Manual processes and lack of data standards plague clinical trials
Lack of clinical information is the biggest threat to genomic research
Ideal state

Electronic health record

~PCORI
~All of Us

Automatic extraction

Sequencing

Genomics

Rich analyses

Tumor / Blood / Tissue
Current state for genomic commons

Sparse clinical data

Tumor / Blood / Tissue

Sequencing

Genomics Commons
Current state for genomic commons

Sparse clinical data → Tumor / Blood / Tissue → Sequencing → Genomics Commons → Limited utility
Current state for genomic commons

Sparse clinical data

Tumor / Blood / Tissue

Genomics Commons

Data pull from EHR / data warehouse

Sequencing

Limited utility
Current state for genomic commons

- Sparse clinical data
- Tumor / Blood / Tissue
- Genomics Commons
  - Sequencing
  - Limited utility

- Data pull from EHR / data warehouse

- Organizations are rarely willing to pull data from the EHR
Current state for genomic commons

Clinical trials data → Sponsor (e.g., COG)

Sparse clinical data

Tumor / Blood / Tissue → Sequencing → Genomics Commons → Limited utility

Data pull from EHR / data warehouse

Organizations are rarely willing to pull data from the EHR
Current state for genomic commons

Clinical trials data → Sponsor (e.g., COG)

Sparse clinical data → Genomics Commons

Tumor / Blood / Tissue → Sequencing

Data pull from EHR / data warehouse

Organizations are rarely willing to pull data from the EHR

Limited utility
Current state for genomic commons

Clinical trials data → Sponsor (e.g., COG)

Sparse clinical data

Tumor / Blood / Tissue → Sequencing → Genomics Commons

Data pull from EHR / data warehouse

Data abstraction by statistician

Significant effort and funds required

Limited utility

Organizations are rarely willing to pull data from the EHR
Why do we need data commons?

- Too much data to store
- Takes too long to transfer
- Too expensive to analyze
- Lack of data standardization
What is a commons?

- Computing - cloud or HPC
- Publicly-available datasets
- Software services and tools
- FAIR digital object compliance

F - Findable
A - Accessible
I - Interoperable
R - Reusable
What is a commons?

Patient data
Lab data
Microbiome data

Researcher puts patient data into commons
Researcher puts lab data into commons
Researcher puts genomics data into commons
What is a commons?

Researcher puts patient data into commons

Researcher puts lab data into commons

Researcher puts genomics data into commons

Patient data
Lab data
Viz
Microbiome data
Tools

Researcher can use and analyze all data right in the commons
Genomic data commons portal
Pediatric cancer is rare - making it difficult to study

<table>
<thead>
<tr>
<th>Adult cancers annual incidence</th>
<th>Pediatric cancers annual incidence</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>All</strong></td>
<td><strong>All</strong></td>
</tr>
<tr>
<td>1,688,780</td>
<td>15,780</td>
</tr>
<tr>
<td>Oral</td>
<td>ALL</td>
</tr>
<tr>
<td>49,670</td>
<td>3080</td>
</tr>
<tr>
<td>GI</td>
<td>CNS</td>
</tr>
<tr>
<td>310,440</td>
<td>2780</td>
</tr>
<tr>
<td>Lung</td>
<td>Hodgkin lymphoma</td>
</tr>
<tr>
<td>222,500</td>
<td>1180</td>
</tr>
<tr>
<td>Skin</td>
<td>NHL</td>
</tr>
<tr>
<td>95,360</td>
<td>1040</td>
</tr>
<tr>
<td>Breast</td>
<td>AML</td>
</tr>
<tr>
<td>255,180</td>
<td>730</td>
</tr>
<tr>
<td>Ovary</td>
<td>Neuroblastoma</td>
</tr>
<tr>
<td>22,440</td>
<td>710</td>
</tr>
<tr>
<td>Prostate</td>
<td>Bone</td>
</tr>
<tr>
<td>161,360</td>
<td>820</td>
</tr>
<tr>
<td>Urinary</td>
<td>Thyroid</td>
</tr>
<tr>
<td>146,650</td>
<td>570</td>
</tr>
<tr>
<td>Lymphoma</td>
<td>Wilms</td>
</tr>
<tr>
<td>80,500</td>
<td>510</td>
</tr>
<tr>
<td>Myeloma</td>
<td>Germ cell</td>
</tr>
<tr>
<td>30,280</td>
<td>540</td>
</tr>
<tr>
<td>Leukemia</td>
<td>Rhabdomyosarcoma</td>
</tr>
<tr>
<td>62,130</td>
<td>340</td>
</tr>
<tr>
<td>Source: cancer.org - 2017</td>
<td>Retinoblastoma</td>
</tr>
<tr>
<td></td>
<td>280</td>
</tr>
<tr>
<td></td>
<td>Melanoma</td>
</tr>
<tr>
<td></td>
<td>310</td>
</tr>
<tr>
<td></td>
<td>Other</td>
</tr>
<tr>
<td></td>
<td>2890</td>
</tr>
<tr>
<td>Source: CDC - 2014</td>
<td></td>
</tr>
</tbody>
</table>
Most children are treated on a Children’s Oncology Group (COG) study.
Survival - Pediatric cancer - still a long way to go

- Acute lymphoblastic leukemia
- Hodgkin lymphoma
- Nephroblastoma
- Germ cell tumors
- Acute myelogenous leukemia
- Ewing sarcoma
- Osteosarcoma
- Neuroblastoma
- Rhabdomyosarcoma
- Non-Hodgkin lymphoma
- Brain tumor

Graph showing survival rates for different pediatric cancers from 1950 to 2010.
Neuroblastoma data commons
Neuroblastoma

The most common solid tumor in children.

Children’s Oncology Group (COG)  Germany  Japan  SIOPEN

<table>
<thead>
<tr>
<th>COG</th>
<th>SIOPEN</th>
<th>Japan</th>
<th>Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>Study 1</td>
<td>Study 4</td>
<td>Study 7</td>
<td>Study 10</td>
</tr>
<tr>
<td>Study 2</td>
<td>Study 5</td>
<td>Study 8</td>
<td>Study 11</td>
</tr>
<tr>
<td>Study 3</td>
<td>Study 6</td>
<td>Study 9</td>
<td>Study 12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>COG</th>
<th>SIOPEN</th>
<th>Japan</th>
<th>Germany</th>
</tr>
</thead>
<tbody>
<tr>
<td>Consensus</td>
<td>Consensus</td>
<td>Consensus</td>
<td>Consensus</td>
</tr>
</tbody>
</table>

Consensus standard

<table>
<thead>
<tr>
<th>Group</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG</td>
<td>4235</td>
</tr>
<tr>
<td>Germany</td>
<td>1938</td>
</tr>
<tr>
<td>Japan</td>
<td>470</td>
</tr>
<tr>
<td>SIOPEN</td>
<td>936</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8800</strong></td>
</tr>
</tbody>
</table>

>20 high-impact publications that changed clinical practice
>20 high-impact publications that changed clinical practice

<table>
<thead>
<tr>
<th>Group</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>COG</td>
<td>4235</td>
</tr>
<tr>
<td>Germany</td>
<td>1938</td>
</tr>
<tr>
<td>Japan</td>
<td>470</td>
</tr>
<tr>
<td>SIOPEN</td>
<td>936</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>8800</strong></td>
</tr>
</tbody>
</table>
Neuroblastoma Commons Cohort Discovery

Immediately see cohort counts

Links to external data sets

Publicly available at inrgdb.org
Neuroblastoma Commons Cohort Discovery

This used to take weeks. Now it can be done in seconds.

Example: Favorable biology, tissue available
Neuroblastoma data commons growth

<table>
<thead>
<tr>
<th>Year</th>
<th>COG</th>
<th>SIOPEN</th>
<th>GPOH</th>
<th>Japan</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>2004</td>
<td>4235</td>
<td>2157</td>
<td>1938</td>
<td>470</td>
<td>8800</td>
</tr>
<tr>
<td>2012</td>
<td>6127</td>
<td>2504</td>
<td>1938</td>
<td>470</td>
<td>11039</td>
</tr>
<tr>
<td>2013</td>
<td>11642</td>
<td>2504</td>
<td>1938</td>
<td>470</td>
<td>16554</td>
</tr>
<tr>
<td>2015</td>
<td>13060</td>
<td>2504</td>
<td>1938</td>
<td>470</td>
<td>17972</td>
</tr>
<tr>
<td>2016</td>
<td>13937</td>
<td>2664</td>
<td>1938</td>
<td>470</td>
<td>19009</td>
</tr>
<tr>
<td>2018</td>
<td>14425</td>
<td>3397</td>
<td>2154</td>
<td>470</td>
<td>20446</td>
</tr>
</tbody>
</table>

Data upload can be automated using a standardized data dictionary with error and consistency checking.
Data standardization is the **most important** goal in building a data commons.
International Soft Tissue Sarcoma Commons (INSTRuCt)

- Discussions began 1/2017
- First meeting in May, 2017 (Copenhagen)
- Second meeting in October 2017 (Chicago)
- Third meeting in March 2018 (Amsterdam)
- Fourth meeting planned September 2018 (Tübingen)
- Executive and informatics calls every 1-2 months
International Soft Tissue Sarcoma Commons (INSTRuCt)

October 2017 - Chicago

March 2018 - Amsterdam
Rhabdomyosarcoma data standardization

<table>
<thead>
<tr>
<th>HISTOLOGY</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alveolar rhabdomyosarcoma</td>
</tr>
<tr>
<td>2</td>
<td>Embryonal rhabdomyosarcoma</td>
</tr>
<tr>
<td>3</td>
<td>Botryoid rhabdomyosarcoma</td>
</tr>
<tr>
<td>4</td>
<td>Not otherwise specified</td>
</tr>
<tr>
<td>5</td>
<td>Undifferentiated sarcoma</td>
</tr>
<tr>
<td>6</td>
<td>Sarcoma, not classifiable</td>
</tr>
<tr>
<td>7</td>
<td>Spindle cell sarcoma</td>
</tr>
<tr>
<td>8</td>
<td>Ectomesenchymoma</td>
</tr>
<tr>
<td>9</td>
<td>Other</td>
</tr>
<tr>
<td>10</td>
<td>Mixed rhabdomyosarcoma</td>
</tr>
<tr>
<td>99</td>
<td>Unknown</td>
</tr>
</tbody>
</table>
Rhabdomyosarcoma data standardization

<table>
<thead>
<tr>
<th>HISTOLOGY</th>
<th>Description</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Alveolar rhabdomyosarcoma</td>
<td>C0206655</td>
</tr>
<tr>
<td>2</td>
<td>Embryonal rhabdomyosarcoma</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Botryoid rhabdomyosarcoma</td>
<td>C1306573</td>
</tr>
<tr>
<td>4</td>
<td>Not otherwise specified</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Undifferentiated sarcoma</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Sarcoma, not classifiable</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Spindle cell sarcoma</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Ectomesenchymoma</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Other</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Mixed rhabdomyosarcoma</td>
<td></td>
</tr>
<tr>
<td>99</td>
<td>Unknown</td>
<td></td>
</tr>
</tbody>
</table>

- Alveolar rhabdomyosarcoma (ARMS) | C0206655
- Botryoid rhabdomyosarcoma (BRMS) | C1306573
- Embryonal rhabdomyosarcoma (ERMS) | C0206656
- Pleomorphic rhabdomyosarcoma (PRMS) | C0334480
- Rhabdomyosarcoma (RMS), not classifiable | C0035412
- Rhabdomyosarcoma (RMS), inadequate tissue for classification | C1709053
- Rhabdomyosarcoma (RMS), w Mixed Embryonal and Alveolar Features | C1709053
- Spindle cell | C0205945

The table provides a classification system for rhabdomyosarcoma with codes for different types and features.
Cohort discovery tool

**Cohort Search**

**Search:** New

**Add a filter:**
- **HISTOLOGY**
  - Negative
  - Positive: PAX3-FOXO1
  - Positive: PAX7-FOXO1
  - Positive: PAX3-Other
  - Positive: Other
  - Unknown

**RESULTS**

- **Patients in INSTRuCT database:** 2028
- **Patients matching filters:** 54

**HISTOLOGY**

- Alveolar rhabdomyosarcoma (ARMS)
- Botryoid rhabdomyosarcoma (BRMS)
- Embryonal rhabdomyosarcoma (ERMS)
- Pleomorphic rhabdomyosarcoma (PRMS)
- Rhabdomyosarcoma (RMS), not classifiable
- Rhabdomyosarcoma (RMS), inadequate tissue for classification
- Rhabdomyosarcoma (RMS), with Mixed Embryonal and Alveolar Features
- Spindle cell
Pediatric cancer commons - Goals

Clinical trials data → Clinical trials data commons

Standardization
Harmonization

Common identifier
Rich analyses

Tumor / Blood / Tissue → Genomics

Sequencing

This is already being done.

This is what we are building

This is the payoff
Pediatric Acute Leukemia Initiative (PedAL)
Goals for PedAL

• Commons for all children being treated for leukemia
• Linkages to multiple disparate data sources
• Real-time access to data to make informed precision medicine decisions
• Meet sponsors’ compliance needs
Sample COG data flow / timeline

- New COG patient
- Project EveryChild
- Phase III COG trial
- Match to a trial within 48 hours
- Progression
- Relapse
- Further data collection
- Data access for researchers

Regular data flow into PedAL - clinical, genomic, IF, flow cytometry

PedAL
PedAL infrastructure

- COG
- Foundation Medicine
- Hematologics
- Kids First
- GDC
- St. Jude
- MSKCC
- Other consortia
- Other partners (e.g., Tempus, Flatiron)

Inputs

PedAL Commons
Interacting commons for clinical data

GDC

Commons interface

St. Jude’s

Kids First
Guiding informatics principles

• Registration for any child with leukemia
• Ingestion of data from multiple stakeholders
• Relentless adherence to accepted data standards
• FAIR digital object compliance
• Robust data governance and provenance
• Regulatory compliance through all phases of the data lifecycle
• Near real-time availability of data for clinicians and researchers
• Cooperation with existing data commons and platforms
Summary / Call to action

• Harmonized data leads to shared data

• Data and samples must have universal identifiers

• We must envision data collection and sharing at all stages of care

• The goal is all data from all patients at all times