A Content Analysis of Stakeholder Interviews on Developing Pathways for Community-led Research with Big Data

Shira Grayson

A thesis

submitted in partial fulfillment of the

requirements for the degree of

Master of Public Health

University of Washington

2019

Committee:

Nanibaa' A. Garrison

Joon-Ho Yu

Megan Doerr

Program Authorized to Offer Degree:

Public Health Genetics

University of Washington

Abstract

A Content Analysis of Stakeholder Interviews on Developing Pathways for Community-led Research with Big Data

Shira Grayson

Chair of the Supervisory Committee:
Nanibaa' A. Garrison, PhD, Graduate Faculty
Department of Pediatrics, Division of Bioethics and Palliative Care
Institute for Public Health Genetics

In an era where Big Data (BD) informs nearly all aspects of human life, there are profound opportunities to translate BD research to guide physicians, epidemiologists, health policy experts, and community leaders in making data-driven decisions to improve health outcomes and reduce health disparities. In order to derive the most benefit from BD research, opportunities are needed to facilitate and support a movement of community-led scientists to design BD research and derived products that intentionally and effectively reduce health disparities and promote health equity within their own communities. The aims of this study are to capture meaningful uses of BD ("use cases") from the perspectives of community leaders and to identify what is needed for community members to engage with BD. We conducted a qualitative content analysis of semi-structured key informant interviews with 16 community leaders. Based on our analysis findings, we developed a BD Engagement Framework that illustrates the pathways and various forces for and against community engagement in BD research, as described by our informants.

We hope that our Framework will promote concrete, transparent dialogue between communities and researchers about barriers and facilitators of authentic community-engaged BD research. Findings from this study will inform the subsequent phases of a multi-phased project with the ultimate aim of organizing fundable frameworks for BD projects within community settings.

Acknowledgements

I owe tremendous gratitude to Joon-Ho Yu, Megan Doerr, and Nanibaa' A. Garrison, for their incredible patience and devotion to mentorship throughout my thesis project. I am appreciative of their eagerness to provide thoughtful guidance as needed, while also inspiring me to make this project my own. I would also like to acknowledge the Institute for Public Health Genetics for its continued support and interdisciplinary training. Finally, thank you to my family and friends for their continuous encouragement throughout this process.

Table of Contents

1. Introduction

Tremendous national and international investments in accelerating precision medicine research are underway to uncover novel targeted approaches for improving disease prevention and treatments (1,2). The *All of Us* (AoU) Research Program is poised to become the largest national initiative, funded by the United States (U.S) National Institutes of Health (NIH), that aims to collect genomic, environmental, and lifestyle data from one million or more people to enable precision medicine for all Americans (3). Though the definition of precision medicine is everchanging and is a continuous source of conversation, there is general agreement that precision medicine is data intensive and requires Big Data (BD) sets comprising a range of health-related variables (4–10). BD informs nearly all aspects of human life, and lays a foundation to provide profound opportunities to translate BD research to help physicians, epidemiologists, health policy experts, and community leaders make data-driven decisions to improve health outcomes and reduce health disparities (11–16). Thus, this powerful vision for accelerating precision medicine to improve the health and well-being of all Americans through BD studies such as AoU, relies not only on the volume and breadth of the data, but also on the representativeness of the datasets and the processes by which BD research is translated into practice (17–19).

Recruitment, engagement, and retainment of communities who have been historically underrepresented in biomedical research (UBR), such as those from racial and ethnic minorities, socio-economically disadvantaged backgrounds, or with disabilities, remain an area for improvement across all research disciplines (20,21). As a result, this historical underrepresentation in research can lead to a limited understanding of community needs,

exacerbate vulnerabilities to poor health outcomes, and prohibit access to healthcare services (22,23).

Studies designed to explore the attitudes, beliefs and perspectives of UBR communities suggest that existing barriers for participation include a lack of awareness about research opportunities (24,25), a belief that the research is not relevant to themselves or their community (26), limited comprehension of the research purpose or procedures (27), and frustrations due to poor dissemination of research findings back to community members (24). There is also a mistrust and unwillingness among some UBR groups such as indigenous communities (including American Indians (AI), Alaska Natives (AN), and Native Hawaiians (NH)), to share personal health information with researchers due to trauma experienced from recent and historical research malpractice (28,29).

Progress is certainly being made to include a more diverse spectrum of research participants in datasets through a range of Community-Engaged Research (CEnR) strategies (3,30). While diverse inclusion in biomedical research is important, community engagement is vital at all stages of the research process to increase the quality, relevance, and efficacy of research translation to improve individual and population health (30,31). A few organizations, including Sage Bionetworks, a non-profit biomedical research and technology organization committed to open science, are piloting citizen science/qualified researcher platforms to encourage more inclusive and collaborative engagement in all aspects of scientific research by making data sources accessible to interested members of the public (32). While this collective effort is a promising start on the path to empower groups who have been disenfranchised by the research

enterprise, much more must be done within this space to determine how best to transfer the power, autonomy, and ownership of datasets back to data donors, and to encourage authentic engagement of communities in the entire process of scientific inquiry and research.

As the potential and national investment in BD research to transform health outcomes and reduce health disparities continues to grow (33), it will be important to ensure that research inclusion efforts do not prematurely halt at the recruitment phase. In order to derive maximum benefit from the data shared by participants from UBR populations, opportunities are needed to facilitate and support (i.e., opportunities to build capacity) a movement of community-led citizen scientists to design BD research and derived-products that intentionally and effectively reduce health disparities and promote health equity within their own communities.

Thus, it is imperative that in this process of including the broadest community of solvers to ask their own research questions of the datasets, we do not dictate what role communities should play in this landscape of BD research or in the research enterprise as a whole, nor should we assume what tools and supports communities will need to facilitate this involvement. Instead, in order to unearth communities' perspectives on these key topics, a trustworthy environment and approach are required in which community leaders and stakeholders can share how BD research may or may not intersect with and support community priorities.

1.1 Study Aims

The aims of this study are to capture meaningful uses of BD ("use cases") from the perspectives of community leaders and to identify what is needed for community members to engage with Big

Data (BD) including why these elements are necessary and how best to deliver them. We define "use cases" in this study as specific situations in which BD could be used by a community for their own benefit. This study constitutes the first phase of a larger, ongoing multi-phased project. Findings from this first phase will inform the second phase of the project that is focused on identifying short and long-term tools and supports to increase communities' capacity for community-driven BD research. The goal of the final phase of the project is to then translate evidence from the earlier phases to develop a robust framework for BD projects within community settings.

2. Methods

2.1 Study Design

We planned semi-structured key informant interviews with the goal of capturing examples or "use cases" that illustrate the ways in which communities might interact with BD research and use BD for their benefit. We developed our interview guide iteratively, informed by literature focused on community-engaged research, patient/citizen-led research, precision medicine, community perspectives on BD, and UBR populations. The study protocol was reviewed by the University of Washington Human Subjects Division and determined to qualify for exempt (category 2) status (STUDY00006646).

2.2 Participant Eligibility and Recruitment

Key informants were identified from our professional networks and invited by email to participate in a one-hour interview. Our professional networks primarily include research and

community partners who we interact with through community-engaged research, ethics and policy work, and open science collaborations. We intentionally selected participants with whom we had established relationships to maximize the potential for trust and openness in the interview, increasing the likelihood of collecting authentic data from community leaders.

We asked interviewees to provide referrals, which led to snowball sampling and additional recruitment of eligible participants. The populations of focus included executive directors and research directors of community-based organizations serving underserved or underrepresented communities, and leaders of communities involved in citizen science, patient-led research, or the quantified self-movement, which refers to the use of wearable digital tracking devices that enable users to track information about themselves and obtain real-time feedback (34). All prospective participants were invited by email to join the study and asked if they would be willing to participate in a recorded interview that would later be transcribed, de-identified, and used for data analysis. Participants were informed that no compensation would be offered in exchange for their participation. Those who responded affirmatively to the email were considered part of the study informant pool and they were scheduled for an interview with a member from the research team.

2.3 Interview Procedures

One of two researchers (MD, JHY) conducted each interview; 10 of 16 interviews were also attended by one or more additional members of the research team. We used this methodology to ensure there weren't any significant differences between the interviewing styles, to allow the interviewer to focus their attention while other members took detailed notes, and for training

purposes. Interviews were scheduled for one hour and were audio-recorded for transcription. Our interview guide prompted participants to describe their community or personal research priorities, the ways in which the use of BD might intersect with these research priorities, and the types of knowledge, skills, tools, and supports needed by their communities to achieve these goals. Each question was elaborated with definitions, examples, and/or follow-up probes to encourage further elaboration. For the purposes of our discussions with key informants, we defined BD as any complex dataset that has genomic and diverse phenotypic data (e.g. Fitbit data or environmental data), regardless of the number of people involved in the dataset. Given our diverse group of community stakeholders and our intention to allow for as much flexibility as possible for stakeholders to draw on related prior experiences, we did not refer to any specific dataset during our interviews, rather we referred to BD more broadly as a concept (See **Appendix A** for full Interview Guide).

Following their interview, we sent a short demographic survey via email to each participant, collecting information such as educational degree, profession, self-identified race and gender, age, and familiarity with genetic research and community engagement. To promote respect for how people self-identify, we encouraged participants to cross out existing options and write in new options as needed (see **Appendix B** for full Demographic Survey). We asked participants to send us back the completed surveys as soon as possible and we sent reminder emails to participants if we did not receive their survey within a week and a half of their interview.

2.4 Data Analysis

Interview transcripts were de-identified using participant identification numbers with other identifiable information redacted. We analyzed the transcripts using content analysis, a form of

qualitative inquiry that seeks to identify, distill, and characterize themes, ideas, and topics from various text sources (35). First, the interview transcripts and audio files were reviewed by the investigative team to resolve any transcriptional errors and to capture broad themes reflected by the whole dataset of interviews. Next, a directed content analysis approach was used to derive a coding framework that was informed by the *a priori* research questions, the interview guide, and review of relevant background literature. The coding framework consisted of six *a priori* content codes: "Use Cases", "Desires & Visions", "Tools & Supports", "Barriers", "Facilitators", and "Attitudes" (defined in **Table 1**). These content codes were used to structure our approach for identifying the range of responses and themes within passages that pertained to each content area of interest. Emergent themes were also extracted from the interviews that reflected novel or recurrent insights that did not fall within any of the content codes, but appeared relevant to the overall project aims. The primary coder (SG) was responsible for coding all transcripts and consulted with the two other members of the analysis team on a weekly basis to discuss codes, findings, and themes over a 3-month time period.

The coded dataset was then analyzed via an iterative process of decontextualization and recontextualization (36). First, the coded content from each individual interview was extracted to a data matrix that linked the coded quotations with the corresponding content code and participant ID. Next, we aggregated the coded content from all participants and sorted the coded quotations into separate summary documents corresponding to each content code. Investigator SG took detailed notes in each of these documents, noting the range of themes represented within each of the datasets (decontextualization). Finally, we recontextualized our findings by considering how the context of the participant, their experiences, and their identities, may aid in

our interpretation of the quotations they contributed with respect to other similarly situated individuals (recontextualization). For example, when interpreting a prevailing theme within a particular content code, we considered if there were any identifiable relationships between certain experiences of participants and the types of themes or viewpoints they contributed to the dataset.

Table 1. *A Priori Content Codes, Definitions, and Examples*

Content Codes	**Definitions**	**Examples of Themes Reflected by Coded Content**
Use Cases	Examples of communities interacting with BD research or using BD for their benefit	Use BD to improve the precision and quality of screening, treatment & prevention options
Desires & Visions	Community visions/desires related to engagement in BD research	To develop *authentic* collaborations between communities and outside entities
Tools & Supports	Current, desired, or previously used tools/supports related to community engagement in BD research	Data tools and trainings
Barriers	Anticipated or experienced barriers/challenges related to community engagement in BD research	Competing community priorities and limited resources
Facilitators	Desired or experienced facilitators related to community engagement in BD research	Trust in community leaders and researchers
Attitudes	Community attitudes related to engagement in BD research	BD currently lacks bidirectional benefit

3. Results

3.1 Participant Interviews

Twenty-one key informants were contacted by email between February 21-27, 2019 to participate in a one-hour interview. Eighteen informants responded affirmatively, two of whom agreed to participate if needed, but asked not to be prioritized as informants due to their relative unfamiliarity with the subject. We received no responses from three of the 21 invited informants. In total, 16 interviews were completed between March 5, 2019-May 22, 2019, ranging in length from 44 minutes to 71 min (mean: 59 min; median: 60.5 min). Due to a technical error, one interview was not recorded or transcribed, so detailed notes taken during the interview by investigators MD and SG were used in place of the transcript for the analysis. Our final list of informants included five additional contacts who we did not contact but we intended to contact if our target number of interviews (n=15) was not reached after the first round of email invitations.

3.2 Participant Characteristics

Demographic information from 14 of the 16 key informants are shown below in ***Table 2.*** We did not receive surveys from two of the 16 informants. The informants ranged in age from 30 to 60 years old, with an average age of 46 years old. The majority of key informants self-identified their gender as female (64%), their ethnicity as non-Hispanic/Latino (79%), and their race as white (57%). All informants were high school graduates and more than half had completed a post-graduate degree (57%). The informant pool represented a wide range of occupations, including community health and patient advocate, researcher, data organizer, physician, professor, and consultants, among others. All informants had at least some knowledge of genetics and 71% stated they felt well-informed about genetics and 29% stated they remembered

some genetic information from school. A total of 43% of informants stated they worked at a community-based organization, and the populations served by these organizations were fairly equally distributed across American Indian/Alaska Native (38%), Asian (38%), Black or African American (23%), Native Hawaiian or other Pacific Islander (38%), White (23%), Hispanic or Latino (31%), and "Other" (23%) communities.

Table 2. *Demographics of Community Leader Key Informants*

		n=14	
Characteristic		**Frequency**	**Percent (%) of Study Population***
1. Age	Mean age (range; SD)	46 (30-60); 9.5	
	32 and under	1	7%
	33-41	3	21%
	42-54	2	14%
	55 and over	2	14%
2. Gender identity	Male	4	29%
	Female	9	64%
	Transgender	1	7%
3. Ethnicity	Hispanic or Latino	3	21%
	Not Hispanic or Latino	11	79%
4. Race	American Indian/Alaska Native	1	7%
	Asian	3	21%
	Black or African American	1	7%
	Native Hawaiian or other Pacific Islander	1	7%
	White	8	57%
	Other	2	14%
5. Occupation	Medical/Genetics/Public Health Researcher; Community Health Advocate; Community Data Organizer; Web Developer; Independent Researcher; Policy Advisor/Deputy Director; Full-time Volunteer for Rare Disease Foundation; Consultant; Patient Advocate; Attorney; Physician; Board Member of Rare Disease Foundation; College Professor		
6. Highest level of education completed	Did not complete high school	0	0%
	High School graduate/GED	1	7%
	Some college	0	0%

	College graduate	5	36%
	Post-graduate (e.g. M.A., M.S., M.D., PhD)	8	57%
7. Statement that best describes knowledge of genetics	I know nothing about genetics	0	0%
	I remember some information about genetics from school	4	29%
	I am well informed about genetics	10	71%
8. Community-based organization?	Yes	6	43%
	No	5	36%
	Doesn't apply	3	21%
9. Description of community or stakeholders	Community health centers and organizations that prioritize underserved AIANHPI populations; Online communities of previvors and survivors who have hereditary cancer mutations; Caregivers & patients with a rare genetic syndrome; chronic disease community and other health communities interested in doing data-driven work; Patients residing in Southern California comprised of 82% Hispanic/Latino heritage, 75% living at or below 100% of Federal Poverty Level, and 30% uninsured; A number of tribal communities within the United States; Native Hawaiian and Pacific Islanders in the US Pacific, Hawaii and Continental US; Rare genetic disease foundation and all families and patient members served by this US non-profit organization; LGBT, and transgender health communities.		
10. Number of employees at organization	<10	4	29%
	11 to 50	3	21%
	>50	4	29%
	Doesn't apply	3	21%
12. Organization's approximate annual operating budget	<1 million	4	29%
	1 to 3 million	3	21%
	>3 million	3	21%
	Doesn't apply	4	29%
13. If applicable, populations served and/or represented by organization	American Indian/Alaska Native	5	36%
	Asian	5	36%
	Black or African American	3	21%
	Native Hawaiian or other Pacific Islander	5	36%
	White	3	21%
	Hispanic or Latino	4	29%
	Other	3	21%

14. Ever had genetic test?	Yes	9	64%
	No	3	21%
	Don't know	1	7%
	Pending	1	7%
15. Ever participated in biomedical research?	Yes	8	57%
	No	5	36%
	No response	1	7%
16. Ever participated in genetic research?	Yes	5	36%
	No	8	57%
	No response	1	7%

**Some percentages add to greater than 100% due to the "select all that apply" option.*

3.3 Analysis Findings and Framework Description

Key findings from the directed content analysis are summarized by each *a priori* content code in **Table 3.** Analysis findings were extrapolated from the dataset of interviews using our analytic approach to develop a conceptual framework that details the components and pathways involved in the authentic engagement of communities in BD research. This BD Community Engagement Framework, depicted in **Figure 1,** contextualizes the processes by which communities move from the first stage of *organizing their members and gathering information to consider how BD could aid or intersect with their research priorities* to the last stage of *translating evidence generated from BD research for direct community benefit.*

Framework facilitators are depicted in green with plus signs and represent those factors that positively reinforce actions along a pathway. *Framework barriers* are depicted in red with negative signs and represent those factors that negatively reinforce factors along a pathway, based on our analysis findings. The *framework facilitators* and *framework barriers* were directly informed by the "Facilitators" and "Barriers" coded content in combination with the other "Tools

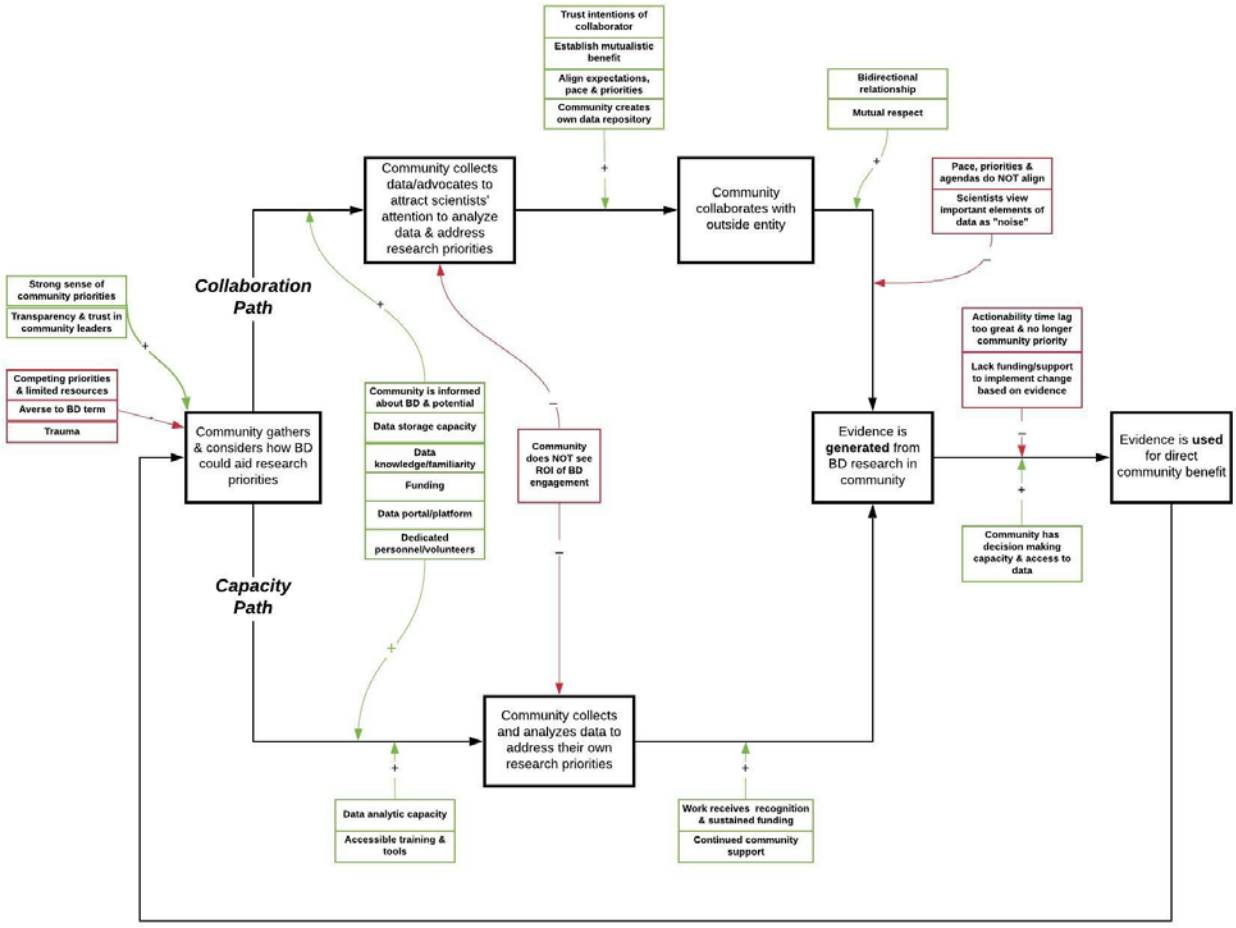
& Supports", "Desires & Visions", "Attitudes" and "Use Case" coded content. For example, a *framework facilitator* such as *the existence of transparency and trust in community leaders* was informed both by coded quotations that illustrate how the lack of transparency and trust in community leaders serve as barriers for community engagement with BD ("Barrier" coded quotations), and by quotations that illustrate how transparency and trust in community leaders facilitate pathways toward community engagement with BD ("Facilitators" or "Desires & Visions" coded quotations). The *Framework stages* (indicated by the six black boxes) were also informed through a holistic picture of all the coded content and notes taken during the interviews by the investigators. The BD Community Engagement Framework is elaborated with supporting quotations in the subsequent sections.

Table 3. *Overview of Key Themes Reflected by a Priori Content Codes*

Use Cases	• Improve screening, treatment & prevention options • Nuanced risk prediction & decision-making tools • Investigate broad range of health determinants • Provide community agency via data access
Desires & Visions	• Legitimize career paths for citizen scientists & patient advocates • Improve science literacy • Consolidate data resources • *Authentic* collaborations • Accessible & affordable genetic tests
Tools & Supports	• Mentorship & partnerships • Reliable funding streams • Data tools & trainings • Educational toolkits for community members • Data sharing platforms
Barriers	• Disconnect between healthcare, research & community needs • Data capacity challenges • Competing community priorities • Unfamiliarity with BD & research • Fear, trauma & mistrust
Facilitators	• Collaboration frameworks & shared resources

	• Interoperability & centrality of data • Trust in leaders • Political will
Attitudes	• Data are valuable & personal • BD currently lacks bidirectionality • Superficial community engagement • Data interpretation requires cultural context

Figure 1. *Big Data (BD) Community Engagement Framework*



3.3.1 Community consideration of BD

The first identified step represented by this framework is the process by which communities convene to discuss and consider how BD could aid or intersect with community research priorities and needs. Based on responses from informants, convening processes varied based on the type of community. For example, within the rare disease community we heard accounts of family members who actively sought out BD as a potential solution for uncovering as much information as possible to address essential questions about their family member's rare disease. Other groups nucleated around a particular question or topic of interest such as a shared culture or interest in citizen/community science or biohacking, and then later identified a way in which BD could aid their community's priorities.

For some communities, this convening process around BD formalized over time through a series of organized meetings, summits or trainings. One community leader in particular spoke about the value of convening annually in a formalized setting to determine how BD could intersect or aid community research priorities,

> *"**We have this gathering every year so that we can talk to the community [to hear] their priorities**. And then the other interesting thing is what we do is that we also […] **put out that data, you know, all the major data. And then after that, there will be a plenary session on gaps, research gap, program gaps, you know, in response to the big data and then we have those community town hall type of session** where communities respond. You know what I mean? 'Cause it's not just conversation. It's like a dialogue." [P10]*

As represented in the framework, before a community can even begin to consider how BD might intersect with their priorities, groups must internally align their community priorities and community leaders must demonstrate trustworthiness through practices such as transparency. Though we did not hear any explicit commentary on the process by which communities crystalize their community priorities, we can infer from our dataset that this is an organic process

for some communities and occurs in a more formalized context for other communities. One community leader explains how the absence of this community vision prevents full investment in a particular research direction right from the start,

> *"Until there is a **sense of congealing and collective sense of vision, and goals, and priorities, then we actually are this kind of multi-armed octopus that's just kind of going in so many different research directions** that perhaps are not aligned towards a common aim or a common focus." [P3]*

One barrier informants described was how competing community priorities and limited community resources halted any progression past this initial stage of the process. Specifically, if community leaders were focused on sustaining the basic operations or funding streams to keep their communities afloat, then they may appreciate the idea of engaging their communities in BD research but find it unfeasible from a pragmatic standpoint. For example, an informant explains,

> *"From a very idealistic perspective, I think having that in-house capacity for community-based organizations to be able to generate, contribute, analyze, and **utilize their own data to contribute to the broader research field would be really, really powerful. But I also from a pragmatic perspective recognize that organizations are just struggling to stay afloat.** And so, finding resources to support staffing to actually just provide the services is a very real challenge for our communities." [P3]*

This same informant goes on to describe how difficult it is to envision how their community might engage with BD amidst all the other individual competing priorities of community members,

> *"**I have a hard time envisioning what that looks like when it comes to big data when so many members of our communities are already having a hard enough time just navigating their health care coverage** or navigating when they have a particular issue. And so, I mean, well, you might be interested in seeing what the data says or seeing what the research says. **The day of day of trying to just make it through another day with what's about you with your health about you, I think it's a huge obstacle to doing that level of reading and analytical research on an individual level to be able to form those pieces."** [P3]*

Similar sentiments were echoed by another informant who explained how even if the community may be intrigued by the potential of BD research, UBR communities can be too preoccupied

with addressing pressing day-to-day concerns to even begin to consider what engagement with BD might look like,

> *"I mean, what's so hard in UBR populations [is] they're not only worried about health. They're worried about stuff that a lot of us don't worry about and to get them engaged and to pull them away long enough to consider some of these new opportunities. You know, it is a day-to-day challenge, they face a lot because* ***if you say "I wanna talk to you about big data, they say "Well, I wanna talk to you too about big data." But I wanna talk to you about this, and that, and other because these things are even more pressing than what you wanna talk to me about… I mean, they're worried about their kids. They're worried about their parents. They're worried about their relatives in other countries. They're worried about their day-to-day jobs. They're worried about their day to day safety and neighborhoods are not safe. And you come to talk to me about big data? It's like take a seat. Take a number."*** *[P11]*

Another barrier that surfaced a number of times during the interviews was that "Big Data" is seen as a buzz word and can seem elitist and inaccessible to many communities. Consequently, when a community is averse to this term and related terms, they do not see the value in considering points of intersection between BD research and their community priorities. This phenomenon was bluntly explained by an informant who interjected mid-way through their interview,

> ***"Whenever you say big data, like I just kind of zone out.*** *Big data feels like not relevant to me but now I see that big data is relevant to us.* ***I don't know if that's a problem in terms of how we're trying to go out there. I never think of [Facebook, Google, or Amazon data] as big data. But that's exactly what we would like to be doing."*** *[P15]*

Another community leader details how this term seems daunting and too riddled with expenses and complexities to initiate engagement,

> *"And you know, big data is more mysterious than—* ***You know, it's a very daunting thing for those of us to understand. You know, big data sounds like it's pretty darn expensive and it's pretty darn complex*** *and those two variables need to be addressed upfront as soon as possible." [P11]*

One informant explained the value of breaking down the why, the what, and the how of BD in addition to the need to describe the potential dangers related to privacy and security by using plain language in order for communities to even begin to consider BD engagement,

> *"So, I feel that, first of all,* ***members of the public and rare disease groups and particularly underrepresented biomedical research, they need to first understand the why, what, and how of big data. And we must, you know, kind of deconstruct for them in very simple, plain language what this entity known as big data really is and why so many are engaged in it and*** *also, you know, to be quite transparent about both its potential and also it's dangerous with regard to privacy and security." [P11]*

Another community executive reiterated how familiarizing community-based organizations with the variety of ways BD might intersect or aid their research priorities may be a necessary ingredient for initiating this first step of engagement,

> *"My sense is at a community-based organization level,* ***folks have heard of big data, but may not necessarily be fully aware of what… how big big data is...of what the opportunities could be in terms of being able to help inform their efforts.*** *So, you know, perhaps that's also just a* ***low hanging fruit type of piece in terms of just familiarize community-based organizations and other leaders about just what big data is, where it's headed, and the implications for our communities as well as the opportunities."*** *[P3]*

In contrast, some communities identified that they don't even have enough access to technology to generate BD, so efforts to familiarize them with all the ways BD might aid their priorities or explaining BD concepts in plain language can only go so far and may actually contribute to widening the digital divide,

> *"Well, I was gonna add that maybe even though it's* ***difficult to collect these types of information on tribes, even though there's difficulties in that process, for now, maybe that's not such a bad thing because too often what we see is a rush in the technology and a huge lag in the ethics and the policy discussion afterwards.*** *And at least even there are barriers to collecting information in tribal communities, at least it's not outpacing our discussion." [P16]*

Participants also highlighted that emotional trauma from a lack of transparency and a history of resarch malpractice posed real barriers for communities considering BD engagement. This

included trauma related to superficial collaborations that took advantage of community members, trauma related to the lack of autonomy over use of community data and specimens, as well as "academic trauma" related to a lack of science and health literacy within communities. One informant described how this concept of "academic trauma" functions as a real barrier for members in their community because they were conditioned based on past experiences in school to believe they were not smart enough, could not succeed, and/or had no role interacting within an academic environment,

> *"So many of our people have **academic trauma**. They were told in school they were less than. School is designed to weed people out. Not to educate everybody for the most part. **So then, we come with data and we come with things that's loaded in fraught with all of the trauma that people have about information, and research, and school.** So, if we're committed to making things accessible so people can use it and feel safe, we have to think about that." [P9]*

Two community leaders elaborated on how their communities feel too vulnerable to participate in BD research because of past and current trauma related to the lack of ownership over their data and lack of informed consent for data collection,

> *"The tribes have sovereignty and that's what makes them not only reluctant and distrustful of participating in big data, but it also makes them very vulnerable to big data because it gets collected anyway...And often, **they don't know it's happening or it's getting collected just off the reservation at say a major hospital in the area. And once they collect Native American data, it's completely out of this jurisdiction.** At least with the university or the federal government, they have some chance of getting it back. But when it's a private HMO, there's no chance they're gonna get it back." [P6]*

Another community leader reiterated the challenge of regaining the trust and interest of communities' participation in BD research because the history of research abuse that their community experienced has still never been adequately addressed by the research community,

> *"**Or to trust for the first time. Some of these communities were never informed. So, they were never addressed. They were basically, you know, thrown into a study never been informed. You know, they would just kind of ignore it. Their individual concerns were never part of the equation. So now, we're saying, "Oh, you matter." And they're***

> ***saying, "You never said that before."*** *And so, you're asking them to trust again, so the how and the why are very, very formidable challenges." [P11]*

We recurrently heard from informants during their interview that it felt like a therapy session since it allowed them to finally unload all the pent up trauma, vulnerability, and frustrations that they have experienced in past research efforts. For example, one informant reflected on this toward the end of our discussion,

> *"**I tend to think that I'm not gonna be very useful because I'm not really a professional, but then I was like…They are kind of like therapy for me**. So, thank you for listening." [P15]*

After another informant described a number of efforts they had exerted to make BD research accessible and approachable for their community, they then sheepishly admitted,

> "***This is where I told you this would be therapy."*** [P2]

What is made clear by our analysis is that communities may have been unable to identify intersection points between BD and their community priorities due to some of the described barriers, a lack of some of the described facilitators, or a combination of the two. Some communities may have moved past this first stage of considering BD engagement at one point in time and due to a myriad of experiences, including failed collaborations, re-traumatization, misuse of data, reprioritization of community needs, or broken trust with community leaders, then they find themselves back at the first stage of the process.

3.3.2 To engage or not to engage, that is the question

Communities who move past the first stage and formulate ways in which BD could aid their community priorities will follow the "Collaboration Path", the "Capacity Path", or attempt the "Capacity Path" after a failed attempt at the "Collaboration Path" (shown in ***Figure 1***). The "Collaboration Path" in the framework represents a community who collects their own data and

advocates to attract the attention of qualified scientists to analyze their data and address their research priorities for them. The "Capacity Path" represents a community whose members collect and analyze their own data and independently addresses their own research questions.

Both pathways are facilitated if a community is informed about BD and its potential application within their community, if there is some data storage capacity accessible to the community, if there is some level of funding to sustain the organization as they begin this project, and if there are dedicated personnel and volunteers who are committed to seeing the project through. A defined barrier for both pathways is if a community is unable to recognize the return on investment (ROI) of engaging in BD research. One key informant describes the importance of demonstrating the ROI of BD research engagement and how it can be achieved by illustrating tangible examples for communities and returning high level project results back to participants to keep them engaged and invested in the research,

> *"So, it **will be good to have more specific examples I think for the community [so] that they can easily understand what is the value to be part of a research study**. It's not only to say that you will contribute with your information to create the data to make available for the researchers […] **then what happens? What is the value for me as a participant? So, I think something that will be important [for] the All of Us research program is […] to provide results at the high level of the project to the participant**." [P4]*

The primary difference between the "Collaboration Path" and the "Engagement Path" represented in the framework is the decision made by communities on the "Collaboration Path", to attempt to engage an outside entity (i.e. an organization, health center, data scientist, research team, foundation, etc.) to form a research collaboration. The first step may require attracting the attention of an outside entity to agree to a collaboration which then creates one of the greatest bottlenecks within this framework. Informants repeatedly stressed how many administrative hoops they had to navigate just to get a potential collaborator to take them seriously and consider

what they had to say or what they had to offer. One key informant describes how hard they must continue to work just to get researchers to acknowledge their community's efforts,

> *"We figured out that our researchers would not use our data if we didn't collect it under an IRB. And then now, it's become a thing that everybody in the platform gets an IRB.* ***So, we're doing things by the rules, but I think it would just be— You know, what it takes me 10 hours to do, somebody who had a PhD could probably do in a half hour with their eyes shut.****" [P8]*

Another informant describes their failed attempts at attracting the attention of outside researchers even while investing all that they have to establish a collaboration,

> *"It's just* ***we keep beating our heads against this wall and making new changes with bubble gum and tape and no resources****, no PhDs, and then everybody writes the papers about us.* ***Everybody goes off and builds the platforms based on our ideas. And nobody ever actually partners with us."*** *[P1]*

After attracting the attention of an outside entity, the facilitators for successfully establishing a research collaboration and initiating engagement between a community and the outside entity include trust in the collaborator's intentions, a defined mutualistic benefit between groups, and an alignment of expectations, pace and priorities. One community leader explains how the onus is really on the community to be informed and do their own research about the rules of collaborations in order to appear credible and attractive to the potential collaborator,

> "*But the first thing is that you, as a patient group or as a community,* ***you've got to make yourself approachable, attractive, and a credible partner who's willing to play the give and take game of every relationship.*** *You have to know when you need to compromise and when you cannot compromise. Yeah. Political savvy and scientific aspect.* ***You've got to be really well informed. A really well-informed partner that it all begins with mutual respect and trust."*** *[P11]*

Many informants echoed how difficult it was to establish collaborations with community groups due to the existing research infrastructure that is not designed to support the type of collaboration that communities expect and desire. One community leader described that the community

engagement model is so difficult for traditional researchers to accept because it requires a completely different skillset that is not rewarded by traditional research incentives,

> *"**Doing community engagement is a totally different skillset.** And it's tough and that's **not** how [traditional researchers] get promoted. They get promoted in the academic space based on publications. **And so, the incentives are not lined up to say, "Hey, let's take a risk. Let's try to do this differently and work [directly with communities]."** That's just not how it's set up."[P12]*

Similarly, this particular informant unveiled how patient involvement in research is often implemented in the most superficial ways that are dictated by grant requirements and are not designed with the intention of providing value or autonomy back to the patient communities,

> *"There's been this **kind of idea attached to funding and I'm starting to actually see it like written into grants for researchers that you involve patients, I'm seeing it done in the most superficial ways.** And so, I don't know how funders should manage that, but it has been frustrating. You know, I'll go like an NIH meeting for organizations like mine or a rare disease day meeting and **people talk about how they're involving patient input, but I don't actually think that's happening. It's being implemented in the most superficial ways and it's very frustrating to watch because it's like you're being told you have something, but you don't actually have it.** So, because you're talking to funders, that's the thinking about how to **make sure that these relationships are meaningful to the groups that they're supposed to be serving."** [P15]*

After a community research collaboration is established, the real challenge expressed by informants lies in maintaining and sustaining an *authentic* research relationship. Informants stated that the sustainment of these relationships requires a continued bidirectional and respectful relationship between the community and the collaborator. If the pace, priorities, and agendas do not ultimately align once the collaboration has begun, or if the collaborators and community do not share the same perspectives regarding the important elements of the collected datasets, then evidence that directly benefits the communities will rarely be generated from these BD research projects. An example of a misalignment of research priorities is illustrated by an informant who had already curated an entire BD set for researchers to use and provided free tools and tutorials

on how to utilize the dataset. This informant describes how as a member of a patient group, they are doing everything in their power to expedite the research process to ask the questions that are most pertinent for their community, but the traditional researchers are not accustomed to the richness of the dataset or the types of valuable questions that can be asked of the dataset,

> *"And **we have this whole long list of questions that's so interesting and then traditional researchers are just like way back like 3 years ago in this journey of like "Oh, hey, this data exists.** It's not too crazy to use it." **Like we can start asking questions of it, but they're 3 years ago of where we were in terms of understanding how rich the questions you can ask now that you have this rich data.** And like I said, **they're kind of just like getting accustomed to how rich the data is and they still haven't really understood all the potential of it**. And part of that is us just being on the bleeding edge like I get that, but it's been like a long uphill battle pulling them uphill with me to like try to get them to the next level of understanding "Okay, we've cleared out the noise for you. Here's what you can study and here's what you can learn from it." Is it like universally going to be true? Not necessarily, but that's why you do research. Figure out. Here's the hypothesis. Here's what we've learned. Can we test this? What population does this work for?" [P2]*

Another patient advocate community leader describes how much effort it requires to try to align perspectives regarding the important elements of the datasets and analytic approaches for answering the questions that are most relevant to the patient population,

> *"But what we learned is it is a **huge amount of work to onboard a data scientist to really understand the [disease] context and the data context and also data scientists who are traditionally trained have traditional methods and want things to fit in nice boxes and be clean** like, well, this is the way we do the analysis because this kind of the problem. What we learned is, well, especially when we're talking about clearing away the noise, **we have new problems that people haven't been able to analyze before. So, this is going to involve new methods of analyses**." [P1]*

A common thread expressed among community leaders on the "Collaboration Path" was this interest in improving the authenticity of the research collaborations so communities don't find themselves coerced into a collaboration that does not have their community's best interest in mind. One informant describes their experience advising their colleague against this type of inauthentic collaboration,

> *"I pointed out the clause about patient involvement and I was like this grant is requiring patient involvement.* ***The researcher who came to [my colleague] had already designed and conceived of the entire thing. They haven't actually like worked with her to develop the idea. They had the idea. They had fleshed it out and then I think they felt like they were gifting her with the opportunity to partner on it."*** *[P15]*

Another informant explains that there is a real art to engaging rare disease groups or UBR populations in research collaborations that requires both parties to be realistic about their desires and expectations and maintain a mutual respect,

> *"So,* ***the researchers have priorities and objectives that are equally important as ours. It's not about separatism. In my mind, it's about asking us what our priorities are, what the researcher's priorities are, and then figuring out a way to align interest and resources towards both****...It's truly we just need to figure out a framework for partnership and equal footing that says here are the community's goals, here are the researcher's goals. And there's gonna be a give and take and shared set of resources that also meets the priorities of the patient community. Not just the priorities of the researchers." [P1]*

Communities on the "Capacity Path" who choose not to engage an outside entity and decide to analyze their own data can ultimately generate evidence from the research they are conducting if their work receives recognition, if there's sustained funding, and if the community remains engaged in the work. Without these facilitators, communities may be unable to translate the data they've collected to address their community research priorities. One informant described their difficult experience along the "Capacity Path" working to translate their research and receive recognition for their work in the absence of outside collaboration,

> *"All of us in the broader e-patient movement know each other.* ***And we've all done this work. And we clawed our way to like fine some sense of a path for ourselves that is never ever valued.*** *I mean, we've created all this value in the healthcare system through patient engagement.* ***So, for us to wake up and recognize how little that is actually respected or how little rights we really have after all the work we've done, it's hard for everybody I'll tell you.*** *And so, in my mind, we do need to create a path to legitimacy." [P1]*

We also heard a number of informants who attempted the "Capacity Path" after a failed or frustrating attempt at the "Collaboration Path". One community leader of a patient advocate group spoke about how they were burned so badly in the past by research collaborators who

would not cooperate with them, that they no longer had any interest in investing in research collaborations,

> *"**Like I really don't have a lot of interaction with them. And at this point, like I don't even trust them. I've been working at this for long enough that if they just turned around and suddenly wanted to work with me, I don't really wanna work with them at this point, but that's been a huge bottleneck for us.** And so, it's part of the reason that we've had trouble building capacity because I went to them early on. And I said, **"You know, we need some discreet projects to fundraise around so that we can build the financial capacity of the organization." I couldn't get anything out of them.** So, we have not been able to say we're funding research because we really weren't. **And it wasn't because we weren't trying. It was because they wouldn't cooperate with us."** [P15]*

Thus, the overwhelming feedback we accumulated from informants supported the observation that the decision between the "Collaboration Path" and the "Capacity Path" is not necessarily calculated. Most communities are interested in the "Collaboration Path" at the start, yet only a select few are able to successfully establish the type of authentic collaborations with outside entities necessary for continuation on this pathway.

3.3.3 From evidence generation to direct community benefit

Our analysis findings reinforce that just because evidence may be generated from BD research in or from communities, this alone does not ensure that the evidence is ultimately translated for direct community benefit. This final step is facilitated only if the community has decision making capacity or access to their data. An excellent example of this phenomenon is explained by a community leader who details the situation when communities do not have possession over the data collected from their communities,

> *"Well, I think that, you know, **just like the old adage that possession is 9/10 of the law. Possession of the data gives you a lot of leverage. And what the tribes don't have is any leverage to enforce, like I said, intervention.** And if the federal government doesn't wanna create that venue for intervention, then they should be creating a place to build capacity. **And so, the tribes often don't have capacity to utilize or analyze the data for***

> ***themselves. So, somewhere, you know, the dam has to give way either towards intervention or towards capacity building so that they can use it for their own benefit."*** *[P6]*

This same community leader elaborates on how a lack of ownership leads to a lack of "bidirectional texture" of the data which ultimately prevents communities from leveraging BD for their own benefit,

> *"Maybe it's just an Orwellian fear that everybody has, but* ***I think it's becoming more real than not, is that all of this big data, it doesn't have any bidirectional texture to it. And so, for this stuff, even if we were to participate in a study about herbicides, and pesticides, and genomics, we could have all the data in the world. But if we don't have decision making capacity or we don't have the ability to have our own copy, somebody will say, "Well, that data doesn't exist." Even though it does exist, they just made the data disappear."*** *[P6]*

Another key barrier at this stage arises if communities lack the funding or support to plan, install, and implement change based on the evidence generated. One common example of this type of barrier expressed by informants was the lack of resources and capacity to plan for, troubleshoot, and deal with the lack of data interoperability. As a community leader describes below, just generating data does not ensure that the data are accessible in a useful format for deriving community benefit,

> *"****It's the interoperability of the data and the real time in retrospective access to it. Those were the key components that unlock everything else. And I feel like those are the ingredients for the recipe of the success for any community,*** *even one with less resources and even ones that are considered to be less data driven. I think actually there's going to be a lot of data that will uncover from these communities about symptom tracking, medication tracking, things like that* ***where we'll get a ton of insight if we can free the data to the right people who can say, "Aha, here's the problem and here's the solution based on what the data is telling us."*** *[P2]*

A second important barrier our informants identified is when the time lag between research initiation and evidence generation is so great, the evidence may no longer be relevant to community priorities and/or accessible to community members. The quote below shared by a

community leader demonstrates a community's experience of harm when there is a considerable time lag,

> *"And so,* ***if you don't have the data to show that that's actually causing, you know, worse health outcomes and that kind of stuff, then you can't go to Congress and lobby on your own behalf.*** *There's 5 tribes that went to court against the army corps of engineers.* ***The reason why they lost is they couldn't find the data.*** *There's data out there, but it wasn't accessible to the tribes even though it was collected on reservation land and it was presumably for reservation benefit." [P6]*

Ultimately, only a small minority of key informants expressed that they had ever reached this final stage of returned community benefit from BD research. Community stakeholders provided important insights as to potential reasons or impediments to this lack of translational benefit. First, informants frequently referenced the fundamental importance of being aware and understanding of culture and context in order to translate data into community benefit, otherwise, the data are just numbers. This concept was eloquently stated by a community health leader in our informant pool,

> *"****Data isn't enough. Just because you could have that data but if you don't have a sense of self-efficacy or self-determination to be able to advocate it for yourself,*** *if you are not the type of person or* ***if your culture teaches you to face and respect positions of power or authority****, in this case a doctor, to raise questions or concerns about what your doctor is prescribing, then having that data again is of no use." [P3]*

Another insight made by a community leader was that UBR communities feel a certain level of vulnerability when engaging with BD research because of the cultural context that is inextricably linked to their data and the perceived risks of being targeted and profiled based on this information,

> *"Obviously, there's been* ***increasing concern about data tracking, about social media, about the ways in which so-called gang databases are put together using big data. So, there's like increasing mistrust.*** *Right? And in one hand, people are super connected in using these social media tools.* ***On the other hand, you know, it feels really scary for them. Right? It feels really vulnerable. And for young people of color who are at risk of being targeted and profiled, it feels really dangerous in some ways****. So, you know, there's*

> *some real concerns. I think any effort to engage what big data can do…**And when you think about data, and you think about translation, and you think about use and application, you know, culture matters and cultural competency matters. Right? So, I guess that's one thing."** [P9]*

As the final stage in the framework illustrates, if the evidence generated from BD research in communities is ultimately used to generate direct community benefit (whether to inform the implementation of targeted public health programs, to lobby policy makers on the community's behalf, or to leverage further funding for community needs, etc.), then this reinforces the process back to the first step in the framework which is *organizing their members and gathering information to consider how BD could aid or intersect with their research priorities*.

Specifically, once communities experience the ROI of engagement in BD research at this final stage, then this has the ability to build communities' sense of autonomy, self-governing capacity, and empowerment which reinforces communities' interest to reengage in this pathway again. One informant sums up this reinforcing loop by describing how community ownership and self-governing capacity over their data inspires a sense of autonomy to utilize BD as a tool to make data-driven decisions as a community,

> *"And **the only thing we can control is our own choices. And to me, that's what autonomy is.** And our individual choices and our collective choices as a group are the one thing I'm trying to protect right now to be honest... **If we don't like have autonomy over our data, over our community, and all of these things in big data start making decisions about us without us**. That is the opposite of autonomy. And we don't want a future or medicine like that. We want to be the one to have the path to learning and who are the beneficiaries of the knowledge so that we can make our choices and our decisions." [P1]*

4. Discussion

In this semi-structured content analysis of community stakeholder interviews we sought to characterize meaningful uses of BD ("use cases") from the perspectives of community leaders and identify what is needed by communities to increase capacity for community-driven BD research. Based on our review of relevant literature and our prior work in community-engaged research, ethics and policy research, and open science collaborations, we did not expect to uncover one perfect solution from our interviews that would apply universally to all communities in every context. We did expect to hear common themes and experiences from community stakeholders about barriers to using BD. Specifically, we anticipated hearing that BD barriers include privacy and trust concerns with personal identifiable information (PII) (37), fears of data misinterpretation due to missing information (38), barriers to data access and sharing (39), lack of data science training and resources for data analysis and interpretation (40), and trauma from a lack of transparency and a history of resarch malpractice (28). Our findings reinforce the perception that these factors pose significant barriers for community engagement and autonomy within BD research. Our interviews also provided a more granular understanding of the underlying factors that perpetuate and reinforce these barriers. Two of the nuanced findings captured by our analysis that dominated our discussions and created the greatest bottlenecks for communities in the Framework included issues related to the "digital divide" that persisted at all levels of BD research and issues related to establishing authentic collaborations between communties and researchers.

The concept of the "digital divide" is especially relevant to BD research given the various data sources and technologies used for BD collection and analysis including electronic health record

(EHR) data, clinical registries, lab tests, insurance claims, and genomic data (41). The "digital divide" usually describes inequalities that arise as a result of an uneven distribution of access to information or technology. In the context of BD, the digital divide has also been used to refer to the inequalities that exist between data donors and those who analyze and interpret the data (42–44). This definition is particularly salient to our research, as the vast majority of the communities we sampled reported they did not have the necessary infrastructure or the advanced analytic training to fully understand how their community's data would be used or analyzed, leading to mistrust of researchers, a disinterest in BD engagement, and also a fear that the community's data might be used inappropriately, thus leading to harm. Most informants discussed how these gaps in data/scientific literacy posed significant barriers for generating interest in BD research within their communities and left them feeling disenfranchised from the research enterprise as a whole.

One finding that we did not anticipate from our interviews was that even the term "big data" itself reinforced the digital divide. Community leaders described "Big data" as elitist, inaccessible, and not relevant, further deterring engagement in BD research. Another novel finding was a "positive" interpretation of the digital divide: that a community's lack of access to technology might mean there is little BD for outsiders to mine. This access lag renders communities less "vulnerable", and can afford communities the opportunity to build their internal knowledge and capacity prior to engaging with BD research. Conversely, while avoiding the potential harms of BD may be a reasonable protectionist strategy in the early-phases of BD research, an absence from BD could indefinitely defer potential benefits from BD insights in an age of precision medicine.

A second concept that community stakeholders frequently discussed was the challenge of establishing and sustaining authentic research collaborations. Research findings support that community involvement at all stages in the research process increases the quality, relevance, and efficacy of research translation to improve individual and population health (30,31). However, determining how best to engage public participation in research remains an ongoing challenge for researchers who are generally not trained to identify, recruit, and convene stakeholders to facilitate this type of engagement (45). Our informants frequently described how naïve and self-serving attempts by researchers to facilitate community and patient engagement efforts led to communities feeling disenfranchised and discouraged to reengage in research collaborations in the future. Another finding from our interviews that aligned with relevant literature, is that current research infrastructure and incentives are not aligned to support community engagement work and this is often a root cause of failed community engagement efforts (46,47).

Some groups aim to address this issue by creating more structured and accountable methods for obtaining input from stakeholders on research design, conduct, and dissemination of findings (27,30,47). However, our community stakeholders described how these existing support systems to encourage community/patient engagement in research (e.g. via grant requirements), do not necessarily translate to *authentic* collaborations. Instead, community leaders spoke about how on paper their communities might appear to be engaged in research collaborations, but in many cases, researchers superficially seek communities' engagement to fulfill a particular research requirement, sometimes even after the study design has already been completed. The number of informants who described the interviews as "therapy sessions" demonstrated just how traumatizing these parasitic research relationships could be for communities. Ironically, the

"research parasite" trope has been used by some researchers in the context of open data sharing platforms to describe citizen/community scientists who use a research group's data for their own ends (48). Diverse community leaders shared stories of researchers acting as the "parasites", dropping into communities to sample only what interested them, ignoring or forgetting the potential or expressed community benefits. With only this partial picture of communities' needs, this type of "collaboration" prevented researchers from designing studies that considered the full picture of community priorities. These insights suggest a need to redefine the scope and parameters of community relationships with researchers who typically hold the power and the locus of control in research and to think more critically about the types of "soft-skills" or trainings needed for traditional researchers to work in community settings.

The BD Community Engagement Framework we developed from our analysis findings illustrates the pathways and various forces for and against community engagement in BD research, as described by our informants. Our research findings reinforce BD engagement as an iterative process that occurs in multiple stages along different trajectories, as opposed to a destination that can be permanently reached or obtained. A minority of key informants expressed they had ever reached the final stage of returned community benefit from BD research. This finding highlights the fallacy that when data is collected it intrinsically provides value. Our framework was intentionally designed to reflect the mobility of communities forward, as well as in retreat, along BD engagement pathways. The Framework reflects both positive and negative community experiences with BD in order to broadly contextualize the arc of the engagement process. Representing our research findings in a framework format highlights that the greatest potential for returning value to communities can be realized if we allow communities' complete

stories to be heard and appreciated rather than viewing their perspectives and concerns as disparate data fragments. Our intention is for the Framework to serve as tool to promote concrete, transparent dialogue between communities and researchers about barriers and facilitators of authentic community-engaged BD research.

5. Conclusion

It was striking how diverse informants across the fields of citizen science, patient-led research and community-based organizations shared with us such comparable stories of frustration and challenge in engaging in research across vastly different health issues and constructs of community. In order for communities to derive benefit from and find autonomy within BD research, their voices, their values, and their priorities must be represented throughout the entire research process, from data collection to research translation. We hope that the Framework will amplify some of the key challenges relevant to community engagement and autonomy within BD research and encourage further investment, anticipation, and mitigation of these factors by program developers, funders, researchers, and community leaders.

Appendices

Appendix A: *Semi-structured Interview Guide*

UNIVERSITY OF WASHINGTON
Developing pathways for community-led research with big data (DeCLR)
Interview Guide

Introduction: Thank you for taking time to speak with [us] today. As [team member] told you, we are speaking with stakeholders to better understand pathways to building community capacity for engagement in and autonomy within big data research. Our aim is to develop a series of examples, or use cases, that will illustrate the ways in which communities might interact with big data research and use big data research to their benefit. We will be recording the interview for transcription and transcript analysis. There are no right or wrong answers. If you feel uncomfortable at any time during the conversation we can stop and move to another question or end our call. Do you have any questions before we get started with our conversation?

Interview guide: What are your community research priorities? Is there a problem that you would like to solve (that could be solved with the assistance of big data)? What are your personal health research priorities?

When we are asking you this, we are asking how you would use data about your community to develop insights that can be used by your community.

Example case study: Smoking cessation in your community. There are a number of social determinants of smoking and desire to quit. We can gather "big data" about those social determinants, for example, measures of income, measures of stress, measures of economic insecurity including housing and food. Additionally, we can learn about the pharmacogenomic variants of small communities [explain more if necessary] that influence the body's nicotine uptake pathways which impacts how difficult it is to quit and how effective smoking cessation interventions, like nicotine substitutes, are for different people. As you can see, there may be both individual level and community level benefit from understanding these different influencers of nicotine dependence and different ways that a community organization like yours could leverage these data to the benefit of those that you serve.

[Through the case study, give a sense of the types of data and how they are collected, give a sense of the types of analyses and what they might return, distinguishing individual level info and the aggregate level info. May need to further elaborate on data types, probe about interest, spend time exploring to bring interviewee forward to this point in brainstorming/community case development.]

So, now, let us return to the initial community research priorities that you identified. Now that we have told you about the potential offered by Big Data, what intersection do you see? Does any of this seem to intersect with your research priorities?

- If yes, dig in
    - Capture example
    - Probe example
        - What if you did this? What skills would you need? What knowledge? What technical tools?
- If no, start with example giving
    - What is the knowledge and skills that are needed by your community

        Probes: what is meant by skills -- give examples

    One set of skills people often start with are data visualization skills. Data visualization is a process of taking data and putting it in to a format where you can look at it and see if there are any patterns, missing things, things that stand out. It is often the first step in helping researchers ask questions about data.

    - What type of access to technical tools would you need?

        Probes: what is meant by technical tools -- give examples

    First, people often need access to data sets. Computers and cloud space (the computing support to house and/or work with the data). Another tool that people often need to start Big Data research is access to the data and software to do data visualization.

    - Get to example (if at all possible) connecting their goals to skills, knowledge, technical tools

- If no, not excited then go to what is past experience with bringing resource (or not) into your community
    - Good example/bad example based on your community's past experience of bringing resource (or not) into your community
    - Walk from here to skills, knowledge, technical tools

If interviewee expresses that “it’s not our job” to do the research/unwilling to imagine building community capacity for research, then:

- What would enable us to reach these community priorities if the skills do not get housed within the community themselves?

Appendix B: *Demographic Survey for Community Leader Key Informants*

UNIVERSITY OF WASHINGTON

Developing pathways for community-led research with big data (DeCLR)

Demographic Survey for Community Leaders

Participant ID:_____

Date:_____

1. What is your age? _____

2. What is your gender identity?
☐ Female ☐ Male ☐ Transgender: _____

3. How do you describe your ethnicity? (check one)
☐ Hispanic or Latino ☐ Not Hispanic or Latino

4. How do you describe your race? (check one or more)
☐ American Indian /Alaska Native ☐ Native Hawaiian or Other Pacific Islander

☐ Asian ☐ White

☐ Black or African American ☐ Other:
_____

5. What is your occupation? _____

6. What is the highest level of education that you have completed? (check one)
☐ Did not complete high school ☐ College graduate

☐ High school graduate/GED ☐ Post-graduate (e.g., M.A., M.S., MD., PhD)

☐ Some college

7. Which statement best describes your knowledge of genetics? (check one)
☐ I know nothing about genetics.

☐ I remember some information about genetics from school.

☐ I am well informed about genetics.

8. Is your organization a community-based organization? (check one)

☐ Yes ☐ No ☐ Don't know ☐ Doesn't apply

9. If applicable, briefly list or describe your community or stakeholders.

_____

_____

10. If applicable, how many employees does your organization employ? (check one)

☐ <10 ☐ 11 to 50 ☐ >50

11. What is your organization's approximate annual operating budget? (check one)

☐ <1 million ☐ 1 to 3 million ☐ >3 million dollars

12. If relevant, what population(s) does your organization typically represent or serve? (check one or more)

☐ American Indian /Alaska Native ☐ Native Hawaiian or Other Pacific Islander

☐ Asian ☐ White

☐ Black or African American ☐ Hispanic or Latino

☐ Other (please describe): _____

13. Have you ever had a genetic test? (check one)

☐ Yes ☐ No ☐ Don't know

14. Have you ever participated in biomedical research?

☐ Yes ☐ No ☐ Don't know

15. Have you ever participated in genetic research?

☐ Yes ☐ No ☐ Don't know

References

1. Collins F, Varmus H. A new initiative on precision medicine. N Engl J Med. 2015;372(9):793–5.
2. The AACR Project GENIE Consortium. AACR Project GENIE: Powering Precision Medicine through an International Consortium. 2017;7(8):818–31.
3. Sankar PL, Parker LS. The Precision Medicine Initiative's All of Us Research Program: An agenda for research on its ethical, legal, and social issues. Genet Med [Internet]. 2017;19(7):743–50. Available from: http://dx.doi.org/10.1038/gim.2016.183
4. President's Council of Advisors on Science and Technology. Priorities for Personalized Medicine. Rep Pres Counc Advis Sci Technol [Internet]. 2008;(September):1–63. Available from: http://www.ostp.gov/galleries/PCAST/pcast_report_v2.pdf
5. Rubin M. Make precision medicine work for cancer care. Nature. 2015;520:290–1.
6. Khoury M. Planning for the Future of Epidemiology in the Era of Big Data and Precision Medicine. Am J Epidemiol. 2015;182:977–9.
7. Robinson PN, Mungall CJ, Haendel M. Capturing phenotypes for precision medicine. Mol Case Stud. 2015;1(1):a000372.
8. Hamburg M, Collins F. A Path to Personalized Medicine. N Engl J Med. 2010;363:301–4.
9. McGrath S, Ghersi D. Building towards precision medicine: Empowering medical professionals for the next revolution. BMC Med Genomics [Internet]. 2016;9(1):1–6. Available from: http://dx.doi.org/10.1186/s12920-016-0183-8
10. Jameson J, Longo D. Precision Medicine- Personalized, Problematic, and Promising. N Engl J Med. 2015;372:2229–34.
11. Sessler D. Big Data - and its contributions to peri-operative medicine. Anaesthesia. 2014;69(2):100–5.
12. Khoury M, Ioannidis J. Big data meets public health: Human well-being cold benefit from large-scale data if large-scale noise is minimized. Science (80-). 2014;346:1054–5.
13. Goytia CN, Kastenbaum I, Shelley D, Horowitz CR, Kaushal R. A Tale of 2 Constituencies: Exploring Patient and Clinician Perspectives in the Age of Big Data. Med Care. 2018;56(10 Suppl 1):S64-9.
14. Sammani A, Jansen M, Linschoten M, Bagheri A, de Jonge N, Kirkels H, et al. UNRAVEL: big data analytics research data platform to improve care of patients with cardiomyopathies using routine electronic health records and standardised biobanking. Netherlands Hear J [Internet]. 2019; Available from: http://link.springer.com/10.1007/s12471-019-1288-4
15. Wu PY, Cheng CW, Kaddi CD, Venugopalan J, Hoffman R, Wang MD. -Omic and Electronic Health Record Big Data Analytics for Precision Medicine. IEEE Trans Biomed Eng. 2017;64(2):263–73.
16. He KY, Ge D, He MM. Big data analytics for genomic medicine. Int J Mol Sci. 2017;18(2):1–18.
17. Williams DR, Bonham VL, Rehm HL, Landry LG, Ali N. Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Aff. 2018;37(5):780–5.
18. Popejoy A, Fullerton SM. Genomics is failing on diversity. Nature. 2016;118(24):6072–8.
19. Ramos E, Callier S, Rotimi C. Why personalized medicine will fail if we stay the course.

2013;9(8):839–47.
20. Bonevski B, Twyman L, Paul C, Hughes C, Randell M, Chapman K, et al. Reaching the hard-to-reach: A systematic review of strategies for improving health and medical research with socially disadvantaged groups. BMC Med Res Methodol. 2014;14(1).
21. Giuliano AR, Mokuau N, Hughes C, Tortolero-Luna G, Risendal B, Ho RCS, et al. Participation of Minorities in Cancer Research. Ann Epidemiol. 2000;10(8):S22–34.
22. Adams-Campbell LL, Ahaghotu C, Gaskins M, Dawkins FW, Smoot D, Polk OD, et al. Enrollment of African Americans onto clinical treatment trials: Study design barriers. J Clin Oncol. 2004;22(4):730–4.
23. Baird KL, Baird KL. The NIH and the FDA: Medical Research Policies and Gender Justice. Gend Justice Heal Care Syst. 2019;1(3):119–59.
24. Locock L, Smith L. Personal benefit, or benefiting others? Deciding whether to take part in clinical trials. Clin Trials. 2011;8(1):85–93.
25. Kauffman KS, Dosreis S, Ross M, Barnet B, Onukwugha E, Mullins CD. Engaging hard-to-reach patients in patient-centered outcomes research. J Comp Eff Res. 2013;2(3):313–24.
26. George S, Duran N, Norris K. A systematic review of barriers and facilitators to minority research participation among African Americans, Latinos, Asian Americans, and Pacific Islanders. Am J Public Health. 2014;104(2):16–31.
27. Erves JC, Mayo-Gamble TL, Malin-Fair A, Boyer A, Joosten Y, Vaughn YC, et al. Needs, Priorities, and Recommendations for Engaging Underrepresented Populations in Clinical Research: A Community Perspective. J Community Health. 2017;42(3):472–80.
28. Mello MM, Wolf LE. The Havasupai Indian Tribe Case — Lessons for Research Involving Stored Biologic Samples. N Engl J Med. 2010;363(3):204–7.
29. Bowekaty MB, Davis DS. Cultural issues in genetic research with American Indian and Alaskan Native people. IRB Ethics Hum Res. 2003;25(4):12–5.
30. Michener L, Cook J, Ahmed SM, Yonas MA, Coyne-Beasley T, Aguilar-Gaxiola S. Aligning the Goals of Community-Engaged Research. Acad Med [Internet]. 2012;87(3):285–91. Available from: https://insights.ovid.com/crossref?an=00001888-201203010-00014
31. Wilkins CH, Spofford M, Williams N, Mckeever C, Allen S, Brown J, et al. Community Representatives' Involvement in Clinical and Translational Science Awardee Activities. Clin Transl Sci. 2013;6(4):292–6.
32. Sage Bionetworks. About Synapse.[online] Available at http://sagebionetworks.org/
33. Zhang X, Perez-Stable EJ, Bourne P, Peprah E. Big Data Science: Opportunities and Challenges to Address Minority Health Disparities in the 21st Century. Ethn Dis. 2017;27(2):95–106.
34. Ajana B. Digital health and the biopolitics of the Quantified Self. Digit Heal. 2017;3:205520761668950.
35. Hsieh H-F, Shannon SE. Three approaches to qualitative content analysis. Qual Health Res [Internet]. 2005;15(9):1277–88. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16204405
36. Ayres L, Kavanaugh K, Knafl K. Within-Case and Across-Case Approaches. Qual Health Res. 2003;13(6):871–83.
37. Mittelstadt BD, Floridi L. The Ethics of Big Data: Current and Foreseeable Issues in Biomedical Contexts. Sci Eng Ethics. 2016;22(2):303–41.

38. Cox DR, Kartsonaki C, Keogh RH. Big data: Some statistical issues. Stat Probab Lett. 2018;136:111–5.
39. Mardis ER. The challenges of big data. Dis Model Mech. 2016;9(5):483–5.
40. Van Horn JD. Opinion: Big data biomedicine offers big higher education opportunities. Proc Natl Acad Sci. 2016;113(23):6322–4.
41. Nunan D, Di Domenico M. Market Research and the Ethics of Big Data. Int J Mark Res. 2013;55(4):505–20.
42. Andrejevic M. The big data divide. Int J Commun. 2014;8(1):1673–89.
43. Crawford K, Miltner K, Gray ML. Critiquing Big Data politics, ethics, epistemology.pdf. Int J Commun [Internet]. 2014;8:1663–72. Available from: http://ijoc.org.
44. Tene O, Polonetsky J. Big Data for All: Privacy and User Control in the Age of Analytics [Internet]. Vol. 11, Northwestern Journal of Technology and Intellectual Property. 2013. 239 p. Available from: http://scholarlycommons.law.northwestern.edu/njtip/vol11/iss5/1
45. McCloskey D. Principles of Community Engagement. 2011;1–193. Available from: http://www.atsdr.cdc.gov/communityengagement/pdf/PCE_Report_508_FINAL.pdf
46. Staley K. Information about INVOLVE [Internet]. 2009. Available from: www.twocanassociates.co.uk
47. Joosten YA, Israel TL, Williams NA, Boone LR, Schlundt DG, Mouton CP, et al. Community Engagement Studios. Acad Med. 2015;90(12):1646–50.
48. Longo D, Drazen JM. Data Sharing. N Engl J Med. 2017;247–66.