Where Do Our Graduates Go? A Tool Kit for Tracking Career Outcomes of Biomedical PhD Students and Postdoctoral Scholars

Elizabeth A. Silva,[†] Alicia B. Mejía,[†] and Elizabeth S. Watkins[†‡§*]
[†]Graduate Division, [‡]Department of Anthropology, History, and Social Medicine, and
[§]Student Academic Affairs, University of California, San Francisco, San Francisco, CA 94158

To the Editor:
Universities are at long last undertaking efforts to collect and disseminate information about student career outcomes, after decades of calls to action. Organizations such as Rescuing Biomedical Research and Future of Research brought this issue to the forefront of graduate education, and the second Future of Biomedical Graduate and Postdoctoral Training conference featured the collection of career outcomes data in its final recommendations (Hitchcock *et al.*, 2017). More recently, 48 institutions have assembled as the Coalition for Next Generation Life Science (CNGLS), committing to ongoing collection and dissemination of career data for both graduate and postdoc alumni. A few individual institutions have shared snapshots of the data in peer-reviewed publications (Silva *et al.*, 2016; Mathur *et al.*, 2018) and on websites. As more and more institutions take up this call to action, they will be looking for tools, protocols, and best practices for ongoing career outcomes data collection, management, and dissemination.

Here, we describe the development and implementation of a methodology for collecting, examining, and reporting graduate and postdoctoral career outcomes data at University of California, San Francisco (UCSF). As a service to the community, we describe and share all tools we have developed, and we provide calculations of the time and resources required to accomplish both retrospective and annual data collection and reporting. We also include practical advice for implementation by other institutions, which we hope will increase the feasibility of this endeavor.

DATA OVERVIEW

We have developed and are maintaining two distinct data sets, one for PhD alumni (Figure 1) and one for postdoctoral alumni (not shown). Our PhD alumni data set includes every student who began a PhD program at UCSF since 1996. A record is created for each student as he or she matriculates to the program. Our postdoctoral alumni data set includes every postdoctoral scholar (postdoc) who left the institution since 2011. In both cases, we include all available demographic information, previous education, program and degree information, and job titles and employers. Data are transferred from the student information system (PhD alumni, via application programming interface [API]) and the Office of Institutional Research (postdoctoral alumni, based on Human Resources records). Career information is collected annually for up to 15 years after a student or postdoc leaves the institution and is displayed on the public website in 5-year increments and/or 5-year aggregates (https://graduate.ucsf.edu/program-statistics; https://postdocs.ucsf.edu/postdocs-ucsf). A full description of all metadata for the PhD data set is provided in Supplemental Material S1 and for the postdoc data set in Supplemental Material S2.

CBE Life Sci Educ December 1, 2019 18:le3
DOI:10.1187/cbe.19-08-0150
*Address correspondence to: Elizabeth S. Watkins (elizabeth.watkins@ucsf.edu).

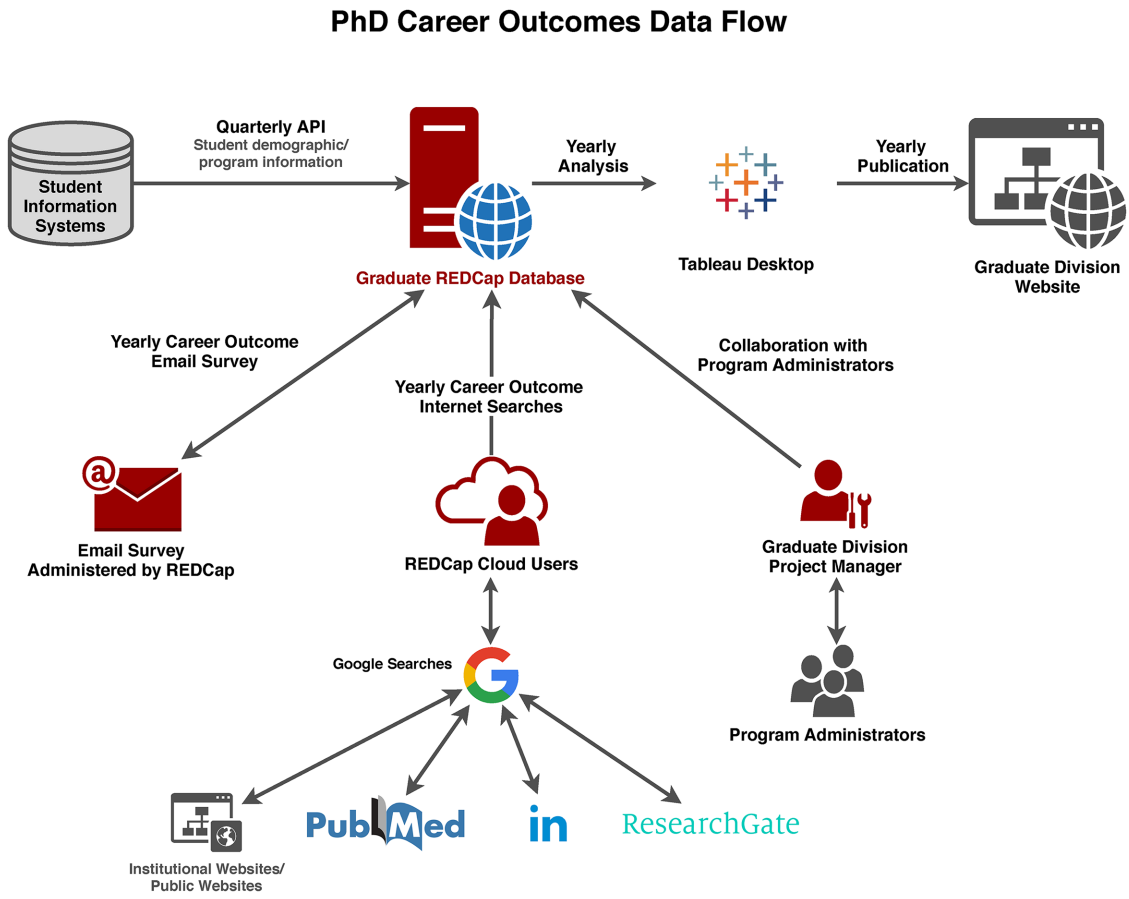

FIGURE 1. Overview of data flow for the PhD alumni career outcomes. Postdoctoral outcomes data flow (not shown) is similar, except where noted. Basic demographic and degree information is transferred to REDCap from the student information system (postdoctoral data come from Human Resources). Annually, staff administer a onetime survey requesting current employment information from the PhD alumni and conduct online searches for those who do not return a survey (postdocs and PhDs). Employment data are recorded in REDcap. Data are then uploaded to Tableau for public display on the graduate division (or postdoctoral) website. PhD data are also shared with the PhD program staff: staff provide updates to the project manager about program alumni, and data collected by our team are shared with the program staff.

REPOSITORY DESIGN

We considered multiple systems and platforms for our data collection, management, curation, and archiving, including Microsoft Excel, Microsoft Access, Smartsheets, Salesforce, and REDCap. We considered the following features in our analysis:

Required:

- Cloud based to allow multiple users
- Compatible with Mac and PC to allow multiple users
- No requirement for individual user license to enable access by multiple users
- Export and import (e.g., comma-separated value [.CSV] files) in a format that enables data to be used on other platforms for analysis or dissemination and data sets from other sources (e.g., student information system) to be uploaded to the database
- Flexible data fields to enable us to add and remove fields as we build the data set and as we refine the uses for the database
- Features that ensure data stability and integrity, such as versioning or protection against data corruption when multiple users access the system

Recommended:

- Variable user permissions to allow access for stakeholders
- Open source or otherwise accessible for additional development work
- Survey option in which survey response fields are directly linked to database fields, to reduce the amount of time for Internet searches
- Custom report builder for providing data to stakeholders

Ultimately, we determined that a database that could provide the paradoxical qualities of flexibility and stability were most crucial to this undertaking. That is, while we had an idea of the breadth of careers and organizations in which our alumni might be employed, we anticipated there would be scenarios that we could not predict. Therefore, we needed to

be able to adjust the fields (flexibility) as we gained a better understanding of the data without risk of corrupting it (stability). We opted for REDCap, developed by Vanderbilt University, which is free and open source and includes all of our required features (www.project-redcap.org). REDCap allows for "data access groups," which means different users can be given different access to subsets of data as defined by the administrator, including view-only access. This flexibility allows various stakeholders on campus to access the data sets relevant to them without violating the Federal Education Rights and Privacy Act and without risking corruption of the data. REDCap also has a survey function (more on that later) and can generate reports that can be exported as .CSV files. Finally, REDCap allows for the development of APIs for data import and export. We took advantage of this feature and updated our student demographic and enrollment data directly from UCSF's student information system. We have included our REDCap data dictionaries for our graduate student and postdoc outcomes databases as Supplemental Material S3 and S4, respectively. These data dictionaries can be used to re-create an empty database in REDCap, which can then be modified according to institutional needs and interests.

RETROSPECTIVE AND ONGOING DATA COLLECTION

We considered there to be two phases to our data-collection effort: retrospective and ongoing. Retrospective data collection, in which we attempt to identify past positions of alumni who graduated previously, is much more labor intensive than ongoing collection, in which only current positions for alumni are collected. We describe each separately, knowing that some institutions may wish to skip a retrospective phase.

We began retrospective data collection in 2017, relying entirely on Internet searches, as previously described (Silva *et al.*, 2016). Our aim was to record one position per year for up to 15 years after leaving the institution, as an annual snapshot taken roughly around June–August each year. We chose to forgo collection start and end dates, because they are unreliably available and would necessitate a more complicated database structure. Note that occasionally we would fail to document a very briefly held position. Since our earlier published study, we have found that LinkedIn is a superior platform for gathering career information, particularly for individuals in the private sector. Yet Google is a superior search engine; a Google search for [First Name] [Last Name] "LinkedIn" is more likely to yield relevant results. Additionally, Google's search results can be influenced by logging into a LinkedIn account for a user who is well connected to your alumni. Google's enriched results incorporate user information (sites visited, current location, log-ins to social media accounts) such that, when a user is logged into a LinkedIn account with more connections to institutional alumni, Google is more likely to return top hits for institutional alumni. Author E.A.S. has 800+ LinkedIn connections, many of whom are UCSF staff, students, or alumni. When she was logged into her account as the searcher, Google was more likely to return the correct individual as the top result. Furthermore, identification of individuals as second- or third-order connections to the LinkedIn user serves as verification and helps disambiguate individuals with similar names.

We launched ongoing data collection in 2018, recording the most recent position for each alumnus. For our PhD alumni data set, we introduced survey results into our data-collection method. We sent the following five-item survey to all alumni for whom we had an email address:

- What is your job title?
- What is the name of your organization/institution/company?
- City
- State (or country if not the United States)
- Salary (optional)

This survey was sent to 1732 alumni with a functioning email address, and 800 responses were received, for a return of 43% and representation of 30% of our alumni. We attribute the high response rate to two factors: 1) brevity of the survey and 2) an appeal to the cause. The email invitation stated that the survey would take less than 1 minute to complete and explained that the data collected would be used for transparent and thorough reporting of career outcomes. Respondents were also assured that data would be displayed anonymously in aggregate. We included a link to our public display of retrospective data so that prospective participants could see how the data were used. Once the survey was complete, we updated the remaining 70% of PhD alumni using the same approach as for retrospective data collection.

Although we publicly display results in 5-year increments, we identified three significant advantages to annual data collection. First, we found it easier to locate and update each individual annually via Internet searching. Second, we believed that annual updates would reveal more nuanced career trajectories, which would assist our student and postdoctoral services staff as they advise trainees on career exploration and decision making. Third, the National Institutes of Health require that institutional training grants (T32) awardees provide annual updates of the career outcomes of funded trainees, for which these data can serve as a source.

We have been unable to locate some individuals by email or Internet searching. For example, individuals who are unemployed rarely identify as such. We observed that those working in clinical practice are disproportionately difficult to find online, because they neither use LinkedIn nor have comprehensive profile pages on institutional websites. Alumni who left the institution more recently are easier to find, and current position is easier to find than any past-held position. In Table 1 we summarize the proportion of PhD alumni for whom we were unable to find information (unknowns) in our retrospective study, comparing current position (2017) to past-held positions (1996–2016).

CAREER CLASSIFICATION

To classify graduate student and postdoctoral alumni careers, we use the taxonomy developed collectively in 2017 by representatives of universities with NIH Broadening Experiences in Scientific Training awards, members of Rescuing Biomedical Research and the founding institutions of the CNGLS. Classification terms are applied by our staff, rather than the alumni themselves, in an effort to ensure consistency. We find that most positions fall clearly into categories for career type and sector; however, many jobs do not fall clearly in a specific career category for job function. When a position does not clearly fall into

TABLE 1. Number and percent of PhD alumni for whom data were missing in our initial retrospective study (2017)

Cohort	Count of trainees	% No job title[a] 1996–2016	Completely unknown[b] 1996–2016	% Completely unknown[b] 1996–2016	% No job title[c] 2017	Completely unknown[d] 2017	% Completely unknown[d] 2017
1996	83	39	3	4	17	2	5
1997	77	45	6	8	17	3	6
1998	102	35	14	14	3	3	3
1999	98	38	6	6	16	9	10
2000	122	35	10	8	19	13	11
2001	108	23	2	2	12	3	3
2002	142	28	12	8	16	22	15
2003	147	22	5	3	12	7	5
2004	127	24	6	5	13	11	9
2005	153	15	0	0	10	2	1
2006	128	16	11	9	10	13	10
2007	137	16	1	1	9	1	1
2008	116	10	1	1	9	4	3
2009	96	13	0	0	11	0	0
2010	70	13	2	3	10	2	3
2011	50	18	3	6	12	3	6
2012	38	18	3	8	19	4	11
Total	1794	24	85	5	12	102	6

[a]Job title not found for previous years.
[b]No information found for previous years (job title, organization, location).
[c]Job title not found for 2017 (current position at the time of search).
[d]No information found for 2017 (current job title, organization, location at the time of search).

a category, we discuss its best placement as a group and then add notes to the definitions associated with each category to clarify how the categories should be applied (Supplemental Material S5).

Each year, once initial classification is completed, we randomly assign a subset of records for re-review by coders—those who applied the classifications to the alumni. In the retrospective phase of our study, 200 individuals were assigned to each of three reviewers. Using a basic spreadsheet, each reviewer indicated records that might require review and provided notes describing the issue. In this process, we identified a few errors, but more importantly, we identified inconsistencies in coding that could be rectified in bulk. For example, a number of institutions, including UCSF, have fellows' programs that provide a pathway from graduate school to independent research, effectively skipping the postdoctoral stage. Our team had discrepant understandings of whether to classify these positions as training positions or independent faculty-like positions ("faculty, tenure-track not applicable"). The audit highlighted the discrepancy and prompted classification decisions. Any necessary reclassifications were then extended to the full data set.

A summary of our audits (from both the retrospective and ongoing studies) is provided in Table 2. We list the type of correction made in order of frequency of occurrence. For PhD alumni, by far the most frequent inconsistency was in classification of faculty as tenure track. In the biomedical academic workforce, it is notoriously difficult to ascertain whether a faculty position is tenure/tenure track from a LinkedIn profile, university or lab website, or curriculum vitae. For postdoctoral alumni, the most common correction was in updating the Internet links where job information could be found for individuals.

RESOURCES NEEDED

The scope and scale of this project demanded significant staff time. Here, we estimate the amount of time required and describe the roles and responsibilities of the primary personnel. We also provide a more detailed summary of our timeline, milestones, and team members in our charter document (here and in Supplemental Material S6).

The primary personnel for the project are the project sponsor, the project/data manager, and the project support staff. The project sponsor makes decisions for the overall project and directs the data collection and analysis. The sponsor is also the primary person responsible for auditing. The sponsor should typically be a dean, associate/assistant dean, or director of a relevant unit, in our case author E.A.S. The project/data manager documents project goals, documents and communicates project status, tracks time and effort spent, identifies roles and responsibilities, and monitors other project details. Secondary roles for the manager include data collection, consolidation and management in REDCap, database administration, and data-quality audits and cleanup (author A.B.M.). Project support staff are those who are primarily responsible for searching and documenting the career outcomes in REDCap and classifying the job titles and employers. In our case, we hired undergraduate and graduate student interns to do this work.

The data collection and classification for our 15-year retrospective study of PhD student alumni, undertaken in 2017, was completed in 3 months (June 15 to September 15). Through the remainder of 2017 and into 2018, a project sustainment plan was developed and implemented by the project manager, and the project was expanded to include retrospective data collection of the postdoc population. An update of all PhD and postdoc alumni outcomes was

TABLE 2. Summary of data audit for PhD and postdoctoral alumni data

Postcollection data audit statistics	PhD alumni	Postdoctoral alumni
Total trainees	2557	2355
Total entries	16,084	12,921
Total trainees reviewed	546	531
Total entries reviewed	3536	2576
Total trainees corrected	49	191
Total number of corrections identified (entries)	153	538
Correction type and frequency:	Tenure track: 89 Other data entry: 20 Other classification error: 20 Added/corrected link: 19[a] Other miscellaneous: 2 Trainee information from Student Information System: 1 Group leader: 1 Entrepreneur : 1	Added/corrected link: 199[a] Other classification error: 95 Tenure track: 77 Other data entry: 51 Group leader: 50 UCSF associate/assistant specialist/researcher: 24 UCSF title: 21 Trainee information from Office of Institutional Research: 12 Entrepreneur: 8 Other miscellaneous: 1
Total number of corrections accepted (entries)	73	427
% Corrections accepted to sample size	2	17
% Corrections accepted to population size	0.45	3

[a]URL for LinkedIn, institutional website, or other Internet site where alumnus information is available.

completed in the three summer months of 2018. In Supplemental Material S7, we provide a worksheet that estimates the resources that would be required at other institutions for a retroactive data search and for an annual update of the alumni outcomes. We found that it took an average of 10 minutes to complete data collection for each individual trainee in our retrospective study and an average of 5 minutes per trainee in our annual update. The total number of hours required for the project (data collection by support staff plus project oversight and management by the sponsor and manager) will depend on the size of the cohorts and the number of cohorts the institution wishes to track.

GLOBAL RECOMMENDATIONS

Many institutions report that they have delayed commitment to these projects due to concerns about the resources required. Having done the work to implement systems for retrospective and ongoing data collection, we share all of our materials and resources here to motivate other institutions to take up this call to action. Transparency in career outcomes for PhD students and graduates is an achievable goal and, we argue, a responsibility that universities must fulfill. We end with three global recommendations in order to encourage other institutions to adopt or adapt our approach with their own alumni.

1. *Don't let the perfect be the enemy of the good.* We acknowledged from the outset that we would not be able to find all alumni or to categorize every job title with precision. We chose a repository with sufficient flexibility so that post hoc adjustments to the data would be feasible. We also made our peace with missing information in the retrospective study, knowing that the quantity and quality of data will improve as we add new graduates to the data set in the ongoing collection.
2. *Develop a project charter.* A project charter sets boundaries on the scope and scale of the project, articulates the roles of the personnel, and includes a timeline and description of the project milestones. This document was critical for ensuring the project progressed at an acceptable pace and for preventing "mission creep"—unplanned expansions that present barriers to completion. A charter was particularly important for our postdoc data set. For decades there has been a dearth of data about postdocs (National Academy of Sciences, National Academy of Engineering, and Institute of Medicine, 2014). As the project grew, so did the enthusiasm for expanding to include additional data points that were not directly relevant to the objectives of the project (e.g., date of birth, time in previous postdoc). While these data points are interesting and valuable, defined limits on the scope of the project are necessary for completion. We provide our project charter in Supplemental Material S6 as a sample.
3. *Collaborate with campus stakeholders.* On every campus, career outcomes data may be collected and reported by a variety of stakeholders, often with little coordination of efforts and resources. Coordination with stakeholders offers the opportunity to improve the quality of the data set while reducing the overall institutional resources required. For instance, graduate program staff and faculty often have firsthand knowledge of the current positions of graduates, having maintained personal connections years after graduation. In our experience, graduate programs may have reliable data, but not a reliable platform for storing and analyzing alumni information. Collaboration with the graduate programs involves collecting accurate alumni information and offering a central platform along with user support for accessing the data. Similarly, T32 program directors are required to report first position and

current position for every funded trainee for 15 years. Meeting these reporting requirements is an enormous undertaking and is resource intensive for each individual training program; centralized administration of these efforts reduces overall resource burden by minimizing duplicate efforts and by capitalizing on the expertise of a data management specialist. Equally, a great deal of data can be extracted from the reports. Finally, our alumni relations office generously shared email contact information from alumni in their database to assist with our survey. In exchange, we reported our survey response rates to alumni relations, who reported the responses as successful touchpoints.

REFERENCES

Hitchcock, P., Mathur, A., Bennett, J., Cameron, P., Chow, C., Clifford, P., ... & Engelke, D. (2017). The future of graduate and postdoctoral training in the biosciences. *eLife*, *6*, e32715. https://doi.org/10.7554/eLife.32715

Mathur, A., Cano, A., Kohl, M., Muthunayake, N. S., Vaidyanathan, P., Wood, M. E., & Ziyad, M. (2018). Visualization of gender, race, citizenship and academic performance in association with career outcomes of 15-year biomedical doctoral alumni at a public research university. *PLoS ONE*, *13*(5), e0197473. https://doi.org/10.1371/journal.pone.0197473

National Academy of Sciences, National Academy of Engineering, and Institute of Medicine. (2014). *The postdoctoral experience revisited*. Washington, DC: National Academies Press. https://doi.org/10.17226/18982

Silva, E. A., Des Jarlais, C., Lindstaedt, B., Rotman, E., & Watkins, E. S. (2016). Tracking career outcomes for postdoctoral scholars: A call to action. *PLoS Biol*, *14*(5), e1002458. https://doi.org/10.1371/journal.pbio.1002458