Data Management in the NIH HEAL Initiative

Executive Summary

The Helping to End Addiction Long-term℠ Initiative, or NIH HEAL Initiative℠ Research Plan describes the wide variety of studies aiming to discover scientific solutions to the crises of pain, opioid misuse, addiction, and overdose. The breadth and depth of the HEAL's research portfolio is unparalleled – over time, generating an extensive collection of findings and datasets that will be the basis of future research in many different biomedical domains. The complexity will present data-related challenges, yet data sharing is a high priority to ensure HEAL-funded research investments best address this urgent public health crisis. The HEAL Data Ecosystem aims to transform research data, findings, and publications into a virtual, annotated, searchable catalog in which datasets and findings from different studies can be analyzed, compared, and combined.

The NIH HEAL Initiative is building a foundation for the HEAL Data Ecosystem through multiple assessments and data-gathering efforts. These fact-finding missions have shed light on how HEAL researchers create data, what data standards they use, and how these researchers might collaborate with others to maximize the value of HEAL data – not only for researchers but for people and communities on the front line of the pain and opioid crises. To begin to understand data needs and related issues, 37 current HEAL awardees were interviewed about their knowledge about and use of scientific data standards, data sharing, and reuse of research data. Key findings and analyses from these interviews are summarized below, and they are detailed in the full report. Although forthcoming assessments will gather comprehensive input from the full HEAL community, these initial findings are instructive toward guiding development of the centrally accessible HEAL Platform, which will support the HEAL mission by maximizing access to and use of HEAL findings, datasets, analysis tools, and other resources.

Key Findings:

- Most HEAL clinical awardees (90%) stated that they are aware of data standards in general and would use them as part of their HEAL research. By contrast, only 12% of the preclinical researchers stated that they would use some data standard as part of their HEAL research.
- "Metadata" is a novel term to most HEAL researchers, even though they may routinely use metadata in practice.
- HEAL clinical researchers are highly reliant on databases for data storage, whereas preclinical researchers generally store data in formatted files created by lab instruments or analysis tools, rather than in databases.
- Some clinical and preclinical HEAL researchers will produce "big data" as part of their HEAL research. Roughly one third of the researchers expect to create data on the petabyte scale (1 petabyte is 1 million gigabytes) or terabyte scale (1 terabyte is 1,000 gigabytes).

- All HEAL researchers surveyed rely on statistical and graphing tools, but clinical researchers were more reliant on relational platforms for their analyses, while preclinical researchers used many instrument-specific analysis tools.
- HEAL respondents were generally enthusiastic about HEAL's efforts to promote open science through its cloud-based platform, for both collaboration and hosting of analysis tools.
- HEAL researchers are generally aware of the HEAL Public Access and Data Sharing plan, but are looking for additional guidance about its implementation.
- Roughly one-third of HEAL respondents will be using an existing public repository as an archive for their research data; for clinical researchers, this may be a condition of their award.

What is Data Management and Why Does It Matter?

Maximizing the value of scientific data can be generally informed by use of FAIR principles that describe how data is ideally Findable, Accessible, Interoperable, and Reusable. In the realm of clinical studies where many of the HEAL Initiative programs operate, FAIR principles are also conditioned by requirements to abide by established governance and privacy laws, as well as to be respectful of individual and community consent.

FAIR principles describe a set of goals, but not the processes needed to meet those goals. Groups aiming to create data that meets general FAIR requirements will always be directing their efforts in context-specific ways, to satisfy those requirements according to how their own communities use data, and what tools and platforms they use. In this context, for the HEAL Initiative's community of researchers, interoperability and reusability requirements are fundamentally addressed by maximizing the uniformity of the HEAL catalog of data sets. Such uniformity of content enables data reuse across data sets, comparable to how a single currency or language enables commerce or communication across communities or nations.

Some of this uniformity is achieved through use of shared scientific data standards. A data standard is a defined and agreed-upon set of terms used in a scientific data set. For example, a data standard may describe a tissue, cell type, or gene name determined by the community itself or by another standards-setting group. Data uniformity is also addressed by establishing shared experimental protocols or methods across research programs, enabling generation of uniform data from different laboratories. Finally, data uniformity is advanced by sharing the same methods of data collection, which in a clinical setting may require, for example, subject self-reporting using defined questions and survey instruments, which would depend on predefined common data elements (CDEs).

In summary, HEAL data management can be broadly defined as the processes (both human expertise and software technologies) that can receive structured data submissions from a wide variety of basic, preclinical, and clinical research programs, and can create catalogs of data that can be searched and reused. These processes will work optimally when the HEAL research community and data managers can agree on data standards, ideally before the research data sets are generated.

Assessment Background

Assessments of the requirements for a HEAL Initiative scientific data strategy were conducted in 2019 and 2020, based on interviews with clinical and preclinical researchers, HEAL teams leading data coordination efforts for HEAL clinical studies, NIH program officers, and NIH leadership. The goal of those assessments was to broadly categorize HEAL research and its data types, describe the data management challenges that HEAL may face, deliver technical and operational options and recommendations to HEAL on how to build cloud-based data platforms, and staff the team that would be required to maintain those platforms. Selected findings from that assessment are summarized as follows:

- The HEAL Initiative will be gathering an incredibly diverse set of research data.
- The degree of compliance with a given data standard across HEAL projects had not been pre-determined, as such even researchers in the same research domain may not have been using the same standards.
- Many awardees are looking to the HEAL Initiative for guidance on data standards and how to use them. An example of this guidance was the effort led by NIH (for HEAL Initiative use) to create CDEs for self-reported pain.
- Many HEAL awardees and NIH program staff were unsure about the level of effort that would be required to standardize and submit HEAL data.
- There are a number of important existing or emerging data standards of different types that the HEAL Initiative can use.

This report, *Data Management in the NIH HEAL Initiative,* builds on the findings of the interviews and landscape assessments previously conducted by Bioteam and simultaneously extends that work with an increased focus on individual awardees and their data management practices. Below is a summary of the findings from a request for data (20 questions about data management, spanning December 2020 to February 2021) to 37 HEAL awardees in the clinical and preclinical research areas.

HEAL Data Call Findings

It is useful to consider the Data Call findings with respect to research domains:

- **Clinical trials** ranging in scope from those with a small number of healthy human participants in Phase I exploring *in vivo* tolerance and pharmacokinetics to Phase II and III trials exploring efficacy in randomized controlled studies with longitudinal participant tracking.
- **Observational studies** in which participant treatments and outcomes are studied over a period of time, but no intervention is made to affect the outcomes. These studies are important for understanding differences between populations experiencing similar and differing environmental or community-level stressors.
- **Therapeutics development** working toward validation of new drugs, treatments, or devices on the path to approval for human use, and which may focus on human subjects and animal models.
- **Preclinical research** in human subjects and animal models on the mechanisms of pain and opioid use disorder. This HEAL work frequently focuses on drug target identification and validation, the study of novel biological mechanisms, and discovery and validation of potential biomarkers for pain and opioid use disorder.
- **Basic research** on mechanisms of pain and opioid use disorder in *in vitro* systems and cell lines, as well development of novel assay methods, computational methods for diagnosis, and research on biomarkers.

Each of these research domains tends to generate and collect data in characteristic ways, and thus may employ different data standards. These different approaches pose unique data management challenges.

Research of the Data Call Respondents

The HEAL Data Call respondents were asked to characterize their research using *one or more* research domains in order to understand the full diversity of data being generated by HEAL (i.e., clinical trial, observational study, therapeutics development, preclinical research, basic research).

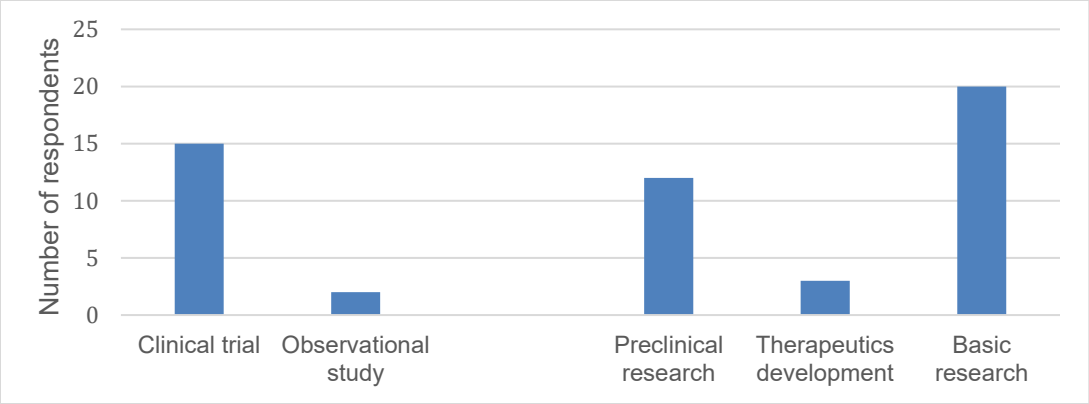

Figure 1 – Number of data call respondents working in the HEAL research domains.

Roughly two-thirds of respondents selected just one of the five possible domains, and the remaining third selected two or three domains (Fig. 1). Choosing multiple domains is consistent with the translational nature of HEAL research, which has the ultimate goal of providing more effective treatments for opioid use disorder and pain, starting with basic research. These results allowed us to categorize the Data Call respondents into two groups:

- 15 clinical trial and observational study researchers (defined as “clinical” in this report)
- 22 preclinical, therapeutics development, and basic researchers (defined as “preclinical” in this report)

The clinical respondents were part of nine HEAL programs:

- HEALing Communities Study
- ERN (Effectiveness Research Network)
- STOP Trial (Subthreshold Opioid Use Disorder Prevention)
- EPPIC-Net (Early Phase Pain Investigation Clinical Network)
- Clinical Research in Pain Management
- ACT NOW (Advancing Clinical Trials in Neonatal Opioid Withdrawal)
- PRISM (Pragmatic Studies for Pain Management Without Opioids)
- BRIM (Behavioral Research to Improve Medication-Based Treatment)

The preclinical respondents were part of these HEAL programs:

- NIDA SBIR
- Target Discovery and Validation

Use of Data Standards

The HEAL Data Call respondents were asked about data standards in order to comprehensively evaluate the variety of HEAL standards.

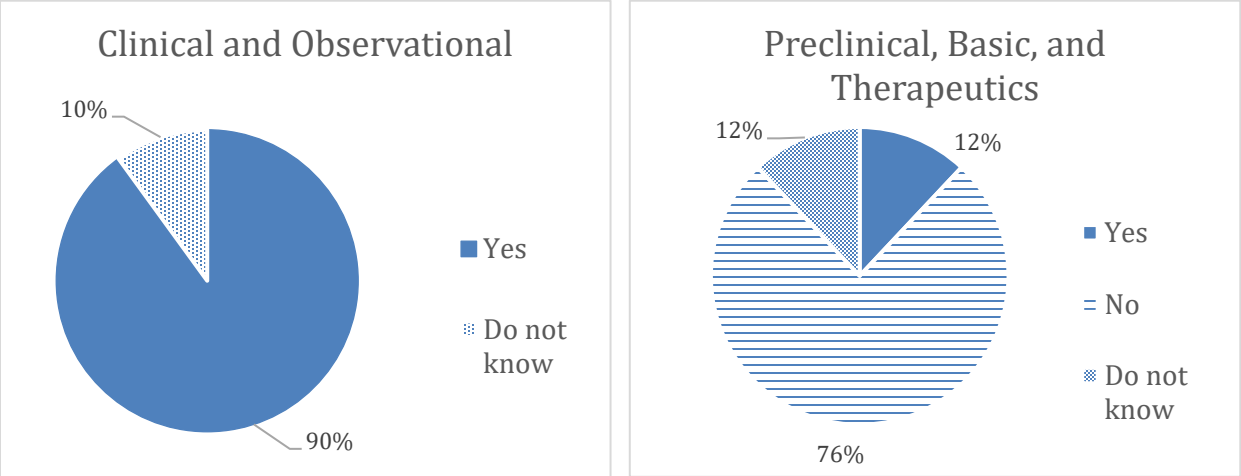

Figure 2 – Use of data standards by HEAL researchers.

As shown in Fig. 2, most (90%) clinical researchers said they would use data standards in their HEAL research, including LOINC, RxNORM, NDC, ICD, MedDRA, WHODrug, CDISC, CDISC SDTM, CDISC CDASH, CDISC ADaM, PhenX, DICOM, SNOMED, NCI Thesaurus, MIAPE, Drug Attributes Data, MED-File, and CTCAE. Some HEAL clinical researchers also noted their use of CDEs, for example from PROMIS or NINDS. Researchers also described how their use of certain standards was dictated by the terms of their HEAL awards.

Use of data standards is much less common in basic and preclinical research, as shown above (reported by only 12% of respondents). The data standards mentioned by HEAL preclinical and basic researchers were imzML and DICOM.

Use of Metadata Standards

HEAL Data Call respondents were asked about metadata (Fig. 3), which is usually defined as "data describing data" (e.g., HEAL study name, researcher name, HEAL award number). Examples include Define-XML (CDISC) and Metadata Acquired from Clinical Case Reports. There is no shared, global definition of metadata and any given research group may have functional, local definitions. For example, a minimal and restricted definition of metadata includes the source of some data – comparable to the small set of terms used to define a scientific publication. An example of a broader definition of metadata may include all the field names used in a given data set. Therefore, it is not surprising that some researchers did not identify their use of metadata – they may be using metadata without realizing it (and as such have not defined it) or are not calling it metadata by name.

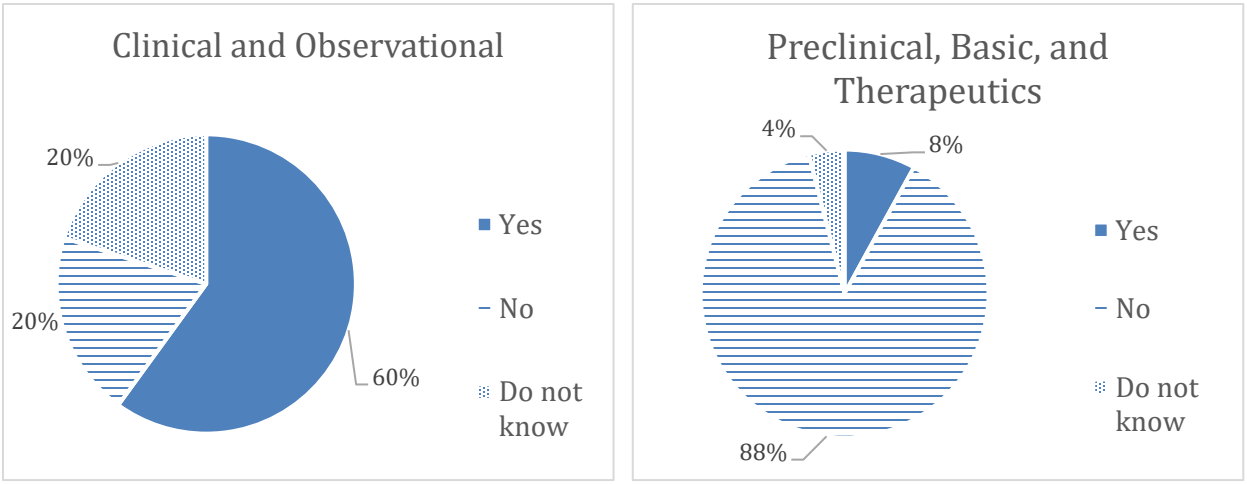

Figure 3 – Use of metadata standards by HEAL researchers.

For HEAL clinical researchers, the most frequently mentioned clinical metadata standards were internally developed vocabularies, presumably used by the organization or the laboratory over a number of studies. However, the public metadata standards mentioned most frequently were CDISC Define-XML and CDISC SDTM (some standards can be considered both data and metadata standards). Some respondents also mentioned their use of REDCap as a metadata repository.

HEAL preclinical researchers mentioned those metadata standards used to describe various data types/techniques such as RNA-Seq, fMRI, and MSI (mass spectrometry imaging), either derived from some open source initiative (i.e. the Proteomics Standards Initiative) or developed by the instrument manufacturer.

Data Formats

HEAL Data Call respondents were asked about one or more general data formats they use in order to understand how their data might be stored and exchanged across HEAL. The following general definitions could apply:

- A relational database is software for data storage where granular data is organized into linked “tables” of related data. Examples include Oracle, MySql, or platforms using relational databases such as REDCap.
- Files could be single documents or document collections organized hierarchically as directories. Examples include Excel spreadsheets, raw output from lab instruments, or images.
- A document database is software for data storage that accepts files and structured data as input and allows searching over all the content. Examples include MongoDB and DynamoDB.
- Object storage is software for data storage where files and file collections are tagged with useful metadata and stored non-hierarchically and retrieved based on custom identifiers. Object storage is the default storage mode in cloud-based environments such as Amazon Web Services and Microsoft Azure.

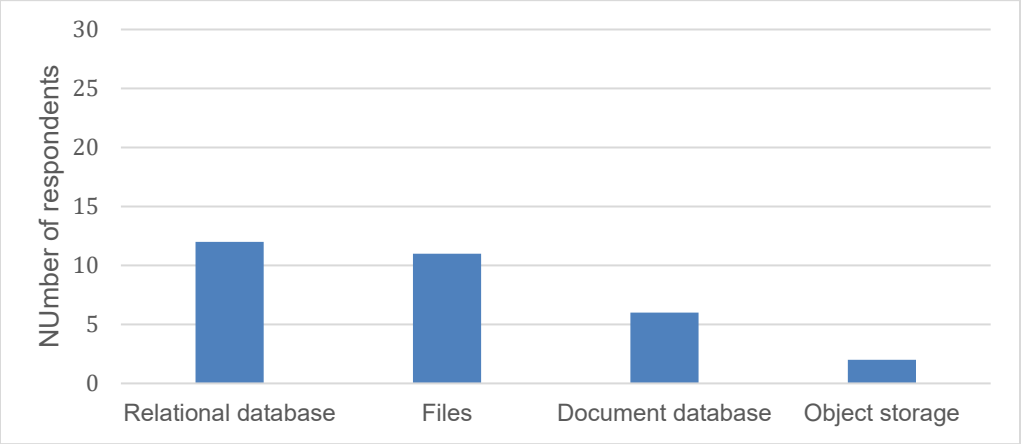

Figure 4 – Data formats used by clinical and observational researchers.

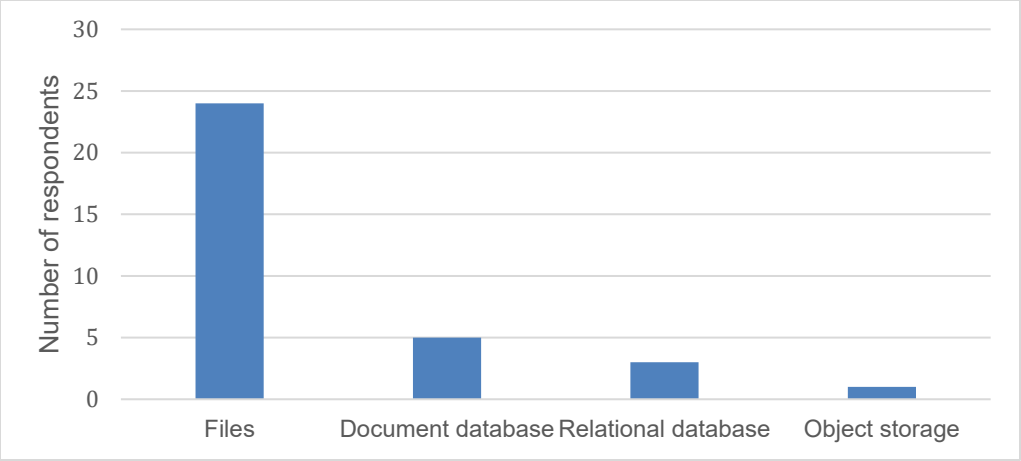

Figure 5 – Data formats used by preclinical, basic, and therapeutics researchers.

As shown in Fig. 5, the Data Call's main finding is that clinical and observational researchers are highly reliant on databases, which are used to store study data, and could also be used for subsequent data mining and analytics. By contrast, preclinical and basic researchers rarely use databases, and store their data primarily in formatted files, frequently created by lab instruments or instrument-specific analysis tools. These choices could also be influenced by security requirements, where study or trial data can be more effectively protected if stored in a database.

The Role of Consent

HEAL Data Call respondents were asked about the role of consent in order to obtain information on requirements related to data security and privacy (Fig. 6).

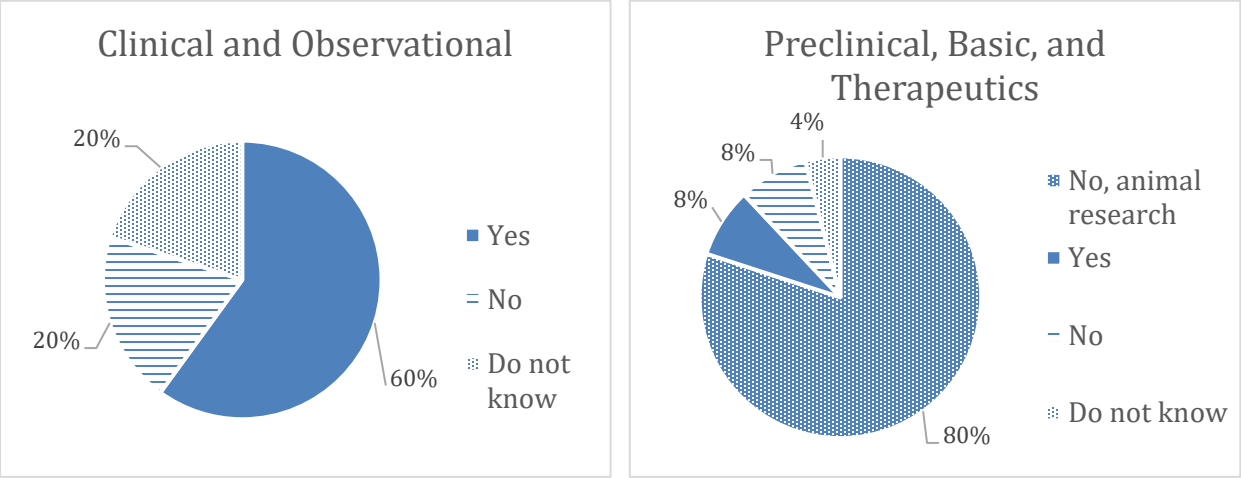

Figure 6 – Percentage of researchers using consent information to control data access.

Attention to consent-related processes vary by research type (see Fig 6). Most preclinical and basic researchers use animal models, obscuring the relevance of informed consent. By contrast, most clinical researchers expect that consent information will be used to manage access to their study data.

Predicted Data Volumes

HEAL Data Call respondents were asked to estimate the amount of data that their HEAL research might create (Fig. 7). These estimates will be useful for anticipating cost and data complexity in the HEAL platform.

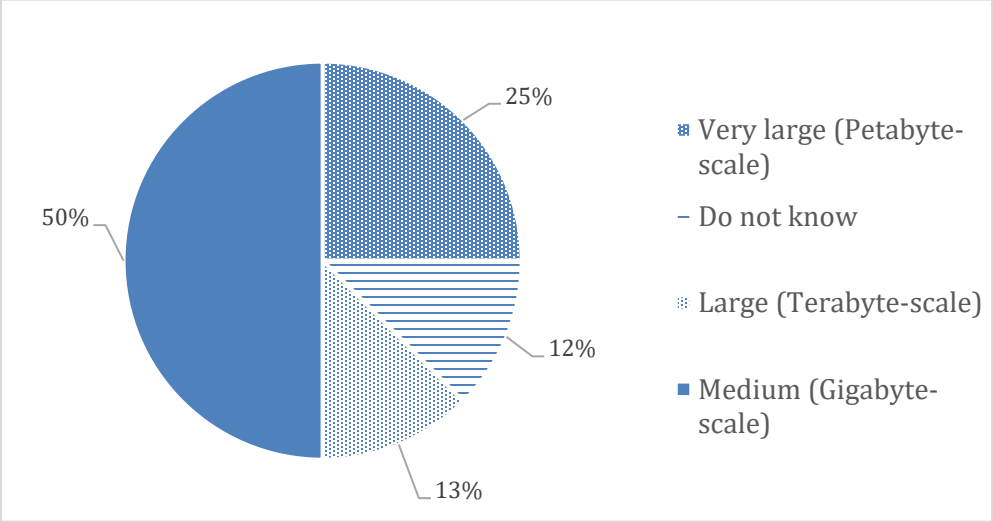

Figure 7 – Predicted HEAL data volumes.

These results are in good agreement with findings from prior HEAL assessments, where most researchers stated they had only modest storage requirements. However, some researchers were operating at the petabyte or terabyte data scales. The HEAL researchers (both clinical and preclinical) that expect to generate very large data sets will generate various file types from a range of experimental methods, including TENS (transcutaneous electrical nerve stimulation), electrophysiology, microscopy, genomic/transcriptomic sequencing, cell sorting, medical/DICOM images, fMRI, MSI , and EEG/EKG recording.

HEAL Computing Resources

HEAL Data Call respondents were asked about computing resources they use in order to anticipate possible needs in the HEAL platform.

Figure 8 – Computing resources used by clinical and observational researchers.

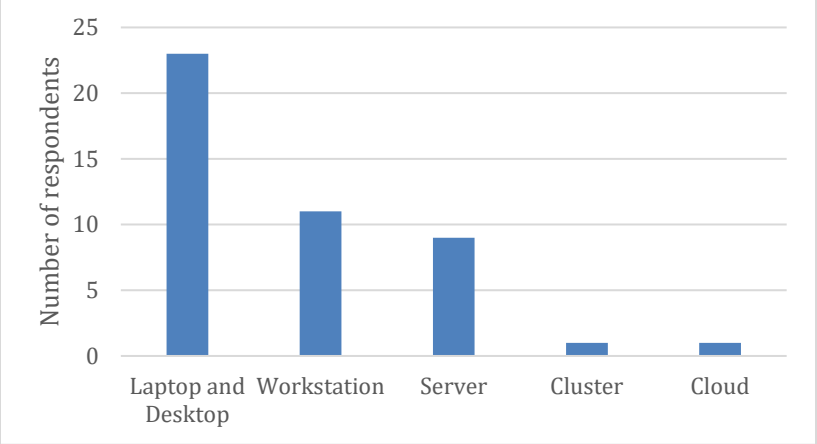

Figure 9 – Computing resources used by preclinical, basic, and therapeutics researchers.

There do not appear to be significant differences between the computing resources used by clinical and preclinical researchers (Fig. 9). Both groups mostly rely on computing resources that can be used in the laboratory (or in some local location), as opposed to supercomputing cluster resources or cloud-based resources.

Widely Used Applications

HEAL Data Call respondents were asked to list their “top 3” research applications (Figs. 10, 11), and these answers may inform the choice of analytic applications that might be supported by the HEAL platform.

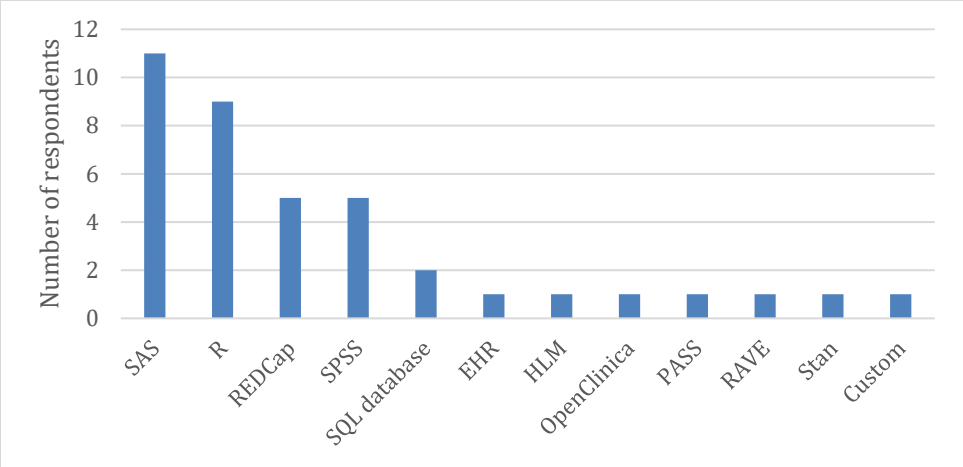

Figure 10 – Key applications used by clinical and observational researchers.

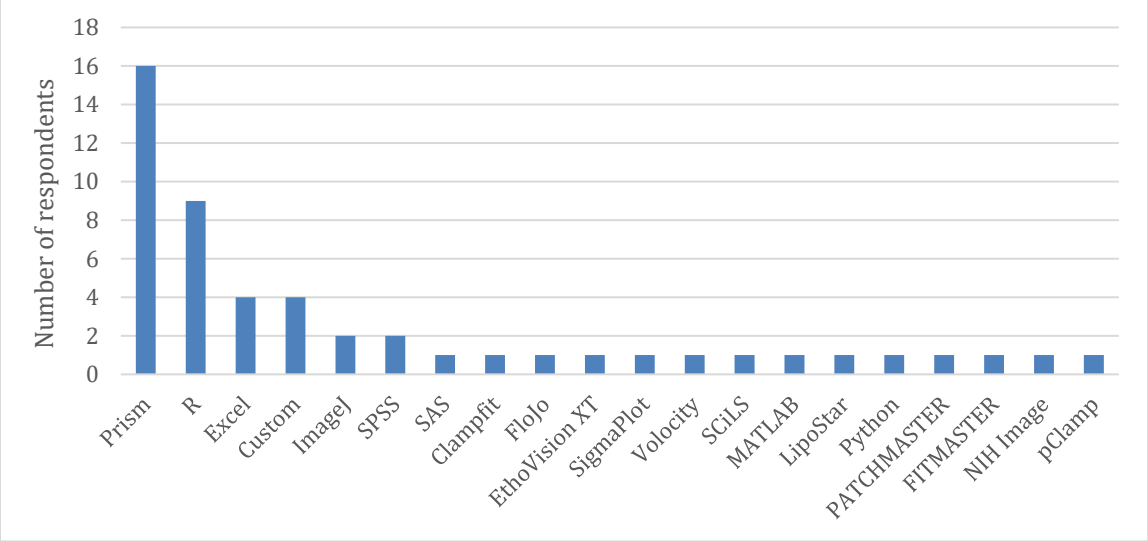

Figure 11 – Key applications used by basic, preclinical, and therapeutics researchers.

Reliance on statistical and graphing tools and languages is shared by all researchers in this Data Call. Clinical groups who responded use database platforms and relational database management tools (e.g., REDCap, “SQL”, EHR, Medidata Rave); while preclinical researchers who responded use instrument-specific analysis tools.

Use of the HEAL Platform

The HEAL platform will provide search and browse functions to connect to HEAL data and cloud-based workspaces for data analysis. HEAL Data Call respondents were asked about possible uses of the HEAL platform in order to inform the ongoing design and build of the platform (Fig. 12).

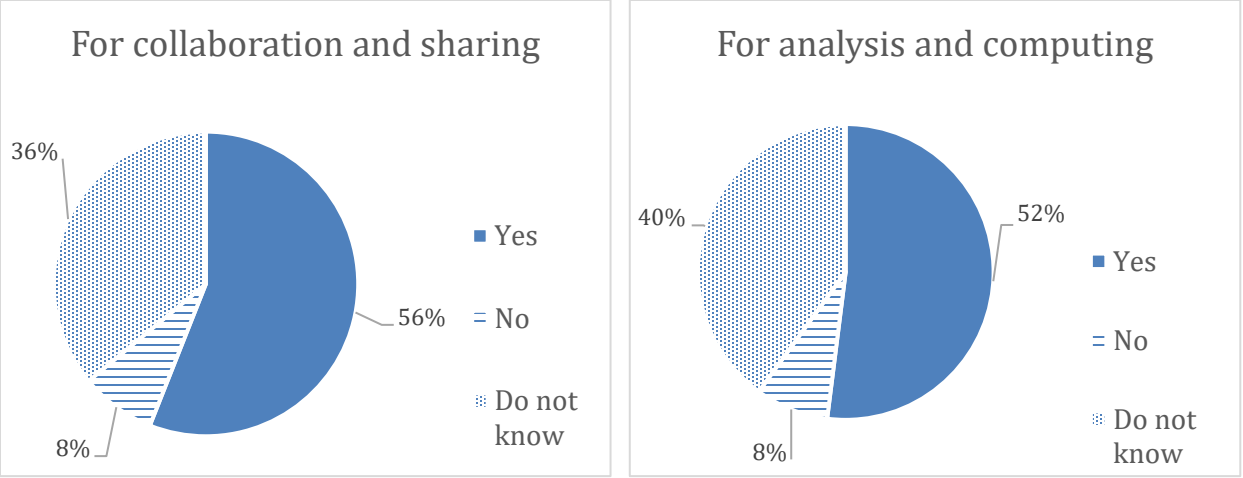

Figure 12 – Percentage of researchers intending to use the HEAL platform.

In previous assessments and in this Data Call, many respondents expressed enthusiasm when contemplating the shared resources that could be built by the NIH HEAL Initiative. For many respondents, a secure cloud-based platform that offers opportunities for collaborative research, or data that can be accessed to further their own research, is both novel and a potential step forward for their own research as well as for biomedical science broadly.

HEAL Data Sharing

HEAL Data Call respondents were asked about how they have planned to implement HEAL's Data Sharing Policy (Fig. 13).

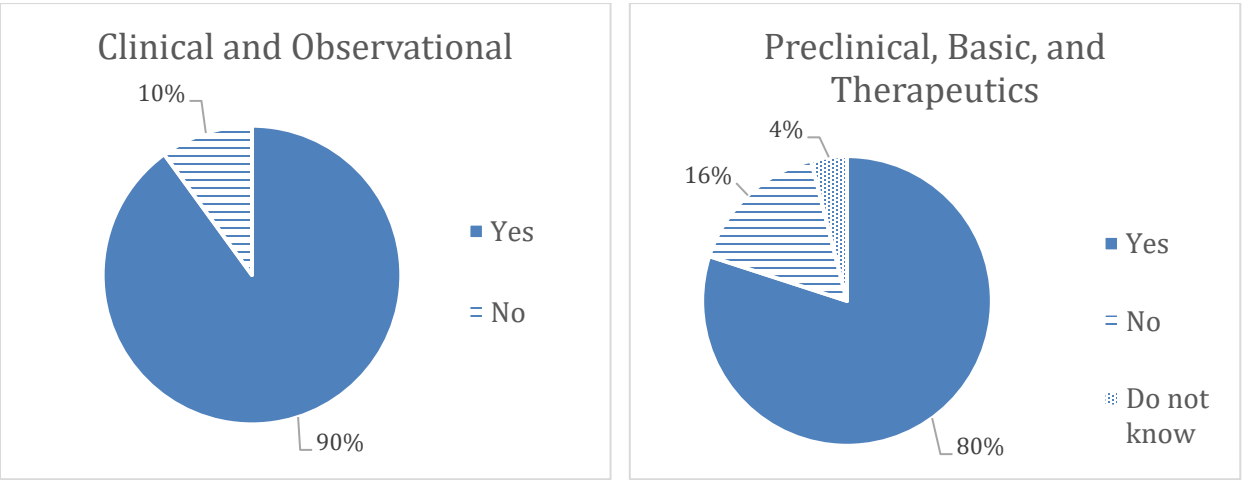

Figure 13 – Percentage of researchers with a HEAL data sharing plan.

Most HEAL researchers are aware of the general requirement for data sharing and have thought about their own data sharing plan, but when asked, only a few researchers described their data sharing plan in detail. Respondents were also asked for their general comments on HEAL data management, and some respondents indicated they are looking forward to working with HEAL to satisfy the data sharing requirements, for example:

- "We would like for the [HEAL] data team to be very explicit with sites in order to know how to prepare the data and upload the data for harmonization purposes."
- "We look forward to partnering with the HEAL initiative and are committed to data sharing. If HEAL has a specific data model that can be shared with us, we can use it to build an ETL process to conform to the applicable segments of the data model."

The Role of Public Repositories

HEAL Data Call respondents were asked about public repository use (Fig. 14) in order to understand where data will be stored across HEAL (public repositories known to the biomedical community in general include FigShare, clinicaltrials.gov, and dbGaP).

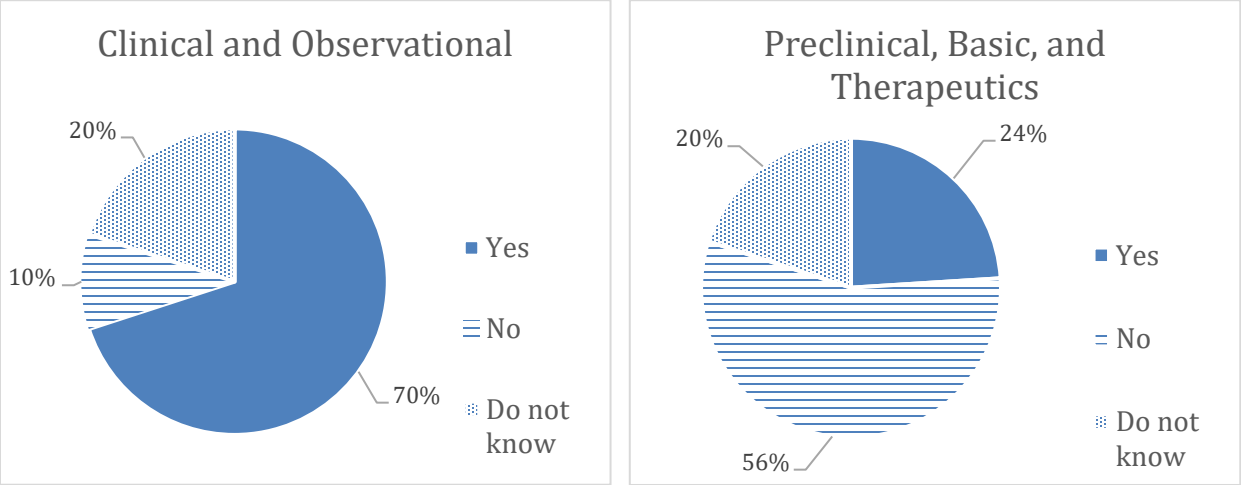

Figure 14 – Percentage of researchers intending to use existing public repositories for HEAL data.

Most clinical researchers will use an existing public repository. The public repositories mentioned by HEAL clinical researchers were NIDA Data Share (clinical trials), NCI NCORP (clinical trials), clinicaltrials.gov, and NICHD DASH (deidentified data from NICHD studies). Many researchers noted that they were required to deposit their data in these repositories according to the conditions of their awards.

Most preclinical researchers who responded have no plan to use an existing public repository. Seven preclinical researchers did say they would use specialized public repositories, including NCBI Sequence Read Archive (SRA, for RNA-Seq data), NCBI Gene Expression Omnibus (GEO, for RNA-Seq data), Dryad (microscopy images), Metaspace (MSI data), and NeuroVault (fMRI data). It is likely that these researchers routinely use these repositories to archive their research data, as these repositories are both widely used and specialized, and may even host analysis tools relevant to a specific data type. It is possible that a given repository would only hold a fraction of the data for a HEAL research publication. For example, a hypothetical HEAL researcher performing target validation might only deposit RNA-Seq data into GEO – not other data such as microscopy or related data files.

Conclusion

This report, *Data Management in the NIH HEAL initiative*, provides a limited snapshot of data management and stewardship activities across the initiative. HEAL will use this information as a basis for a more broad and thorough landscape analysis, with the goal of reaching out to all investigators across HEAL for information about their data-management plans, data-related activities and needs. Collectively, these findings will inform data-management plans for the centrally accessible HEAL Platform, which will support the HEAL mission by maximizing access to and use of HEAL findings, datasets, analysis tools, and other resources.

Appendix

The following sections contain information relevant to this assessment.

Data Call Questions

The following questions were asked in the Data Call:

1. What data domains are you working in for HEAL?
2. Do you use some consistent form of metadata to describe your HEAL data?
3. Do you use any public data standards in your HEAL work?
4. What general resources will be used for storage of your HEAL data?
5. Will you be storing any data in commercial clouds?
6. Will you be storing any HEAL data in public repositories?
7. Are you developing APIs to allow access to your HEAL data?
8. Are you using or planning to use subject consent information to control access to your HEAL data?

9. Are you using or planning to use some form of user authentication to ensure secure access to your HEAL data?
10. Are you using or planning to use some form of data authorization to ensure secure access to your HEAL data?
11. Will you be creating any of these large data files as part of your HEAL work?
12. What kind of computing resources are you using for HEAL data analysis?
13. What are the Top 3 applications that are central to your HEAL work?
14. Do you have a plan to share your HEAL data upon publication?
15. Do you need or have a process for receiving, reviewing and approving Data Access Requests?

Questions 7, 9, and 10 about APIs, authentication, and authorization were only asked if the respondent was part of a HEAL data coordinating team (ten respondents total), and these questions are not discussed in this report. Five of the questions (2, 3, 5, 6, 14) were conditional and asked for further details if the response was positive.