

Precisely Practicing Cancer Medicine from 700 Trillion Points of University of California Health Data

Atul Butte, MD, PhD

Chief Data Scientist, University of California Health (UC Health)

Priscilla Chan and Mark Zuckerberg Distinguished Professor

Director, Bakar Computational Health Sciences Institute, UCSF

atul.butte@ucsf.edu ▪ @atulbutte

Conflicts of Interest

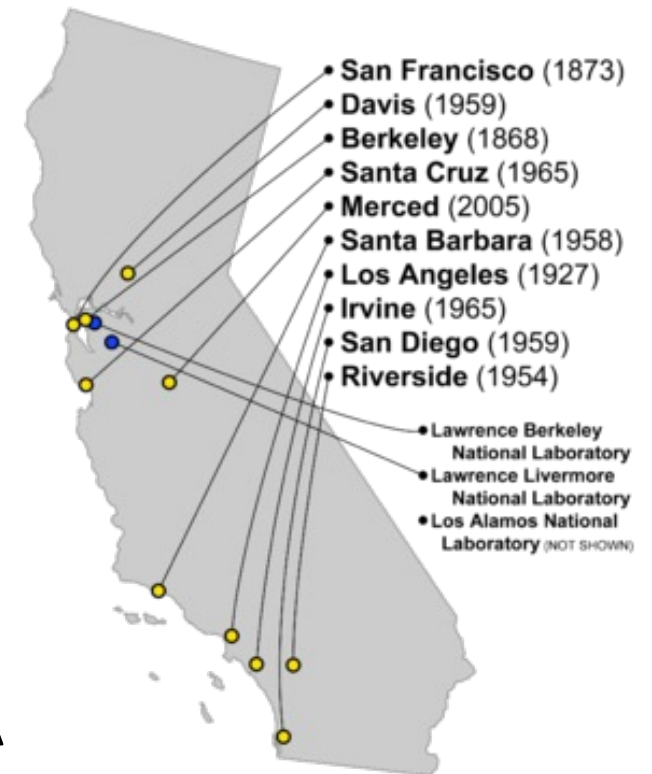
- Scientific founder
 - Personalis
 - NuMedii
 - Carmenta (Progenity)
 - Genstruct
 - Honoraria for talks
 - Lilly
 - Pfizer
 - Siemens
 - Bristol Myers Squibb
 - AstraZeneca
 - Roche
 - Genentech
 - Warburg Pincus
 - CRG
 - AbbVie
 - Westat
 - Past or present consultancy
 - Personalis
 - NuMedii
 - Lilly
 - Johnson and Johnson
 - Roche
 - Genstruct
 - Tercica
 - Ecoeos
 - Helix
 - Ansh Labs
 - uBiome
 - Prevendia
 - Samsung
 - Assay Depot
 - Regeneron
 - Verinata (Illumina)
 - Pathway Diagnostics
 - Geisinger Health
 - Covance
 - Wilson Sonsini Goodrich & Rosati
 - Orrick
 - 10X Genomics
 - GNS Healthcare
 - Gerson Lehman Group
 - Coatue Management
- Shares or Ownership
 - Johnson and Johnson
 - Optum
 - NuMedii (major)
 - Personalis (major)
 - Apple
 - Facebook
 - Alphabet (Google)
 - Microsoft
 - Amazon
 - Snap
 - 10x Genomics
 - Illumina
 - Nuna Health
 - Assay Depot (Scientist.com)
 - Vet24seven
 - Regeneron
 - Sanofi
 - Royalty Pharma
 - AstraZeneca
 - Moderna
 - Biogen
 - Paraxel
 - Sutro
 - Other corporate relationships
 - Northrop Grumman
 - Genentech
- Speakers' bureau
 - None
 - Companies started by students
 - Carmenta
 - Serendipity
 - Stimulomics
 - NunaHealth
 - Praedicat
 - MyTime
 - Flipora
 - Tumbl.in
 - Polyglot
 - Iota Health
 - Ongevity Health

University of California

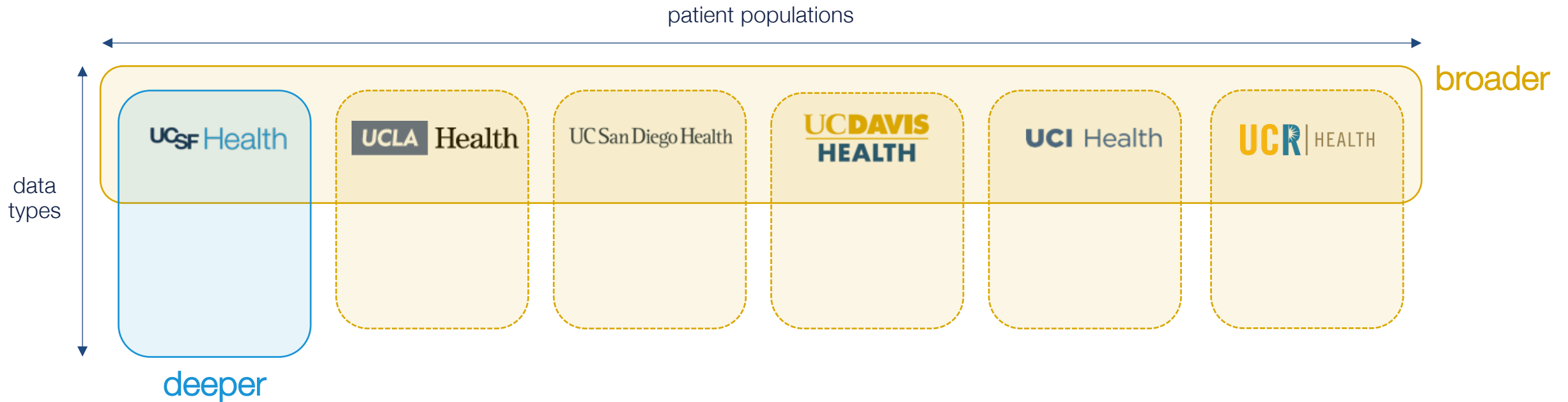
- 10 campuses and 3 national labs
- ~200,000 employees, ~250,000 students/yr

UC Health

- 20 health professional schools (6 med schools)
- Train half the medical students and residents in California
- ~\$2 billion NIH funding
- \$13+ billion clinical operating revenue
- 5000 faculty physicians, 12000 nurses
- UCSF and UCLA are in US News top 10
- 5 NCI Comprehensive Cancer Centers, 5 NIH CTSA
- IRB reliance, centralized contracting



Working with UCSF and UC-wide data



- Access to deidentified, limited, and identified structured and unstructured clinical data
- Access to additional datasets that are or can be linked to clinical data (e.g., imaging, 'omics, waveforms)

- Access to deidentified structured clinical data from a larger patient population ()
- Access to additional data sets possible (---) but requires additional work and time (e.g., recruitment and coordination of PIs at multiple sites, increased contractual complexity)

The University of California has an incredible view of the medical system

- Combined EHR data from UCSF, UCLA, UC Irvine, UC Davis, UC San Diego, and UC Riverside
- Central database built using open-source OMOP as a data backend
 - First EHR installation was January 2012
 - Structured data from 2012 to the present day
 - **8.7 million patients with “modern” data**
 - 378M encounters, 1.0B procedures, 1.3B+ med orders, 44M device uses, 1.4M providers, 1.1B diagnosis codes, 5.2B+ lab tests and vital signs
 - “From Tylenol to CAR-T cells...”
 - Merged with California state data, pathology and radiology text elements, CA death index
 - Claims data from our self-funded plans now included
 - Continually harmonizing elements
- Safe, respectful, regulated, responsible use of clinical data

UNIVERSITY OF CALIFORNIA
HEALTH

UCSF Health

UC DAVIS
HEALTH

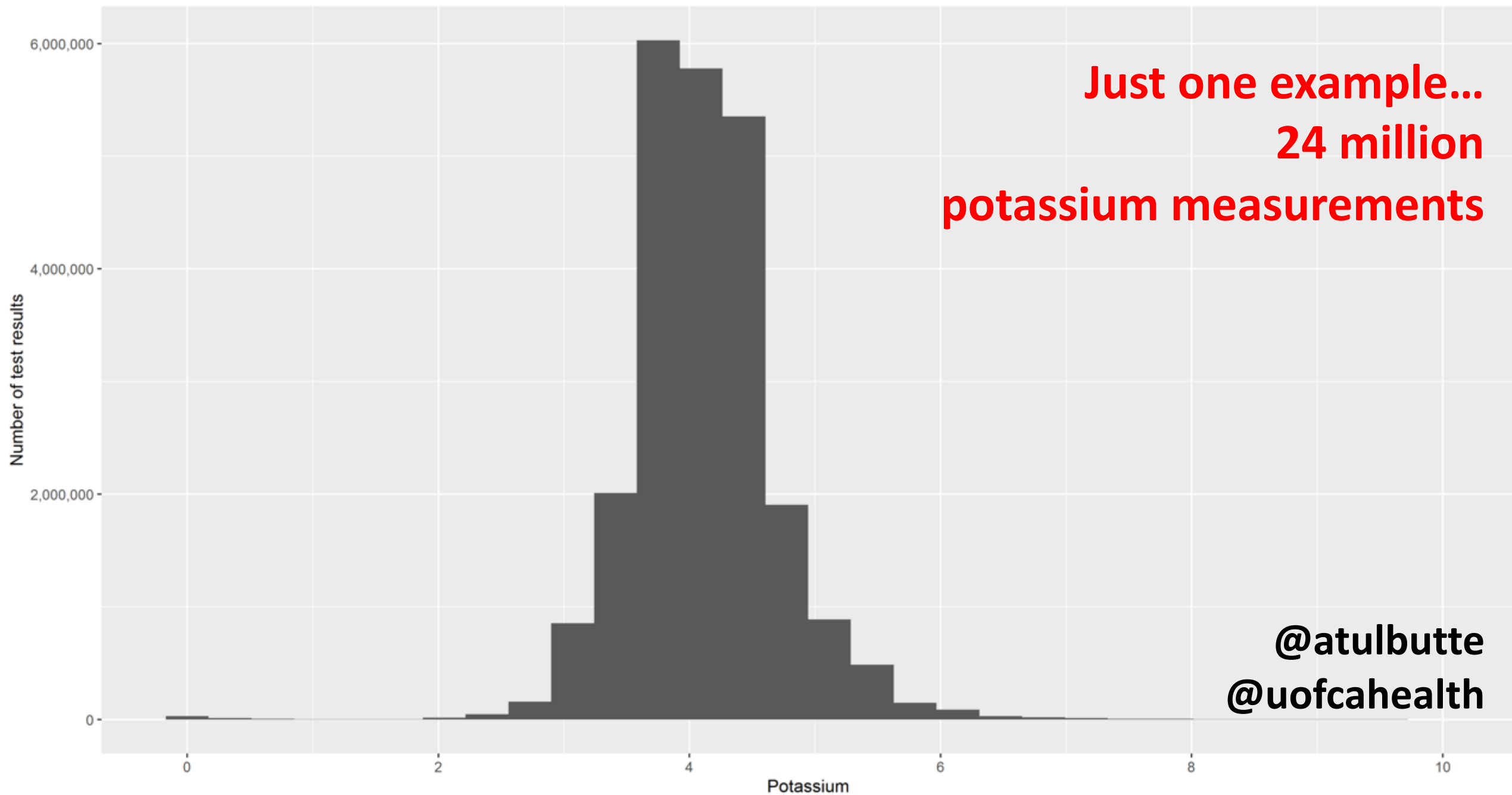
UCLA Health

UCI Health

UCR | HEALTH

UC San Diego Health

University of California Health Laboratory Test Results



**Just one example...
24 million
potassium measurements**

**@atulbutte
@uofcahealth**

100,000+ active cancer patients across UC Health

UNIVERSITY OF CALIFORNIA
HEALTH

UC Cancer Demographics

as of December 31, 2022

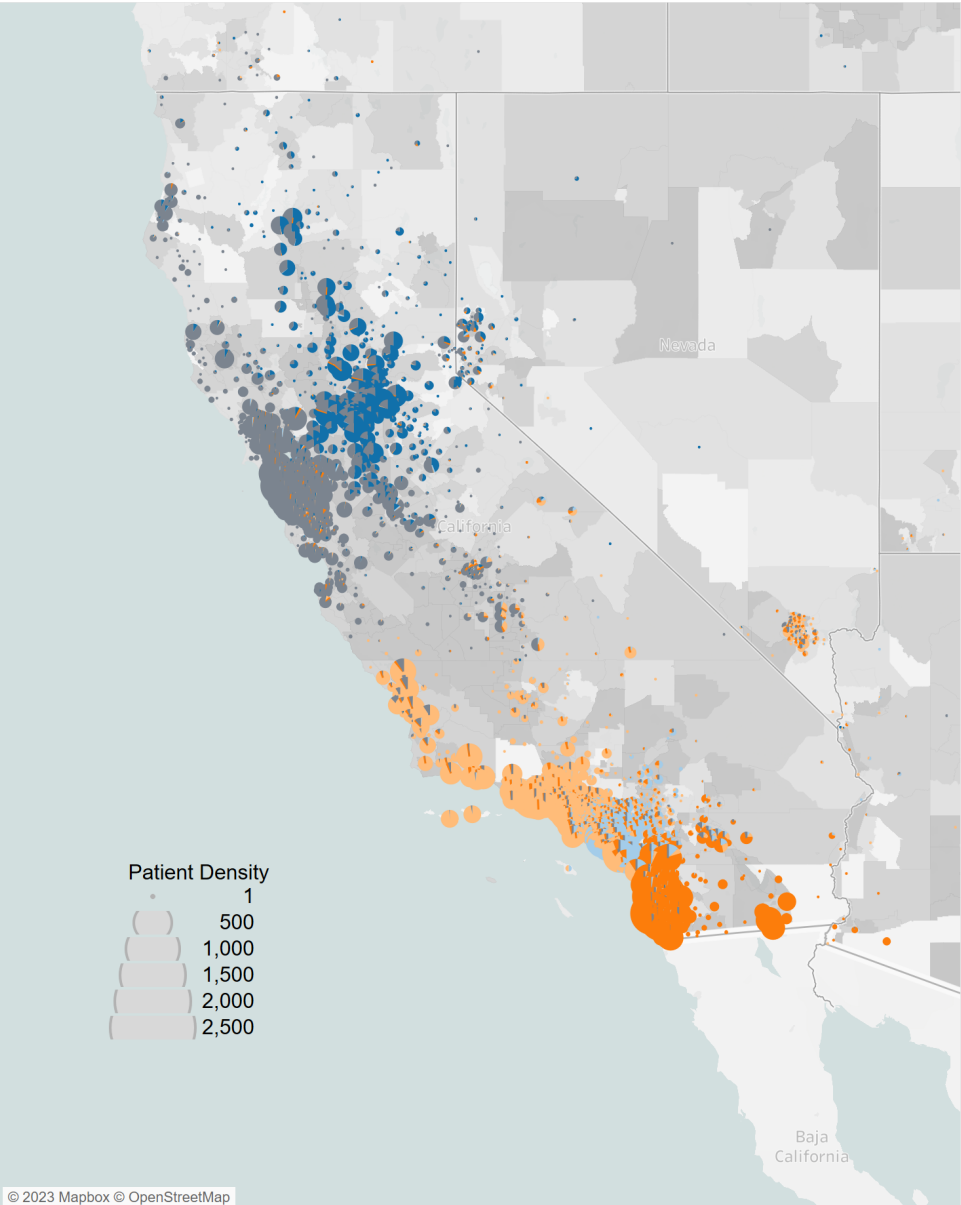
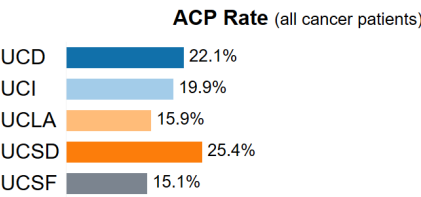
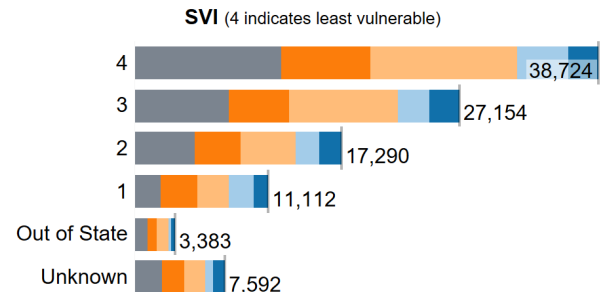
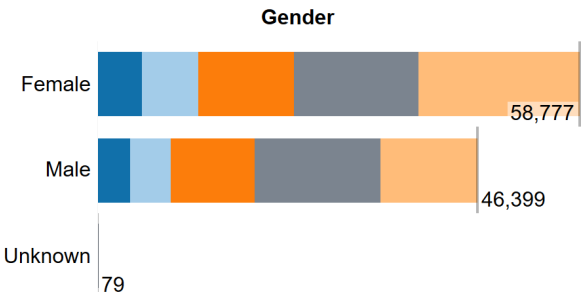
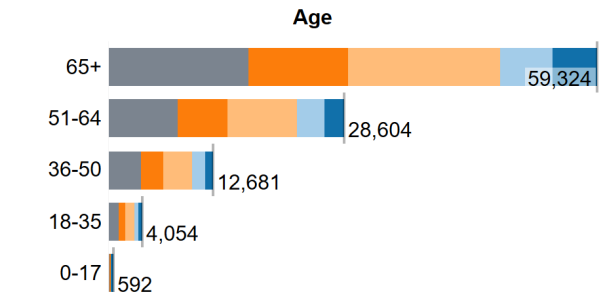
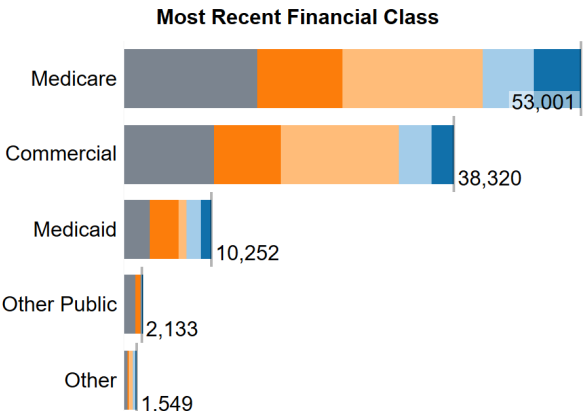
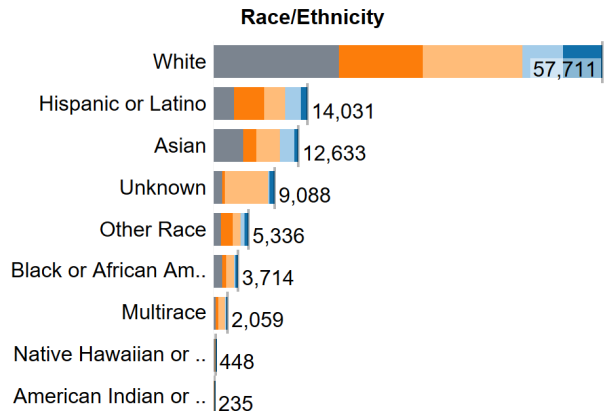
UC site
All

Department
All

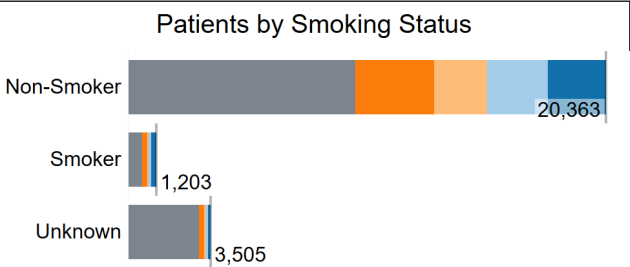
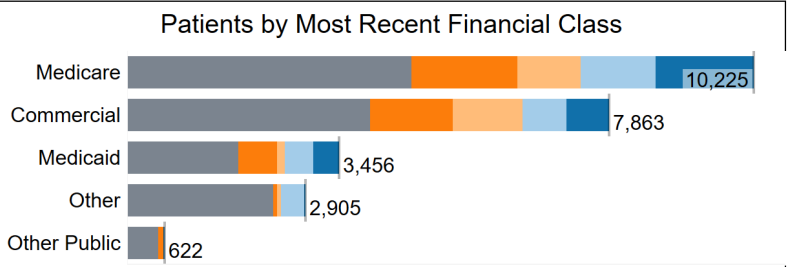
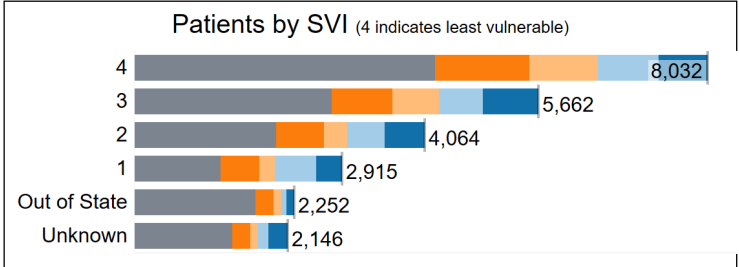
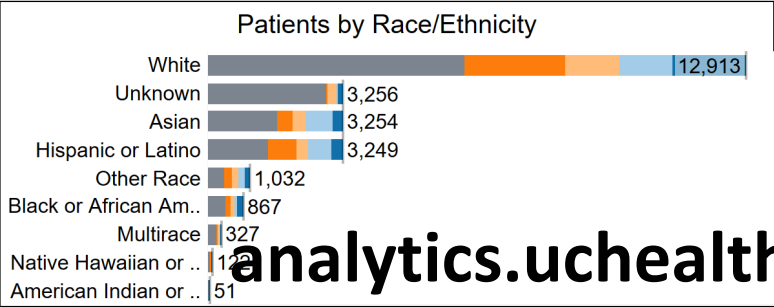
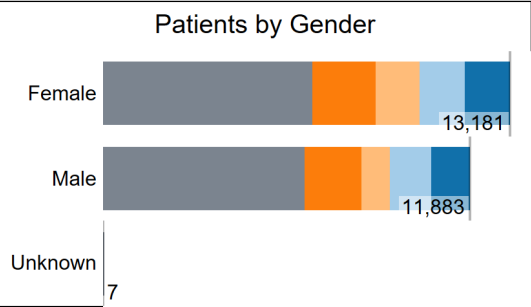
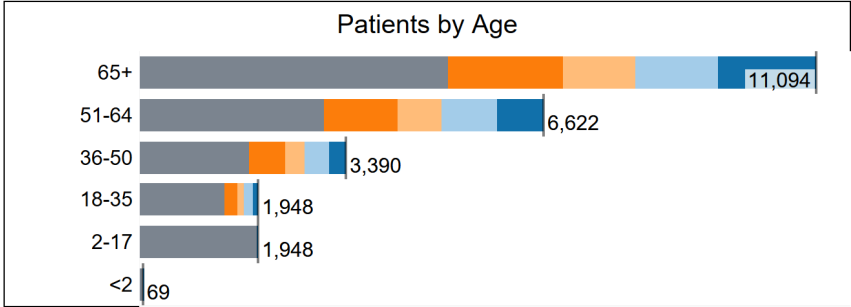
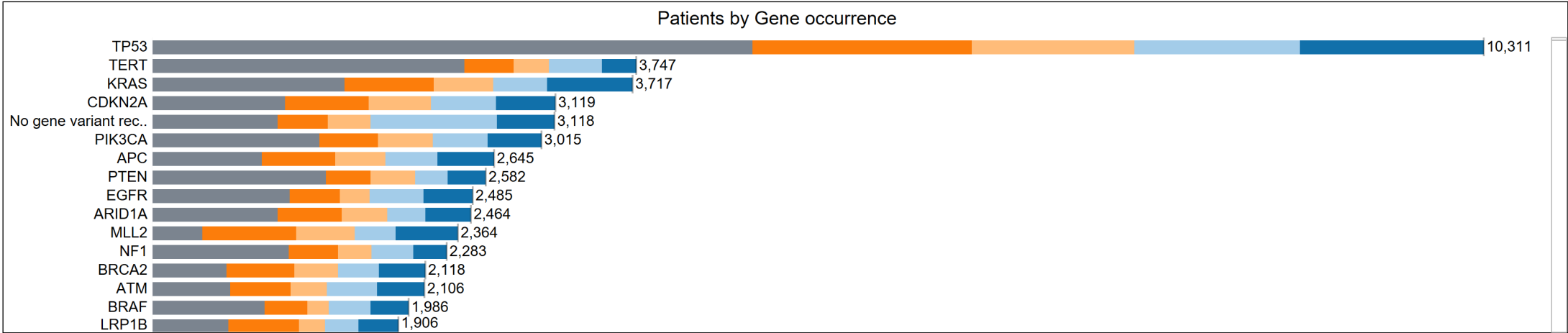
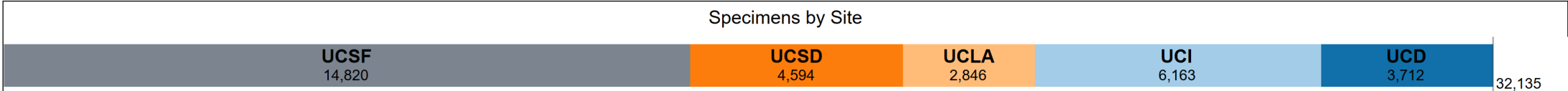
Cancer Group
All

Disease Lookup
All

105,255 Total UC patients with a cancer diagnosis and
Two Visits in Prior Year OR Two Consecutive Annual Visits



32+ thousand cancer genomic reports incorporated into the same database



View by cancer or by gene mutation

UNIVERSITY OF CALIFORNIA
HEALTH

Cancer Genomics Explorer - Disease Occurrence

Updated through December 31, 2022. 32,135 samples included.

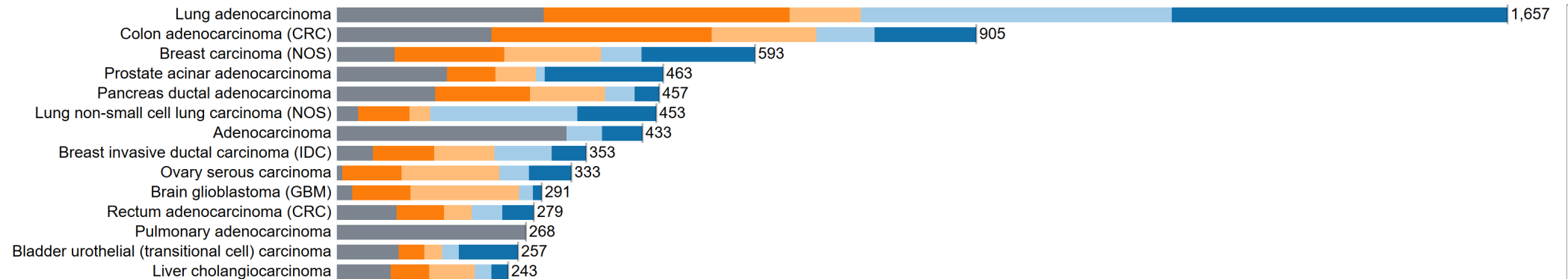
All Genes All Testing Labs

Specimens by Site

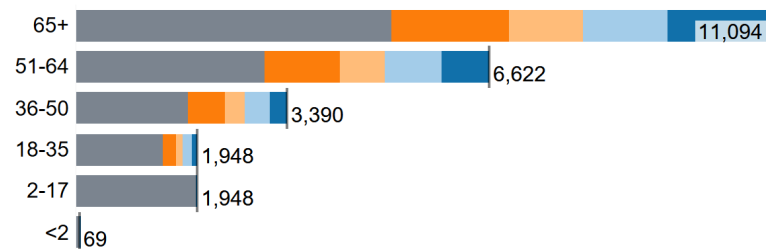


Disease occurrence

Click a disease to filter demographics



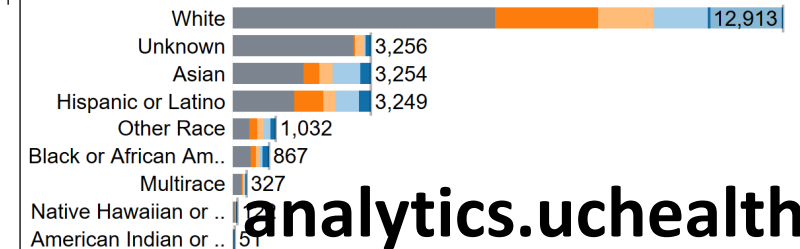
Patients by Age



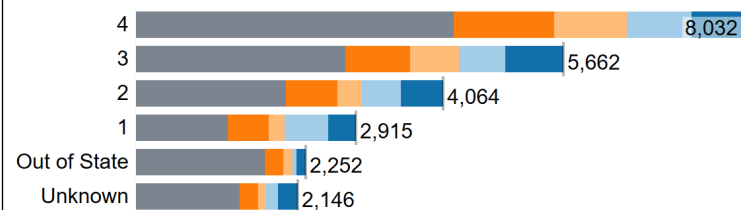
Patients by Gender



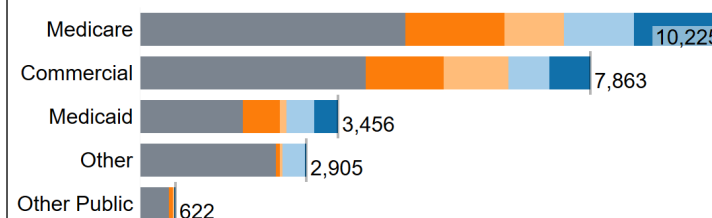
Patients by Race/Ethnicity



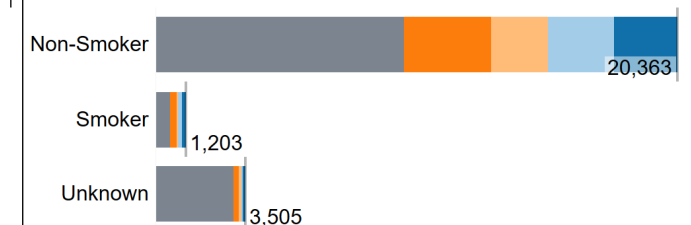
Patients by SVI (4 indicates least vulnerable)



Patients by Most Recent Financial Class



Patients by Smoking Status



analytics.uchealth.edu

For example, only the cancers with an NF1 mutation

UNIVERSITY OF CALIFORNIA
HEALTH

Cancer Genomics Explorer - Disease Occurrence

Updated through January 31, 2023. 32,797 samples included.

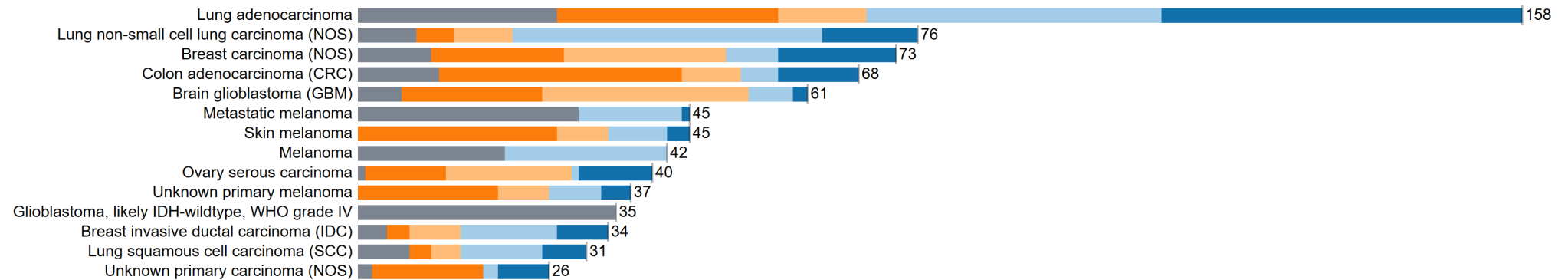
Genes
NF1 All Testing Labs

Specimens by Site

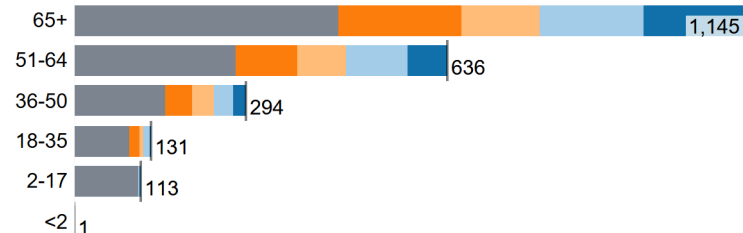


Disease occurrence

Click a disease to filter demographics



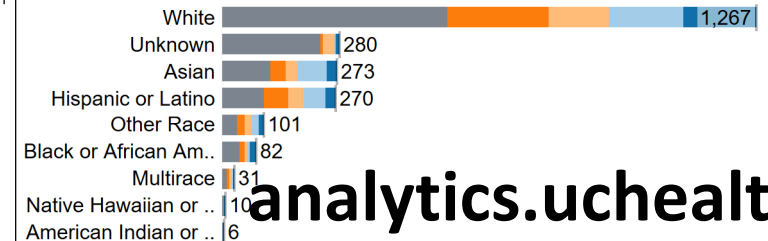
Patients by Age



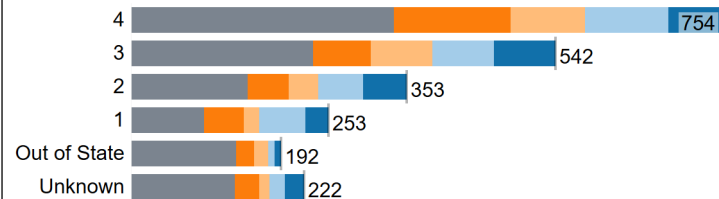
Patients by Gender



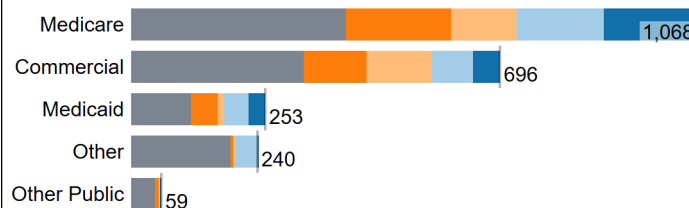
Patients by Race/Ethnicity



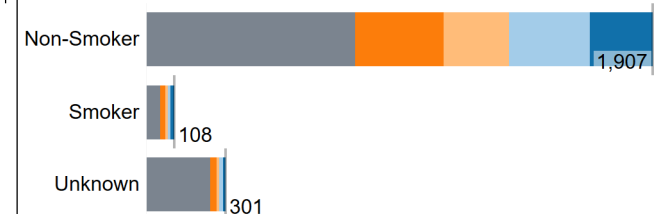
Patients by SVI (4 indicates least vulnerable)



Patients by Most Recent Financial Class

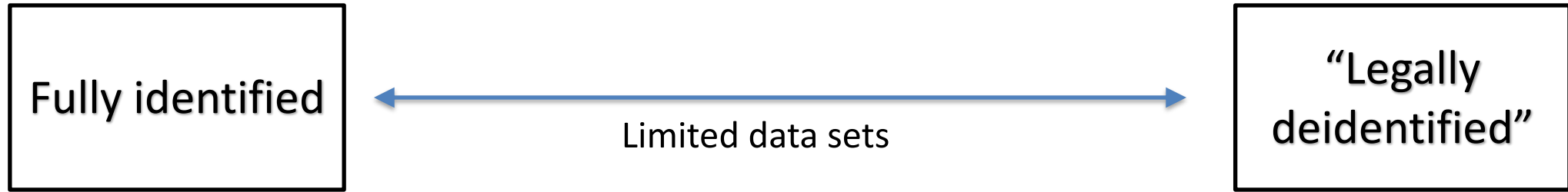


Patients by Smoking Status



analytics.uchealth.edu

How does research access work across UC Health?



- Researchers should first write and optimize OMOP SQL queries locally, on their own campuses
- When ready to scale (and authorized), we spin up a virtual machine for the researcher, populated with common tools
 - Electronically sign a UC Health data use agreement
 - R, Tableau, Jupyter Notebooks, Julia, SQL, Windows or Linux available
- Upload your scripts and run, but cannot download data
- Safe, respectful, regulated research use of clinical data

Safe, Respectful access to Deidentified Data now available through Cloud-based Databricks (single sign on with UC Health campus credentials)

Microsoft Azure

 databricks

Search

⌘ + P

uchdw-db-prod-worksp...

?

atul.butte@ucsf.edu

Main UCHDW queries

SQL

File

Edit

View

Run

Help

Last edit was 24 days ago

Give feedback

Connecting

Schedule

Share

Count encounters

Cmd 15

1 %sql

2

3 select visit_concept_id, concept_name, count(*) as count

4 from omop_deid.visit_occurrence as vo join omop_deid.concept as co on vo.visit_concept_id = co.concept_id

5 group by visit_concept_id, concept_name

Cancel

Waiting for cluster to start: Starting Spark

Table

	visit_concept_id	concept_name	count
1	9202	Outpatient Visit	30143864
2	9203	Emergency Room Visit	3550274
3	9201	Inpatient Visit	1890330
4	581477	Office Visit	93381430
5	0	No matching concept	2

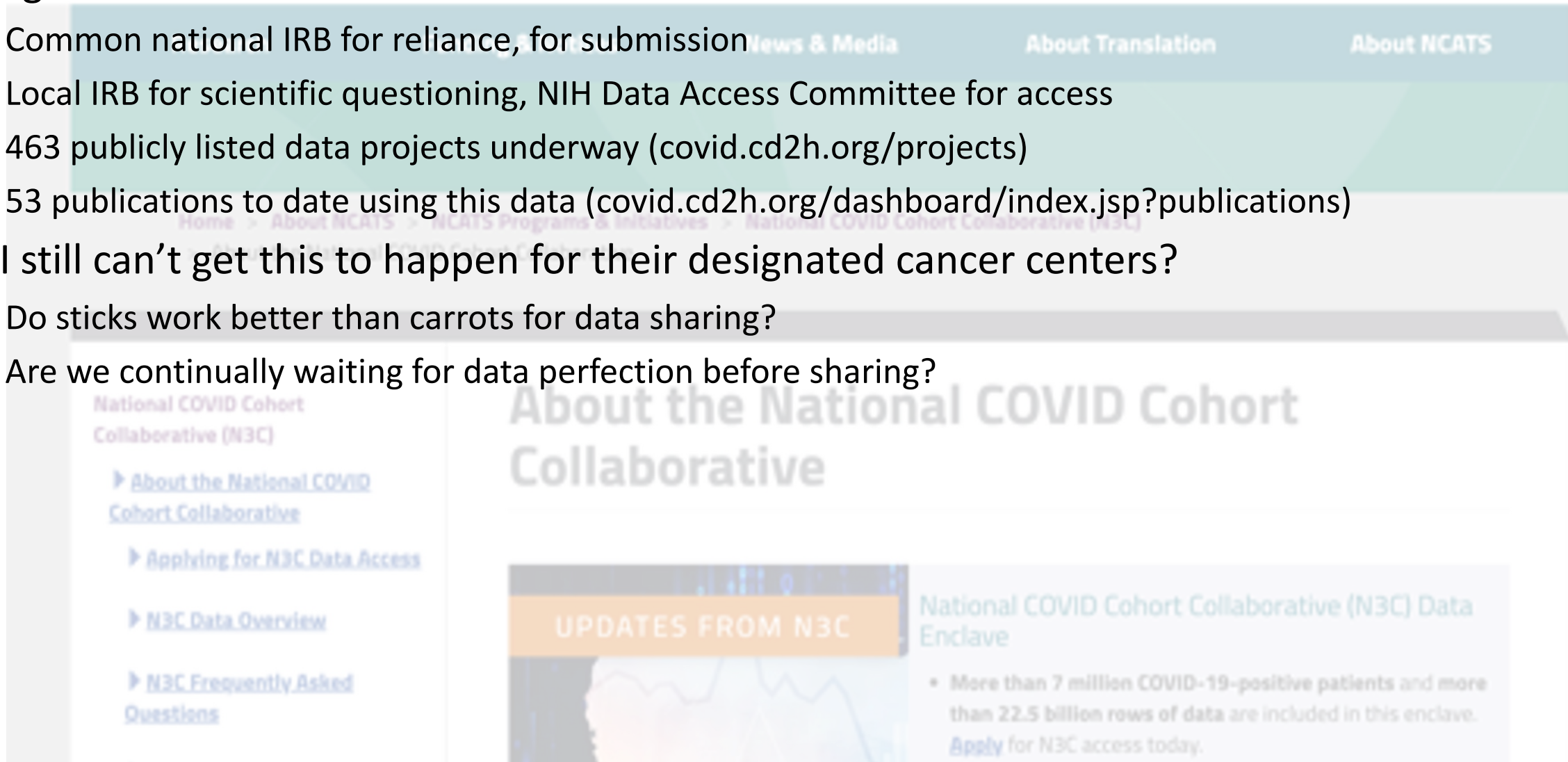
Showing all 5 rows.

analytics.uchealth.edu

Refreshed now

NCATS/NIH National COVID Cohort Collaborative (N3C)

- How was NCATS able to motivate 100+ health systems (especially their CTSA programs) to sign and transfer deidentified EHR data around SARS-CoV-2 tested individuals?
 - Common national IRB for reliance, for submission
 - Local IRB for scientific questioning, NIH Data Access Committee for access
 - 463 publicly listed data projects underway (covid.cd2h.org/projects)
 - 53 publications to date using this data (covid.cd2h.org/dashboard/index.jsp?publications)
- NCI still can't get this to happen for their designated cancer centers?
 - Do sticks work better than carrots for data sharing?
 - Are we continually waiting for data perfection before sharing?



N3C Dashboards

Reliable High-Velocity COVID-19 Insights Brought to you by the N3C.

[Explore Dashboards](#)

[Publications](#)



The N3C Data Enclave represents one of the largest secure collections of harmonized clinical health data in the United States.



Sites: 77



Persons: 18.2 million



COVID+ Cases: 7,152,385



of Rows: 22.9 billion



Clinical Observations:
2.1 billion



Lab Results: 10.9 billion



Medication Records:
3.6 billion



Procedures: .9 billion



Visits: 1.2 billion

Explore Our Dashboards

Overview



Members



Tracking



Cases



Medications



Diseases



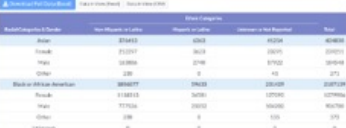
Institutions Contributing Data

The N3C allows medical sites within the United States to securely transfer anonymized data into the Enclave. The average interval for data transfer from our partners is once a week. To explore the geographic coverage of our current partners, please see the map below.

Related Dashboards



Regional Distribution of COVID+ Patients

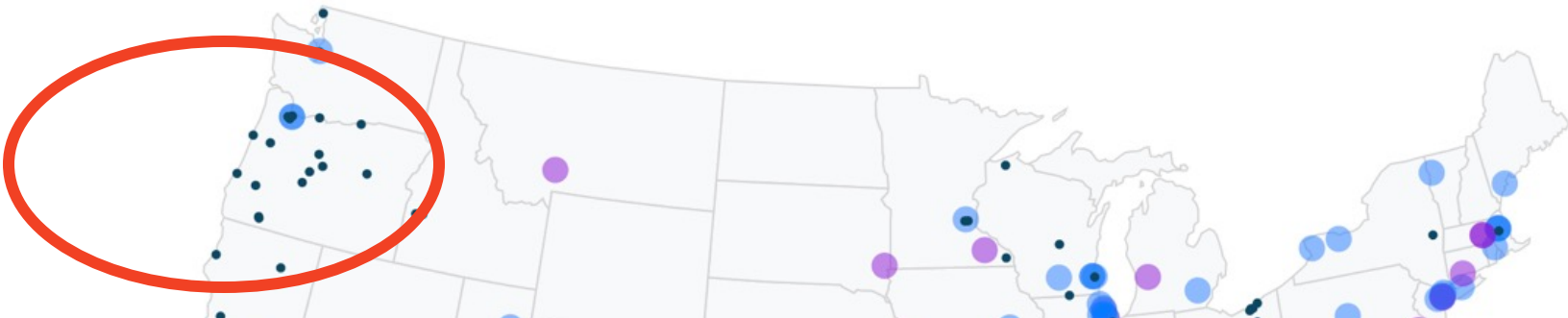


Demographics	Number of Sites	Number of Patients	Number of Sites
Age	21,013	1,025	1,024
Gender	21,013	1,025	1,024
Race	21,013	1,025	1,024
Ethnicity	21,013	1,025	1,024

Demographics Table for IRB Submission

N3C Contributing Sites

- Data Available
- Data transfer signed, pending availability
- OCHIN contributing site



Demographics

Learn more about the N3C Data and explore demographics, comorbidities, and vaccination counts.

Select a Topic to Explore:

Demographics of COVID+ Patients

Display



Filters

- Age +
- Race +
- Ethnicity +
- Sex +
- Severity +

Related Dashboards

Long COVID

Demographics Table

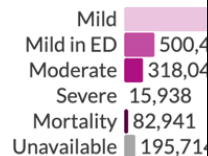
Total COVID+ Patients ¹
+ 7.15M

Data as of Mar 02, 2023 (v113)

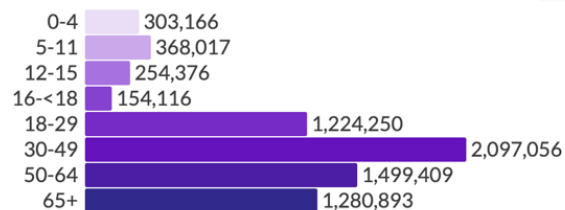
Patients in Dashboard ¹
7.18M

% of Total COVID+ Patients in Dashboard ¹
100.40%

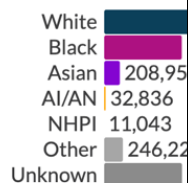
Severity



Age



Race



Sex



Ethnicity



Sample: COVID+ patients ¹ in the N3C Data Enclave. Data aggregated by Age, Race, Ethnicity, Sex, and limitations below.

covid.cd2h.org/dashboard

Mortality

This dashboard provides data on patients with dates of death occurring after hospitalization for COVID or Long COVID within the Enclave.

Select a Topic to Explore:

COVID+ and Hospitalization

Display



Filters

- Age +
- Race +
- Sex +
- Ethnicity +

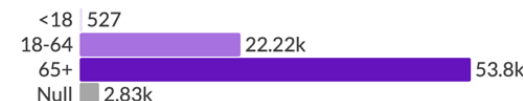
Related Dashboards

Total COVID+ Patients ¹
+ 7.15M

Data as of Mar 02, 2023 (v113)
[* See Limitations Below](#)

Mortality Total*
79.37k

Counts of Mortalities by Age



Counts of Mortalities by Race



Counts of Mortalities by Sex



Counts of Mortalities by Ethnicity



Ongoing challenges and opportunities

- Health data interoperability/interchange is solvable, especially if there's a business reason for it
- Structured data elements from unstructured data is still a challenge
 - Expensive human curators now, artificial intelligence soon enough
- Mapping patient care trajectories to actual protocols (and deviations) is still hard
- Just because you have all the clinical and molecular data in one place does not mean people use it
 - Need better (easier) tools on top of these datasets, ideally within other tools being used (EHR)
 - Need to integrate the use of computational/data tools into the molecular cancer workflow
- Still hard to inspire cross-campus clinical trials: incentives not enough to push against all the resistance?
- We essentially have zero cancer precision medicine tools for patients
- Can we learn from the NIH/NCATS N3C experience?



RICK LARSEN
OKSANA GOLOGORSKAYA
NELSON LEE
SANA SWEIS
THOMAS HURAY

UC San Diego Health

JENNIFER HOLLAND
DOUGLAS MACLEOD
PETER RYAN
HIRAM CARDOZA
ANDY LUCAS
Priscilla Jayaprakash
CALVIN FONG

UCI Health

DAVID MERRILL
NORA LEWIN
KATHY PICKELL
LEANIE MAYOR
MELODY HILL
NEAKTISIA LEE

UNIVERSITY OF CALIFORNIA HEALTH

CARRIE BYINGTON
ATUL BUTTE
CORAN HAN
PAGAN MORRIS
MONTE RATZLAFF
MIKE KILPATRICK
LISA DAHM
ANDENET EMIRU
JENNIFER BENBOW
EMERICA AGOSSA
AYAN PATEL
AIDEN BARIN
CHAYA MOHN
RAY PABLO
TIM HAYES
DAVID GONZALEZ
ROB FOLLETT
TEJU YARDI
NADYA BALABANOVA

UCLA Health

ALBERT DUNTUGAN
ANDREW WEAVER
Yael BERKOVICH
JAY SHAH
VAJRA KASTURI
PALLAVI MYNAMPATI
SAAJID FAZIL
BILL LAZARUS
EDGAR TIJERINO
BILL CINNATER
JOSH PELINO

The CIO Team

ELLEN POLLACK (UCLA)
JOSH GLANDORF (UCSD)
SCOTT JOSLYN (UCI)
JOE BENGFORT (UCSF)
ASHISH ATREJA (UCD)
TOM ANDRIOLA (UCI)

UC DAVIS HEALTH

KENT ANDERSON
PUNEET GILL
HEMANATH TATIPARTHI
SUPRAJA RADHAKRISHNAN
JODI NYGAARD
JEFFREY STERETT
RALPH PERRIN
THOMAS AMI

Support

- University of California, San Francisco
 - Priscilla Chan and Mark Zuckerberg
 - Barbara and Gerson Bakar Foundation
 - NIH: NIAID, NLM, NIGMS, NCI, NHLBI, OD; NIDDK, NHGRI, NIA, NCATS, NICHD
 - Food and Drug Administration
 - California Governor's Office of Planning and Research
 - Howard Hughes Medical Institute, California Institute for Regenerative Medicine
 - March of Dimes, Juvenile Diabetes Research Foundation
 - Hewlett Packard, L'Oreal, Progenity, Genentech, Janssen
 - Intervalien Foundation, Leon Lowenstein Foundation, Scleroderma Research Foundation, Clayville Research Fund, PhRMA Foundation, Stanford Cancer Center, Bio-X, SPARK
-
- Tarangini Deshpande
 - Kimayani Butte
 - Carrie Byington, Talmadge King, Mark Laret
 - Jack Stobo, Sam Hawgood, Keith Yamamoto
 - Isaac Kohane

Admin and Tech Staff

- Boris Oskotsky
- Andrew Jan
- Mounira Kenaani
- Andrew White
- Pam Allarde
- Amber Nolan