Fred Hutch Data Needs Assessment

Version 1.0 || Jun 7, 2023

Executive Summary

Fred Hutchinson Cancer Center was formed from the merger of Fred Hutchinson Cancer Research Center and the Seattle Cancer Care Alliance in April 2022. Fred Hutch is an independent, nonprofit organization that also serves as the cancer program for UW Medicine. This relationship allows for enhanced care coordination with one of the world's leading integrated health systems.

In May 2022, Fred Hutch hired Jeff Leek as Chief Data Officer and J Orin Edson Foundation Professor of Biostatistics. Previously, he was a professor of Biostatistics and Oncology at the Johns Hopkins Bloomberg School of Public Health and co-director of the Johns Hopkins Data Science Lab. Dr. Leek's responsibility as CDO in the first year was to identify the data needs of the newly merged Fred Hutch and to design a comprehensive data strategy to address those needs. To conduct this study he recruited two individuals - Dr. Sean Kross whose expertise is in the area of human-computer interaction focusing on data science and Dr. Amy Paguirigan who has been a member of the Fred Hutch research community with experience across wet lab, clinical assay development, bioinformatics and computing.

The Fred Hutch data needs assessment ran from April 2022 to April 2023. The needs assessment used a mixed methods approach encompassing informal conversations, formal interviews, and ongoing collaborations addressing acute data needs. Both quantitative and qualitative data were collected from individuals across all divisions and all levels of personnel at Fred Hutch. Recruitment for the needs assessment used multiple channels including announcements on public websites, on Centernet, on social media, outreach to leadership of each division and program, user groups, a Consortium oriented Slack workspace, and through direct collaboration requests. The result of this outreach was a collection of 213 conversations representing 319 individuals and over 1,200 hours of conversation - spanning Fred Hutch research and clinical personnel, as well as a collection of administrative personnel who support research and clinical activities.

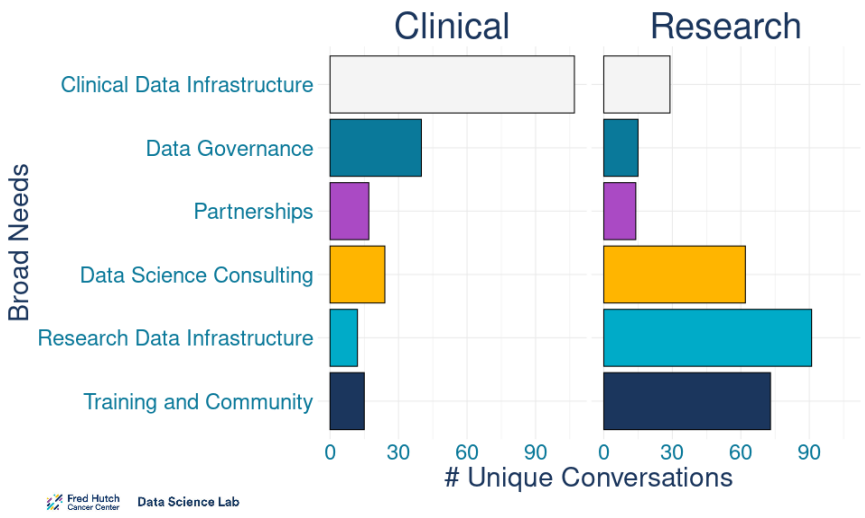

Figure 1 Summary of Data Conversations during Fred Hutch Data Needs Assessment. *This figure represents the number of unique conversations where specific broad areas of need were raised by members of Fred Hutch's community. Needs were separated into two groups - individuals who use Fred Hutch patient data for clinical, research, or operations purposes and the broader Fred Hutch research community who leverage a broader range of data resources. Needs were assigned to categories based on a controlled vocabulary with six broad categories (shown here and defined in Appendix A) and 26 specific categories (shown in later figures and defined in Appendix A).*

The initial high level take-away of both the quantitative and qualitative data from our assessment show there are two key patterns of need in the Fred Hutch's data community that broadly reflect our hybrid institution: a Fred Hutch patient clinical data community focused on the research, clinical, and operational use of Fred Hutch patient data and a broader Fred Hutch research data community with diverse researcher needs (**Figure 1**).

Fred Hutch Patient Clinical Data Community

The Fred Hutch patient clinical data community need distribution was driven by challenges centered around the clinical data collected about Fred Hutch patients. The Fred Hutch patient clinical data community is most interested in data related to clinic operations, clinical care, philanthropy, and clinical research such as clinical trials. The top needs are:

1. *Ease of access to comprehensive clinical data in a timely manner*
2. *Appropriate data governance systems for access to these data*
3. *Clinical data analysis support*
4. *Appropriate clinical data storage, management, and analysis infrastructure*
5. *Partnerships to support clinical care, trials and research via clinical data sharing/data use agreements*

The FHCC clinical data are highly regulated and thus administrative needs include expanded information security and compliance support and training, as well as expanded clinical data analytic support. This component of Fred Hutch's data community can be best addressed by creating and managing a centralized data infrastructure, defining clear guidelines for data governance for multiple uses, and leveraging data partnerships to augment data sources.

Broader Fred Hutch Research Data Community

The Fred Hutch research data community need distribution reflected the diversity of Fred Hutch research groups. The clinical research program as we have defined it for this study is focused on clinical research using data about Fred Hutch patients and will be served by Fred Hutch's patient clinical data program. However, there is a broad range of individual research programs and a correspondingly high diversity of data needs that span everything from individual wet labs, to large epidemiological cohort studies, to genetic studies, to massive national and international scale data coordinating centers. The top cross-lab data needs at Fred Hutch include:

1. *Access to training in data science, machine learning and reproducible research*
2. *Research data storage and computing technology and access*
3. *Research data management support to reduce costs, comply with requirements, and archive data*
4. *Data science consulting to adopt modern cloud, software, machine learning and data technologies*
5. *Data science communities for sharing and supporting best practices*
6. *Product development support for R packages, Python libraries, workflow management and Shiny apps*
7. *Partnerships coordination for philanthropy and technology*
8. *Support for data sharing agreements including execution and governance.*

This component of Fred Hutch's data community can be best addressed by creating a collection of smaller, interoperable data products and services that researchers can mix and match to best address their data needs. Due to the long tail of unique data needs for this community beyond these shared groupings, addressing the evolving, fine-grained needs of the community will require active engagement and community connections to continue to adapt to fluidly changing needs over time.

While we have reported here on the initial needs identified during our conversations over the past year with Fred Hutch's community there will be ongoing and new data needs over time. We plan to continue to engage the community through conversation and collaboration to ensure that Fred Hutch's data program is responsive to the diversity of needs of Fred Hutch's community in an ongoing way.

Summary and Next Steps

Fred Hutch's data community is both strong and diverse. Integration and support of all scientists, providers, and administrative personnel to leverage data will have significant positive impacts on patient care, clinical trials, research productivity, and institutional function. Fred Hutch's Strategic Data Plan 2023 outlines how this support can be delivered through an integrative approach leveraging existing data expertise across the enterprise.

Fred Hutch Data Needs Assessment

Version 1.0 || Jun 7, 2023

Background

Fred Hutchinson Cancer Center was formed from the merger of Fred Hutchinson Cancer Research Center and the Seattle Cancer Care Alliance in April 2022. Fred Hutch is an independent, nonprofit organization that also serves as the cancer program for UW Medicine. This relationship allows for enhanced care coordination with one of the world's leading integrated health systems. The mission statement of Fred Hutch is:

> *Fred Hutchinson Cancer Center unites innovative research and compassionate care to prevent and eliminate cancer and infectious disease. We're driven by the urgency of our patients, the hope of our community and our passion for discovery to pursue scientific breakthroughs and healthier lives for every person in every community.*

The research and clinical missions of Fred Hutch are supported by local, national, and international data programs. These programs cover not only the clinical data about Fred Hutch patients, but the HIV Vaccine Network, the Covid Clinical Trials Network, the Southwest Oncology Group, and the Women's Health Initiative among others. Moreover, the more than 250 independent faculty groups at the Hutch manage data ranging from hundreds of terabytes of cellular imaging data, to epidemiological-scale cohorts of genetic data, to bespoke assays measuring biochemical and biomolecular processes.

The data program at Fred Hutch is as diverse as the clinical and research missions we pursue to prevent and eliminate cancer and infectious disease. The data needs also go well beyond those of just our scientific and clinical programs. Data empowers and supports the entire Fred Hutch enterprise, from people management, to financial planning, to philanthropic development. Supporting this diverse collection of providers, clinicians, researchers, trainees, and professionals requires careful identification of key data needs and development of programs to address those needs.

In May 2022, Fred Hutch hired Jeff Leek as Chief Data Officer and J Orin Edson Foundation Professor of Biostatistics. Previously, he was a professor of Biostatistics and Oncology at the Johns Hopkins Bloomberg School of Public Health and co-director of the Johns Hopkins Data Science Lab. His group develops statistical methods, software, data resources, and data analyses that help people make sense of massive-scale genomic and biomedical data. As the co-director of the Johns Hopkins Data Science Lab he helped to develop massive online open programs that have enrolled more than 8 million individuals and partnered with community-based non-profits to use data science education for economic and public health development. He is a Fellow of the American Statistical Association and a recipient of the Mortimer Spiegelman Award and Committee of Presidents of Statistical Societies Presidential Award.

As a component of his recruitment, Dr. Leek asked for a year to discover the data needs of the newly merged Fred Hutch and to design a data program to address needs across the institution. The goal of this discovery process was to determine key strengths, opportunities, weaknesses and threats to the data program at Fred Hutch and to design a cross-institutional data program to address the ongoing needs of the institution. In the modern era of biomedical research, data and computing resources play a critical role in driving progress towards the development of new treatments and cures for disease. The need for data and computing resources is only increasing as the complexity of scientific problems and the amount of data generated by experiments continues to grow. Therefore, it is an essential goal of the Data Science Lab to ensure that the resources we have available are both sufficient and directed to meet the current needs of our researchers, clinicians and staff, and to prepare us for exciting futures of data-enabled care, treatment, and interventions,

This document represents the results of this year-long discovery process and a summary of recommendations for the data program at Fred Hutch. In addition to these general recommendations, this assessment suggested

the need for a cross-functional data laboratory, responsible for developing and maintaining an overall data strategy, filling in the gaps in the current Fred Hutch data program and facilitating cross-institutional collaboration and coordination.

This data needs assessment represents our first pass at understanding how data does and does not work at Fred Hutch. However, one of our key findings is that there is an ongoing need to identify evolving data challenges, prioritize and highlight those challenges, and coordinate cross-functional groups to tackle them in order to maintain an effective data strategy for the institution as a whole. We therefore are releasing this needs assessment as the first version of what will be an ongoing Data Science Lab project to continue to surface data needs at Fred Hutch in partnership with the community, integrate those needs into our overall data strategy, and then communicate to the community about those needs and their impact on our data strategy over time.

Methods

Primary Goals of the Needs Assessment

The driving purpose of the data program at the Fred Hutch is to be an enabling effort to (1) make data easier for people across the Fred Hutch to incorporate into their scientific or medical practice, (2) to improve data literacy and capability, and (3) to identify future technologies and data sources that will allow new innovation in the modern data era of biomedicine. Understanding the data needs of Fred Hutch personnel is critical to enabling the next generation of data innovation. To that end, the primary goals of the needs assessment were to find out from the Fred Hutch community:

1. What is working well with data?
2. What is not working well?
3. What would work well with more resources or support?

Given the broad range of roles and “data domains” at Fred Hutch, it is not surprising that there are a broad range of data needs ranging from basic data management, to complicated scalable machine learning. The needs assessment was therefore also focused on identifying and distinguishing needs that impacted a large number of personnel from those that were more bespoke. The intention was to identify those needs that could be addressed to maximum positive effect for people all across Fred Hutch.

Personnel

To comprehensively identify data needs across the institution Dr. Leek recruited two people to lead the Fred Hutch data needs assessment. Dr. Sean Kross was recruited from UC San Diego where he received his PhD in cognitive science with a focus on understanding how scientific researchers leverage data analysis and computer programming in pursuit of their research goals. Dr. Kross’ background is in the design and execution of qualitative and quantitative studies of data science behavior. To complement Dr. Kross’ general human-computer data interaction experience, Dr. Leek also recruited Dr. Amy Paguirigan who has a 15 year history of experience in the Fred Hutch data community spanning the wet lab, clinical assay development, bioinformatics, computing and data science. Dr. Paguirigan has a track record of documentation and synthesis of data processes and procedures, as evidenced by her leadership of the Fred Hutch SciWiki (https://sciwiki.fredhutch.org/) and collaborations with many Fred Hutch groups.

Timeline

The data needs assessment began with informal conversations between Dr. Leek and individuals or groups of individuals at Fred Hutch in May of 2022. Dr. Kross and Dr. Paguirigan were recruited in July 2022. The first stage of the needs assessment involved informal conversations between Fred Hutch employees and members

of the data needs assessment team that ran from April 2022 through April 2023. Later, more formal interviews were introduced into the process between September 2022 and April 2023.

The merger of the two institutions happened in April 2022 and the data needs assessment has been expanded to include initial interviews with clinical personnel about their data needs. The first draft of the data needs assessment was produced in April 2023, however, there are ongoing conversations across Fred Hutch that we intend to continue to enable us to release future versions of needs assessment documentation. It is important to note that this first version of a needs assessment study may not be exhaustive, and we acknowledge that there are additional data and computing needs that have not been captured in this study. If a member of the Fred Hutch community feels that their needs are not sufficiently represented in the context of this assessment, they can directly reach out to Fred Hutch Data Science Lab staff at: hutchdatascience.org/needs to engage in ongoing conversations about the needs of the Fred Hutch community. While there will be ongoing and newly developed data needs, this study represents the initial information that informed the design of Fred Hutch's data strategy. The results of this study are a valuable starting point for further discussion and action towards meeting the data and computing needs of the Fred Hutch community.

Data Collection Process

The design of this study employed both quantitative and qualitative research methods. This mixed methods approach, common in the analysis of complex sociotechnical systems, integrates diverse data collection and analysis techniques, and offers researchers a more comprehensive understanding of complex phenomena. Much of the data collected in this study was collected via the semi-structured interview format, which allows for the exploration of participants' experiences, beliefs, and opinions in a flexible yet guided manner. Semi-structured interviews are characterized by their use of open-ended questions, which encourage participants to elaborate on their thoughts and feelings. This approach allows researchers to gather rich, in-depth information that can reveal important patterns and themes. The majority of the data in this study came from one of two sources: notes and quotations recorded by the researchers during a series of formal, semi-structured interviews, and a corpus of notes taken during meetings and conversations that the research team had that were contemporaneous with the formal interview study.

The formal, semi-structured interviews were conducted both in-person and online to gain in-depth knowledge of individual perspectives with respect to participants' data and computing experiences at Fred Hutch. During these interviews, we took detailed notes and recorded some conversations, later transcribing the recordings to obtain accurate and complete textual data. The full list of questions used by researchers during these interviews is included in Appendix B of this document.

Data collection for this study also included gathering notes from meetings and conversations, which allowed us to capture real-time insights and observations from participants. By combining these data collection methods, we were able to gather rich information that captured diverse viewpoints and experiences, providing a strong foundation for our qualitative analysis.

Other mechanisms of data collection included a Data Science Lab Roadshow on August 30th, 2022, where we visited each of the divisions of Fred Hutch individually and provided open discussion hours for those groups to meet with our team; presentations to divisions and multiple administrative groups about the NIH Data Sharing and Management Plan; participation and presentations at Division and IRC retreats; ongoing Data House Calls consulting opportunities; and Single Cell, R User group and Effective Computing user groups/drop in hours between April 2022 and April 2023.

By using multiple data sources we aimed to achieve data source triangulation, a technique used in qualitative research to enhance the validity of study findings by drawing upon multiple sources of data. By comparing and cross-referencing information from these various sources, including the interviews, conversations, notes from office hours, roadshows, and other events, we seek to identify common patterns and minimize potential biases associated with a single data source.

The needs assessment was publicly announced at www.hutchdatascience.org on October 28th, 2022 and solicitation of participation was sent to members of the HELM leadership team for Fred Hutch on October 12th and 13th, with individual follow-ups in the following weeks. The needs assessment was also announced in the Fred Hutch Data Slack channel on August 29, 2022. Follow up messages were sent to individual community members to participate via user groups, office hours, presentations at group events and during direct conversations with individuals.

Data Analysis Process

Our data analysis was based on establishing a Grounded Theory, a systematic qualitative research methodology aimed at generating theory from empirical data. This method is an inductive approach, where theories emerge from the data rather than being imposed upon it. Grounded Theory is widely adopted in disciplines that study sociotechnical systems, as it enables researchers to develop contextually relevant and empirically grounded explanations for complex social phenomena.

We used open coding, a fundamental component of grounded theory, in the initial phase of our data analysis. In this stage we organized the raw data into discrete concepts and categories. A close reading of the collected data (described in detail in the section above), which included interview transcripts, meeting notes, and other documents, allowed us to identify key ideas, recurring themes, and patterns. We then labeled each of these observations with codes, which represent the underlying meaning or essence of these data segments. As this process of coding continued, we iteratively compared and contrasted the emerging codes to identify similarities, differences, and relationships between codes. This constant comparative method facilitates the development of more abstract and higher-order categories, which ultimately formed the basis of our theoretical framework. Our goal was to reach theoretical saturation, a point where no new insights or categories were emerging from the data, which would indicate a comprehensive and well-grounded theory.

By the end of this analysis process, twenty-six intermediate themes surfaced, which were ultimately consolidated into six overarching themes: Clinical Data Infrastructure, Research Data Infrastructure, Training, Partnerships, Data Governance, and Data Science Consulting. The full list of categories of needs that we identified can be found in the appendix.

Results

Needs assessment population

In the first phase of the needs assessment, we performed outreach to a broad cross-section of the personnel at Fred Hutch. During this phase we were able to speak with 319 individuals representing all major research divisions, clinical providers, clinical and research administration, and clinical operations. Our interviews encompassed all levels of Fred Hutch staff from trainees, to administration, to faculty, to providers and resulted in 213 conversations representing 319 individuals between April 2022 and April 2023.

We developed a two-level controlled vocabulary for data needs at Fred Hutch. The first level of the vocabulary translated data needs into one of 26 fine grained data needs categories which we then clustered into one of 6 broad data needs categories (see **Appendix A** for detailed descriptions of both broad and specific needs). We then calculated both the direct number of data conversations (informal interviews, formal interviews, group interviews, collaborative conversations, or workgroup conversations) where each of the broad needs was expressed.

One of the primary discoveries of the initial needs assessment is that these broad needs seemed to represent two distinct data communities within Fred Hutch. One data community is focused on the data collected about and for Fred Hutch patients for a variety of uses. This Fred Hutch patient clinical data program is based around the unifying feature of Fred Hutch patient data - even though the diversity of needs for this data are significant

and vary across clinical, research, financial, operational, and philanthropic applications. The other data community has a more diverse collection of needs since it represents the broader Fred Hutch research community. Unlike the community focused on Fred Hutch patient data, the broader research data community does not have a single unifying data source or need and therefore requires a more diffuse set of supports and infrastructure.

For the remainder of this assessment we will separate these two communities for clarity in the needs description; recognizing that there are many individuals who participate in both communities. Appendix A contains a number of specific examples of needs that correspond to the controlled vocabulary we use in this document.

Fred Hutch Clinical Data Community Needs

The clinical community at Fred Hutch has unique data needs related to the collection, storage, management, and analysis of clinical data. Our impression is that the community is focused on the clinical data collected about Fred Hutch patients, specifically data related to clinical operations, clinical care, philanthropy, and clinical research including clinical trials. To address the data needs of this community, we have identified several key areas of need, including ease of access to comprehensive clinical data, appropriate data governance systems for access to these data, clinical data analysis support, appropriate clinical data storage, management, and analysis infrastructure, and partnerships for clinical data via clinical data sharing and associated data use agreements. In addition, our clinical data are highly regulated, so secondary needs include expanded information security, compliance support, and training, as well as expanded clinical data analytic support. In the sections below, we discuss each of these areas of need in detail.

Fred Hutch Patient Clinical Data as a Unifying Data Resource

Fred Hutchinson Cancer Center was formed from the merger of Fred Hutchinson Cancer Research Center and the Seattle Cancer Care Alliance in April 2022. Fred Hutch is an independent, nonprofit organization that also serves as the cancer program for UW Medicine. This relationship allows for enhanced care coordination with one of the world's leading integrated health systems.

One of the most significant Fred Hutch data communities is all of the individuals who rely on Fred Hutch patient data for their roles. Given both the research and clinical missions of the institution, the range of roles represented within this community includes clinical providers, clinical operations professionals, clinical administration, researchers, research administration, philanthropic and development professionals, financial professionals and Fred Hutch leadership. Each of these

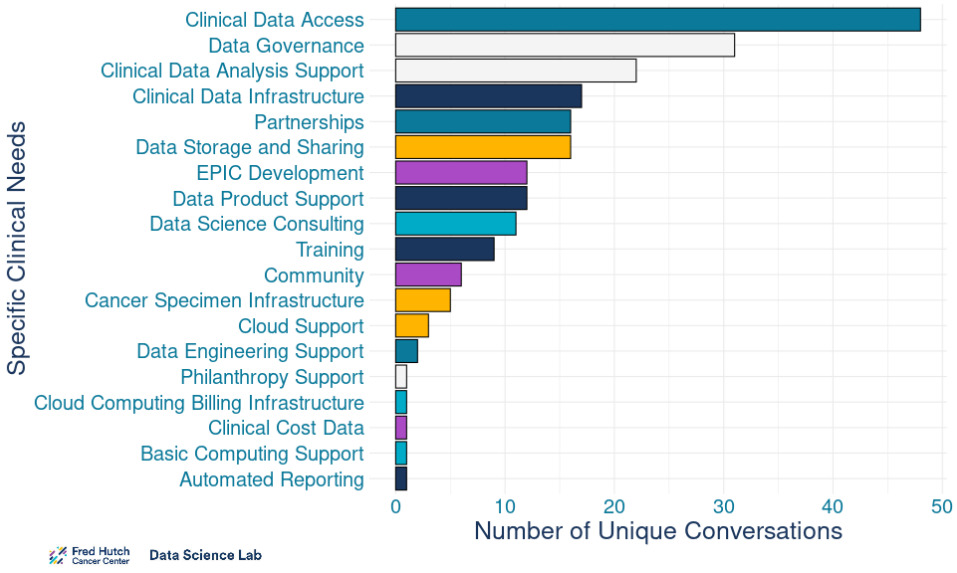

Figure 2 Summary of specific needs for the Fred Hutch patient data community. *This figure represents the number of unique conversations where specific areas of need were raised by members of Fred Hutch's clinical, research, and operations communities focused on Fred Hutch patients. Needs were assigned to categories based on a controlled vocabulary with six broad categories (shown above in Figure 1 and defined in Appendix A) and 26 specific categories (shown here for the clinical community and defined in Appendix A). The consistent theme of the Fred Hutch patient focused data community was need for access, governance, and infrastructure for Fred Hutch patient data.*

members relies on access to up to date patient information to fulfill their roles.

A consistent theme was the need for access to Fred Hutch patient clinical data for uses like clinical trials as well as translational research (**Figure 2**). The most impactful limited access was to comprehensive and timely electronic health record (EHR) data for translational research uses as this lack has resulted in significant burdens requiring data abstraction and prevented research completely in cases such as new therapeutic evaluations (due to years long time lag in data availability for research). Both EHR data, as well as multi-model laboratory data was part of this need to allow for enrollment of patients into clinical trials as well as other types of translational research. This included data access that required interactions with partner systems (Seattle Children's, UW) and external clinical vendor systems (Tempus, Caris, among many others) as well as data governance and analytics support. Participants also identified challenges in understanding if and how clinical data is (or is not) available at all due to challenging to navigate systems as well as poorly documented access instructions.

Data Access Agreements

As a part of the merger, Fred Hutch developed data exchange and sharing policies with Seattle Children's Hospital and the University of Washington. To ensure access to clinical data between institutions multiple data sharing agreements were developed with differing sets of requirements based on the relationships between the institutions.

Fred Hutch entered into a clinical/operations data sharing agreement and a philanthropic data sharing agreement with the University of Washington as a component of the merger to allow Fred Hutch to manage the adult oncology program. More recently, a data exchange agreement was developed between the University of Washington and Fred Hutch covering clinical operations, financial, and research use of patient data in April 2023. This expanded agreement was referenced as a critical need by many staff as the lack of legal oversight of data use for research caused significant challenges in executing data governance.

Multi-modal data sources integration

In addition to the core EHR data, a variety of other data sources are collected about Fred Hutch patients. In particular, panel genetic testing is performed by the University of Washington (Oncoplex) and a range of outside providers (Tempus, Caris, Color Genomics, and others). Additional pathology data is collected including imaging, slides which may be digitized, flow cytometry, and other laboratory assays. Radiology and other imaging data are further collected and stored in digital form. More bespoke assays are often performed within the context of individual research studies representing a broad diversity of genetic and clinical data.

Streamlined access to multi-modal data types were identified as a core need for genetic counseling, clinical care, clinical trials, and clinical and translational research.

Clinical data infrastructure

Clinical data infrastructure represents a critical component of Fred Hutch's needs. Managing multi-modal clinical data presents a significant challenge due to regulatory requirements, governance requirements, data size, and the range of tools needed to interact with these data. Specifically, the community needs streamlined access to clinical data across clinical, operations, and research applications, as well as appropriate tooling for data visualization, analysis, cleaning, and processing. The infrastructure should also be designed to ensure compliance with regulatory and ethical standards for data privacy and security. Addressing these needs would enable clinicians to make better data-driven decisions.

Universal Electronic Consent

Universal electronic consent would allow for streamlined access to patient clinical data for a variety of research projects. Several tumor programs have already developed their own broad-based consent forms, but a universal electronic consent system would enable researchers to access patient data across all tumor

programs at Fred Hutch. This type of consent form was identified as a core need by multiple clinical researchers, as it would speed up research projects and facilitate collaboration across different programs. As a feature for a universal electronic consent system, the community also identified the need that the system should enable researchers to filter patients by consent status, in addition to other criteria. Electronic recoding of consent would allow researchers to quickly identify eligible patients for their studies, helping to streamline the research process.

Data Governance

There are a large number of clinical patient data managed by the University of Washington, Seattle Children's Hospital, and Fred Hutch that require careful data governance to comply with data use agreements, regulatory requirements, and legal requirements. Insufficient data governance support was identified as a frequent hurdle for clinical trials and clinical research. Data governance is a particularly high priority given the new regulatory requirements created by the establishment of Fred Hutch as a HIPAA hybrid entity at the completion of the merger.

Clinical Trials Support

The clinical trials team at Fred Hutch has reported the need for improved patient identification for enrollment in trials, data abstraction to allow for more complete patient selection, and data access and sharing for reporting requirements. Additionally, the clinical trials team has identified the need for automated reporting to streamline the reporting process and reduce the administrative burden on researchers. By addressing this need, Fred Hutch can improve the efficiency and effectiveness of its clinical trials program, enabling researchers to conduct high-quality trials.

Clinical and Clinical Research Analytics

Clinical analytics is provided by a coordinated Fred Hutch clinical analytics team that was a part of the former Seattle Cancer Care Alliance. Clinical Research Analytics is supported by a range of groups, from the Gateway program, to clinical Biostatistics, to individual lab research analytics support. Despite these resources, there is still significant demand for clinical and research clinical analytics support above and beyond the existing resources at Fred Hutch.

Linkage of Specimen Data to EHR and MultiModal Data

Biological specimens are sampled from Fred Hutch patients both in Fred Hutch clinical facilities as well as in UW laboratories. These biological specimens represent a broad range of healthy and tumor tissue and are stored in both institutional and research group specific repositories. The Office of Translational Research (OTR) has been developing a system for specimen acquisition and tagging based on the Open Specimen platform. However, there is a need to catalog specimens and link them systematically to patient EHR data identified by multiple tumor groups within Fred Hutch.

Training

Considering the diversity of clinical needs, there is a proportional need to provide comprehensive training to ensure that clinicians have the knowledge and awareness they need to both work effectively with clinical data, and to understand how to work within the clinical data ecosystem at Fred Hutch. Training needs span a range of topics, including how to operate within data governance frameworks, how to integrate multi-modal data sources, how to interact with clinical data infrastructure, how to leverage consent processes, awareness of the availability of clinical trials support, and how to engage with clinical research analytics. This need for training reflects the need for the clinical data program to be responsive to the needs of the community.

Fred Hutch Research Data Community Needs

A diverse collection of research data programs

Fred Hutch's clinical data program can be organized around the data collected for and about Fred Hutch patients. However, the Fred Hutch research program encompasses a diverse and broad ranging set of data. The types of data being collected, managed, analyzed, reported, and shared by Fred Hutch researchers spans everything from unprotected and ulta large cellular imaging data, to epidemiology scale genetic data, to multinational clinical trials, to longitudinal data repositories. Moreover, individual research labs collect a wide range of data from very low resolution to high resolution and analyze them using a broad range of computing, statistical, and machine learning tools.

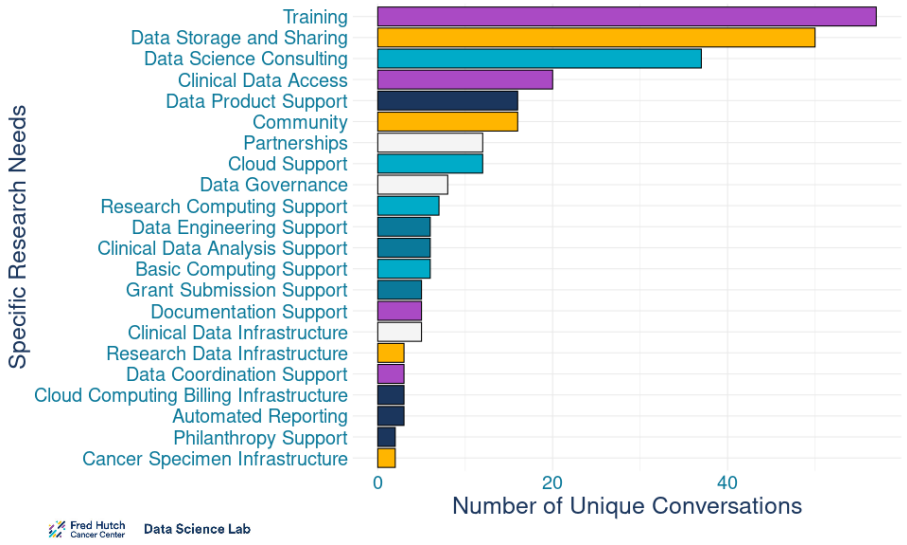

Figure 4 Summary of specific needs for the broader Fred Hutch research community. *This figure represents the number of unique conversations where specific areas of need were raised by members of Fred Hutch's broader research community. Needs were assigned to categories based on a controlled vocabulary with six broad categories (shown above in Figure 1 and defined in Appendix A) and 26 specific categories (shown here for the research community and defined in Appendix A). The consistent theme of the broader Fred Hutch research community was need for training, data consulting, and data storage and sharing (data management) support.*

It is difficult to think of this wide ranging collection of data operations as a single "data program". Using a common analogy for Fred Hutch Research, it is easier to think of the research data program as a shopping mall, or a collection of small, medium, and large enterprises - each with their own needs. We have aggregated this wide ranging set of data needs into some common "clusters of need" that span across multiple groups (**Figure 4**). However, there is a long tail of data needs that represent significant challenges for individual groups which are captured within the heading of "data science consulting" - representing everything from cloud readiness, to data management support, to advanced machine learning collaboration.

Training

Training represents a critical need among the research community at Fred Hutch, as data and computational competence is directly related to researchers' ability to realize their research goals. The need for training is particularly acute given the rapid pace of technological change, as well as the high demand for talented data and computing professionals in industry. Given this demand, the community faces a constant challenge in maintaining a high level of research data and computational competence. This requires ongoing training in a range of topics, including computer programming, data science, computational biostatistics, data visualization, computational genomics, and other related skills.

Community

Considering the diversity of research happening in individual labs at Fred Hutch, it can appear that labs have little in common. However, a unifying thread across many of these labs lies in their data needs, data types, and the data analysis and computer programming tools they employ. Researchers at Fred Hutch have expressed a strong desire for a data community that fosters collaboration and communication among researchers. Researchers want communities to provide a platform for discussing the use of specific tools, techniques for interacting with various data types, and leveraging unique datasets. These conversations could encompass topics such as different file formats, handling large-scale data, and awareness of access for specific data

repositories. Researchers have emphasized the importance of creating dedicated spaces for them to engage in dialogue about interacting with data specifically within the context of Fred Hutch, promoting knowledge exchange and collaboration across the institution.

Data Lifecycle Strategy

Researchers at Fred Hutch often find themselves in the midst of a complex data flow process. Researchers collect data from an array of sources including lab experiments, sophisticated equipment such as genome sequencing machines and advanced microscopes, medical images, electronic health record systems, as well as collaborators and scientific data repositories. This data influx forms the basis for a variety of downstream activities.

Once researchers gain access to data from these streams, the next phase typically involves processing and analysis. This process often results in the generation of more data, the integration of data from various sources, or the production of artifacts like data tables, figures, and academic publications. Furthermore, researchers use this data to create statistical models, machine learning models, scientific data packages for interacting with the collected data, computational pipelines for result reproducibility, and web applications for science or data interaction.

Consequently, researchers express the need for a well-defined strategy to manage the lifecycle of their data. They need to understand the origin of their data, decide on storage and access methodologies for active use in tasks like academic paper writing, machine learning model building, etc., and consider archival storage. Additionally, they need to plan for data sharing to comply with regulatory requirements, such as the NIH data sharing requirements. This necessitates a comprehensive data lifecycle strategy that encapsulates data acquisition, use, storage, and sharing.

Data Storage and Sharing

Researchers at Fred Hutch have highlighted the need for clarity regarding their data storage options, both on-premises and those offered by cloud computing vendors. They wish to understand the trade-offs between these options to make informed decisions that best suit their specific requirements. Additionally, data sharing is a significant aspect of their work, involving sharing within individual labs, with collaborators both inside and outside of Fred Hutch, and at times, with the broader scientific community.

Data sharing needs to be secure, and researchers must have the capability to manage data permissions in accordance with legal, ethical, and Institutional Review Board (IRB) guidelines. Moreover, when sharing data more publicly, it should be done in a manner that enhances discoverability by potential collaborators.

This ties into the data lifecycle strategy outlined previously, as both data storage and sharing form crucial components of this strategy. Researchers also expressed a need to comprehend the cost models associated with data storage and sharing. Understanding these costs will enable them to plan and budget their projects more effectively.

Cloud Computing Support

Within the Fred Hutch research community, the need to leverage cloud computing capabilities is steadily growing. Researchers find that cloud computing offers a variety of products and configurations not available on-premises, thereby broadening their analytical capabilities. At times, these researchers need to utilize cloud computing and data resources set up for them by other institutions.

Moreover, cloud computing vendors often express a willingness to donate computing resources, storage resources, and engineering resources to scientists. However, researchers require assistance in articulating their specific cloud computing needs to these vendors, to ensure that they receive the most suitable resources for their work.

Understanding pricing models for cloud computing services is also a pressing need. Researchers require clear cost models to effectively forecast their cloud computing costs, ensuring their projects remain financially viable.

These needs resonate strongly with the previously discussed data storage and sharing, as well as data lifecycle strategy needs. Cloud computing not only offers researchers diverse options for storing and managing their data but also supports the sharing and collaboration aspect of their work. Furthermore, cloud computing is intertwined with data lifecycle strategies, offering solutions for data storage, access, and archival purposes.

Data Science Consulting

Researchers expressed a significant need for data science consulting to better plan their data analysis, software usage, and data lifecycle management for their studies. This need for consultation serves as a venue for discussions on how to apply data science methods to their research effectively, thus enhancing the quality and impact of their work.

Researchers require assistance in various areas, including scientific software development, code review for data analysis and processing, consultations about reproducible research methods, and building customized systems for automated reporting. Furthermore, researchers often need expert advice on software tools specific to genomics and proteomics analysis.

Additionally, this need for consultation extends to publishing in data science-focused venues and journals, creating effective lab data management strategies, and formulating compliant NIH data sharing plans. Essentially, researchers are seeking experts in data science to collaborate with, someone who can help them understand and apply the latest data science methods to advance their research and increase its innovation potential.

Data Product Support

Folks in the research community also identified a need for data product support. Researchers often need assistance with the initial evaluation of scientific software and platforms, in order to determine whether the tool or platform is worth the cost or commitment required. This need arises from the aggressive marketing of software and platforms by various companies, and researchers require expert help during calls and meetings to evaluate whether these products meet their specific needs and are worth the investment.

Once a tool or platform has been deemed valuable, researchers have expressed a need for support in deploying these applications within the Fred Hutch infrastructure, including support for ensuring that these tools meet existing compliance requirements. Additionally, researchers often require assistance in building custom web applications that can cater to the unique needs of their scientific community. This task is becoming increasingly important, as these applications are often a significant deliverable for grants. Example application types include applications that facilitate data exploration and discovery. As the number of online projects that researchers publish grows, they additionally need support in setting up and measuring analytics related to the usage of these online resources.

Moreover, researchers have indicated a need for assistance in leveraging off-the-shelf artificial intelligence tools in their research. These tools include technologies such as image recognition and automated labeling, image segmentation, speech-to-text translation, and technologies that use transformer models like ChatGPT. This array of needs illustrates the importance of data product support in enabling researchers to conduct their work.

Documentation

The Fred Hutch research community has clearly articulated the need for robust and comprehensive documentation. This need was emphasized by the widespread usage and reliance on the current Biomedical Data Science Wiki, which researchers frequently have open as a constant reference when interacting with Fred

Hutch's data and computing infrastructure. This wiki has proven to be an invaluable resource, providing guidance on a multitude of topics related to data science within the context of Fred Hutch's systems and resources.

However, despite the utility of the existing wiki, researchers have indicated that additional documentation is required. They need detailed guidance about the various data and computing systems at Fred Hutch, suggesting that the current documentation may not fully cover all aspects of these systems or may not delve into the level of detail required by researchers. This need for documentation extends across multiple areas previously discussed, including data storage and sharing and data lifecycle strategy. It also relates to the effective use of specific software and platforms.

Understanding and navigating the diverse and complex data types, large-scale data repositories, and specific computational tools, among other things, is a critical part of research work at Fred Hutch. Comprehensive, accessible, and up-to-date documentation can help researchers maximize the utility of these resources and tools, as well as foster a collaborative data community. Thus, a consistent emphasis on improving and expanding the existing documentation appears to be a significant need within the Fred Hutch research data community.

Grant Support

The research community has expressed the need for comprehensive data science support when it comes to grant applications. Researchers are finding themselves at the intersection of diverse data streams, advanced technologies, and stringent regulations, all of which need to be effectively communicated in their grant proposals.

With the ongoing evolution of NIH requirements, grant stipulations, and data repository prerequisites, researchers need to articulate data sharing plans that adhere to these guidelines. The challenge of conveying these details in their grant applications is an ongoing need within the community, and an area where targeted data science support can make a significant difference.

Moreover, the rise of artificial intelligence in research has introduced another layer of complexity to grant applications. It is increasingly the case that grant proposals are more competitive if they include a strong AI component. Therefore, researchers need to convincingly demonstrate that Fred Hutch not only possesses the technical capacity for AI but also has the necessary staff support to implement AI effectively in their projects.

Finally, another need that has been highlighted is the requirement for strong letters of support, evidencing Fred Hutch's commitment to supporting the researchers' data science-related ambitions. These letters become an integral part of the grant application process, bolstering the credibility of the proposed research and assuring funding bodies of the institutional backing for the project's data science endeavors.

Partnerships Support

The Fred Hutch research community is privileged to be connected to a diverse network of partnerships spanning cloud vendors, big tech companies, pharmaceutical entities, scientific research collaborators, universities, and philanthropic organizations. These relationships play a vital role in the dynamic ecosystem of research, development, and innovation at Fred Hutch.

However, when it comes to data and data science concerns, researchers have identified a gap in the organization and management of these relationships. There is currently no centralized entity at Fred Hutch that oversees data-related interactions with these partners, leading to potential fragmentation and lack of coordination in these collaborations.

Furthermore, researchers have expressed a need for democratization of access to these partnerships. Ideally any researcher could be introduced to and initiate collaborations with these partners, irrespective of their research domain or seniority. Such a system could potentially enhance the equitable distribution of

opportunities across the research community, ensuring that all researchers can benefit from these partnerships in their data-driven pursuits.

Data Sharing Agreements

The Fred Hutch research community also identified a need for guidance and support in navigating data sharing agreements. Currently, researchers are uncertain about whether they are responsible for creating such agreements when they wish to share their own data. Further, when they receive data sharing agreements, they often find themselves unsure of the implications and potential obligations these entail.

This uncertainty extends to concerns about what rights they may inadvertently relinquish under these agreements and how they are permitted to use the data governed by such contracts. Questions also arise regarding the necessary security measures and data stewardship requirements in order to comply with these agreements.

Beyond data sharing, researchers require assistance with software licensing, whether it involves their own software or that of others, and with establishing collaboration agreements with data partners where Fred Hutch data may be integrated into third-party systems. Researchers want to ensure that such collaborations are permissible under existing data use agreements and regulations that Fred Hutch or its researchers have already consented to.

In essence, researchers are seeking a cohesive integration of business, compliance, and data aspects of these agreements to ensure they are productive and beneficial for the research community. They need a resource that can provide comprehensive guidance to ensure that these agreements are not just legally sound, but also practically feasible and advantageous for their research endeavors.

Fred Hutch Administrative Needs

Administrative Community Supporting Clinical and Research Communities

The administrative staff of Fred Hutch form a support community for the clinical and broader research communities. We spoke with administrators spanning the research, clinical, and general operations of Fred Hutch. When the administrators were primarily supporting clinical care or clinical research for Fred Hutch patients we identified them as Clinical Administration; when they were primarily supporting non-Fred Hutch patient research we identified them as Research Administration; and when they were supporting general financial, operations, IT, or philanthropic missions of the organization we classified them as General Administration. Overall the needs closely reflected the needs of the respective communities - Clinical Administration data needs were tightly correlated with Clinical community needs and Research Administration needs were tightly correlated with Clinical community needs (**Figure 5**). However, there were also specific data needs that were more common within the administrative community at Fred Hutch.

Clinical Data Access

The administrative aspects of Fred Hutch involve many needs for clinical data access. The Finance department, for instance, requires access to clinical data for a better understanding of medical billing. This data is crucial for understanding and optimizing financial operations related to healthcare services.

Similarly, clinical operations needs the integration of clinical data with scheduling data. This integration is key to optimizing scheduling for patients who need to see healthcare providers, undergo medical tests, or receive cancer treatments, ensuring that patients' needs are met.

The IT Department requires improved data from existing HR systems like Workday for efficient authentication and user identity management within Fred Hutch IT systems. Additionally, IT requires support in designing a

fair and transparent chargeback system for data and computing resources, allowing researchers to make informed decisions based on their anticipated data needs.

The Shared Resources Core presents a unique need for enhanced data access to biospecimen processing combined with clinical metadata about those specimens. This access is essential for facilitating laboratory-based research on biospecimen data, or running a biospecimen repository with sufficient clinical annotations. Without such access, using specimens for research becomes considerably challenging, and the potential of a biospecimen repository to facilitate research is significantly hindered.

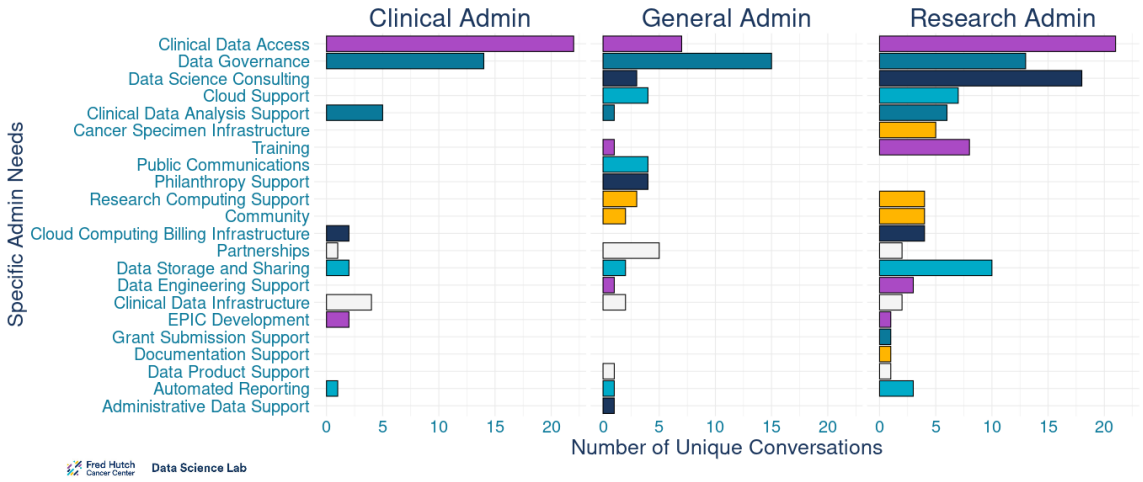

Figure 5 Summary of specific needs for the Fred Hutch administrative community *This figure represents the number of unique conversations where specific areas of need were raised by members of Fred Hutch's administrative community. Needs were assigned to categories based on a controlled vocabulary with six broad categories (shown above, Figure 1 and defined in Appendix A) and 26 specific categories (shown here for the administrative communities, defined in Appendix A). The administrative needs closely reflected the communities the administrators supported, with some unique needs in communication, finance, and philanthropy that were specific to the administrative community.*

Philanthropy

The philanthropy and fundraising team at Fred Hutch needs access to a specific subset of clinical data, which they can then use, within regulatory constraints, to foster relationships with potential donors. Moreover, the philanthropy office is actively involved in coordinating fundraising for data and clinical or cancer care initiatives. This task requires a fine balance of aligning fundraising goals with the scientific objectives of these initiatives, and the team needs effective support to manage these activities. In addition, there's a growing interest from external companies to establish data partnerships with Fred Hutch, with the goal of aiding fundraising for cancer and disease research. The philanthropy team is instrumental in vetting these potential partnerships, ensuring they align with the mission of Fred Hutch and bring value to its research endeavors. The team's role in these partnerships necessitates expert support to ensure effective evaluation and facilitation of such collaborations.

External Partnerships

At Fred Hutch, the BD and IT departments play a crucial role in evaluating potential partnerships, however they need significantly more support, especially when these evaluations are related to data issues. These partnerships fall into a few categories, including those that involve major cloud providers such as Amazon and Microsoft. One goal BD and IT has is to determine whether any vendor offerings are useful at Fred Hutch, a task that requires a comprehensive understanding of both the institution's IT infrastructure and the potential cloud partners' capabilities. Another type of partnership pertains to Software as a Service (SaaS) companies interested in selling their products to Fred Hutch. The BD and IT department need support in evaluating the potential utility of these products for Fred Hutch users and determining their worthiness for investment.

Additionally, the BD and IT departments also need support in assessing partnerships that researchers specifically request for their labs or data-related needs. The Fred Hutch Shared Resources Core, for instance, requires assistance in evaluating potential cloud providers for projects such as the Data Portal.

The Precision Oncology effort at Fred Hutch presents another layer of complexity, as it involves vetting partners in the genomics and sequencing space, such as Caris, Tempus, and Illumina. This effort also necessitates vetting vendors for tasks like handling specimens and managing patient consent systems. In each of these areas, the ability to effectively vet and manage partnerships is a vital need within the administrative structure of Fred Hutch.

Internal Communications

The ever-evolving landscape of data and information technology at Fred Hutch presents its own unique set of administrative needs, particularly in the realm of internal communications. As data systems undergo frequent changes and upgrades, these adjustments necessitate shifts in user behavior. This represents a significant challenge for both the Data and IT teams at Fred Hutch due to the need for both large scale communication of upcoming technological shifts and change management to help support teams to adopt new technologies.

In addition, when significant decisions or large-scale changes are on the horizon, the Data and IT departments seek to garner consensus and support from the research and clinical communities. However, effectively reaching these communities to communicate these needs and changes can be challenging, due to fragmented communication channels, information overload, and difficulties in communicating technological shifts to broad audiences.

Therefore, there is a clear need for improved communication channels that can effectively reach all members of the Fred Hutch community. The goal is to ensure that important information about Data and IT system changes and upgrades is disseminated effectively, promptly, and with the opportunity for feedback from the community. The recent M365 changeover strategy offers examples of how to effectively communicate with the broader Fred Hutch community, but this type of overcommunication needs to be carefully strategized across projects.

External Communications

External communications, led by our Communications and Philanthropy departments, significantly shape the public's perception and interactions with Fred Hutch. Occasionally, the communications or philanthropy departments need informed commentary from data science professionals on biomedical data science initiatives or efforts underway at Fred Hutch. These quotes and comments can serve to articulate the ongoing work, translating complex scientific endeavors into public understanding or donor interest. However, identifying the right person to provide such commentary was identified as an ongoing need.

Data Governance Infrastructure

The Fred Hutch research community operates within a complex web of commitments related to data governance, influenced by a diverse array of ethical considerations, regulatory requirements, patient privacy concerns, IRB stipulations, and data use agreements. These commitments are critical to the integrity of the research and clinical work conducted at the institution, but they also present a significant administrative challenge.

From an infrastructure perspective, understanding the details of these commitments is essential for ensuring that Fred Hutch's technological systems can effectively comply with all agreements. This involves managing data permissions and visibility in accordance with the stipulations laid out in these agreements. Without a clear understanding of these commitments, the Data and IT teams cannot guarantee that the infrastructure is equipped to meet these requirements.

Fred Hutch's legal and compliance teams require detailed information about these data governance commitments to effectively track them, ensure that they do not conflict with one another, and understand how they are implemented within the institution. With this information, they can ensure that researchers and clinicians are in compliance with the various agreements and contracts that govern their data.

Thus, there exists a clear need for a robust data governance infrastructure that can capture and manage this complex array of data-related commitments, serving as a crucial resource for the technological, egal and compliance teams.

Legal and Compliance

In the intricate landscape of data governance at Fred Hutch, the business development, legal, and compliance teams are central figures. They hold the responsibility of making sure that everyone at the institution is within the boundaries of data sharing agreements, a task that requires intricate understanding of the data ecosystem and its nuances. They also need support in executing the demanding task of documenting and managing data use agreements, a process that involves capturing detailed information about a multitude of agreements and ensuring they are effectively implemented and adhered to.

Beyond managing these agreements, there is a critical need for training to help individuals across the institution understand how to comply with and operate within the boundaries set by these data use and data sharing agreements. Furthermore, these teams play a vital role in supporting researchers as they navigate the complexities of writing IRB protocols where data use is involved. This involves understanding the specific requirements of these protocols and ensuring they align with existing data use and sharing agreements.

This underscores the need for a robust data governance infrastructure, as previously discussed, that can provide the necessary support to these teams. This would enable them to more efficiently meet the multifaceted needs of the Fred Hutch community while ensuring adherence to all data-related commitments.

Documentation

The Fred Hutch administrative community is also in need of better documentation support. Many of these departments need comprehensive, up to date, and accessible documentation that can be shared with the broader community. This documentation would encompass the range of available tools and processes, acting as an essential guide to the resources and services that the IT, BD, Philanthropy, Finance,and other administrative departments provide. Similarly, the Enterprise Analytics team, which is responsible for leveraging data to drive decision-making and strategic planning, has expressed a need for support in both developing and sharing documentation. This suggests a requirement for clear, thorough, and accessible records that can effectively communicate how to best interact with the team.

Discussion

Fred Hutch sits at the epicenter of dual revolutions in data collection and computational science. The price of data is falling precipitously and new data collection technologies are coming to market at a staggering rate. At the same time, the explosion of interest in AI is just one indication of the breakneck pace of development in machine learning, statistics, and data science. Here we have summarized the immediate term data needs at Fred Hutch which reflect the fact that data and data science are now part of the day to day professional lives of all of our staff. Addressing these needs will create a solid foundation upon which Fred Hutch personnel can leverage data more easily and in new and creative ways; build data resources and technologies that will power the next generation of discovery and care; and build the careers and data culture at Fred Hutch. This is the groundwork that will make Fred Hutch a leader in the ongoing changes in a field driven by data, machine learning, and new technologies by attracting world-class talent drawn to a world-class data ecosystem.

Appendix A - Controlled Vocabulary with Examples

Here we describe the controlled vocabulary we used for the data analysis and figures throughout the Fred Hutch needs assessment through representative examples. As is to be expected with a community as diverse as Fred Hutch there are a broad range of data needs. Some of these needs are highly specific to a few individuals, but consistent themes emerged from conversations across all the personnel at Fred Hutch. To identify coherent needs across individuals we developed a controlled, hierarchical vocabulary consisting of 6 broad data needs and 26 more specific needs within those broad categories.

Broad Need: Clinical Data Infrastructure

We labeled a need as a clinical data infrastructure need when an individual expressed the need for a piece of technology, hardware, or organized data related to the clinical data of Fred Hutch patients. These needs were further subdivided into the following specific needs.

Specific Need: Cancer Specimen Infrastructure

Example 1: A researcher studying cancer needs access to a biobank that collects and stores cancer tissue specimens along with associated clinical data. The biobank provides services such as nucleic acid extraction, tissue microarray construction, and histology services to support research studies. This researcher needs access to technology platforms for high-throughput molecular analysis of specimens such as next-generation sequencing or proteomics analysis.

Specific Need: Clinical Cost Data

Example 1: A researcher studying a new cancer treatment needs to estimate the cost-effectiveness of the treatment compared to the standard of care. They need access to clinical cost data such as hospital charges, insurance reimbursement rates, and patient out-of-pocket expenses to calculate the total cost of care for each patient in their study.

Example 2: An administrator needs access to clinical cost data to optimize resource allocation and manage costs across different departments and programs. They need to track costs such as personnel expenses, equipment costs, and supply costs to ensure that resources are being used efficiently and effectively.

Specific Need: Clinical Data Access

Example 1: A researcher needs access to clinical data that contains information on cancer patients' demographics, diagnosis, staging, treatment, and outcomes. This clinical data also includes information on laboratory results, imaging studies, and pathology reports. With access to this data, the researcher could conduct a retrospective cohort study to investigate the relationship between a particular treatment and patient outcomes, or to identify potential risk factors associated with a specific cancer type.

Example 2: A researcher needs to identify eligible patients who meet the inclusion and exclusion criteria for a clinical trial. This will require access to clinical data that contains information on patient demographics, medical history, diagnosis, and treatment status as well as data that may not be in the EHR or in a laboratory report (such as genomic data not clinically reported). This researcher also needs to monitor patient safety and track

adverse events throughout the trial, which requires access to clinical data such as laboratory results, vital signs, and medication records.

Specific Need: Clinical Data Analysis Support

Example 1: A researcher needs clinical data analysis support to conduct analyses of clinical data to gain insights into patient outcomes and treatment efficacy. This researcher is studying the effectiveness of a new cancer treatment, and they need help with the statistical analysis of clinical trial data to identify differences in outcomes between the treatment and control groups. This researcher also needs help with data visualization to communicate the results of their analysis to a broader audience.

Specific Need: Data Storage and Sharing

Example 1: A researcher needs to be able to store and share de-identified patient clinical data with a local Fred Hutch research repository to integrate that data with external data sourced from governmental, not-for-profit, or for-profit data producers.

Example 2: A clinical operations administrator needs shared access to provider schedule templates and to link them to appointment and visit information available in EPIC.

Specific Need: EPIC Development

Example 1: A genetics counselor needs summarized clinical genomics information from both internal and external laboratories organized so that it can be visualized longitudinally without having to dig through and manually extract data from pdf reports on sub-tabs of EPIC.

Example 2: A researcher wants to implement a machine learning enabled clinical decision support tool for passive and active validation within EPIC.

Broad Need: Research Data Infrastructure

We labeled a need as a research data infrastructure need when an individual expressed the need for a piece of technology, hardware, or organized data within the context of the broader Fred Hutch research data space beyond Fred Hutch patient clinical data. These needs were further subdivided into the following specific needs.

Specific Need: Administrative Data Support

Example 1: A finance professional would like to integrate and automate the synthesis of multiple data types to streamline reporting of financial information to payers.

Example 2: An administrative professional needs support in assembling data, processing data, and developing a dashboard with key performance indicators (KPIs).

Specific Need: Basic Computing Support

Example 1: A researcher needs help understanding how to use OneDrive for collaboration with other groups across Fred Hutch.

Example 2: A researcher needs help installing and using Slack and Teams for communicating with others around Fred Hutch.

Specific Need: Cloud Computing Billing Infrastructure

Example 1: A researcher wants to understand their up-to-date cloud spend, how it breaks down by individuals within their lab, and the projected costs at current rate of spend for the remainder of the year.

Specific Need: Cloud Support

Example 1: A lab needs to archive data from an instrument that will be sizeable but only rarely accessed. They will need to store that data for a period of time because it is part of a clinical trial. No one in the lab knows how to use cloud data storage and needs help knowing how to put the data in the cloud, then how to access it when they need it.

Specific Need: Data Coordination Support

Example 1: A researcher needs help coordinating data from multiple outside entities including help with the data use agreements, infrastructure, and data cleaning to perform an integrative analysis.

Example 2: A lab needs a systematic way to coordinate the receipt, processing, and analysis of data from an outside center for an ongoing project.

Specific Need: Data Engineering Support

Example 1: A clinical fellow needs help with developing a cloud-based system to organize, store, and share a de-identified set of clinical data with outside groups.

Specific Need: Data Storage and Sharing

Example 1: A lab group generates a substantial amount of genomic data through many different projects over time, that come from many different sequencing labs, and they want to be able to store the data somewhere where their lab members can use it, but also work in collaboration with others in other Fred Hutch labs and with external collaborators. They want to know how best to store their large data to make sure it's cost effective but also allows for granular, flexible data access and sharing.

Specific Need: Documentation Support

Example 1: A translational researcher has generated some datasets in the lab and their staff is struggling to figure out what software and tools they can use to analyze the data. Because the lab hasn't historically done large scale data analysis as their main focus, no one in the lab knows where to go to get help or where to find training on topics they need to learn to effectively analyze their data.

Example 2: A new faculty member is setting up their lab and needs to be able to rapidly direct their new staff to resources available for their data and computing needs. They want to set their lab up to be able to get help when they need it, effectively use the data storage and computing shared resources without having to wait for classes, or seminars or by having to spend a lot of time learning by trial and error.

Example 3: A lab has generated a sizable amount of documentation about how to analyze the particular data they work with and are happy to share that information out with other scientists both at Fred Hutch and beyond. They are interested in getting feedback on those documents from other members in the community in order to create community-based best practices in a novel research space.

Specific Need: Grant Submission Support

Example 1: Due to the new NIH data sharing requirements for new NIH grants, an investigator who traditionally has submitted modular grants that don't generate genomic data is now struggling to know how to address this new requirement.

Specific Need: Research Computing Support

Example 1: A new postdoc has learned that the Fred Hutch has a high performance computing cluster named "gizmo" but they don't know how to access and use it. They have limited knowledge about using a computing cluster but need to get onboarded and be able to move their project forward rapidly without having to learn a lot of Fred Hutch specific details about infrastructure.

Specific Need: Research Data Infrastructure

Example 1: A lab wants to run an analysis on some data that are stored in the cloud but the computing required will be substantial. They need to test their workflows first before running them in the cloud but don't know how to use docker containers on the Fred Hutch cluster which will be a requirement when they move to the cloud.

Broad Need: Data Governance

We labeled a need as a data governance need when an individual expressed the need for access to data or for a process for understanding how and under what conditions they could use the data. These needs were further subdivided into the following specific needs.

Specific Need: Clinical Data Access

Example 1: A physician researcher needs access to EHR data on a subset of patients meeting a specific set of criteria for analysis in a paper.

Example 2: A clinical trialist needs access to demographics, baseline characteristics, health records, and genomic data for evaluating patients against trial inclusion criteria.

Specific Need: Data Governance

Example 1: A clinician wants to access EPIC Slicer/Dicer to extract clinical information into a research data set. They want to know where the data is allowed to be stored and managed and what they have to do to be able to get access.

Example 2: A researcher has signed a data use agreement for a data set which includes restrictions on the type of computing infrastructure they can use; they want to know if they can use the local computing cluster, the cloud, or both to analyze the data.

Broad Need: Partnerships

We labeled a need as a partnerships need when an individual expressed the need for help in developing, keeping track of, or building partnerships or relationships with outside organizations. These needs were further subdivided into the following specific needs.

Specific Need: Partnerships

Example 1: A researcher wants to partner with a local technology company to analyze data collected by the company. They want support to manage and negotiate the terms of the deal to minimize costs, data governance risks, and comply with Fred Hutch policies.

Example 2: A large number of researchers want to partner with a local technology company. However, many of them have the same need and would get more efficient support if they were coordinated. They need help organizing their needs into one coherent ask for the technology partner.

Specific Need: Philanthropy Support

Example 1: A member of the philanthropy team needs support in meeting with a prospective donor who has a data and technology background and would like to identify personnel at Fred Hutch with similar data and technology interests.

Example 2: A researcher at Fred Hutch wants to develop a high-impact technology platform but does not have sufficient funding or technology expertise to complete the project, they want help identifying a potential donor for their project.

Specific Need: Public Communications

Example 1: A member of the communications team would like background on a new data science research project and would like background on how that project could potentially impact cancer care.

Example 2: Members of the communications and business development teams need a data-oriented individual to attend an event for visiting governmental partners.

Broad Need: Training and Community

We labeled a need as a training and community need when an individual expressed the need for help in learning a data-related topic or finding people to interact with around a data-related topic including technical skills, data generating technologies, executive skills, or basic data literacy. These needs were further subdivided into the following specific needs.

Specific Need: Training

Example 1: A laboratory wants to adopt GitHub for version control, but they don't have any members of the lab who currently use that tool who can teach them how to adopt it.

Example 2: A clinical fellow wants to integrate machine learning into a project they are working on and needs support in developing the technical skill set.

Example 3: An administrator wants to learn about chatGPT at a high level to determine if it is a tool that could be adopted within their division to speed up a manual process.

Specific Need: Community

Example 1: A new grad student is working on a project that involves a data type they've never worked with and would like to know if there are other people in other divisions or groups that have experience with it that they could contact for help. Since no one in their lab or on their floor has any experience with the data, they aren't sure how to even go about finding other people to talk to about their work.

Broad Need: Data Science Consulting

We labeled a need as a consulting need when an individual wanted support from a trained data or technology professional to help them work on a project. These needs were further subdivided into the following specific needs.

Specific Need: Automated Reporting

Example 1: Multiple laboratories want automated reporting of data about the people and environment of the Fred Hutch catchment area for research projects, reporting, and grant submissions.

Specific Need: Basic Computing Support

Example 1: A researcher wants support in developing a file naming and management system for their lab.

Specific Need: Data Engineering Support

Example 1: A researcher wants to integrate a particular data type like digital pathology images with electronic healthcare records to create a research quality data set.

Example 2: A clinician wants to move a data set from Google Cloud to our on premises computing infrastructure.

Specific Need: Data Product Support

Example 1: A physician researcher wants to prototype a dashboard that visualizes both electronic healthcare record data and genomics data for supporting clinical care.

Example 2: A public health researcher wants support in developing an R package to make new methodology available to the broader research community.

Data Science Consulting

Example 1: A research laboratory wants 10% of a data scientist to work in their lab to help manage data, perform analyses, and apply new machine learning methods to augment their team's skills.

Specific Need: Grant Submission Support

Example 1: A researcher is submitting a grant for a coordinating center where a requirement of the grant is the use of a cloud computing platform and they need a letter of support.

Example 2: A researcher needs help with writing a data sharing and management plan in support of their NIH grant submission and wants to understand the costs of the corresponding plan to write them into their grant.

Appendix B

Overall, what are you trying to accomplish in your research?

- How would you describe the scientific goals of your research work?
- What are the final products of your research work? Publishing papers? Sharing datasets? Maintaining software? Launching a drug? Completing a clinical trial?
- What does success look like for you and your research?

What types of tasks make up your day-to-day research hours?

- How much time do you spend in different physical settings? (in front of a computer, at the lab bench, in meetings, in the clinic with patients)
- What degree of administrative or future planning work (project planning, grant writing, etc) do you do versus implementing plans or research processes (experiments, analysis, publication, distribution of results, etc)?
- How much time do you spend in a mentoring or advising capacity either as it relates to career development or to performance of the research itself?

What do the mechanics of research workflow processes work in your group?

- Where does your data originate? In your lab, from a clinical trial, medical records, external datasets?
- Who is involved in receiving/generating and managing the data assets you rely on? Your staff, IT, shared resources, external collaborators?
- What role does data analysis play in your research work and who performs it? Lab staff analyzing their own generated data, statistical collaborator,
- How do you communicate your research outputs or results? Publication, pass curated data sets to collaborators for final analysis, share software, etc.

What hardware and software tools do you use in your typical research work? How do you use these tools?

- Do you interact with data via a PC, a shared computing cluster, web-based resources, etc.
- What software do you use during your workflow processes and where do you get it from/how do you pay for it/how do you manage it?
- If you generate your own software, scripts or code, do you have an example artifact from a recent analysis, like a code file, that we could look at together?

What software tools do you use to collaborate and communicate with partners involved in your research? How do you use these tools?

- Are there different tools for internal and external partners?

What human resources do you leverage during your research?

- Primarily only your own staff, Hutch shared resources, IT partners, external collaborators, etc.
- What communities do you engage with for support, advice or learning during the execution of your workflow? How?

What is the organizational structure of your research work?

- How do you store and organize: Data? Code and software? Project documentation?

What processes are involved in your research workflow that you feel result in unnecessary delays, staffing/expertise challenges, or prevent you from executing projects you'd like to do?

- Do you encounter any particularly difficult or chronically reoccurring challenges in your data analysis and coding work?
- Data access challenges either in knowing what is available, or knowing how to leverage what is available?

When these challenges exist, to whom do you look to for support, or who do you believe should be able to or wish was able to support you?

- When you encounter challenges, what resources do you use to seek help?

- To what extent does the data and computing infrastructure affect your research productivity?
- What is your perception/relationship with resources that are available to you? What do you wish you could get help with?