Enhancing Secondary-use of Electronic Health Records for Geospatial-temporal Population Health Research

Jimmy Phuong

A dissertation

submitted in partial fulfilment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2020

Reading Committee:

Sean Mooney, Chair

Elaine Faustman

Andrea Hartzler

Christina Bandaragoda

Program authorized to offer degree:

Biomedical and Health Informatics

University of Washington

Abstract

Enhancing Secondary-use of Electronic Health Records for Geospatial-temporal Population Health Research

Jimmy Phuong

Chair of Supervisory Committee:

Sean D. Mooney

Department of Biomedical and Health Informatics

For almost three decades, the United States Department of Human and Health Services, Center for Disease Control, and the World Health Organization have recognized the role of social and environmental determinants of health in understanding the health of populations. Community and population health is a function of each individual's health and wellness, determined in large by their socioeconomic status, environmental factors, and access to healthcare services. In disastrous times, spatiotemporally-relevant information escalate in importance as health systems strive to address emergent concerns, pre-existing needs, population migration, while experiencing disruption in available resources and infrastructure. With their adoption by hospitals and health systems, Electronic Health Records (EHRs) contain a richness and diversity of information about patients. EHRs could inform where and how to prepare for population-scale

patient needs in future disaster scenarios with a timely, equitable, and data-driven approach; however, the ability to apply spatiotemporal reasoning with EHRs have remained an underrepresented capacity. Informatics innovations would need to account for the operational, technical, and ethical constraints felt by those who study the health of populations. In this dissertation, I focus on three areas for building capacities to use of geospatial-temporal information to address population health needs. The aims are to: 1) assess information needs and priority use-cases for population health research in hydrologic disaster preparedness, 2) design spatiotemporal use-case workflows to survey trends and anomalies for regional areas using gridded hydrometeorological data products, a surrogate for structured multivariate datasets, and 3) develop an approach for spatiotemporal inferential statistics of EHR patient diagnosis information. This work incorporates flexible design and secondary-use of data for population health research and geographic inferences in preparation for future disasters.

ACKNOWLEDGEMENTS

I humbly give my thanks to these many great people who have granted me warmth, wisdom, support, and mental support along this Odyssey. And coffee. Lots and lots of coffee.

First, I'd like to thank my committee members for taking a chance on me. A special thank you to my chair and advisor, Sean Mooney, for his guidance, access to resources, and trust in me to lead on multiple scientific fronts. Another special thank you to Christina Bandaragoda for being inspirational, resourceful, and daring to run towards hurricanes and real-world problems. I am so grateful to Andrea Hartzler for the guidance, structure, and clarity in how to communicate my questions and goals as a researcher. Finally, I am so thankful to have Elaine Faustman who brings a thrilling excitement to the room, encouragement, and pragmatism as my graduate student representative.

Next, I'd like to extend a special thank you to members of Mooney group: Vikas Pejaver, Steve Mooney, Tim Bergquist, Chethan Jjejevarapu, Christian Bock, Rishabh Jain, Don Smith, Abhi Pratap, Sicheng Song, Noah Hammarlund, Houda Benhlabib, and Yao Yan. Our lab meetings are flush with profound scientific discourse and your creative perspectives are diversely enriching as we each have a pursuit in a different problem spaces. Thank you to the staff members at eScience institute, the Institute for Translational Health Sciences, and the Univ of Washington Department of Civil Engineering have offered critical help and resources to address complex data science problems. I received invaluable advice from UW Biomedical and Health Informatics faculty and staff Neil Abernethy, John Gennari, Shawn Banta, Heidi Kruegger, Akiyo Kodera, Lora Brewsaugh, Jill Fillmore, and a number of alumni. I am so appreciative of you all. It does take a village to raise a good child.

I had an amazing opportunity to interview interdisciplinary researchers in population health. I would like to thank all my respondents who were enthusiastic and willing to help and the participants that took time away from their schedules for an interview with me. These individuals are truly incredible human-beings. They confided to me wisdom about human health, devastating disasters, and hope. Population health is visible and emotional, social and systematic, and shook by forces of nature and man-made burdens to each individual. I'm reminded of Maya Angelou's words: "*We are more alike, my friends, than we are unalike*." It's absolutely redeeming to believe that so many experts have the desire to improve health locally and across the world. A very special thank you to Kari Stephens for her guidance in designing the study, recruitment strategy, and advice in the scientific presentation. I am very grateful to Derek Fulwiler (Univ of Washington Population Health Initiative), Bryant Karras (Washington State Department of Health), Graciela Ramirez-Toro (Inter American University of Puerto Rico), and Patricia Ordonez (University of Puerto Rico - Rio Piedras) for their help in participant recruitment.

This PhD journey has truly been an odyssey. From day one, it has been one devastation after another, but I was fortunate to receive emotional support and enthusiasm from my fellow cohort members, classmates, and friends. For that, I'd like to thank Harkirat Sohi, Aakash Sur, Will Kearns, Jin Qu, Tressa Hood, Xiyao Wang, Esther Wu, Nicole Boyle, Claire Beveridge, and Beaker Hood. I'd like to specifically thank friends and colleagues within biomedical and health informatics Shefali Haldar, Ross Lordon, Sonali Mishra, Calvin Apodaca, Regina Casanova, Daehyun Lee, Nikhil Gopal, Nick Robison, Graham Kim, Ahmad Aljadaan, Abdul Alshammari, Maher Khelifi, Lucy Wang, Sean Mikles, Laura Kneele, and Yong Choi. These creative people have given me excellent advice and challenged me to writing sessions with them, to design

engaging research strategies, to think about what I want in life, to remember what is important to me, and to keep cool and find humor through the odyssey.

Lastly, I want to thank my friends and family as my motivators to reach for the PhD and a reminder to humbly give back. My mom, Annie, my dad, Peter, and my brothers Henry, Brandy, and Larry are emblems of resilience. A special thank you to my uncles, aunts, and cousins, and close confidants Garrett, Wes, David, Luoping, Lana, Alex, and Matt. These people showered love and coaching to work harder, think faster, attack aggressively, and get back up when knocked down. I can proudly say I am the first doctorate within my family because of this village.

Dedicated for mom, dad, and my brothers

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1 - Introduction

1.1 BACKGROUND

Demand for healthcare that is safe, timely, effective, efficient, equitable, and patient-centered (mentioned as the Institute of Medicine Six Aims) has spurred a societal transition towards medicine that is preventative and quality-oriented [1]. At the same time, population health has emerged as a field and perspective of health focused on the distribution of human health outcomes, the determinants of health, and policy interventions to promote improvements to the health of the population [2–4]. It seeks to understand whether there are issues that precede onset of disease, such as access to healthcare, capability to eat a healthy diet, ability to manage stressors, etc. as they relate to actionable goals. Determinants of health help to explain the burden of disease and likelihoods towards positive health outcomes within the population [2–4]. While healthcare systems, such as hospitals and clinics, regularly collect information to understand the patient experience of healthy and unhealthy patient populations, this information is captured within EHRs and researchers have had limited computational capabilities to interact with it. Although the United States (US) Health Information Technology for Economic and Clinical Health (HITECH) act of 2009 provisioned incentives to help US healthcare systems adopt EHRs, for almost a decade, health systems continue to work towards achieving equitable access, meaningful use of health information, and better quality information. Even so, EHRs have had perceptions by care providers of reduced productivity, dynamic data standards and quality, which make analyses with the collected data non-trivial [5]. While population health and public health research have historically acted independently of the information captured within EHRs, recent

cyberinfrastructure and virtual environment technologies have made it possible to engage in population-scale, multidisciplinary health research, provided that data interoperability and information standards may be harmonized for analysis.

The Center for Research on the Epidemiology of Disasters report that annual hydrometeorological natural disasters, such as hurricanes and floods, have intensified in frequency and severity of economic damage worldwide since 1950 [6]. Within the US during 2017 and 2018, category 4 and 5 Hurricanes Harvey, Irma, Maria, Florence, Michael, Lane, and Walaka have devastated the states of Texas, Florida, the territory of Puerto Rico, the US Virgin Islands, North Carolina, South Carolina, and Hawaii. Meanwhile, several states experienced the ravaging effects of extreme heat, wide-spread wild-fire, and drought. Several of these affected areas remain in a state of disaster recovery. Aside from physical infrastructure damage and estimated economic losses, these natural disasters afflicted physical and mental health burden upon the populations, where more research is available for acute phase compared to chronic human health outcomes. In some cases, health systems have gone offline resulting in data gaps. This is due in part to electricity outage and telecommunication failure but also to operational priorities to deliver care rather than spend time on documentation and data collection. Human resources may migrate away forming shortages in workers, technical skill sets, and subject matter expertise. Natural disasters are expected to become a grander challenge as climate change scenarios evolve. The issues are both local and global, and it depends on how well-prepared individuals, communities, health systems, and emergency services may respond.

The disaster risk research community have well-characterized models and protocols for response from lessons learned with past hydrologic natural disasters. However, these protocols often focus on short-term basic survival priorities and may ignore the diverse needs of the

population [7,8]. The recent research in disaster risk reduction highlight the need to consider not only response to wind and inundations hazards but all hazards for a measure of community resilience—the ability for communities to recover and become better through adverse situations like disasters be it natural, manmade, or biological [8,9]. Gridded hydrometeorological data products have provided the hydrometeorological community a means to consider simulations of meteorological and hydrologic effects on the landscape. Recent studies have combined geo-environmental and demographic surveys to yield gridded population data sets and new insight to enhance population-modeling resolution and accuracy [10,11]. It can be expected that gridded data products will continue to increase in abundance and complexity, and these data products may provide insights into geographic and environmental impacts of historical events. These integrative studies look at the influences upon infrastructure and resource dependencies for community resilience, which have timeliness and geographic relevance. In this era of internet-enabled technologies, it is unknown how well new information through cyberinfrastructure and analytics tools may improve disaster response, situation awareness, and community resilience research without impeding health delivery or deepening the reliance on electricity. Since 2015, the National Science Foundation (NSF), National Institutes of Health (NIH), National Library of Medicine (NLM), and National Institute of Environmental Health Sciences (NIEHS) have made informatics in disaster response research one of the top funding priorities [12,13]. So, despite the abundance of prior disaster risk research, multiple institutions recognize the broad-based need for formative operational improvements, user-centered design for unmet needs in disaster preparedness, and leveraging existing knowledge-bases for actionable population health uses.

In disaster preparedness scenarios and even into the immediate aftermath of a disaster, EHRs could potentially inform responders understand where and how to prepare for population-scale

patient needs [9,14,15]. The problem is that this information needs to quickly retrieved and health responders and researchers must perceive that it improves their operations and decision-making. Disaster can spur population migration, and in the absence of skilled workers or those that have the necessary capacities, skills must be retrained as needed. Methods such as spatiotemporal modeling, discovery of determinants of health, and secondary-use of EHR data for research are generally outside the scope of delivery of Medical care. Hence, without a means to build capacities for spatiotemporal analyses and data stewardship with health information, these personnel and expertise will be difficult to acquire.

Alongside the increasing demand and availability for information online, there exist emerging concerns about the ethical use and treatment of the patient information. The US Health Insurance Portability and Accountability Act (HIPAA) of 1996 prohibits the disclosure and/or publication of patient health information that may violate patient privacy and authority. This includes the potential for re-identification by geographic means. Hence, although there is a need for spatiotemporal analyses with health information, the practice must be upholding ethical and protective use criteria, take into consideration the social consequences of spatiotemporal findings, and acknowledge the limitations in spatiotemporal inference and representation from secondary-use EHR information.

1.2 DISSERTATION AIMS

This work includes three studies to address the following aims:

AIM 1: Assess qualitative information needs and use-cases for considering population health research in disaster preparedness.

I will advance the state of understanding about information needs and use-cases to improve population health research in hurricane and flood disasters.

AIM 2: Design workflows to survey trends and anomalies for regional areas using structured spatial-temporal information.

I will develop use-cases and workflows designed and implemented using distributed computing to enhance task efficiency and operated within cloud computing environments, analogous to HIPAA-aligned enclave virtual environments.

AIM 3: Develop and evaluate spatiotemporal inferential statistics through secondary-use of patient diagnoses within EHRs.

I incorporate modeling approaches from statistics and spatial epidemiology to make inferences of geographic and temporal trends in patient population healthcare needs using diagnosis information from EHRs.

1.3 DISSERTATION OVERVIEW

This work advances the state of knowledge about sociotechnical needs, barriers, and use-cases for population health research in hurricane and flood disasters. It advances the state of user interactions for research use of spatio-temporal data products and geographic analyses with electronic health records.

CHAPTER 2 - Information needs and use-cases to improve population health research in future hurricanes and floods: a research focus for disaster preparedness

2.1 INTRODUCTION

Historic hurricanes, like Katrina and Sandy, are remembered for the catastrophic flooding, devastation, and publicized limitations of the political response. However, the effects on the populace can be summed by the burden of disease, homelessness and displacement, and the anxiety of subsequent disasters [16,17]. From these past disasters, the public health community has adopted the population health framework to understand and promote the health of the population, relating the vulnerabilities to adverse health outcomes associated with social-environmental determinants of health and mitigating policies [9,16,17]. While various studies define a myriad of emergent hazards associated with hurricanes and floods [9,16,18,19], the major disasters during the 2017 hurricane season (July-September) revealed gaps in responding to emergent concerns and sharing of information for understanding population health. For almost two decades, research has called for refocusing disaster preparedness phase of an anticipated hurricane or flood (e.g., up to 10 days prior) to prioritize collecting multidisciplinary information and maintaining stakeholder engagement as the situation unfolds [6,8,12,14,20,21]; little is known about what information is meaningful to capture, at what scale, and the degree of complexity needed to support population health researchers.

In response to the 2017 hurricanes, post hoc data archives have proliferated to help document key factors before, during, and aftermath of hurricanes. These archives are developed during the aftermath to provide access to spatial-temporal information critical for forming exposure-

outcome relationships at different time frames of a hurricane (e.g., empirical observations of drinking water analyses, simulations of hurricane and hydrology, case-study reports of public health concerns) [22–24]. Aside from this retrospective view based on archived information, informatics opportunities with these archives for public health practitioners, population health analyses, and community preparedness for future events are largely unexplored [12,20,25–28]. To date, a dearth of research considers the information needs of diverse population health researchers and how information can be useful in the time-course of a hurricane.

In this chapter, I describe a study conducted while I was a Graduate Research Assistant within the Puerto Rico Water Studies group. Although I was the Graduate Student Lead on the study, I will use "we" to refer to efforts and decisions made by myself, the research team, and the collaborating co-authors.

2.2 OBJECTIVE

With the goal of informing the design of population health research tools in disasters preparedness, we conducted a needs assessment to information needs, barriers, and priority use-cases of population health researchers associated with disaster preparedness phase, prior to hurricanes and floods. We characterize expertise and readiness for disaster research to identify insights not usually leveraged within hurricane or flood disaster management, and share what tasks, questions, and information should be considered for designing future tools.

2.3 METHODS AND MATERIALS

2.3.1 Study design

We conducted a mixed-method needs assessment comprised of interviews and think-aloud card sorting. Interviews explored participant information needs and barriers in population health research and the participant's readiness and strategic approach for health research in future disasters. Card sorting provided participants a means to envision use-cases by grouping information cards, through which we learned about how they prioritize information for disaster preparedness. Study procedures were approved by the University of Washington Institutional Review Board.

2.3.2 Setting and participants

We define "population health researchers" as professionals who use health information, policy, and determinants of health data to study health outcomes for populations (or subgroups) that reside within geographic areas [14]. Email snowball-sampling was used to recruit from diverse work settings (i.e., academic, government, non-profit) and geographies affected (Territory of Puerto Rico) and not affected (Washington state) by 2017 hurricanes. Respondents were eligible if they were English-speaking and currently conduct health research about populations.

We conducted 15 individual sessions (14 in-person and one video conference) lasting 80-minutes on average (range: 40min to 4hrs) consisting of two parts, a semi-structured interview then card sorting. Card-sorts for three sessions were conducted electronically using the

OptimalSort® web-interface while the remaining 12 sessions used physical cards. Sessions were audio-recorded and transcribed using Temi® then coded with Dedoose®. Data collection was completed once thematic saturation was achieved [29,30].

2.3.3 Interviews

The semi-structured interview guide (Appendix 1-1) focused on four key constructs (Table 2-1), based on the major categories of challenges for conducting disaster research [12]. Participant described their research role and expertise (construct 1) and shared their attitudes towards two socio-technical aspects critical to disaster management: collaborative research (construct 2) and data and technology adoption (construct 3) [12,20,21]. Finally, participants described their readiness and strategic approach for population health research when anticipating hurricanes and floods up to 10 days in the future (construct 4). Interviewer memos were used to verify transcripts for accuracy.

Table 2-1: Summary of constructs within the interview guide

Key constructs	Concepts for capture	Example questions
1. Research role and expertise	Research training, subject matter expertise, analytic methods, current occupational setting, context about prior use of spatial-temporal data (i.e., temporal scale, geospatial scale, place of focus)	What is your research role, and what are your key components to conducting research?
2. Collaborative research	Positive and negative attitudes towards prior research collaboration	What are some positive and negative experiences you've had with collaborations in your research role?
3. Data and technology adoption	Positive and negative attitudes towards prior experience with data sets, software, and technology adoption experiences	What are some positive and negative experiences you've had with data and technology adoption in your research role?
4. Readiness and strategies for future disaster research	Readiness for a role within future disasters, prior experiences with disaster management, strategic approach and envisioned needs and barriers	What approaches would you take to research the population health impacts, if you could anticipate a hurricane or flood to occur in the next 10 days?

2.3.4 Card-sort

After the interview, we conducted a closed card-sort [31] using 31 hurricane- and flood-related information cards (Appendix 2-2). Components of a card are shown in Figure 2-1. Card labels represent keywords intended for indexing the Hurricane Maria post-hoc data archive [24] and include concepts from geographic indicators of community resilience [10,17,20,21], health research from prior hurricanes and floods [9,15,16,18,19,32] and known disaster research concerns [12,14]. Participants were asked to envision themselves as the key actor for population health research within the disaster preparedness phase of an anticipated hurricane or flood up to

10 days in the future. Participants were told to think-aloud as they rated each card by what they perceived as useful for their research into population health into usefulness categories (i.e., "useful", "maybe useful", "not useful") and then grouped cards rated as "useful" and "maybe useful" according to their order of priority for addressing the anticipated theoretical event. Finally, we asked participants to nominate information not already represented by any card that should be added. Photos of the completed physical card-sorts and OptimalSort® card-sort reports were used to verify transcripts for accuracy.

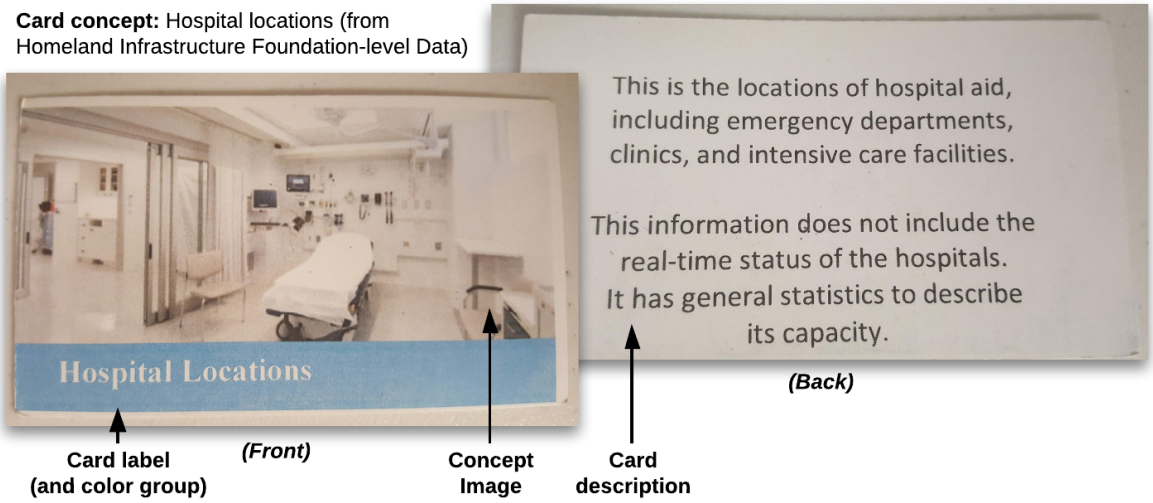


Figure 2-1: Sample information card.

Cards were assembled into 6 color-coded groups for communication purposes: spatial resource information (blue), logistics and context information (purple), training and decision support (gray), environmental quality and testing (yellow), demographics and health outcomes (green), and geospatial shape visualizations (orange).

2.3.6 Analysis of interviews

We conducted a template-based thematic analysis [33] of interviews to identify themes for each construct, described as the concepts for capture (Table 2-1). Barriers and facilitators were coded respectively with negative and positive attitudes towards prior experiences (construct 2-3). For construct 4, participant readiness was assigned one of three codes: 1) self-identified as planned and prepared for a role in a future disaster, 2) currently not prepared but had some experience with disaster management, or 3) no experience with disaster management and expressed need for more knowledgeable collaborators. Finally, we examine participant strategies for population health research in an anticipated hurricane or flood disaster (construct 4) across sessions to highlight similarities in operations and approaches to barriers and facilitators.

The two coders achieved a moderate inter-coder Kappa agreement of 0.67 on a 10% random sample of double coded interview excerpts, followed by discussions to resolve disagreements in code definitions and interpretations until consensus was reached [34,35].

2.3.7 Analysis of card-sorts

I estimated perceived usefulness scores for each card [36], then analyzed for groups of cards that received similar scores. Here, for each card, perceived usefulness scores were calculated as three percentages for each rating (i.e. “useful”, “maybe useful”, “not useful”), where each percentage is the number of participants of that rating divided by the total participants that rated the card. Pearson’s correlation assessed the pairwise similarities between cards based on variations in perceived usefulness scores. Hierarchical agglomerative clustering with Euclidean distance and Ward’s minimum variance method [37] detected clusters based on Pearson’s

correlations. The optimal *k* clusters is defined using the "elbow" method, where additional clusters after *k* clusters contributes a minimal reduction to the explained inertia [38].

A use-case is a goal addressed with one or more types of information identified through qualitative thematic analysis of card-sorts. To identify priority use-cases, use-cases were ranked by the number of participants that discussed it. For each use-case, we calculated the fraction of card associations, calculated as the number of participants that associated each card with the use-case divided by the total participants that discussed the use-case. We conducted a preliminary thematic analysis [33] to develop a codebook consisting of card labels, nominated information names, use-cases and definitions, and comments about the use-case or information card. A second coder (SH) independently examined one full-length card-sort transcript to identify excerpts and apply the codebook. The two coders then discussed code definitions and reconciled discrepancies until consensus was reached. Each participant card-sort was summarized into a sparse matrix that applies the qualitative use-case themes to inform quantitative analyses. Each information card or nominated information (row) has annotations for the participant rating (categorical) and comments about the card (free-text). Each use-case discussed (column) has attributes for the comments about the use-case (free-text) and given a score of 1 for each card associated at least once with the use-case (numeric). Matrices were constructed as spreadsheets (.xlsx) then analyzed with Python v3.6.

2.4 RESULTS

The following sections describe findings among participant characteristics (construct 1), readiness and strategies for future disasters (construct 4), the barriers and facilitators (construct 2-3), clusters observed among the perceived information usefulness and use-cases.

2.4.1 Participant characteristics

Between June 2018 to December 2018, we emailed at least 43 researchers to participate in the study. 15 eligible participants (P01-P20) enrolled. Participants resided in the Territory of Puerto Rico (n=2) or Washington State (n=13), and primarily worked in academia (n=10) or government agencies (n=5).

Participants self-identified with a total of 13 researcher types and described 22 subject matters of study as part of their work (Appendix 2-3). On average, participants self-identified with two researcher types (range: 1-4), where the top 4 include epidemiologist, environmental health researcher, global health researcher, and emergency management researcher. On average, participants described researching six subject matters (range: 1-12) within their career, where the top 4 included all-hazards emergency management, exposure hazard agents, health outcomes related to disasters, and disease surveillance systems. Participants described methods for primary data analysis (n=5), secondary analysis (n=4), both primary and secondary (n=2), or was not described (n=4). Contexts of research focus varied broadly but often having one or more temporal scales (n=11), geospatial scales (n=8), and/or places of focus (n=9).

2.4.2 Readiness for future hurricanes and floods

Participants who self-identified as ‘planned and prepared’ for a role in a future disaster, including five government staff and one academician (40%; n=6). Some participants had experience with disaster management but are currently not prepared (47%, n=7), while the remaining participants had no experiences with disaster management and expressed need for

more knowledgeable collaborators (13%, n=2). Figure 2-2 indicates that amongst participants who self-identified as 'planned and prepared' the most frequent subject matter expertise included: Disease surveillance systems, Foodborne and communicable diseases, All-hazards emergency management, and Exposure hazard agents. Participants with other skilled capacities in population health require additional efforts to be prepared for future disaster research.

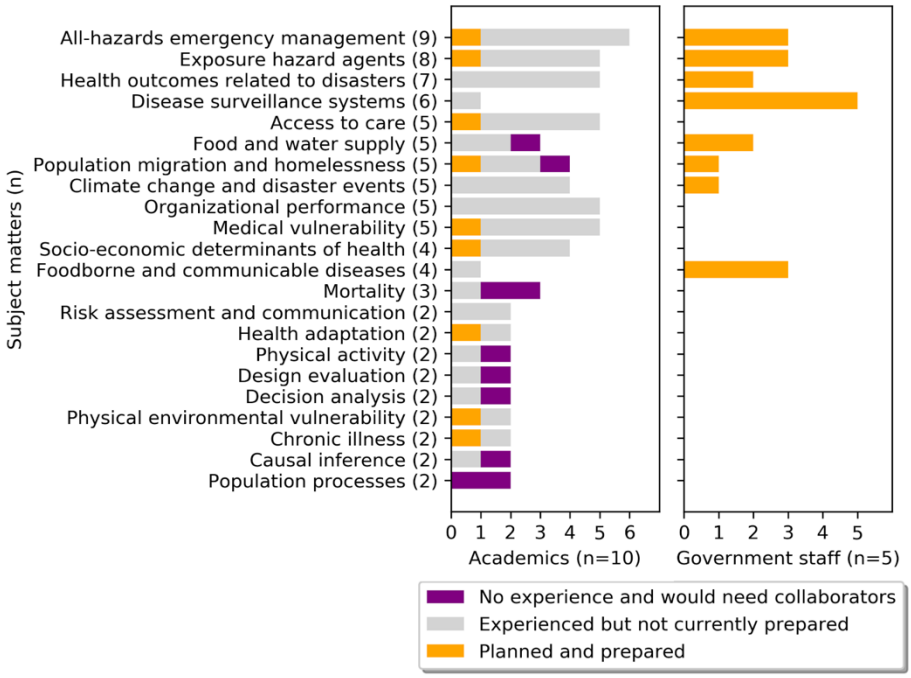


Figure 2-2: Participant subject matter expertise.
Stratified by current occupational setting and readiness categories, the distribution displays the subject matters that are currently part of disaster management by government staff versus areas of research by academics.

2.4.3 Current strategies for future disaster research

The following notable trends were observed between participants with versus without plans and preparations for disaster management, part of their strategy to consider population health in an anticipated hurricane or flood (up to 10 days in the near future).

For participants with plans and preparations, their strategy includes their agency emergency management plan and the Incident command system. Participants described 1) their role within a team, 2) their team contributes certain capacities within the All-hazards approach, 3) status reporting of situational information, and 4) the coordination with local health jurisdiction and community collaborators. Participants were affiliated with three types of teams: Emergency medical services, Situation awareness teams, and Reconnaissance teams (e.g., Environmental health strike team, Epidemiology response team), where reconnaissance teams focus on applied research information related to population health and hazards. While some teams are activated continuously throughout disasters, reconnaissance teams are invited on short rotations (e.g., 3-weeks) for particular capacities typically during the disaster recovery phase. Invited teams rely on Unified command and Situational awareness teams for daily "*report[s] that characterizes the situation by zone*" *(P20)* and collaborations with local responders and community members to gain insights on community status, understand local nuances, and build human resources and capacities. As such, collaboration between teams, local health officials, and civilian points-of-contact are critical mechanisms within the emergency management plans.

By contrast, participants without plans for disaster management primarily sought to form new collaborations in support of disaster recovery. Various administrative barriers precede research in disaster settings and participant had different approaches to navigate them. Some identify interested collaborators to propose solutions for unmet needs declared by officials and funders; others take a community-guided approach to concurrently identify unmet needs and appeal to funders; others look to join collaborators to refine research questions. As new information arrives, this process iterates into attack plans with defined collaborator roles (e.g., community stakeholders, data collector, analytics, point-of-contacts with local familiarity). There

is agreement that these process limitations are necessary precautions to build relationships, considering information and methods for reliability and usability, appraisal of ethical practice, obtaining authorizations, and coordination to avoid impeding recovery efforts or concurrent projects. The following questions were recurrent among these participants as prerequisites to project planning:

- Who is available for research collaborations? Community collaborators?
- Who's in-charge of what and where?
- What are the major health vulnerabilities in the area?
- What are the communication needs with citizens in the region?
- Who has access to good quality data or new data to be collected?
- What are the existing research efforts and questions?

2.4.4 Barriers and facilitators of population health researchers

Seven barriers and six facilitators emerged from interviews (Figure 2-3). Representative quotes are provided within Table 2-2, 2-3, and 2-4.

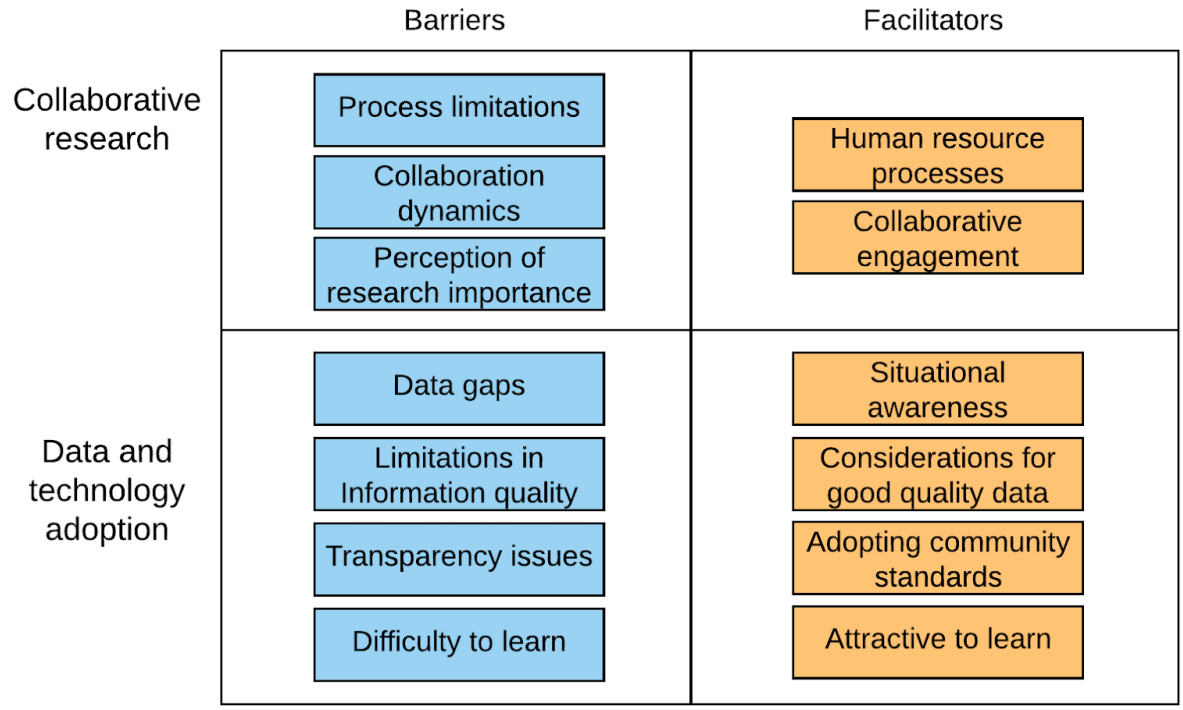


Figure 2-3: Barriers and facilitators of Collaborative research and Data and technology adoption.

2.4.5 Barriers to collaborative research

Process limitations (noted by 12 participants) are the burdens of rate-limiting logistics, observed here in seven subthemes (Table 2-2). *People migrating away* due to evacuation or driven by the *"economic situation" (P01)* emerged as a noteworthy finding. Having fewer people than anticipated exacerbates collaborator searches and recruitment of *"populations willing to be studied" (P5)*. Though most process limitations are consistent with prior research [12,20], concerns of population size and capabilities confer barriers to population-based research projects and resource allocation decisions but are seldom mentioned.

Collaboration dynamics (noted by 5 participants) may decline as tensions between collaborators escalate. Collaborators who are *"too confident that their discipline is the solution"*

(P08) can create rifts in acknowledgement and interpersonal interactions. Territorial behavior was referenced when collaborators run *"out of time" (P06)* to complete their contribution leading into subsequent problems aimed to *"control people's time" (P05)*. These burdensome experiences present challenges in maintaining collaborations, stemming from inflexible time allocations and communication of priorities without recognizing the differences in values.

Perception of research importance (noted by 5 participants) relates to the public view of what is disaster preparedness apart from improvisation. Participants felt the public relate disasters with first responders more so than with public health efforts in pre-disaster planning and post-disaster recovery. There was a distinction between flexibility and improvisation. Flexibility is still operating within the guidance of a plan. Improvisation, seen as operating without plans or straying from plans, fragmented coordinated efforts like *"developing strategies in the wrong phase [of disaster management]" (P02)*. The desire to improvise may be based on misconceptions about the vulnerabilities and the consequences of error. These perceptions influence the preparations that are performed, but they can also be divisive between collaborators and coordinated efforts.

Table 2-2: Barriers and facilitators of collaborative research within population health in disasters

Subthemes in barriers	*Key representative quote*
Process limitations	
Interest and momentum	*"[…] these relationships take a while to build, and to find the right people to work with." (P13)*
Funding and capacity	*"[...] how are we going to be sustainable [...] And, I mean sustainable in all of the sense of the word. Economically in terms of resources, personnel, and as an administration." (P02)*
Duplicate efforts	*"we work on the same things [...] just very narrow lane differences." (P18)*
People migrate away	*"with the economic situation [...], I wouldn't be surprised if some of these people [new graduates] just migrated elsewhere." (P01)*
Study recruitment	*"populations willing to be studied." (P05)*
Gap in updates	*"nobody knows what's going on in other parts of the department." (P16)*
Time-sensitive events	*"[...] once an event happens, it's like a little late [...] it would be great, you know, if this is the preparedness side of things, like if we were better prepared by sharing information before there's an event." (P06)*
Collaboration dynamics	
Collaborator coordination	*"[…] we recruited a community partner, and they ended up not implementing the thing that we had wanted them to implement […] they were like "we don't have time, we don't have time, we didn't have time" and then they just -- you know, we ran out of time." (P06)*
Territorial authority	*"[…] there were sort of territorial constraints over who got to control people's time and allocating work. Um, and the human-level interactions got ugly as a result." (P05)*
Perception of research importance	
Culture of improvisation	*"[…] they were trying to go directly to the recovery phase without going to the response phase. The response phase is always immediate need. [...] And, they were talking about developing strategies for the long-term. No, sir. […] We are now in the response phase, and people need essential services to survive. […] most of the agencies were trying to develop strategies in the wrong phase." (P02)*
Relatability	*"So, the boys in blue do response. Police, fire, EMTs -- that's your normal response. […] the part that's unsung is recovery. So public health, in my mind and some of my counterparts' minds, is we're the day after people. […] you know, that's where it gets to be difficult, because that's where we see a difference in what we do and versus what first responders [do]." (P18)*
Subthemes in facilitators	
Collaborative engagement	
Reputation	*"I think reputation of team is important. So, if they've success--successfully pulled this kind of thing off before, that's--that adds to my confidence and probably, for better and for worse, diminishes the amount of time I would spend checking up on them somewhere else." (P05)*
Trust	*[...] I think if the state had just done that in a vacuum, I don't think the results would have been as well received by other local health jurisdiction. So by having a peer actually do that more -- that kind of study, it validated and gave maybe a little bit more credence I think with local decision makers. (P18)*
Respect each other's lanes	*'[...] it's only when they have exhausted their resources that they call us at this State, because we're a home rule state. So, they operate with impunity in their own health districts, and we don't get involved unless they're like, "hey, we're --we're totally overwhelmed. Come help us."' (P18)*
Focus on a mutual goal	*"[…] they were kind of relating their personal experiences with the goals of the project." (P13)*
Know when to conduct independent versus collaborative research	*"I know my limits. When I find my limitations, I'm like 'wait a minute, I need someone.'" (P01)*
Support from non-profits and NGOs	*"[…] there is a community-based organization that is called COSAU […] they identified those needs for their communities [before the hurricane]." (P02)*
Soliciting feedback	*"[...] you have to have both a sort of an appreciation of the other person's perspective, but also an appreciation of their approach. Their methodological training that's relevant for them, how they think about the problem and -- and, respect. I mean, it requires a great deal of mutual respect." (P08)*
Human resource processes	
Prospective research partnerships	*"[...] you'd have to plan in advance for the data that you're going to need and make sure that you've got the partnerships in place so that you can get the access to the data and do the analysis quickly to get a sense of how big were the health consequences and where were the biggest health consequences." (P11)*
Access to data collectors	*"[…] so we have a lot of plans, but a lot of it has to be exercised. And that's where you really are only going to find the gaps. So, the source of the data would be important and making sure that can be sustainable." (P20)*
Local familiarity and context	*"[...] turns out it comes down to documentation practices at that facility, and then it comes down to specific nurses and specific providers who are documenting not in the triage notes field that we collect, but in a different field." (P16)*
Identifying missing skill sets	*"We are concerned that people are still going out in these disciplinary teams and they're not incorporating public health or behavioral health or behavioral sciences or social sciences." (P08)*
Develop staff capacities	*"[…] it would be interesting for all physicians, as part of their annual continued medical education, to have one hour or two hours on water quality and how those parameters are associated to health. […] We're not taught about that in medical school." (P01)*
Access to subject matter expertise	*"So, we have collaborations on opioid overdose. [...] sexual and intimate partner violence [...] injury and violence prevention. [...] Older adult falls [...] They have program specialists who do this stuff. So, the subject matter expertise to help us surveil for these conditions and interpret our findings. Huge. We rely on collaborations within the department and there are subject matter experts for that." (P16)*
Organizational infrastructure	*"[...] oftentimes, a local health jurisdiction will go into what they call incident command and activate their incident management team, and that's when you have at least an organized body to fulfill the work that is required for the outbreak or the incident." (P20)*

2.4.6 Facilitators to collaborative research

Human resource processes (noted by 14 participants) facilitate building the teams with the right traits. Once the nature of a disaster is known, collaborators are deductively searched for based on skills and capacities: technical decision-making, field methods, subject matter expertise, access to data, analytical skills, and/or community familiarity. To address mid-operation staff shortages, disaster management agencies use organizational infrastructure to find teams for rotations, while reconnaissance teams seek to build community capacities, referring to *"students" (P07)* and "*biology majors*" *(P18)* as potential community collaborators with prerequisite training. Despite some reference to organizational infrastructures, academic participants often found collaborators informally or via social network platforms like LinkedIn. Mechanisms to find subject matter experts and trainable personnel have been beneficial but the information in such resources need to be maintained up-to-date.

Collaborative engagement (noted by 14 participants) was crucial to fostering working relationships pro-actively built on trust, respect, autonomy, and flexibility. Reputation can give *"credence" (P18)* for peer adoption of the product when it models how the community or decision-makers might approach the problem. Alternatively, reputation of having "*successfully pulled this kind of thing off before*" *(P05)* shapes trust and confidence in the collaborator and their quality of work. With metaphors like mutual "*exchange" (P04)* or *"dance" (P20)*, participants assert that collaborators must know when to conduct independent research versus work collaboratively, how to respect each other's lanes (or boundaries), and how to receive and solicit feedback that shows appreciation for *"how they think about the problem" (P08)*. The intent with getting leadership buy-in and building rapport should be to achieve indicators of

"*deep collaboration*" *(P06)*. Understanding how collaborators use these interactions enables more meaningful developments and supportive experiences.

2.4.7 Barriers to data and technology adoption

During disasters, electrical and telecommunication infrastructure can be unreliable and impact health systems and data collection systems downstream. However, for understanding health in the population, many of the barriers with data products and technology platforms start during normal circumstances as problems by design or non-alignment with research needs. These barriers separate into four themes (Table 2-3).

Data gaps (noted by 14 participants) diminish the cross-sectional representations of reality. To understand health outcomes, participants often leveraged platforms adopted by their stakeholders, sustaining anomalies from incompatible data standards and legacy systems. During normal circumstances, data gaps often refer to *"incomplete entries" (P06)* or when systems *"stopped acquiring data" (P02)*. Disease surveillance systems often look for the *"plummet" (P17)*, indications of localized data flow anomalies. During disasters and power loss, medical services strive to continue, so documentation take second priority and may revert to paper records at-best. There was agreement that exposures (e.g., contaminant type, concentration, and temporality), mental health outcomes, and disruption in care remain gaps without sufficient baseline characterizations to study causality from disasters. Existing efforts may close the gap on environmental baselines through *"ephemeral data collection" (P19)* and social endpoints using population-based sampling strategies, but many endpoints remain uncertain and new gaps may present over time.

Table 2-3: Barriers in data and technology adoption for population health research in disasters

Subthemes in barriers	Key representative quote
Data gaps	
Different standards	*"One problem is that sometimes Puerto Rico is excluded from data generated for the United States, because it's focused on the states." (P01)*
Concurrent circumstances	*"[...] 'yeah, this is definitely giardia'. uh, "this is not." [...] "they missed three – three weeks of their metformin and now they're back on it." Those are the kinds of things that the emergency responses that doesn't like separate or disaggregate for us." (P18)*
Data capture is second priority	*"[...] if you lose power for three months or more, you have to make sure that these [hospital] systems are -- these [hospital] systems keep going [...] I would expect healthcare personnel to focus on healthcare delivery, rather than on recording data." (P01)*
Data for status characterization	*"one of the things that came out of this forum is we have to really step up guidance to local people on the ground to collect baseline data, so we know the before. And we know what it needs to get back to right after a disaster. [...] biodiversity, a density of product, in some cases the baseline chemical analysis to, you know, how -- what baseline contamination is." (P19)*
Collecting exposure information	*"Part of the challenge is figuring out actually who was exposed to what in a disaster. So, you've got challenges there, because you don't know where people were, necessarily. [...] maybe you could even model how much they were exposed to certain air pollutants, but you still wouldn't actually know what was in it." (P12)*
Important endpoints not collected	*"So, say somebody went to some, um, volunteer health station and got treatment for something. That's probably not recorded, it's probably not captured anywhere, so you won't get that information." (P12)*
Incompleteness	*"Those all have problems with missingness that are frustrating." (P06)*
Sensitive to disruption	*"[...] connectivity or web connections is almost unheard of. [...] my cell phone got service like when I was on a hill and the wind was blowing in the right direction. [...] That is very common after a major disaster, like a hurricane." (P18)*
Difficulty to learn	
Relearning	*"[...] when the outbreak of leptospirosis occurred in Puerto Rico [...] they didn't know how to deal with the paper work or how to collect the data after -- days after the hurricane without a system, even if it was considered maybe in the plan." (P02)*
Poor design and usability	*"[...] there was links to the same map from different websites, but some of them worked and some of them didn't. Or, they use some sort of platform [...] It was kind of frustrating. Some of them never would load." (P17)*
Adopting prerequisite systems	*"[...] until December, almost 40 percent of the population lack of electricity and [while] the data is available on the internet. [...] And, we store everything in the cloud. [...] it's becoming one of the essential services to continue normal operation conditions." (P02)*
Siloed by design	*"[...] It was also, you know, in a totally separate application. So, I had to kind of like impute [laugh] from this little map over here. And figure out where to look at." (P17)*
Subject matter literacy	*"If you're an expert, -- you can say, 'well, these are toxic because they belong to this category thing and they had this type of effect. [...] But -- So, that requires a great deal of expertise to use that. [...] unless you dig in, you sometimes can't tell which is which." (P08)*
Limitations in information quality	
Clarity about sampling representation	*"[...] one thing you don't have that you always really need is good information. We're almost always finding ourselves in the position of having to make decisions based on limited -- Know -- knowingly inaccurate or absent information a lot." (P16)*
Challenges with quality and validity assurance	*"[...] they require a correction factor. And, it's unknown if the correction factor works in really bad events [...]" (P13)*
Geographic relevance	*"[...] it is not geo fenced. So, we can't get data just for Houston or we can't get data just for Santa Rosa. We get data for everyone and we don't know where they're from, so that wasn't very helpful." (P13)*
Inconsistent display	*"[...] there's a policy that says that you cannot publish rates where the numerator is 9 or lower. [...] The problem was that that warning was not placed at the website where I took the information from." (P01)*
Low resolution and sample size	*"There's lots of wood, wood stoves, and burning, and agricultural burning that causes bad air quality. [...] um--but, they're only giving one reading for the whole valley or two readings for the whole valley." (P13)*
Transparency issues	
Access to data	*"[...] there's not a lot of ways to track that data because it goes into the emergency response realm. It doesn't come back to the public health realm." (P18)*
Assurance of ethical use	*"[...] even right now, you can't publish data with information about health conditions that have small counts, where people could be identified." (P12)*
Lag	*"[...] we do have private entities and there can be issues with accessing data in a timely way. So, I think that access to data eventually sometimes happens, but when you need it the most is, you know." (P20)*

Limitations in information quality (noted by 14 participants) reduces the trustworthiness and usability of the information. Early into a crisis, situational information may be scarce and *"knowingly inaccurate" (P16).* While participants err on precautionary principles, there a consensus to avoid making data-driven claims without knowledge of the sampling validity and reliability, referring to positive predictive value for true events. Although geographic maps are important communication tools, maps are ineffective with information stripped of place references, pictures without coordinate references, and where it is unclear if a pattern exists among raw data. Two participants encountered inconsistent displays of data attribution and data-use policies at data access portals, highlighting interface design flaws that are obstructive to research data users. More research is needed to understand factors that limit usability and opportunities to improve measurement validity.

Transparency issues (noted by 12 participants) were characterized by concerns of ethical practice and timely access to data. Here, access to data was specific to data that should be accessible but are absent, not timely, or improperly handled. Participants were unclear where to find up-to-date information about vulnerable populations when evacuations occurred. Populations may seek help from urgent care tents and places of refuge, but it was unclear if situational information were collected into organizational *"silos" (P18)* that are unprepared to share at the time of need. Some participants discussed location registries based on medical status, but there were disagreements about registries in the ethical impacts on patient privacy, the difference in timeliness and accuracy compared with disease surveillance systems, and perpetuating biases to populations with continued access to care. These barriers affect what researchers decide as pursuable research questions and how research is conducted.

Difficulty to learn (noted by 11 participants) was prominent in efforts to learn or change workflows. Changes in workflow prompts relearning, where moderate examples are software updates but extreme examples are the loss of electronic- or internet-based tools. Adopting new tools and workflows can have a steep "*learning curve*" *(P01)* for the terminologies, tasks, decision processes that be recognized as *"not intuitive" (P11),* and the platform dependencies. Collaborations with communities often need to consider the diverse personal abilities, communication strategies, and overall sustainability of the proposed strategy. Further research should examine how drastic changes in workflows can subject users and teams efforts to process relearning.

2.4.8 Facilitators of data and technology adoption

Given the hurdles in adopting new data products and technologies, four themes arose as redeeming factors for considering adoption (Table 2-4). Incorporating these themes into tools and strategic design ensures pragmatic appeal to population health and disaster researcher workflows and settings.

Situational awareness (noted by 13 participants) was described as understanding a situation through the relationship between contextual factors. Participants responded positively to data visualization on geographic maps, referring to "*flooding dashboards*" *(P17)*, community discussions facilitated with "*scenarios of climate change" (P02)*, and "*future urbanization plans*" *(P11)* as existing communication strategies for action that adopted maps. Participants also preferred obtaining situation information through collaborators with authoritative knowledge, reliable key informants, and "*free text" (P20)* reports from credible sources, but expressed concerns about summarize raw text or images without loss of interpretability.

Table 2-4: Facilitators of data and technology adoption for population health research in disasters

Subthemes in facilitators	Key representative quotes
Situational awareness	
Tools for visual communication	*"[...] we needed some way of showing "awareness" and "activation" in different parts of the forecasts and warning system. [...] It's very hard to show something this complex in a way that you can get an overview of. And so we wanted a graphic representation that will still give people some idea of the content " (P08)*
Relevance to understand contextual factors	*"We'll track how many mosquitoes we have detective West Nile Virus in. Then, we'll look at how many birds, and we'll count the horses, and then that gives an idea of the probability of a person being exposed already getting it." (P18)*
Relevance for priority action	*"we're usually looking for acute impacts. […] E. coli, Salmonella or something, you know, like something that's going to have a short term impact, we might monitor for an outbreak in patients residing in those areas." (P17)*
Awareness through reliable sources	*"With local health jurisdictions, they know what's going on in their communities in a way that we never could. They make observations all the time that we wouldn't or couldn't with the technical experts all around our agency [...]" (P16)*
Considerations for good quality data	
Validity and trustworthiness	*"[…] it's highly structured. It's hand entered by a registrar. Like it's, it's a highly reliable, robust dataset." (P16)*
Have a designated analyst to assure data quality	*"These studies always include a data analyst who's at least half time on the project, whose job is to ensure that the data is coming in properly based on the software that we're using for defining that collection." (P04)*
Preference for highest resolution possible	*"[…] when you're talking about weather and health, you've got more data availability for weather data and it's more granular. [...] The challenge is usually finding health data at that granular level. So, the health data tends to determine the scale at which you can do your analysis. [...]" (P12)*
Adopting community standards	
Use validated toolkits	*"probably the most popular one is CASPER [...] It's published by the CDC. It's been well validated in a variety of different disasters." (P13)*
Avoid data silos	*"[…] there's some really cool things already that integrates across geotechnical and structural data, but they don't integrate social science data yet. So we're trying for that." (P08)*
Start with open access data	*"[…] for some kinds of research [questions] you need data that is identified but for many other things you may just need de-identified data." (P01)*
Attractive to learn	
Flexible for different projects	*"[...] has its own disaster response research protocol that also collects exposure data. […] things like air quality or water quality or soil quality, hazardous material exposure, and then also some health information. You can customize the tools." (P13)*
Having control and process automation	*"With epiR, you can adjust 20-, 30-, whatever number of rates by age, which used to be a real hassle. Once you learn how to do that, you just have a script ready." (P01)*
Direct communication with data experts	*"[...] they have one staff person specifically who will engage with you in like trying to do something in R with the data. And, he'll like share with you his approaches to using R. And, like, visualizations, where he's done some text-based analysis, um, and share his code and how he's using it." (P17)*
Easy to find help trouble shooting	*"[…] if I have a problem, I go to the web. There's so much on R that I can usually solve my issues just by looking." (P01)*
Offline capacity	*"[...] they bring this kit that is -- allows you to -- to do incubation for fecal enterococci and coliforms In this in situ system. [...] it was very useful for communities to know if it was safe for them to use that water [...] Most of the labs were damaged [...]" (P02)*
Learning prerequisites	*"Once you work with this, you're better prepared to get involved with specific packages." (P01)*
Easy to learn for communication	*"[…] 'No, you shouldn't eat shellfish at this timeframe', etc. Etc. So, some of that's pretty clear […] it's designed to be able to warn the public […]" (P18)*

Considerations for good quality data (noted by 8 participants) was seen as objective proof that the data collection method has best-practices to ensure validity. Data sets are created fit-for-purpose and participants looked for indicators that the records are *"robust" (P16)* for replication. Three participants assert that collaborative projects should have analysts designated for quality control, someone capable of appraising methods and determining data engineered are in a usable state. Heuristics like *"triangulation" (P05)* show agreement across methodologies to the extent of each data set's limitations. Participants also noted hopeful sentiments about real-time technologies in generating high-resolution data sets that better represent situational contexts in time and space.

Adopting community standards (noted by 7 participants) was described as aligning with community best-practices to avoid duplicating efforts. There were concerns that research tools were *"recreated all the time" (P06)*, recapitulations without superior design to current tools. Participants recommended validated toolkits like Community Assessment for Public Health Emergency Response (CASPER) and Disasters Response Research (DR2) surveys; these tools were designed with a field sampling strategy, multiple languages, a focus on health effects, portable dashboards, and collective inputs about human factors engineering from uses in past disasters. If such conditions exist, leveraging community standards minimizes efforts to learn workflows that may not transfer.

Attractive to learn (noted by 9 participants) represents the value proposition for users to gain desirable traits. There was a consensus desire that efforts taken to learn technologies have high promise in return on investments, like flexible uses and improvements on steps that were *"real hassles" (P01)*. Participants responded positively about direct access to experts, emphasized as a

necessary collaboration through customer service to reach non-early adopters and users who have low bandwidth for elective learning. Separately, few tools remain adaptable without electricity but, like portable water testing kits, those that do provide flexible means for situational awareness. Participants also referred to CASPER and DR2 surveys as offline tools, where data are collected on *"analogous paper systems" (P18)* then input and analyzed on local dashboards that have passive syncing capabilities. These features appeal to broader groups of users and help motivate researchers to commit into adopting tools.

2.4.9 Perceived information usefulness

Figure 2-4A summarizes the 31 information card labels ranked from left-to-right by least "not useful" ratings. No information card was unanimously rated in any way, though *Time constraint annual* received equal ratings across the usefulness categories. The cards separate into six clusters based on correlations in perceived usefulness scores (Figure 2-4B), where cluster1 contains the 13 most useful cards whereas clusters 5 and 6 includes the 5 least useful cards.

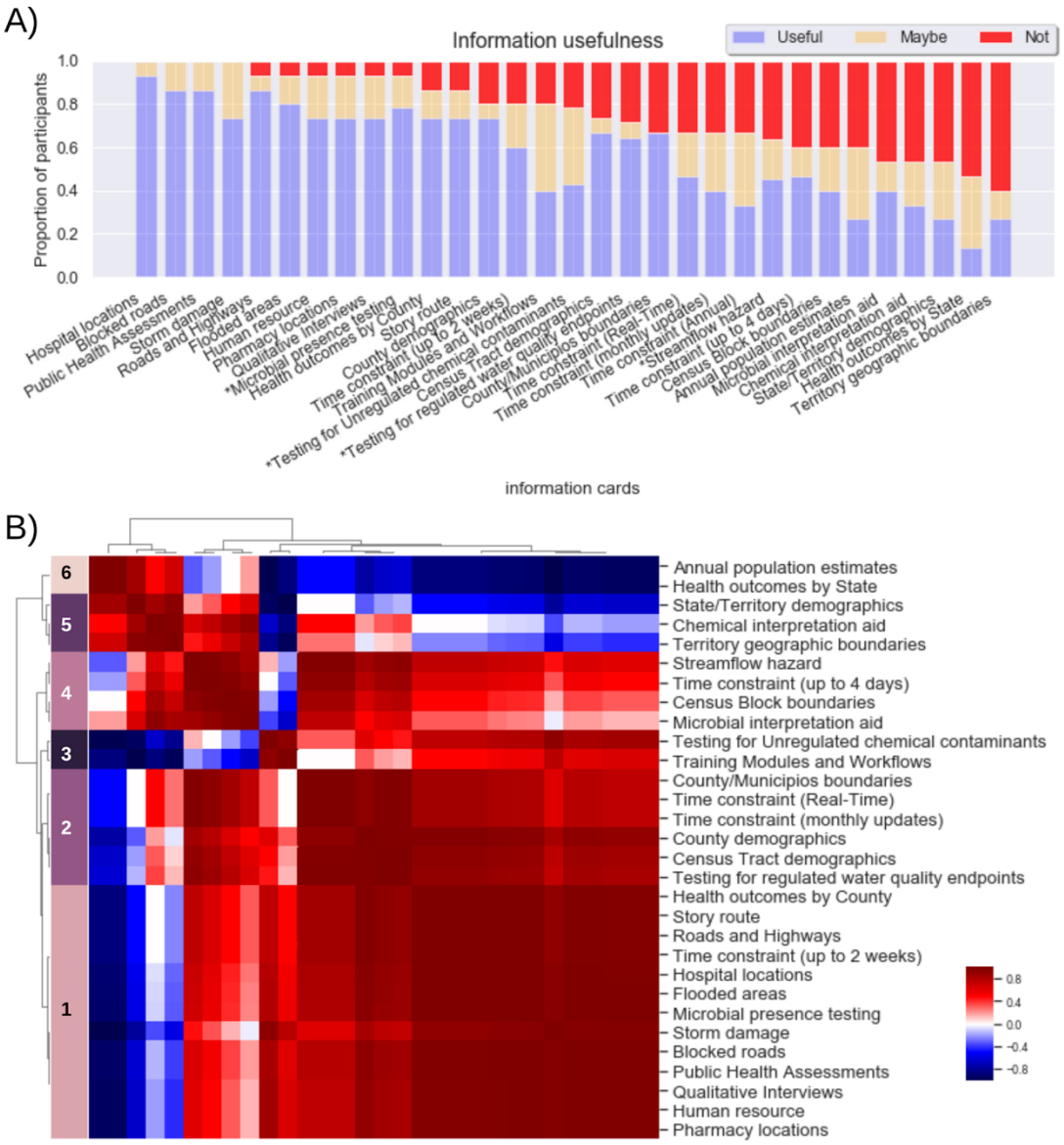


Figure 2-4: Perceived usefulness scores and card clusters of similar usefulness ratings.

A) Usefulness of information cards and B) Clusters and correlation between information cards by perceived usefulness. Cards with asterisks(*) were sorted by at least 11 of the 15 participants. Time constraint (annual) was excluded from correlation analysis due to lack of difference between usefulness categories.

2.5.10 Information use-cases

Table 2-5 reports the top 15 use-cases discussed by five or more participants, out of 79 use-cases detected. Major pre-disaster characterizations focused on pre-existing medical needs of vulnerable populations (UID 2), environmental hazards and structural risk factors (UID 3,6,7), and estimating the population at-risk (UID 9). Several use-cases begin pre-disaster but lead into disaster response and recovery, looking at food, water, and community health status (UID 1,4,8,10,11), identifying skilled people and developing community capacities (UID 14,15), accessibility (UID 12), and information sharing (UID 5,13). While most card-sort loadings generally reflected the participant's perceived usefulness ratings, some card-sorts, like with P18, reveals existing needs and priorities at different time-phases, where six of the top 15 use-cases can be recognized (Figure 2-5).

Table 2-5: Top 15 use-cases discussed by 5 or more participants

UID	Use-cases (number of participants)	Definition	Information cards used (fraction of participants)	Relative time to event
1	Assess access to care and treatments (9)	Access to usual treatment was discussed with regards to chronic disease patients	Hospital locations (0.78) | Blocked roads (0.44) | Pharmacy locations (0.44) | Flooded areas (0.33) | Roads and Highways (0.33) | County demographics (0.11) | Territory geographic boundaries (0.11) | Storm damage (0.11) | State/Territory demographics (0.11) | Qualitative Interviews (0.11) | Public Health Assessments (0.11) | Health outcomes by State (0.11) | Health outcomes by County (0.11)	Before and after
2	Locate and prioritize areas with vulnerable populations (8)	The elderly, children, the frail, and people who depend on devices to live. This may include knowledge of where are the nursing homes, intermediate care facilities, and the allied specialized healthcare facilities.	Census Tract demographics (0.5) | County demographics (0.38) | Territory geographic boundaries (0.12) | Flooded areas (0.12) | Health outcomes by County (0.12) | Hospital locations (0.12) | Public Health Assessments (0.12) | Qualitative Interviews (0.12) | Pharmacy locations (0.12) | Time constraint (Real-Time) (0.12) | Time constraint (up to 2 weeks) (0.12) | Time constraint (up to 4 days) (0.12)	Before
3	Identify risk factors that will impact people in the area (8)	Based on prior information about hazards, identify what risk factors are present on-site and how they may have direct or indirect effects in understanding causal outcomes within the community.	Testing for regulated water quality endpoints (0.38) | Testing for Unregulated chemical contaminants (0.38) | Streamflow hazard (0.38) | Public Health Assessments (0.38) | County demographics (0.38) | Flooded areas (0.38) | Qualitative Interviews (0.25) | Microbial presence testing (0.25) | Census Tract demographics (0.25) | Hospital locations (0.25) | County/Municipios boundaries (0.25) | Story route (0.25) | State/Territory demographics (0.12) | Annual population estimates (0.12) | Census Block boundaries (0.12) | Human resource (0.12) | Microbial interpretation aid (0.12) | Pharmacy locations (0.12) | Time constraint (Annual) (0.12) | Storm damage (0.12) | Territory geographic boundaries (0.12) | Time constraint (up to 4 days) (0.12) | Time constraint (Real-Time) (0.12) | Time constraint (up to 2 weeks) (0.12)	Before
4	Characterize potential harms in the water systems (7)	Prioritized with knowledge about the affected areas, water system tests can be performed to get a sense of the chemical and microbial exposure harms to the local population.	Testing for regulated water quality endpoints (0.57) | Microbial presence testing (0.57) | Testing for Unregulated chemical contaminants (0.43) | Chemical interpretation aid (0.29) | Roads and Highways (0.29) | Microbial interpretation aid (0.29) | Blocked roads (0.29) | Flooded areas (0.29) | Storm damage (0.14) | Streamflow hazard (0.14) | Time constraint (up to 2 weeks) (0.14)	After
5	Coordinate efforts with the administrative governance (7)	Engage with administrative entities to understand their information needs and collaborative opportunities to make research useful to them. Consider how to plan project work and disaster management steps to avoid impeding each other.	County/Municipios boundaries (0.71) | County demographics (0.29) | Health outcomes by County (0.29) | Territory geographic boundaries (0.29) | Census Tract demographics (0.14) | Census Block boundaries (0.14) | State/Territory demographics (0.14) | Qualitative Interviews (0.14) | Human resource (0.14) | Health outcomes by State (0.14)	Before and after
6	Assess the pre-event conditions (6)	Consider the status of resources vulnerable to change or damage before the disaster occurs.	Health outcomes by County (0.67) | Microbial presence testing (0.33) | Census Tract demographics (0.33) | County demographics (0.33) | Health outcomes by State (0.17) | Human resource (0.17) | Time constraint (Annual) (0.17) | Streamflow hazard (0.17) | Story route (0.17) | State/Territory demographics (0.17) | Roads and Highways (0.17) | Public Health Assessments (0.17) | Annual population estimates (0.17)	Before
7	Characterize areas with limited accessibility (6)	In the absence of telecommunication, areas with limited physical access would have difficulty seeking help. Document the status of roadways and transportation routes to characterize areas at risk of geographic isolation if damaged or blocked.	Roads and Highways (0.83) | Blocked roads (0.67) | Flooded areas (0.33) | Storm damage (0.17) | Pharmacy locations (0.17) | Hospital locations (0.17)	Before

Table 2-5: Top 15 use-cases discussed by 5 or more participants

UID	Use-cases (number of participants)	Definition	Information cards used (fraction of participants)	Relative time to event
8	Make regular assessments of individual and community well-being (6)	Collect information on the physical health, mental health status, and hazards affecting communities. Make cross-sectional measures of amount of damage and disrepair, the number of people that died or sustained health issues, and the prevelance of coping for emotional stressors.	Public Health Assessments (0.83) | Qualitative Interviews (0.83) | Time constraint (up to 2 weeks) (0.33) | Time constraint (monthly updates) (0.33) | Time constraint (Annual) (0.17) | Microbial presence testing (0.17) | Time constraint (Real-Time) (0.17) | Time constraint (up to 4 days) (0.17)	After
9	Estimate the effect denominators for number of people affected (5)	Based on the anticipated event, the number of people at-risk should be estimated based on the population size that are residents, work in the area, and the rate and frequency of changes in such estimates.	County demographics (0.6) | Census Tract demographics (0.6) | Annual population estimates (0.4) | Health outcomes by State (0.2) | County/Municipios boundaries (0.2) | Census Block boundaries (0.2) | State/Territory demographics (0.2)	Before
10	Identify major acute health concerns (5)	Surveil for concerns that may intensify into risks of acute death. Observing these health concerns would need preparation in order to recognize, mitigate, and contain early warning signs.	Public Health Assessments (0.8) | Microbial interpretation aid (0.4) | Qualitative Interviews (0.4) | Microbial presence testing (0.2) | Testing for Unregulated chemical contaminants (0.2) | Testing for regulated water quality endpoints (0.2) | Time constraint (Real-Time) (0.2) | Health outcomes by County (0.2) | Chemical interpretation aid (0.2) | Time constraint (up to 2 weeks) (0.2)	After
11	Estimate the expected rates of health outcomes (5)	Based on knowledge of the population size and prior health outcome events, estimate the expectation for possible health outcomes and use those to compare with the rates of occurence.	Health outcomes by County (0.6) | Public Health Assessments (0.6) | Annual population estimates (0.4) | Census Tract demographics (0.4) | Census Block boundaries (0.2) | County demographics (0.2) | County/Municipios boundaries (0.2) | Health outcomes by State (0.2) | Qualitative Interviews (0.2) | State/Territory demographics (0.2) | Territory geographic boundaries (0.2)	Before and after
12	Identify roads that are operational (5)	Identify roads and highways that were not damaged. This status is conditional, but the information is necessary as access ways for responder deployment and routing decisions.	Roads and Highways (0.8) | Blocked roads (0.8) | Flooded areas (0.4) | Storm damage (0.2)	After
13	Characterize the place and people in the community of focus (5)	Assemble a debrief about the place and community situation. This can include the languages and choices for communication strategies, the demography and population size, and cohesiveness. This is prerequisite knowledge to start human-centered efforts.	County demographics (0.6) | Qualitative Interviews (0.4) | Health outcomes by County (0.4) | Census Tract demographics (0.4) | Annual population estimates (0.2) | Human resource (0.2) | State/Territory demographics (0.2) | Story route (0.2)	Before and after
14	Identify skilled personnel (5)	Consider what kind of human resources are available and skill level. This could be anywhere and brought in or locally in the affect zone.	Human resource (1.0) | Training Modules and Workflows (0.2)	After
15	Provide just-in-time training (5)	During planning, the key players and human resources need to be identified. The skilled human resources may not be experts, but they may know enough to be trained quickly for technical tasks.	Training Modules and Workflows (1.0) | Human resource (0.6) | Chemical interpretation aid (0.4) | Microbial interpretation aid (0.4) | Census Tract demographics (0.2) | County demographics (0.2) | Health outcomes by State (0.2) | Health outcomes by County (0.2) | State/Territory demographics (0.2)	Before and after

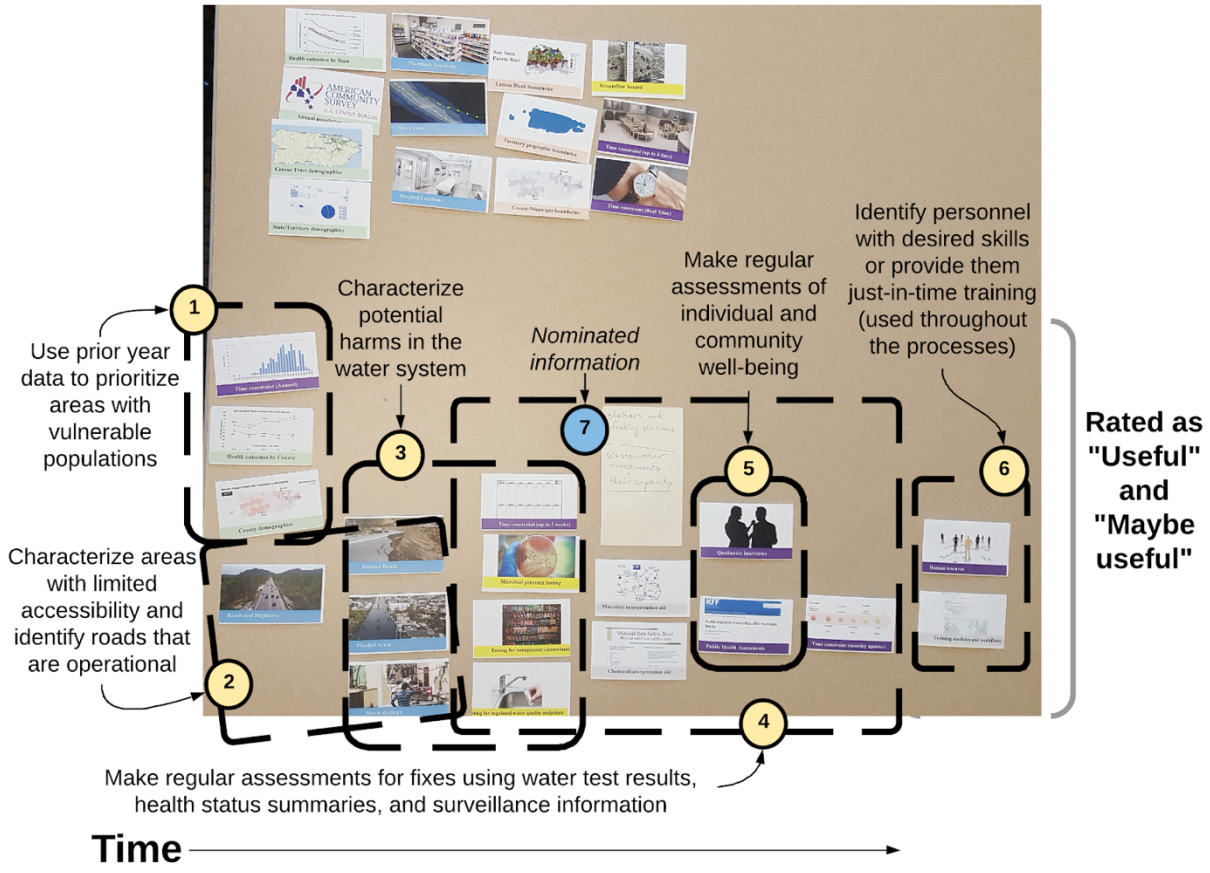


Figure 2-5: Card-sort with P18.

Think-aloud with the cards at-bottom revealed 18 potential use-cases, of which 6 clusters prioritized by time-sequence were visible in the loadings (1-6). For environmental health reconnaissance functions during disaster recovery, (1) and (2) summarise information about the accessible paths and the location of vulnerable populations. (3-5) establish preliminary assessments for fixes and disease surveillance needs. Throughout the deployment, (6) provides a means to find and build research capacities. “Shelter and feeding stations” and “Wastewater treatment [plants] & their capacity” were written in as nominated information (7).

The distribution of use-cases per information card is summarized as boxplots, colored based on sample size of participants (Figure 2-6). Participants discussed a median of 14 use-cases per session (range: 5-29) distributed across information cards. Some participants discussed far more use-cases than others, whereas there were many instances where participants rated information

cards as “useful” but did not verbally discuss a use-case. County demographics received the highest median number of use-cases and highest interquartile distribution. These distributions reveal heterogeneity in how different participants perceive information as appropriate for research use-cases when preparing for an envisioned hurricane or flood up to 10 days in the future.

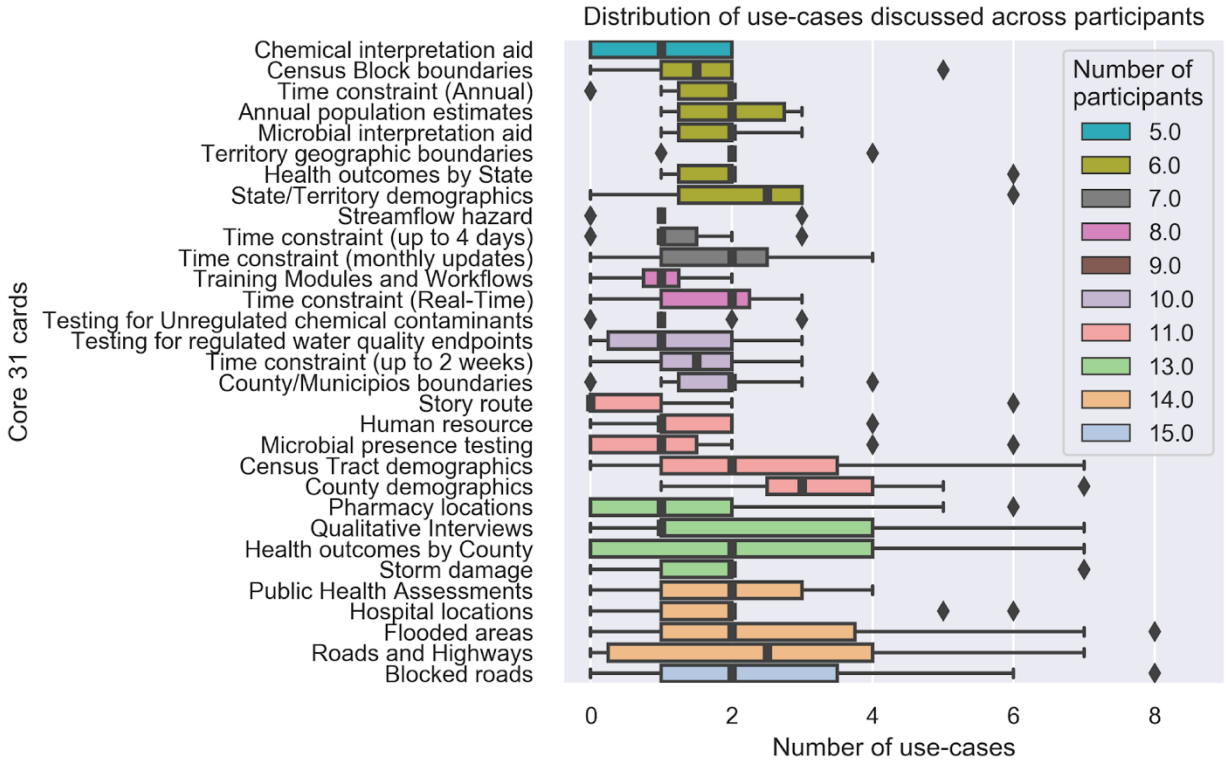


Figure 2-6: Distribution of unique use-cases by information card.
Each participant that discussed usage with the card are summarized as a point within each boxplot.

2.6 DISCUSSION

A clear separation was observed among strategies for future disaster research, generally due to their primary occupational appointment and readiness. Prior studies have discussed process

limitations and collaborator engagement to some extent [12,20,39], where it has been suggested to "*let the community teach the responders"* [20]. However, the status quo for research by reconnaissance teams generally begins at the end of disaster response phase and academic participation in disaster research is more characteristic to the disaster recovery phase. There exist barriers to population-based research projects if the population size and capabilities are diminished severely. While studies have proposed post hoc approaches for estimate long-term migration and evacuation, local displacement from areas with poor connectivity remains an issue [27]. The top use-cases are generally consistent with the priorities in data collection about a disaster site [18,39]; our approach detected variabilities in what population health researchers perceive as relevant information to address use-cases. Updating vulnerability assessments have been a point of concern for communications with the population at-risk, responders, and humanitarian volunteers [12,13,20,21]. With a larger participant pool, it is possible that the heterogeneity in broad use-cases could diverge into subset use-cases, but would require a different approach in identifying use-cases. Though our exploratory approach was able to detect gradients in participant perceived usefulness and envisioned use-cases, further research should consider the entirety of the technology acceptance model [36].

Information quality and learnability were recurrent themes within the needs and barriers in adopting data and technology. Towards the use of data archives, documenting longitudinal observations that can be continued even with an intermittent disaster timeline requires new efforts to generate multidisciplinary big data sets. Interpretability and usability arose as prominent concerns to clarify methods of collection and the relevance to response actions. Our study participants acknowledged these issues during their normal operations in population health but anticipate heightened difficulties trusting data within a disaster setting. Some concerns

originate from the transparency of evidence-based and validity of information, given changing standards of interpretation. Future work should consider validation methods within data-generating steps in greater detail.

Collaboration, context, and situational information are critical in disaster management and in conducting research. Workflows change when challenged by unreliable resources, siloed data, or loss of necessary facilities (electricity, communication systems, laboratories), which cause concerns about low quality data during workflow transitions. Collaborations complement the various barriers in data use and tools adoption, which can vary depending on training and expertise. Participants repeatedly mentioned concern for collaborators who normally operate near-capacity, stressed by the status-quo public health funding mechanisms. It is clear that various persistent factors add burden and barriers to researchers and collaborative interactions; however, it is unclear how these factors may prevent researchers from engaging in disaster-related research, if they or their domain have not had experience within disaster research before.

2.7 LIMITATIONS

Although this small but diverse sample had yield rich information, our findings might not generalize to the broader population health or other disaster research communities. Snowball sampling helped us reach the target researchers, but could introduce selection bias, unbalanced sampling, and under-representation e.g., researchers from the non-profit sector and local health jurisdictions were not represented). Card-sort findings were reported as descriptive statistics as it was unclear how to set thresholds for statistical enrichment between emergent use-cases and hypothetical information. Card labels were often referred to by pronouns and short names, leading to some uncertainties in reproducing reasoning from transcripts. Cards selected as useful

but not described within a use-case may be indicative of initial information biases. Low frequency use-cases may be relevant to domain-specific research operations, but it was not feasible to account for these subgroups given our sample size. For future studies, we recommend purposive recruitment to achieve balanced sampling for focused subgroups within the population health research community.

2.8 CONCLUSION

This chapter identifies the diversities within population health research and the various areas of research needs to understand population health within hurricane and flood disasters. Population health research is undertaken by a breadth of subject matter domains and researchers of distinguishing interests. Qualitative accounts highlighted key differences in research strategy, generally related to their readiness and primary occupational appointments. A mixed-method approach with card sorting helped us to investigate the heterogeneity in perceptions of information usefulness and learn of the variety of potential use-cases, expanding the knowledge gained than by qualitative semi-structured interview alone. While environmental, health, and geographic information may provide crucial contexts for situational use (15 common priority use-cases identified), health research users experience profound barriers in interacting with data products during hurricane and flood disruptions given the gaps in baseline knowledge. Users need information tools designed for transferable applications across science domains, proof of best-practices, and low barriers in adopting new technology. Resources to build cross-functional technical teams are needed throughout disasters and should be coupled with interactions that develop supportive collaboration engagements. Collaboration is central to current strategies for population health research in future disasters.

CONTRIBUTIONS FROM CHAPTER 2

In this chapter that is focused on Aim 1, I present a mixed-method analysis to explore the information needs, barriers and use-cases of population health researchers for incorporating health research into disaster preparedness. Two key takeaways were learned from the analysis. First, seven barriers and six facilitators emerged as important factors for further consideration in the design of better data sets, tools, and collaborative research solutions to support population health researchers and strategies for health research in future disasters. Second, 15 priority use-cases were identified as the top use-cases of concern for coordinated research efforts in future hurricanes and floods as well as an initial survey of what information would be informative towards those use-cases. Future studies focused on integrating population health research with disaster research should incorporate these initial design criteria and set of use-cases for the strategic planning of information sharing in future hurricane and flood disasters.

In the following chapter, I explore how recent advances in geospatial data visualizations, data sharing, and research computing environments can be integrated to study big data sets by geographic places of focus.

CHAPTER 3 - Automated retrieval, preprocessing, and visualization of gridded hydrometeorology data products for spatial-temporal exploratory analysis and intercomparison

3.1 INTRODUCTION

Gridded data products are extensively used in Earth Science research [40–42], social vulnerability analysis [10] and population risk and estimate studies [11]. Gridded hydrometeorological data products are produced by interpolating local observations to predetermined spatial-temporal resolutions. The purpose of developing gridded products is to extend spatial information beyond point locations and provide space and time dimensions to observations such that spatial-temporal variability can be analyzed. Gridded data products also provide a means to compare and validate numerical weather prediction outputs (short term forecasts and long term climate change). Prior studies in hydrometeorology have highlighted the growing usefulness of gridded data products in Earth science modeling (most recently reviewed in Henn *et al.,* 2018) [43].

In this chapter, I introduce the Observatory for Gridded Hydrometeorology (OGH) open-source python toolkit, a collaborative product designed and developed while I was a research assistant within the University of Washington (UW) Civil and Environmental Engineering Watershed Dynamics Group. I will use "we" to refer to the research team within the Watershed Dynamics Group. I represent the approach and conceptual diagramming as shown in Phuong *et al.*, (2019) [44].

OGH was designed to support watershed scale science applications. The purpose was to streamline processes for interacting with gridded hydrometeorological data products at a user-

defined spatial scale of interest (single location to regional watershed). This tool fills a model pre-processing gap of processing large regional datasets (~1000 km2) for smaller scale geographic subsets (~1 km^2 - 100 km^2).

In the conterminous United States (CONUS), since the introduction of Parameter-elevation Regressions on Independent Slopes Model (PRISM) [45], gridded meteorological data products are routinely interpolated using daily measurements from over 20,000 NOAA COOP observation stations [46,47], with similar products in development for other regions around the world [48]. In these data products, each grid cell contains observation-interpolated, multivariate time-series [46]. It is common practice for the hydrologic community to incorporate gridded meteorological time-series variables as inputs to land surface hydrologic models such as the Variable Infiltration Capacity (VIC) model [49]. A wealth of modeled land surface hydrologic states (e.g, soil moisture) and fluxes (e.g., latent and sensible heat) have been developed and used in Earth science research [47,50]. In mountainous regions, where ground-level observation collection is not feasible, the Weather Research Forecasting (WRF) atmospheric model has been downscaled and similarly used to produce hydrologic model outputs; however, this process inevitably requires bias-correction based on observational products [51,52]. Recently, gridded hydrometeorological data was combined with geo-environmental and demographic surveys to yield gridded population data sets and new insight to enhance population-modeling resolution and accuracy [10,11]. It can be expected that gridded data products will continue to increase in abundance and complexity in how they represent the impacts of geography and landscape morphology.

Before the potential usefulness of gridded data products can be realized for watershed-scale actionable research, several data and metadata access challenges need to be addressed.

Continental-scale gridded data products, such as those from Livneh *et al.,* (2013; 2015) [47,50] and Salathé et al., (2014) [52], are increasingly being published as NetCDF files. For watershed researchers who are interested in studying physical processes, NetCDF files used for regional and continental-scale gridded products (1000 km^2 - 10,000 km^2) contain information that far exceeds the geographic extent needed for local, watershed-scale research (*e.g.,* 1-100 km^2 catchment area), adding computational resource burden in exploratory research. One alternative data product format is the 1D ASCII time-series files for a geographically-specific gridded cell. 1D ASCII time-series files may not be self-described with column names, time-series dates, or value units, so annotations will need to be extracted from source files and publications in order to perform analyses with the published files, as observed in Livneh *et al.* (2013; 2015) and Salathé *et al.* (2014) data products [47,52]. Information to locate and use the data files may be confined to elusive publications and documentation files. Even so, the data files may be hosted in management structures for their study convenience (*e.g.,* Universal Transverse Mercator boundaries), making manual data retrieval non-trivial and not intuitive by human interpretation. Hence, annotating data provenance, metadata provenance, and file management structure are crucial steps towards making gridded data products findable, accessible, interoperable, and reusable (FAIR) [53,54] for secondary analyses.

Currently available tools for water data, such as WaterML, WOFpy, GSFLOW, Geoknife, offer access to time-series of observation data [55–57]. However, in areas where observations are unavailable, such as heterogeneous landscapes at high elevation locations, data sparsity can be addressed with krigged and model-interpolated data products. Python libraries such as OpenClimateGIS offer access to NetCDF gridded data products, but these functionalities exclude legacy data sets provided in 1D ASCII time-series format. More importantly, aside from data

access, it is challenging to recognize differences between gridded data product—such as aggregation into different gridded cell schemas, the temporal resolution, time period, or long-term trends—for use in future modelling efforts and model validation operations.

To streamline the processes needed for interacting with gridded hydrometeorological data products in a FAIR manner, promote use in Earth modeling and interdisciplinary domains, and support decision-making for selecting gridded cells in a study site, we designed OGH as an open-source python toolkit to select gridded cells in a study site, data download, spatial-temporal analyses, and provide data visualization. OGH is comprised of functions written to simplify data access and processing of gridded data product for research use. Users begin with ESRI shapefiles that describe their study site (e.g., HUC12 units, county-boundaries, state-boundaries). Combining the study site shapefiles with shapefiles that describe the centroid point of the grid cell schema, the user generates a comma-separated value table to manage grid cells of interest across across gridded data products of the same grid cell schema. Users can then retrieval data files in-parallel from a number of gridded hydrometeorological data products with automated file management for analyses. These functions are flexible to integrate new gridded products and publishing standards. To address the absence and/or variability in describing online gridded data products, a set of minimum annotation criteria (metadata fields) was proposed for describing ASCII gridded hydrometeorological data products and the decision steps needed to access these gridded datasets.

In the Methods section, we describe OGH software design for gridded cell selection and visualization, data download, spatial-temporal calculations, and applied statistics functionalities. OGH was designed to incorporate climatological and hydrometeorological gridded data products with comparable data structures to ASCII and NetCDF formats. Here, we emphasize watershed-

scale applications using 1D ASCII time-series data products. In the Results section, we demonstrate these functionalities using three watersheds of end-member climatologies in CONUS: two high alpine glacierized watershed in the high-end of the precipitation gradient in the CONUS (> 3,000 mm), Sauk-Suiattle, WA [58] and Elwha river basin, WA [59]; and a desert watershed with a large precipitation gradient from valleys to peaks (200 mm - 3500 mm), Upper Rio Salado, NM [60]. We compute precipitation and temperature spatial-temporal statistics and exceedance probability calculations using gridded hydrometeorological data products from Livneh *et al.,* (2013) and Salathé *et al.,* (2014) [47,52]. OGH v.0.1.11 is publicly accessible at https://github.com/Freshwater-Initiative/Observatory and available by conda installation.

3.2 METHODS

OGH is a python library designed to perform gridded cell selection, data download, data processing for desired space-time analytics, and visualization of spatial-temporal data. A proposed set of metadata annotations was developed to provide default information about data product capacities using two case study gridded products. We provide OGH examples reproducible on HydroShare, a cyber-infrastructure for sharing data and models [61–63]. OGH is not dependent on HydroShare, though in the example use-cases, HydroShare is a platform providing dockerized or local server environments for community software, including those used by OGH operations to manage, compute, and store directories of files. User workflows, scenario use-cases, and key socio-technical needs were developed through key informant interviews, diagramming, and rapid prototype testing sessions [64,65].

3.2.1 Software Design

OGH was written using Python v.3.6. OGH functions were designed as modular components that leverage class structures and methods from various Python libraries, including Pandas time-series analysis [66]; Geopandas, Fiona, and Shapely geospatial visualizations [66–69]; and Multiprocessing and Dask distributed computing processes [70]. Functions script sequential operations into wrapped functions, which can then be applied in distributed computing practices.

OGH is intended to perform operations within computing environments, where input and output files are managed within data sharing platforms and community data repositories (e.g., HydroShare). While HydroShare has many features, we make use of the docker or server environment, where file storage, migration, and computation can be performed with multi-core resources. HydroShare (www.hydroshare.org) is a collaborative platform that supports data sharing and model reproducibility in hydrologic research, and it provides a cloud-computing environment through the Consortium of Universities for the Advancement of Hydrologic Science, Inc. (CUAHSI) JupyterHub server [62]. The HydroShare REST API Python library (*hs_restclient*) is used to migrate files contained in HydroShare resources in and out of the JupyterHub docker environment [71]. File storage within the computing environment is intended to be temporary storage for the duration of the computations. The research workflows presented in this paper are designed to be guided by Jupyter notebooks, sharable code-execution interfaces that operate within the JupyterHub docker environments [63].

3.2.2 Gridded data product annotations

We annotated seven daily, 1/16^{o} (~6 km) gridded data products from three studies (Livneh *et al.,* 2013, 2015; Salathé *et al.,* 2014) [47,50,52]. The datasets published by Livneh *et al.,* (2013)

provided an interpolated climate-station meteorology for CONUS and a meteorology data product that was bias-corrected to the Columbia river basin regional climatology in the time-span from 1915 to 2011 [47]. Expanding on Livneh *et al.,* (2013), Livneh *et al.,* (2015) includes a PRISM-calibrated interpolated climate-station meteorology extends from Mexico to limited regions of Canada [50]; however, the period has 33 total years less data with a time-span of 1950 to 2013. Both interpolated meteorology data products were used to predict macro-scale hydrologic fluxes at the $1/16^{o}$, daily resolution by the VIC model. The WRF gridded data product provides model downscaled daily precipitation, maximum and minimum air temperature, and wind speed for the Columbia river basin for the period of 1950 to 2010 [52]. Each data products can be differentiated by features and variables, the start date and end date of annotated data, the type of analysis, reported spatial coverage, and their reference publication among other metadata of potential use (Table 3-1).

Table 3-1: Summary of seven daily, 1/16° gridded data product.

Data set	Features and variables (in order)	Start date	End date	Analysis type	Spatial coverage	Publication
Climate station meteorology						
dailymet_livneh2013	PRECIP, TMIN, TMAX, WINDSPEED	1915-01-01	2011-12-31	raw	CONUS	[Livneh et al., 2013]
dailymet_bclivneh2013	PRECIP, TMIN, TMAX, WINDSPEED	1915-01-01	2011-12-31	bias-corrected	Columbia river	[Livneh et al., 2013]
dailymet_livneh2015	PRECIP, TMIN, TMAX, WINDSPEED	1950-01-01	2013-12-31	raw	CONUS	[Livneh et al., 2015]
WRF-NNRP model meteorology						
dailywrf_salathe2014	PRECIP, TMIN, TMAX, WINDSPEED	1950-01-01	2010-12-31	raw	Columbia river	[Salathé et al., 2014]
dailywrf_bcsalathe2014	PRECIP, TMIN, TMAX, WINDSPEED	1950-01-01	2010-12-31	bias-corrected	Columbia river	[Salathé et al., 2014]
Variable Infiltration Capacity						
dailyvic_livneh2013	YEAR, MONTH, DAY, EVAP, RUNOFF, BASEFLOW, SMTOP, SMMID, SMBOT, SWE, WDEW, SENSIBLE, LATENT, GRNDFLUX, RNET, RADTEMP, PREC	1915-01-01	2011-12-31	Physics-based model	CONUS	[Livneh et al., 2013]
dailyvic_livneh2015	YEAR, MONTH, DAY, EVAP, RUNOFF, BASEFLOW, SMTOP, SMMID, SMBOT, SWE, WDEW, SENSIBLE, LATENT, GRNDFLUX, RNET, PETTALL, PETSHORT, PETNATVEG	1950-01-01	2013-12-31	Physics-based model	CONUS	[Livneh et al., 2015]

The annotations describe the gridded data products published as ASCII files, where each file contains the gridded cell historic time-series data. Annotation features include the data set short name, information to locate the ASCII files, information about the file structure and sources of metadata, and metadata about the file variables (Table 3-2). File locations can be represented or reconstructed given by the web protocol (e.g., ftp, https), web domain and subdomain, decision steps within the subdomain to locate the data file subdirectory (e.g., centroid latitude given the the spatial resolution, bounding box bins), the filename structure, and the file format. The file structure is described by the variable list (left-to-right column order), time-series date range,

temporal resolution, file delimiter, and the data types and unit increment for each variable. Full annotations are provided in the ogh_meta module.

Table 3-2: Minimum annotation criteria for gridded data products.

Metadata	Metadata descriptions
File location	
1. Dataset	name of the gridded data product
2. Spatial resolution	the distance between gridded cell centroids
3. Web protocol	the data transfer protocol
4. Domain	the web domain
5. Subdomain	the subdomain path
6. Decision steps	the file organization for locating data files
7. Filename structure	the standard components to the filename
8. File format	the file type at download
File structure	
9. Start date	the start date of the time-series
10. End date	the end date of the time-series
11. Temporal resolution	the unit increment for time-steps
12. Delimiter	the column separator within each line of data
13. Variable_list	the list of variables in order of appearance
14. Reference	the sources of metadata
Variable structure	
15. Variable_info	
● desc	the long name of the variable
● dtypes	the expected data type
● units	the unit increment of the data

3.2.3 Example use-cases

We present four example use-cases in the form of Jupyter Notebooks to demonstrate the OGH operations (Figure 3-1). In the first use-case, we identify the subset gridded cells of interest for three watershed study sites using the *treatgeoself* function (Figure 3-1A). The shapefiles for these watersheds are stored within a public HydroShare resource for ease of collaborative use [Sauk-Suiattle river basin [58], Elwha river basin [59], and Upper Rio Salado basin [60]]. Watershed boundaries were defined in ArcGIS® using 12-digit Hydrologic Unit Code polygons from the National Watershed Boundary Database. In the second use-case, the time-series data files are retrieved, cataloged, then summarized for data availability (Figure 3-1B). In the third

use-case, we focus on the Sauk-Suiattle watershed to determine the monthly meteorological spatial-temporal statistics computed using the Livneh *et al.,* (2013) Meteorology versus the Salathé *et al.,* (2014) WRF model output data products (Figure 3-1C). Finally, we compute potential runoff values using the VIC hydrologic data product from Livneh et al., (2013) to approximate the 10% exceedance probability thresholds based on the daily time-series in each dataset (Figure 3-1D).

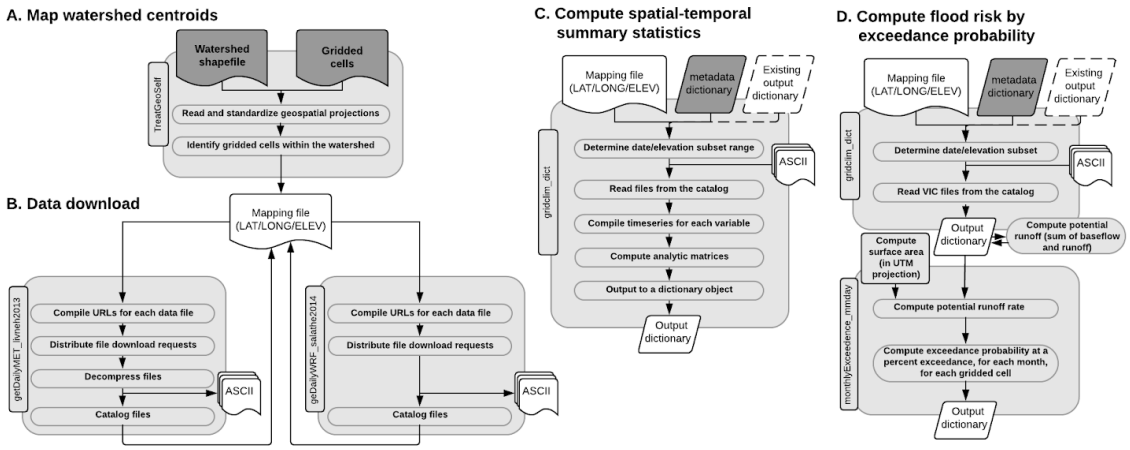


Figure 3-1: Scenario use-cases for OGH operations in cloud environments.

A) With the user-defined watershed shapefile, spatial intersection with the 1/16o gridded cell centroid shapefile identifies the target gridded cell centroids, which are then documented within the mapping file output. B) The mapping file provides the Lat-Long coordinates for data download operations to produce a localized folder of ASCII files and a file catalog appended within the mapping file. C) For each gridded data product, spatial-temporal summary statistics are computed from the mapping file, file structure metadata, and ASCII files. The output dictionary can be reused to collect summary statistics for multiple data products. D) The hydrologic gridded product is used to compute exceedance probability at a statistical threshold for each gridded cell.

For illustration, five reference locations are used in the Sauk-Suiattle watershed examples. One gridded cell is identified at the highest average elevation value, 2216 meters above sea level. Two gridded cells were identified at the lowest average elevation value, 164 meters above sea

level. The Darrington ranger station (COOP station 451992) is used as the reference source of meteorological observations, with data collected at 167 meters above sea level from Jan 1 1931 through Dec 31 2005 [72]. Sauk River Near Sauk, WA (USGS-12198500) used as the reference source of observed streamflow discharge using data collected between Jan 1 1950 through Dec 31 2011 [73,74].

3.2.4 General workflow and required files

Workflows for the use-cases executed in the JupyterHub environment (Figure 3-2) are illustrated in detail in Figure 3-1. The general workflow begins with three HydroShare resources as sources of input files (Figure 3-2). Resource A - a HydroShare resource that contains a Jupyter Notebooks to execute code for each example use-case presented in this paper, Resource B - a HydroShare resource with a user-defined shapefile representing the region of interest (e.g., a watershed), and Resource C - a file of point-locations describing CONUS gridded cell centroids (only pre-requisite for 'mapping watershed centroids'). From the web page for HydroShare Resource A, the Jupyter Notebooks are launched in the JupyterHub docker environment, wherein the HydroShare REST API functions migrate in requisite data files from Resource B and C (Figure 3-2). Use-case notebooks 1 through 4 progress through OGH operations in Figure 3-2: 'identify watershed gridded cell centroids' (map watershed centroids),'download and display data availability' (data download), 'summarize monthly meteorology' (data processing), and 'compute exceedance probabilities' (another form of data processing) (Figure 3-2). Each use-case notebook produces output data files, plots data visualizations, and finally migrates these output to new shareable HydroShare resources to conclude the use-case demonstration.

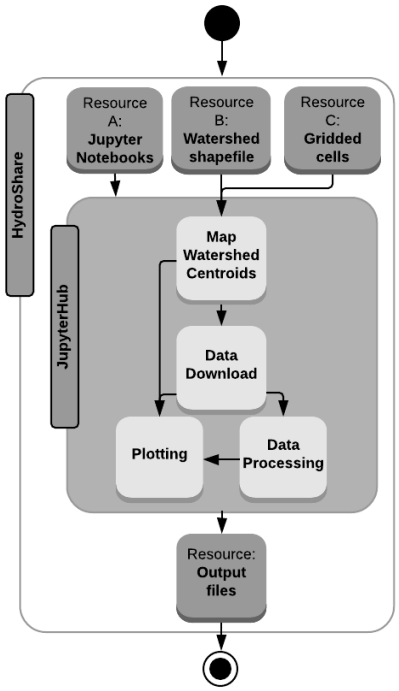


Figure 3-2: The general workflow for OGH in cloud-computing environments.
HydroShare is the collaborative platform to facilitate file storage (Resources A, B and C), and HydroShare API rest-client migrates files in-and-out of the JupyterHub docker environment. Jupyter notebooks guide users through different use-cases like decision-making steps (e.g., map watershed centroids), data download, data processing, or generating visualization products. At the close of each notebook, research data products are migrated to HydroShare as new sharable resources.

3.2.5 Map watershed gridded cell centroids

For each of the three example watersheds, we generate a mapping file with the gridded cell centroids that spatially intersect these study sites (Figure 3-1A). Shapefiles were transformed into the 1984 World Geodetic System (WGS84) Lat-Long coordinates system as the standard projection. The study site was given a buffer region (default buffer distance of 0.06^{o}) to include adjacent gridded cells. CONUS $1/16^{o}$ (i.e., 0.06250^{o}) gridded cell ESRI shapefile identifies each gridded cell by the 5-digit centroid latitude-longitude [47,75]. Average elevation in the gridded

cell (in meters above sea level) are based on the CONUS digital elevation model described in Livneh *et al.,* (2013) [47].

The output from this use-case includes three mapping files, each denoting the latitude, longitude, and average elevation within the gridded cell. Other outputs include spatial visualizations of the study site (maps) and the elevation gradient among the gridded cells, and plots showing the data for select grid cell traces.

3.2.6 Summarize data download and availability

The mapping file guides data download from the seven gridded data products (Figure 3-1B). Target gridded cell files identified within the mapping file are web requested using data download wrapper functions (e.g., the *getDailyMET_livneh2013*). Request operations are distributed using multiprocessing pool operations. Downloaded files are cataloged into the mapping file (using *addCatalogToMap*). Data availability is determined for each gridded data product and watershed study site (using *mappingfileSummary*). Files that do not exist for retrieval are excluded from the catalog.

The output from this use-case is a summary table that describes data availability and seven folders containing the downloaded files for all the watersheds.

3.2.7 Summarize monthly meteorology

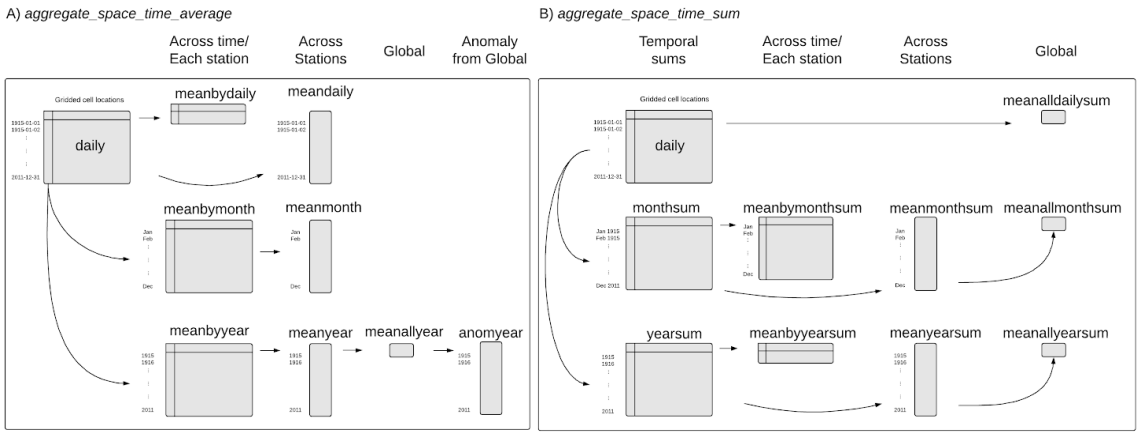


Figure 3-3: Spatial-temporal calculations (total sum and average).
The *gridclim_dict* function reads and applies A) *aggregate_space_time_average* for each variable in the gridded data product. Variables to be considered by periodic sums (e.g. total annual precipitation) can be processed with B) *aggregate_space_time_sum*.

With the Livneh *et al.,* (2013) interpolated meteorology and Salathé *et al.,* (2014) WRF files, we compared monthly meteorology variables for the Sauk-Suiattle river watershed using the 61-years of data from their mutual time-series period (i.e., Jan 1, 1950 through Dec 31, 2010). Using the Sauk-Suiattle mapping file, each variable from the ASCII gridded cell time-series are compiled into data frames, where rows are the daily time-series and columns are denoted with the gridded cell centroids. Temperature trends are interpreted using monthly mean, yearly mean, and global mean expected values (Figure 3-3A). Annual anomaly from the global mean value is used to identify years with extreme events (highest and lowest data values). Precipitation trends are interpreted using period sums (e.g., month-yearly sums and yearly sums) and mean of period sums (e.g., mean monthly sums, mean yearly sums, and the global mean of monthly sums) (Figure 3-3B).

The *gridclim_dict* function is a series of wrapped operations to return a dictionary object of spatial-temporal values across the ASCII gridded cell time-series (Figure 3-1C). *Gridclim_dict*

provides parameters to specify elevation ranges or time-period selection, where defaults are all gridded cells and the full time-series. *Gridclim_dict* wraps *read_files_in_vardf*, which performs distributed file reading to generate a variable data frame for each variable in the data product, then applies *aggregate_space_time_average* to compute summary statistics using the prefix-suffix conventions for each variable (Figure 3-3A). The suffix represents the gridded data product, which can be user-defined or default to the annotated gridded data product dataset name (e.g., 'dailymet_livneh2013'). The first prefix appended to the suffix by underscore separation is the data product variable (e.g., 'PRECIP'). The second prefixes represent the statistical averages computed using the gridded cell dimensions (columns) and the temporal groupings (rows). The *aggregate_space_time_sum* produces outputs with the following second prefixes: "meanbydaily" (daily averages by each gridded cell), "meanbymonth" (daily averages by each month and gridded cell), "meanbyyear" (daily averages by each year and gridded cell), "meandaily" (average values across gridded cells by each date), "meanmonth" (daily averages across gridded cells for each month), "meanyear" (daily averages across gridded cells by each year), "meanallyear" (global mean of daily values across all years and gridded cells), and "anomyear" (the residual between each yearly mean and the global mean). To consider trends by period sums of daily events, the *aggregate_space_time_sum* function computes summary statistics of month-yearly and yearly sums (Figure 3-3B). The second prefixes here include "monthsum" (month-yearly sum of daily values by gridded station), "yearsum" (annual sum of daily values by gridded station), "meanbymonthsum" (mean of monthly sums for each calendar month and gridded cell), "meanbyyearsum" (mean of annual sums for each gridded cell), "meanmonthsum" (mean of month-yearly sums across gridded cells), "meanyearsum" (mean of annual sums across gridded cells),"meanalldailysum" (global mean of the daily sums across all gridded cells),

“meanallmonthsum” (global mean of month-yearly sums across gridded cells), ”meanallyearsum” (global mean of annual sums across gridded cells). Variations to these outputs are influenced by the gridded cells and time-period parameters.

The output for this use-case is a JSON dictionary object containing analytical data frames and data series, as shown in Figure 3-3, for each variable within a given time-frame. Other outputs include maps and monthly boxplots of the corresponding grid cell values.

3.2.8 Compute exceedance probabilities

For the Sauk-Suiattle study watershed, we approximate the 10% exceedance probability threshold using unrouted daily runoff for each calendar month and each gridded cell (Figure 3-1D). This is useful for visualizing the 10% highest daily streamflow generated for each grid cell in the dataset, which is a combined function of climate data, soils, land cover, and other model parameters in each grid cell. Potential runoff rates are computed as the sum of baseflow rate (mm/s) and surface flow runoff rate (mm/s) from Livneh *et al.,* (2013) VIC model outputs, converted the units to millimeters per day (mm/day) for comparison with daily precipitation rates. The same general operations were applied to Livneh *et al.,* (2015) VIC model outputs. For each calendar month (e.g., January) and each gridded cell (*e.g.,* centroid Lat-Long at 48.8723, -121.8974), daily potential runoff rates are compiled into a cumulative distribution function using data from 1 Jan 1950 through 31 Dec 2011 (62-years), the mutual/overlapping time-series period between Livneh *et al.,* (2013) and Livneh *et al.,* (2015) data products. Each distribution has approximately n=1800 VIC modeled observations. The 10% monthly exceedance probabilities (peak runoff threshold) for each gridded cell is estimated by linear interpolation as the 90th-percentile of the respective cumulative distributions [76]. An exceedance probability developed

from a population of daily runoff in a given month should not be confused with annual flood statistics, which are developed by fitting a statistical distribution to a population of annual maximum daily streamflow. The 10% exceedance probability of observed streamflow discharge measured at Sauk River Near Sauk, WA (USGS-12189500) can be plotted with the modeled streamflow to provide an observed reference based on routed streamflow for relatively high flows.

The output for this use-case includes low, average, and high elevation analytical data frames at the 10% exceedance threshold for VIC results, compared to observations. Other outputs include maps and monthly boxplots of the exceedance probability for each gridded cell.

3.3 RESULTS

3.3.1 Map Watershed Centroids

Functions used:
reprojShapefile, treatgeoself, multiSiteVisual, griddedCellGradient

Sauk-Suiattle, Elwha, and Upper Rio Salado watersheds were processed to generate mapping files and gridded cell gradient visualizations (Figure 3-4). Ninety-nine grid cells were identified for the Sauk-Suiattle river watershed, displaying the largest elevation difference (162 m - 2246 m) among the three watersheds (Table 3-3). Sauk-Suiattle river watershed is located in the northwestern region of the Cascade mountains in Washington state, USA, ranging from multiple high elevation areas in the southeast to a single outlet in the northwestern gridded cells. Fifty five gridded cells were identified for the Elwha river watershed, which has a comparable elevation difference to Sauk-Suiattle. The Elwha river watershed is located on the northern region of the

Olympic Peninsula, where the elevation gradient (36 m - 1642 m) descends from the southern gridded cells by a single river draining to the northern gridded cells (Figure 3-4). Thirty one grid cells were identified for the Upper Rio Salado watershed, with a higher elevation (1962 m - 2669 m) grid cells than the other two watersheds (Table 3-3). Upper Rio Salado's elevation gradient descends from the southwest-most to the northeast-most gridded cell.

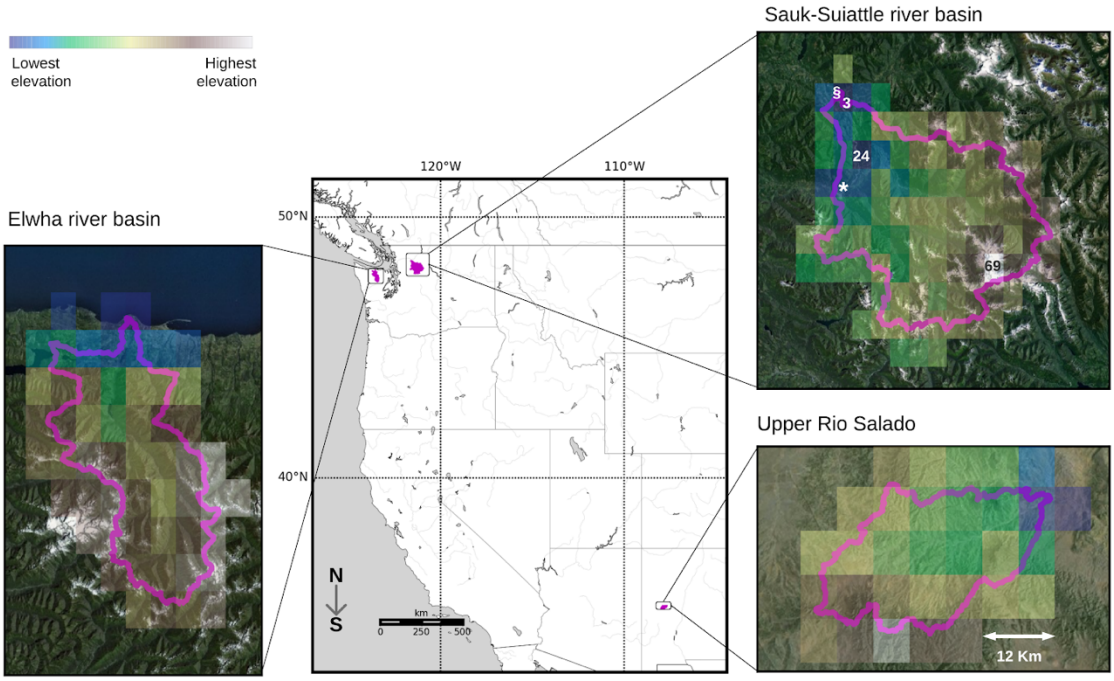


Figure 3-4. Aerial view of watersheds and gridded cells.

The Sauk-Suiattle (164-2216 m), Elwha (36-1642 m), and Upper Rio Salado (1962-2669 m) watersheds located in western United States were visualized using the *multisiteVisual* function with an EPSG:3857 geospatial projection. Each watershed (outlined in magenta) and their gridded cells are visualized using the *griddedCellGradient* function at the 1/16° spatial-resolution (~6 km). In the Sauk-Suiattle watershed, five reference markers denote the highest elevation gridded cell (gridded cell 69, elevation: 2216 m), the lowest elevation gridded cells (gridded cells 3 and 24, elevation: 164 m), the Darrington Ranger Station site (*; COOP station 451992, elevation: 167 m) for observed meteorology data, and the Sauk River Near Sauk, WA streamflow gauge (§, USGS-12189500, elevation: 81 m) for observed streamflow discharge measured at the downstream-most tip of the watershed. While the numeric distributions can be conformed to a single scale, each watershed map uses a different numeric colorbar legend, so this figure is intended to provide a qualitative impression of the elevation gradient.

3.3.2 Summarize data download and availability

Functions used:
getDailyMET_livneh2013, getDailyMET_bcLivneh2013, getDailyMET_livneh2015, getDailyVIC_livneh2013, getDailyVIC_livneh2015, getDailyWRF_salathe2014, getDailyWRF_bcsalathe2014, mappingfileSummary

Among the seven gridded data products, 1D ASCII time-series files were fully-represented for Sauk-Suiattle, mostly represented for Elwha, and substantially limited in representation for Upper Rio Salado (Table 3-3). Download tasks for the full time-series ASCII files were distributed across 5-10 parallel worker CPUs. Computation efficiencies consisted of 693 Sauk-Suiattle files (10.0 Gb disk space) downloaded in 3 min 56 s wall time, 375 Elwha files (5.4 Gb) took 1 min 59 s, and 124 Upper Rio Salado files (2.7 Gb) took 48.8 s. All files were cataloged into their respective mapping files, organized by gridded data product short name and gridded cell centroid.

Elwha is located in the northwestern-most region of Washington state. Three gridded cells were available for all seven gridded data products, although they were available for the bias-corrected Livneh *et al.,* (2013) meteorology and Livneh *et al.,* (2015) meteorology and VIC model output products. Differences among the elevation gradient suggest that these three gridded cells were the northern-most low-elevation gridded cell, on the boundary of CONUS and Columbia River Basin extents (Figure 3-4). This poses certain limitations if multiple gridded data products for Elwha were used for intercomparison. These limitations are more obvious with Upper Rio Salado, which is located outside of the Columbia River Basin.

Livneh *et al.,* (2013) and (2015) gridded products were consistently spatially available in each of the watersheds, but the gridded products differed in the temporal extent (historic time-series included). The overlap period between Livneh *et al.,* (2013) and (2015) data products is

Jan 1 1950 through Dec 31 2011 (62-years). Livneh *et al.,* (2013) and Salathé *et al.,* (2014) share the Jan 1 1950 through Dec 31 2010 (61-years). Despite the spatial availability of time-series data within a watershed, gridded data product intercomparisons should consider the historic time period represented as well as data variabilities such as correction methods and algorithms used to generate the gridded product.

Table 3-3: Counts of gridded cell ASCII files for each watershed by gridded data product. For the seven gridded data products, the downloaded files are summarized for each watershed as an inventory of the data availabilities and potential gaps due to spatial extent of the gridded data product.

	Watersheds		
	Sauk-Suiattle river	Elwha river	Rio Salado
Median Elevation in meters [range] (Number of gridded cells)	1171[164-2216] (n=99)	1020[36-1642] (n=55)	2308[1962-2669] (n=31)
dailymet_bclivneh2013	1171[164-2216] (n=99)	1120[36-1642] (n=55)	0
dailymet_livneh2013	1171[164-2216] (n=99)	1146[174-1642] (n=52)	2308[1962-2669] (n=31)
dailymet_livneh2015	1171[164-2216] (n=99)	1120[36-1642] (n=55)	2308[1962-2669] (n=31)
dailyvic_livneh2013	1171[164-2216] (n=99)	1146[174-1642] (n=52)	2308[1962-2669] (n=31)
dailyvic_livneh2015	1171[164-2216] (n=99)	1120[36-1642](n=55)	2308[1962-2669] (n=31)
dailywrf_salathe2014	1171[164-2216] (n=99)	1142[97-1642] (n=53)	0
dailywrf_bcsalathe2014	1171[164-2216] (n=99)	1142[97-1642] (n=53)	0

3.3.3 Summary monthly meteorology

Functions used:
findCentroidCode, overlappingDates, gridclim_dict, aggregate_space_time_sum, valueRange, saveDictOfDf, renderValueInBoxplot, renderValueInPoints

The function *gridclim_dict* generates a JSON dictionary for Sauk-Suiattle that contains 36 analytical data frames for Livneh *et al.,* (2013) meteorology and Salathé *et al.,* (2014) WRF outputs. Each data frame was named according to the analysis method (second prefix), variable (first prefix), and gridded data product short name (suffix). In Figure 5, average monthly total

precipitation (*i.e., meanbymonthsum_PRECIP_dailymet_livneh2013* and *meanbymonthsum_PRECIP_dailywrf_salathe2014*) are depicted as boxplots to represent the distribution of values across the 99 gridded cells. The Livneh *et al.,* (2013) interpolated meteorology (Figure 3-5, top-left) indicates a greater variability of average monthly precipitation during the November through January months, while Salathé *et al.,* (2014) WRF model outputs (Figure 3-5, bottom-left) shows a higher median and greater variability from April through September. Average monthly precipitation for targeted high and low elevation grid cells (Figure 3-4) are plotted alongside the boxplots, as well as the point observations from Darrington Ranger Station (Figure 3-5). Comparison of observations and modeled precipitation shows that observed precipitation (at low elevations) is less than the monthly averages modeled by Salathé *et al.,* (2014) during spring and summer. Spatial variations observed with Livneh *et al.,* (2013) show large deviations between neighboring cells, especially in comparison to the smoother spatial trends organized with the elevation gradient can be observed from the Salathé *et al.,* (2014) (Figure 3-5, right).

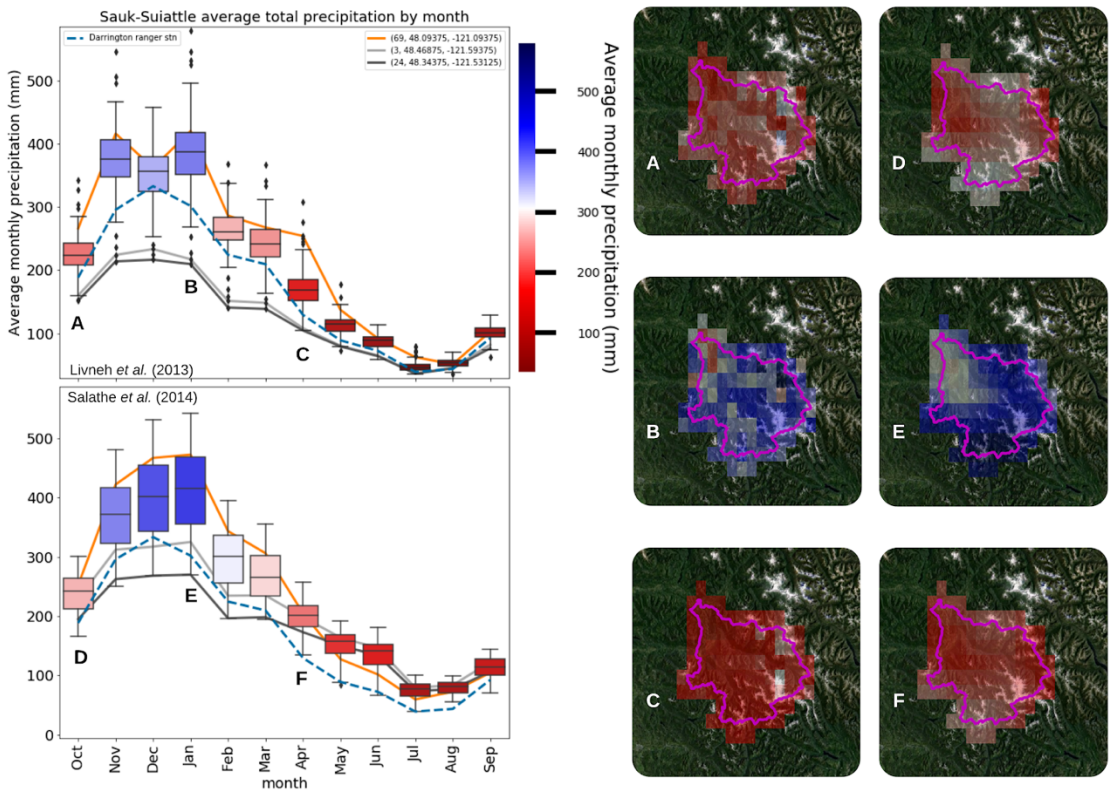


Figure 3-5: Comparison of the average monthly total precipitation for each gridded cell in the Sauk-Suiattle watershed.

The boxplots compare the statistical distribution of the average monthly total precipitation (inches) to the spatial distribution of precipitation in each gridded cell (A-F), using data from Jan 1 1950 through Dec 31 2010. Created using the *renderValueInBoxplot* function, the boxplot colors represent the median value of the gridded cell distributions. Reference trend lines were included to illustrate Sauk-Suiattle's highest elevation gridded cell (#69; orange) and the lowest elevation gridded cells (#3 and #24; light and dark gray), found using the *findCentroidCode* function. The gridded cell distributions are rendered spatially with a basemap using the *renderValueInPoints* function. The spatial distribution of gridded cell values are rendered using the *renderValueInPoints* function for Livneh *et al.* (2013) interpolated meteorology for A) October, B) January, and C) April, compared with to Salathé *et al.* (2014) WRF model outputs for D) October, E) January, and F) April. All maps and boxplots use the same colorbar legend and numerical distribution shown in the top-left.

Monthly temperature statistics were computed for each grid cell using the daily minimum and maximum temperature between Jan 1, 1950 through Dec 31, 2010 (Figure 3-6). The

distribution of mean maximum temperature shows that Livneh *et al.,* (2013) interpolated meteorology has greater variability than Salathé *et al.,* (2014) WRF model outputs. This effect is also observed when comparing the distribution of mean monthly minimum temperature, noting that Livneh *et al.,* (2013) has more extreme hot and cold trends, sometimes up to 5^{o}C difference compared with the Salathé *et al.,* (2014) WRF model outputs. Reference meteorological observations from the Darrington Ranger Station closely resemble the Livneh *et al.,* (2013) interpolated meteorology. Livneh *et al.,* (2013) values are dependent on source observations clustered around low elevation gridded cells (light and dark gray) with COOP stations; sparse observations limit the performance assessment for high elevation gridded cells. While average daily minimum temperature seems to be comparable, Salathé *et al.,* (2014) predicts colder maximum temperatures for low elevation areas for all months, and warmer temperatures for higher elevation areas (orange line) from November through April.

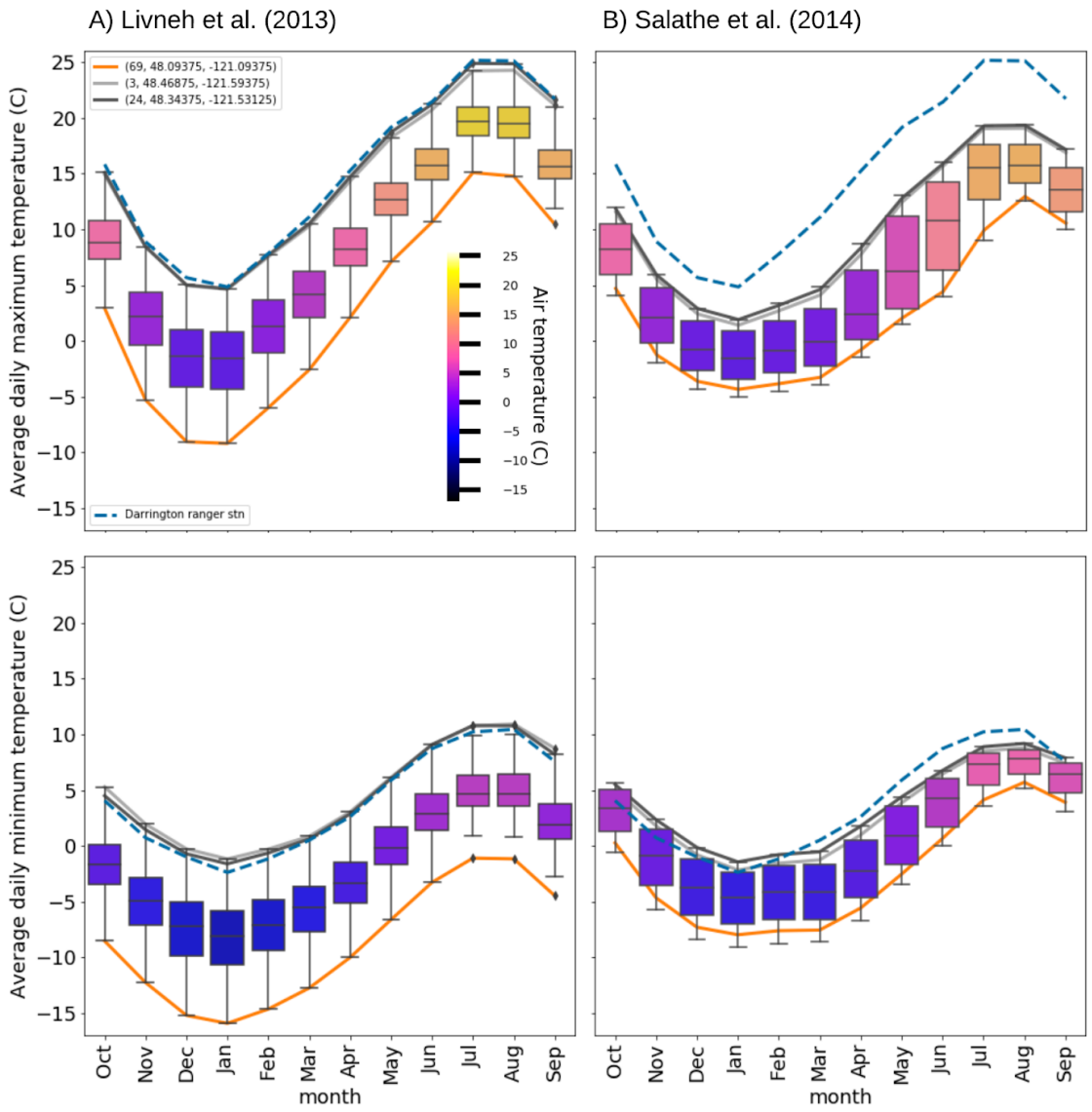


Figure 3-6: Comparison of monthly mean of daily minimum and maximum temperature.

The monthly mean of daily maximum (top) and minimum (bottom) temperatures (in Celsius) were computed for each of the 99 Sauk-Suiattle gridded cells. The boxplots represent the observations from Livneh *et al.,* (2013) meteorology (left) and Salathé *et al.,* (2014) WRF model outputs (right). Reference trend lines were included to represent the highest elevation gridded cell (orange) and the lowest elevation gridded cells (light and dark gray) in Sauk-Suiattle. The field observations (blue dashed line) measured at Darrington Ranger Station (elevation: 167 m) indicates that maximum daily temperature (top) are more closely represented by Livneh *et al.,* (2013) in the Sauk-Suiattle watershed, while there are no remarkable differences observable for minimum daily temperature (bottom).

3.3.4 Compute exceedance probabilities

Functions used:
monthlyExceedence_mmday, computeSurfaceArea, cfs_to_mmday

Figure 3-7 displays the monthly 10% exceedance probability and average (50%) thresholds for unrouted potential runoff using two VIC gridded data products for the Sauk-Suiattle watershed. Not to be confused with approximations like the 10,000-year flood which are based on empirical streamflow values, the probabilities generated by this function are based on empirical unrouted model outputs, which have limited numeric range and interpretation thereof. Each point in the distributions represent the potential runoff threshold at which there is only a 10% chance expectation of exceeding that value in that month. Both gridded data products display a slight increase in November where the muted impact of the highest extreme winter flood events on lower average monthly flows, followed by a high peak flow in June/July (from snowmelt runoff), using the same color scale and axis ranges (Figure 3-7, right side). The major contribution to potential runoff in any given year is from the Cascade mountain ranges during the snowmelt season (June-July). Although the 10% exceedance probability at the highest elevation are comparable between data products, the trends at the low elevation stations indicate that Livneh *et al.,* (2015) produces more potential runoff between November through April months than Livneh *et al.,* (2013). Livneh *et al.,* (2015) applies a bias correction (using PRISM data as a proxy for observations), which produces more precipitation from winter rainfall season compared to results without a bias correction. A comparison between the two gridded data products illustrates that Livneh *et al.,* (2015) boxplots have approximately 2-3 mm median increase in potential runoff between October to May compared to Livneh *et al.,* (2013) (Figure 3-7).

The monthly 10% exceedance probability with Sauk River Near Sauk, WA, discharge observations range from 4.2 to 15.3 mm/day across the calendar months. The exceedance probability threshold peaks in two months, November and June, corresponding with fall atmospheric river rainfall-dominated storm events, and early summer snowmelt. The first smaller peak is observed in November, which aligns with both Livneh *et al.,* (2013) and (2015) the boxplot distributions. The second larger peak occurs in June, where the mean decreases though the variance increases in July. The Sauk River Near Sauk gauge is approximate 81 meters above sea level, half of the average elevation in gridded cell 3 where it is located. The spatially averaged (mm/day) routed streamflow dynamics in Sauk River Near Sauk can be expected to fall within the elevation mean of unrouted VIC modeled runoff using the lower resolution $1/16^{o}$ gridded cells. The observed data at the outlet is provided for context, and the modeled results is provided to demonstrate the spatial variability of the grid cells contributing to the modeled streamflow at the watershed outlet.

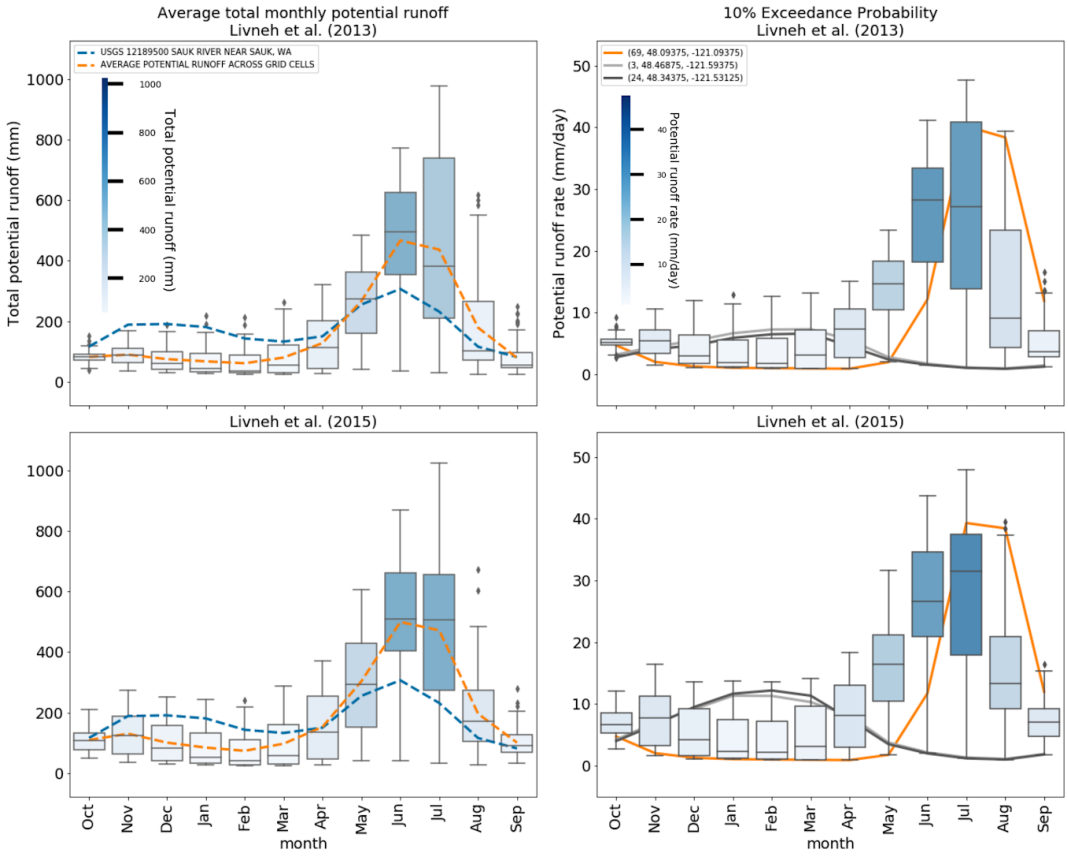


Figure 3-7: Average total monthly potential runoff (mm) and 10% exceedance probability for each monthly unrouted potential runoff (mm/day) within the Sauk-Suiattle watershed.

The boxplots are comprised of 99 gridded cell values for each month. Peak of average total monthly potential runoff (left) occurs in November and June months shown by the observed USGS streamflow discharge (blue dash), and observable by the spatial average of the gridded cell (orange dash). The 10% exceedance probability for each gridded cell (right) is a function of the spatial average of peak flow occurs in November and July. The snowmelt season is the major period for expected runoff for highest elevation gridded cell (orange line), while the rainfall season is the major period contributing to runoff for lowest elevation gridded cells (light and dark gray lines).

3.4 DISCUSSION

The primary step in the workflow presented here is *treatgeoself*, which enables users to control gridded cell inclusion and exclusion using shapefile-guided data selection. It generates the mapping file to guide distributed computing for data download and data processing; this catalogue allows for machine reading, selection and sorting of available data. The examples use

watersheds defined by HUC12 boundaries, but the shape can be user-defined (*e.g.,* census block, legislative boundaries, fish species migration spatial clusters). Geopandas operations enable *treatgeoself* to transform shapefiles of varying spatial projections and to include buffer regions. In the early development stage, *treatgeoself* applied unfiltered spatial intersection with each shape polygon in the shapefile, resulting in slow mapping performance and interpretation difficulty when buffer regions were included. At present, study sites with multiple subpolygons are merged by spatial union into a single MultiPolygon object, simplifying *treatgeoself* into a first-order loop. Intercomparison of gridded data products of different gridded parcel schemas are expected to be enabled and more efficient with the use of the projection alignment and cross-mapping functionalities.

Data download operations are functionalized for distributed computing, but the concurrent queue and transfer rate are limited by the computing resources allocated by the user and the data content provider. Data download was found to be rate-limited to approximately 5 concurrent web requests to the Livneh *et al.,* (2013) web domain. All other gridded data product hosts enabled 10 or more concurrent web requests. A rate-limiter for the number of parallel data retrieval tasks was incorporated into the data download functions, but not for local data processing operations. The rate of data transfer would need to be assessed before OGH could integrate data servers with RESTful API such as ERDDAP, which could expedite mapping and retrieval of gridded data products and metadata [77]. Other limits include nuanced issues of data maintenance by the data publisher/provider. For example, during production and testing of workflow and functions, data products mentioned in Livneh *et al.,* (2013) were migrated to a new web domain, resulting in misdirected requests. Annotations and data retrieval functions may need updating over time.

We qualitatively described differences between two gridded data products of the same empirically estimated statistical exceedance probability approach. The VIC-modeled gridded data products originate from unrouted flow modeling. In contrast to empirically estimated 10,000-year flood at-stream gauges, empirically estimated exceedance probability for grid cells without flow routing may be limited in interpretations to potential runoff. It is unclear how the different model simulations may affect these interpretations. These concerns regarding model comparison would merit further research and development of functionalities for in-watershed stream gauge selection and quantitative determinations for the goodness-of-fit between routed and unrouted modeled values relative to those estimated from at-stream observations.

While OGH was designed specifically for users who import ASCII-formatted files into hydrologic and earth surface model software e.g. Landlab [78], a noteworthy limitation of using ASCII file format is that as NetCDF adoption is increasing as a data standard, ASCII time-series may not be available for newer gridded data products. This limitation is addressed using the proposed minimum information criteria and having initial criteria for conducting gridded data product intercomparisons. NetCDF files are embedded with metadata, while ASCII files are unannotated. To inform the structure and use of the ASCII files, the proposed minimum information criteria serves as a road map for locating gridded data product files and considers the schema of the file organization and the features within each file. Gridded data products published by Livneh *et al.,* (2013) partitioned files by spatial bounding box subfolders denoted by the file prefix, West, East, South, and North cardinal limits. For Livneh *et al.,* (2013), *scrape_domain* and *mapToBlock* functions were designed to abstract the bounding boxes then decide the subfolder identity by spatial intersection. The spatial bounding box for gridded cells within

British Columbia, Canada did not follow this folder naming structure; thus, a separate annotation was provided for the spatial boundary in British Columbia, Canada.

Among the annotated ASCII gridded data products, we've observed a variety of file organizations; different gridded cell schemas, spatial resolution, or NetCDF file organizations may be adaptable. Retrieval and data management of NetCDF files in cloud computing environments would benefit from further design assessments, as it is not yet clear how to conduct or evaluate NetCDF-to-ASCII intercomparison without *a priori* format preferences that may result in information loss. In addition, the development of a user-centered reference of controlled vocabulary would improve the usefulness and adoption of a minimum information criteria that can be used across data formats. These may help adapt climate and water resource information for researching interdisciplinary questions with other data products such as Air Quality or Population data sets [11,79]. For use by researchers who are not hydrometeorology analyst, NetCDF files contain data outside the study area extent; 1D ASCII time-series files may be the preferred format for small study areas (1-100 km^2).

We developed and tested the examples using HydroShare for the computing and data sharing environment. An important benefit of HydroShare is that it hosts a REST API that enables data migration and the creation of new shareable data objects. Additionally, as a community repository for hydrologic science, FAIR publication of hydrologic data sets and software execution with reproducible workflows is demonstrated with the use-cases developed in this work. OGH operations are technically independent of HydroShare, and minor changes would allow the code to operate in other similar computing and data sharing environments such as local servers, cloud servers (Amazon AWS, Microsoft Azure), and dockerized virtual environments with a Jupyter instance (data.world, DataOne, PanGeo, ESIPhub).

3.5 CONCLUSIONS

OGH is a toolkit that makes download and processing of large climate datasets more efficient by leveraging distributed computing for watershed scale research and intercomparison of ASCII gridded data products, which extends climate modeling products to represent otherwise sparsely observed parts of the landscape. The mapping file output is the key data management tool, which catalogs the watershed gridded cells and downloaded files as a lens across gridded data products. Along with the proposed minimum information criteria to annotate ASCII gridded data products, these data management tools enable multiprocessing and dask-distributed operations comparable to the efficiency of Xarray for NetCDF gridded data products. This metadata component improves the standardization of gridded hydrometeorology products published for use by third-party researchers and scientists. The dictionary of analytical data frames is a key data management device that enables key-value pair retrieval and exporting of summary outputs. To address user needs for exploratory data analysis and visual control, various data frames were rendered into different geographic and temporal modes of human-readable visual inspection. Overall, OGH is equipped with metadata framework and workflow that makes it a useful introduction and training tool for watershed studies using gridded data products and ASCII time-series data sets. The data summary capabilities increase the efficiency of comparing multiple gridded hydrometeorology products without discontinuous use of different software. OGH and the four use-cases demonstrated are available for interactive use on HydroShare (https://www.hydroshare.org/resource/87dc5742cf164126a11ff45c3307fd9d) and also available for open development from the University of Washington Freshwater Initiative Observatory repository: https://github.com/Freshwater-Initiative/Observatory).

CONTRIBUTIONS FROM CHAPTER 3

In this chapter focused on Aim 2, in collaboration with hydrometeorologists - researchers that focus on modeling climate and water-resource processes, I designed a workflow with four commonly practiced use-cases to process structured hydrometeorological gridded time-series data. I annotated metadata for multiple hydrometeorological gridded modeling outputs (gridded data products), compartmentalized processing tasks into use-cases for automation, and incorporated statistical and data visualization operations for users. More importantly, climate research data products can be made accessible and introduced to new researcher through a minimum set of metadata for communication. One product of this work is the creation of an open-source python library that enable these operations and transfer between computing environments. The workflow was designed for user interaction with data within remote virtual machine resources, similar to the computing architecture of HIPAA-aligned enclave computing environments.

In the following chapter, I explore medical diagnosis information for associations with geographic place, wherein various data visualization operations developed here would be incorporated for communicating findings on geographic maps.

CHAPTER 4 - Identifying geographic influxes in patient attendance and health outcomes in the State of Washington from the perspective of the Electronic Health Records

4.1 INTRODUCTION

Due in part to climate change, an unprecedented number of extreme climate catastrophes have tested the weaknesses in health systems and community resilience across the globe [80]. Adoption of electronic health records (EHRs) have been the most important innovation providing opportunities to study health system resilience and improve patient care. Since the HITECH Act of 2009, over $19 billion have been invested to advance health information technologies and improvements to EHRs across the United States [81]. Meanwhile, policies like the Affordable Care Act continue to reduce the rate of uninsurance in the United States [82]. EHRs hold a wealth of information about people who seek medical care and this information holds increasing potential to inform population health questions as lack of insurance becomes less of a barrier to care. Although primarily a tool for care delivery operations, EHR information could also benefit research focused on population health and disaster preparedness. It could describe the geography of medically vulnerable patient populations, inform strategic positioning of emergency resources, and provide characterizations of diverse patient needs to mitigate disruption to care.

Despite the availability of health information documented in EHRs, it is uncertain which geographic or subgroup populations are best represented within secondary-use analyses. At any given time in Washington state, patients may seek healthcare from any of the hundreds of hospitals, health clinics, and provider groups. Provided that patients have proximity to care and means for access to care, EHRs document care provided and not the care provided elsewhere.

The UW Medicine network adopted EHRs into use in the early 1990s and has expanded the in-network connectivity to prevent siloing of patient care information within individual EHRs. This presents opportunities to gain insights about the geographic distribution of diverse health outcomes in the patient population. Studies have inquired about researcher needs of EHRs to pursue meaningful research questions in healthcare [83], practices for geospatial analyses to be mindful of patient privacy and confidentiality [84], heterogeneity in diagnosis coding practices [85,86], and the barriers conferred in extrapolating perspectives from a single EHR system to county or state-level inferences [87]. To the best of our knowledge, while some of these limitations have been acknowledged, the catchment areas served by UW Medicine where patient health is proportional to understand population health has not been explored.

In recent years, the population health and spatial epidemiology communities have developed methods to explore spatio-temporal research questions with medical diagnoses. Prior studies often characterized county-level geographic variability in health outcomes using state registry data sets [87–89] and few studies focus on the spatio-temporal relationships of multiple health outcomes [89,90]. More and more, spatial scan approaches are used to model spatio-temporal patterns and identify clusters of elevated outcome incidences [91,92]. These approaches have been used to estimate trajectory and size of communicable disease outbreaks [91–93], scan for hotspots in cancer incidence [94], and aid cancer care delivery systems to identify their patient catchment area [95,96]. While various geospatial use-cases proposed for the study of social determinants of health could potentially benefit from spatial scan and other spatial analysis methods with EHRs [97], little is known about how analytical workflows can scale to study granular local patterns of health effects in a large health care delivery system's patient population, while adhering to HIPAA-aligned computing practices. It is also unknown how

cancer hotspot and care delivery analysis could be adapted for use in disaster preparedness and response.

In this chapter, with biomedical and population health researchers as the primary actors, two research use-cases explore the spatio-temporal trends in UW Medicine patient attendance, diagnosis outcomes, and geographic units. Each use-case method analyzes patient records within a HIPAA-aligned virtual machine environment. Although I was the primary lead on the study, I will use "we" to refer to efforts and decisions made by myself, the research team, and the collaborating co-authors. First, we identify hotspot clusters among zip-codes areal units where the annual patient attendance deviates from prior year patterns more than the surrounding geographic areas. Second, we identify zip-codes that have statistical enrichment patterns for categories of medical diagnoses (hereafter referred as "diagnosis code families") and whether they were recent findings or consistent with enrichment analyses of prior time periods. Finally, we contrast the information learned about hotspot clusters in patient attendance with the information about enrichment and depletion associations.

4.2 MATERIALS AND METHODS

4.2.1 Computing architecture

This study focuses on use-cases with de-identified patient diagnosis and zip-code information. The study was determined by the University of Washington Human Subjects Division as minimal-risk non-human subjects research and exempt from Institutional Review Board approval [98]. All patient diagnosis data files were deidentified, extracted from the UW Medicine clinical data repository, and transferred into a HIPAA-aligned x86 64-bit redhat-linux-

gnu virtual machine environment by the University of Washington Medicine Institute of Translational Health Sciences honest broker (Figure 4-1). The virtual machine environment was also maintained by the honest broker and access permissions were limited to the study team. Additional files for analyses such as the zip-code shapefile, annual population estimates, the diagnosis code ontologies were imported into the environment. The study team developed analytical scripts within this environment using Bash *v.4.2.46*, Python *v3.6* and R *v.3.5.1*. Statistical analyses and geographic visualizations were conducted using Python with the pandas (*v.0.25.1*), numpy (*v.1.17.2*), matplotlib (*v.3.1.1*), networkx (*v.2.3*), geopandas (*v.0.6.1*), statsmodels (*v.0.10.1*), and dask (*v.2.6.0*) python libraries. Spatial scan statistics analysis was conducted using R with the scanstatistic (*v.1.0.1*) packages. Only data visualizations and summary outputs were exported from the system.

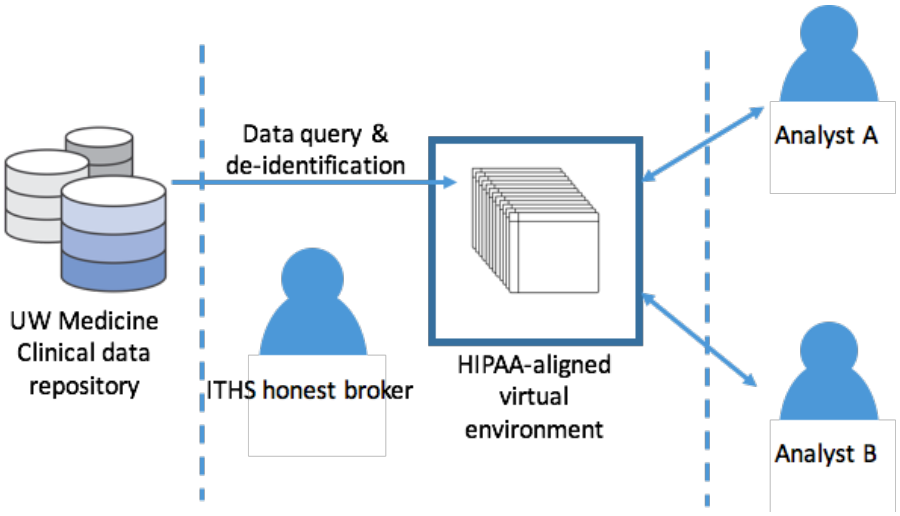


Figure 4-1: Data analysis and architecture.
Data are extracted and migrated into the virtual environment by the honest broker. The study team performs data cleaning, manipulation, and analysis with the data solely through the system by Bash Python, and R.

4.2.2 Patient diagnosis data set

Data availability: De-identified patient diagnosis information was requested for patient visits between January 01 1993 through November 22 2017 to the UW Medicine network, including in-patient and out-patient visits to multiple hospital centers (e.g., UW MC, Harborview MC, etc.) and clinics. Data availability included 2 million unique patients, 25 million patient visits, and 75 million individual diagnoses, accessed from UW Medicine Clinical Data Warehouse.

Data features and cleaning: Patient records were represented by demographic information (Race and ethnicity, death date, sex, and most recent occupational status) and obfuscated Patient IDs. Patient visits were represented by the visit start date, the medical facility code, patient age (in years) and zip-code on file at the time-of-visit, and all annotated diagnoses determined from that visit instance as relevant to the medical phenomena and clinical procedures provided. Date and time entries were standardized to Pacific Standard Time notation for Washington State (Universal Transverse Mercator Zone 10 in YYYY-MM-DD HH:MM:SS.%3f format). Patient visits were removed if they were determined as "no shows", where no new information, annotated diagnosis code or procedure information was determined.

Diagnosis code relationships: International Classification of Disease version 9 and 10 (ICD9/10) patient diagnosis codes within the data set were documented from multiple providers and multicenter sources. While codes are assumed to be validated by clinical teams at diagnosis and at billing operations, we did not verify code validity towards the coded medical phenomena. Codes within ICD9CM and ICD10CM hierarchical tree networks were retained as valid codes for analysis [99,100]. Diagnosis codes were mapped to 22 high-leveled diagnosis code families (DxF) by crosswalk between ICD9CM and ICD10CM hierarchical structures (Appendix 4-1),

where parent-child subClassOf relationships were represented as shown within their respective ICD9CM and ICD10CM versions (circa February 2017) [99,100]. No code IDs were mutual between ICD9CM and ICD10CM. We transformed the crosswalk into a key-value dictionary for mapping purposes.

4.2.3 Geographic data and population estimate data

Census geographic polygons and population information: Located in the US Pacific Northwest region, Washington is a home rule state with 39 self-governing counties and local health jurisdictions. We mapped patient 5-digit zip-codes on-file at the time-of-visit to the 598 Washington state ZCTA, areal units established in the U.S. Census 2010 [101] to be comparable with the US Postal Service 5-digit zip code regions, excluding Post Office Box numbers. Annual population estimates for ZCTA used estimates provided by the Washington State Office of Financial Management Small Area Estimate Program (SAEP), version released in 2019 that includes estimates from 2000 through 2018 [102,103]. Shapefile geometry were standardized to EPSG:2927 spatial reference (units in meters) and NAD83 geodetic system as the preferred visual projection for Northern Washington.

Inclusion and exclusion of zip-codes: Three criteria were imposed to reduce risks of potential patient reidentification: 1) patient zip-codes that are international or external to Washington state, 2) zip-codes with attendance below 20 unique patients within a full calendar year, and 3) patient visits without a zip-code on-file -- were remapped to “other”, a non-geographic placeholder for low frequency zip-code categories.

4.2.4 Use-case 1: Identify Most Likely Clusters of patient population attendance

Goal and metric: Hotspot clusters or Most Likely Clusters (MLCs) have been used to identify the catchment areas for patient attendance [95,96] and hotspots of various public health and medical concerns [92–94]. MLCs are defined here as the spatial areas with the highest likelihood ratio between observed and expected counts as compared with the surrounding areas. MLCs can be determined by the Poisson spatial scan statistic, a relative risk score from the likelihood-ratio test for a scanning window *W* calculated by Monte Carlo simulation of observations as a Poisson random variable. The goal with computing MLCs is to determine the adjacent groups of ZCTA where patient attendance has endured deviation from prior year expectations the most relative to surrounding areas.

Key factors for the hypothesis-generating and testing: The likelihood-ratio test considers the alternative hypothesis that there exists a *W* where the observed count of patients within *W* is higher than would be expected by chance than outside of *W* [92]. To detect small area clusters with high relative risks, we used the circular scanning window approach, where each scanning window *W* can contain k=[1,2… 20] ZCTA nearest-neighbors based on distance between ZCTA centroids (in meters) [92]. Here, spatial scan performs a Monte Carlo simulation to get a confident measure of the observed counts based on the Poisson distribution model, which is compared with the expected counts modeled from the training time-period data. Multiple MLCs were identified using the sequential deletion procedure described in Zhang *et al.* [104]. Under the null hypothesis, where there is no cluster, all scanning windows have observed counts with a mean value *p* equal to the Poisson-based maximum likelihood estimate for expected counts *q*. Under the alternative hypothesis, there exists a scanning window *W* where the ratio between *p*

and q is greater than 1 and greater than the surrounding areas, indicating an elevated relative risk of the event within W than outside of W.

We separated a 17-year time-series of annual patient attendance into the earlier 9 years (training period: 2000-01-01 thru 2008-12-31) and the latter 8 years (test period: 2009-01-01 thru 2016-12-31). For each calendar year, patients were mapped to their most recent zip-code on-file to remove outdated patient representations to place. Following convention, we assumed that the expected count of patients associated to any ZCTA geographic unit is proportional to the estimated population in residence based on the Poisson distribution model. Using the training data, we fit a Poisson-based generalized linear regression model to predict the count of patient attendance as a function of the population size, where the corresponding 9-years of SAEP annual population estimates data were used for model fitting and corresponding 8-years for predicting the estimated patient attendance (Figure 4-2). For a more confident estimate of the observed patient attendance, we performed n=9999 Monte Carlo replications with the observed counts within the spatial scan function. Comparison between replicate observations and expected counts yield the relative risks and p-values, which test the hypothesis that the proportion of the population at-risk within W is statistically greater than outside W [92]. The sequential deletion procedure removes the spatial area for prior most likely clusters, then remodels and recomputes expected and observed counts for subsequent MLCs [104]. Spatial scan statistics analyses was conducted using the scanstatistic R package.

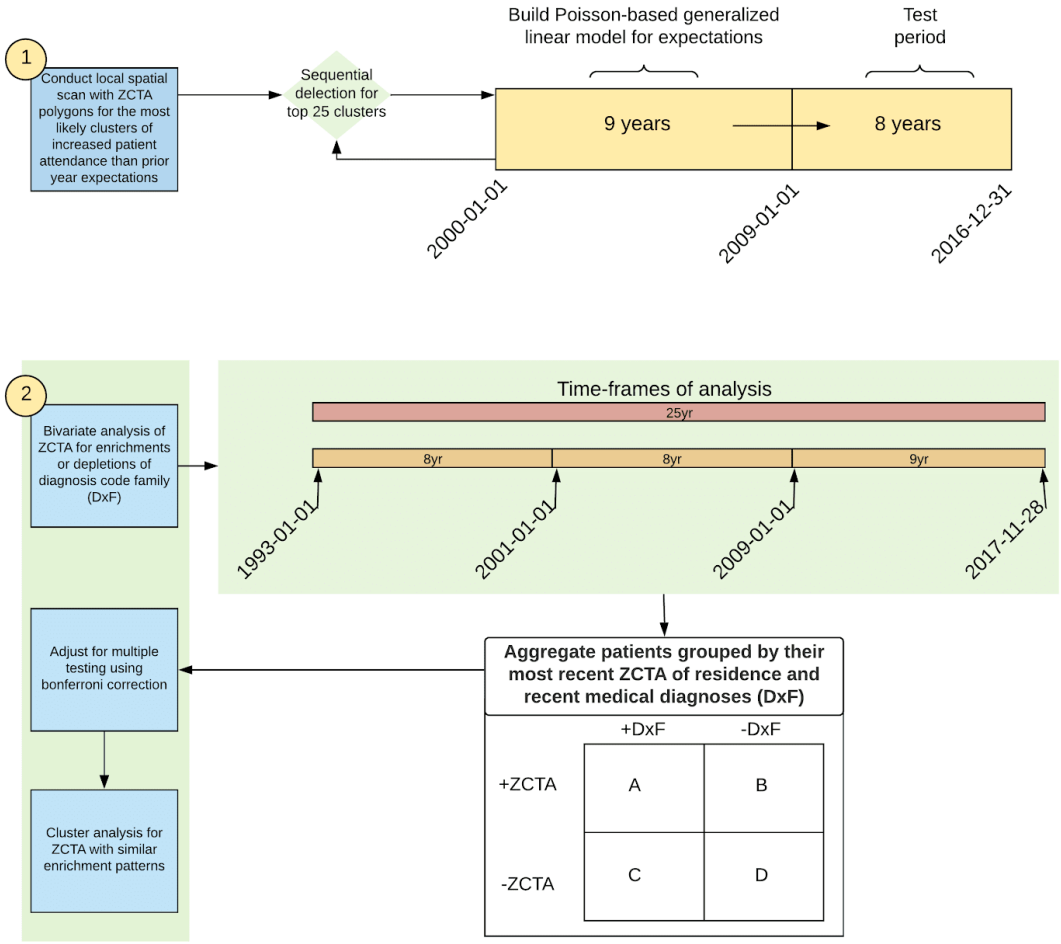


Figure 4-2: General workflow of the analysis.

Two research use-cases are incorporated to explore spatio-temporal trends in patient attendance and diagnosis. In use-case (1), spatial scan identifies the most likely clusters using the earlier 9 years as training for expected counts and the latter 8 years for hypothesis testing. Subsequent most likely clusters are recomputed after removing prior most likely clusters from the training and test set. In use-case (2), UW Medicine patient diagnosis records are separated into cross-sections to test for statistically significant patient population associations related to zip-code on-file.

4.2.5 Use-case 2: Identify statistical enrichments between zip-code on-file and diagnosis code families

Goal and metric: We conducted a series of cross-sectional bivariate analyses to estimate the odds ratio (OR) and 95% confidence intervals as measures of association and test for statistical enrichment of patient medical diagnoses within a Washington ZCTA compared to outside of that ZCTA . The goal is to identify enriched associations between patient medical diagnoses and ZCTA geography that may be consistent or recent enrichments.

Key factors for the hypothesis-generating and testing: Pairwise comparisons were made for 22 DxFs and 590 Washington ZCTA for UW Medicine patient zip-codes on-file. Each pairwise comparison was analyzed in four cross-sectional time-frames using UW Medicine patient visits and diagnoses within the time frame: the 25-years historic time frame (i.e., 1993-01-01 through 2017-11-29), two eight-year and one nine-year timeframes (i.e., 1993-01-01 through 2000-12-31, 2001-01-01 through 2008-12-31, and 2009-01-01 through 2017-11-29). For each hypothetical association, *p*-value test statistics were computed in the form of 2x2 contingency table and Fisher Exact Test for independence (Figure 4-2) [37]. The contingency table includes cross-tabulation of patients by their most recent zip-code on-file and whether they received a diagnosis within the timeframe to be considered a member of the diagnosis code family. Unlike prevalence estimates for a particular disease definition, we exclude adjustments for patient death within the odds ratio calculations.

Two methods were used to consider the evidence to reject the null hypothesis. Bonferroni correction was applied to adjust *p*-values with the family-wise error rate (standard threshold set at alpha=0.05). Separately, the 95% confidence intervals are evaluated for their evidence to reject the null hypothesis; if an odds ratio confidence interval crosses the null effect threshold (OR=1),

there is a probable likelihood that the odds ratio effect is random and thereby insufficient evidence to reject the null hypothesis. Odds ratios for pairwise tests that did not reject the null hypothesis are nullified (set OR=1) prior to cluster analysis.

Pattern identification: Hierarchical agglomerative clustering with Euclidean distance and Ward's minimum variance method [37] detected clustering organizations based on the log-base-2 odds ratios. The optimal *k* clusters in the bicluster is defined using the k-means clustering "elbow" method, where additional clusters after *k* clusters contributes a minimal reduction to the explained inertia [38]. The clustering configuration from the 25-year analysis are used as the frame of comparison for other time-frame analyses. After adjustment for multiple testing and evaluating the confidence intervals for evidence to reject the null hypothesis, associations are enrichments if OR>1 (or $\log_2 OR>0$), depletions if OR<1 (or $\log_2 OR<0$).

4.3 RESULTS

4.3.1 Use-case 1: Most Likely Cluster discovery

Figure 4-3 displays a preliminary analyses of the patient attendance at UW Medicine since EHRs were first integrated into clinical use in 1994. Total patient load steadily increased until 2009, where upon attendance escalated across age groups. While roughly 2 million uniques patients were provided care at UW Medicine in the recent 25 years of documented history, approximately 600 thousand patients received care in the first 8 years (1993-2000), 824 thousand patients in the second 8 years (2001-2008), and 1.1 million patients in the final 8 years and 11 months of the data set (2009-2017). While 2005 through 2008 displays a plateau, the segmented

8-year summary indicates that patient attendance has been on a steady growth, where events in 2008 onwards are observed in the marked increases in patient attendance.

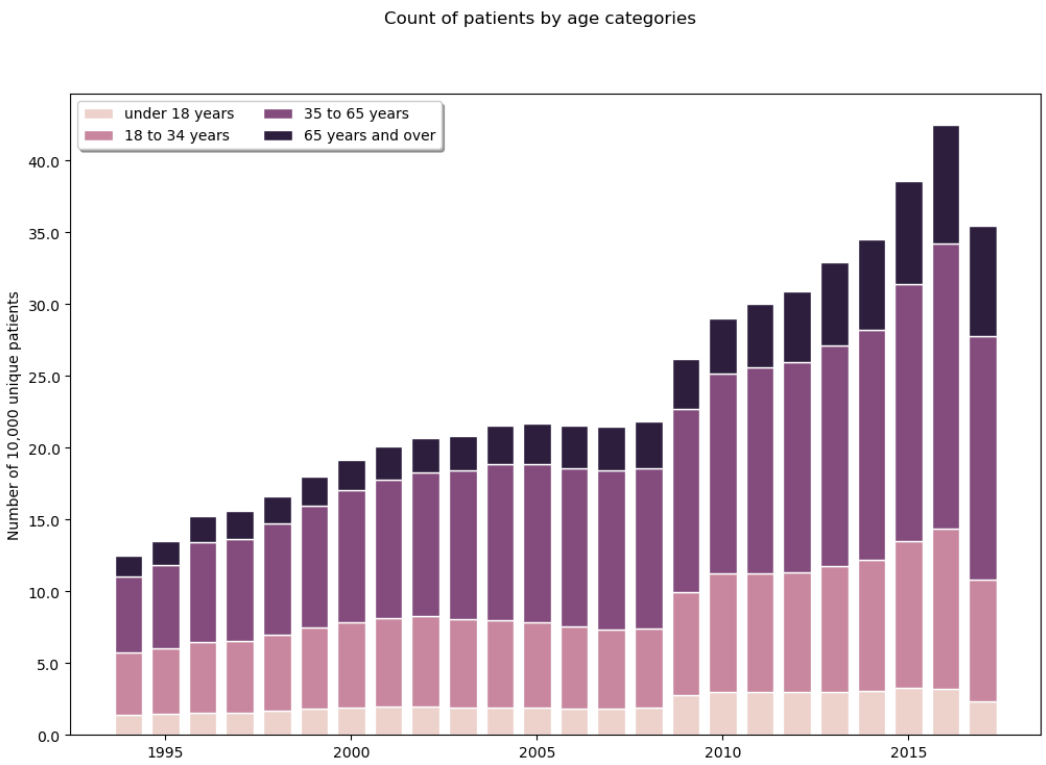


Figure 4-3: Annual summary of unique patient attendance by age-groups.
From 2009 onwards, marked increases in patient attendance were observed among adults ages 35 to 65. By the end of 2016, more than twice as many patients were provided care compared to 2009.

Between 2009 through 2016, the top MLC contained 20 ZCTA locations within Seattle King County and Snohomish neighborhoods as hotspots within the influx in patient attendance (Figure 4-4). Compared to trends in patient attendance between 2001 to 2008, the top 5 MLCs indicate King County and Snohomish County as the highest areas of increased total attendance and the top 25 MLCs show coverage for 460 of the 598 Washington ZCTA. The increase in patient attendance was 5.94 times as likely to be associated to ZCTA in MLC1 than outside of MLC1

over the 8 year period. The sequential deletion procedure identified MLC2 through MLC9 indicate King county and Puget Sound area, related to proximal distance from medical facilities in the UW Medicine network. MLC6 has visually larger areal coverage from ZCTA with lower population density, whereas MLC7 encompass ZCTA in multiple counties have a water barrier as the shortest route for access to UW Medicine. The relative risk diminishes until MLC17, indicating that the higher patient attendance areas to UW Medicine have already been represented within prior MLCs.

MLC	Duration	Relative risk compared to remaining areas	MLC size
1	8	5.94	20
2	8	4.27	20
3	8	4.07	20
4	8	3.77	20
5	8	3.47	20
6	8	2.93	20
7	8	2.40	20
8	8	2.16	20
9	8	1.91	20
10	8	2.40	19
11	8	2.07	20
12	8	2.14	20
13	8	2.42	19
14	8	2.47	20
15	8	2.77	20
16	8	2.30	19
17	8	5.28	18
18	8	2.97	19
19	8	4.29	19
20	8	3.97	19
21	8	2.97	6
22	8	2.71	18
23	8	1.95	20
24	8	2.57	8
25	7	1.66	16

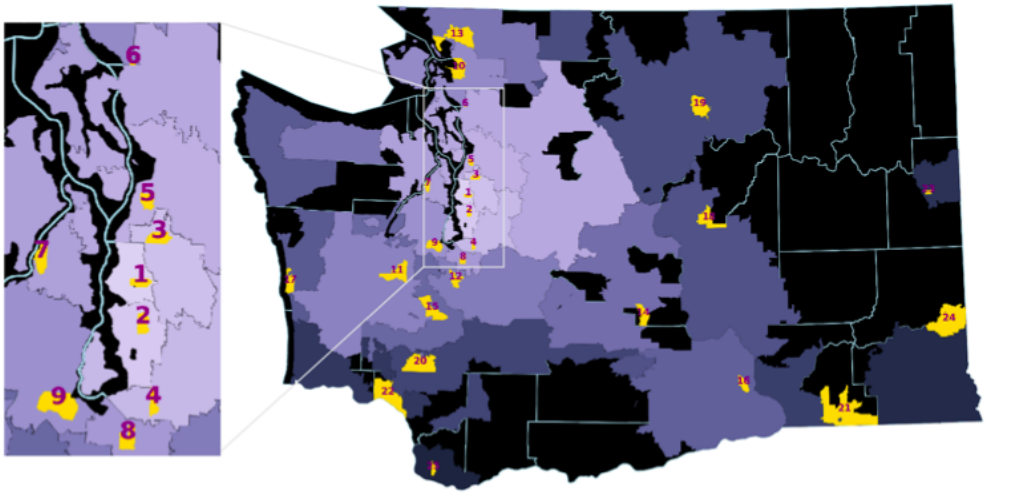


Figure 4-4: The top 25 most likely cluster of UW Medicine patient attendance between 2009 to 2016 among Washington ZCTA locations overlaid on Washington counties.
Here, the top 25 sequential identified MLC, where spatial windows exceeded expected proportions of attendance based on the population size, are shown with a darkening gradient. The ZCTA with the highest relative score within each sequential spatial scan are highlighted as yellow within their clusters. The relative risk descends until MLC17, where the ZCTA remaining reflect a lower expectation in patient attendance, considering the majority of high attendance areas have been excluded.

4.3.2 Use-case 2: Cross-sectional enrichment analysis

Figure 4-5 depicts the 9-year analysis for enrichments in patient diagnoses between 2009-2017, after multiple testing correction and evaluation of the confidence intervals, clustered

according to the 25-year analysis cluster organizations. Out of the 598 Washington ZCTA, "other" and 343 ZCTA had statistically significant associations with odds ratios indicative of enrichments and depletions patterns to at least one DxF in the 25-year analysis. Cluster detection found 5 DxF clusters and 10 ZCTA clusters as the optimal *k* clusters to distinguish the major variabilities observed among the odds ratios of significant associations. D3 or NEOPLASMS display depletions (blue) ranging down to $\log_2 OR=-1.42$ in Z9 and Z10 and enrichments (red) ranging up to $\log_2 OR=1.3$ in various Z1, Z3, Z4, Z5, Z7, and Z8 (Figure 4-5). A geospatial view of NEOPLASMS reveals enrichments with ZCTA located in the upper Puget Sound, Eastern Olympic peninsula, and along the Cascade mountain range (Figure 4-6A). In contrast, D3, D4, and D5, which are generally considered less frequent types of medical diagnoses, exhibit enrichments across ZCTA clusters, indicating greater proportions of patients received services for these medical needs and diagnoses than the background rates. ZCTA within Z1, Z3, Z4, Z5, and Z7 have higher rates of referral patients for particular diagnostic needs and procedures. The distribution of enrichments and depletions across DxF suggest that ZCTA within Z9, Z10, and a subset of Z6 may have higher proportions of patients receiving routine-care than in other locations, resulting in an accurate depiction of cancer and other rare diagnostic types as rare events.

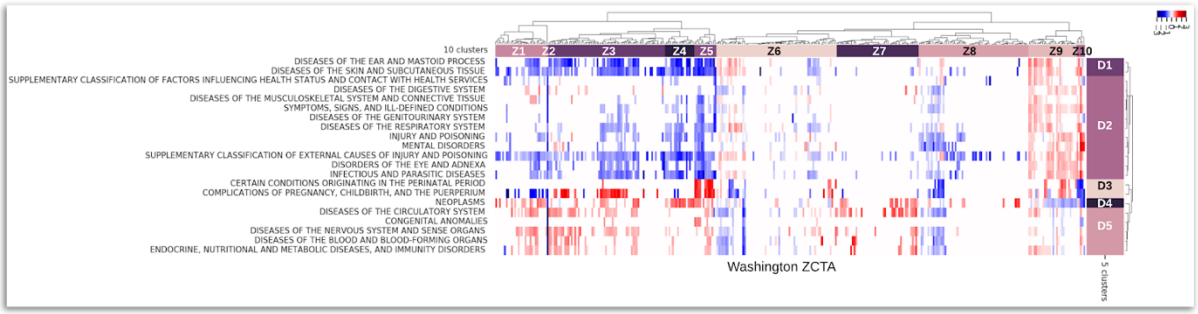


Figure 4-5: Odds ratios for statistically significant association between diagnosis code families and most recent zipcode on record.

10 clusters among the 344 Washington ZCTA (column) and 5 cluster among the 22 diagnosis code families suggest similarities within group over 25-years of UW Medicine patient diagnoses.

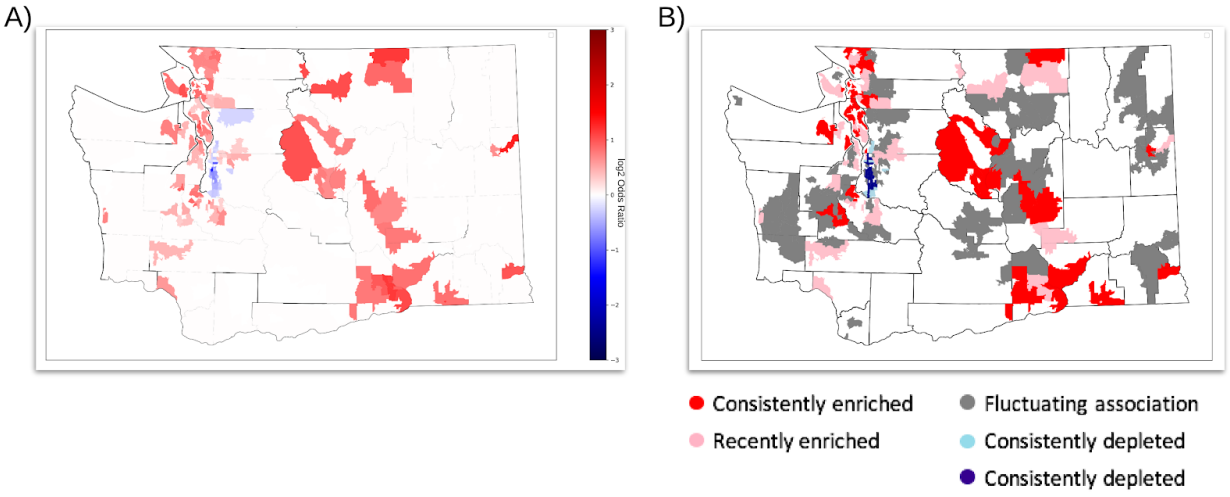


Figure 4-6: Statistically significant enrichment and depletion associations between Washington ZCTA and NEOPLASMS in A) the 2009-2017 time-frame and B) qualitative associations across time-frames.

ZCTA within MLC1 and MLC2 had depletion associations relative to the rates in surrounding areas, which were consistent across time-frames. Without examining the magnitude of the Odds ratios beyond a conclusion of enrichment, depletion and no association, various ZCTA identified in 2009-2017 were consistent findings (red/blue), recent statistical enrichments (pink/skyblue), or displayed fluctuations between time-frames (gray).

A number of DxF displayed statistical associations that varied between consistent findings, recent findings, and associations that fluctuated between time-frames. Figure 4-6B summarizes the five categories of associations with NEOPLASMS on a geospatial view across time-frames.

Visual observation can detect ZCTA and DxF clusters where the Odds Ratio associations hold consistent magnitude in the measure of association (Figure 4-7).

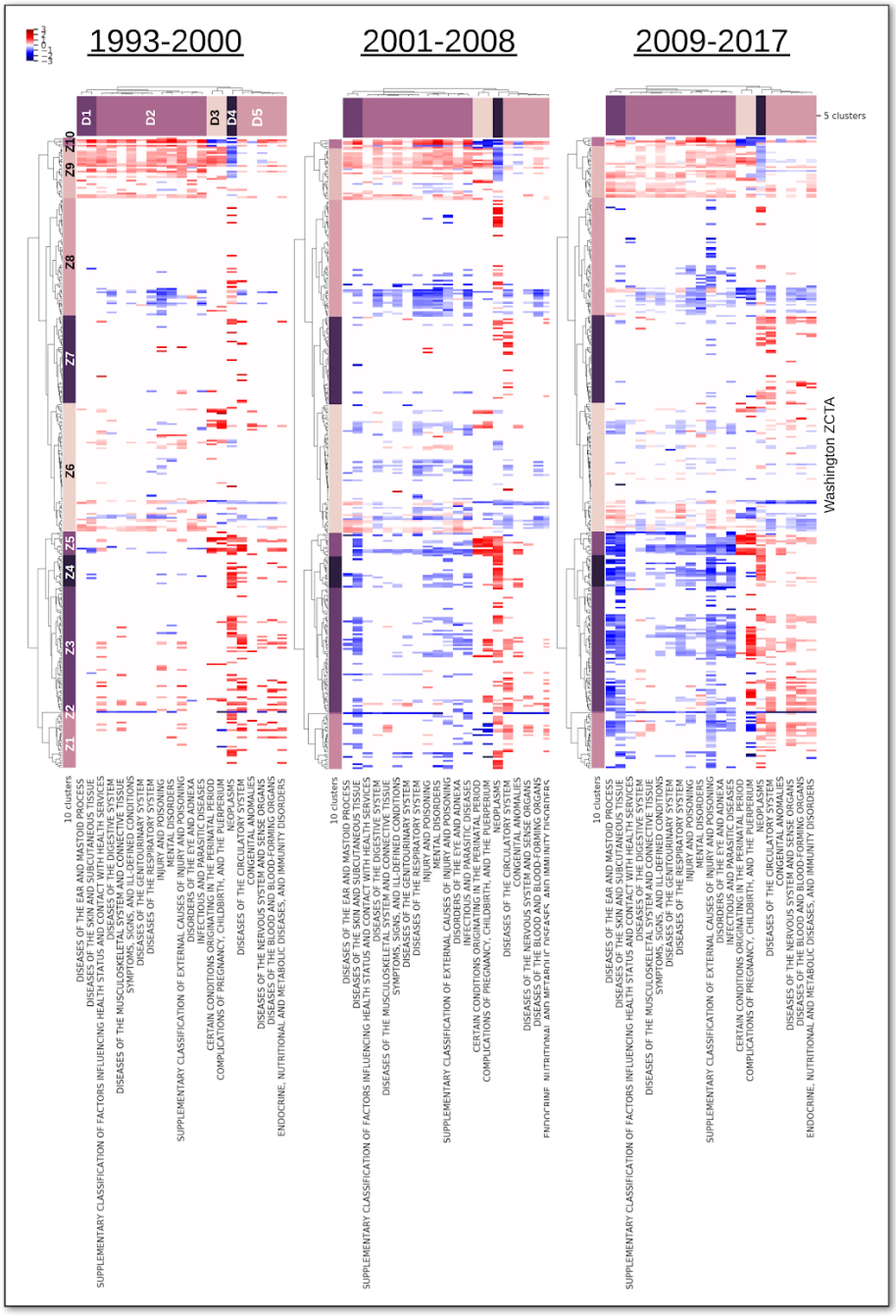


Figure 4-7: Heatmap of odds ratios for statistically significant association between diagnosis code families and Washington Census 2010 ZCTA for three segments of UW Medicine EHR history.

Each heatmap was biclustered according to the 25-year analysis cluster organizations.

Across the three time-frames, 1993-2000, 2001-2008, and 2009-2017, the total unique patients within each analysis has progressively increased from 600935, 824566, to 1146208, respectively. 286 enrichments and 144 depletions remained consistent pairwise findings of statistically significant associations. 438 enrichments and 871 depletions arose as recent pairwise findings in the 2009-2017 or since 2001-2008 timeframes where no statistical relationships were found in prior time-frames. 1335 associations fluctuated between statistical significance and directionality in the measure of association. Table 4-1 cross-tabulates these findings by DxF. As time progressed, Z1, Z3 and Z4 gradually transitioned towards depletion relationships for D1 and D2, indicating that medical diagnoses of those categories continued but at much lower proportions than with other locations (Figure 4-7). Various enrichment patterns in Z6 and Z8 fluctuated into depletion patterns, and vice versa. Various enrichment associations in Z9 and Z10 diminished towards non-statistical significance while enrichments increased over time in Z1, Z3, Z4, and Z5 for specialty care issues.

Table 4-1: Summary of pairwise associations across the three time-frames.

Consistent findings suggest statistical significance and similar directionality across time-frames. Recent findings indicate emergent statistically association of a single directionality in either the most recent or two most recent time-frames. Fluctuating associations display inconsistent statistical significance and/or directionality.

Diagnosis code families	Consistently enriched	Recently enriched	Fluctuating associations	Recently depleted	Consistently depleted
Diseases of the ear and mastoid process	10	17	64	76	1
Disease of the skin and subcutaneous tissue	18	11	61	109	6
Supplementary classification of factors influencing health status and contact with health services	5	11	89	23	2
Diseases of the digestive system	14	20	46	17	8
Diseases of the musculoskeletal system and connective tissue	11	15	54	23	2
Symptoms, signs, and ill-defined conditions	18	12	54	40	5
Diseases of the genitourinary system	6	11	41	27	2
Diseases of the respiratory system	7	8	66	61	8
Injury and poisoning	17	9	73	45	23
Mental disorders	16	8	77	34	13
Supplementary classification of external causes of injury and poisoning	21	12	78	131	3
Disorders of the eye and adnexa	14	11	55	53	2
Infectious and parasitic diseases	19	6	60	82	10
Certain conditions originating in the perinatal period	14	7	46	9	5
Complications of pregnancy, childbirth, and the puerperium	14	41	58	35	7
Neoplasms	38	47	97	13	25
Diseases of the circulatory system	24	53	88	22	9
Congenital anomalies	2	5	40	7	2
Diseases of the nervous system and sense organs	3	65	55	18	6
Diseases of the blood and blood-forming organs	3	26	59	19	3
Endocrine, nutritional and metabolic diseases, and immunity disorders	12	43	74	27	2
Total association by group	**286**	**438**	**1335**	**871**	**144**

4.3.3 Comparison between use-case 1 and use-case 2

The top 5 MLCs indicate that the major influxes in patient attendance relate to zip-code areas in King County and Snohomish County. Based on the depletion pairwise associations with NEOPLASMS, ZCTA in Z9 and Z10, and subsets of Z6 overlap with MLC1 through MLC4, where there is more influx of patients and UW Medicine has more diagnostic knowledge for a broader category of diagnoses in D1 and D2 (Figure 4-5). ZCTA enriched for NEOPLASMS in 2009-2017, areas outside of King County but along the Puget Sound (MLC5 through MLC13), Cascade mountain range (MLC16, MLC18, and MLC19), and Lewis and Cowlitz County (MLC20 and MLC22) saw more depletion associations in D1 and D2. The multiple time-frame approach would recognize many of these enrichments as consistent with the previous 16-years of patient diagnoses or new findings related to the 2009-2017 or since 2001-2008 (Figure 4-6B). While the spatial scan results would indicate influxes in patient attendance from each of these areas, medical diagnostic knowledge for patients associated to those areas are limited to specialty care categories in D3, D4, and D5 and statistically significantly less about D1 and D2 categories (Figure 4-7).

4.4 DISCUSSION

Recent efforts to address the rate of uninsured populations and access to healthcare also continue to fill the knowledge gaps about health in the population [82]. An influx in patient attendance can be observed in the UW Medicine EHRs coinciding with the timing of the Affordable Care Act of 2008. Use-case 1 applied spatial-scan to identify small area clusters associated with the recent influx, incorporating 460 Washington ZCTA in up to MLC25. MLC1 through MLC4 include 80 ZCTA in the King County and Snohomish County out of 598 ZCTA

within Washington, suggesting that the predominant areas of increased patient attendance to UW Medicine are potentially related to the urban and suburban areas of high population growth and proximal geographic distance from the health system. As UW Medicine is one of many health systems in the State of Washington, additional research is needed to consider the relative patient catchment areas and amount of recent knowledge about patients at a local level between neighboring health systems and community health clinics.

The way that the spatial scan analyses were configured may have unexplored influences on the interpretation of the relative risks. Although spatial scans configured for long duration associations may find long-term patterns, short duration approximations may provide more insights for emergency preparedness about recent attendance trends. Studies have used spatial scan to detect patient catchment areas configured for as short as 1-2 year duration clusters [95,96]. Separately, it is possible that the sequential deletion procedure reached a saturation point, where additional clusters merit alternative interpretations about the configuration, the resulting estimate error and the optimal *k* cluster size [104]. One simulation study has explored the role of statistical power and configuration of the Poisson spatial scan towards bias in the relative risk estimates of disease outbreak clusters [105]. More research is needed to understand the optimal configurations for cluster detection in patient attendance catchment area and other categories of diagnosis concern.

It is feasible that estimate error in the expected counts may originate from artifacts in the Census-based SAEP annual population estimates. SAEP interpolates intercensal ZCTA annual population estimates [102], wherein changes in US postal service operations may have occurred in response to recent population growth [106,107]. Discrepancies in ZCTA boundaries and exclusion of PO Box usage areas have been recognized as sources of uncertainty in the

representation of population size over space and time [106]. In Washington state, rural areas have polygon sizes that are too large for precise coordination whereas urban areas are subdivided into many small polygons and high commuter migration. SAEP does not account for cumulative migration in and out of Washington areas or the transient population that experience housing barriers, as the objective of SAEP is to estimate population size based on housing capacities in buildings [102]. Despite that these artifacts, Census-based products provide spatial polygons and population estimates where no other available resources are as comprehensive. More research is needed to improve geocoding to meaningful place-based insights relevant to neighborhoods and communities instead of zip-code or urban/rural characteristics, while account for commuting and migration patterns.

Discrepancies exist in the secondary-use of patient zip-code on-file, which is the preferred contact information for communications via mailing address. While it may be assumed that the mailing address relates with current patient residence, that is not always the case [106,107]. Patients may use PO Box numbers, provide no zip-code on-file, or use mailing addresses related to workplace and reasons other than residence. In the former two reasons, here, patients would be mapped to the "other" non-geographic label, where cross-sectional enrichment could still be performed but spatial scan would exclude these patients. For the latter reason, the difference in purpose such as relocation or homelessness cannot be determined as is. Zip-code on-file may be out-dated if the patient has not had recent contact with UW Medicine or provided contact information updates. Various circumstances may confound the idea that zip-code on-file indicates the patient general geographic location or that the zip-code area explains the rates in medical diagnoses [108,109]. Further studies should consider verifying reasons for zip-code

annotations or use alternative representations for the places where they spend significant amounts of their time.

The cross-sectional enrichment analysis highlighted general differences in UW Medicine patients by diagnosis and zip-code associations, but it does not qualify the knowledge about patients relative to the general population. With the current set of methods, we cannot account for the relative extent of knowledge on patients or populations related to a given zip-code compared to other health systems. For rare diagnosis phenomena like neoplasms, zip-codes exhibiting odds ratios indicating it is not rare can be observed more obviously as areas where patients are received through referral for particular care reasons. As such, knowledge about patient health in those areas is accepted as limited. However, some proportion of the patients associated with ZCTA in King County and Snohomish County may also be referred from other health systems. For various reasons such as socio-economics and insurance coverage, proximity to access healthcare, care seeking behavior or care options, patients may be provided the majority of their healthcare from another health system or community health network. Thereby, the majority of diagnostic knowledge about those patients reside with their primary care providers [97]. Some proportion of the population may continue to experience barriers to receiving healthcare. Integrating EHRs with public information about climatological features, environmental hazard, and indices for community vulnerabilities may distinguish priority areas for considering the relative recent knowledge about patients [7,10,110]. Future studies should explore metrics for cross-sectional baseline characterization across health systems and community health information systems that would be interpretable to those in the emergency preparedness sector. Such research should be wary of the potential for residential fallacies and gaps in recent knowledge.

Finally, although multiple hypothesis testing correction was incorporated, it is possible that false positives and false negatives persisted through the analyses. Across the 25 years of UW Medicine EHR usage, disruptions from software updates, changes in diagnosis coding practices, and retention of legacy EHRs may have introduced data gaps and annotation artifacts [85,86]. If patients had died outside of the health system, that information may not be documented within EHRs, a hurdle in computing prevalence metrics. The use of high-leveled diagnosis concepts may alleviate specific coding preferences at the terminal end of the ICD9/10 code structure, but it is not free of medical misdiagnosis errors or missing annotations for true cases [86]. Separately, Bonferroni correction is known to impose a conservative filter upon the alpha-threshold, where weak but positive associations may have been removed [111]. False-discovery rate or False coverage rate could be viable alternatives to correct for multiple hypothesis testing among *p*-values or confidence intervals, respectively. There was some congruence or consistency in enrichments between time-frames at the high-leveled concepts. Though it was not explored here, future analyses could incorporate more hierarchical relationships in ICD9/10 diagnosis coding for mid-level code enrichments or the tentative crosswalk with the ICD11CM ontology structure. Additional research could examine enrichment patterns with more granular time-frames and incorporate additional descriptors for patient demography, validation of diagnosis code annotations, and triangulating logic produced or confirmed over multiple visit clinical notes or diagnostic tests.

4.5 CONCLUSION

In this study, we incorporate two analytical methods to explore the spatio-temporal relationships related to recent influxes in patient attendance. Using spatial scan statistics, we

identified the most likely clusters of ZCTA attributable to the influx of patient attendance. A variation to spatial scan statistic was able to find subsequent hotspot clusters as the next most likely cluster of where patients attendance may be attributable. As a separate use-case, the cross-sectional enrichment analysis was able to establish statistically significant enrichment and depletion associations between individual Washington zip-code geographies and patient diagnostics. Comparison across multiple time-frames of analysis distinguish associations that were consistent, recent, or fluctuating findings. Though comparisons across the use-cases were limited by the different use-case objectives, findings from the multiple time-frame comparison could potentially distinguish where findings have been consistent despite the influx of patients and where enrichment patterns may be recent effects. Towards emergency preparedness, these methods could gauge not only the baseline count of patients but also the relative proportions in care and diagnostic services provided. The spatial and temporal window of analysis could be expanded as data flow and patient influx progresses, but higher resolution than zip-code and comparisons across health systems is needed to estimate how much patient health is proportional to population health.

CONTRIBUTIONS FROM CHAPTER 4

In this chapter that is focused on Aim 3, I incorporate epidemiological methods to explore space-time clusters and associations between geographical trends in patient population healthcare needs using diagnosis information from EHRs. This aim was separated into two research use-cases seeking 1) to identify geographic distributions for recent influx in patient attendance and 2) to identify the statistical enrichments between patient zip-code on-file and categories of medical diagnoses. Through spatial scan, I explored the top 25 most-likely clusters of Washington Zip Code Tabulated Areas (ZCTA) contributing to recent influxes in patient attendance and where those clusters exist. I showcase a series of cross-sectional enrichment analyses using up to 25-years of de-identified UW Medicine patient diagnosis records to detect geospatial health trends and enrichment patterns among individual Washington ZCTA.

I demonstrate analyses to gain insights about the geographic distribution of health needs among patient population. Multiple methods were applied within a HIPAA-aligned virtual machine environment similar to the docker environment in Aim 2. The research use-cases could potentially inform population health and disaster preparedness endeavors as it provides period proportions of medical diagnoses and care provided to patients, potentially informative as a method for baseline characterization of diverse patient health needs by individual geographic units. These methods could tentatively scale for less and more granular units of analysis in space and time as well as incorporate more features to explain patient health, subcategories in medical diagnoses of increased vulnerability, and avenues to conduct such research while protecting patient privacy and confidentiality.

CHAPTER 5: Conclusion

To conclude this dissertation, I summarize the contributions in fulfillment of the dissertation aims by reviewing the advances in knowledge (Section 5.1) and acknowledging the limitations and opportunities for future work (Section 5.2).

5.1 SUMMARY OF CONTRIBUTIONS

This body of work comes at a time when growing global populations, changing climate, land use and urban, and extremes events, and natural hazards in environmental systems can lead to a broad variety of health outcomes related to overwhelmed health systems, dynamic geographic-based needs for customized information and limited disaster management capacities. Various scientific domains and research consortia are seeking to better understand climate and environmental anomalies and patient diagnostics related to natural disasters, but with limited knowledge transfer and often without the capacity for data integration.

Three dissertation aims separately address these issues with the goal of enabling data-driven inquiries of population health for disaster preparedness, as shown in Figure 5.1. The three aims are related by contributing methods, tools, and knowledge for improved future population health models that integrate environmental observations with electronic health records and develop software designed for collective usability by earth, health, and disaster researchers.

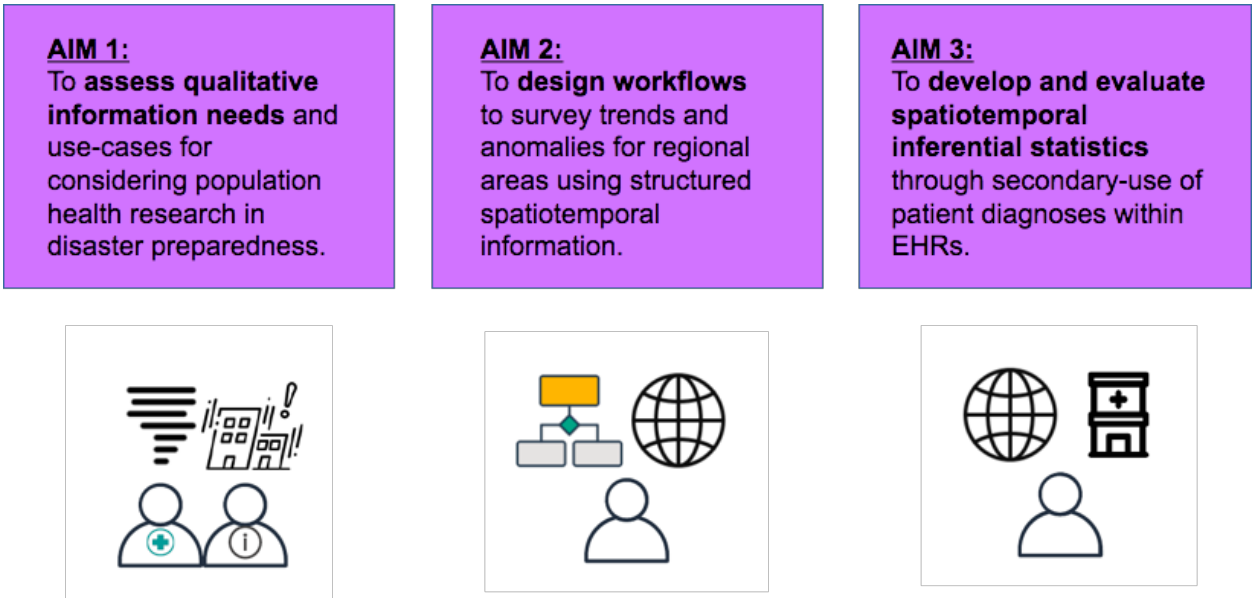


Figure 5-1. Dissertation aims.

5.1.1 Aim 1 summary

For aim 1, I interviewed 15 population health researchers using a mixed-method approach for needs assessments, including a semi-structured interviews followed by think-aloud closed card sorting. In the interviews, I asked open-ended questions about their research training and focus, the barriers and facilitators they have experienced in collaborative research and instances of data and technology adoption. I also asked about their readiness for future disasters and what they would do to consider the health of a population if anticipating a hurricane or flood in the near future. Following the interview, I asked the participant to perform a closed card sort activity in which they were asked to think-aloud as they rated information cards for their usefulness for research and what groupings and use-cases they may be envisioned to serve.

Analysis of the interview responses found major differences in strategy for future disasters, generally related to whether they were planned and prepared for disaster management roles and their primary occupational appointment. Thematic analysis found that seven barriers in past

experiences with collaborative research and in data and technology adoption, where the barriers are more challenging in a disaster situation. In contrast, six facilitators have been beneficial factors in conducting population health research, factors that should be incorporated into the design of tools for population health research. Card sorting identified a gradient in which information cards were perceived as *most* versus *least* useful for research use. Although heterogeneous perceptions of information use were observable, thematic analysis of the card sorts found 15 research use-cases discussed by 5 or more participants out of 78 use-cases detected. Our results show 15 use-cases which prioritize population health researcher information needs: (1) produce baseline characterizations to address gaps about population health before disasters, (2) modes for real-time data collection and communication for situational awareness of community health and environmental status, and (3) options to identify and train collaborators of diverse capacities.

Overall, population health research users experience profound barriers in interacting with data products during hurricanes and floods, especially given the usual gaps in baseline knowledge. Users need information tools designed for transferable applications (e.g., offline scenarios with loss of telecommunication), proof of best-practices, and having low barriers in learning the tool. Resources to build cross-functional technical teams are needed throughout disasters, and should be coupled with interactions that develop supportive collaboration engagements. Collaboration is central to the strategies for population health research in future disasters.

In Chapter 2, this work contributes to the population health field as follows:

- The needs assessment identified key facilitators, barriers, and current strategies to approach research about the state of population health if anticipating a hurricane or flood up to 10 days in the near future.
- Through analysis of the card-sorts, I identified the 15 priority use-cases, the most commonly discussed use-cases, as recommendations to improve health research within disaster preparedness.
- The heterogeneous relationship in how population health researchers perceived the usefulness of different hurricane and flood disaster-related information and their potential function to address a research use-case.

5.1.2 Aim 2 summary

For aim 2, I collaborated with Geoscientists (e.g. hydrologist, ecohydrology, geology, atmospheric science, civil and environmental engineers, and computational hydrology research software developers) in the development of the Observatory for Gridded Hydrometeorology (OGH) python library. We recognized that spatially-distributed time-series data provide a range of interdisciplinary opportunities to integrate environmental modeling into other data research efforts. However, standardized tools for acquiring interpolated hydrometeorological data into analyses have not been readily available for watershed scale research. I used scenario-based design and iterative prototyping to compartmentalize use-cases to accomplish tasks required to interact with gridded data sets. During the design, architecture, and application of OGH, four commonly practiced use-cases with gridded time-series data at watershed scales were used to guide the order of operations towards the fulfillment of the use-case. Our approach involved annotating metadata to make gridded data products discoverable and usable within the software,

enabling interoperability and reproducibility of models that use the data. The resulting work is published as an open source python library and publically accessible via python or anaconda to support users to fetch, manage, analyze, and visualize distributed data processed from regional and continental-scale gridded hydrometeorology products.

Overall, this work addresses a critical software gap for interactions with small- and large-scale geospatial data. Our prototype was designed for use with the Jupyterhub instance hosted by HydroShare. It demonstrates the feasibility to interact and do analysis with large-scale geographic data within dockerized virtual environments.

My contributions towards the use geospatial data for analysis and interpretations by research users (Chapter 3):

- The minimum annotation criteria for gridded data products serve as a set of design recommendations to simplify the annotation process and expedite the communication and usage of gridded data to new data users.
- The workflow and use-case design within a dockerized virtual environment demonstrates an approach to enable data access, analysis of geographic environmental data sets, and replicable process automation within cloud computing environments.
- The Observatory for Gridded Hydrometeorology open-source python library provides the public with access to the pythonic operations and solutions developed to enable user interactions with gridded data products.
- The Observatory for Gridded Hydrometeorology open-source python library is used in classroom education and has already been cited in three peer-reviewed journal articles within the first year of paper publication by research teams not associated with this dissertation.

5.1.3 Aim 3 summary

For aim 3, I conducted two research use-cases with patient geographic information. In one use-case, I conducted a local spatial scan to characterize hotspot clusters (or Most Likely Cluster) among the recent UW Medicine patient attendance. In another use-case, I analyze for bivariate associations between each category of patient medical diagnoses (also referred as "diagnosis code family", DxF) and patient zip-code on-file, which I mapped to Washington zip-code tabulated area (ZCTA) as the geographic surrogate for US Postal Service 5-digit zip-code. The patient data were separately analyzed as the 25-year analysis and three time-frames of analysis, the earlier two are 8-year time-frames and the latter a 9-year time-frame. The top most-likely cluster includes 20 North King County and Snohomish county ZCTA locations observing the greatest proportion increase of patient attendance between 2009 to 2016. The 25-years cross-sectional analysis of patient attendance indicates 5 clusters among the DxF, indicating consistent differences in medical services provided relative to identifying with certain geographic locations. The 344 ZCTA with enriched associations at least one DxF can separate into 10 ZCTA clusters, indicating marked differences between groups in terms of medical diagnosis and care provided. The ZCTA clusters reflect differences in populations that seek routine care versus specialty care referrals. Comparison between the three time-frames of analyses show various similarities in enrichment and depletion patterns, but the fluctuating associations merit further research and considerations for patterns in diagnosis and approaches to determine thresholds for statistical significance given multiple testing scenarios.

Overall, this work showcases the capacity to abstract geospatial-temporal insight from patient attendance and diagnosis information. The approach incorporates multiple statistical reasoning to

test the presence of possible hotspot clusters and test hypothetical associations between patient diagnosis trends with regards to space- and time-relationships. This work was performed within a HIPAA-aligned enclave virtual computing environment and demonstrates the potential to use electronic health records to address geospatial health use-cases while conducting in accordance to secure patient privacy.

In support of the dissertation aims, my contributions in this work focus on patient information from electronic health records to address population-scale questions that could inform disaster preparedness (Chapter 4):

- The application of expectation-based Poisson spatial scan with patient geospatial information demonstrates a feasible procedure to account for relationships in space and time for identifying hotspot clusters, where patient health may be represented within electronic health records.
- The cross-sectional enrichment analysis approach incorporates multiple parameters for the statistical evaluation of association between place and categories of diagnosis.
- I demonstrate that such operations can be accomplished with a HIPAA-compliant enclave virtual machine environment.

5.2. LIMITATIONS AND FUTURE WORK

Although the composite body of work within this dissertation expands knowledge across interdisciplinary subjects (i.e., population health research in disasters, user interactions to extract insights from big data in hydrological models, geographic modeling with de-identified patient biomedical information), there are a number of limitations within each study and across studies.

Overall, two different research stakeholders were the key partners to guide design and analysis. Geoscientists (e.g. hydrologist, civil, environmental and water resource engineers) confer different design criteria and goals within the resulting product than population health researchers (Chapter 2). In Phuong et al., (2019; Chapter 3), our methods were based on design needs, use-cases and objectives of watershed scale physical process understanding (1-100 sq km areas) with iterative prototyping oriented on data processing. This contrasted to the Chapter 2 methods with one-on-one interview sessions and limited data collection with each population health researcher. With each stakeholder, there may be information gaps related to the choice of research method used, and the findings may vary beyond the scenario and context of the study. The following sections will cover the limitations specific to each aim.

5.2.1 Aim 1 limitations and future work

In Aim 1, the goal of the recruitment strategy was to gather a diverse sample of population health researchers. However, I cannot discount the possibility of sample biases which may limit the generalizability of the findings. The responding participants were represented largely by *Epidemiologists* and *Environmental health researchers* and the majority described their research as related to *All-hazards emergency management*, *Exposure hazard agents*, *Health outcomes related to disasters*, and *Surveillance systems*. I note that these researcher types and subject matter expertise refer to broad disciplines with various methodological focus that may not be represented within the participant pool. Only two participants interviewed resided in Puerto Rico, a geography recently affected by hurricanes and floods, and we were not able to reach researchers within non-profit and local health jurisdictions within the recruitment period. Due to concerns of sample biases and small sample size, we could not explore relationships researcher

needs and barriers stratified by researcher role, subject matter of focus, geography of residence. Future studies should consider alternative recruitment strategies for a balanced representation.

The card sorting activity had a few limitations associated with the research instruments used. The 31 cards were developed from review literature about health vulnerabilities from hurricanes and floods. Many participants had difficulty interpreting the time constraint cards with regards to the context of disaster preparedness. Participants nominated another 30 cards as information they would like to have in addition. This suggests that there were some design issues with the information cards.

During the card sorts, participants frequently used pronouns (e.g.,'this' or 'those') and short names (i.e., 'time constraint' instead of 'time constraint (annual)') to refer to cards. From the audio transcription, some references remain unclear. Similarly, participants who performed card sorting online using OptimalSoft took issue with the manual nature of the software interaction. It is unclear whether the software had introduced stress or distraction, but participants who used the physical cards engaged in think-aloud much more fluidly. I recommend that future studies conduct card sorting using physical cards on a table-top and to employ video recording to get a clear record of the card sort progression.

In future studies, in addition to the recommendations already mentioned, a broader approach for recruitment and multiple methods of data collection should be considered for gathering information needs from population health researchers. A balanced representation among subject matters should be furthered explored, but continued approaches for diverse sampling may find additional disciplines not yet represented within this study. Combining semi-structured interviews and card sorting has been very successful as complementary methods for individual participants to engage with information needs and usage in different ways. Including focus

groups and community meetings could strategically gather consensus opinions of priorities, though the main disadvantage to this method is that of dominating voices. Certain participatory design activities could help mediate these imbalances by written ideation exercises or role-playing in scenarios of collaboration, like pairing tools and expertise and accounting for missing expertise towards local research preparedness. Finally, as there were a number of nominated information concepts learned in Aim 1, I want to follow-up to determine readable or preferred concept labels that could be used to map existing or prospective data resources for intuitive user access.

5.2.2 Aim 2 limitations and future work

In Aim 2, early design choice was made in creating the metadata criteria for a baseline annotation of the data sets. Working with the hydrometeorology partners, we assessed for consistency in metadata and structural similarities between seven gridded hydrometeorological data sets. These data sets used ASCII data format for data storage, providing a tractable data format that requires comparably low disk storage burden. That said, the advantages this confers was built into our metadata design and data processing strategies, where ASCII data sets had a standard data schema that could be recognized for increments of time, space, and features that are characteristic to hydrometeorological modeling. The disadvantage was the annotation step and metadata criteria may not transfer towards other data formats. ASCII data sets are also notably easier to interact with than NetCDF data sets, a recent standard geospatial data representation with a built-in metadata configured to more closely resembles hypertext markup language (HTML). NetCDF is one of many data standards with new ones being developed to

navigate the limitations of the existing options. Metadata design that is transferable between data formats is a reasonable concern as the alternative follows the one-format-one-standard approach.

There are some noteworthy limitations to consider with regards to parallelization algorithms as it is unclear how to balance computing performance against changing computing hardware, operating systems, and new formats for data representation. I have not explored the limitations between computing capacity and maximum allowable information complexity. That is, it remains to be explored when processing performance becomes suboptimal while accounting for the simultaneous number of files, numbers of features, and the resolution of time and space. Moreover, I have not explored how to translate operations for structured geographic units to uneven, semi-structured geographic parcels.

The system as a whole has not been explored for the difficulties towards usability and adoption. My research team and I chose to use HydroShare, Jupyter notebooks, Python programming, Github, and ESRI shapefiles to leverage the broad adoption that these tools have already gathered. However, with this many software as part of the approach, people who are not programmers or who are not familiar with any one of these components may experience a profoundly challenging learning curve. While the python library Observatory for Gridded Hydrometeorology has gathered attention and demand, it is not clear how the different user experiences may have led to adoption or rejection. Thus far, I have relied on direct communications of software errors or issue reporting from users through github, though this should be recognized as a more advanced set of users. I recommend that future work incorporate user experience evaluations to consistently consider the various components of the Technology Acceptance Model and develop formative assessments for further fixes and developments.

Preferably, these evaluations could occur by multiple methods to efficiently reach users who may get lost looking for help online.

In future studies, in addition to the recommendations already noted, I would also want to consider alternative visualization options. Javascript-based tools like Tableau, Shiny.io and CalEnviroScreen incorporate interactive geographic platform, where the user can easily load or select layers and subsets of information. Similarly, 3D-visualizations and virtual simulations open a new realm of possibilities to interact with data and what it means to any place. These technologies still have performance limitations towards high-resolution information, but it would enable a broader audience of users to engage with geographic tools without needing foundations in programming.

5.2.3 Aim 3 limitations and future work

In Aim 3, the approach used patient diagnosis information and methods to compute spatial health trends and inference of spatial-temporal effect clusters. Use of patient health information in research carries a number of ethical sensitivities and must remain compliant with HIPAA guidelines to protect patient privacy. In the Institutional Review Board (IRB) application for Aim 3, the goal of the study was to consider spatial health trends and the decision to use postal service zip-codes was to reduce the potential for patient re-identification, reduce dimensionality to broad units of place, and enable integration with data sets focused on social determinants of health. As an additional constraint to reduce risks of re-identification, pairwise tests with a numerator below 20 were excluded. Moreover, ZCTA are decadal Census products and zip-codes are units for postal service operations that change over time. Prior studies have highlighted erroneous Census ZCTA polygon extents in recognizing local environments, reporting varying levels of uncertainty

and mismatch between ZCTA, zip-codes, and census block associations [106]. For these reasons, the study and geographic components at zip-code or ZCTA to understand population demography are subject to design errors and objective biases, where population size and polygon shape can distort the understanding of population density and land use. Alternative units for geocoding to enable inferences of patient diagnosis trends at scales that are more meaningful for public health and community emergency preparedness.

Secondary analysis of patient diagnosis information are also subject to the disadvantages of the clinical primary data collection process. Electronic Health Records (EHRs) are tools for documenting care provided. The objective is not to document every observation unless it has relevance towards patient care. Though the use of high-level categories in medical diagnoses may alleviate specific coding practices at the terminal end of the ICD9/10 code structure, but it is not free of medical misdiagnosis errors or missing annotations of true cases. I do not include severity, early or late diagnosis, and it is unclear if some diagnoses and medical anomalies are not represented with ICD9/10 tree ontologies. It cannot be determined whether the use of a single or combination of diagnosis codes reflect the same diagnostic and disease definition or different phenotypic phenomena. Over the 25-years of patient diagnosis information, transitions between EHR platforms and diagnosis coding practices may have introduced period where coding annotations are suspect for low quality. I analyze for spatial health trends using high-leveled categories of diagnostics to encompass general categories by the ICD9 and ICD10 structure, but I acknowledge that the clinical decision making towards disease diagnosis and management can influence interpretations in this retrospective analyses. Future studies should consider methods to verify the code use, triangulating logic produced or confirmed over multiple visit clinical notes or diagnostic tests.

Various Census-based products were used to represent geographic boundaries and annual population estimates. The Washington State Office of Financial Management Small Area Estimate Program (SAEP) annual population estimates were used to calculate the spatial scan statistic, where I took a standard assumption that the estimates are a reasonable reflection of the population size in residence. However, SAEP interpolates for intercensal population estimates based on housing capacity, and it may not provide an accurate estimate of the true population, account for cumulative migration in and out of Washington area, or the transient population that experience housing barriers. For patients without a place of residence or zip-code on-file, I assigned them to zip-code value of 'other', a non-geographic label. This action retained the patients within the cross-sectional enrichment analysis, but they would be excluded from the spatial scan statistics as the label does not have geographic relevance for quantitative spatial clustering. Zip-code on-file should be interpreted as the hospital's means to contact the patient via mailing address. It is not always the case that the mailing address refer to their place of residence, because patients may use the mailing address of their workplace or reasons other than temporary or permanent residence. To avoid misinterpretations of residential fallacy, the methods used are not intended to make claims of causality, but rather to understand diagnosis proportions within certain time-frames within a zip-code as opposed to the diagnosis proportions with patients associated with elsewhere.

In future studies, several operations could be done to incorporate more information on patients, information about their facilitators and barriers in seeking healthcare, and use information beyond one health system. Patients have multiple options to seek care from other health systems and provider networks in the United States, provided they have proximity and access to care. Barriers to care such as uninsurance remains an issue. It is unclear how such

factors are or are not being documented, but it could be incorporated into a pairwise analysis to understand the differential rates by place. In addition, data sets that report the social vulnerability index or social deprivation index can serve as a proxy for the prevalence of factors related to social determinants of health. As ICD11 and Census 2020 are looming, these methods provide a means to compare baseline estimates in this 2019 analysis with the prospective coding structure and new geography in 2020. The present set of methods incorporates the knowledge from a single health system. While that is currently limited, new technology in HL7-FHIR could bridge the access to patient health data to gather the perspectives from multiple EHRs. Similar to a syndromic surveillance approach, a synthesis across health systems could identify not only hotspot clusters but the coldspot clusters where there is a paucity of knowledge about patient health.

REFERENCES

1 Schiff GD, Young QD. You Can't Leap a Chasm in Two Jumps: The Institute of Medicine Health Care Quality Report. *Public Health Reports* 2001;**116**:8.

2 Kindig D, Stoddart G. What is population health? *American journal of public health* 2003;**93**:380–3.

3 Kindig DA. Understanding Population Health Terminology: *Understanding Population Health Terminology*. *Milbank Quarterly* 2007;**85**:139–61. doi:10.1111/j.1468-0009.2007.00479.x

4 Kindig DA. A Population Health Framework for Setting National and State Health Goals. *JAMA* 2008;**299**:2081. doi:10.1001/jama.299.17.2081

5 Mennemeyer ST, Menachemi N, Rahurkar S, *et al.* Impact of the HITECH Act on physicians' adoption of electronic health records. *Journal of the American Medical Informatics Association* 2016;**23**:375–9. doi:10.1093/jamia/ocv103

6 Leaning J, Guha-Sapir D. Natural Disasters, Armed Conflict, and Public Health. *New England Journal of Medicine* 2013;**369**:1836–42. doi:10.1056/NEJMra1109877

7 Cutter SL. The landscape of disaster resilience indicators in the USA. *Natural Hazards* 2016;**80**:741–58. doi:10.1007/s11069-015-1993-2

8 Maini R, Clarke L, Blanchard K, *et al.* The Sendai Framework for Disaster Risk Reduction and Its Indicators—Where Does Health Fit in? *International Journal of Disaster Risk Science* 2017;**8**:150–5. doi:10.1007/s13753-017-0120-2

9 Lowe D, Ebi K, Forsberg B. Factors Increasing Vulnerability to Health Effects before, during and after Floods. *International Journal of Environmental Research and Public Health* 2013;**10**:7015–67. doi:10.3390/ijerph10127015

10 Cutter SL, Ash KD, Emrich CT. The geographies of community disaster resilience. *Global Environmental Change* 2014;**29**:65–77. doi:10.1016/j.gloenvcha.2014.08.005

11 Lloyd CT, Sorichetta A, Tatem AJ. High resolution global gridded data for use in population studies. *Scientific Data* 2017;**4**. doi:10.1038/sdata.2017.1

12 Miller A, Yeskey K, Garantziotis S, *et al.* Integrating Health Research into Disaster Response: The New NIH Disaster Research Response Program. *International Journal of Environmental Research and Public Health* 2016;**13**:676. doi:10.3390/ijerph13070676

13 Rosen J, Miller A, Hughes J (Chip), *et al.* National Institute of Environmental Health Sciences Worker Training Program: Perspectives on the Health and Safety of Workers, Volunteers, and Residents Involved in the Cleanup and Rebuilding of New York City

Housing Damaged by Hurricane Sandy. *Environmental Justice* 2015;**8**:105–9. doi:10.1089/env.2015.0008

14 Lindsay JR. The Determinants of Disaster Vulnerability: Achieving Sustainable Mitigation through Population Health. *Natural Hazards* 2003;**28**:291–304. doi:10.1023/A:1022969705867

15 Simpson CL, Novak LL. Place Matters: The problems and possibilities of spatial data in electronic health records. *AMIA Annual Symposium Proceedings* 2013;**2013**:1303–11.

16 Greenough PG, Lappi MD, Hsu EB, *et al.* Burden of Disease and Health Status Among Hurricane Katrina–Displaced Persons in Shelters: A Population-Based Cluster Sample. *Annals of Emergency Medicine* 2008;**51**:426–32. doi:10.1016/j.annemergmed.2007.04.004

17 Lane K, Charles-Guzman K, Wheeler K, *et al.* Health Effects of Coastal Storms and Flooding in Urban Areas: A Review and Vulnerability Assessment. *Journal of Environmental and Public Health* 2013;**2013**:1–13. doi:10.1155/2013/913064

18 Quinlisk P, Jones MJ, Bostick NA, *et al.* Results of Rapid Needs Assessments in Rural and Urban Iowa Following Large-scale Flooding Events in 2008. *Disaster Medicine and Public Health Preparedness* 2011;**5**:287–92. doi:10.1001/dmp.2011.82

19 Alderman K, Turner LR, Tong S. Floods and human health: A systematic review. *Environment International* 2012;**47**:37–47. doi:10.1016/j.envint.2012.06.003

20 O'Sullivan TL, Kuziemsky CE, Toal-Sullivan D, *et al.* Unraveling the complexities of disaster management: A framework for critical social infrastructure to promote population health and resilience. *Social Science & Medicine* 2013;**93**:238–46. doi:10.1016/j.socscimed.2012.07.040

21 Gall M, Nguyen KH, Cutter SL. Integrated research on disaster risk: Is it really integrated? *International Journal of Disaster Risk Reduction* 2015;**12**:255–67. doi:10.1016/j.ijdrr.2015.01.010

22 Arctur D. Harvey Flood Data Collections. Arctur, D. (2018). Harvey Flood Data Collections, HydroShare, https://doi.org/10.4211/hs.12e69ee668124fdf833b29b5167e03c3

23 Arctur D. Irma Flood Data Collections. https://doi.org/10.4211/hs.db5883d16e874ee3b7edd666dbad7d03

24 Bandaragoda C, Phuong J, Leon M. Hurricane Maria 2017 Collection. http://www.hydroshare.org/resource/97a696e7202d4ca98349a0742a725451

25 Lober WB. Roundtable on Bioterrorism Detection: Information System-based Surveillance. *Journal of the American Medical Informatics Association* 2002;**9**:105–15. doi:10.1197/jamia.M1052

26 Johanning E, Auger P, Morey PR, *et al.* Review of health hazards and prevention measures for response and recovery workers and volunteers after natural disasters, flooding, and water damage: mold and dampness. *Environmental Health and Preventive Medicine* 2014;**19**:93–9. doi:10.1007/s12199-013-0368-0

27 Maas P. Facebook Disaster Maps: Aggregate Insights for Crisis Response & Recovery. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*. Anchorage, AK, USA: : ACM Press 2019. 3173–3173. doi:10.1145/3292500.3340412

28 Revere D, Turner AM, Madhavan A, *et al.* Understanding the information needs of public health practitioners: A literature review to inform design of an interactive digital knowledge management system. *Journal of Biomedical Informatics* 2007;**40**:410–21. doi:10.1016/j.jbi.2006.12.008

29 DeCuir-Gunby JT, Marshall PL, McCulloch AW. Developing and Using a Codebook for the Analysis of Interview Data: An Example from a Professional Development Research Project. *Field Methods* 2011;**23**:136–55. doi:10.1177/1525822X10388468

30 Ando H, Cousins R, Young C. Achieving Saturation in Thematic Analysis: Development and Refinement of a Codebook. *Comprehensive Psychology* 2014;**3**:03.CP.3.4. doi:10.2466/03.CP.3.4

31 Paul CL. A Modified Delphi Approach to a New Card Sorting Methodology. 2008;**4**:24.

32 Lindell MK, Prater CS. Assessing Community Impacts of Natural Disasters. *Natural Hazards Review* 2003;**4**:176–85. doi:10.1061/(ASCE)1527-6988(2003)4:4(176)

33 Braun V, Clarke V. Using thematic analysis in psychology. *Qualitative Research in Psychology* 2006;**3**:77–101. doi:10.1191/1478088706qp063oa

34 Topf M. Three estimates of interrater reliability for nominal data." Nursing research (1986). *Nursing research* 1986;**35**:253–5. doi:http://dx.doi.org/10.1097/00006199-198607000-00020

35 McHugh ML. Interrater reliability: the kappa statistic. *Biochemia Medica* 2012;:276–82. doi:10.11613/BM.2012.031

36 Holden RJ, Karsh B-T. The Technology Acceptance Model: Its past and its future in health care. *Journal of Biomedical Informatics* 2010;**43**:159–72. doi:10.1016/j.jbi.2009.07.002

37 Fraley C. How Many Clusters? Which Clustering Method? Answers Via Model-Based Cluster Analysis. *The Computer Journal* 1998;**41**:578–88. doi:10.1093/comjnl/41.8.578

38 Salvador S, Chan P. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: *16th IEEE International Conference on Tools with Artificial Intelligence*. Boca Raton, FL, USA: : IEEE Comput. Soc 2004. 576–84. doi:10.1109/ICTAI.2004.50

39 Lillibridge SR, Noji EK, Burkle FM. Disaster assessment: The emergency health evaluation of a population affected by a disaster. *Annals of Emergency Medicine* 1993;**22**:1715–20. doi:10.1016/S0196-0644(05)81311-3

40 King AD, Alexander LV, Donat MG. The efficacy of using gridded data to examine extreme rainfall characteristics: a case study for Australia. *International Journal of Climatology* 2013;**33**:2376–87.

41 Gampe D, Ludwig R. Evaluation of gridded precipitation data products for hydrological applications in complex topography. *Hydrology* 2017;**4**:53.

42 Ledesma JL, Futter MN. Gridded climate data products are an alternative to instrumental measurements as inputs to rainfall–runoff models. *Hydrological Processes* 2017;**31**:3283–93.

43 Henn B, Newman AJ, Livneh B, *et al.* An assessment of differences in gridded precipitation datasets in complex terrain. *Journal of hydrology* 2018;**556**:1205–19.

44 Phuong J, Bandaragoda C, Istanbulluoglu E, *et al.* Automated retrieval, preprocessing, and visualization of gridded hydrometeorology data products for spatial-temporal exploratory analysis and intercomparison. *Environmental Modelling & Software* 2019;**116**:119–30. doi:10.1016/j.envsoft.2019.01.007

45 Daly C, Neilson RP, Phillips DL. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of applied meteorology* 1994;**33**:140–58.

46 Maurer EP, Wood A, Adam J, *et al.* A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of climate* 2002;**15**:3237–51.

47 Livneh B, Rosenberg EA, Lin C, *et al.* A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *Journal of Climate* 2013;**26**:9384–92.

48 Livneh B, Rajagopalan B. Development of a gridded meteorological dataset over Java island, Indonesia 1985–2014. *Scientific data* 2017;**4**:170072.

49 Liang X, Lettenmaier DP, Wood EF, *et al.* A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *Journal of Geophysical Research: Atmospheres* 1994;**99**:14415–28.

50 Livneh B, Bohn TJ, Pierce DW, *et al.* A spatially comprehensive, hydrometeorological data set for Mexico, the US, and Southern Canada 1950–2013. *Scientific data* 2015;**2**:150042.

51 Mote PW, Salathé EP. Future climate in the Pacific Northwest. *Climatic change* 2010;**102**:29–50.

52 Salathé Jr EP, Hamlet AF, Mass CF, *et al.* Estimates of twenty-first-century flood risk in the Pacific Northwest based on regional climate model simulations. *Journal of Hydrometeorology* 2014;**15**:1881–99.

53 Wilkinson MD, Dumontier M, Aalbersberg IjJ, *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 2016;**3**.

54 Mons B, Neylon C, Velterop J, *et al.* Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud. *Information Services & Use* 2017;**37**:49–56.

55 Kadlec J, StClair B, Ames DP, *et al.* WaterML R package for managing ecological experiment data on a CUAHSI HydroServer. *Ecological Informatics* 2015;**28**:19–28.

56 Gardner MA, Morton CG, Huntington JL, *et al.* Input data processing tools for the integrated hydrologic model GSFLOW. *Environmental modelling & software* 2018;**109**:41–53.

57 Read JS, Walker JI, Appling AP, *et al.* geoknife: reproducible web-processing of large gridded datasets. *Ecography* 2016;**39**:354–60.

58 Bandaragoda C. *Sauk-Suiattle HUC12 17110006, HydroShare*. 2017. http://www.hydroshare.org/resource/c532e0578e974201a0bc40a37ef2d284 (accessed 13 Aug 2018).

59 Beveridge C, Bandaragoda C, Phuong J. *Elwha Observatory- Public, HydroShare*. 2017. http://www.hydroshare.org/resource/1de72928f573433290f6c8bb393523df (accessed 13 Aug 2018).

60 Bandaragoda C. *Upper Rio Salado Watershed Boundary - Cibola National Forest, HydroShare*. 2017. http://www.hydroshare.org/resource/5c041d95ceb64dce8eb85d2a7db88ed7 (accessed 13 Aug 2018).

61 Heard J, Tarboton DG, Idaszak R, *et al.* An architectural overview of hydroshare, a next-generation hydrologic information system. 2014.

62 Horsburgh JS, Morsy MM, Castronova AM, *et al.* Hydroshare: Sharing diverse environmental data types and models as social objects with application to the hydrology domain. *JAWRA Journal of the American Water Resources Association* 2016;**52**:873–89.

63 Castronova AM, Brazil L, Seul M. Cloud-based Jupyter Notebooks for Water Data Analysis. 2017.

64 Baxter G, Sommerville I. Socio-technical systems: From design methods to systems engineering. *Interacting with computers* 2011;**23**:4–17.

65 Dwivedi MSKD, Upadhyay MS, Tripathi MAK. A working framework for the user-centered design approach and a survey of the available methods. *International Journal of Scientific and Research Publications* 2012;:12.

66 McKinney W. pandas: a foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing* 2011;**14**.

67 Jordahl K. GeoPandas: Python tools for geographic data. *URL: https://github com/geopandas/geopandas* 2014.

68 Gillies S. Fiona is OGR's neat, nimble, nononsense API. *URL https://github com/Toblerity/Fiona Accessed: April* 2017.

69 Gillies S, Bierbaum A, Lautaportti K, *et al.* Shapely: manipulation and analysis of geometric objects. *Toblerity org* 2007.

70 Rocklin M. Dask: Parallel computation with blocked algorithms and task scheduling. *In Proceedings of the 14th python in science conference* 2015;:130–6.

71 *hs_restclient: HydroShare REST API Python client library*. http://hs-restclient.readthedocs.io/en/latest/

72 *Western Regional Climate Center - Darrington Ranger Station monthly climate summary*. https://wrcc.dri.edu/cgi-bin/cliMAIN.pl?wa1992 (accessed 13 Aug 2018).

73 *National Water Information System. Daily streamflow discharge from the Sauk River Near Sauk, WA (USGS-12189500) from Jan 1 1950 through Dec 31 2011*. https://waterdata.usgs.gov/nwis/dv?cb_00060=on&format=rdb&site_no=12189500&referred_module=sw&period=&begin_date=1950-01-01&end_date=2011-12-31 (accessed 13 Aug 2018).

74 Konrad CP, Voss FD. Analysis of streamflow-gaging network for monitoring stormwater in small streams in the Puget Sound Basin, Washington. *USGS Scientific Investigation Report WA* 2012.

75 Livneh B. *Gridded climatology locations (1/16th degree): North American extent, HydroShare*. 2017. http://www.hydroshare.org/resource/ef2d82bf960144b4bfb1bae6242bcc7f (accessed 13 Aug 2018).

76 Vogel RM, Matalas NC, England Jr JF, *et al.* An assessment of exceedance probabilities of envelope curves. *Water resources research* 2007;**43**.

77 Simons R, Mendelssohn R. ERDDAP-A Brokering Data Server for Gridded and Tabular Datasets. 2012.

78 Hobley DE, Adams JM, Nudurupati SS, *et al.* Creative computing with Landlab: an open-source toolkit for building, coupling, and exploring two-dimensional numerical models of Earth-surface dynamics. *Earth Surface Dynamics* 2017;**5**:21.

79 Wohlstadter M, Shoaib L, Posey J, *et al.* A Python toolkit for visualizing greenhouse gas emissions at sub-county scales. *Environmental modelling & software* 2016;**83**:237–44.

80 Below R, Wallemacq P. Natural Disasters 2017. Brussels: : Centre for Research on the Epidemiology of Disasters (CRED) 2018. https://cred.be/sites/default/files/adsr_2017.pdf (accessed 26 Nov 2018).

81 Blumenthal D. Launching HITECH. *New England Journal of Medicine* 2010;**362**:382–5. doi:10.1056/NEJMp0912825

82 Blumenthal D, Collins SR. Health Care Coverage under the Affordable Care Act — A Progress Report. *New England Journal of Medicine* 2014;**371**:275–81. doi:10.1056/NEJMhpr1405667

83 Stephens KA, Lee ES, Estiri H, *et al.* Examining Researcher Needs and Barriers for using Electronic Health Data for Translational Research. In: *Proceedings - AMIA Joint Summits on Translational Sciences Proceedings*. 2015. 168–72.

84 Gutmann MP, Witkowski K, Colyer C, *et al.* Providing Spatial Data for Secondary Analysis: Issues and Current Practices Relating to Confidentiality. *Population Research and Policy Review* 2008;**27**:639–65. doi:10.1007/s11113-008-9095-4

85 Botsis T, Hartvigsen G, Chen F, *et al.* Secondary Use of EHR: Data Quality Issues and Informatics Opportunities. *Summit on Translational Bioinformatics*;**2010**:5.

86 O'Malley KJ, Cook KF, Price MD, *et al.* Measuring Diagnoses: ICD Code Accuracy. *Health Services Research* 2005;**40**:1620–39. doi:10.1111/j.1475-6773.2005.00444.x

87 Hibbert JD, Liese AD, Lawson A, *et al.* Evaluating geographic imputation approaches for zip code level data: an application to a study of pediatric diabetes. *International Journal of Health Geographics* 2009;**8**:54. doi:10.1186/1476-072X-8-54

88 Basara HG, Yuan M. Community health assessment using self-organizing maps and geographic information systems. *International Journal of Health Geographics* 2008;**7**:67. doi:10.1186/1476-072X-7-67

89 Helbich M, Plener PL, Hartung S, *et al.* Spatiotemporal Suicide Risk in Germany: A Longitudinal Study 2007–11. *Scientific Reports* 2017;**7**. doi:10.1038/s41598-017-08117-4

90 Dwyer-Lindgren L, Bertozzi-Villa A, Stubbs RW, *et al.* Trends and Patterns of Geographic Variation in Mortality From Substance Use Disorders and Intentional Injuries Among US Counties, 1980-2014. *JAMA* 2018;**319**:1013. doi:10.1001/jama.2018.0900

91 Jung I, Kulldorff M, Klassen AC. A spatial scan statistic for ordinal data. *Statistics in Medicine* 2007;**26**:1594–607. doi:10.1002/sim.2607

92 Neill DB. An empirical comparison of spatial scan statistics for outbreak detection. *International Journal of Health Geographics* 2009;**8**:20. doi:10.1186/1476-072X-8-20

93 Li L, Xi Y, Ren F. Spatio-Temporal Distribution Characteristics and Trajectory Similarity Analysis of Tuberculosis in Beijing, China. *International Journal of Environmental Research and Public Health* 2016;**13**:291. doi:10.3390/ijerph13030291

94 Allévius B, Höhle M. An expectation-based space-time scan statistic for ZIP-distributed data. *arXiv:171209188 [stat]* Published Online First: 26 December 2017.http://arxiv.org/abs/1712.09188 (accessed 13 Nov 2019).

95 Wang A, Wheeler DC. Catchment Area Analysis Using Bayesian Regression Modeling. *Cancer Informatics* 2015;**14s2**:CIN.S17297. doi:10.4137/CIN.S17297

96 Onyile A, Vaidya SR, Kuperman G, *et al.* Geographical distribution of patients visiting a health information exchange in New York City. *Journal of the American Medical Informatics Association* 2013;**20**:e125–30. doi:10.1136/amiajnl-2012-001217

97 Comer KF, Grannis S, Dixon BE, *et al.* Incorporating Geospatial Capacity within Clinical Data Systems to Address Social Determinants of Health. *Public Health Reports* 2011;**126**:54–61. doi:10.1177/00333549111260S310

98 University of Washington HIPAA policy for minimal risk data sets. 2018.http://depts.washington.edu/comply/comp_103/#103_VI

99 *International Classification of Diseases, version 9 - Clinical Morphologies (ICD9CM)*. 2017. https://bioportal.bioontology.org/ontologies/ICD9CM

100 *International Classification of Diseases, version 10 - Clinical Morphologies (ICD10CM)*. 2017. https://bioportal.bioontology.org/ontologies/ICD10CM

101 *U.S. Census Bureau. 2010 Census 5-digit ZIP Code Tabulated Areas by county parts, 2010 TIGER/Line Shapefile.* Washington State Office of Financial Management 2010. https://www.ofm.wa.gov/washington-data-research/population-demographics/gis-data/census-geographic-files

102 Mohrman M, Kimpel T. Small Area Estimate Program User Guide. Washington State Office of Financial Management 2012. http://www.ofm.wa.gov/pop/smallarea/docs/saep_user_guide.pdf

103 *Small Area Estimates Program (SAEP) for ZIP Code Tabulated Areas by County Parts*. Washington State Office of Financial Management (OFM) Small Area Estimates Program (SAEP) 2019. https://www.ofm.wa.gov/washington-data-research/population-demographics/population-estimates/small-area-estimates-program

104 Zhang Z, Assunção R, Kulldorff M. Spatial Scan Statistics Adjusted for Multiple Clusters. *Journal of Probability and Statistics* 2010;**2010**:1–11. doi:10.1155/2010/642379

105 Prates MO, Kulldorff M, Assunção RM. Relative risk estimates from spatial and space-time scan statistics: are they biased?: RELATIVE RISK ESTIMATES FROM SPATIAL AND SPACE-TIME SCAN STATISTICS. *Statistics in Medicine* 2014;**33**:2634–44. doi:10.1002/sim.6143

106 Grubesic TH, Matisziw TC. On the use of ZIP codes and ZIP code tabulation areas (ZCTAs) for the spatial analysis of epidemiological data. *Int J Health Geogr* 2006;**5**:58. doi:10.1186/1476-072X-5-58

107 Krieger N, Waterman P, Chen JT, *et al.* Zip Code Caveat: Bias Due to Spatiotemporal Mismatches Between Zip Codes and US Census–Defined Geographic Areas—The Public Health Disparities Geocoding Project. *American Journal of Public Health* 2002;**92**:3. doi:10.2105/AJPH.92.7.1100

108 Chaix B, Duncan D, Vallée J, *et al.* The "Residential" Effect Fallacy in Neighborhood and Health Studies: Formal Definition, Empirical Identification, and Correction. *Epidemiology* 2017;**28**:789–97. doi:10.1097/EDE.0000000000000726

109 Delgado-Rodriguez M. Bias. *Journal of Epidemiology & Community Health* 2004;**58**:635–41. doi:10.1136/jech.2003.008466

110 Bazemore AW, Cottrell EK, Gold R, *et al.* "Community vital signs": incorporating geocoded social determinants into electronic records to promote patient and population health. *Journal of the American Medical Informatics Association* 2016;**23**:407–12. doi:10.1093/jamia/ocv088

111 Mcdonald JH. *HANDBOOK OF BIOLOGICAL STATISTICS*. 2nd ed. University of Delaware: : Sparky House Publishing http://www.biostathandbook.com/

Appendix 2-1: Interview Guide
Introduction [5 min]

Hi, I'm an informatics student working with the **Puerto Rico Water Studies - Population Health Research workgroup**. Thank you very much for making time to talk with me today.

The reason I'd like to interview you is because our team really wants to hear about your **experience considering the health of populations affected by Hurricane and flood-related disasters**. By talking one-on-one with researchers like yourself, we hope to engage you and learn what your experience has been like **gathering information and problem-solving**.

From these interviews, our team will try to **create tools to improve how population health researchers access and use information for preparation or recovery from hurricane and flood-related disasters**. So, with that in mind, I'd like to talk with you for 90 minutes in a recorded session and ask you several open-ended questions.

Here is a consent form that explains this IRB approved study. We'll be sharing your responses only with the **Puerto Rico Water Studies - Population health research workgroup**. We will remove links between your name and your responses. When we report results of this study, only overall findings will be presented, without identifying any individuals. During the interview, you do not have to answer any questions you do not feel comfortable answering and you are free to stop at any time if you don't want to continue.

Do you have any questions for me?

Since this is intended to be an audio-recorded interview session, I'd like to get your permission to continue with audio recording. **It will help us get a perfect record of the conversation and so we can transcribe your responses for analysis**. Is it ok with you if we record this interview?

[If yes] Great. I'll go ahead and start the recording now.
[If no] Okay, that's fine too. We'll go ahead without recording.

I would like to ask that you read this study consent form. When you are ready, please read the final paragraph aloud acknowledging that you have read and understood the consent form, then sign it.

Health Research Participation [10 min]
Question 1:

Question Goal: To evaluate participant's role and motivation for taking part in this needs assessment (If not applicable - skip to Qx 2)

1. What is your role in population health research or management?
    a. How involved do you feel with population health research operations?

    b. Do you focus on a particular region, patient population, and/or health outcome?
        i. Which ones?

2. What are key components for you to conduct your research in population health?

3. How has hurricane or flood affected your ability to conduct research?

4. How has hurricane or flood affected your population of interest?

Researcher's Attitudes Towards Technology for their Health Research [10 min]
Question 2:

Question Goal: To learn about participant's attitudes toward technology, and what they're comfortable using.

1. Can you describe a positive experience you've had with new technology in your population health research?
    a. Why was it positive experience?

2. How about a more negative experience with new technology in your health research?
    a. What happened to make it negative?

3. Among the technologies you've just mentioned, how often do you use them now?
    a. How comfortable are you using these technologies now, if you are still using them?

4. What time-scale or regional-scale of information does this technology provide you?

5. What do you hope to get out of new tools for population health research with regards to natural disasters?
    a. Any expectations?

Researcher's Attitudes towards integrative researcher networks for their research [10 min]
Question 3:

Question Goal: To see if the participant is engaged in other research projects networks, and how they feel about those.

1. What research communities and networks are you involved with?
    a. Why did you decide to participate in these?
2. Can you describe to me a positive experience you've had with one of these networks/communities?
3. What about a not so positive experience (or a more negative experience)?
4. Overall, how do you feel about patient networks and communities? Do you feel as though these networks/communities help you in some way?
    a. In what way?

Population Health Research Processes and Contexts [20 min]
Question 4:

Question Goal: To understand what health questions the participant might have, and to determine if those health questions can be answered by the data we have.

1. Do you ever find yourself wondering about the health status within your patient population?
   [If yes]:
    a. How often would you say this happens? And Why?
    b. Are there any specific questions that you would try to answer first?
    c. If you wanted to answer these questions right now, what would you do?
    d. How does this change during post-hurricane or flood conditions, if at all?
    e. How do you consider information quality between these two scenarios?

[If no]:

a. Why not?
    - Is there a reason for this, or do you simply just not think about it?

b. Would you say that your information resources have kept you informed in your operations?

Question 5:

Question Goal: To understand what workflows or protocols might exist? and in hurricane and flood scenarios?

1. Suppose that a hurricane/flood occurs tomorrow. Is there a standard protocol in how you would respond?
   [If yes]:
   a. Can you describe your workflow or standard protocol?

   b. Are there any considerations that you pay more or less attention?

   [If no]:
   a. How would you respond to the needs of your patient population?

   b. What questions would you ask?

   c. Are there particular data sources or information providers that you would use to answer these questions?

   d. Are you well-informed to navigate and use these data sources and information providers?

2. If you were provided regional water quality information, would that improve your decision making?
   [If yes]:
   a. How do you make sense of water quality measurements? to health outcomes?

b. How often do you need it?

c. What things should be prioritized to the front?

[If no]:

a. Why not?
    - Is there a particular reason for this, or do you simply just not think about it?

b. What about your participation in other researcher networks and communities? Has that helped?

3. If you were provided with regional health statistics, would that improve your decision making?
   [If yes]:
   a. How often do you need it?

   b. What things should be prioritized to the front?

   [If no]:
   a. Why not?
       - Is there a particular reason for this, or do you simply just not think about it?

   b. What about your participation in other researcher networks and communities? Has that helped?

Question 6:

Question Goal: To understand how participant currently looks up health information, and what a positive experience of this was like.

1. As I mentioned earlier, I also want to learn a little bit about what your experience has been like looking up health information. So with that in mind, can you walk me through a specific time when you really wanted to look up population health statistics, and you ended up finding what you needed successfully?

2. Where were you when you looked up this info? (e.g. home, work, commute, clinic setting)

3. Why did you seek out this information? (e.g. illness, health concerns, curiosity)

4. What was the actionable outcome of this research, if any?

5. Did you have questions before you looked up this information?

6. What was your impression of the process to find your answer?
   [*If negative]:*
   a. Why was this hard for you? What got in the way?

   [If *positive]:*
   a. What made it easy? What helped?

7. Do you remember what you used to find the information? (e.g. smartphone, laptop, WebMD, Google, Mayo Clinic, Medical Journals, some other resource?)

8. Would you say what you just described to me is pretty typical for you?

Question 7:

Question Goal: To facilitate a brainstorming session with the participant to see what they would need or want from health research pertinent to their patient population's health conditions.

1. Imagine you have access to a website that shared up-to-date health research about people living with health conditions in the area. Do you see yourself using this website?
   [If yes]:
   a. Why would you want to use this?

   b. What kinds of health information would make this website useful for you?

   c. How would you want this information to be presented?

   d. Are there certain questions you would want a website like this to answer?

   e. What do you imagine yourself doing with this information?

   [If no]:
   a. What else/alternates would you prefer?

   b. What would be needed to make this useful for your own health research needs?

2. Imagine you have access to a website that shared up-to-date storm information, wind and water burden, and the density of people with medical conditions requiring access to care in those areas. Do you see yourself using this website?

   [If yes]:
   a. Why would you want to use this?

   b. What kinds of health information would make this website useful for you?

   c. How would you want this information to be presented?

d. Are there certain questions you would want a website like this to answer?

e. What do you imagine yourself doing with this information?

[If no]:

a. What else/alternates would you prefer?

b. What would be needed to make this useful for your own health research needs?

Demographic questions

Before we end today, I'd like to ask you a few questions about yourself.

1. What is your gender?
    c. Male
    d. Female
    e. Other
    f. Decline to answer

2. Are you:
    g. 18-44
    h. 45-64
    i. 65+

3. Which of these occupational roles would best describe you?
    j. Data analyst for population health management
    k. Population health support staff but not an analyst
    l. Community health worker (e.g., public health nurse)
    m. Program analyst (e.g., program management and reporting)

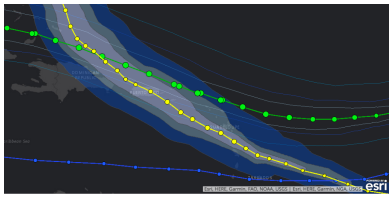
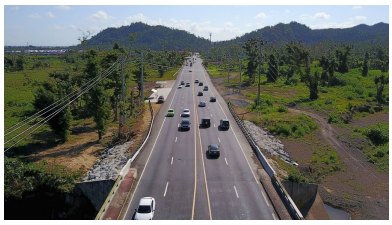
4. Please select all occupational functions that apply to you.
    n. Project management
    o. Data visualization on a spatial-scale
    p. Descriptive statistics of public health concerns
    q. Communication of patient health trends
    r. Statistical data analysis and modeling to discover health determinants

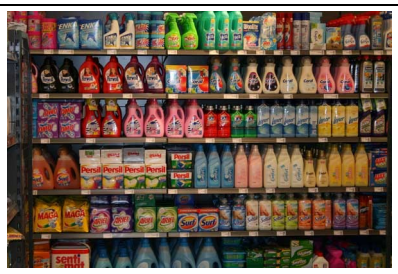
5. What kind of natural disaster did you provide community/population health services?
    s. region affected by Hurricane
    t. region affected by flood
    u. region affected by both

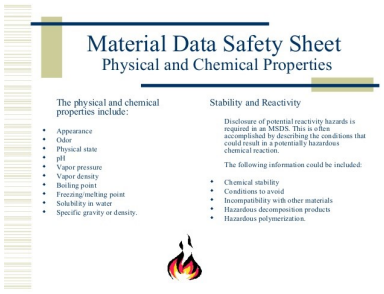
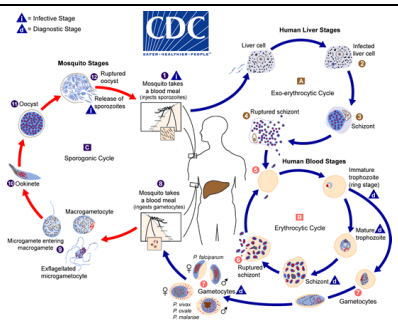
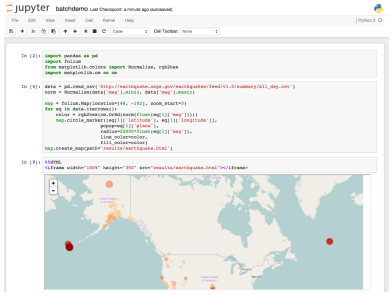
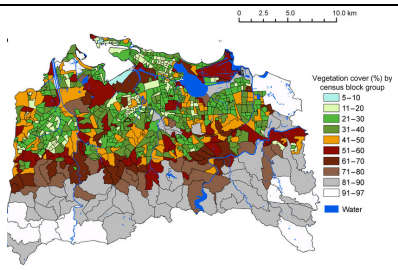
All right, that's all the questions I have for you. Is there anything else you'd like to share with me today?

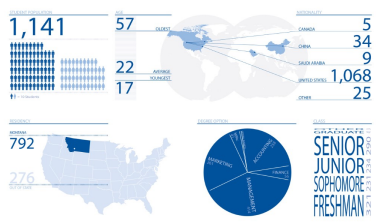
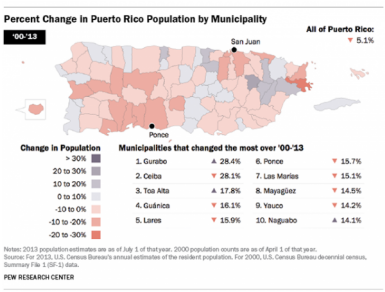
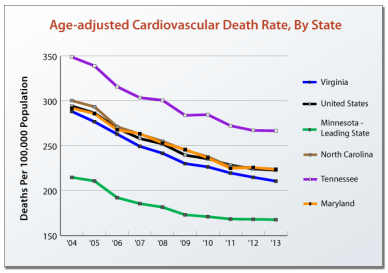
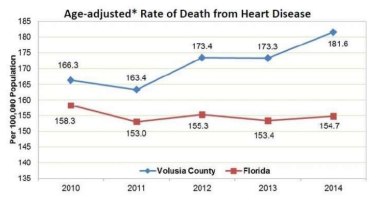
Thank you so much. All the information you've given our team today has been really helpful. I really appreciate your time and I hope you have a great rest of your day! Take care.

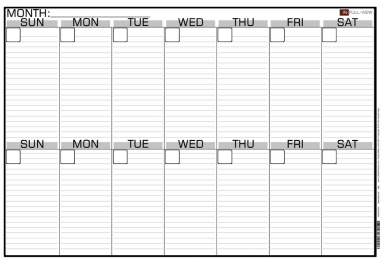
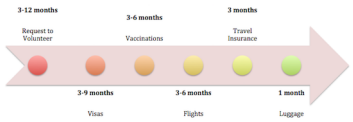
Appendix 2-2. Hurricane and Flood-related Information Cards

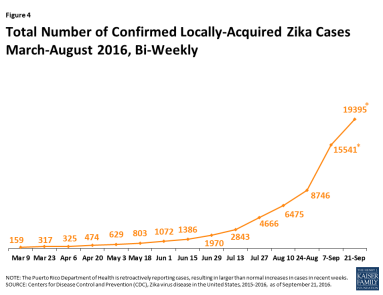
Card themes	Card label (card front)	Image (card front)	Card description (card back)
Storm route - story map	Storm route		A story map describing the directional trajectory, storm size, and sequential time-steps of a storm.
Storm damage - real-time geocoded photos from NAPSG	Storm damage		Geocoded photos uploaded in real-time via Twitter, damage surveyors, and news reporters.
Hospital locations - HIFLD	Hospital locations		This is the location of hospital aid, including emergency departments, clinics, and intensive care facilities. This information does not include the real-time status of the hospitals. It has general statistics to describe its capacity.
Pharmacy locations - HIFLD	Pharmacy locations		These are the locations for all accessible pharmacies and pharmaceutical supplies. This information does not include the real-time status of the facility.
National roads linkages - HIFLD	Roads and Highways		This file maps the roads and their connections with one another. While not updated in real-time, it can be used to understand the shortest route between two locations.
Blocked roads - FEMA	Blocked roads		This geocoded data set describes blocked roads surveyed by FEMA. This data set is a one-time survey at the beginning of emergency response, and may not represent events that follow.

Flooded areas - FEMA	Flooded areas		This geocoded data set describes the areas documented by FEMA to have flood waters. This data set is a one-time survey at the beginning of emergency response, and may not represent events that follow.
Streamflow hazard - sensors and gauges	Streamflow hazard		Real-time information from streamflow sensors and groundwater gauges provide instantaneous geographic insight about the water flow influence by Hurricane Maria.
Microbial presence - water quality testing	Microbial presence testing		Infection to microbial agents in drinking water may result in recurring water-borne illnesses. Some agents may be effectively disinfected. Others require more information about their presence before actions could be taken.
Chemical levels (regulated contaminants) - water quality testing	Testing for regulated water quality endpoints		A set of policies established water quality endpoints and chemical contaminants are part of routine testing. These endpoints may follow standard operating protocols and established benchmark limits.
Chemical levels (unregulated contaminants) - water quality testing	Unregulated chemical contaminants		Some unregulated chemical contaminants are suspect for environmental health concerns. Sometimes they show up where they are not expected (like drinking water), drawing greater concern about their exposure, safety, and hazard.

Chemical level interpretation aid	Chemical interpretation aid		Reference guides help provide structure to interpret information about environmental hazards, like chemical hazards. These guides can simplify what to do and what not to do.
Microbial interpretation aid	Microbial interpretation aid		Reference guides help provide structure to interpret information about environmental hazards, like microbial hazards. These guides can simplify what to do and what not to do.
Story notebooks	Training Modules and Workflows		These data files can be used for training purposes for example tasks or analyses.
Spatial polygons by State - Census	Territory geographic boundaries		When combined, this geographic data set can show point location or road segments relative to the entirety. (1 polygon).
Spatial polygons by County - Census	County/Municipios boundaries		When combined, this geographic data set can show different information relative to the County/Municipio boundaries (78 polygons).
Spatial polygons by Census blocks - Census	Census Block boundaries		Blocks are the smallest available geographic unit to compare Census information, almost to neighborhoods. Spatial analyses commonly use blocks as the smallest unit of aggregated information.

Demographic profile by State - Census	State/Territory demographics		Census information from 2000 or 2010 were collected as Census blocks, then aggregated up to a State-level summary.
Demographic profile by County/Municipality - Census	County demographics		Census information from 2000 or 2010 were collected as Census blocks, then aggregated up to County/municipios-level summaries.
Demographic profile by Tract - Census	Census Tract demographics		Census information from 2000 or 2010 were collected as Census blocks, then aggregated up to County/municipios-level summaries.
Population estimates - Census ACS	Annual population estimates	AMERICAN COMMUNITY SURVEY U.S. CENSUS BUREAU	Between mass Census surveys, the Census Bureau reports use annual birth and death registries to make annual population estimates. This information may not be available at smaller scales than county/municipios.
Health outcomes by State	Health outcomes by State		Annual rates of health outcomes are commonly summarized for the whole state/territory. Changes in resource allocation and policies can thereby be compared with reference to prior rates.
Health outcomes by County	Health outcomes by County		County-level health outcomes before and after a disaster may indicate uneven distribution of the burden of disease. Considering health outcomes by county may identify counties and sub-counties with health disparities.

Public Health needs assessment	Public Health Assessments		These reports often summarize a number of public health concerns at a single point in time from the perspective of a public health practitioner. Information from prior scenarios and literature may be used to give recommendations on early warning practices.
Qualitative interviews	Qualitative Interviews		The first-person perspective of their experiences. This can give insights into how people perceive and feel in a situation. This can include their way of making sense of the new information, what is perceived as the problem, their conventional response, their priorities, and their value-driven decisions.
Professional collaborations - Human resources	Human resource		Community coordination and collaboration with field expertise may resolve problems faster than addressing learning curves. Also consider where they are and what they need for the situation.
Time-frame constraints (up to 4 days)	Time constraint (up to 4 days)		Some things add stress and should not exceed 4 days delay. Information of this sort has immediate decision making influence and the situation becomes more urgent with delay (e.g., intensive care services, food and water supply).
Time-frame constraints (up to 2 weeks)	Time constraint (up to 2 weeks)		Some information takes a week or more to retrieve. This can be the case with operational processes in progress, or where prior sampling events or dependencies are not yet in place.
Time-frame constraints (up to 6 months)	Time constraint (monthly updates)		Monthly updates can highlight expected or unexpected trends specific to a certain period. This can be the case when a cumulative set of events need to happen first (e.g., assessments, repairs, inspections).

Time-frame constraints (up to 1 year)	Time constraint (Annual)		Milestone updates might be summarized for all data over an entire period of time. This may include prior year information.
Time-frame constraints (real-time)	Time constraint (Real-Time)		Some information needs to be updated immediately as it is made known. Information of this sort has immediate decision making influence, and should be given the greatest priority.

Appendix 2-3A: Participant Readiness for future disaster management

PID	Readiness
P01	Experienced but not planned or prepared
P02	Planned and prepared for a role
P04	Experienced but not planned or prepared
P05	No experience and would need collaborators
P06	Experienced but not planned or prepared
P07	No experience and would need collaborators
P08	Experienced but not planned or prepared
P11	Experienced but not planned or prepared
P12	Experienced but not planned or prepared
P13	Experienced but not planned or prepared
P16	Planned and prepared for a role
P17	Planned and prepared for a role
P18	Planned and prepared for a role
P19	Planned and prepared for a role
P20	Planned and prepared for a role

Appendix 2-3B: Type of researchers

PID	Biomedical informatician	Decision analyst	Demographer	Emergency management researcher	Environmental health researcher	Epidemiologist	Geographer	Global health researcher	Operations researcher	Physician	Policy researcher	Reconnaissance researcher	Risk management researcher
P01	1	0	0	0	0	1	0	0	0	1	0	0	0
P02	0	0	0	0	1	0	1	0	0	0	0	0	0
P04	0	0	0	1	0	1	0	1	0	0	0	0	0
P05	0	0	0	0	0	1	0	0	0	0	0	0	0
P06	0	0	0	0	0	1	0	0	1	0	0	0	0
P07	0	0	1	0	0	0	0	0	0	0	0	0	0
P08	0	1	0	0	0	0	0	0	0	0	1	0	1
P11	0	0	0	0	0	1	0	1	1	0	0	0	1
P12	0	0	0	0	1	0	0	1	0	1	0	0	0
P13	0	0	0	0	0	0	0	0	0	0	1	1	0
P16	0	0	0	1	0	0	0	0	0	0	0	0	0
P17	0	0	0	0	0	1	0	0	0	0	0	0	0
P18	0	0	0	0	1	0	0	0	0	0	0	1	0
P19	0	0	0	0	1	0	0	0	0	0	0	0	0
P20	0	0	0	1	0	1	0	0	0	0	0	0	0

Appendix 2-3C: Participants by Type of Researcher and Current Occupational Setting

	Codes	PID	
group		Academic	Government
1	Epidemiologist	P01, P04, P05, P06, P11	P17, P20
2	Environmental health researcher	P02, P12	P18, P19
3	Global health researcher	P04, P11, P12	
4	Emergency management researcher	P04	P16, P20
5	Risk management researcher	P08, P11	
6	Reconnaissance researcher	P13	P18
7	Policy researcher	P08, P13	
8	Physician	P01, P12	
9	Operations researcher	P06, P11	
10	Geographer	P02	
11	Demographer	P07	
12	Decision analyst	P08	
13	Biomedical informatician	P01	

Appendix 2-3D: Analysis types

PID	Apply theoretical frameworks	Driven by a research question	Field sampling and assessment	Hypothesis-generation and exploratory analysis	Mixed method	Observation	Policy assessment	Primary data analysis	Secondary data analysis	Spatial-temporal modeling	Statistical methods	Stratification by group	Survey methods and interviews	Geospatial scale of focus	Temporal scale of focus
P01	0	0	0	0	0	0	0	0	1	0	1	1	0	0	1
P02	1	0	0	0	0	0	0	0	1	1	0	0	1	1	1
P04	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
P05	0	1	0	1	0	0	0	0	1	1	1	1	0	1	1
P06	0	1	0	0	1	1	0	1	0	0	0	0	1	1	1
P07	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0
P08	1	0	0	0	1	0	0	1	0	0	0	0	1	0	0
P11	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
P12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P13	0	1	0	0	1	0	1	1	0	1	0	0	1	0	1
P16	0	1	1	0	0	0	0	1	1	1	0	0	0	1	1
P17	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
P18	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
P19	0	0	1	0	0	0	0	1	0	0	0	0	1	0	1
P20	1	0	0	0	0	0	0	1	1	0	0	0	1	1	1

Appendix 2-3E: Participants by Analysis types and and Current Occupational Setting

	Codes	PID	
group		Academic	Government
1	Survey methods and interviews	P02, P06, P07, P08, P13	P19, P20
2	Primary data analysis	P04, P06, P08, P13	P16, P19, P20
3	Secondary data analysis	P01, P02, P05, P07	P16, P20
4	Driven by a research question	P05, P06, P11, P13	P16
5	Spatial-temporal modeling	P02, P05, P13	P16
6	Apply theoretical frameworks	P02, P08, P11	P20
7	Mixed method	P06, P08, P13	
8	Stratification by group	P01, P05	
9	Statistical methods	P01, P05	
10	Field sampling and assessment		P16, P19
11	Policy assessment	P13	
12	Observation	P06	
13	Hypothesis-generation and exploratory analysis	P05	

Appendix 2-3F: Participant subject matters of focus

PID	All-hazards emergency management	Causal inference	Chronic illness	Climate change and disaster events	Decision analysis	Design evaluation	Disease surveillance systems	Access to care	Exposure hazard agents	Food and water supply	Health outcomes related to disasters	Medical vulnerability	Physical environmental vulnerability	Population migration and homelessness	Socio-economic determinants of health	Foodborne and communicable diseases	Health adaptation	Mortality	Organizational performance	Physical activity	Population processes	Risk assessment and communication
P01	0	1	1	0	0	0	1	1	0	1	1	1	0	0	1	0	0	1	1	0	0	0
P02	1	0	1	0	0	0	0	1	1	0	0	1	1	1	1	0	1	0	0	0	0	0
P04	1	0	0	0	0	1	0	1	1	0	0	1	0	0	0	1	0	0	0	0	0	0
P05	0	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0
P06	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0
P07	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	1	0
P08	1	0	0	1	1	0	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1
P11	1	0	0	1	0	0	0	1	1	1	1	1	1	1	1	0	1	0	0	0	0	1
P12	0	0	0	1	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0
P13	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	1	0	0
P16	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
P17	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
P18	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	1	0	0	0	0	0	0
P19	1	0	0	0	0	0	1	0	1	1	1	0	0	1	0	1	0	0	0	0	0	0
P20	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0

Appendix 2-3G: Participants by Subject matter of focus and Current Occupational Setting

	Codes	PID	
group		Academic	Government
1	All-hazards emergency management	P02, P04, P06, P08, P11, P13	P16, P19, P20
2	Exposure hazard agents	P02, P04, P11, P12, P13	P18, P19, P20
3	Health outcomes related to disasters	P01, P06, P08, P11, P13	P19, P20
4	Disease surveillance systems	P01	P16, P17, P18, P19, P20
5	Access to care	P01, P02, P04, P11, P12	
6	Food and water supply	P01, P07, P11	P18, P19
7	Population migration and homelessness	P02, P07, P11, P12	P19
8	Climate change and disaster events	P06, P08, P11, P12	P18
9	Organizational performance	P01, P06, P08, P12, P13	
10	Medical vulnerability	P01, P02, P04, P08, P11	
11	Socio-economic determinants of health	P01, P02, P08, P11	
12	Foodborne and communicable diseases	P04	P16, P18, P19
13	Mortality	P01, P05, P07	
14	Risk assessment and communication	P08, P11	
15	Health adaptation	P02, P11	
16	Physical activity	P05, P13	
17	Design evaluation	P04, P05	
18	Decision analysis	P05, P08	
19	Physical environmental vulnerability	P02, P11	
20	Chronic illness	P01, P02	
21	Causal inference	P01, P05	
22	Population processes	P05, P07	

Appendix 2-3H: Participants by Context of research focus and Current Occupational Setting

	Codes	PID	
group		Academic	Government
1	Temporal scale of focus	P01, P02, P05, P06, P11, P13	P16, P17, P18, P19, P20
2	Place of research focus	P01, P02, P04, P05, P06, P07, P12	P17, P19
3	Geospatial scale of focus	P02, P05, P06, P07	P16, P17, P18, P20

Appendix 4-1: Crosswalk between ICD9 and ICD10 code families

ICD9 code family	ICD10 code description	ICD10 code family
001-139.99	Certain infectious and parasitic diseases	A00-B99
140-239.99	Neoplasms	C00-D49
280-289.99	Diseases of the blood and blood-forming organs and certain disorders involving the immune mechanism	D50-D89
240-279.99	Endocrine, nutritional and metabolic diseases	E00-E89
290-319.99	Mental, Behavioral disorders	F01-F99
320-389.99	Diseases of the nervous system	G00-G99
360-379.99	Diseases of the eye and adnexa	H00-H59
380-389.99	Diseases of the ear and mastoid process	H60-H95
390-459.99	Diseases of the circulatory system	I00-I99
460-519.99	Diseases of the respiratory system	J00-J99
520-579.99	Diseases of the digestive system	K00-K95
680-709.99	Diseases of the skin and subcutaneous tissue	L00-L99
710-739.99	Diseases of the musculoskeletal system and connective tissue	M00-M99
580-629.99	Diseases of the genitourinary system	N00-N99
630-679.99	Pregnancy, childbirth and the puerperium	O00-O9A
760-779.99	Certain conditions originating in the perinatal period	P00-P96
740-759.99	Congenital malformations, deformations and chromosomal abnormalities	Q00-Q99
780-799.99	Symptoms, signs and abnormal clinical and laboratory findings, not elsewhere classified	R00-R99
800-999.99	Injury, poisoning and certain other consequences of external causes	S00-T88
E000-E999.9	External causes of morbidity and mortality	V00-Y99
V01-V91.99	Factors influencing health status and contact with health services	Z00-Z99