A clinical data mesh for quality improvement and research in healthcare

Mike Hogarth
MD, Clinical Research Information Officer, UC San Diego Health

Tom Covington
CEO of Tag.bio

Mark Mooney
VP Customer Success, Tag.bio
Agenda (50 minutes)

UCSD and Tag.bio history 15 min
What is a Data Mesh? 10 min
Nightingale demo 20 min
Q & A 5 min
UCSD

Mike Hogarth, MD, Clinical Research Information Officer, UC San Diego Health
The problem (use case) we are trying to solve

- Provide clinical data (protected health information - PHI) to UCSD biomedical researchers:
  - Securely
  - Swiftly
  - Simply
  - Standardized
  - Semantically consistent
Data (and analyses) should be FAIR.

**Findable**
“I know where all our data is”

**Accessible**
“I can access any of the data that I need”

**Interoperable**
“I use one language for all my requests”

**Reusable**
“I use the existing data to answer new questions”
UCSD Health Secure Research Cloud in AWS

- What is it?
  - Integrating Data for Analysis, anonymization, and SHaring (iDASH) - v2.0
  - A secure computing environment for sensitive data
  - HIPAA compliant
  - Health system CSO approved for use involving protected health information (PHI)
  - Includes virtual research desktops and VMs in a locked-down AWS VPC without the ability for users to connect to the “outside”

- Why we needed it
  - c2017 - EHR data (PHI) for research routinely given to investigators through a download -- some data ended up shared it with external entities without data use agreements, no visibility into the location/use of the data, the ADCS situation along with ~800 other AWS accounts by UCSD Health staff/faculty without visibility or controls
UCSD health secure research cloud with nodes in a VPC
THE VIRTUAL RESEARCH DESKTOP (VRD)

- It is a modified version of the Amazon Web Services (AWS) Windows 10 “Workspace” virtual machine
- Runs in the protected UCSDH Secure Cloud in AWS
  - in the AWS HIPAA environment
  - approved by UCSDH CSO for PHI
- Provisioned with:
  - SPSS
  - R/RStudio
  - Python/PyCharm
  - Java 8 JDK
- Depending on approval, access to internal databases – ie, UC CORDS
  - tag.bio based access to available databases
Diagram of the Causes of Mortality
in the Army in the East.

The areas of the blue, red, & black wedges are each measured from the centre as the common vertex.
The blue wedges measured from the centre of the circle represent annual variates & the red wedges measured from the centre the deaths from smallpox & the black wedges measured from the centre the deaths from all other causes.
The black line across the red triangle in Nov. 1854 marks the boundary of the deaths from all other causes during that month.
In October 1854 & April 1855 the black area concudes with the red, in January & February 1856 the blue concudes with the black.
The entire areas may be compared by following the blue, the red & the black lines enclosing them.

Florence Nightingale, c. 1860
History

Tom Covington, CEO Tag.bio
How did the data mesh platform arise?

- Jesse Paquette (CSO Tag.bio) worked at UCSF Helen Diller, Family Comprehensive Cancer Center (2007-2010)
  - Working with Oncology researchers he realized that enabling them to answer their own questions would speed the turnaround time of question to answer
  - Created an initial software called EGAN (Exploratory Gene Association Networks)
  - Formed Tag.bio in 2014 with Tom Covington (CEO) and built first versions of what were then called Flux Capacitors or FC’s but became data nodes.
  - Began projects with UCSF Med Center on billing, encounters and claims data in 2018.
  - Realized that the architecture we were working on was an implementation of a “data mesh” after reading Zhamak Dhegani’s article: [How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh](#)
What’s The History?

- In January 2020 met Tag.bio at Precision Medicine World Conference
  - Initial discussion about work with UCSF on value based healthcare.
  - Set up a visit to UCSD in February
  - Initiated research collaboration with Mike Hogarth in March

- The Pandemic
  - Realized there was an immediate need to make COVID data accessible
  - Built first COVID registries in April

Building a range of patient registries at the present time
What is a Data mesh?

Mark Mooney, VP of Customer Tag.bio
Data nodes deployed and registered in a decentralized **Data Mesh**

Similar to an app store or a library of data products.

Zhamak Dehghani (Thoughtworks):  [How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh](#)
UCSD health secure research cloud with nodes in a VPC
Tying 3 technical components...

...into a domain-driven data node

Data

Algorithms

Visualization

Campus LYSAN
NODE: Data Map

Types
- Proprietary data (e.g. patients, billing, etc.)
- Public data (e.g. TCGA, CDC, etc.)
- Data services (e.g. annotation, usage, etc.)
- Emerging data types

Sources
- Siloed data
- Data lake
- Data node

Formats
- CSV
- JSON
- SQL

UC CORDS COVID-19

Combining healthcare data from across the six University of California medical schools and health systems

Aug 14 2020
- 175,517 COVID tested patients
- 6,056 COVID+ patients
- all labs, meds, vitals, 29 ICU data elements
- 319,952,837 “data points”

de-identified data
Ingested in 5 hours
NODE: Algorithms

Classic statistical methods
PCA, tSNE, UMAP,
t-test, Mann-Whitney,
Hypergeometric,
Cox regression,
Paired analysis over time,
Pathway analysis, etc.

Integration with
Machine Learning / AI
NODE: Smart API/UX

Analysis Apps
Made by the data scientist
Used by the researcher

Smart API
Data Map
Algorithms

N-Month Survival
Cox Survival
Expression
Data Quality
PCA
Expression Cutoff
DNA Mutation/Variation

Campus LYSAC
data node
data node
data node
NODE: Fast MVP
robust CI/CD

NODE: DIY
"protocol_definition": {
    "name": "elastic_net_crossvalidation",
    "visible": true,
    "title": "Elastic net cross validation",
    "description": "This protocol performs elastic net cross validation."
},
"category": "Python plugins",
"argument_sets": [
    "protocols/argument_sets/background_cohort_argument_set.json",
    "protocols/argument_sets/model_outcome_argument_set.json",
    "protocols/argument_sets/model_cross_validation_input_argument_set.json",
    "protocols/argument_sets/elastic_net_cross_validation_parameters_argument_set.json"
],
"tags": [
    "Python",
    "Elastic net",
    "Cross validation"
},
"method": "external",
"sdk": "python",
"plugin": "python/elastic_net_cross_validation.py",
"output_type": "html",
"background": "protocols/argument_references/background_cohort_reference.json",
"analysis_variables": [
    "protocols/variables/auto/patient_id_collection.json",
    "protocols/argument_references/model_outcome_variable_reference.json",
    "protocols/argument_references/model_input_variables_reference.json",
    "protocols/argument_references/plos_one_model_cv_input_variables_reference.json"
]
Nightingale registries
(a de-identified, automated, OMOP data product)

Mark Mooney, VP of Customer Tag.bio
### Tag admin console

Here you can configure users and datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gene Annotation</td>
<td>Deploys gene annotation information from public domain</td>
</tr>
<tr>
<td>Head and Neck Cancer (TCGA)</td>
<td>A dataset with clinical and multi-omics data for 590 head and neck cancer patients (from TCGA)</td>
</tr>
<tr>
<td>Synthetic COVID-19 demo data</td>
<td>Using synthetic, SARS-CoV-2 infected COVID-19 data to enable analysis demonstration and review</td>
</tr>
<tr>
<td>TCGA Pan-Cancer Atlas [catalyze]</td>
<td>Combined data from 33 cancer types from the 2018 TCGA Pan Cancer Clinical Data Resource</td>
</tr>
<tr>
<td>TCGA Pan-Cancer Atlas and the Immune Landscape of Cancer</td>
<td>Combined data from 33 cancer types from the 2018 TCGA Pan Cancer Clinical Data Resource</td>
</tr>
<tr>
<td>Tag.bio Analysis History</td>
<td>A dataset containing protocol usage history for all users</td>
</tr>
</tbody>
</table>
## Synthetic COVID-19 demo data

**Using synthetic, OMOP mapped, COVID 19 data to enable analysis demonstration and review.**

### Dataset privacy status:
- **Private**
- **Public**

### Dataset requires authentication:
- **Requires auth**
- **Does not require auth**

### Dataset details

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong></td>
<td>Synthetic COVID-19 demo data</td>
</tr>
<tr>
<td><strong>Singular entity name</strong></td>
<td>patient</td>
</tr>
<tr>
<td><strong>Plural entity name</strong></td>
<td>patients</td>
</tr>
<tr>
<td><strong>Description</strong></td>
<td>Using synthetic, OMOP mapped, COVID 19 data to enable analysis demonstration and review.</td>
</tr>
<tr>
<td><strong>Cloud Provider</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Link</strong></td>
<td><a href="https://github.com/syntheicheck/synthea/issues/679">https://github.com/syntheicheck/synthea/issues/679</a></td>
</tr>
</tbody>
</table>

### Users

- Add an existing user to this dataset
- Add all site users to dataset
- Remove all users from dataset
Select a dataset

After selecting a dataset, you will see a list of available protocols. Protocols are workflows designed to answer questions about that dataset.

UCSD COVID Patient Registry

UCSD COVID-19 Research Registry - All Tested Patients History View

UCSD COVID-19 Research Registry - Positive Patient History View

CORDS - UC COVID Patient Registry

UC CORDS Research Registry - All Tested Patients History View
Synthetic COVID-19 demo data

1625 patients

Using synthetic, OMOP mapped, COVID-19 data to enable analysis demonstration and review.

Overview Apps

Overview of Data
This protocol provides basic summaries for patient demographics and visit data.

Click to run

COVID-19 Summary Apps

Summary of COVID-19 Positive Patients
Summary of Specific Variables for COVID-19 Positive Patients
Summary of COVID-19 Positive Patients with Pre-Existing Conditions
COVID-19 Comparison Apps

Comparison of COVID-19 Positive Patients
This protocol compares clinical outcomes for all COVID-19 positive patients with a defined sub-cohort of COVID-19 positive patients.

COVID-19 Specialty Apps

Cox Survival

Cox Survival Using Specific Variables
Comparison of COVID-19 Positive Patients

This protocol compares clinical outcomes for all COVID-19 positive patients with a defined sub-cohort of COVID-19 positive patients.

COVID-19 Positive Cohort
Select additional criteria to further define the COVID-19 positive cohort.

- Build cohort
- Use saved cohort

Selected parameters
Sort attribute: Tag.score
Sort direction (default): >

Define a maximum number of results to view across all analyzed variables: 5000

Data from Entire Patient Record to Analyze
Select data to analyze.

- Outcomes recorded only after COVID positive status
- Visit collections
- Patient-related collections
- Activity collections
- Examination collections
Cohort builder

Visit Details

Patient Details

Gender
- Select all
- FEMALE
- MALE

Ethnicity

Race

Age range
- 40
- 80

412 patients  Cancel  Save & use
Comparison of COVID-19 Positive Patients

This protocol compares clinical outcomes for all COVID-19 positive patients with a defined sub-cohort of COVID-19 positive patients.

Instructions

COVID-19 Positive Cohort
Select additional criteria to further define the COVID-19 positive cohort.

- Build a new cohort
- Use saved cohort

Selected parameters
COVID-19 Positive Cohort
COVID-19 Positive cohort: f-40 to 80

Sort Results
- Sort attribute: Tag score
- Sort direction (default): ▼

Define a maximum number of results to view across all analyzed variables: 2500

Data from Entire Patient Record to Analyze
Select data to analyze.

Outcomes recorded only after COVID positive status

- Select all
- Filter values...
- Condition - After COVID Positive Diagnosis
- Drug - After COVID Positive Diagnosis
- Measurement - After COVID Positive Diagnosis
- Observation - After COVID Positive Diagnosis
- Procedure - After COVID Positive Diagnosis
Comparison of COVID-19 Positive Patients

This protocol compares clinical outcomes for all COVID-19 positive patients with a defined sub-cohort of COVID-19 positive patients.

Variable collection #2 of 3

**Condition - After COVID Positive Diagnosis**

- **Respiratory distress**
  - 27% Respiratory distress as a percentage of patients in the specified cohort (119 / 412 patients)

- **Vomiting co-occurrence and due to infectious disease**
  - 42% Vomiting co-occurrent and due to infectious disease as a percentage of patients in the specified cohort (175 / 412 patients)
## Comparison of COVID-19 Positive Patients

**Synthetic COVID-19 demo data**

**Date:** Aug 27, 2020 9:33:40 AM

**Notes:**

Click here to add notes.
Share this analysis

Enter one or more email addresses below.

Mhgarthi@health.ucsd.edu

Could you look at this comparison and give me your opinion?

Share  Cancel

Notes:
This is interesting because these patients seem to have a higher incidence of respiratory distress
Gene expression UMAP and clustering

This protocol allows you to perform UMAP embedding for selected samples and selected genes, followed by clustering with k-means or DBSCAN.

Instructions

- Background cohort
- Gene subset
- Low gene expression filter
- Dimension reduction parameters
- Clustering parameters

Selected parameters

- Low gene expression filter
  - Remove genes without expression in at least this many samples
- Dimension reduction parameters
  - Nearest neighbors parameter for UMAP
  - Number of dimensions to produce in projection
- Clustering parameters
  - Clustering method: k-means
  - Number of clusters to produce
  - DBSCAN radius

Run protocol
Cohort builder

Demographics

Cancer

Cancer type
Use these checkboxes to reduce the background cohort to only subjects from selected Cancer type variables.

- [ ] Select all
- [ ] Invasive Breast Carcinoma 1084 samples
- [ ] Non-Small Cell Lung Cancer 1053 samples
- [ ] Colorectal Adenocarcinoma 594 samples
- [ ] Glioblastoma 592 samples
- [ ] Endometrial Carcinoma 586 samples
- [ ] Ovarian Epithelial Tumor 585 samples
- [ ] Head and Neck Squamous Cell Carcinoma 523 samples
- [ ] Esophageal Adenocarcinoma 514 samples
- [ ] Diffuse Glioma 513 samples
- [ ] Renal Clear Cell Carcinoma 512 samples

10967 samples  Cancel  Save & use
TCGA pan-cancer gene expression UMAP and clustering

This protocol analyzed 10,967 samples using UMAP embedding, followed by clustering.
TCGA pan-cancer gene expression UMAP and clustering

This protocol analyzed 10,967 samples using UMAP embedding, followed by clustering.

Scatterplot cohort

Please specify a name for this cohort:

Inv Brst

Cancer type: Invasive Breast Carcinoma

1084 samples

Cancel  Save cohort
TCGA pan-cancer gene expression UMAP and clustering

This protocol analyzed 10,967 samples using UMAP embedding, followed by clustering.
<table>
<thead>
<tr>
<th>Tag score</th>
<th>Tag</th>
<th>Expression</th>
<th>Effect size</th>
</tr>
</thead>
<tbody>
<tr>
<td>750</td>
<td>SCGB2A2</td>
<td>+7.99</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>PIP</td>
<td>+6.33</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>LMX1B</td>
<td>+6.27</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>GATA3</td>
<td>+5.87</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>ANKR030A</td>
<td>+5.87</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>AZGF1</td>
<td>+5.86</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>RPS4Y1</td>
<td>-5.62</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>TFAP2B</td>
<td>+5.40</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>SPDIF1</td>
<td>+5.39</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>PBLR</td>
<td>+5.39</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>MUC1</td>
<td>+5.14</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>AARD</td>
<td>+5.12</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>ADIPOD</td>
<td>+5.07</td>
<td></td>
</tr>
<tr>
<td>750</td>
<td>FOXA1</td>
<td>+4.95</td>
<td></td>
</tr>
</tbody>
</table>
Differential expression

Variable collection
Expression

Type here to filter 100 results

**SCGB2A2**

Expression

- **750**
  - p = 0
  - Higher than expected

9.51 Average expression of SCGB2A2 for patients in the focus cohort (1082 patients)

1.52 Average expression of SCGB2A2 for all patients in the background cohort (8414 patients)

0.342 Average expression of SCGB2A2 for patients NOT in the focus cohort (7332 patients)

**PIP**

Expression

- **750**
  - p = 0
  - Higher than expected

**Dimension 1**

**Dimension 2**
SCGB2A2

Expression

Tag score: 750

Expression:

9.51 Average expression of SCGB2A2 for patients in the focus cohort (1082 patients)

1.52 Average expression of SCGB2A2 for all patients in the background cohort (8414 patients)

0.342 Average expression of SCGB2A2 for patients NOT in the focus cohort (7332 patients)
SCGB2A2

Annotation

- **Chromosomes**: 11
- **Cytoband**: 11:0123
- **Ensembl RNA ID**: ENST000000227818.3, ENST000000225380.1
- **Ensembl gene ID**: ENSG00000110484
- **Ensembl protein ID**: ENSP00000227818.2, ENSP00000225380.1
- **GO Pathway**: GO:0030521: androgen receptor signaling pathway, GO:0008150: biological_process
- **GO Cellular component**: GO:0005615: extracellular space, GO:0005575: cellular_component
- **GO Molecular function**: GO:0030674: molecular_function, GO:0005515: protein_binding
- **Gene ID**: MGI
- **Gene name synonym**: mammaglobin A
- **Gene symbol**: SCGB2A2
- **Gene synonym**: VGB2, MGB1, PSBP1
- **Gene type**: protein_coding
- **Genbank accession**: AR015224
- **HGNC previous symbol**: MGB1, PSBP1
- **HGNC symbol**: SCGB2A2
- **HGNC synonym**: MG071974, UGS2
- **Locus group**: protein_coding gene
- **Locus type**: gene with protein product
- **Modification date**: 2013/01/13
- **Mouse genome database ID**: MGI:03780828
- **Orientation**: Positive strand
- **Other database ID**: HUGO HUGO:7030, MIM:605852, Ensembl:ENSG00000110484
- **Other gene designation**: prostate steroid binding protein 1, mammaglobin A, mammaglobin 1
- **PubMed ID**: 17102571, 12821983, 27477018, 28416596, 31347781, 22790808, 26227677, 16117060, 22343676, 2523111, 20050026, 17070415, 21744998, 1888421, 16205799, 20092039, ... (44 more)
- **Refseq RNA**: XM_005274005.9, XM_0052411.4
- **Refseq gene**: SCGB2A2
The Cancer Genome Atlas (TCGA) Datasets

TCGA Pan-Cancer Atlas and the Immune Landscape of Cancer

Combined data from 33 cancer types from the 2018 TCGA Pan Cancer Clinical Data Resource

METABRIC Breast Cancer


Head and Neck Cancer (TCGA)

A dataset with clinical and multi-omics data for 530 head and neck cancer patients (from TCGA).
Summary

Mike Hogarth MD, Clinical Research Information Officer, UC San Diego Health
A tool for data exploration and analysis

- We have installed the tag.bio system in our research cloud and it has access to data sets in our ‘secure data commons database’
- The Nightingale portal provides population level access and ability to perform analysis
- A user can ‘slice’ the cohort and select specific analyses (demographic, survival, comparison between cohorts)
- Planned, pending approval, provide ‘download’ of limited data set (LDS) row-level data from selected data set into the investigator’s virtual research desktop for further analysis
Help us evolve the mesh

● What other registries should be available?
● How would you like to query them?
● Are there public data sources you would like to see here?
● Could we use the mesh for other data sources?

Please contact Mike Hogarth at mihogarth@health.ucsd.edu with suggestions or comments.
Thank You!

Questions?
Next presentation

Come see our next UC TECH Presentation 9/03:

Email Overload: Practical Tools for Influencing Email Volume in the Age of Telecommuting

Are you overwhelmed by the number of emails you receive daily? Has email management become a burdensome core task that monopolizes your time? When volume exceeds 200+ emails a day, generic email tips & tricks for email management simply won’t cut it. This session will go beyond email platform use to focus on email management from a behavior modification and process improvement perspective. We will cover practical tools and strategies for actively managing virtual work and interactions with your co-workers more effectively, giving you the ability to actually influence the volume of emails you receive.

Speaker:

Loralyn Cross, Office of Research Affairs, UC San Diego
Reference slides
Enabling doctors to provide instant answers

9 years of billing and encounter data

- All inpatient and outpatient data in a combined dataset

→ Over 2,000 analyses performed by value improvement physicians in the past year

“The ability to have this kind of on-demand information completely changes the culture. I can’t imagine doing my job without the Tag.bio platform.”

- Jahan Fahimi, Director of Value Improvement at UCSF Health

Jahan Fahimi, MD, PhD,
Associate Professor of Emergency Medicine,
Director of Value Improvement at UCSF Health
## What does this enable?

<table>
<thead>
<tr>
<th>Node</th>
<th>Development</th>
<th>Collaboration</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Data</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Is a Data Product</td>
<td>Build a rapid Data Product</td>
<td>SMART API</td>
</tr>
<tr>
<td>Domain-specific analysis</td>
<td>Iterative dev cycle</td>
<td>Secure data</td>
</tr>
<tr>
<td>Immutable data and transient node</td>
<td>Integrate other functions (R, Python, ML)</td>
<td>Secure deploy</td>
</tr>
<tr>
<td>Deploy anywhere</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Mesh</strong></td>
<td>Transfer apps between nodes</td>
<td>Distributed querying</td>
</tr>
<tr>
<td>Many Nodes, many data types</td>
<td>Federated functionality</td>
<td>Centralized analysis</td>
</tr>
<tr>
<td>Node functional diversity</td>
<td>No down time</td>
<td>Public/Private nodes</td>
</tr>
<tr>
<td>Distributed analyses</td>
<td>An ecosystem of nodes</td>
<td></td>
</tr>
<tr>
<td>Network effect on data value</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Portal</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Publish data with analyses</td>
<td>Rapidly populated by an admin</td>
<td>Reproducible, replayable analyses</td>
</tr>
<tr>
<td>Track analyses through history</td>
<td>Cohorts into Nodes</td>
<td>Share analyses</td>
</tr>
<tr>
<td>Create COHORT/UDAT</td>
<td>Transactional Nodes</td>
<td>Versioned resources</td>
</tr>
<tr>
<td>Reference &amp; Annotation nodes</td>
<td>Usage allows evolution of mesh</td>
<td>De-silo analysis types</td>
</tr>
</tbody>
</table>