

Breakout Session 1: Track A

Strategies for Improving the Readiness of Large-Scale Cohort Data for AI/ML

Dr. James V. Lacey Jr.
Professor, City of Hope



Strategies for Improving the Readiness of Large-Scale Cohort Data for AI/ML

James V. Lacey, Jr., Ph.D.
Professor
City of Hope

**“A More Perfect Union: Leveraging Clinically Deployed Models
and Cancer Epidemiology Cohort Data to Improve AI/ML
Readiness of NIH-Supported Population Sciences Resources”**

3U01-CA199277-08S2: J. Lacey & M.E. Martinez, MPIs

Disclosures



James V. Lacey, Jr., Ph.D.
Professor
City of Hope, Duarte, CA

- **NIH grant funding**
 - **MPI on U01-CA199277**
 - **PI on 3U01-CA199277-09S1**
 - **Co-I on NCI & NIEHS R01s**
- **Named co-inventor on pending patent for cohort-selection web application**

Problem Statement

- **Cancer Epidemiology Cohorts (CECs):** Unique yet underutilized NIH-funded resources
 - *NCI has invested over \$1B in over 30 CECs over last few decades*
 - *Millions of high-quality data points on lifestyle, PROs, environment, biospecimens, claims, & more*
- **Three big challenges:** Most* CECs' data are...
 - *Local & siloed, rather than FAIR*
 - *Stored & structured for manual & individual analyses*
 - *"Shared" primarily as limited datasets*

Epidemiology and Genomics Research Program

[Home](#) [Funding & Grants](#) [Research Areas](#) [Research Resources](#) [News](#) [Events](#)

Cancer Epidemiology Cohorts

[Home](#) / [Research Resources](#) / Cancer Epidemiology Cohorts

RESEARCH RESOURCES

[All Research Resources](#)

[Biospecimens](#)

[Cancer Epidemiology Cohorts](#)

[NCI Cohort Consortium](#)

[Cancer Registry Resources](#)

[Consortia](#)

[Dietary Assessment Resources](#)

[Genomic Summary Results for Cancer Research Studies](#)

[Physical Activity Assessment Resources](#)

[Research Highlights](#)

[Statistics](#)

[Surveys](#)

On this page...

- [Overview](#)
- [Funded Projects](#)
- [Related Research Resources](#)
- [Related Workshops and Webinars](#)

Overview

Cohort studies are one of the fundamental designs for epidemiological research. Cancer epidemiology cohorts are large observational population studies in which groups of people with a set of characteristics or exposures are prospectively followed for the incidence of new cancers and cancer-related outcomes. Data from cohort studies have helped researchers to better understand the complex etiology of cancer, and have provided fundamental insights into key environmental, lifestyle, clinical, and genetic determinants of this disease and its outcomes.

California Teachers Study (CTS)

www.calteachersstudy.org

ABOUT US

The California Teachers Study (CTS) was founded in 1995 when 133,477 teachers, administrators, school nurses, and other members of the California State Teachers Retirement System (CalSTRS) agreed to provide information about their health and behaviors to CTS researchers. In the years that have followed, participants have continued to provide important information about their health and lives.

In addition to enabling widespread research on breast cancer, the data provided by participants has allowed CTS researchers to study the causes of multiple other cancers and diseases.

1995

Year
Established

133,477

Participants
Enrolled

4

Partner
Institutions

200+

Academic
Publications

Vast, Diverse, & High-Quality Participant Data



Over 500,000 surveys from 133,477 women

Lifestyle, family history, social support, physical activity, diet, & PROMIS from up to 6 surveys each



25+ year residential history

Linked individual-level geospatial data on the built environment, Census data, air pollution, oil & gas exposures, climate change, & other environmental exposures



Biospecimens from over 23,000 participants

Over 660,000 research-ready aliquots of serum, plasma, RBCs, & clots

Over 1700 WGS, plus genotyping arrays



Over 40,000 cancers

Type, date, tumor details, and initial treatment



Over 780,000 hospital visits

Detailed diagnostic, procedure, and healthcare utilization data from inpatient, emergency department, and ambulatory surgery visits



Over 1 million linked Medicare records

Provider, claims, home health, & skilled nursing for FFS & Medicare Advantage



Over 38,000 linked death records

Date and causes of death

Innovative & Cloud-Based Secure Collaboration Workspace



Suite of statistical analysis & GIS software



Custom interactive visualizations



AI/ML know-how, workflows, tools, & resources



Secure data warehouse & data lakes



Integrated project management & tracking



Self-service cohort-selection tool (Pat. Pend.)

Open & Democratic Approach



Equal access for everyone everywhere



Interactive visualizations on Tableau Public



Extensive documentation, code, and templates



User-friendly & efficient data-sharing tools



Github for broad dissemination & sharing



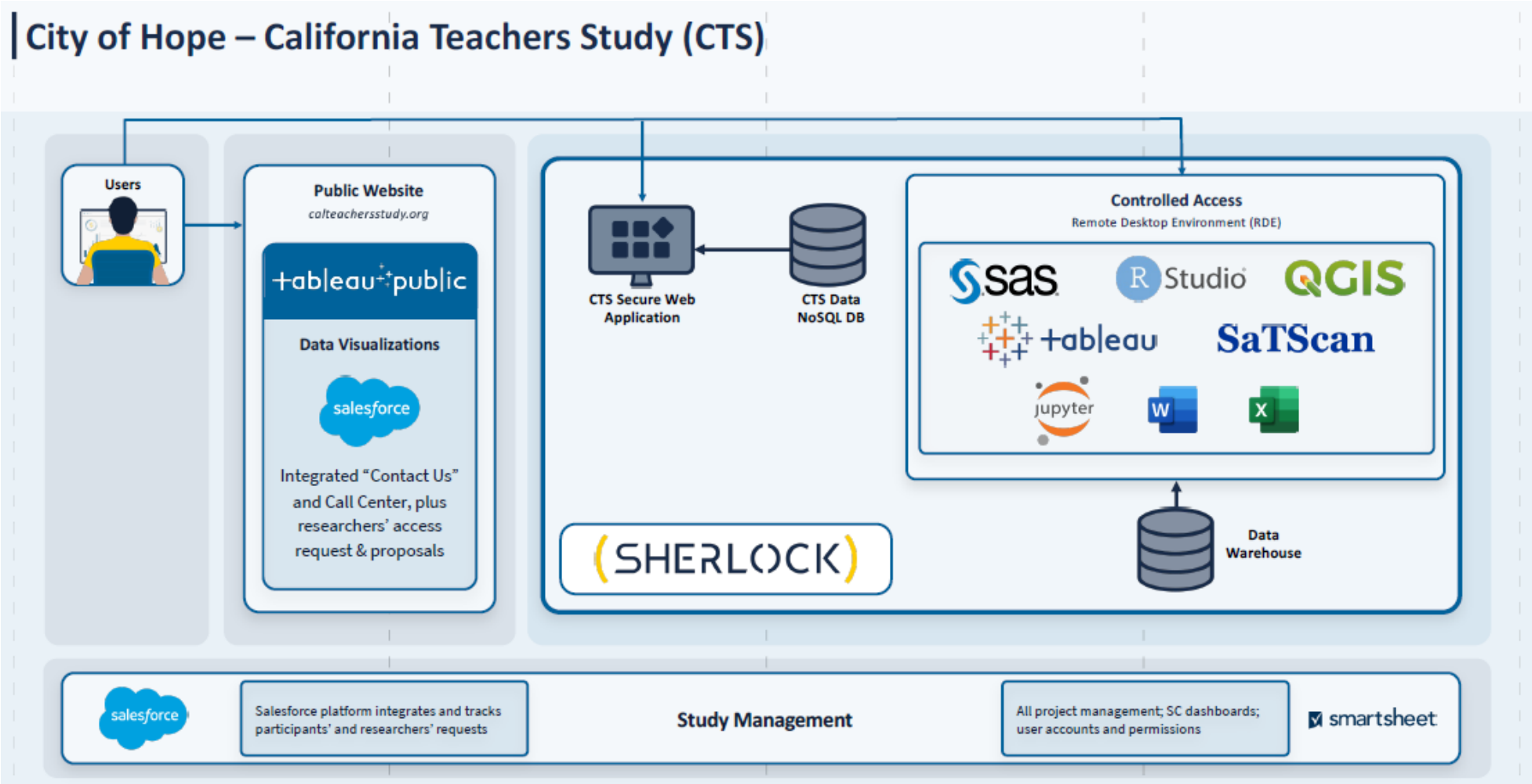
Community of over 200 users worldwide

Project Summary and Goals



- **Objective:** Increase the AI/ML readiness of CTS data
- **Specific Aim #1:** Deploy a secure and scalable computing environment for CTS data
 - With partners at the San Diego Supercomputer Center's (SDSC) Sherlock Cloud, expand our CTS data commons to include an AWS SageMaker environment for AI/ML applications
 - *Goal: Increase the readiness of our CTS environment for AI/ML*
- **Specific Aim #2:** Generate low-dimensional latent representations, or embeddings, for complex CTS survey data
 - Visualize clusters from over 5000 data columns for over 133,477 participants
 - *Goal: Identify additional phenotypes and help inform cohort discovery*
- **Specific Aim #3:** Test a City of Hope readmissions model in CTS data
 - Revised: train two new models focused on participants' survey response & lifespan
 - *Goal: Real-world model training on CTS data in new environment by data scientists*

Highlights: Aim #1



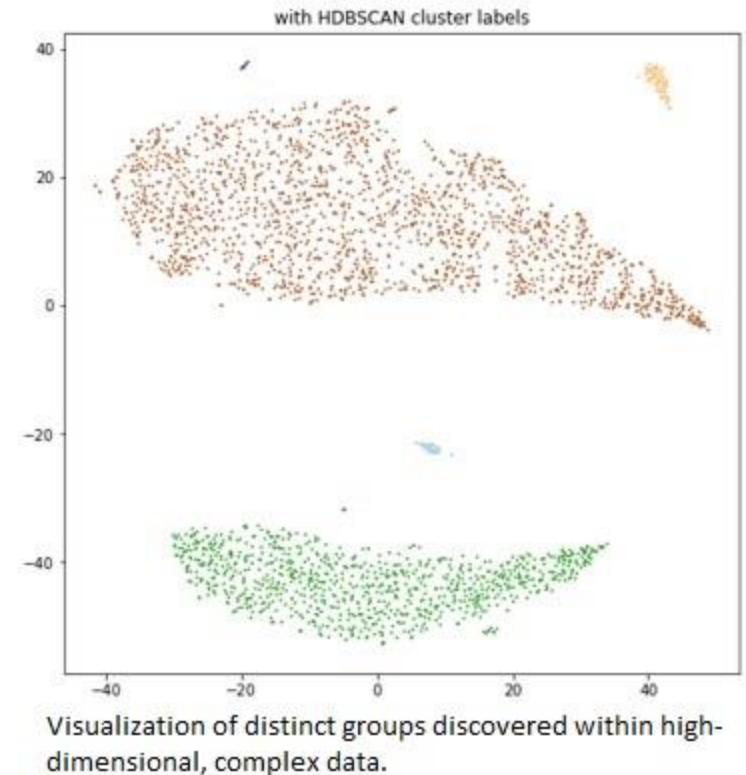
Expanded and Enhanced Analytics Infrastructure

City of Hope – California Teachers Study (CTS)



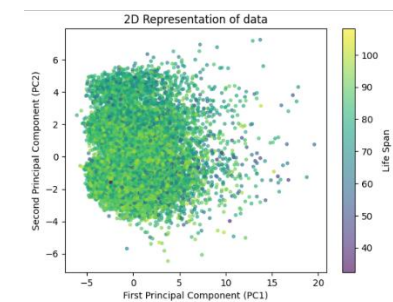
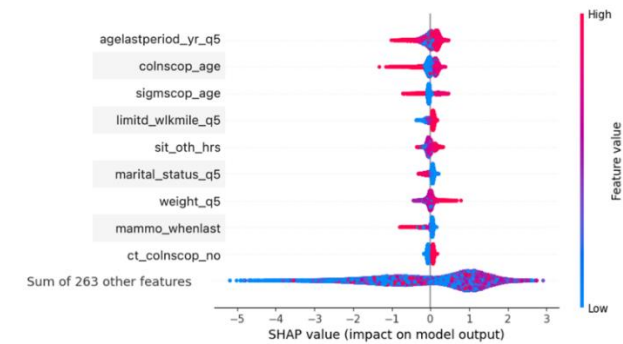
Highlights: Aim #2

- Focused on subset of CTS population with minimum amount of missing data
 - Over 27,500 participants (observations) who completed all 6 CTS surveys
 - 89 data columns (features) that were completed by all participants
- t-SNE: t-distributed stochastic neighbor embedding on 10% sample
 - Hierarchical density-based clustering to partition 2D t-SNE embedding into spatial groups
 - Used learned cluster labels to train a random forest classifier to predict labels in validation set
 - Trained classifier then interpreted using SHAP values



Highlights: Aim #3

- Demonstrate potential downstream use of CEC data by training 2 models: participant survey response & life expectancy
 - Split dataset into 80:20 training & testing sets; reweighted b/c of class imbalances
- Who completes next mailed surveys?
 - Separate models for all 5 follow-up surveys; including gradient boosting classifier
- Can diet & smoking predict death during follow-up?
 - 3 models: 1) age-based, 2) XGBoost regression, & 3) DNN
 - Transformed data into principal components (PC)



Challenges & Lessons Learned



Environment & Data

- Extensive & sensitive CEC data
- Valuable to surface heretofore hidden data & compute costs



Missing Data

- By design vs. data flaws
- CEC documentation subpar; need CDMs, CDEs, & standards



Data Diversity in CECs

- A strength & AI/ML weakness
- Organize & document in ways that convey context better

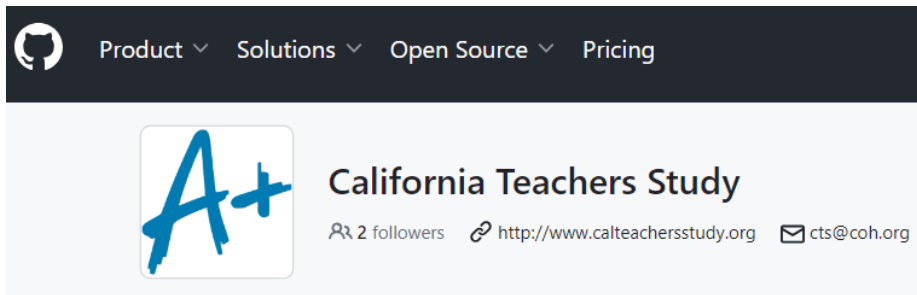


Closed vs. Open Cohorts

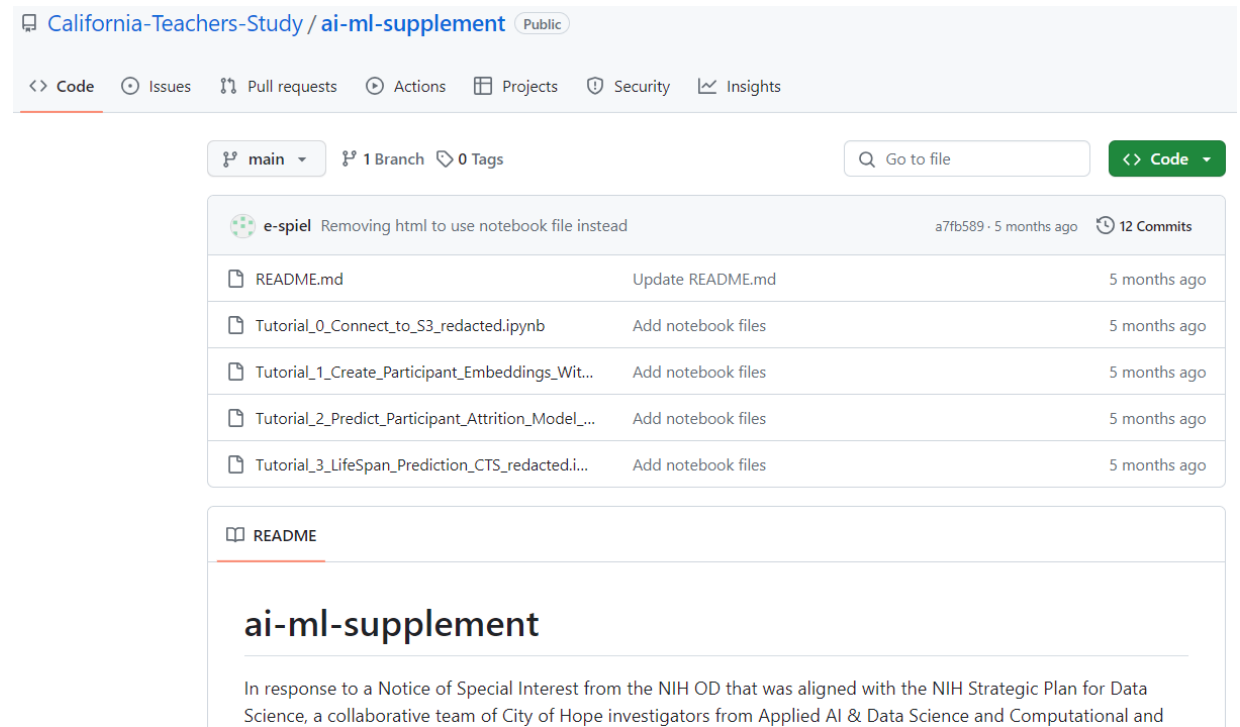
- CECs don't replenish population
- Major questions about models over time & age in CEC data

Links & Future Work

- 4 tutorial notebooks on CTS GitHub: github.com/California-Teachers-Study/ai-ml-supplement



- Incorporating embeddings into ongoing CTS work, e.g., geospatial exposures
- With SDSC Sherlock, accelerating transition of CTS infrastructure to AWS
- Strategic AI/ML plans for new cohorts



The CTS: A different approach



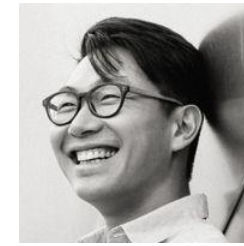
Equal access for everyone to the same CTS data & resources

All users, data, tools, software, & documentation in one place

Thorough security framework for all users, data, & the environment

Standardize, automate, & eliminate barriers to success

Thank You



Kingson Man, PhD



Alec Wong, PhD



Mohsen Nabian, PhD

Commitment to Diversity, Equity, and Inclusion

The California Teachers Study team is committed to equity, diversity, and to honoring our participants' contributions through excellence in research. We are dedicated to creating an inclusive and equitable research environment in which all CTS researchers, staff, partners, and participants can participate fully.

Public health research is often at the intersection of science and social justice, and those who work in this discipline recognize racism as a public health crisis. The California Teachers Study acknowledges that social and institutional inequalities experienced based on race, ethnicity, gender identity, and sexual orientation continue to disproportionately affect the health outcomes of marginalized groups.

Actively addressing disparities and racism is essential in our roles as health researchers. The California Teachers Study investigators and research staff are committed to advancing equity in research through:

- Continued internal education and discourse about diversity and inclusion
- Collaborating with other researchers to advance research on cancer and health conditions that disproportionately affect underrepresented populations
- Actively building partnerships to increase the diversity of researchers who use CTS data or join the CTS team



Caroline Thompson



Sophia Wang



Jim Lacey



Elena Martinez



Cheryl Anderson



Jennifer Benbow



Sandeep Chandra



Jessica Clague DeHart



Christine Duffy



Hannah Lui Park



Kristen Savage



Emma Spielfogel