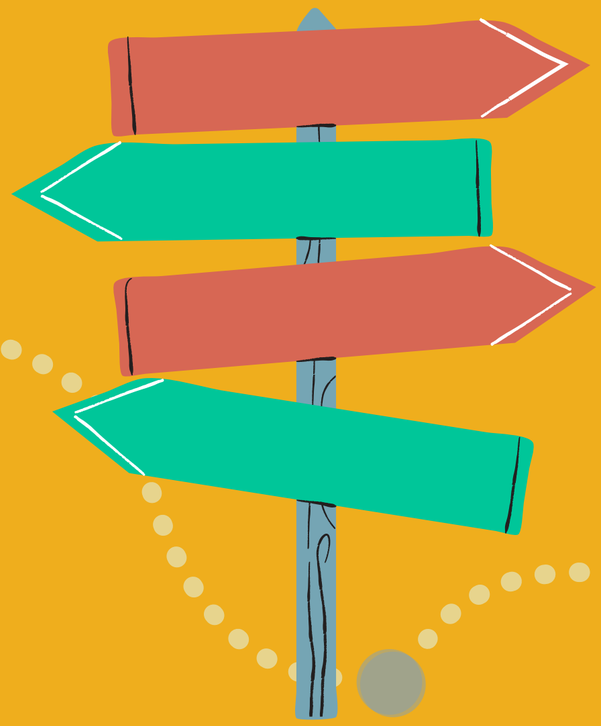
Outline

- About Me & My Journey to Bioinformatics
- Quick Primer on Biomedical Data & Metadata
- Case Study #1: RNA-Seq Data & Metadata
- Case Study #2: Clinical Data & Metadata
- Implications for Curators

About Me



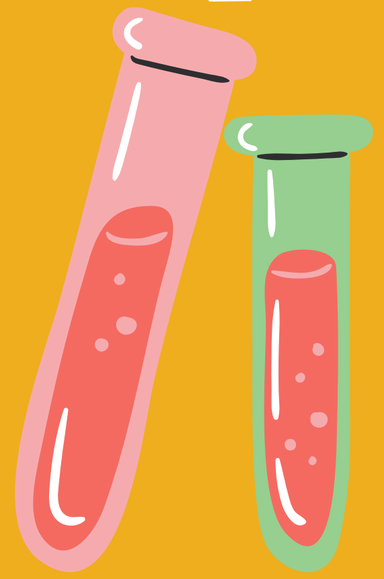
My Journey to Bioinformatics



What's the most complex data you've worked with before?

Why was it challenging?

QUICK PRIMER ON BIOMEDICAL [META]DATA



What makes biomedical data unique?

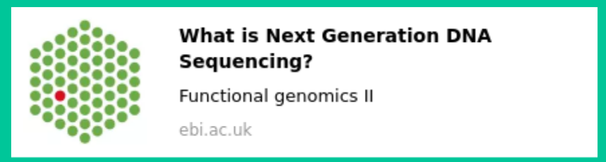
- Data and metadata are often incredibly heterogeneous
  - Data generation usually involves many intermittent steps
- Metadata can be bespoke to a particular experiment, protocol, even a lab
- It may be some of the most controlled data, subject to privacy and other regulations
- Curators may never see the actual data

(Subramanian, et al., 2020)

'Next Generation' (NGS) or 'High Throughput' Methods

- Faster
- Cheaper
- Much larger and more complex data outputs

(Huerta & Burke, 2020)



Common Biomedical Data Types, Known as " 'omics Data "

Genomic

(*+ Epigenomic, Transcriptomic*)

Proteomic

Metabolomic

(Very) Simplified Breakdown of How These Relate

Genes

transcription
expression
translation

Proteins

function

Traits

Example Biomedical Data Types

- Genetic sequencing (DNA-Seq, RNA-Seq, WGS, WES, PCR)
- Gene expression counts
- Genetic variation

- Mass spectrometry for protein profiling
- Flow cytometry (CyTOF)
- Protein microarrays

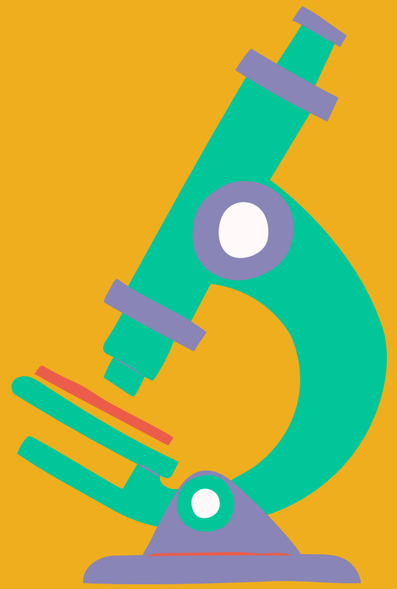
- High performance liquid chromatography (HPLC)

Common Data Use Cases

- Using biomarkers as early onset indicators
- Identifying which genotypes lead to certain phenotypes
- Explicitly targeting cells, genes, and/or proteins to treat disease
- Further unraveling our genomes to reveal medical insights

(NIEHS, 2021; CDC, 2021)

CASE STUDY #1: RNA SEQUENCING



What is RNA Sequencing?

"set of experimental procedures that generates cDNA molecules derived from RNA molecules, followed by sequencing-library construction and massively parallel deep sequencing"

(Huerta & Burke, 2020)

Photo by National Cancer Institute on Unsplash

RNA-Seq Data & Metadata

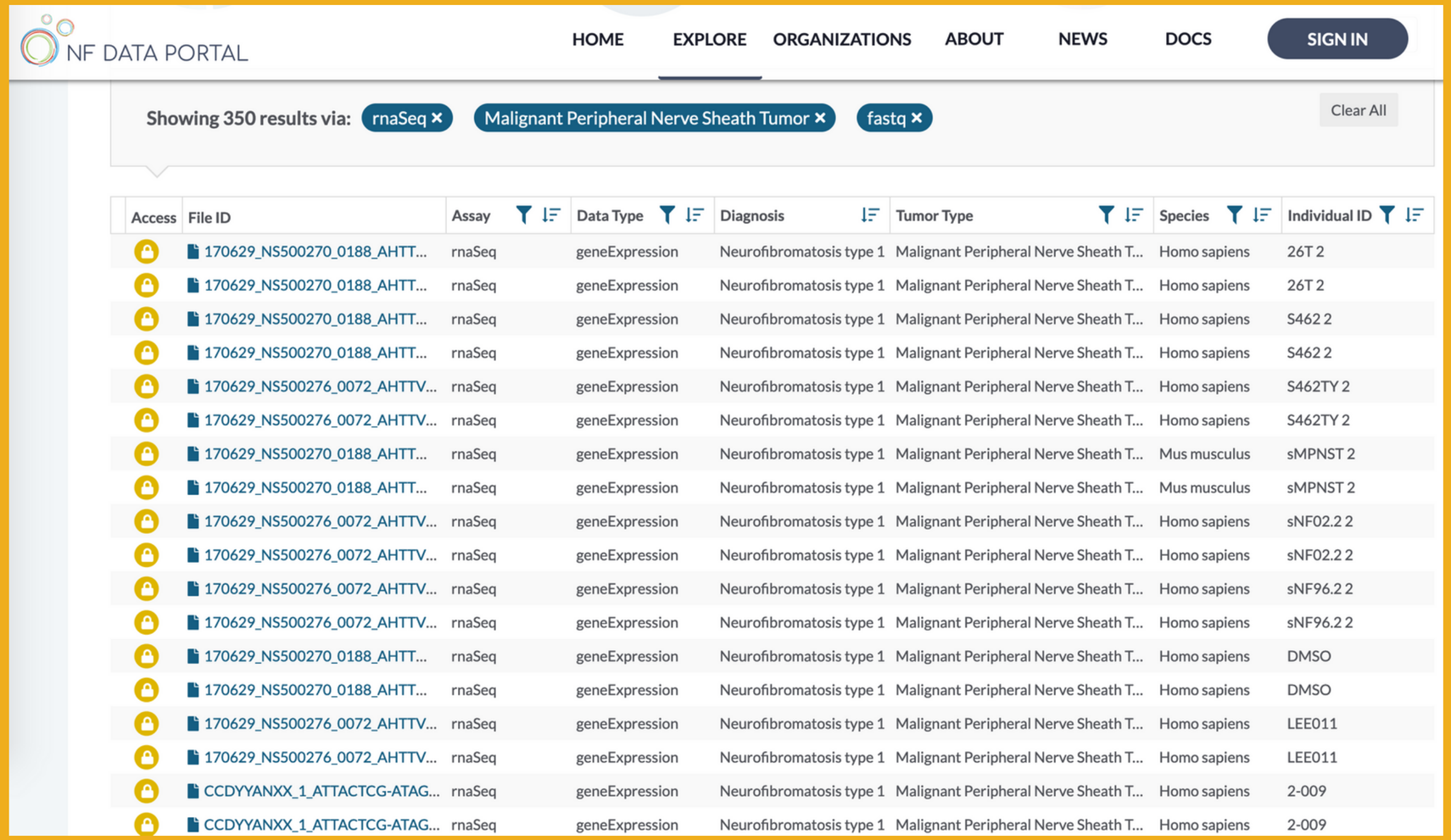
- Most common raw format is a **FASTQ** file
- FASTQs are typically **processed** through several levels before analysis
- Different metadata may be collected depending on the processing stage
- FASTQ metadata typically includes at least the following:
  - **Sample** information
  - **Experiment** information
  - **Platform** information
  - **Quality control** information

(Illumina, 2020)

RNA-Seq Data & Metadata

nf.synapse.org

RNA-Seq Data & Metadata



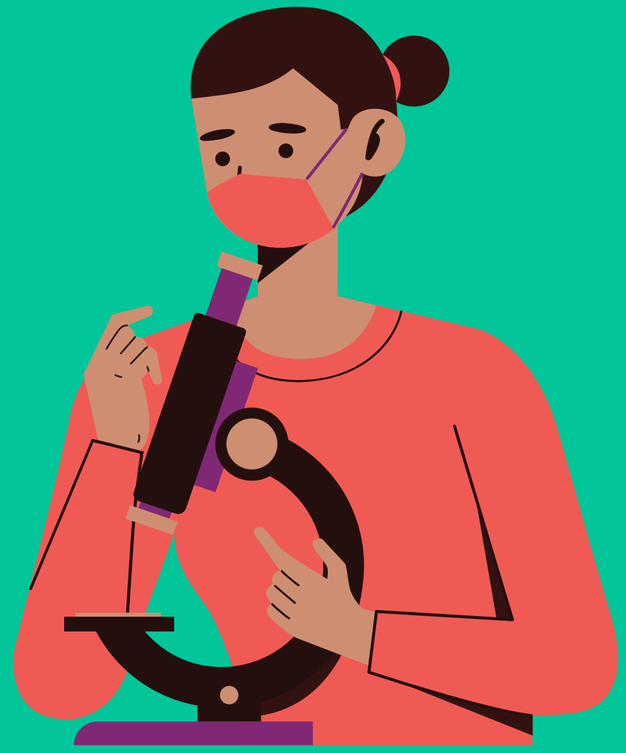
Access	File ID	Assay	Data Type	Diagnosis	Tumor Type	Species	Individual ID
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	26T 2
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	26T 2
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	S462 2
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	S462 2
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	S462TY 2
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	S462TY 2
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Mus musculus	sMPNST 2
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Mus musculus	sMPNST 2
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	sNF02.2 2
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	sNF02.2 2
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	sNF96.2 2
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	sNF96.2 2
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	DMSO
	170629_NS500270_0188_AHTT...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	DMSO
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	LEE011
	170629_NS500276_0072_AHTTV...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	LEE011
	CCDYYANXX_1_ATTACTCG-ATAG...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	2-009
	CCDYYANXX_1_ATTACTCG-ATAG...	rnaSeq	geneExpression	Neurofibromatosis type 1	Malignant Peripheral Nerve Sheath T...	Homo sapiens	2-009

nf.synapse.org

Common Standards

- Genomic Data Commons (GDC)
- ENCODE
- dbGap Submission Guidelines

(See resources at end)



Stretch Break

+

Q & A

CASE STUDY #2: CLINICAL INFO

What is clinical data?

"Clinical data is a staple resource for most health and medical research. Clinical data is either collected during the course of ongoing patient care or as part of a formal clinical trial program. Clinical data falls into six major types:

- Electronic health records
- Administrative data
- Claims data
- Patient / Disease registries
- Health surveys
- Clinical trials data"

(UWHSL, 2021)

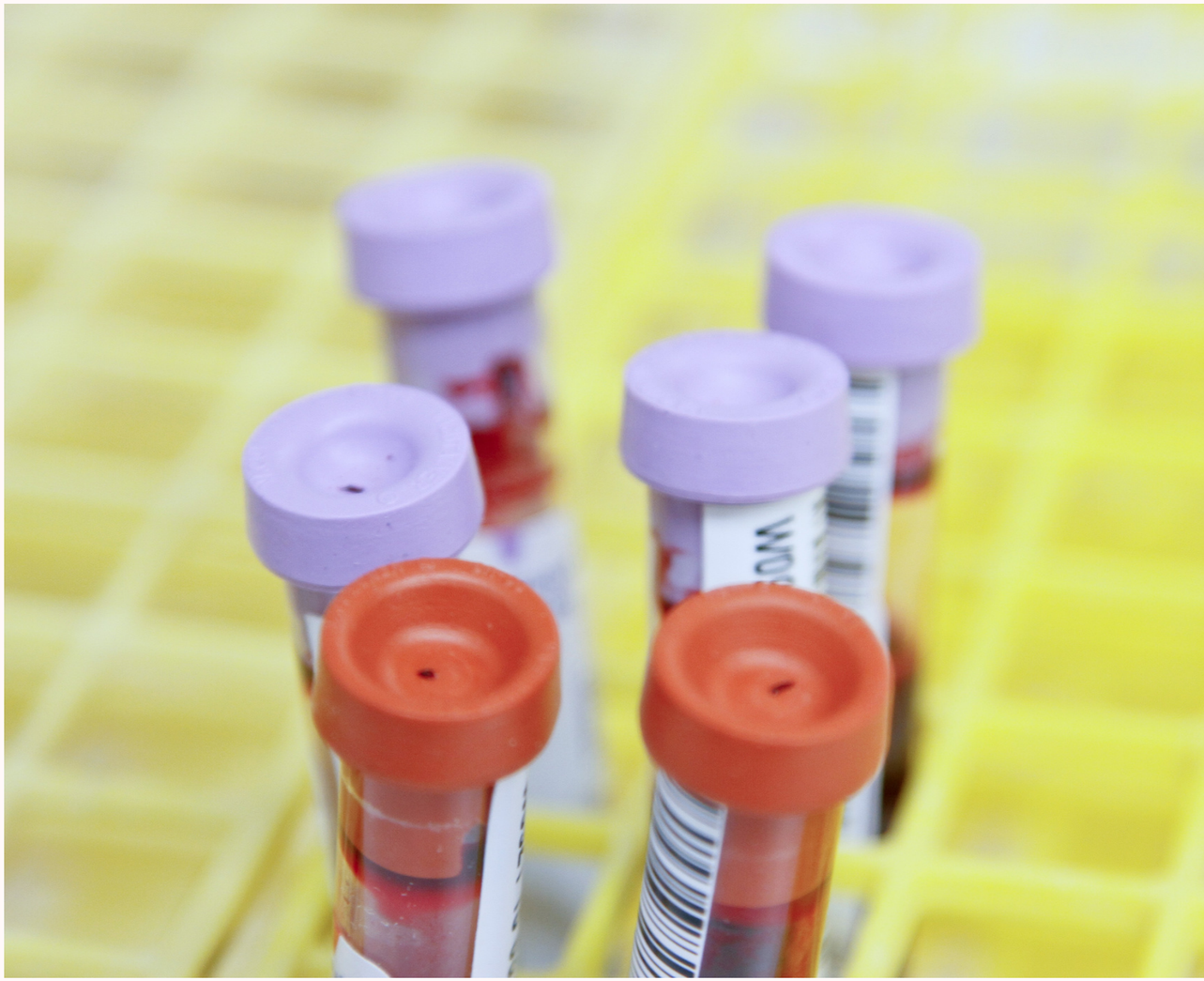


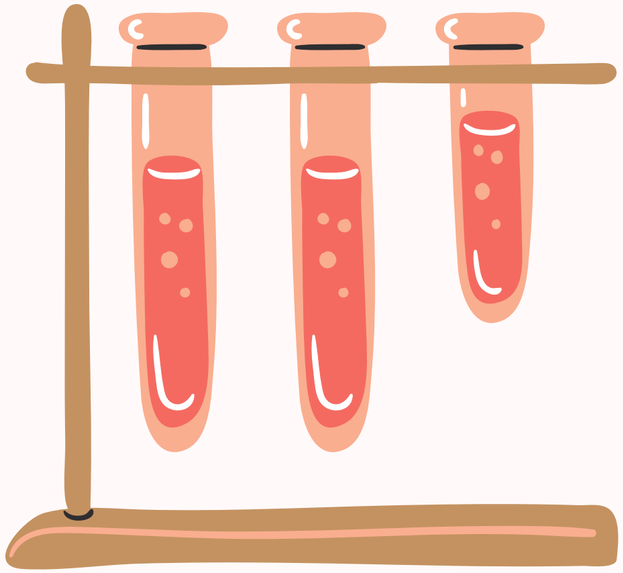
Photo by National Cancer Institute on Unsplash

Why Clinical Data?

- While 'omics data helps us understand what's happening on a cellular and biological level, clinical data fills in the rest of the picture — a subject's clinical progression from diagnosis to treatment to, hopefully, cure.

Common Clinical Data Types

- Demographic
- Diagnosis
- Treatment
- Outcome
- Finding
- Medical History & Comorbidities
- Concomitant Medications
- Family Medical History
- Insurance & Payment



NIH NATIONAL CANCER INSTITUTE GDC Data Portal

Home | Projects | Exploration | Analysis | Repository | Quick Search | Manage Sets | Login | Cart 0 | GDC Apps

Cases | Clinical | Genes | Mutations «

Search Clinical Data

Search...

Only show fields with values (84 fields shown)

Demographic — Expand All

- Gender
- Race
- Ethnicity
- Vital Status
- Cause Of Death

6 More...

Diagnoses — Expand All

- Age At Diagnosis
- Year Of Diagnosis
- AJCC Clinical Stage

Clear | Is Cancer Gene Census | IS | true

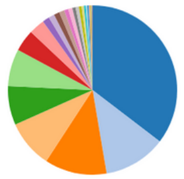
Cases (13,399) | Genes (576) | Mutations (164,704) | OncoGrid

View Files in Repository

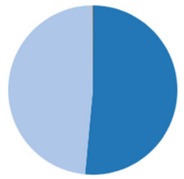
Primary Site

Project

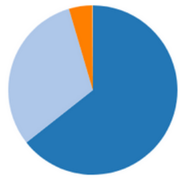
Disease Type



Gender



Vital Status



Showing **1** - **20** of **13,399** cases

Biospecimen | Clinical | JSON | TSV | Save/Edit Case Set

Case ID	Project	Primary Site	Gender	Files	Available Files per Data Category							# Mutations	# Genes	Slides
					Seq	Exp	SNV	CNV	Meth	Clinical	Bio			
TCGA-A5-A0G2	TCGA-UCEC	Corpus uteri	Female	60	4	5	16	7	1	10	17	1,856	481	(3)
TCGA-EO-A22U	TCGA-UCEC	Corpus uteri	Female	59	4	5	16	7	1	10	16	1,234	445	(2)
TCGA-FI-A2D5	TCGA-UCEC	Corpus uteri	Female	60	4	5	16	7	1	11	16	1,103	437	(2)
TCGA-B5-A3FC	TCGA-UCEC	Corpus uteri	Female	59	4	5	16	7	1	10	16	1,180	436	(2)
TCGA-AX-A2HC	TCGA-UCEC	Corpus uteri	Female	67	6	10	16	7	2	10	16	1,095	434	(2)
TCGA-EO-A22R	TCGA-UCEC	Corpus uteri	Female	61	4	5	16	7	2	10	17	1,100	429	(3)
TCGA-IB-7651	TCGA-PAAD	Pancreas	Female	58	4	5	16	7	1	8	17	1,062	422	(3)
TCGA-2W-A8YY	TCGA-CESC	Cervix uteri	Female	58	4	5	16	7	1	9	16	1,013	409	(2)
TCGA-AX-A1CE	TCGA-UCEC	Corpus uteri	Female	60	4	5	16	7	1	10	17	1,022	402	(3)

Clinical Data Standards & Common Data Models (CDMs)

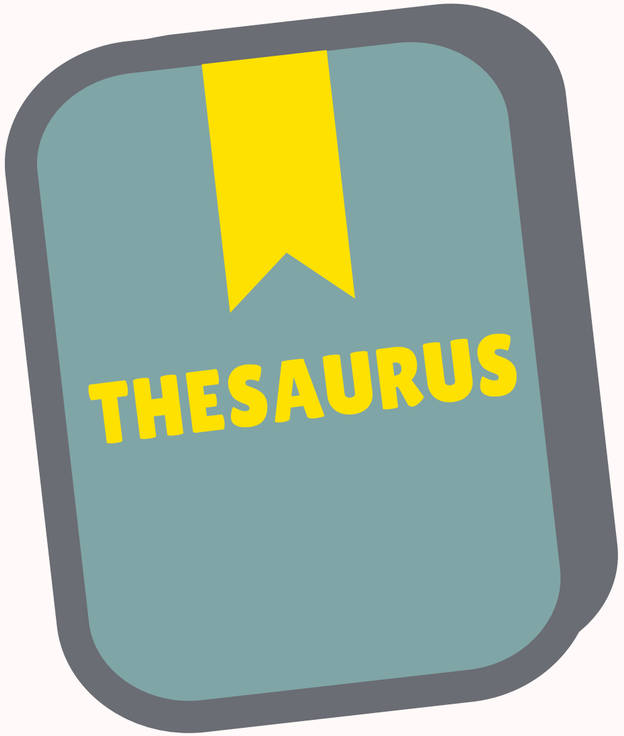
- Clinical Data International Standards Consortium's (CDISC) Study Tabuluation Data Model (SDTM)
- Observational Health Data Sciences and Informatics' (OHDSI) Observational Medical Outcomes Partnership (OMOP)
- Sentinel
- National Patient-Centered Research Network (PCORnet)
- ... and *many* more.

(See resources at end)

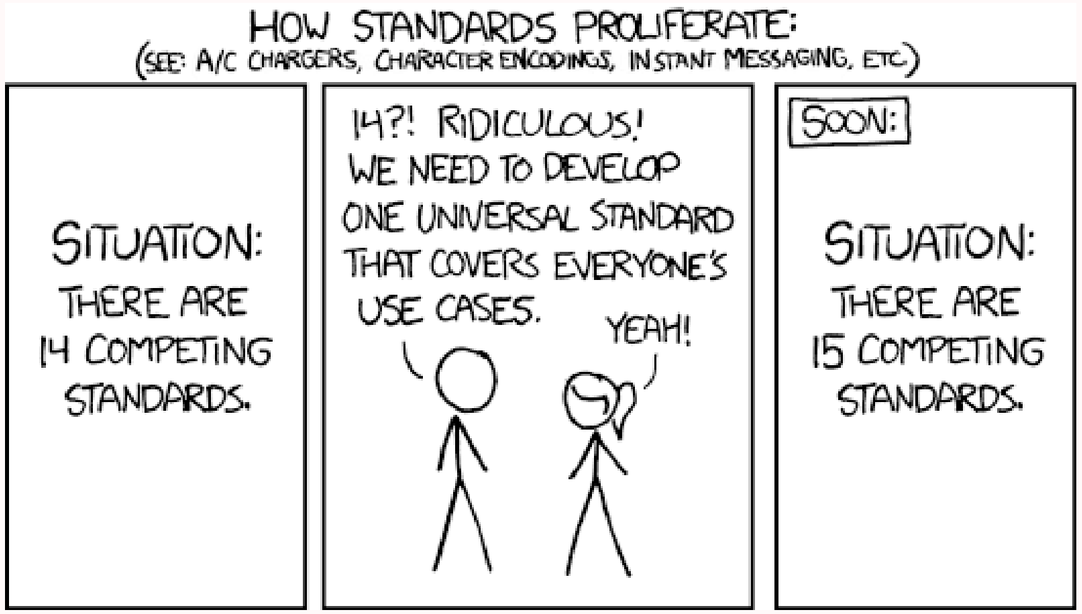
Vocabularies / Ontologies & Coding Systems

- SNOMED
- LOINC
- UMLS
- RxNorm
- CDEs
- NCI Thesaurus
- … and many more.

(See resources at end)



Why so many standards?



xkcd: https://xkcd.com/927/

Divergent Data Consumers

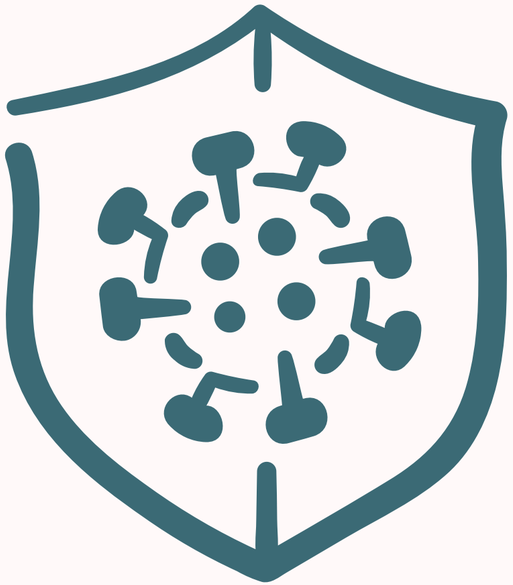
- Clinicians & Researchers
- Regulatory Authorities
- Insurers
- Patients & Citizen Scientists

Not surprisingly, these audiences have different questions to ask of clinical data.

Divergent Privacy Regulations

- HIPAA (US)
- GDPR (EU)
- PIPA (S. Korea)

This effects everything from what and how data can be collected, to how data is structured.



Disease Divergence

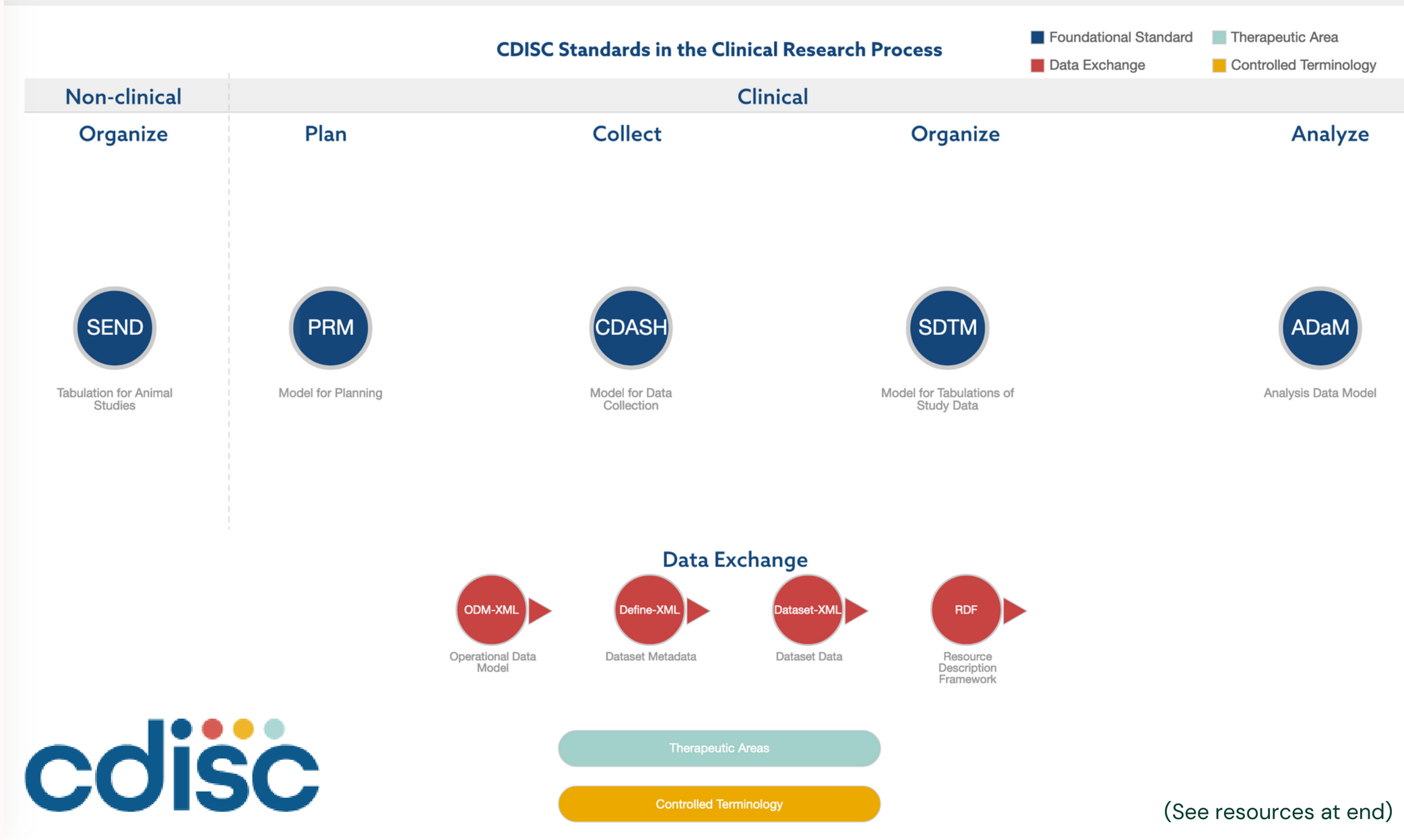
- Standards proliferate to accomodate various diseases and other medical anomalies and niches

Global Consortia & Harmonization

- Increase in global projects that combine the challenges of the first two items - disparate audiences and regions
- Increase in harmonization proejcts that combine both 'omics and clinical datasets to answer specific research questions

Upstream Challenges

- 'Raw' clinical data, otherwise known as 'patient data', 'in clinic data', or 'point of collection' data, has a variety of challenges:
  - Not always electronic
  - Not always standardized
  - Often needs to be abstracted
  - Outputted in a variety of formats



(See resources at end)

Rising Popularity of Mapping Tools

- Fast Health Interoperability Resources (FHIR)
- MetaMap
- etc.

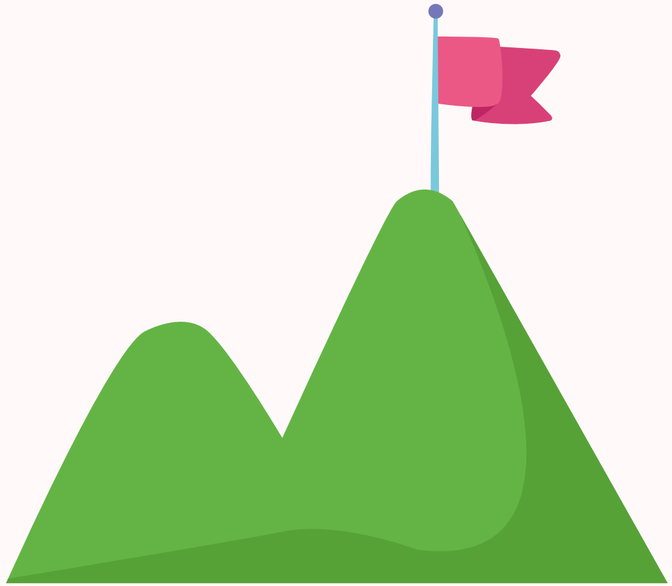
(See resources at end)

WHERE DATA CURATORS COME IN

Key Opportunities & Challenges

Challenges

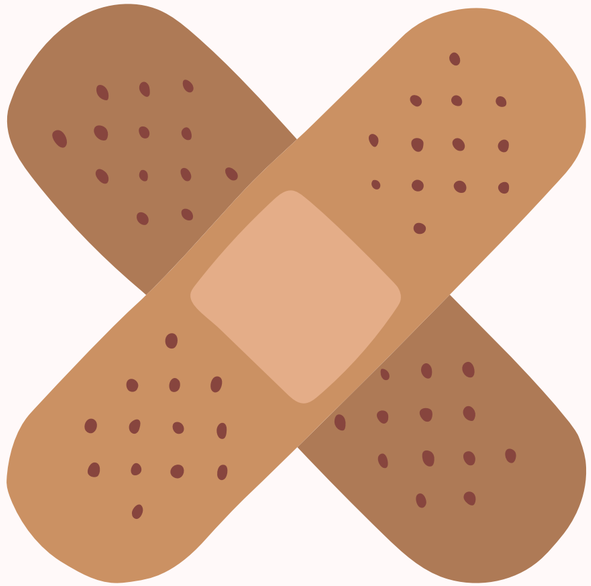
- Size, Scale, Complexity of Biomedical Data & Metadata Infrastructure
- Competing Interests & Standards
- Funding for 'Novel' Research
- Over-Reliance on Tech



Information Science Solutions

(not just band-aids)

- Standards & Ontologies
- Data Modeling
- Strong Documentation (& Computational Methods)



MLIS-Specific Skills

- Information Organization, Behavior, & Architecture
- Policy Development & Adherence
- Partnership & Consensus Building
- User Outreach & Reference
- Equitable & Free Information Access

Q&A

References

Allaway, R.J., La Rosa, S., Verma, S. et al. (2019). Engaging a community to enable disease-centric data sharing with the NF Data Portal. Scientific Data (6), 319. https://doi.org/10.1038/s41597-019-0317-x

Centers for Disease Control and Prevention (CDC). (2021). Understanding mRNA COVID-19 vaccines. https://www.cdc.gov/coronavirus/2019-ncov/vaccines/different-vaccines/mRNA.html

Garza, M., Del Fiol, G., Tenenbaum, J., Walden, A., & Zozus, M. N. (2016). Evaluating common data models for use with a longitudinal community registry. Journal of biomedical informatics, 64, 333-341. https://doi.org/10.1016/j.jbi.2016.10.016

Huerta, L. & Burke, M. (2020). *Functional genomics II*. EMBL-EBI Training. https://doi.org/10.6019/TOL.FunGenII-c.2016.00001

Illumina. (2020). FASTQ files explained. https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html

References

National Institute of Environmental Health Sciences (NIEHS). (2021). Biomarkers. https://www.niehs.nih.gov/health/topics/science/biomarkers/index.cfm

NF Data Portal. (n.d.). https://nf.synapse.org/

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics & Biology Insights, 14*, 117793221989905l. https://doi.org/10.1177/117793221989905l.

University of Washington, Health Sciences Library (UWHSL). (2021). Data resources in the health sciences: Introduction to clinical data. https://guides.lib.uw.edu/hsl/data/findclin

Note: all images free from either Canva or Unsplash. Presentation developed using Canva.

Resources: Standards

'omics Data Standards

- Genomic Data Commons (GDC): https://gdc.cancer.gov/
- ENCODE: https://www.encodeproject.org/help/data-organization/
- dbGap Submission Guidelines: https://www.ncbi.nlm.nih.gov/gap/docs/submissionguide/#astart
- Cancer Immunologic Data Commons: https://cimac-network.org/cidc/

Resources: Standards

Clinical Data Standards

- Clinical Data International Standards Consortium's (CDISC) Study Tabuluation Data Model (SDTM): https://www.cdisc.org/ + https://www.cdisc.org/standards/foundational/sdtm
- Observational Health Data Sciences and Informatics' (OHDSI) Observational Medical Outcomes Partnership (OMOP): https://www.ohdsi.org/ + https://www.ohdsi.org/data-standardization/
- National Patient-Centered Research Network (PCORnet): https://pcornet.org/

Resources: Standards

Clinical Data Standards

- Sentinel: https://www.fda.gov/safety/fdas-sentinel-initiative
- NLM Electronic Health Records Tool Guide: https://www.nlm.nih.gov/healthit/meaningful_use.html

NLM Resources on Metadata & Standards

- National Library of Medicine's (NLM) Metadata Overview Page: https://nnlm.gov/data/metadata
- NLM Data Tools: https://nnlm.gov/data/data-tools

Resources: Standards

Controlled Terminologies & Thesauri

- SNOMED-CT: https://www.nlm.nih.gov/research/umls/Snomed/snomed_main.html
- LOINC: https://loinc.org/
- UMLS: https://www.nlm.nih.gov/research/umls/index.html
- RxNorm: https://www.nlm.nih.gov/research/umls/rxnorm/index.html
- Common Data Elements (CDEs): https://www.nlm.nih.gov/research/umls/index.html
- NCI Thesaurus: https://ncithesaurus.nci.nih.gov/ncitbrowser/

Resources: Standards

Mapping Tools

- Fast Health Interoperability Resources (FHIR - pronounced 'fire'): http://www.fhir.org/
- MetaMap: https://metamap.nlm.nih.gov/

Additional Resources

Genetics / Genomics

- *The Gene: An Intimate History* | Siddhartha Mukherjee (ISBN: 978-1432837815)
  - **Highly recommended, regardless of your interest in science!** I wish this was required reading for everyone. Dr. Muhkerjee makes genetics approachable and relevant and fascinating.
- A Gentle Introduction to RNA-Seq | StatQuest | YouTube
  - Did RNA-seq confuse, but also interest, you? This scientist explains it nicely!

Clinical Standards

- The Book of OHDSI
  - If you want to read an open-science-forward clinical data standard that's also well-documented and well-explained, this is for you!

Additional Resources

Biology & Bioinformatics

- Bioinformatics for Beginners | Coursera
- Data Management for Clinical Research | Cousera
- EMBL-EBI Training
  - Excellent, well-vetted, free resources on 'big data in biology.'
- Biostars | A bioinformatics forum
- OMGenomics | YouTube channel about bioinformatics
- *The Social Life of DNA: Race, Reparations, and Reconciliation After the Genome* | Alondra Nelson (ISBN: 978-0807033029)
- *A Brief History of Everyone Who Ever Lived: The Human Story Retold Through Our Genes* | Adam Rutherford (ISBN: 978-1615194186)

Contact

Kaitlin Throgmorton, *MLIS*

kaitlin@kaitlinthrogmorton.com

github.com/kthrog

these (DCI) lecture materials:
github.com/kthrog/LIS_545_guest_lecture

my DCII lecture materials:
github.com/kthrog/LIS-546-guest-lecture