The Pacific Innovations, Knowledge, and Opportunities (PIKO) Program: A Data Lifecycle Research Experience

Rylan Chong PhD; Laura Tipton PhD

Abstract

Pacific evidence-based clinical and translational research is greatly needed. However, there are research challenges that stem from the creation, accessibility, availability, usability, and compliance of data in the Pacific. As a result, there is a growing demand for a complementary approach to the traditional Western research process in clinical and translational research. The data lifecycle is one such approach with a history of use in various other disciplines. It was designed as a data management tool with a set of activities that guide researchers and organizations on the creation, management, usage, and distribution of data. This manuscript describes the data lifecycle and its use by the Biostatistics, Epidemiology, and Research Design core data science team in support of the Center for Pacific Innovations, Knowledge, and Opportunities program.

Keywords

data lifecycle, health, evidence-based, clinical, translational, Native Hawaiian health, Pacific Islander health, Filipino health, data science, PIKO

Abbreviations and Acronyms

API = application programming interface
BERD = Biostatistics, Epidemiology, and Research Design
CTR = clinical and translational research
NIH = National Institutes of Health
PIKO = Pacific Innovations, Knowledge, and Opportunities

Introduction

The Pacific Innovations, Knowledge, and Opportunities (PIKO) program is grant funded by the National Institute of General Medical Sciences (U54GM138062), and its primary aim is to improve the health of Native Hawaiians, Pacific Islanders, Filipinos, and people who are medically underserved in Hawai‘i.[1] This program is made up of 7 cores that function independently and collaboratively to provide support to PIKO researchers, defined as a researcher who is either going through the process of applying for PIKO pilot project funding support or has received PIKO funding support. PIKO researchers are often, but not exclusively, early-stage investigators , who are within 10 years of their terminal degree and have not had a substantial National Institutes of Health (NIH) independent research award. One of the cores that is integral to PIKO researchers' success is the Biostatistics, Epidemiology, and Research Design (BERD) core. The broad goals of this core are to support PIKO researchers through all stages of their study and for the PIKO researchers to develop competencies in Pacific evidence-based clinical and translational research (CTR). Core goals are met through consultation, training, and mentoring from each of the BERD components: biostatistics, epidemiology, research design, data science, psychometrics, mixed methods, economics, and database design.[1] In this column, the traditional Western research process is introduced as the context for BERD‘s innovation in integrating the data lifecycle research process. This is followed by a description of how PIKO researchers were exposed to data lifecycle competencies through the lens of the PIKO BERD core data science team.

Traditional Western Research Process

The traditional Western research process, described in 7 steps in **Figure 1**, is an important process to extend the current body of knowledge in most scientific disciplines.[2] Individuals who are interested in research are primarily exposed to the research process in graduate school through a class project, grant funded project, a master's thesis, or a doctoral dissertation.[2] However, not all researchers' journeys are the same and not everyone goes through a graduate school research experience. Some researchers, including some PIKO pilot project awardees, are exposed to research after finishing their education during their career. For example, a community health practitioner who has no formal research experience could join PIKO to learn how to conduct research. Individuals who experienced the research process are expected to obtain the competencies illustrated in **Figure 1**.

Through the BERD core, the PIKO program provides support for PIKO researchers to help develop competencies on the traditional Western research process in 3 phases. The first phase is the design and submission of an abstract proposal, in which a researcher is exposed to the Activities 1-3 of the research process in Box 1 on the left-hand side of **Figure 1**. Next, researchers whose abstract proposals are selected to continue in the program are invited to submit a full proposal for funding, in which they revisit the same 3 activities in greater detail. In the third phase and Box 3 of the PIKO program, a funded full proposal is executed, and the PIKO researcher performs the remaining activities of the research process.

Performing CTR in the Pacific is greatly needed as there is an acknowledged lack of research in all areas of health for Native Hawaiians, Pacific Islanders, Filipinos, and people who are medically underserved in Hawai‘i.[3] However, several challenges stem from the data that contribute to the lack of CTR in the Pacific. These include: (1) small sample sizes; (2) limited availability, access, and usability of the data; (3) lack

of awareness of the various types of data that can be collected and transformed; and (4) analytics that will produce useful and meaningful results. The data lifecycle process is a data science data management tool that addresses these challenges and is used in various disciplines, including as biology, environmental sciences, economics, cybersecurity, library sciences, business, political science, and social science.[4,5] The data lifecycle offers further insights into competencies that can address some of the data challenges of the Pacific.

Data Lifecycle

Launching the data lifecycle begins with understanding the processes that make up this framework. Illustrated in **Figure 2**, the data lifecycle focuses on what happens to the data from the formation of the question through the end of the project. This process is used to guide the actions of the PIKO BERD core data science team.

As with the traditional Western research process, the 7 data lifecycle activities can be aligned with the PIKO program phases. The abstract proposal and full proposal mostly encompass planning, which is the first lifecycle activity. The remaining lifecycle activities (collect, process, analyze, preserve, share, and determine the course), all happen in the full proposal execution phase.

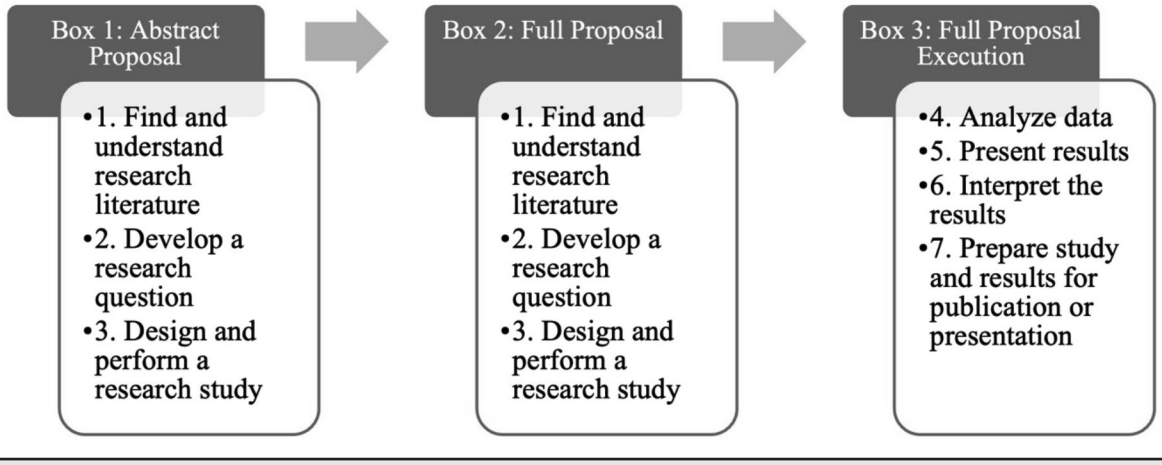

Figure 1. Three Phases of the PIKO Program and How they Align to the Traditional Western Research Process[2]

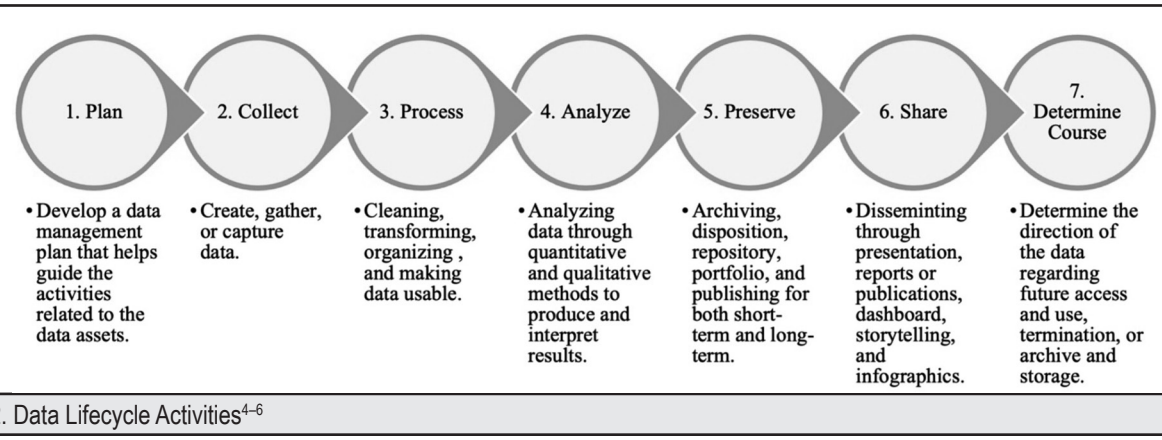

Figure 2. Data Lifecycle Activities[4–6]

BERD Data Science Data Team Lifecycle Experience

This section provides details on how the data lifecycle process is utilized in the BERD core data science team with PIKO researchers. All stages of the data lifecycle are relevant. The primary role of the BERD data science team is to provide guidance and support while working with PIKO researchers.

Activity 1, "plan" occurs during the abstract proposal and the full proposal phases of the PIKO program. During these phases, a PIKO researcher develops a research plan that includes all components of a NIH proposal abstract, including data to be collected (Activity 2 of the data lifecycle) and how that data will be processed (Activity 3) and analyzed (Activity 4). Activities 5-7 of the data lifecycle, preservation, sharing, and determining the course of the data, are not described in the proposal. The data science team interacts with the PIKO researchers, usually after the abstract proposal phase, to discuss and provide recommendations on the reviewer comments or to provide any support that will benefit the researcher's project going forward. We found the most valuable tool to facilitate discussions with a researcher during the planning activity is a data management plan. Discussions can then cover project information that is not accurately described in the proposal: data types, formats, and sizes; related tools, software, and code to use with the data; collection activities and timeline; file formats and storage; processing of the data; and any standards and policies applied to the data.[7–9]

Once a PIKO researcher's plan is funded, PIKO researchers start their collection of structured, unstructured, and semi-structured data (Activity 2). Structured data is highly organized and has a tabular structure.[10] An example is survey data where values are stored in rows and columns. Conversely, unstructured data, such as images, documents and records, and videos, cannot be organized in rows and columns.[10] Finally, semi-structured data is a hybrid of structured and unstructured data. Examples of semi-structured data include social media Application Programming Interface (API) data, web scrapped internet articles, open-ended survey questions, and interviews and focus group narratives.[10] In the first few years of PIKO, there were some cases where PIKO researchers needed support to explore the more complex, or less traditional forms of data that could be collected to explain quantitative results, to address the lack of data for a particular question, or to investigate a question. Based on the case at hand, the data science team needed to be creative with their approach by suggesting to explore unstructured and semi-structured data and related tools to collect the data, which included documents and records, images, web scrapped internet articles, and social media API data.

After the data are collected, PIKO researchers begin Activity 3 of the data lifecycle and processed the data through three activities: cleaning, transforming, and organizing. Data cleaning includes removing, validating, modifying, aggregating, and subsetting the data. Transforming the data includes converting data from one format or structure to another. Data organization includes categorizing and classifying data to make it usable for the analysis activity.[4,6] The primary activity the data science team performed was to validate the PIKO researcher's data usefulness and to provide suggestions on transforming the data values from character to numeric for an easier analysis.

The analysis activity of the data lifecycle, Activity 4, is the area where the data science team and the PIKO researchers interact the most. The data science team currently offers 4 types of analysis support for researchers, including data exploration (eg, descriptive statistics, charts, figures); data modeling using machine learning techniques (eg, regression, naïve Bayes, decision tree); geospatial or mapping; and text mining (ie, exploring themes and patterns of text using coded algorithms). Data exploration has been the primary type of support offered by the data science team. The data science team provides mentoring and advice, performs 1-on-1 programming exercises, provides code, reviews the accuracy of code, and assists with programming statistics and figures. The other areas include data modeling using regression analysis and addressing inquiries about text mining methods for researchers who are performing qualitative studies.

The remaining Activities 5-7 of the data lifecycle are preserving, sharing, and determining the course of the data, which are primarily performed by PIKO researchers with their teams. Yet, the data science team supports these activities as well, including developing infographics and publications. Regarding infographics, the data science team provides guidance on applications and suggestions on how to create infographics that will communicate the information and results of a project to Pacific stakeholders. For manuscripts submitted for publication, the data science team performs some statistical analyses, confirms results, cleans data, provides recommendations for tables and figures to be used in the publication, suggests the additions of approaches and concepts in the methods section, and assists on responding to reviewers' comments.

Data Lifecycle Competencies

Through working with the data science team and the data lifecycle, the goal is that PIKO researchers obtain 1 or more the following competencies:

- Able to understand and evaluate PIKO pilot project through a data management plan.
- Able to identify and understand unstructured and semi-structured data collection approaches.
- Able to identify and understand data usefulness and transformation.
- Able to identify and understand geographic information system (GIS) and/or text mining analysis methods.

- Able to evaluate data exploratory and data modeling analysis methods.
- Able to communicate information and results through infographics and publication.

While each of these competencies is valuable to PIKO researchers, 3 stand out from the rest through observation and working with PIKO researchers. First, the competency to use a data management plan as a framework to guide a PIKO project discussion is evident in the first phase and second phase of the PIKO program. Second, an understanding of collection approaches for unstructured and semi-structured data, how to make the data useful, how to transform the data, and how to analyze the data are competencies developed during the last phase of the PIKO program in the execution of their full proposal. Lastly, researchers who work with the data science team on infographics and publication learn to transfer and share results during the last phase of the PIKO program.

Conclusion

The PIKO program was established to support culturally responsible and evidence-based clinical and translational research to improve the health of Native Hawaiians, Pacific Islanders, Filipinos, and other people who are medically underserved in Hawai'i.[1] The BERD core is one of the 7 cores of this program and the only core to include a data science component that is meant to support PIKO researchers through all stages of both the traditional Western research process and the data lifecycle. After completing the PIKO program, a PIKO researcher is expected to publish his or her work and start the process of applying for a larger grant. Even if they do not publish their results or apply for a larger grant, the experience of working with the data lifecycle introduces new methods, tools, and resources to early-stage investigators that can support development of competencies in Pacific-related research.

Conflict of Interest

None of the authors identify a conflict of interest.

Acknowledgement

This work was partially supported by Grant number U54GM138062 (PIKO) of the National Institutes of Health (NIH) and by Grant number 2217242 (ALLSPICE) of the National Science Foundation (NSF). The content is solely the responsibility of the authors and do not necessarily represent the official views of the NIH and the NSF.

Authors' Affiliation:
- School of Natural Sciences and Mathematics, Department of Data Science, Analytics and Visualization, Chaminade University of Honolulu, Honolulu, HI

Corresponding Author:
Rylan Chong PhD; Email: rylan.chong@chaminade.edu

References

1. John A. Burns School of Medicine. Center for Pacific innovations, knowledge, and opportunities. Published 2022. Accessed December 20, 2022. https://piko.jabsom.hawaii.edu/
2. Dark M, Stuart L. Innovation in cybersecurity research traineeship in the INSuRE project. In: *Innovative Approaches to Teaching Cybersecurity*. 2015:1-7.
3. Fong M, Braun KL, Tsark JU. Improving Native Hawaiian health through community-based participatory research. *Calif J Health Promot*. 2003;1(SI):136-148. doi:10.32398/cjhp.v1iSI.565
4. Goben A, Raszewski R. The data lifecycle applied to our own data. *J Med Libr Assoc*. 2015;103(1):40-44. doi:10.3163/1536-5050.103.1.008
5. Griffin PC, Khadake J, LeMay KS, et al. Best practice data lifecycle approaches for the life sciences. *F1000Research*. Published online 2018:1-28.
6. Ball A. *Review of Data Management Lifecycle Models*. University of Bath, UK; 2012:1-15. Accessed January 18, 2019. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.224.4219&rep=rep1&type=pdf
7. MIT Libraries. Write a data management plan. Data management. Published 2022. Accessed December 20, 2022. https://libraries.mit.edu/data-management/plan/write/
8. National Institutes of Health. Writing a data management and sharing plan. NIH Scientific data sharing. Published 2022. Accessed December 20, 2022. https://sharing.nih.gov/data-management-and-sharing-policy/planning-and-budgeting-for-data-management-and-sharing/writing-a-data-management-and-sharing-plan
9. U.S. Geological Survey. Data management plans. USGS Science for a changing world. Published 2021. Accessed December 20, 2022. https://www.usgs.gov/data-management/about-usgs-data-management-website
10. Praveen S, Chandra U. Influence of structured, semi-structured, unstructured data on various data models. *Int Jl Sci & Engg Res*. 2017;8(12):67-69.