

1 **Automated self-service cohort selection for large-scale population sciences and**  
2 **observational research: The California Teachers Study Researcher Platform**

3 James V. Lacey, Jr.<sup>1\*</sup>, Emma S. Spielfogel<sup>1</sup>, Jennifer L. Benbow<sup>1,2</sup>, Kristen E. Savage<sup>1</sup>, Kai Lin<sup>3</sup>,  
4 Cheryl A.M. Anderson<sup>4,7</sup>, Jessica Clague-DeHart<sup>5</sup>, Christine N. Duffy<sup>6</sup>, Maria Elena Martinez<sup>4,7</sup>,  
5 Hannah Lui Park<sup>8,9</sup>, Caroline A. Thompson<sup>10</sup>, Sophia S. Wang<sup>1</sup>, Sandeep Chandra<sup>3</sup>

6 1. Department of Computational and Quantitative Medicine, Beckman Research Institute, City  
7 of Hope. Duarte, CA, USA.

8 2. Center for Data-driven Insights and Innovation, University of California Health. Oakland, CA,  
9 USA.

10 3. San Diego Supercomputer Center. University of California San Diego. La Jolla, CA, USA.

11 4. Herbert Wertheim School of Public Health & Human Longevity Science, University of  
12 California San Diego, San Diego, CA, USA.

13 5. School of Community & Global Health. Claremont Graduate University. Claremont, CA,  
14 USA.

15 6. Department of Epidemiology and Biostatistics. University of California San Francisco. San  
16 Francisco, CA, USA.

17 7. Moores Cancer Center, University of California, San Diego, La Jolla, CA, USA.

18 8. Department of Pathology and Laboratory Medicine, School of Medicine, University of  
19 California, Irvine, CA, USA.

20 9. Department of Epidemiology and Biostatistics, Program in Public Health, University of  
21 California, Irvine, CA, USA.

22 10. Department of Epidemiology, Gillings School of Global Public Health, The University of  
23 North Carolina at Chapel Hill, Chapel Hill, NC, USA.

24 \*Corresponding author; email: [jlacey@coh.org](mailto:jlacey@coh.org)

25 Short title: Automated self-service cohort-selection for observational & real-world data

## 26 **Abstract** (300 words)

27 **Objective.** Cohort selection is ubiquitous and essential, but manual and ad hoc approaches are  
28 time-consuming, labor-intense, and difficult to scale. We sought to automate the task of cohort  
29 selection by building self-service tools that enable researchers to independently generate  
30 datasets for population sciences research.

31 **Materials and Methods.** The California Teachers Study (CTS) is a prospective observational  
32 study of 133,477 women who have been followed continuously since 1995. The CTS includes  
33 extensive survey-based and real-world data from cancer, hospitalization, and mortality linkages.  
34 We curated data from our data warehouse into a column-oriented database and developed a  
35 researcher-facing web application that guides researchers through the project lifecycle; captures  
36 researchers' inputs; and automatically generates custom and analysis-ready data, code,  
37 dictionaries, and documentation.

38 **Results.** Researchers can register, access data, and propose projects on the CTS Researcher  
39 Platform via our CTS website. The Platform supports cohort and cross-sectional study designs  
40 for cancer, mortality, and any other ICD-based phenotypes or endpoints. User-friendly prompts  
41 and menus capture analytic design, inclusion/exclusion criteria, endpoint definitions, censoring  
42 rules, and covariate selection. Our platform empowers researchers everywhere to query,  
43 choose, review, and automatically and quickly receive custom data, analytic scripts, and  
44 documentation for their research projects. Research teams can review, revise, and update their  
45 choices anytime.

46 **Discussion.** We replaced inefficient traditional cohort-selection processes with an integrated  
47 self-service approach that simplifies and improves cohort selection for all stakeholders.  
48 Compared with manual methods, our solution is faster and more scalable, user-friendly, and  
49 collaborative. Other studies could re-configure our individual database, project-tracking,  
50 website, and data-delivery components for their own specific needs, or they could utilize other

51 widely available solutions (e.g., alternative database or project-tracking tools) to enable similarly  
52 automated cohort-selection in their own settings. Our comprehensive and flexible framework  
53 could be adopted to improve cohort selection in other population sciences and observational  
54 research settings.

## 55 Introduction

56 Observational research using real-world data makes vital contributions to biomedical research  
57 (2). Large cohort studies of volunteers whose data are tracked and aggregated for research play  
58 an especially important role and often become community resources (26). The largest cohorts,  
59 including the NIH All of Us Research Program (2), UK Biobank (12), and Million Veteran  
60 Program (21), can include hundreds of thousands of participant partners (i.e., volunteers), last  
61 for decades, and support a wide range of future research projects (34) and broad data sharing  
62 (32).

63  
64 Individual research projects rarely require all the data a large cohort has assembled. “Cohort  
65 selection” refers to the process of generating project-specific datasets that give researchers  
66 what they need while protecting participants’ privacy and confidentiality. Cohort selection  
67 includes specifying the study design; applying eligibility, inclusion, and exclusion criteria;  
68 operationally defining key study parameters and endpoints; and choosing specific covariates.  
69 Growing use of Electronic Health Records (EHRs) (30) and research data warehouses and  
70 repositories (22) relies on cohort selection. The CONSORT (4) and STROBE (35) statements  
71 recommend that reports of study results thoroughly document the cohort selection process.

72  
73 When data are private or proprietary, data providers can spend considerable time and energy  
74 helping data requestors understand the data, optimize requests, and perform cohort selection  
75 (26). Even with modern computing infrastructure, manual cohort-selection takes too long and  
76 cannot scale (17); cohort selection is often a bottlenecking event. Even after deploying a cloud-  
77 based data commons specifically designed to improve data access and sharing (19), our  
78 prospective California Teachers Study (CTS) cohort (5) struggled with manual cohort selection.

79

80 Other cohorts (37) and enterprises (17) provide self-service query tools. We sought to enable  
81 complete and comprehensive self-service cohort selection, including data delivery. This report  
82 describes our development and deployment of the CTS Researcher Platform, an innovative tool  
83 that empowers researchers everywhere to query, choose, review, and automatically and quickly  
84 receive custom data, analytic scripts, and documentation for their research projects.

85

## 86 **Materials and methods**

### 87 **The California Teachers Study (CTS)**

88 The CTS is an NCI-funded, multi-site prospective cancer epidemiology cohort (CEC) study (5).  
89 It began in 1995-1996, when 133,477 adult women completed a survey and consented to future  
90 data collection and use of their data for research (9). Participants completed up to five follow-up  
91 surveys that covered diverse health, lifestyle, and environmental exposures (10). Ongoing  
92 annual data linkages with the California Cancer Registry (CCR), Department of Health Care  
93 Access and Information (HCAI; formerly Office of Statewide Health Planning and Development  
94 (OSHPD)), and Department of Public Health Vital Records (CDPH-VR) have identified over  
95 36,000 participants with cancer, over 108,000 participants who were hospitalized, and over  
96 38,000 participants who died during follow-up (6). Linkages with the Centers for Medicare and  
97 Medicaid Services (CMS) identified detailed provider and claims data for over 98,000  
98 participants. With additional biospecimens, biomarkers, and linked geospatial data, the CTS's  
99 survey and real-world data can enable hundreds of potential research projects (8).

100

### 101 **Ethics**

102 Based on the study invitation they received, participants who completed the baseline CTS  
103 survey were considered to have provided informed consent, with a waiver of written informed

104 consent based on return of the completed baseline survey. CTS data are under controlled  
105 access to protect participants' privacy and confidentiality. The first and last completed baseline  
106 surveys were received on Oct. 27, 1995, and Aug. 20, 1999, respectively.

107

## 108 **CTS data environment and previous cohort selection**

### 109 **methods**

110 Until 2015, we used entirely manual cohort selection: data managers received individual  
111 requests, worked with requestors, and created datasets with accompanying dictionaries, using  
112 locally stored files and desktop software. In 2016, our data commons (19) replaced those silos  
113 and brought users, data, software, and tools together in one secure shared environment that  
114 includes 1) a data warehouse with data marts designed for analysis; 2) file and extract-  
115 transform-load (ETL) servers; 3) software, tools, metadata, and documentation; and 4) a remote  
116 desktop environment that serves as the collaborative workspace.

117

### 118 **Cohort selection: Still a rate-limiting step**

119 The CTS data commons (19) did not eliminate manual cohort selection. We standardized the  
120 process by asking users to specify their inclusion/exclusion criteria, endpoints, covariates, and  
121 other details using drop-down menus and open-ended text in Excel worksheets. CTS data  
122 analysts then manually inserted those choices into SAS Proc SQL templates (over 90% of  
123 projects and analysts used SAS (Cary, NC) software) to join data from the CTS warehouse  
124 and/or marts. Researchers then used those customized templates to call and analyze their  
125 project-specific data in our secure CTS workspace. Even with cloud-based data, data calls, and  
126 workspaces, manual entries by researchers and manual customization of data calls by the CTS  
127 team were unsustainable and unscalable.

128

## 129 **New challenge: Self-service cohort selection at scale**

130 We needed to eliminate manual cohort selection. Instead of giving their choices to CTS  
131 analysts, researchers should be free to directly interact with CTS data, automatically apply their  
132 choices, and generate their own data. Our data commons already provided a secure workspace  
133 with the data, tools, and documentation users needed (19), but it needed three new features: 1)  
134 flexible, comprehensive, and robust menu-driven workflows that include the full range of cohort  
135 selection choices without limiting researchers' options; 2) a web-based application for applying  
136 those choices to our data source; and 3) workflows for automatically generating the required  
137 deliverables, including datasets, data dictionaries, and analysis scripts.

138

## 139 **Workflows: Robust yet flexible**

140 The CTS is primarily a “risk” CEC, rather than a “survivorship” CEC (23); this shapes our cohort  
141 selection choices. Using our existing CTS templates, we identified the full range of potential  
142 study designs (case-control, cohort/time-to-event, cross-sectional), analytic outcomes (incident  
143 cancer, mortality, hospital- or ICD-based outcomes/phenotypes), and exposure data (self-report  
144 from surveys, geospatial data, biospecimen-based data) that are typical of risk cohorts. Our  
145 initial solution focused on the most common type of CTS project to date: a cohort design with an  
146 individual cancer type as the analytic outcome and self-reported survey data as the main  
147 exposures (8).

148

149 We articulated detailed user stories (38) to capture design requirements across the entire  
150 process. These included the ability to select multiple outcomes and multiple exposures; choose  
151 covariates individually or by hierarchical categories; set specific start-of-follow-up and end-of-  
152 follow-up dates; establish inclusion and exclusion criteria; apply analytic censoring rules; and

153 review real-time dashboards that displayed frequencies based on users' inputs. Another  
154 essential user requirement was the ability to revise any individual component of the cohort  
155 selection—e.g., to add independent variables or modify censoring criteria—during a project  
156 without requiring users to completely start over. We identified the types of deliverables users  
157 would need (e.g., custom data dictionaries with only the covariates they had selected; a  
158 summary of their design decisions, etc.). We decided to configure datasets that could be  
159 analyzed using open-source (e.g., Posit, formerly R Studio) or commercial (e.g., SAS) analysis  
160 software packages. We managed user stories and project management via smartsheet.com,  
161 mock-ups via figma.com, and communication and collaboration via Slack.

162

163 We also identified combinations of design choices that initially seemed too complex to  
164 automate, such as idiosyncratic matching criteria for controls in nested case-control projects  
165 and clinical phenotypes based on complex combinations of ICD codes and hospitalization  
166 patterns. We excluded those workflows from the initial solution.

167

## 168 **Source data: High-performance data management from** 169 **multiple domains**

170 CTS data linkages update cancer, hospitalization, and mortality data annually (19), but survey  
171 data are static. This allowed us to extract data from our warehouse and avoid having the self-  
172 service tool directly query our data marts. We also considered missing-by-design data: in large  
173 prospective studies, study censoring (e.g., participants who die do not complete subsequent  
174 surveys) and rare outcomes (e.g., low frequencies of multiple-primary cancers) create valuable  
175 information but are inefficient from a traditional SQL database perspective, because they create  
176 large numbers of columns, but many of those columns contain empty cells. To increase  
177 flexibility, scalability, and speed, we added a columnar (column-oriented) online analytical



178 processing (OLAP) database management system designed for high-performance (ClickHouse,  
179 Redwood City, CA) as the data source that would be available to the self-service web  
180 application.

181  
182 We wrote a script that extracts data from our data warehouse from three domains—participants,  
183 cancers, and surveys—into a large and wide OLAP database. Participant data included  
184 essential characteristics (e.g., date of birth, date of death, cause of death, race/ethnicity, vital  
185 status) and follow-up information (e.g., dates of study enrollment, follow-up surveys, last follow-  
186 up, etc.). Cancer data included detailed site, stage, grade, diagnosis date, etc. for all cancers  
187 during follow-up. Survey data included approximately 1200 covariates (i.e., columns) from the  
188 CTS questionnaires, with at least one column for every question, plus other existing derived  
189 covariates derived (e.g., body mass index (BMI) based on self-reported height and weight). All  
190 questionnaire covariates were previously tagged to facilitate identification by question number,  
191 questionnaire section, or questionnaire number (6). The workflow for creating and updating this  
192 presentation database leveraged the efficient schema of our data warehouse to generate  
193 individual participant records across all domains.

194

## 195 **Data and code availability**

196 All CTS data and analytic code are available in the CTS Researcher Platform (7), which is also  
197 available via the “For Researchers” tab on the CTS website (5). CTS data are not publicly  
198 available because they include extensive identifiable, sensitive, and confidential information, but  
199 any researcher who agrees to protect and use CTS data responsibly for research can access all  
200 CTS resources through the Researcher Platform (7). The underlying code for the self-service  
201 cohort-selection application is available upon request via the Researcher Platform (7).

202

## 203 **Results**

### 204 **Developing the web application: User-friendly menus and** 205 **interactive visualizations**

206 User stories and the potential study design and analysis choices drove the web application  
207 development. We wanted to emulate the user-friendly menus and buttons in tools such as the  
208 NCI SEER\*Explorer Application (SEER\*Explorer) (27). In our experience, all researchers who  
209 ask to use data from cohorts like the CTS for their projects understand the basics of cohort  
210 selection. We designed the application for novice users who did not have any prior experience  
211 analyzing CTS data; i.e., the application needed to walk users through every step of the process  
212 in easily understandable ways.

213

214 We identified six steps in the cohort selection process and created a sequential task-based  
215 page in the application for each step: 1) select cancer endpoint; 2) select start of follow-up; 3)  
216 select censoring rules; 4) select questionnaire data; 5) enter biospecimen data; and 6) review  
217 summary. The application includes drop-down menus and search functions to help users  
218 choose their cohorts and data. Users can choose cancer endpoints by ICD-O-3 site or SEER  
219 Site Group Recode values (31). Additional choices can be made by histology codes. Users can  
220 start their follow-up at any of the dates that participants completed a CTS survey or enter a  
221 different start-date in the form of MM/DD/YYYY. Users can choose whether to censor  
222 participants who develop cancers other than their analytic endpoint. When users choose breast,  
223 uterine, or ovarian cancer endpoints, censoring rules automatically incorporate censoring at the  
224 date of bilateral mastectomy, hysterectomy, or bilateral oophorectomy, respectively; all are  
225 available through the CTS's linked hospitalization data. As users choose their data, interactive  
226 visualizations display and update, in real-time, frequencies and distributions of their choices.

227 The application automatically saves users' interim progress and allows all members of a project  
228 team to make or modify choices. Each page includes a "back" button that returns users to the  
229 previous page, where choices can be modified. Figure 1 shows screenshots from a sample  
230 cohort design with a cancer endpoint and survey-based exposures.

231

232 Fig 1. Screenshots of the self-service cohort selection application for a cohort analysis with a  
233 cancer endpoint and survey-based exposure data.

234 (a) Select cancer endpoint. (b) Select start of follow-up. (c) Select censoring rules. (d) Select  
235 questionnaire data. (e) Summary.

236

## 237 **Deliverables: Immediate access to data and continuous CTS**

### 238 **support**

239 The final screen includes a "Generate Data" button. When users click that button, the  
240 application saves all of the inputs and generates six deliverables: 1) a custom \*.csv dataset  
241 based on the user's choices; 2) a SAS-specific formats file matching the custom \*.csv dataset  
242 (SAS is the primary software used in the CEC community); 3) a SAS data call that brings the  
243 custom \*.csv dataset into SAS, using the formats file to automatically apply the appropriate data  
244 formats for analysis; 4) an Posit (R Studio) script that reads the custom \*.csv dataset into a new  
245 R session for analysis; 5) a custom data dictionary that includes all of the covariates selected  
246 and omits all CTS covariates that were not selected; and 6) a PDF summary of all the cohort  
247 selection choices. The dataset and formats files are automatically written to a read-only  
248 directory and the data calls, dictionary, and summary file are automatically written to that  
249 project's dedicated directory, all within the CTS's remote desktop environment (19).

250

251 Generating these deliverables typically requires less than 30 seconds; users essentially get  
252 immediate access to the data, tools, and documentation they need to analyze their data. Writing  
253 datasets to a read-only drive facilitates data governance, data lineage, and version control, and  
254 it preserves data fidelity for every output dataset back to the CTS data warehouse. Writing the  
255 data calls to project-specific workspaces gives researchers complete control and flexibility over  
256 what they do with their code, scripts, and analytic methods. Because all deliverables reside in  
257 the CTS's shared workspace (19), our CTS team can assist, troubleshoot, or collaborate in real-  
258 time with any researcher on any part of any project.

259

## 260 **Other essential components: User accounts, project** 261 **tracking, and integration**

262 When researchers sign up for a user account on the CTS website (7), their account details are  
263 tracked in the CTS's Salesforce organization. Salesforce also serves as the back-end of the  
264 "For Researchers" page on the CTS website, where researchers can propose, submit, and track  
265 their projects. Both our CTS team and researchers can see status updates as projects move  
266 through the research lifecycle. This tracking also enables researchers to automatically receive  
267 access to the cohort selection web application as soon as their project meets required IRB and  
268 approval criteria (7). We use smartsheet.com to track additional details of every project (Fig. 2).

269

270 Fig 2. Project characteristics that are captured and tracked for all projects in the CTS  
271 Researcher Platform.

272

273 Researchers access the web application through the CTS website (5), but the web application  
274 and column-oriented database are hosted within the San Diego Supercomputer Center's  
275 (SDSC's) secure Sherlock environment (29). The web application bidirectionally integrates the

276 active directory (AD) with our secure remote desktop environment, where all user accounts  
277 have role-based permissions (19), with our Salesforce organization, where projects and project  
278 status are linked to individual users. Figure 3 provides an overview of the Platform.

279

280 Fig. 3. Overview of CTS Researcher Platform and integration with secure CTS environment  
281 hosted at SDSC Sherlock.

282

### 283 **Data scope: Standardized and customized**

284 All datasets automatically include 62 essential covariates (e.g., dates of birth, death, and  
285 baseline survey; BMI; smoking status; etc.). Instead of requiring users to make every decision  
286 from scratch, the application provides default choices on key analytic decisions, (e.g., exclude  
287 participants with prevalent cancer), while also allowing researchers to make alternative choices.

288

289 Some complex data (e.g., geospatial-based exposures, food frequency questionnaires, etc.)  
290 were excluded from the initial database that the application uses. When a project requires those  
291 data, or if a user wants to bring their own data into a project, our CTS Research Data Steward  
292 (E.S.) uses data-call templates to deposit the needed data excerpts in the project team's read-  
293 only directory. The excerpts use the same universal data key to facilitate easy and immediate  
294 joins for any additional or custom data; users also receive updated standard CTS code to join  
295 their custom data.

296

### 297 **Platform: Initial launch in 2021 and ongoing improvements**

298 Design, development, testing, and integration took six months. After a soft launch and additional  
299 refinements (24), we launched the full platform in March 2021. Table 1 shows how cohort  
300 selection has evolved since the CTS began.

301  
 302 Table 1. Evolution of cohort selection methods and procedures in the CTS from its beginning, in  
 303 1995-1996, through the CTS Researcher Platform.

	Original CTS Data Strategy (1995 – 2016)	Initial CTS Data Commons (2016 – 2021)	CTS Researcher Platform (2021 – present)
What are the data sources?	Locally stored SAS datasets	CTS data warehouse (DW) & data marts	NoSQL database extract from DW
Who makes choices?	CTS data managers at CTS sites	CTS-wide data analyst	Researchers who request data
Who writes/edits code?	CTS data managers at CTS sites	CTS-wide data analyst	N/A; web application replaced code
Who executes code?	CTS data managers at CTS sites	CTS-wide data analyst	Researcher clicks "Generate Data" button
Where is code stored?	Locally at CTS sites	Secure shared CTS-wide workspace	N/A; web application saves researchers' choices
How are data generated?	Local CTS manager manually creates dataset	CTS analyst runs template	Web application automatically creates data
Where are data delivered?	Local CTS manager sends data to researcher	CTS analyst deposits in project-specific folder	Web application automatically deposits data
How are projects tracked?	Manually at individual CTS sites	Manually within shared workspace	Integrated CTS website & Salesforce
How are changes managed?	Manually by local CTS data managers	Manually within shared workspace	Automated version control
How much time is required?	Weeks to months	Days to weeks	As little as 5 minutes

304  
 305  
 306 The application initially supported analytic projects with incident cancer as the primary endpoint.  
 307 It now also supports cohort selection for ICD-based mortality and hospitalization-based  
 308 phenotypes (6). For these endpoints, a simpler query lets users select the covariates they need  
 309 after skipping the cancer-related questions. Users enter their endpoint information (e.g., specific  
 310 ICD codes or other requirements, such as length-of-stay) in text-based forms and then click  
 311 "generate data" to create the deliverables described above. This simpler query achieves two  
 312 goals. One, it often produces complete data for mortality projects, because date and cause of  
 313 death are automatically included in all datasets. Two, it lets users immediately begin analyzing  
 314 covariate data for hospitalization projects that require additional work by the CTS team and/or  
 315 researchers to generate clinical endpoints. ICD-based phenotypes and inclusion/exclusion  
 316 criteria from real-world hospitalization claims data can be complex; the application standardizes

317 those decisions by asking users to describe their operational definition for each phenotype;  
318 multiple concurrent phenotypes are allowed. Endpoint data are then deposited, with  
319 accompanying code to join those with the output of the query, as described above.

320

321 As of Dec. 2023, 32 projects with 56 total investigators have used this application (Figure 4).

322 Ten projects were led by students or trainees in academic programs; all generated results and  
323 internal presentations. Five projects are part of multi-study consortia. Five projects produced  
324 published or submitted manuscripts; almost all the others are still analyzing data.

325

326 Most projects have chosen a cohort design with cancer endpoints and survey-based exposure  
327 data, and these projects required no help from our CTS team. Projects with hospitalization-  
328 based endpoints typically require some input from our team because of the complexity of the  
329 hospitalization data. However, all those projects received complete, timely, and analysis-ready  
330 phenotype data as part of their deliverables, even when projects included multiple phenotypes.  
331 In our experience, all users have been able to independently navigate the start and stop dates,  
332 censoring decisions, and covariate selection section of the cohort-selection application.

333

334 Fig. 4. Distribution of study designs and analytic endpoints in CTS Researcher Platform projects  
335 to date.

336

## 337 **Discussion**

338 Cohort selection presents significant challenges for clinical trials, cohort studies, disease  
339 registries, disease networks, enterprise-wide clinical data, and data repositories. For data  
340 providers and data requestors (26), identifying the right patients or research volunteers and then  
341 selecting the right data for those cohorts often bottlenecks the Research Data Management

342 Lifecycle (25). Cohort selection also encounters a negative feedback loop: larger and more  
343 diverse data resources can support a wider range of research projects, but the difficulty of  
344 cohort-selection increases as the breadth, depth, and complexity of the data sources increases.

345

346 Cohort selection today usually occurs one of two ways. In one, researchers submit requests to  
347 providers, who then manually curate and deliver output. Most NCI-funded cancer epidemiology  
348 cohort (CEC) studies do this; our CTS used this approach for twenty years. This method often  
349 relies on labor-intense manual workflows, requires significant back-and-forth between  
350 investigators, and cannot scale to meet contemporary data-sharing requirements. In the other  
351 approach, providers make large source datasets available for exploration, query, and selection.  
352 This appears to be more common for enterprise-wide data providers, including electronic health  
353 records (EHRs), but also typically requires manual and project-specific assistance (13). Even  
354 forward-looking and innovative enterprise-wide query approaches, such as the Duke University  
355 Enterprise Data Unified Content Explorer (DEDUCE), struggled to provide service and data at  
356 scale (17). Two recent reviews described the challenges associated with the preliminary step of  
357 leveraging data to identify “computable clinical” (15) or “digital” (11) phenotypes and concluded  
358 that new, more efficient, and automated approaches are needed to accelerate research.

359

360 We developed a novel self-service cohort selection approach designed to eliminate manually  
361 curated datasets. We configured widely available products and software—from Microsoft  
362 Windows, Salesforce, smartsheet, and ClickHouse—and developed one new custom web  
363 application. Our integrated platform empowers users to choose and automatically receive the  
364 data and documentation they need to conduct their research; facilitates efficient collaboration  
365 and sharing; and enables us to fully track, manage, and support every user, team, and project.

366



367 Because no two cohorts are ever identical, cohort selection is typically cohort-specific.  
368 Nonetheless, our automated approach has potential for broad reusability. Regardless of where  
369 or how cohort selection occurs, it entails the same fundamental components. Cohort selection in  
370 time-to-event analyses must operationally define clinical endpoints, specify follow-up intervals,  
371 determine censoring rules, establish inclusion and exclusion criteria, and choose exposure and  
372 covariate data. Cohort selection for cross-sectional analyses requires three identical steps:  
373 define clinical endpoints, establish inclusion and exclusion criteria, and choose exposure and  
374 covariate data. We designed our approach around these common steps that are reusable  
375 across different cohort selection settings. This modular approach to cohort-selection workflows  
376 enabled us to efficiently expand our platform's scope from just cancer endpoints to also  
377 hospitalization and mortality endpoints, while reusing other components.

378

379 The long track record of CECs successfully sharing their data in consortia, such as the NCI  
380 Cohort Consortium (32), denotes the broad potential applicability and reusability of our cohort  
381 selection approach. Dozens of CECs worldwide (20) regularly harmonize and share individual-  
382 level data for consortia projects. In those projects, every participating CEC performs cohort-  
383 selection on its data using a set of common criteria established by the consortia. The source  
384 data in each cohort are similar enough that they can be harmonized for individual-level pooled  
385 analyses (i.e., rather than meta-analyses). When similar yet independent cohorts can all  
386 execute a common cohort selection workflow to yield interoperable data that are harmonized  
387 and pooled at the level of individual participants, then the upstream cohort selection process is  
388 inherently standardizable.

389

390 Data providers' environments, architectures, and strategies affect cohort selection. Recent  
391 papers (1,13,16-18,36) describe different approaches for leveraging research data warehouses,  
392 data marts, and repositories for research. Before developing our Researcher Platform, we used

393 a combination of spreadsheets and SQL templates to select cohorts directly from our CTS data  
394 warehouse and data marts. We added an OLAP database as the data source for our cohort-  
395 selection application both to accelerate performance and to simplify the development of the  
396 application. Designing the application to query data from this middle layer, rather than directly  
397 from our data warehouse and data marts, also simplified our data governance strategy, because  
398 it created a buffer between CTS users and CTS source data (19). This modular approach to our  
399 solution architecture—i.e., an online database, a project tracking platform, a web-based project  
400 management tool, and a custom-built web application—can also be replicated because  
401 numerous existing tools can perform these tasks.

402

403 Two nationwide cohorts, the NIH All of Us Research Program and the UK Biobank, recently  
404 launched research platforms that make more of their data and resources FAIR (Findable,  
405 Accessible, Interoperable, Reusable). The UK Biobank initially avoided the cohort selection  
406 problem by allowing users to download source data. As the UK Biobank grew, that  
407 unsustainable strategy, which also required over a year to deliver data (14), gave way to a  
408 centralized research platform that includes analytic capabilities. The UK Biobank Research  
409 Analysis Platform (RAP) (33) utilizes a DNANexus platform that offers a variety of tools,  
410 software, and options (e.g., Spark SQL, JupyterLab, Jupyter notebooks) that researchers can  
411 use to configure and analyze data. The NIH’s All of Us Research Hub (3) includes a Data  
412 Browser for publicly available data and a Researcher Workbench for controlled-access  
413 “Registered Tier” data. This Workbench splits cohort selection into two steps: the “Cohort  
414 Builder” uses inclusion and exclusion criteria to identify a population, and the “Dataset Builder”  
415 chooses data for that cohort. Users with R or Python experience can then analyze those data in  
416 Jupyter Notebooks (3). As nationally supported cohorts, both the UK Biobank and All of Us  
417 operate at scales much larger than an individual cohort like the CTS. Within our CEC  
418 community, Python and SQL skills are not yet commonplace on research teams, and this

419 influenced our decision to create a new point-and-click web application as the primary user  
420 interface for CTS cohort selection. Despite those differences, the NIH All of Us Researcher  
421 Workbench, the UK Biobank RAP, and our CTS Researcher Platform provide three distinct  
422 examples, at different scales, of successful web-based cohort selection.

423

424 These three first-generation cohort-selection tools reveal common themes. Robust processes  
425 and comprehensive workflows, even for historically open-ended tasks like cohort-selection,  
426 pave a path from manual methods to scalable self-service. For cohort selection, dividing the  
427 overall process into more modular units, whether the Cohort and Data Builders of the  
428 Researcher Workbench or the approach our CTS Research Platform took, works. These  
429 concepts align with emerging best practices for self-service tools more broadly: build data  
430 culture, prioritize data literacy, ensure governance, and target specific business goals (28).

431

432 Development of tools like this requires tradeoffs and design choices. We tackled cohort-  
433 selection for CECs, but our cohort also includes hundreds of thousands of linked hospitalization  
434 records that support research on other chronic disease and clinical phenotypes (6). We  
435 leveraged our study-specific and cloud-based data warehouse, but our use of a columnar OLAP  
436 database provided a simple option for making large-scale cohort data easily available to a web  
437 application. For practical reasons, we omitted our CTS genomic, geospatial, and raw dietary  
438 (from two food frequency questionnaires) data from the self-service portion of our platform, but  
439 those data can be easily and quickly joined with our platform's outputs when needed. Some  
440 study design components, such as control matching in case-control studies, might be too  
441 complex to automate or convert into menu-based choices. We will continue to learn from our  
442 user community and improve our CTS Researcher Platform.

443

## 444 **Conclusion**

445 Data providers and data requestors continue to struggle with contemporary cohort selection.

446 Greater use of large-scale survey-based and real world data for research to improve health

447 outcomes will continue to strain today's manual and labor-intense cohort-selection workflows.

448 The CTS appears to be the first long-running observational cohort to replace its legacy manual

449 cohort-selection methods with a comprehensive web-based self-service application that lets all

450 researchers independently and directly choose, review, receive, and modify the custom data

451 they need for their research. Automated self-service improves efficiency, scalability, and

452 sustainability of data sharing, but ongoing evaluation and community feedback will be essential

453 to identify the right balance of common standards and cohort-specific features and

454 configurations that enable efficient reusability. The CTS Researcher Platform demonstrates that

455 automated and user-friendly self-service cohort selection is practical, even for large and

456 complex data sources, and can be deployed using widely available tools and approaches.

457

## 458 **Acknowledgements**

459 The collection of cancer incidence data used in the California Teachers Study was supported by

460 the California Department of Public Health pursuant to California Health and Safety Code

461 Section 103885; Centers for Disease Control and Prevention's National Program of Cancer

462 Registries, under cooperative agreement 5NU58DP006344; the National Cancer Institute's

463 Surveillance, Epidemiology and End Results Program under contract HHSN261201800032I

464 awarded to the University of California, San Francisco, contract HHSN261201800015I awarded

465 to the University of Southern California, and contract HHSN261201800009I awarded to the

466 Public Health Institute. The opinions, findings, and conclusions expressed herein are those of

467 the author(s) and do not necessarily reflect the official views of the State of California,

468 Department of Public Health, the National Cancer Institute, the National Institutes of Health, the

469 Centers for Disease Control and Prevention or their Contractors and Subcontractors, or the  
470 Regents of the University of California, or any of its programs.

471

472 The authors would like to thank all California Teachers Study participants and the Steering  
473 Committee that is responsible for the formation and maintenance of the Study within which this  
474 research was conducted. A full list of California Teachers Study team members is available at  
475 <https://www.calteachersstudy.org/team>.

476

477

## 478 **References**

- 479 1. Abrahão MTF, Nobre MRC, Gutierrez MA. A method for cohort selection of cardiovascular  
480 disease records from an electronic health record system. *Int J Med Inform.* 2017  
481 Jun;102:138-149. doi: 10.1016/j.ijmedinf.2017.03.015. Epub 2017 Mar 30. PMID: 28495342.
- 482 2. All-of Us Research Program Investigators The All of Us research program. *New England*  
483 *Journal of Medicine* 381, 668–676 (2019).
- 484 3. All of Us / Research Hub / Researcher Workbench. [https://www.researchallofus.org/data-](https://www.researchallofus.org/data-tools/workbench/)  
485 [tools/workbench/](https://www.researchallofus.org/data-tools/workbench/)
- 486 4. Begg C, Cho M, Eastwood S, Horton R, Moher D, Olkin I, Pitkin R, Rennie D, Schulz KF,  
487 Simel D, Stroup DF. Improving the quality of reporting of randomized controlled trials. The  
488 CONSORT statement. *JAMA.* 1996 Aug 28;276(8):637-9. doi: 10.1001/jama.276.8.637.  
489 PMID: 8773637.
- 490 5. California Teachers Study. [www.calteachersstudy.org](http://www.calteachersstudy.org).
- 491 6. California Teachers Study: California Teachers Study Data.  
492 <https://www.calteachersstudy.org/cts-data>

- 493 7. California Teachers Study: Researcher Platform.  
494 <https://calteachersstudy.my.site.com/researchers/s/>
- 495 8. California Teachers Study: Study Findings. <https://www.calteachersstudy.org/study-findings>
- 496 9. California Teachers Study: Study Population. <https://www.calteachersstudy.org/study->  
497 [population](https://www.calteachersstudy.org/study-population)
- 498 10. California Teachers Study: Past Questionnaires. <https://www.calteachersstudy.org/past->  
499 [questionnaires](https://www.calteachersstudy.org/past-questionnaires)
- 500 11. Capurro D, Barbe M, Daza C, Santa Maria J, Trincado J. Temporal Design Patterns for  
501 Digital Phenotype Cohort Selection in Critical Care: Systematic Literature Assessment and  
502 Qualitative Synthesis. JMIR Med Inform. 2020 Nov 24;8(11):e6924. doi:  
503 10.2196/medinform.6924. PMID: 33231554; PMCID: PMC7723741.
- 504 12. Conroy MC, Lacey B, Bešević J, Omiyale W, Feng Q, Effingham M, Sellers J, Sheard S,  
505 Pancholi M, Gregory G, Busby J, Collins R, Allen NE. UK Biobank: a globally important  
506 resource for cancer research. Br J Cancer. 2023 Feb;128(4):519-527. doi: 10.1038/s41416-  
507 022-02053-5. Epub 2022 Nov 19. PMID: 36402876; PMCID: PMC9938115.
- 508 13. Danciu I, Cowan JD, Basford M, Wang X, Saip A, Osgood S, Shirey-Rice J, Kirby J, Harris  
509 PA. Secondary use of clinical data: the Vanderbilt approach. J Biomed Inform. 2014  
510 Dec;52:28-35. doi: 10.1016/j.jbi.2014.02.003. Epub 2014 Feb 14. PMID: 24534443; PMCID:  
511 PMC4133331.
- 512 14. Data Access Quick Guide to UK Biobank: April 2020. [https://md.catapult.org.uk/wp-](https://md.catapult.org.uk/wp-content/uploads/2020/05/Data-Access-Quick-Guide-UK-Biobank-0420.pdf)  
513 [content/uploads/2020/05/Data-Access-Quick-Guide-UK-Biobank-0420.pdf](https://md.catapult.org.uk/wp-content/uploads/2020/05/Data-Access-Quick-Guide-UK-Biobank-0420.pdf)
- 514 15. He T, Belouali A, Patricoski J, Lehmann H, Ball R, Anagnostou V, Kreimeyer K, Botsis T.  
515 Trends and opportunities in computable clinical phenotyping: A scoping review. 2023. J  
516 Biomed Informatics; 140.
- 517 16. He W, Kirchoff KG, Sampson RR, McGhee KK, Cates AM, Obeid JS, Lenert LA. Research  
518 Integrated Network of Systems (RINS): a virtual data warehouse for the acceleration of

- 519 translational research. J Am Med Inform Assoc. 2021 Jul 14;28(7):1440-1450. doi:  
520 10.1093/jamia/ocab023. PMID: 33729486; PMCID: PMC8279787.
- 521 17. Horvath MM, Rusincovitch SA, Brinson S, Shang HC, Evans S, Ferranti JM. Modular design,  
522 application architecture, and usage of a self-service model for enterprise data delivery: the  
523 Duke Enterprise Data Unified Content Explorer (DEDUCE). J Biomed Inform. 2014  
524 Dec;52:231-42. doi: 10.1016/j.jbi.2014.07.006. Epub 2014 Jul 19. PMID: 25051403; PMCID:  
525 PMC4335712.
- 526 18. Hurst JH, Liu Y, Maxson PJ, Permar SR, Boulware LE, Goldstein BA. Development of an  
527 electronic health records datamart to support clinical and population health research. J Clin  
528 Transl Sci. 2020 Jun 23;5(1):e13. doi: 10.1017/cts.2020.499. PMID: 33948239; PMCID:  
529 PMC8057430.
- 530 19. Lacey JV Jr, Chung NT, Hughes P, Benbow JL, Duffy C, Savage KE, Spielfogel ES, Wang  
531 SS, Martinez ME, Chandra S. Insights from Adopting a Data Commons Approach for Large-  
532 scale Observational Cohort Studies: The California Teachers Study. Cancer Epidemiol  
533 Biomarkers Prev. 2020 Apr;29(4):777-786. doi: 10.1158/1055-9965.EPI-19-0842. Epub  
534 2020 Feb 12. PMID: 32051191; PMCID: PMC9005205.
- 535 20. Membership of the NCI Cohort Consortium. [https://epi.grants.cancer.gov/cohort-](https://epi.grants.cancer.gov/cohort-consortium/members/)  
536 [consortium/members/](https://epi.grants.cancer.gov/cohort-consortium/members/)
- 537 21. Million Veteran Program. <https://www.mvp.va.gov/pwa/>
- 538 22. Murphy SN, Visweswaran S, Becich MJ, Campion TR, Knosp BM, Melton-Meaux GB, Lenert  
539 LA. Research data warehouse best practices: catalyzing national data sharing through  
540 informatics innovation. J Am Med Inform Assoc. 2022 Mar 15;29(4):581-584. doi:  
541 10.1093/jamia/ocac024. Erratum in: J Am Med Inform Assoc. 2022 Jul 12;29(8):1445.  
542 Erratum in: J Am Med Inform Assoc. 2022 Dec 13;30(1):209. PMID: 35289371; PMCID:  
543 PMC8922176.

- 544 23. PAR-20-294: Core Infrastructure Support for Cancer Epidemiology Cohorts.  
545 <https://grants.nih.gov/grants/guide/pa-files/PAR-20-294.html>
- 546 24. Push Button Data Sharing: Web-Based Self-Service and Automated Data Delivery in the  
547 California Teachers Study. [https://epi.grants.cancer.gov/cohort-consortium/cohort-](https://epi.grants.cancer.gov/cohort-consortium/cohort-events.html)  
548 [events.html](https://epi.grants.cancer.gov/cohort-consortium/cohort-events.html). Jan 12, 2021.
- 549 25. Research Lifecycle. <https://researchsupport.harvard.edu/research-lifecycle>
- 550 26. Rolland B, Geiger AM. Addressing Challenges in Converting Grant-Funded Infrastructures  
551 to Broadly Used Research Resources. *Cancer Epidemiol Biomarkers Prev.* 2019  
552 Oct;28(10):1559-1562. doi: 10.1158/1055-9965.EPI-19-0043. Epub 2019 Aug 28. PMID:  
553 31462397.
- 554 27. SEER\*Explorer: An interactive website for SEER cancer statistics [Internet]. Surveillance  
555 Research Program, National Cancer Institute; 2023 Apr 19. [cited 2023 May 26]. Available  
556 from: <https://seer.cancer.gov/statistics-network/explorer/>. Data source(s): SEER Incidence  
557 Data, November 2022 Submission (1975-2020), SEER 22 registries.
- 558 28. Self-Service Analytics: How to Use Healthcare Business Intelligence.  
559 [https://www.healthcatalyst.com/insights/self-service-analytics-how-use-healthcare-business-](https://www.healthcatalyst.com/insights/self-service-analytics-how-use-healthcare-business-intelligence)  
560 [intelligence](https://www.healthcatalyst.com/insights/self-service-analytics-how-use-healthcare-business-intelligence)
- 561 29. Sherlock Cloud Solution & Services. <https://sherlock.sdsc.edu/>
- 562 30. Site Recode ICD-O-3/WHO 2008 Definition.  
563 [https://seer.cancer.gov/siterecode/icdo3\\_dwhohome/](https://seer.cancer.gov/siterecode/icdo3_dwhohome/)
- 564 31. Shortreed SM, Cook AJ, Coley RY, Bobb JF, Nelson JC. Challenges and Opportunities for  
565 Using Big Health Care Data to Advance Medical Science and Public Health. *Am J*  
566 *Epidemiol.* 2019 May 1;188(5):851-861. doi: 10.1093/aje/kwy292. PMID: 30877288.
- 567 32. Swerdlow AJ, Harvey CE, Milne RL, Pottinger CA, Vachon CM, Wilkens LR, Gapstur SM,  
568 Johansson M, Weiderpass E, Winn DM. The National Cancer Institute Cohort Consortium:  
569 An International Pooling Collaboration of 58 Cohorts from 20 Countries. *Cancer Epidemiol*



- 570 Biomarkers Prev. 2018 Nov;27(11):1307-1319. doi: 10.1158/1055-9965.EPI-18-0182. Epub  
571 2018 Jul 17. PMID: 30018149.
- 572 33. The UK Biobank Research Analysis Platform. [https://www.ukbiobank.ac.uk/enable-your-](https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform)  
573 [research/research-analysis-platform](https://www.ukbiobank.ac.uk/enable-your-research/research-analysis-platform)
- 574 34. Vander Weele, TJ. Observational Studies and Study Designs: An Epidemiologic  
575 Perspective. *Observational Studies*. 2015;1(1):223-230.
- 576 35. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP; STROBE  
577 Initiative. The Strengthening the Reporting of Observational Studies in Epidemiology  
578 (STROBE) Statement: guidelines for reporting observational studies. *Int J Surg*. 2014  
579 Dec;12(12):1495-9. doi: 10.1016/j.ijsu.2014.07.013. Epub 2014 Jul 18. PMID: 25046131.
- 580 36. Walters KM, Jojic A, Pfaff ER, Rape M, Spencer DC, Shaheen NJ, Lamm B, Carey TS.  
581 Supporting research, protecting data: one institution's approach to clinical data warehouse  
582 governance. *J Am Med Inform Assoc*. 2022 Mar 15;29(4):707-712. doi:  
583 10.1093/jamia/ocab259. PMID: 34871428; PMCID: PMC8922173.
- 584 37. Women's Health Initiative (WHI) Query Builder. <https://www.whi.org/qb/>
- 585 38. Writing Effective User Stories. [https://tech.gsa.gov/guides/effective\\_user\\_stories/](https://tech.gsa.gov/guides/effective_user_stories/)

## Project: Admin Project

Select Data

Summary Charts

Start Over



Select cancer endpoint



Select start of follow-up



Select censoring rules



Select questionnaire data



Summary

Cancer Site Group

Acute Monocytic Leukemia x



Cancer Site Group	SEER Code	ICD-O-3 Site	ICD-O-3 Histology	Total	<input type="checkbox"/>
Acute Monocytic Leukemia	35031	C421	9891	17	<input checked="" type="checkbox"/>
Total Number of Cancer Records				17	

NEXT

Figure 1a

## Project: Admin Project

Select Data

Summary Charts

Start Over



Select cancer endpoint



Select start of follow-up



Select censoring rules



Select questionnaire data



Summary

The next questions ask about follow-up time and whether to exclude prevalent cancers.

For your analysis, when should follow-up begin? Please select one of the following.

- CTS Baseline, i.e. Questionnaire 1 (1995-1996)
- Questionnaire 2 (1997-1998)
- Questionnaire 3 (2000-2002)
- Questionnaire 4 (2005-2008)
- Questionnaire 5 (2012-2015)
- Questionnaire 6 (2017-2019)
- Other (please specify):


Participants with prevalent cancer at the start of your follow-up can be included or excluded. Please choose whether to exclude participants who had cancer at the start of follow-up:

- Exclude all participants who had a prevalent cancer of any type at the start of follow-up.
- Exclude only the participants who had a prevalent cancer of interest (i.e., the cancer endpoint for your analysis) at the start of follow-up.
- Include all participants, even those with prevalent cancer at the start of follow-up.

BACK

NEXT

Figure 1b

 Project: Admin Project[Select Data](#)[Summary Charts](#)[Start Over](#)

The next questions ask about the end of follow-up and censoring rules.

CTS follow-up data are currently complete through 12/31/2020. Will your analysis include all eligible cases diagnosed through 12/31/2020?

- Yes  
 No

By default, CTS analyses censor participants when they are diagnosed with any other cancer; die; move out of California; if applicable, undergo risk-eliminating surgery (hysterectomy, bilateral oophorectomy, or bilateral mastectomy for analyses of uterine, ovarian, or breast cancers, respectively); or reach the administrative censoring date (12/31/2020). You can choose whether to censor participants who are diagnosed with any other cancer. Please specify your choice for censoring rules:

- Use the default CTS rules. Follow-up time will end at the earliest of the dates described above.  
 Do not censor at diagnosis of any other cancer. Follow-up time will end at the earliest of the other dates described above.

[BACK](#)[NEXT](#)

Figure 1c

 Project: Admin Project

Select Data

Summary Charts

Start Over



Select cancer endpoint



Select start of follow-up



Select censoring rules



Select questionnaire data



Summary

Over 1200 variables are available for your analysis. To make your selections, check the boxes and review them in the window on the right.

The **Sections** view below is organized by the section titles on the physical questionnaires (1-6). If you prefer to review across all questionnaires by topic area, click **View By Topics**. You can also use the search bar to search by key terms.

Every analysis automatically includes the most commonly used CTS variables, which are marked in green in the table below and are already included in your dataset.

Q1
Q2
Q3
Q4
Q4 Mini
Q5
Q5 Mini
Q6

View by Topics

🔍

**My Selections**


Variable Name	Description		
<input checked="" type="checkbox"/>	Section : Background and environment	(10/53)	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Section : Reproductive history	(7/82)	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Section : Health history	(3/42)	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Section : Personal and family medical history	(12/39)	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Section : Physical activity	(2/33)	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Section : Diet	(1/57)	<input type="checkbox"/>
<input checked="" type="checkbox"/>	Section : Alcohol and tobacco use	(5/29)	<input type="checkbox"/>

BACK
NEXT

**Q1**

- adopted
- age\_at\_baseline
- age\_dad\_atbirth
- age\_mom\_atbirth
- alchl\_analyscat
- allex\_hrs\_q1
- allex\_life\_hrs
- birthplace
- birthplace\_dad
- birthplace\_mom
- bmi\_q1
- brca\_selfsurvey
- cancer\_self\_q1

Figure 1d

 Project: Admin Project

[Select Data](#)
[Summary Charts](#)
[Start Over](#)


Select cancer endpoint



Select start of follow-up



Select censoring rules



Select questionnaire data



Summary

## Cancer Endpoint

Site Group Name	SEER ID	ICD O3 CDE	Histologic Type
Acute Monocytic Leukemia	35031	C421	9891

## Start of Follow-up

**Follow-up begins:** CTS Baseline - Questionnaire 1 (1995-1996)

**Participants with prevalent cancer:** Exclude only the participants who had a prevalent cancer of interest (i.e., the cancer endpoint for your analysis) at the start of follow-up.

## Censoring Rules

**Follow-up ends:** Include all eligible cases diagnosed through 12/31/2020.

**Participants with any other cancer:** Do not censor at diagnosis of any other cancer. Follow-up time will end at the earliest of the other dates described above.

## Selected Variables in Questionnaire

Q1	
adopted	3. Were you adopted? From the first CTS questionnaire (1995-1996).
age_at_baseline	2. Age at Q1. Created by subtracting birth date from questionnaire 1 fill date. From the first CTS questionnaire (1995-1996).

Figure 1e part 1

age_from_atbirth	How old was mom when you were born? From the first CTS questionnaire (1995-1996).
alchl_analyscat	83. Cat of alcohol g/d past yr for bc analysis. From the first CTS questionnaire (1995-1996).
allex_hrs_q1	68, 69. Strenuous + moderate exercise hrs/wk over year past 3 yrs. From the first CTS questionnaire (1995-1996).
allex_life_hrs	68, 69. Strenuous + moderate exercise hrs/wk over yr for lifetime. From the first CTS questionnaire (1995-1996).
birthplace	5. Where were you born? From the first CTS questionnaire (1995-1996).
birthplace_dad	5. Where was your father born? From the first CTS questionnaire (1995-1996).
birthplace_mom	5. Where was your mother born? From the first CTS questionnaire (1995-1996).
bmi_q1	65. Body mass index (kg/m**2). From the first CTS questionnaire (1995-1996).
brca_selfsurvey	67. Breast cancer self(survey data only). From the first CTS questionnaire (1995-1996).
cervca_self_q1	67. Have you ever had cervical cancer? From the first CTS questionnaire (1995-1996).
cig_day_avg	84-86. Avg number of cigarettes smoked per day. From the first CTS questionnaire (1995-1996).
colnca_self_q1	67. Have you ever had Colon/Rect cancer? From the first CTS questionnaire (1995-1996).
diab_self_q1	67. Have you ever had Diabetes? From the first CTS questionnaire (1995-1996).
endoca_self_q1	67. Have you ever had endometrial cancer? From the first CTS questionnaire (1995-1996).
fullterm_age1st	27-28. Age 1st full term preg(lb,sb). From the first CTS questionnaire (1995-1996).
hbp_self_q1	67. Have you ever had Hi Blood Press? From the first CTS questionnaire (1995-1996).
height_q1	65. Height today. From the first CTS questionnaire (1995-1996). Derived.
hodg_self_q1	67. Have you ever had - Hodgkins Dis? From the first CTS questionnaire (1995-1996).
leuk_self_q1	67. Have you ever had leukemia? From the first CTS questionnaire (1995-1996).
lungca_self_q1	67. Have you ever had lung cancer? From the first CTS questionnaire (1995-1996).
meln_self_q1	67. Have you ever had Malig Melanoma? From the first CTS questionnaire (1995-1996).
menarche_age	21. Age at menarche. From the first CTS questionnaire (1995-1996).
meno_stattype	34-36,41,56. Menopausal status and type of menopause. From the first CTS questionnaire (1995-1996).
nih_ethnic_cat	6. For grant inclusion enrollment report - Part A-Ethnic Category. From the first CTS questionnaire (1995-1996).
oralcntr_ever_q1	24-26. Ever used OC. From the first CTS questionnaire (1995-1996).
oralcntr_yrs	24-26. Total yrs used OC. From the first CTS questionnaire (1995-1996).
ovryca_self_q1	67. Have you ever had ovarian cancer? From the first CTS questionnaire (1995-1996).
participant_race	6. Participant's race. Chinese, Filipino, Hawaiian, Japanese & Korean are all coded to Asian/Pacific Islander. If White and any other race then the racex variable was coded to the other race. If not White and more than one race was reported then the racex variable was coded to Other/Mixed. From the first CTS questionnaire (1995-1996). Derived.
preg_ever_q1	28. Ever pregnant. From the first CTS questionnaire (1995-1996).
preg_total_q1	28. Total Number of Pregnancies. From the first CTS questionnaire (1995-1996).
smoke_expcat	84, 87, 88. Smoking exposure categories. From the first CTS questionnaire (1995-1996).
smoke_totpackyrs	2, 84-86. Total pack years of smoking. From the first CTS questionnaire (1995-1996).
smoke_totyrs	2, 84, 85. Total years smoked. From the first CTS questionnaire (1995-1996).
thyrca_self_q1	67. Have you ever had Thyroid cancer? From the first CTS questionnaire (1995-1996).
twin	4. Are you a twin? From the first CTS questionnaire (1995-1996).
vit_mulvit_q1	71-72. Multivit and single vit use variable. From the first CTS questionnaire (1995-1996).
weight_q1	65. Weight today. From the first CTS questionnaire (1995-1996). Derived.

Save data as

BACK

GENERATE DATA

Figure 1e part 2

Project Characteristics that are Captured and Tracked for all Projects in the CTS Researcher Platform



### Research Team

- ✓ Personnel
- ✓ Institution
- ✓ Primary contact



### Timeline

- ✓ Start date
- ✓ Status
- ✓ End date
- ✓ Current project stage



### Human Subjects

- ✓ IRB approval status
- ✓ Inclusion of sensitive data
- ✓ IRB project type



### Design Details

- ✓ Study endpoint
- ✓ Biospecimen inclusion
- ✓ Project summary
- ✓ Study design type
- ✓ Geospatial data



### Data Sharing

- ✓ Data-use agreements
- ✓ Consortium or pooling project participation

Figure 2



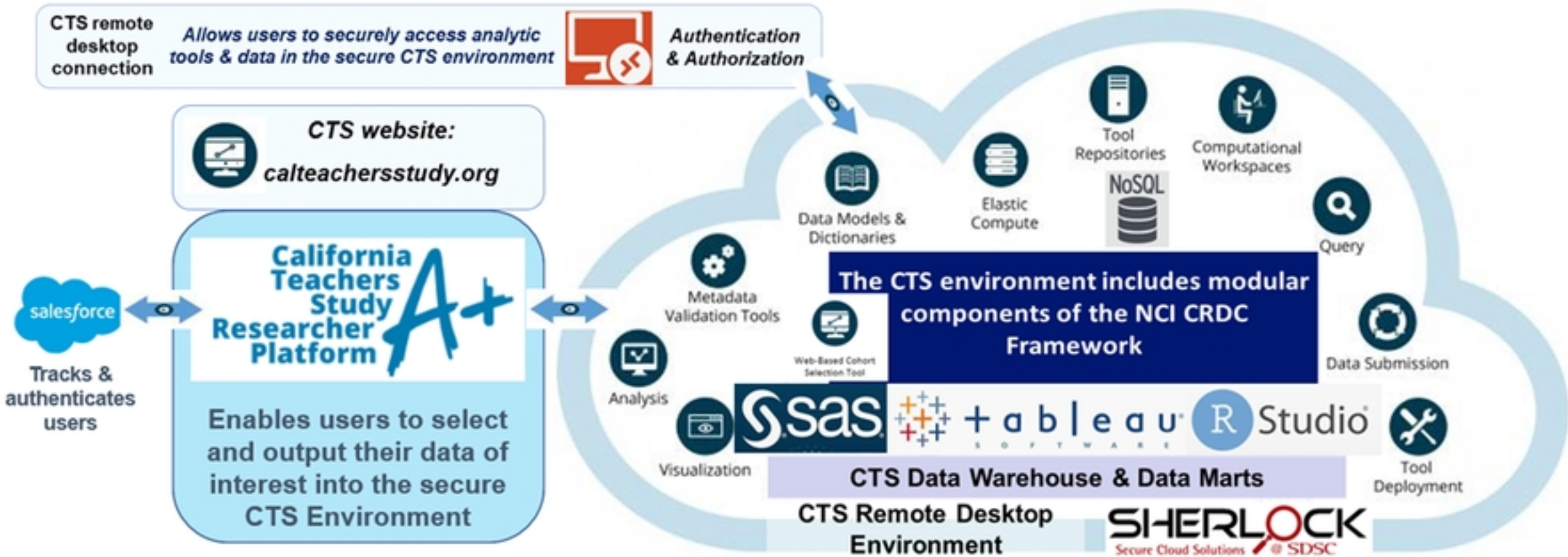
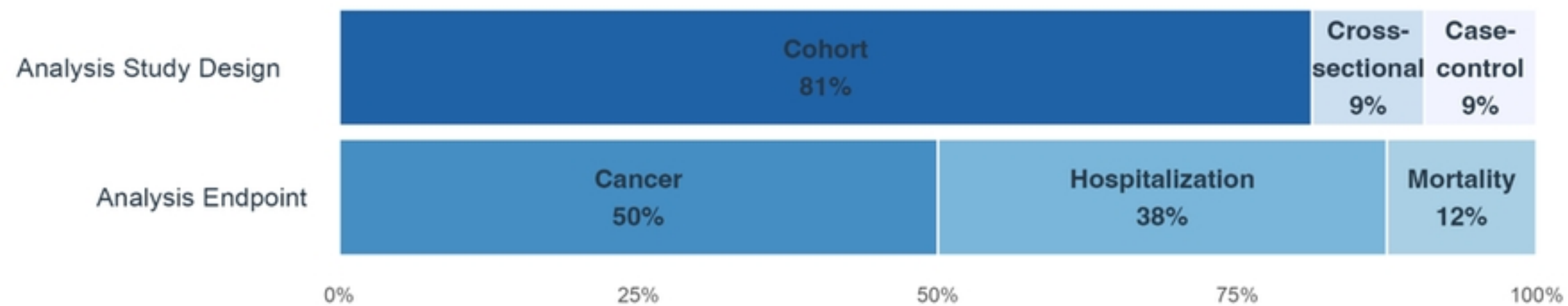


Figure 3



### Types of Exposure Data Included in Projects Using the CTS Researcher Platform



Figure 4