

Novel and Exceptional Technology and Research Advisory Committee

**Data Science and Emerging
Technology in Biomedical
Research Draft Report**

August 2023



NATIONAL INSTITUTES OF HEALTH
**Novel and Exceptional Technology and
Research Advisory Committee (NExTRAC)**



**Data Science and Emerging Technology in Biomedical Research
Working Group**

Roster

Co-Chairs

KHETERPAL, Sachin, MD, MBA*
Associate Dean for Research Information Technology
Professor of Anesthesiology
University of Michigan Medical School
Ann Arbor, MI 48109

OSSORIO, Pilar N., PhD, JD
Professor of Law and Bioethics
University of Wisconsin
Madison, WI 53706

Members

ALBRITTON, Lorraine M., PhD
Professor Emeritus
Department of Microbiology, Immunology, and
Biochemistry
College of Medicine
The University of Tennessee Health Science Center
Memphis, TN 38163

BAKKEN, Suzanne B., PhD, RN, FAAN, FACMI*
Professor of Biomedical Informatics
Alumni Professor of the School of Nursing
Columbia University
New York, NY 10032

BLOSS, Cinnamon, PhD
Professor
Herbert Wertheim School of Public Health and
Human Longevity Science
University of California, San Diego
La Jolla, CA 92093

LESHNER, Alan I., PhD
Chief Executive Officer, Emeritus
American Association for the Advancement of
Science
Potomac, MD 20854

MANJI, Hussein K., MD, FRCPC*
Visiting Professor
Oxford University
Past Global Therapeutic Head for Neuroscience
Janssen Research & Development
Oxford, England

MATHEWS, Debra J.H., PhD, MA*
Associate Director for Research and Programs
Berman Institute of Bioethics
Associate Professor, Genetic Medicine
Johns Hopkins University
Baltimore MD 21205

RILEY, Margaret F., JD
Professor, School of Law
Professor of Public Health Science
School of Medicine
Professor of Public Policy
Batten School of Leadership and Public Policy
University of Virginia
Charlottesville, VA 2290

*Ad Hoc Members

CONTENTS

Executive Summary.....	1
I. Introduction	4
II. Background and Overview	7
III. Research Questions on Data Science.....	9
A. Introduction and Background.....	9
B. Types of Research Questions to Consider Through Public Engagement.....	11
IV. Public Engagement to Inform Implications of Research Questions and Technologies	15
A. Collection of Feedback from Public Engagement and Methods of Analysis	15
B. Community Conversations Across the United States.....	16
C. Workshop: Using Public Engagement to Inform the Use of Data in Biomedical Research.....	22
V. Overall Observations, Findings, and Recommendations for NIH	26
VI. Conclusion.....	32
VII. Appendix 1: Further Description of Phase 2 Efforts	33
A. Schedule of Community Conversations and Webinars	33
B. Case Studies and Analysis of Case Study Conversations in Phase 2.....	33
IX. Appendix 2: Generative AI as a Use Case Study for WG Recommendations	39
A. Introduction	39
B. Case Study.....	39

Executive Summary

Advances in data science hold significant promise for improving our understanding of human health and disease, but certain applications of data science can also undermine research participants’ autonomy, privacy, or the experience of respect when contributing to biomedical research.¹ Used improperly, tools developed using emerging data science could entrench or exacerbate inequities in health or health care. The National Institutes of Health (NIH) seeks to understand the policy and ethical implications of emerging data science technologies, and how emerging technologies enable the use of datasets from disparate sources (e.g., data from research and non-research settings) for health research. NIH also seeks to anticipate potential benefits and risks of emerging data science for research participants, families, populations, researchers, and society. To assist in evaluating recent advances in data science in biomedical research, NIH turned to the Novel and Exceptional Technology and Research Advisory Committee (NExTRAC), a federal advisory Committee that provides guidance to the NIH Director on the use of emerging technologies in biomedical research, to help ensure that research progresses responsibly and meets public expectations for the Nation’s investment in research. The NExTRAC was charged to:

- Define and characterize the types of research questions that require increasing granularity and aggregation of data about individuals that are likely to be addressed through emerging technologies
- For those questions and technologies defined above, consult with stakeholders to discuss and assess the value of and potential implications for individuals, groups, and society

To address the charge, the NExTRAC established the Data Science and Emerging Technologies Working Group (WG). The WG split their effort in two separate phases. As part of Phase 1, the WG met with subject matter experts in health-related data science to discuss current research efforts and solicit input on emerging technologies and ethical issues in the field. Based on these discussions and further analysis, the Committee identified three overarching topic areas and corresponding types of research questions, outlined in the table below:

Topic Areas	Types of Research Questions
Data: Use of novel data from outside of the traditional healthcare system	How are personal health data ² collected from outside of the traditional health system or research setting (e.g., through fitness trackers, apps, and social media posts) being used to study health-related questions and predict health risks, at an individual, family, group, or public health level?
	How can/should other consumer and lifestyle data from non-health-specific sources (e.g., sensors in the home and credit card and consumer rewards data) be used to study health-related questions and predict health risks?
	How can/should health data be integrated with data on social determinants of health (SDOH) to enable better risk prediction and development of predictive algorithms?
Algorithms: Use of models and algorithms	What is the role of computer-based algorithmic technologies, such as artificial intelligence (AI), natural language processing (NLP; including large language models like ChatGPT), machine learning (ML), and automated image analysis in biomedical research and in health decision-making?

¹ Biomedical research is “Research that addresses mechanisms that underlie the formation and function of living organisms, ranging from the study of single molecules to complex integrated functions of humans and contributes profoundly to our knowledge of how disease, trauma, or genetic defects alter normal physiological and behavioral processes.” National Research Council (2005). *Advancing the Nation’s Health Needs: NIH Research Training Programs*. Washington, DC: The National Academies Press. <https://doi.org/10.17226/11275>.

² In this report, the Committee considers “personal health data” to be data collected about individuals or groups and analyzed for the purpose of addressing health-related questions in biomedical research. This broad definition encompasses data that are collected in clinics and health care settings (the “traditional” health system) and person-generated health data (PGHD; i.e., collected by the individual through personal devices designed to collect health-related metrics). These may be observations, opinions, perceptions, motivations, behaviors, etc.

	How can/should NLP be deployed to analyze data held in health systems (e.g., electronic health records (EHRs), health insurance data, and data from pharmacies) to provide insights about patient symptoms and disease classification?
Integration: Data linkage and aggregation of disparate datasets from multiple sources	Are there opportunities to standardize data formats – or deploy standardizing technologies – so that data from different countries and healthcare systems could be aggregated, linked, and shared across populations?
	Which disparate (and potentially conflicting) data sets (e.g., genomics, proteomics, clinical information and clinical imaging) can be linked and combined by (automated) data aggregators?
	How can/should personal health libraries be used to combine individuals’ health information across multiple different data streams to inform health outcomes?
	How can/should privacy-preserving record linkage (PPRL) be used to combine data on individuals from multiple sources and with different identifiers for precision medicine and public health?
	How can/should the research context (e.g., clinical and public health) and participants' consent status affect data linkage and aggregation?

Additional discussions of information gathered during Phase 1 led to the findings that are listed in Table 2 below.

In Phase 2, to inform the NExTRAC and WG discussions, NIH first convened a series of community conversations with members of the public to gather feedback on the research questions from Phase 1 and their potential impact on individuals and communities. The WG deliberately asked NIH to include in these engagements members of communities that have been historically marginalized in biomedical research. These engagements were designed to meet members of the public in their own communities, to open respectful dialogue in ways that resonated with participants, to establish a foundation for sustained relationships, and potentially to advance public trust in biomedical research.

Across the community conversations, three high level themes emerged. First, community members’ support for data sharing and linkage efforts depended on who received the data, whether the recipients were considered trustworthy, and whether there would be significant return of value to individuals and communities. Second, many community members expressed a desire for changes to the informed consent process so people would know more about how and when their data are used (whether initially collected for research or other purposes). Additionally, some individuals wanted further control over uses of data from or about them. Third, while many community members expressed some enthusiasm about novel uses of data across the three research topic areas, many also sought greater transparency about research data use, reassurances of safeguards, and additional benefits to research that would make participation worth the risks. Through deliberations on the feedback noted here, the NExTRAC reports the following findings (Table 2).

Phase of effort	Findings
Phase 1 WG discussions, including with subject matter experts	1. Emerging technologies to capture, store, and analyze personal health data (e.g., wearables, apps, and AI) offer rich resources for research. However, more policy and ethics work is needed to responsibly realize their full benefits.
	2. Emerging methods to harmonize and aggregate data from different studies and across the world show great promise for precision medicine and public health. Still, refinements are needed to limit variation that can reduce significance, accuracy of results, and generalizability of conclusions.
Phase 2 Public engagement with community members and scholars	3. There is misalignment between the expectations of some members of the public and some NIH policies and programs on the use and analysis of personal health data for biomedical research.
	4. It is important to consult with research participants, individuals whose data are used in secondary research, families, and communities throughout the biomedical research process when novel types, uses, and analyses of data are involved.
	5. More transparency is needed around the research process, how data will be collected, analyzed, and used, and the risks and benefits to research participants.

	6. Current regulations and procedures for use of biomedical research data provide ways for data to be reused in research without consent or adequate oversight.
	7. Linkage of data in the absence of robust informed consent raises concerns about data privacy, transparency, and respect for research participants.
	8. Beyond potential promises of precision medicine, research participants deserve greater return of value for themselves and their communities.
	9. Providing research participants with financial benefits and/or increased control over personal data may bolster community support for biomedical research conducted by for-profit entities or in public-private partnerships.

In light of these findings, the Committee reports eight high level recommendations centered on the following overarching themes: **trustworthiness and transparency in research, respect for participants and communities, improving informed consent for and control of data sharing and linkage, and providing return of value for participation in research, data sharing, and linkage** (Table 3).

Table 3: Recommendations Based on Findings	
1.	NIH should catalogue, evaluate, and, as appropriate, coordinate efforts among current initiatives addressing public engagement, ethical, and technological issues in data science.
2.	When NIH invests in public engagement, especially around emerging technologies, the agency should be prepared to update processes and infrastructure and to revise its programs and policies in accordance with public expectations and social justice principles, or clearly explain to the public why the agency may not revise its course.
3.	NIH, in consultation with other federal agencies as appropriate, should align its policies on data and research governance with widely held public values.
4.	NIH should help establish long-term relationships between members of the public and research communities to facilitate sustained public engagement on evolving issues regarding the use, analysis, management, and reuse of personal health data in biomedical research.
5.	NIH should encourage investigators to prospectively address bioethical considerations in emerging areas of data science that are particularly important to participant communities, as reflected in public engagement, in their research programs and proposals.
6.	NIH, in consultation with other federal agencies as appropriate, should develop and disseminate best practices on informing participant communities about how their personal health data are being managed, reused, and linked to other data, especially when these best practices extend beyond regulatory requirements.
7.	NIH should leverage prior research on return of results and support further public engagement to inform best practices for return of results from biomedical research using emerging data science and to determine implementation infrastructure needs.
8.	NIH should support exploration of methods for providing benefits and return of value beyond returning research results to participants and their communities, when appropriate and as consistent with current regulations.

The Committee’s approach, observations, and reasoning in reaching findings and recommendations are described in full below with contextual information in appendices.

I. Introduction

Improving our understanding of human health and disease requires access to and full utilization of data. Increasingly, researchers are looking for these data in places outside of the traditional healthcare system, such as through health apps,³ wearable technologies,⁴ social media,⁵ and other novel sources of data. Many hope that these higher-dimensional datasets will provide a more complete picture of an individual's health and well-being and spur advances in personalized medicine that can tailor treatments to an individual's needs. Novel methods in AI⁶ and ML⁷ enable investigators to wield these datasets and promise new insights for research and faster decisions in the clinic. Researchers have also developed new ways to link and aggregate personal data and make it more useful, including for addressing public health emergencies like the recent COVID-19 pandemic.⁸

While these advances may improve biomedical research, they raise difficult questions for policymakers and members of the public. How do people think about the risks and benefits of sharing their personal health information in these novel contexts? How might individuals and communities be impacted by these new uses of data, and how should data collection and policies around their use be adjusted as a result? Are there adequate privacy protections in place? How might policies need to change to keep pace with emerging science and technology?

To help address questions like these ones, NIH often turns to advisory committees like the NExTRAC. The purpose of the NExTRAC, as noted in the Committee's charter, is to,

“...provide advice to the [NIH] Director on matters related to the conduct and oversight of research involving emerging technologies in biomedical science (also referred to as emerging biotechnologies).

The Committee will address scientific, safety, ethical, and social issues associated with areas of emerging biotechnology research for which the NIH requests advice or guidance.”

Here, NIH seeks to anticipate how emerging technologies may enable the combination and use of human datasets, particularly from disparate sources (e.g., research and non-research settings), to anticipate potential benefits and risks for research participants, families, populations, and society. NIH will consider this information as the agency develops and implements research strategies and policy frameworks, informed by public input, to ensure that science advances responsibly and meets expectations for the Nation's investment in research. In July 2021, the NExTRAC was charged to:

³ Chaturvedi, R. R., Angrisani, M., Troxel, W. M., Gutsche, T., Ortega, E., Jain, M., Boch, A., & Kapteyn, C. (2023). American Life in Realtime: a benchmark registry of health data for equitable precision health. *Nature Medicine*, 29, 283–286.

<https://doi.org/10.1038/s41591-022-02171-w>; Stern, A. D., Brönneke, J., Debatin, J. F., Hagen, J., Matthies, H., Patel, S., Clay, I., Eskofier, B., Herr, A., Hoeller, K., Jaksa, A., Kramer, D. B., Kyhlstedt, M., Lofgren, K. T., Mahendraratnam, N., Muehlan, H., Reif, S., Riedemann L., & Goldsack, J. C. (2022). Advancing digital health applications: priorities for innovation in real-world evidence generation. *Lancet Digital Health*, 4, e200–e206. [https://doi.org/10.1016/S2589-7500\(21\)00292-2](https://doi.org/10.1016/S2589-7500(21)00292-2).

⁴ Abramson, A., Chan, C. T., Khan, Y., Mermin-Bunnell, A., Matsuhisa, N., Fong, R., Shad, R., Hiesinger, W., Mallick, P., Gambhir, S. S., & Bao, Z. (2023). A flexible electronic strain sensor for the real-time monitoring of tumor regression. *Science Advances*, 8, eabn6550. <https://doi.org/10.1126/sciadv.abn6550>; Mannhart, D., Lischer, M., Knecht, S., du Fay de Lavallaz, J., Strebel, I., Serban, T., Vögeli, D., Schaer, B., Osswald, S., Mueller, C., Kühne, M., Sticherling, C., & Badertscher, P. (2023). Clinical Validation of 5 direct-to-consumer wearable smart devices to detect atrial fibrillation: BASEL Wearable Study. *JACC Clinical Electrophysiology*, 9, 232–242. <https://doi.org/10.1016/j.jacep.2022.09.011>.

⁵ Klein, A. Z., Magge, A., O'Connor, K., & Gonzalez-Hernandez, G. (2022). Automatically identifying Twitter users for interventions to support dementia family caregivers: Annotated data set and benchmark classification models. *JMIR Aging*, 5, e39547.

<https://doi.org/10.2196/39547>; Nebeker, C., Dunseath, S. E., & Linares-Orozco, R. (2020). A retrospective analysis of NIH-funded digital health research using social media platforms. *Digital Health*, 6, 2055207619901085. <https://doi.org/10.1177/2055207619901085>.

⁶ U.S. Government Accountability Office and National Academy of Medicine (2022). *Artificial Intelligence in Health Care: Benefits and Challenges of Machine Learning Technologies for Medical Diagnostics*. <https://www.gao.gov/products/gao-22-104629>.

⁷ Nestor, B., Hunter, J., Kainkaryam, R., Drysdale, E., Inglis, J. B., Shapiro, A., Nagaraj, S., Ghassemi, M., Foschini, L., & Goldenberg, A. (2023). Machine learning COVID-19 detection from wearables. *Lancet Digital health*, 5, e182–e184. [https://doi.org/10.1016/S2589-7500\(23\)00045-6](https://doi.org/10.1016/S2589-7500(23)00045-6).

⁸ National Academy of Medicine (2022). *Sharing Health Data: The Why, the Will, and the Way Forward*. Washington, DC: National Academy of Sciences. <https://nam.edu/wp-content/uploads/2022/01/Sharing-Health-Data-prepub-FINAL.pdf>.

- Define and characterize the types of research questions that require increasing granularity and aggregation of data about individuals that are likely to be addressed through emerging technologies, considering:
 - Goals of such research studies and how they advance the NIH mission
 - Emerging technologies that may generate potentially sensitive datasets
 - Data types generated and their sources (e.g., digital health devices and EHR platforms) with an emphasis on exploring new data types or unique sources
 - Data science platforms and tools that facilitate data access, combination, and analysis (e.g., AI and cloud computing)
- For those questions and technologies defined above, consult with stakeholders to discuss and assess the value of and potential implications for individuals, groups, and society, considering:
 - Attitudes and perspectives about sharing participant data to advance biomedical research, specifically through the lens of balancing research risk (e.g., privacy and autonomy) with research deliverables
 - How these perspectives may evolve depending on the context of who is to benefit or assume risk, whether it be at the individual level, through the community, or broader society's expectations for public health advancement

In addressing the charge, the Committee was asked to consult with members of the public including, but not limited to, research participants, patient groups, ethicists and privacy experts, data scientists, technology developers across sectors, and public health officials.

As explained in Section II, a NExTRAC WG was established to pursue the charge in two separate phases: first, identifying types of research questions (Section III) and second, discussing those research questions with members of the public (Section IV; see workflow in Figure 1 below). In addressing the second part of the charge, the WG intentionally included voices that are often underrepresented in biomedical research and heard from members of the public in facilitated conversations in their own communities. Additionally, in support of the effort, NIH hosted a workshop to hear from scholars, developers, and physician-scientists while centering the community feedback. The approach for public engagement is explained in more detail in Section IV. The Committee drew heavily on feedback from these discussions to generate recommendations for NIH to consider (Section V). The goal of these recommendations is to facilitate the development and implementation of policies, programs, and processes that help investigators integrate public perspectives and ethical, legal, and social implications of emerging uses and analysis of data into biomedical research.⁹

⁹ Of note, while NLP and large language models were discussed, discussion about the use of generative AI in biomedical research and society remained limited. ChatGPT (and other similar artificial intelligence technologies) rose to prominence as the WG was completing its charge. The WG has used programs like ChatGPT as a test case for their recommendations to NIH as described in Appendix 2.

NExTRAC Data Science and Emerging Technology Workflow and Output

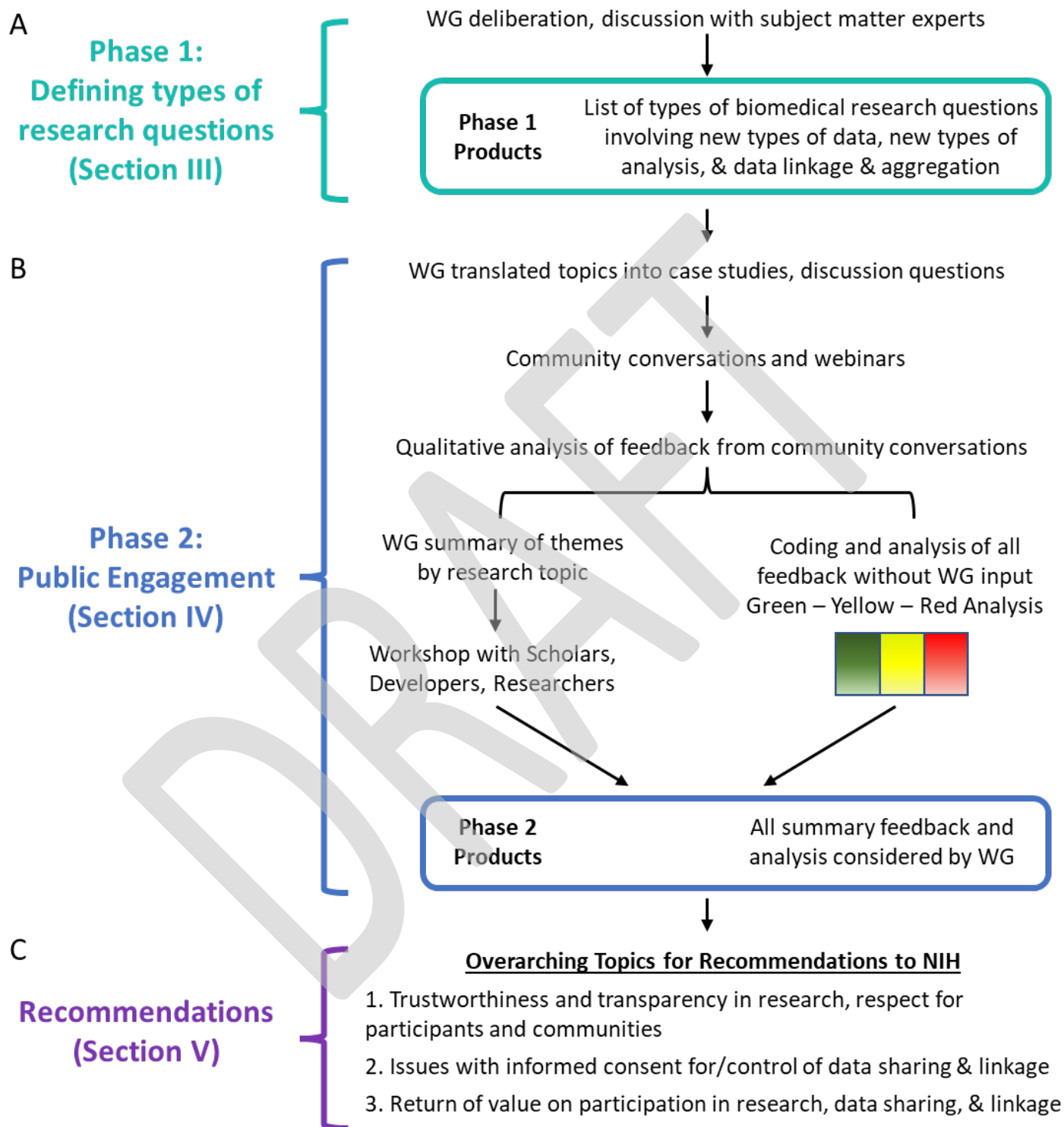


Figure 1: Workflow and Output. **A)** In Phase 1, the WG defined research topics and types of research questions based on internal deliberation and discussions with subject matter experts. **B)** In Phase 2, the WG gathered feedback from community conversations and a workshop with scholars, which was summarized by the WG and coded and analyzed without WG input. **C)** Recommendations for NIH were developed on three overarching topics.

II. Background and Overview

The framework for WG activities was provided by the NExTRAC's charge as well as a previous report of the Committee, the *Report to Establish a NExTRAC Framework*.¹⁰ The charge was divided into two parts, which served to organize the WG's activities. The WG was tasked to first identify types of research questions and second to discuss those research questions with relevant members of the public. As noted in the introduction, the WG pursued these two tasks in separate phases (Phase 1 and Phase 2; Figure 1). To guide deliberations, the WG noted several possible themes that illustrate the core issues to structure their deliberations (Table 4). Using these themes, the WG subsequently identified types of research questions (Phase 1; Section III) and developed materials to evaluate public attitudes (Phase 2; Section IV; Appendix 1).

Themes	Evaluative Question
Novelty	Have the technologies or data been around for a long time?
Context	Are the data being collected, used, or applied in new domains?
Complexity	Are the data particularly difficult to analyze, even in small amounts?
Sensitivity	Do the data reveal personal information about individuals, families, or groups?
Justice	Do the data have the potential to produce an unjust distribution of risks and benefits?
Access	Is there lack of access to the data, the underlying technologies, or downstream benefits from using the data?
Reliability	Are the data predictive and useful for addressing the relevant research questions? Can the data be relied upon to make decisions?
Unintended consequences	Do the data have significant potential to be used in ways different from originally intended and to cause harm from those uses?
Control	Are individuals able to exert control over downstream uses of the data?
Aggregation	Are the data combined with or easily linked to multiple other types of datasets?
Identifiability	Do the data contain identifiable information, either on their own or when combined with other readily available information?
Privacy, Confidentiality	Are the data vulnerable to unauthorized access, disclosure, and revealing of personal information?

The WG focused on three topic areas to inform health-related questions:

- **Data:** Use of novel data from outside of the traditional healthcare system
- **Algorithms:** Use of models and algorithms
- **Integration:** Data linkage and aggregation of disparate datasets from multiple sources

This framework of Data, Algorithms, and Integration was meant to encompass the entire data lifecycle, from data collection to storage, analysis, sharing, and downstream uses. The first component, novel data, addresses new sources of information being used in biomedical and behavioral research. The second component, models and algorithms, asks about novel ways that both new and historical biomedical research data are being analyzed, and in some cases, incorporated in healthcare. The third component, integration, covers the new ways that datasets are being combined, linked, and aggregated to lead to new discoveries. These topic areas are discussed in more detail in Section III.

The WG also sought to avoid duplicating previous and current efforts across the United States government. For example, research questions focused specifically on genomic data were deemphasized, given that there are many other ongoing efforts in this space. Additionally, the WG noted that there were several initiatives focused on AI at NIH and across the United States government and wanted to ensure that any recommendations on AI did not simply repeat the conclusions of previous reports. Finally, the WG placed less emphasis on issues raised by international data sharing.

¹⁰ Novel and Exceptional Technology and Research Advisory Committee (2020, December). *Report to Establish a NExTRAC Framework*. https://osp.od.nih.gov/wp-content/uploads/NExTRAC-Framework-Report_FINAL_508.pdf.

Given the strong role played by United States regulations (e.g., HIPAA and the Common Rule) in determining how data are used in NIH-funded research, the WG instead considered data sharing in a domestic context.

A. Approach for Phase 1

As outlined in the Report to Establish a NExTRAC Framework, the Committee utilizes horizon scanning to identify emerging technologies that will be implemented in the next 5-10 years. As a starting point, the WG held brainstorming meetings to share perspectives on the most pressing issues in data science in biomedical research and to identify gaps in WG knowledge. To bridge these gaps, the WG performed a literature review and identified subject matter experts in topics related to the NExTRAC's charge. Over the course of Phase 1, the WG considered the following topics:¹¹

- Use of social media and other user-generated data for medical diagnosis and public health surveillance
- Use of AI and ML in medical decision-making
- Health data platforms and federated data sharing¹²
- Consumer and lifestyle data from non-health-specific sources
- Novel data streams made possible by wearables, smartphones, and the internet

In July 2022, the WG presented a draft list of types of research questions to the NExTRAC, informed by their analysis as articulated in Table 1 (Executive Summary) and bulleted summary (Section III). As noted above, the WG prioritized a subset of these research questions and translated them into case studies for public engagement in Phase 2.

B. Approach for Phase 2

In Phase 2 of the effort, the WG solicited broad feedback from diverse subsets of the public and engaged communities to build trust and sustainable partnerships. To facilitate WG efforts, NIH supported conversations with people in their communities across the United States and solicited feedback on a set of topical and relatable case studies (Appendix 1; more detailed materials can be found on the NIH Office of Science Policy website¹³ and in Section IV).

Six "community conversations" were held in-person and each lasted four hours. They were relatively small (~25 community members), with breakout sessions to facilitate discussion. Facilitators led community members through the case studies, which were presented in multiple formats, through text, illustrations, and clarifying discussion. The WG garnered feedback from communities that have been historically marginalized in biomedical research. For example, one event (Bronx, NY) was held entirely in Spanish. Community members were also compensated for their time.

At the WG's behest, four additional events were held to include perspectives not captured in the community conversations noted above. Three webinars included advocates from rural health organizations, rare disease networks, and American Indian and Alaska Native groups and followed an analogous agenda to previous community conversations (Appendix 1). Here, participants were professional advocates for their communities and thus were representing a broader perspective than the individuals at the other community conversations. A separate NIH-hosted workshop was held in San Diego, CA with researchers, technology developers, and bioethicists. The goal of this event was to gather feedback on how the results of the community conversations could inform the development of NIH programs and policies. Summaries of all events can be found in Section IV.

¹¹ As noted in the introduction, most of the discussions about relevant research questions were completed before ChatGPT was made publicly available. The Committee would like to reinforce that NIH should consider the implications of this type of generative AI through the lens of the findings and recommendations in this report (as outlined in Appendix 2).

¹² Federated data sharing refers to the practice of using distributed data sources across different locations to store and control data, in contrast to centralized data sharing where data from multiple sources are pooled and controlled at a single source. Thorogood, A., Rehm, H. L., Goodhand, P., Page, A.J.H., Joly, Y., Boudis, M., Rambla, J., Navarro, A., Nyronen, T.H., Linden, M., Dove, E.S., Fiume, M., Brudno, M., Cline, M.S., & Birney, E. (2021). International federation of genomic medicine databases using GA4GH standards. *Cell Genomics*, 1, 2. <https://doi.org/10.1016/j.xgen.2021.100032>.

¹³ National Institutes of Health, Office of Science Policy (2023, February). *Using Public Engagement to Inform the Use of Data in Biomedical Research, Workshop at University of California, San Diego February 27, 2023*. https://osp.od.nih.gov/wp-content/uploads/2023/02/Tab3_NIH_UCSD_Workshop_Community_Conversation_Summary_230227_508.pdf.

III. Research Questions on Data Science

A. Introduction and Background

Data science continues to accelerate the potential for discoveries and breakthroughs in biomedical research. However, these exciting opportunities are tempered by risk to individuals, families, communities, and the public, raising critical questions about how people balance potential risks and benefits. Specific types of research questions and themes that were considered by the Committee include:

- **Increasing volumes of Person Generated Health Data (PGHD)¹⁴ will drive advances in medical science, but clear potential pitfalls remain:** Wearable sensors and devices for health monitoring are being used at increasing rates. Individuals can capture and integrate their personal health information from myriad sources in apps, which may empower them to manage this information themselves. In addition, biomedical investigators would like to incorporate these data in clinical research to generate a more holistic picture of individuals and communities. However, ensuring participant privacy and data security remains a challenge and the cost of these emerging technologies raises concerns for health equity.
- **Cautious enthusiasm for the use of algorithms for interpreting social media data for real-world applications:** Social media data are in the public domain and there has been much interest in harnessing these data for biomedical research and public health. For example, some evidence indicates that NLP and ML can predict mental health crises for individuals and communities.^{15,16} However, people are often surprised by what conclusions can be drawn from these data and do not understand how they may be used. To prevent stigmatization of communities and erosion of trust, more work is needed to define people’s expectations and when and how to get permission for use of their data in biomedical research.
- **Keeping humans in the loop for analysis of biomedical research data through AI:** AI models and algorithms incorporate and amplify biases inherent to the datasets that are used in their development (i.e., in “training” the models and algorithms) and they often identify patterns in data that are not evident to human observers.^{17,18} While these algorithms may augment human experts’ (e.g., clinicians and computer scientists) interpretation of data, they may also reach conclusions that have negative consequences for certain groups. Investigators would benefit from guidelines to better vet, regulate, and interpret the output of a specific technology. Human experts

¹⁴ As noted in footnote 1, person-generated health data are data collected by an individual through personal devices designed to collect health-related metrics (e.g., wearables and apps). These may be observations, opinions, perceptions, motivations, behaviors, etc.

¹⁵ Birnbaum, M. L., Ernala, S. K., Rizvi, A. F., Arenare, E., R Van Meter, A., De Choudhury, M., & Kane, J. M. (2019). Detecting relapse in youth with psychotic disorders utilizing patient-generated and patient-contributed digital data from Facebook. *NPJ Schizophrenia*, 5, 17. <https://doi.org/10.1038/s41537-019-0085-9>.

¹⁶ Choi, D., Sumner, S. A., Holland, K. M., Draper, J., Murphy, S., Bowen, D. A., Zwald, M., Wang, J., Law, R., Taylor, J., Konjeti, C., & De Choudhury, M. (2020). Development of a machine learning model using multiple, heterogeneous data sources to estimate weekly US suicide fatalities. *JAMA Network Open*, 3, e2030932. <https://doi.org/10.1001/jamanetworkopen.2020.30932>.

¹⁷ Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H., & Ferrante, E. (2020). Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *PNAS*, 117, 12592–12594. <https://doi.org/10.1073/pnas.1919012117>.

¹⁸ Wiens, J., Price, W.N., & Sjoding, M.W. (2020). Diagnosing bias in data-driven algorithms for healthcare. *Nature Medicine*, 26, 25-26.

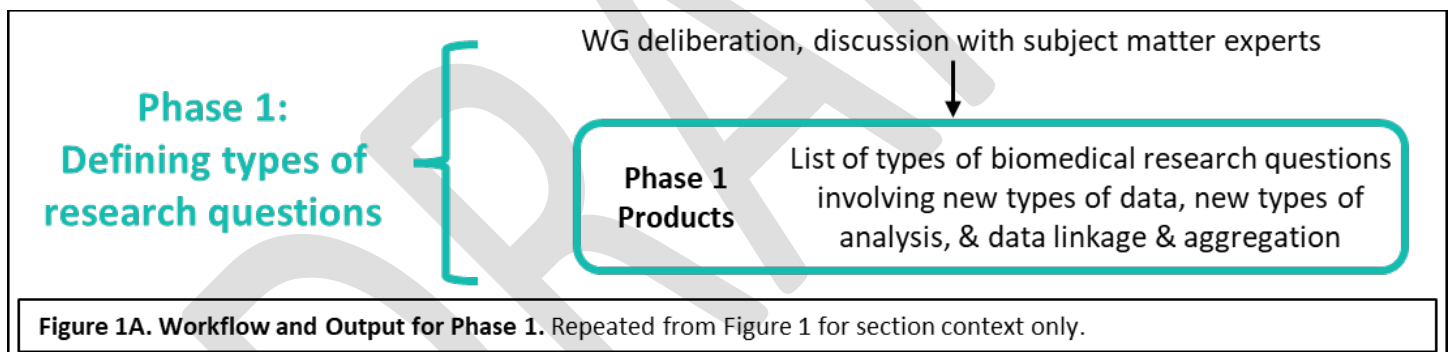
could then evaluate technology as a standardized process and identify limitations to inform the next phase of research and innovation.^{19,20,21}

- **A need for standardized health metrics, data elements, and models:** Increased data sharing may accelerate progress in biomedical research by increasing statistical power and maximizing utility of valuable datasets. To ensure interoperability, data and metadata must be standardized across research studies and jurisdictions. A variety of approaches are available to standardize data for comparison, linkage, and aggregation.^{22,23} A related challenge is that investigators often rely on traditional health metrics that may not reflect the complexity of physiology and wellness (e.g., blood pressure). As new types of personal data are collected outside of the traditional healthcare system, developers could collaborate with clinicians and researchers to ensure that data are easy to collect, are reliably captured by affordable technology, and are predictive of specific health outcomes.

Based on deliberations of these considerations and related types of research questions, the Committee prioritized three major research topics for further exploration (Table 1; Figure 1A):

- **New types of data:** Use of novel data from outside of the traditional healthcare system
- **New types of analysis:** Use of models and algorithms
- **Linkage and aggregation:** Data linkage and aggregation of disparate datasets from multiple sources

Though the Committee addressed the entire data lifecycle, this framework was not comprehensive of all potential issues raised by data science. There is also overlap in the types of research questions across topics (e.g., some novel algorithms are based on novel sources of data). Each type of research question is discussed in more detail below.



¹⁹ Braatz, J., Rajpurkar, P., Zhang, S., Ng, A. Y., & Shen, J. (2022). Deep learning-based sparse whole-slide image analysis for the diagnosis of gastric intestinal metaplasia. <https://arxiv.org/abs/2201.01449>.

²⁰ Holzinger A. (2016). Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain informatics*, 3, 119–131. <https://doi.org/10.1007/s40708-016-0042-6>; Budd, S., Robinson, E. C., & Kainz, B. (2021). A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical Image Analysis*, 71, 102062. <https://doi.org/10.1016/j.media.2021.102062>.

²¹ Bodén, A. C. S., Molin, J., Garvin, S., West, R. A., Lundström, C., & Treanor, D. (2021). The human-in-the-loop: an evaluation of pathologists' interaction with artificial intelligence in clinical practice. *Histopathology*, 79, 210–218. <https://doi.org/10.1111/his.14356>.

²² Taber, P., Radloff, C., Del Fiore, G., Staes, C., & Kawamoto, K. (2021). New standards for clinical decision support: A survey of the state of implementation. *Yearbook of Medical Informatics*, 30, 159–171. <https://doi.org/10.1055/s-0041-1726502>.

²³ NeuroLINCS. <http://neurolincs.org/publications/>.

B. Types of Research Questions to Consider Through Public Engagement

Use of Novel Data from Outside of the Traditional Healthcare System

The Committee looked beyond traditional sources of health data, such as those from hospitals, patient health records, and information from primary research studies. The examples below include data from social media, smartphones, wearables, health apps, consumer data, and a variety of sources on SDOH. Data from these sources are unique because they are typically viewed as data that *could* be used to infer health traits as opposed to primary indicators of one's health status. However, in some cases, data from consumer devices may be just as informative as data obtained from medical devices. Additionally, investigators increasingly use these data in biomedical research and patient care, blurring the traditional definitions of health data. Using “non-health” data to fill in the gaps and in some cases to replace traditional health data is a particularly important trend to consider for biomedical research. General types and specific examples of research questions on this topic include:

1. How are personal health data collected from outside of the traditional health system or research setting (e.g., through fitness trackers, apps, social media posts) being used to study health-related questions and predict health risks at an individual, family, group, or public health level?

Example research questions pertaining to social media data (often publicly available):

- Can patient-generated digital data from Facebook predict or detect relapse in psychiatric disorders?¹⁵
- Can real-time streams of secondary information related to suicide (e.g., tone of language in posts and affiliation with a user groups) be used to accurately estimate suicide fatalities in the United States in real time?¹⁶

Example research questions pertaining to wearables (often data are presumed private, but 3rd party storage of raw and/or processed data may indicate otherwise):

- Can wearable microphones paired with accelerometers provide reliable long-term cardiopulmonary monitoring of patients?²⁴
 - Can chest-worn inertial sensors accurately detect a broad range of physiological signals (e.g., cardiac and respiratory parameters) and behaviors (e.g., posture, sleep, falling, swallowing) for real time tracking?²⁵
 - Can interstitial glucose levels for clinical decision making be measured accurately and precisely through wearable technology?²⁶
2. How can other consumer and lifestyle data from non-health-specific sources (e.g., sensors in the home and credit card and consumer rewards data) be used to study health-related questions and predict health risks?

Example research questions:

- Can credit scores define cardiovascular disease risk?²⁷
- Can personal AI assistants (e.g., Alexa) reliably detect health conditions through changes in speech patterns?²⁸

²⁴ Gupta, P., Moghimi, M. J., Jeong, Y., Gupta, D., Inan, O. T., & Ayazi, F. (2020). Precision wearable accelerometer contact microphones for longitudinal monitoring of mechano-acoustic cardiopulmonary signals. *NPJ Digital Medicine*, 3, 19. <https://doi.org/10.1038/s41746-020-0225-7>.

²⁵ Rahmani, M. H., Berkvens, R., & Weyn, M. (2021). Chest-worn inertial sensors: A survey of applications and methods. *Sensors*, 21, 2875. <https://doi.org/10.3390/s21082875>.

²⁶ Cengiz, E., & Tamborlane, W. V. (2009). A tale of two compartments: interstitial versus blood glucose monitoring. *Diabetes technology & Therapeutics*, 11, S11–S16. <https://doi.org/10.1089/dia.2009.0002>.

²⁷ Israel, S., Caspi, A., Belsky, D. W., Harrington, H., Hogan, S., Houts, R., Ramrakha, S., Sanders, S., Poulton, R., & Moffitt, T. E. (2014). Credit scores, cardiovascular disease risk, and human capital. *PNAS*, 111, 17087–17092. <https://doi.org/10.1073/pnas.1409794111>.

²⁸ Anthes E. (2020). Alexa, do I have COVID-19? *Nature*, 586, 22–25. <https://doi.org/10.1038/d41586-020-02732-4>.

- Can in-home sensors be leveraged to create safer environments for people suffering from cognitive decline?²⁹
3. How can integration of health data with data on SDOH enable better risk stratification of patient populations and development of predictive algorithms?³⁰ SDOH can include socio-economic status, housing status, education status, geographical environments in which people spend time (e.g., crime rates or environmental pollutants in a given neighborhood), and identity factors that advantage or disadvantage health status.

Example research questions:

- Can screening of the status of SDOH in EHRs enable tailored referrals to available community-based agencies?³¹
- Can integration of SDOH data into EHRs facilitate clinical risk prediction and intervention?³²

Use of Models and Algorithms

Recent advances in analysis, especially computational techniques, are pushing the frontiers of biomedical research. For example, new methods in AI, ML, and NLP can identify patterns in an individual’s EHRs that would be challenging for a clinician to uncover. But many questions and concerns remain regarding how those models and algorithms are “trained” and how they are used in the real world. These considerations will undoubtedly remain important and evolve for behavioral and biomedical researchers over the next 5-10 years. General types and specific examples of research questions on this topic include:

1. What is the role of computer-based algorithmic technologies, such as AI, NLP (including large language models like ChatGPT), ML, and automated image analysis in biomedical research and in health decision-making?

Example research questions:

- Can deep learning algorithms (i.e., algorithms that update themselves based on new information) be deployed for effective automated medical image analysis to replace clinical expertise for diagnosis of disease? (e.g., pneumonia³³; stomach or intestinal cancer¹⁹)
 - Does bias in algorithm development impact populations differentially by race, gender, and/or social/cultural identity?¹⁷

²⁹ Vahia, I. V., Kabelac, Z., Hsu, C. Y., Forester, B. P., Monette, P., May, R., Hobbs, K., Munir, U., Hoti, K., & Katabi, D. (2020). Radio Signal Sensing and Signal Processing to Monitor Behavioral Symptoms in Dementia: A Case Study. *The American Journal of Geriatric Psychiatry*, 28, 820–825. <https://doi.org/10.1016/j.jagp.2020.02.012>.

³⁰ Cantor, M. N. & Thorpe, L. (2018). Integrating data on social determinants of health into electronic health records. *Health Affairs*, 37, 585–590. <https://doi.org/10.1377/hlthaff.2017.1252>.

³¹ Gottlieb, L. M., Tirozzi, K. J., Manchanda, R., Burns, A. R., & Sandel, M. T. (2015). Moving electronic medical records upstream: incorporating social determinants of health. *American Journal of Preventive Medicine*, 48, 215–218. <https://doi.org/10.1016/j.amepre.2014.07.009>.

³² Chen, M., Tan, X., & Padman, R. (2020). Social determinants of health in electronic health records and their impact on analysis and risk prediction: A systematic review. *Journal of the American Medical Informatics Association*, 27, 1764–1773. <https://doi.org/10.1093/jamia/ocaa143>.

³³ Irvin, J. A., Pareek, A., Long, J., Rajpurkar, P., Eng, D. K., Khandwala, N., Haug, P. J., Jephson, A., Conner, K. E., Gordon, B. H., Rodriguez, F., Ng, A. Y., Lungren, M. P., & Dean, N. C. (2022). CheXED: Comparison of a deep learning model to a clinical decision support system for pneumonia in the emergency department. *Journal of Thoracic Imaging*, 37, 162–167. <https://doi.org/10.1097/RTI.0000000000000622>.

- When and how should human experts be inserted to monitor and analyze the output of health-related algorithms and models to ensure fidelity? (see overview²⁰; e.g., clinician expert²¹, bioinformatics supervision³⁴)
 - Can algorithms determine prognoses for patients on admission to the hospital?³⁵
 - How can/should the underlying computational strategies used by AI/ML algorithms be tracked and corrected for ensuring accuracy in clinical diagnoses?¹⁸
 - What types of bias are involved in algorithms taking “shortcuts” to reach conclusions?¹⁹ (While these algorithms can find solutions and provide diagnoses based on the data that they have been given, it is often not clear how they came to those conclusions. Algorithms may be responding to spurious correlations or taking “shortcuts” to identify patterns in data.)
2. How can/should NLP be deployed to analyze data held in health systems (e.g., EHRs, health insurance data, data from pharmacies) to provide insights about patient symptoms and disease classification?

Example research questions:

- What symptom information can be found in unstructured narratives in EHRs?³⁶
 - How might large-scale language models exacerbate existing biases in medical research and clinical care?³⁷
 - How do the identities and cultural backgrounds of doctors and patients impact the information derived using NLP on unstructured data?

Linkage and Aggregation of Disparate Datasets from Multiple Sources

Many investigators would like to facilitate reuse of data by linking data from different sources. Until recently, combining different datasets introduced significant privacy concerns or significantly diminished their utility. Two researchers seeking to share data needed data in the same format and needed to ensure that sharing data would not accidentally reveal sensitive information about participants or violate informed consent. With new privacy tools and software (e.g., PPRL), researchers can link datasets while meeting privacy protection standards. Also, large-scale, systematic efforts at standardizing data are being applied to research data. While some privacy and interoperability concerns remain, these new technologies will enable new discovery across datasets over the next 5-10 years. General types and specific examples of research questions on this topic include:

³⁴ Goodwin, T. R. & Demner-Fushman, D. (2020). A customizable deep learning model for nosocomial risk prediction from critical care notes with indirect supervision. *Journal of the American Medical Informatics Association*, 27, 567–576.

<https://doi.org/10.1093/jamia/ocaa004>.

³⁵ Kamran, F., Tang, S., Otlis, E., McEvoy, D. S., Saleh, S. N., Gong, J., Li, B. Y., Dutta, S., Liu, X., Medford, R. J., Valley, T. S., West, L. R., Singh, K., Blumberg, S., Donnelly, J. P., Shenoy, E. S., Ayanian, J. Z., Nallamothu, B. K., Sjoding, M. W., & Wiens, J. (2022). Early identification of patients admitted to hospital for covid-19 at risk of clinical deterioration: model development and multisite external validation study. *BMJ*, 376, e068576. <https://doi.org/10.1136/bmj-2021-068576>.

³⁶ Koleck, T. A., Dreisbach, C., Bourne, P. E., & Bakken, S. (2019). Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association*, 26, 364–379.

<https://doi.org/10.1093/jamia/ocy173>.

³⁷ Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P.S., & et al. (2021). Ethical and social risks of harm from language models. <https://arxiv.org/abs/2112.04359>.

1. Are there opportunities to standardize data formats – or deploy standardizing technologies – so that data from different countries and healthcare systems could be aggregated, linked, and shared across populations?

Example research questions:

- Can large-scale observational research (enabled by standardizing data formats across healthcare systems and countries) outperform established guidelines and expert opinion?³⁸
- To what extent can the interoperability standards currently in use be applied for clinical decision support?²²

2. Which disparate (and potentially conflicting) data sets (e.g., genomics, proteomics, clinical information, clinical imaging) can be linked and combined by (automated) data aggregators?

Example research questions:

- Can multi-modal cancer data (e.g., EHR data, genomic data, health imaging data) be meaningfully pooled from numerous sources to improve its usefulness to the broader cancer research community?²³

3. How can/should personal health libraries be used to combine individuals’ health information across multiple different data streams to inform health outcomes?

Example research questions:

- How can all the behavioral data about a person’s health and lifestyle be combined and managed by the individual to provide an accurate picture of their health (i.e., a “digital twin”)? Can such digital twins be used to improve healthcare service for individuals?³⁹

4. How can/should PPRL be used to combine data on individuals from multiple sources and with different identifiers for precision medicine and public health?

Example research questions:

- Can PPRL identify new research opportunities for previous human research participants without sacrificing privacy?
- Can PPRL safely combine public health data and clinical data without putting individuals at risk?

5. How can/should the research context (e.g., clinical, public health) and participants' consent status affect data linkage and aggregation?

³⁸ Observational Health Data Sciences and Informatics. <https://www.ohdsi.org/>; Schuemie, M. J., Ryan, P. B., Pratt, N., Chen, R., You, S. C., Krumholz, H. M., Madigan, D., Hripcsak, G., & Suchard, M. A. (2020). Principles of Large-scale Evidence Generation and Evaluation across a Network of Databases (LEGEND). *Journal of the American Medical Informatics Association*, 27, 1331–1337. <https://doi.org/10.1093/jamia/ocaa103>.

³⁹ Liu, Y., Zhang, L., Yang, Y., Zhou, L., Ren, L., Wang, F., Liu, R., Pang, Z., & Deen, M.J. (2019). A novel cloud-based framework for the elderly healthcare services using digital twin. *IEEE Access*, 7, 49088-49101.

IV. Public Engagement to Inform Implications of Research Questions and Technologies

A. Collection of Feedback from Public Engagement and Methods of Analysis

Introduction and Background

Public investment in biomedical research has led to remarkable advances in the understanding of fundamental biology that have transformed human health and quality of life. As the largest public funder of biomedical research in the world, NIH has a responsibility to ensure, as much as possible, that the work it supports will benefit everyone and incorporates public concerns and perspectives. Historically, NIH has relied heavily on requests for information, meetings, and workshops with investigators, medical experts, and scholars within the biomedical research ecosystem to solicit public opinion. Noting the charge, the Committee sought to cast a wider net and, to inform these efforts, NIH held conversations with individuals from diverse backgrounds in communities across the United States. Additionally, NIH hosted a workshop with scholars from relevant backgrounds (e.g., ethicists and privacy experts, data scientists, technology developers across sectors, clinician-scientists, and patient advocates) to reflect on the community feedback and contextualize their professional perspectives within these broader discussions (Figure 1B; Subsection IV.C).

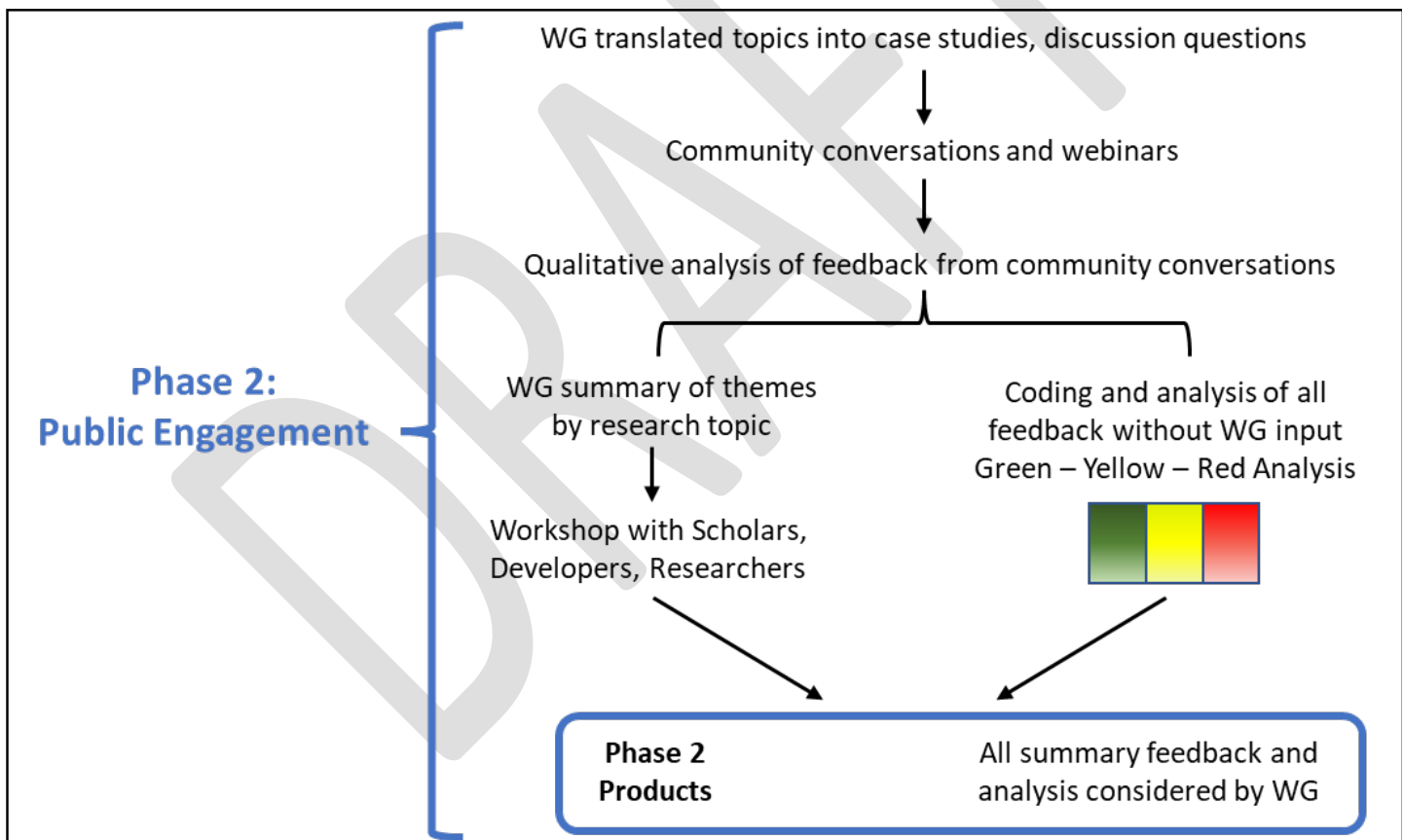


Figure 1B. Workflow and Output for Phase 2. Repeated from Figure 1 for section context only.

NIH partnered with Pyxis Partners, a healthcare advocacy firm with expert facilitators and established relationships in communities across the United States, to recruit community members and develop materials for discussion. Six, four-hour in person conversations were held with ~25 community members each in different geographical locations. In addition, three virtual webinars gathered feedback from rural health advocates, American Indian and Alaska Native populations, and rare disease advocates (Table 6). These conversations focused on case studies and discussion questions

around the types of research questions defined in Phase 1, all developed to maximize accessibility to community members. To cover different elements of the research topics, two different case studies were developed for Topic Area #1 and #2 (Table 6).

The following sections include a summary of facilitated discussions, analyzed in two ways. First, the WG identified key themes by research topic, informed by notes and their experience attending these conversations. Second, Kendra Nervik, from the University of Wisconsin, Madison, coded the science writers' summaries in NVivo and identified factors that impact community members' support for types of research. Ms. Nervik did not attend the community conversations and therefore offered a different perspective than WG members (Table 5; Tables 7,8). Factors identified by Ms. Nervik's analysis are color coded as follows:

- Green denotes factors that increased community member support for the type of research. Overall, the analysis indicates that maximizing these concepts will benefit trustworthiness and public perception of the research.
- Yellow denotes factors that caused community members to pause and seek further information before determining whether to support the type of research. These are opportunities to build trust-based relationships with potential research participants by anticipating their questions and concerns and building resources for ongoing engagement.
- Red denotes factors which strongly concerned community members and detracted from their support of the type of research. Overall, limiting these factors would be more aligned with community members' expressed interests and values. They also present an opportunity for funders and investigators to communicate how these concerns have been considered and addressed or mitigated.

Within color-coded categories, greater color saturation correlates to more observations of a theme in the feedback from the community conversations. These color-coded categories overlap to some extent across research topics. There is also strong concordance between themes identified by the WG and Ms. Nervik's more granular color-coded analysis. Finally, potentially identifying information (e.g., quotes from community members) were removed to protect privacy.

B. Community Conversations Across the United States

Introduction and Background

Across community conversations, webinars, research topics, and case studies, three high level themes emerged:

- Community members' support for data sharing efforts depended on who received the data, whether the recipients were considered trustworthy, and whether there would be significant return of value to individuals and communities.
- Several community members expressed a desire for changes to the informed consent process so people would know more about how their data are used (whether collected first for research or for other purposes) and wanted increased control over uses of data from or about them.
- While community members expressed some enthusiasm about novel uses of data across the three topic areas, they generally wanted greater transparency about research data use, reassurances of safeguards, and additional reasons to believe that the benefits would be worth the potential risks.

These general themes are reflected below in a more detailed description of community conversation feedback on topic areas and types of research questions.

Research Topic Area #1: Biomedical research and technology involving novel data types used outside of the traditional healthcare system (e.g., wearable technology, social media), as typified by the following research questions:

- How are personal health data collected from outside of the traditional health system or research setting (e.g., through fitness trackers, apps, social media posts) being used to study health-related questions and predict health risks, at an individual, family, group, or public health level?

- How can/should other consumer and lifestyle data from non-health-specific sources (e.g., sensors in the home and credit card and consumer rewards data) be used to study health-related questions and predict health risks?
- How can/should health data be integrated with data on SDOH to enable better risk prediction and development of predictive algorithms?

Summary of Community Conversation Feedback on Research Topic Area #1

Community members' support for sharing and use of these novel data in biomedical research was dependent on who was receiving the data and whether they assessed that person or entity to be trustworthy. Many noted that they wanted to receive any health-related results from a trusted clinician with whom they have a long-term relationship. Beyond that, feedback suggested that there is some trust in the healthcare system, some trust in government research and authorities (varied across individuals), and generally very limited trust in industry. While community members expressed some enthusiasm about novel uses of data across the three topic areas, they generally wanted greater transparency in what these uses would entail and reassurances that the benefits would in fact be worth the potential risks. There was a strong emphasis on the importance of respect in approaching communities to build trustworthiness through transparency in use of personal health data. The main themes of conversation around this research topic area included:

- **Trust:** Community members placed high importance on whether individuals or organizations receiving and using data could be trusted. Companies or researchers making money from sharing data were sometimes viewed as untrustworthy.
- **Transparency:** In general, community members wanted greater transparency about where data are going, how data are being used, and the goals of research using their data.
- **Privacy:** Community members wanted greater reassurances that their privacy would be protected, especially for data used for secondary purposes that could potentially stigmatize or otherwise harm individuals and/or their communities. For example, community members identified the possibility that data could be shared from wearables or use of in-home sensors in ways that could affect insurance, reproductive health, immigration status, or generally could cause stigma for certain groups.
- **Control over downstream uses:** Community members generally wanted more control over future uses of their data, especially when data would be identifiable or unprotected. For example, if a study utilized in-home sensors, there was a perception that data from those not participating in the study could be collected and used for other purposes with few options for limiting uses.
- **Return of value to individuals and communities:** Community members wanted more of a return in terms of health benefits to community members and communities. For example, there was strong interest in sharing research information (including results) with health care providers for individuals. Some community members were also skeptical that social media data would be sufficiently accurate that it could translate to health benefits for individuals.
- **Challenges with informed consent:** There was slight concern with researchers obtaining personal health data from public and/or commercial sources (e.g., social media data) without asking individuals for consent to do so, as well as the practice of obtaining data when individuals perceive they have very little choice over whether to consent to data sharing (e.g., rewards cards and clickwrap⁴⁰ on needed services/apps). Some community members seemed aware that these data are often owned by developers or were in the public domain and preferred to consent to or at least receive notification of use of this data in research.

⁴⁰ Clickwrap is a digital prompt that facilitates consent processes by affording users the opportunity to quickly accept or reject digital media policies. Observational Health Data Sciences and Informatics. Obar, J. A. & Oeldorf-Hirsch, A. (2018). The Clickwrap: A political economic mechanism for manufacturing consent on social media. *Social Media + Society*, 4. <https://doi.org/10.1177/2056305118784770>.

Analysis of Community Conversation Feedback on Case Studies 1 and 2

For this topic, qualitative analysis revealed several themes that were present across the community conversations (Table 5). When discussing both case studies, community members were most likely to support research (green factors) if they felt like it would benefit them and their community, the boundaries of the research were clear, there were appropriate safeguards in place for their privacy, there were opportunities to control data use, and if this information was conveyed by someone they considered trustworthy (particularly during the informed consent process). On the other hand, if risks (e.g., with respect to participant privacy or security of data) were perceived as unclear or outweighed the benefits, they would seek more information before deciding whether to support the research (yellow factors). Many also noted they would be more concerned when using data from individuals in protected classes (e.g., minors), more easily identified individuals (e.g., tribal communities in rural areas), and those who identified with groups who have been historically harmed or marginalized in biomedical research. Additionally, while community members recognized the utility of new types of data in biomedical research when they clearly related to health metrics (e.g., step counters, glucose monitors, digital health devices), many felt that the relationship of consumer data to health was tenuous, that it was unlikely that anything relevant could be extracted from social media data, and that there was reason to be worried about loss of security and stigma from the use of such data. Finally, while some community members recognized the need for private industry partnership to drive innovation in research, many were opposed to companies making money from their data (red factors). Community members were concerned about the scope of data collected and potential future uses, as well as the possibility of their data being sold to third party entities (e.g., insurance companies), leading to negative repercussions for them, their family, or their community. For more information on feedback by case study, please see Appendix 1.

Case Study	Green: increased support	Yellow: uncertainty or mixed support	Red: diminished support
Case Study 1: Wearables, Apps, Social Media & Heart Health	Health relevance of wearable data	Transparency of study procedures	Monetization of data
	Familiarity with wearables	Privacy and anonymity concerns	Sharing social media data
	Procedures for ensuring privacy and anonymity	Sharing "non-health" or "personal" information	Risks from downstream usage
	Clear explanations in informed consent	Sharing social media data	
	Trusting relationship with researchers	Data aggregators	
	Clear communication of benefits of participation	Including Tribal communities in research using novel data types	
	Including minors in research using novel data types		
Case Study 2: In-Home Sensors for Older Individuals	Procedures for assuring security and confidentiality	Privacy and security	Accuracy, effectiveness, and reliability of the data
	Transparency around data collection and use	Sharing data for research or commercial purposes	Scope of data collected
	Participant control over data collection	Risk/benefit ratio	Collecting data on visitors
	Compensation to participants	Benefits to individuals and communities	Monetization of data
	Clear informed consent		
	Health benefits for older individuals		
	Assurance of benefits		

Research Topic Area #2: Biomedical research and technology involving new types of analyses (e.g., AI, ML, NLP), as typified by the following research questions:

- What is the role of computer-based algorithmic technologies, such as AI, NLP (including large language models like ChatGPT), ML, and automated image analysis in biomedical research and in health decision-making?

- How can/should NLP be deployed to analyze data held in health systems (e.g., EHRs, health insurance data, data from pharmacies) to provide insights about patient symptoms and disease classification?

Summary of Community Conversation Feedback on Research Topic Area #2

Community members’ support for the use of algorithms and models to analyze personal health data was dependent on the reliability of algorithms over time (e.g., if they accurately predicted what they claimed to predict and could be used to make healthcare decisions). Many noted that if algorithms could demonstrate reliability, they were excited about the idea of sharing their data to receive more accurate diagnoses, better care, and to address public health issues. While community members largely supported the use of AI in clinical contexts, many trusted the decisions of human clinicians more, based on longstanding relationships and existing trust (echoing their preference in hearing from these trusted advisors on the implications of collection and use of novel types of personal data for biomedical research in research Topic #1 above). Many community members were aware of the possibility of bias and misuse of algorithms, particularly that their communities needed to be represented in the data used to develop the algorithm, and that bias of developers could impact these models and lead to stigmatization. Accordingly, they wanted to know more about the possibility of stigma and how else the data and algorithms might be used. Main themes of conversation around this research topic area include:

- **Trust:** While open to the idea of trusting AI in clinical contexts, many community members ultimately wanted to hear from their own trusted health care provider to provide context and explain how AI informed decision-making.
- **Transparency:** When receiving diagnoses from an algorithm, community members said that they would want greater transparency around how the algorithm was developed, who developed it, how it comes to conclusions, and which populations were studied during development. Many community members wanted to hear from both researchers and health care providers to understand more.
- **Accuracy and reliability:** Community members wondered how inaccurate conclusions made by AI might be corrected, whether those conclusions could be harmful, and how those conclusions and any associated data would be used in the future. For example, many community members were enthusiastic about *the potential* for using algorithms to analyze social media data to help address mental health issues but were skeptical that this would translate to reliable interventions.
- **Return of value to individuals and communities:** Many community members strongly wanted the research to be translated into benefits for the people involved in the research and their communities. They also wanted greater involvement from communities (including children as appropriate) in deciding how data would be used. There was some concern that this sort of research could cause stigma to individuals and communities.
- **Privacy risks require significant health benefits:** Many community members perceived a potential loss of privacy in the use of AI but expressed that the tradeoff would be worth it if the health benefits were significant. For example, receiving an early cancer diagnosis may be worth the potential privacy risk.
- **Control over downstream uses:** Community members generally wanted more control over data collected and analyzed by AI in research, especially when they failed to see a connection of those data to the main topic of a study (e.g., social media data). Some community members indicated that they would be happy to receive potential personal health information that AI might have discovered beyond the scope of the research for early detection, but that the possibility that this could happen should be shared in advance and individuals should have more options to control how the data are used.

Analysis of Community Conversation Feedback on Case Studies 3 and 4

Community conversations highlighted varying perspectives on the use of algorithms and models in biomedical research (see additional analysis in Appendix 1, Table 7). Many of the themes from Topic #1 that community members indicated would increase their support of research (green factors) involving novel data types were also mentioned in

relation to the use of algorithms and models. Community members placed a large emphasis on transparency about the research and trusting relationships formed through community partnership. Many community members were aware of AI in the form of algorithms and models, but lacked technical expertise, and sought clear communication about the scope of their use in research and how they were developed. Several wanted to communicate with trusted advisors, in particular their personal health care provider (as opposed to technical experts), to understand how they and their communities would benefit from the research and for reassurance about potential risks to privacy, security, and to the welfare of their communities. These factors were also reflected in some of the areas where community members expressed diminished support (red factors) for the use of AI analyses in research. For example, if community members felt their privacy and anonymity could not be protected, especially in reuse of data beyond the initial research study, they were unlikely to support the research. Some community members were also aware of how bias can be built into algorithms and sought strong reassurance that stopgaps were in place to prevent these analyses from stigmatizing them or their communities.

Community members' attitudes varied depending on how AI was applied and the context of the research (yellow factors). Case Study 3 focused on the use of algorithms to analyze imaging data for better detection of lung cancer. Many of the community members who discussed this case study appreciated and advocated for the utility of AI to "see" things that might be imperceptible to the human expert's eye. They sought more information about how their privacy would be protected and their data managed and secured. They also had mixed reactions to return of their individual results, noting that they would appreciate hearing about expected personal health results (i.e., potential outcomes in scope of the research) from a trusted clinician and would want more information when results went beyond the scope of the initial study design and description. In contrast, Case Study 4, which discussed collection and analysis of public social media data to predict mental health states, evoked strong responses to the intersection of AI analyses and social media data. Some community members were concerned about the reliability of what people share through social media platforms and whether algorithms would be able to interpret any nuances (e.g., when posts are deliberately trying to exaggerate, present information ironically, or require a lot of context to understand its meaning). Many were motivated to pursue avenues that would limit suicide (in line with the goals of the research described in the case study) but were concerned about how community members could be targeted as a result of findings and how communities could be stigmatized.

Research Topic Area #3: Biomedical research and technology involving data linkage and aggregation, as typified by the following research questions:

- Are there opportunities to standardize data formats – or deploy standardizing technologies – so that data from different countries and healthcare systems could be aggregated, linked, and shared across populations?
- Which disparate (and potentially conflicting) data sets (e.g., genomics, proteomics, clinical information, clinical imaging) can be linked and combined by (automated) data aggregators?
- How can/should personal health libraries be used to combine individuals' health information across multiple different data streams to inform health outcomes?
- How can/should privacy-preserving record linkage be used to combine data on individuals from multiple sources and with different identifiers for precision medicine and public health?
- How can/should the research context (e.g., clinical and public health) and participants' consent status affect data linkage and aggregation?

Summary of Community Conversation Feedback on Research Topic Area #3

Community members were skeptical about the process of data linkage but nonetheless many community members supported its use to advance research. Community members were more likely to support linkage if there was greater transparency about how data might be used, if they received detailed information about the possibility of linkage during the informed consent process, and if they could be notified about when and why (i.e., for what research purpose) their

data would be linked for further use. Many noted and supported the use of linkage to expedite investigators' recruitment of community members for a specific study. Some community members, however, thought this was incompatible with the spirit of de-identifying participant data. As with the previous topics, community members had several concerns with the involvement of private companies in data linkage and how many different entities could be making money from participant data. Some community members felt that linkage illustrated how sharing of personal data is already out of the control of individuals and their communities. The WG noted the following main themes of conversation around this research topic area:

- **Challenges with informed consent:** Those who supported linkage tended to endorse the idea that initial broad consent gave permission to conduct data linkage with de-identified personal health data, especially for research that would have a significant impact on particularly urgent health needs. However, community members were generally skeptical of “check the box” consent procedures (e.g., clickwrap) that do not describe future possible uses in detail, including the possibility of linkage. Some community members noted frustration with feeling as if they were acquiescing as opposed to consenting to linkage in many contexts.
- **Transparency:** While some of the benefits of linkage were clear to community members, particularly for addressing issues related to the COVID-19 pandemic, many community members perceived the process as lacking transparency and posing too many risks (e.g., re-identification and data breaches), given the amount of data available for linkage and who else might receive the data downstream from research.
- **Privacy and identifiability:** In general, community members were more willing to support linkage between completely anonymized and aggregated datasets. Linking more personalized data, such as from drugstore rewards cards, was more likely to be viewed as inappropriate, though community members recognized that they may have consented to this when signing up (see note above on acquiescing vs. consenting).

Analysis of Community Conversation Feedback on Case Study 5

As shown in Table 8 in Appendix 1 (green factors), the most important consideration for community members' support of data linkage of de-identified personal health data was informed consent. Community members were more supportive of data linkage if they had explicitly consented to it or if they would be recontacted to provide consent for new data linkage. Both within and beyond consent documents, many community members desired more information about which of their data would be linked, how those data would be shared, how researchers would use the data, and what types of use their consent would cover. In general, they sought clear communication and more involvement in the research process, including notification of new linkage and the ability to revoke linkage of their data at later dates. Community members also believed that it was important for researchers to be trustworthy and some suggested that investigators establish sustained relationships with community members and their communities.

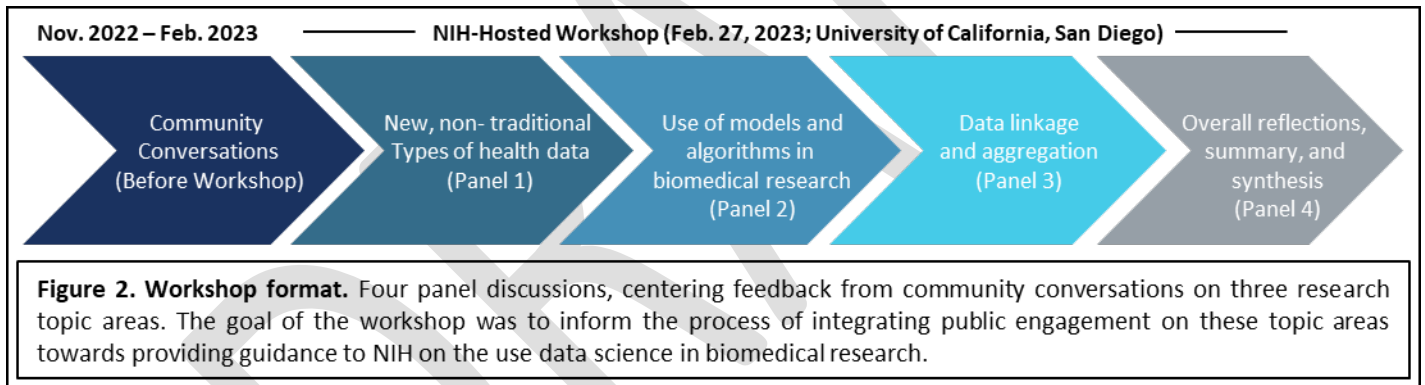
Many community members were somewhat wary of data linkage (yellow factors) because of the perceived lack of transparency about the possibility of future data linkage and use in research. Community members indicated that they would be more comfortable if details were provided in consent forms, rather than generic statements merely indicating that data would be shared, and if important information was repeated in easily digestible notifications. Relatedly, several community members were concerned about language in consent forms that permitted broad categories of future use of data. While many noted that future research is difficult to predict, they wanted consent forms to characterize subsequent data linkage with as much specificity as possible. The de-identification process was also cause for some concern. To an extent, community members were reassured by the de-identification process, but some noted that extensive linking of data made reidentification more likely and sacrificed safeguards for privacy. Lastly, while community members appreciated the potential benefits of data linkage for researchers, they wanted to see more of a return to community members and their communities (which was also a strong theme in Topics 1 and 2). Without more of a return of value, many community members thought that the risks from secondary use of linked data outweighed the potential benefits.

Indeed, community members’ strongest concerns centered on the perceived risks of future uses of linked data (red factors). Those who disapproved of data linkage tended to be concerned about linking sensitive data and linking of data that they didn’t perceive as directly relevant to health (e.g., from grocery store reward cards). Some alluded to the potential for stigma or negative judgements, especially for more easily identified communities (e.g., from Tribes or rare disease communities) and were concerned about the implications of such uses. As with other topics, community members also strongly disapproved of monetizing their data and the possibility that companies were primarily using data linkage to make money.

C. Workshop: Using Public Engagement to Inform the Use of Data in Biomedical Research

Introduction and Background

NIH charged the NExTRAC to “convene consultations with stakeholders including, but not limited to, research community members, patient groups, ethicists and privacy experts, data scientists, technology developers across sectors, and public health officials.” The community conversations described in Sections IVA & IVB gathered feedback from potential research participants and patient groups. To support the Committee’s deliberations and broaden perspectives captured, NIH held a workshop at the University of California, San Diego to contextualize the public feedback from community conversations through the lenses of research, ethics, technology development, and application to clinical and public health contexts (event webpage⁴¹ and videocast recording⁴²). The workshop consisted of four discussion panels, including panelists from Pyxis Partners who facilitated the community conversations to share key points and takeaways, ethicists, biomedical researchers, technology engineers and developers of hardware and software, public health experts, and clinician researchers who might utilize these technologies (Figure 2 below).



Panelists reviewed summary feedback from the community conversations in advance of the workshop and were asked to first provide short reflections on this feedback from their personal and professional perspectives. Thereafter WG members facilitated conversations on the defined research topics to:

- Discuss how designers, developers, and users of novel data consider the value and potential implications of this research for individuals and communities, and how they account for these considerations throughout the innovation and research process
- Contextualize community conversations that NIH held across the United States in a broader bioethical framework to ascertain lessons learned
- Integrate public feedback to assess implications for NIH programs and policies

Summary information from the panels is noted below in reference to the three research topics and towards motivating feedback on how NIH could take future action.

⁴¹ <https://osp.od.nih.gov/events/nih-workshop-using-public-engagement-to-inform-the-use-of-data-in-biomedical-research/> accessed 08.05.2023.

⁴² <https://videocast.nih.gov/watch=49084> accessed 08.05.2023.

Panelist Discussion on Research Topic Area #1: Biomedical research and technology involving novel data types used outside of the traditional healthcare system (e.g., wearable technology and social media)

Panelists discussed the importance of engaging communities, being trustworthy, building capacity for community-engaged research, and returning value to research participants who have shared personal data through emerging technologies for biomedical research. Themes discussed in Panel 1 include:

- **Seeking input from communities and participants early in the research process:** Panelists strongly encouraged researchers and developers to engage with communities early and often, and not just about data collection or novel types of data.
- **Learn how communities define trust, trustworthiness, and transparency:** Trust, trustworthiness, and transparency were prominent themes in the community conversations and panelists encouraged researchers to further explore how their unique participant populations define these terms.
- **A need for greater transparency around how data are used, not just where they originate:** Some panelists expressed that with respect to “new types of data,” the type of data did not matter as much as how the data might be used.
- **Return of benefit and value to participants:** Many panelists endorsed the idea that participants and communities should benefit from sharing data for biomedical research, particularly when considering communities that have been historically harmed or marginalized by biomedical research.

Panelist Discussion on Research Topic Area #2: Biomedical research and technology involving new types of analyses (e.g., AI, ML, NLP)

Panelists from Panel 2 also noted that investigators performing biomedical research involving new types of analyses with AI should engage with communities throughout the research process and should offer greater return to the public. Panelists discussed the need to address potential ethical, legal, and social issues with technology developers and industry researchers early in the research and innovation pipeline before issues arise. Conversations should define challenges and possible solutions to keep up with evolving methods using AI in biomedical research and to better align technology with regulations, policies, and public preferences. Themes discussed in Panel 2 include:

- **Address the unique ethical, legal, and social issues raised by biomedical research involving private industry:** Panelists noted that many community members did not trust for-profit research entities and recommended that investigators and technology developers pay more attention to the different regulations that industry researchers work within and how those regulations can negatively impact research participants (e.g., in using data acquired via end-user license agreements in lieu of research informed consent).
- **Need regular community engagement on data governance for algorithms:** Some panelists recommended that ongoing engagement with communities should inform algorithm development for biomedical research to ensure reliability (e.g., that they can be safely used to make healthcare decisions), cultural appropriateness, and to help address stigma that could arise from their use in society.
- **Return of value to those who contribute data for algorithm development:** Panelists suggested working more closely with communities to understand what they expect to receive in return for sharing their data, as it is likely not as simple as financial compensation or ownership over intellectual property.
- **Give people more control over data sharing:** Some panelists recommended that NIH provide guidance as to how to give participants more control over how their data is shared, as well as to facilitate understanding of the risks, benefits, and tradeoffs involved in sharing data in particular ways.
- **Address the longstanding challenge of staying ahead of AI development:** Some panelists noted that current regulations, policies, and ethical norms are not always well equipped to address the use of AI-based data analysis in biomedical research and recommended that NIH and other agencies lead by developing ways of

tracking algorithms being used in health spaces, as well as defining standards to assess their robustness and the presence of bias.

Panelist Discussion on Research Topic Area #3: Biomedical research and technology involving data linkage and aggregation

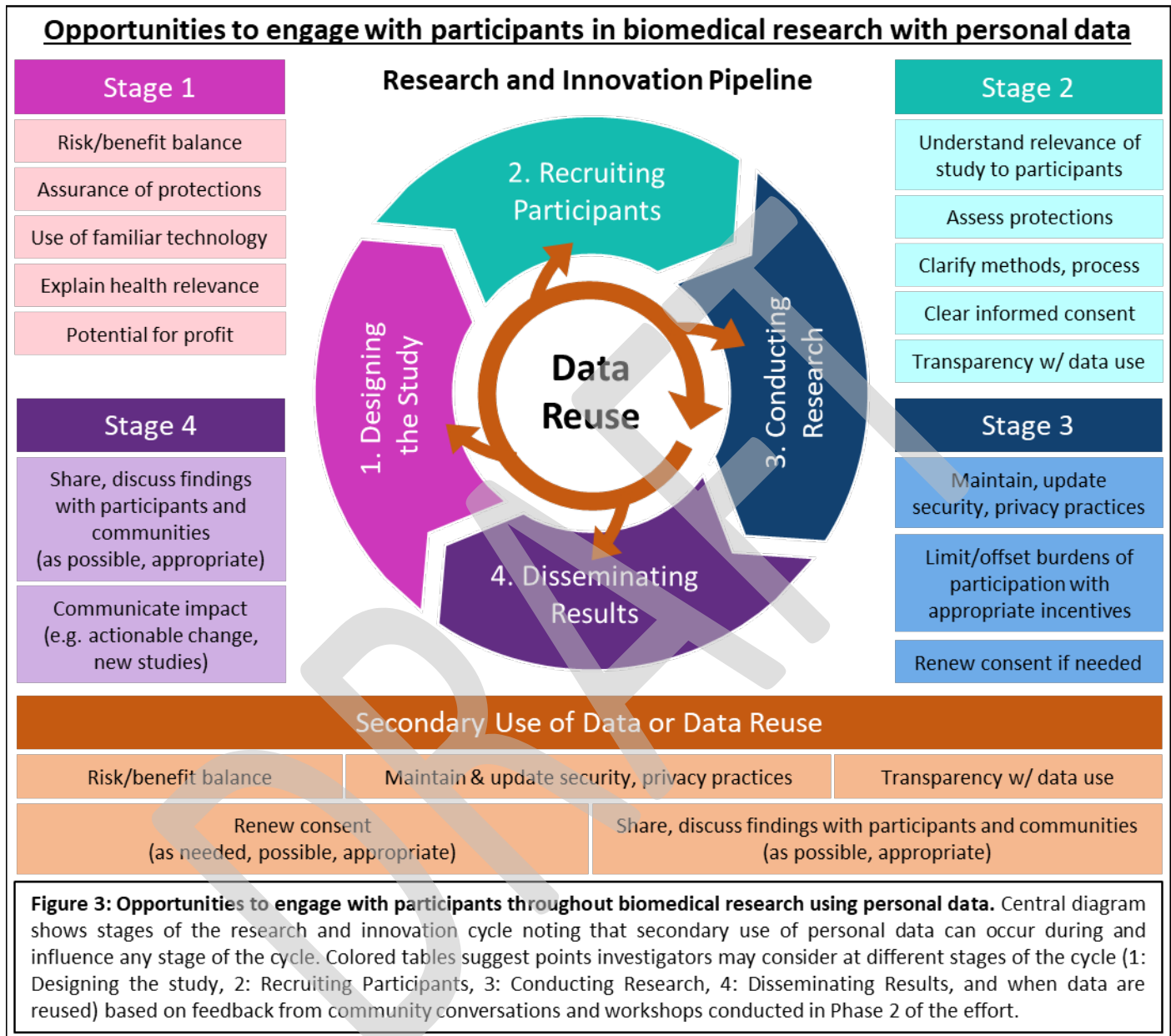
Panel 3 discussed the need for respectful engagement when requesting participation in secondary research, especially through data linkage. This could include better communication during and beyond the informed consent process and in providing support to communities that bear greater risks from data linkage. Panelists also reiterated the need for ongoing engagement with communities throughout the research and innovation pipeline. Themes discussed in Panel 3 include:

- **A need for early and ongoing engagement with communities for complex data use:** Panelists suggested that the complexities and subtleties of data linkage in biomedical research require early engagement with communities, perhaps even in defining and/or prioritizing research questions, as is consistent with a community-based participatory research approach.
- **Illustrating how linkage and aggregation can be beneficial:** Some panelists noted that data linkage and aggregation can be quite beneficial for some communities but that those benefits need to be communicated better to prospective participants, particularly as the complexity of the linkage process has increased.
- **Incorporating both individual and community preferences:** Panelists discussed the occasional tension between community and individual preferences and goals regarding participation in biomedical research, and some suggested that families and communities should be engaged to solicit their preferences for how and when to link or aggregate data in biomedical research.
- **Provide support and communication to communities with greater risks:** Some panelists suggested that community engagement is needed to involve easily identifiable communities (e.g., smaller and/or rural) and communities that have been previously harmed and/or stigmatized by biomedical research (e.g., American Indian and Alaska Native communities) in making decisions about data usage, analysis, and dissemination.
- **Being explicit about privacy risks raised by data linkage:** Some panelists suggested investigators should be fully transparent about the possible privacy risks of linkage with participants to demonstrate trustworthiness, as even technologies advertised as “privacy preserving” (e.g., PPRL) can increase the risk of re-identification.

Overall Reflections, Summary, and Synthesis from Panelists

The final panel discussed overarching points made across the previous panels. First, some panelists suggested that NIH lead by example. Private industry is heavily involved in developing new uses and analyses of data, but NIH can set policies and standards regarding data collection, management, sharing, and linkage in biomedical research, and that may encourage researchers in the broader ecosystem to follow suit. Second, they acknowledged a tension between data control and open data; future research and community engagement should more transparently impart the tradeoffs for individuals and communities between sharing data and restricting their use in biomedical research. Next, panelists identified the need to address several structural barriers to participating in biomedical research using novel types of data, analysis, and linkage. These ranged from limited affordability and access to these technologies to limitations in protecting privacy. Some panelists challenged investigators to speak directly with the individuals and communities they hope to work with at every step of the research process, including when shaping and prioritizing the research performed and after the completion of a specific effort. One way to do this would be to bolster the informed consent process through multiple touchpoints in the research innovation pipeline. Increased community engagement is also one among several mechanisms that investigators can leverage to return value to participants and their communities. Others recommended were return of research results and varying forms of compensation. In sum, panelists suggested opportunities for NIH to support research around and provide guidance on best practices for community engagement, the informed consent process, and returning value to participants. Overall feedback from community conversations and the workshop has been summarized below as points to consider for investigators for opportunities to engage with

research participants in biomedical research with personal data (Figure 3). As secondary uses of data are also an essential part of the research and innovation pipeline and were emphasized in the public engagement, these are noted distinctly.



V. Overall Observations, Findings, and Recommendations for NIH

Introduction and Background

This section describes the Committee’s overall observations and findings on the research topics in data science and biomedical research defined in Phase 1 and explored and enriched through feedback from the community conversations and workshop conducted during Phase 2. The Committee also developed recommendations for NIH, drawn from these observations, findings, and other considerations relevant to policies and programs pertaining to emerging issues in data science.

Two of the Committee’s findings primarily stem from Phase 1 deliberations and encourage NIH to continue supporting and pursuing these areas of research, the details of which are outlined in Section III:

Finding 1. Emerging technologies to capture, store, and analyze personal health data (e.g., wearables, apps, and AI) offer rich resources for research. However, more policy and ethics work is needed to responsibly realize their full benefits.

Finding 2. Emerging methods to harmonize and aggregate data from different studies and across the world show great promise for precision medicine and public health. Still, refinements are needed to limit variation that can reduce significance, accuracy of results, and generalizability of conclusions.

The rest of the observations, findings, and recommendations resulted primarily from Phase 2 deliberations. In considering the full breadth of themes from the community conversations and the public workshop, the Committee identified three topics to organize their observations, findings, and recommendations:

- **Topic I:** Trustworthiness and transparency in research and respect for participants and communities
- **Topic II:** Improving informed consent for and control of data sharing and linkage
- **Topic III:** Return of value for participation in research involving new uses and analyses of data including data sharing and linkage

Across these topics, there are a variety of initiatives—at NIH and elsewhere—that inform how investigators approach public engagement around the ethics of personal data use and analysis in biomedical research. As the world’s largest public funder of this research, NIH is well positioned to identify crosscutting efforts, elucidate best practices and lessons learned, and disseminate this information so that investigators can efficiently incorporate public feedback into their research and data management practices. In this vein, the Committee shares the following general recommendation:

Recommendation 1: NIH should catalogue, evaluate, and, as appropriate, coordinate efforts among current initiatives addressing public engagement, ethical, and technical issues in data science as applied in biomedical research.

The Committee asserts that as NIH addresses this recommendation as well as those outlined by topic below, the agency should also embed methods to evaluate their implementation and impact on biomedical research. Where useful, the Committee has added context under recommendations to guide potential NIH actions.

Topic I: Trustworthiness and Transparency in Research and Respect for Participants and Communities

Observations from community conversations: Community member support for data sharing efforts was dependent on who was receiving the data and whether they assessed that person or entity to be trustworthy. Many noted that they wanted to receive health information from a trusted clinician with whom they have a long-term relationship. Beyond that, feedback suggested some trust in the healthcare system and less in industry. While community members expressed some enthusiasm about novel uses of data across the three research topic areas, they generally wanted greater transparency in what these uses would entail and reassurances that the benefits would in fact be worth the potential risks. There was a strong emphasis on the importance of respect in approaching communities and in use of personal health data.

Observations from public workshop: Panelists encouraged a greater focus on building and demonstrating trustworthiness and improving transparency in not just how data are used, but in research more generally. What researchers and public institutions have control over is whether their conduct is trustworthy. To build trustworthiness, many panelists encouraged greater transparency around the research process, how data will be collected, analyzed, managed, and used, and the risks and benefits to participants. Across the biomedical research ecosystem, and when applying data science and emerging technologies, strong efforts should be made to develop long-lasting relationships with participants and communities through robust community engagement. Such engagement should be tailored to specific communities and to the different stages of the biomedical research and innovation pipeline.⁴³ This is critical in the context of emerging uses of personal data (for individuals and in aggregate) and their analyses in biomedical research, as these present an ever-evolving array of ethical considerations.

Other relevant considerations: Public opinion on the proper use of personal data in research may not always be based on the most recent or accurate information, can vary broadly, and may evolve.⁴⁴ For example, individuals may not be familiar with regulatory structures and mechanisms for protecting data in academic and industry research. Transparency (or the lack thereof) around the different practices in academic and industry research for consent, public engagement, and AI development may also impact their perceived trustworthiness. Occasionally, some members of the research community note that it can be difficult and time-intensive to identify champions, liaisons, and partners within communities, and that it can be challenging to maintain fidelity of information when more people are in the chain of communication around a research project.

Finding 3. There is misalignment between the expectations of some members of the public and some NIH policies and programs on the use and analysis of personal health data for biomedical research.

Recommendation 2: When NIH invests in public engagement, especially around emerging technologies, the agency should be prepared to update processes and infrastructure and to revise its programs and policies in accordance with public expectations and social justice principles, or clearly explain to the public why the agency may not revise its course.

It is important to be responsive to public feedback. Public expectations can vary and at times conflict across different communities, and federal agencies must ultimately make policy and governance decisions based on input from many invested parties. However, as a matter of respect, for members of the public to feel heard and to build trustworthiness, NIH must be prepared to translate public input into meaningful change. In some instances, NIH may chart a different path than that suggested by specific public feedback. When this occurs, NIH should provide reasons for agency decisions (e.g., as NIH often does in response to public comments on new policies).

Recommendation 3: NIH, in consultation with other federal agencies as appropriate, should align its policies on data and research governance with widely held public values.

⁴³ Unertl, K. M., Schaeffbauer, C. L., Campbell, T. R., Senteio, C., Siek, K. A., Bakken, S., & Veinot, T. C. (2016). Integrating community-based participatory research and informatics approaches to improve the engagement and health of underserved populations. *Journal of the American Medical Informatics Association*, 23, 60–73. <https://doi.org/10.1093/jamia/ocv094>.

⁴⁴ Gerdon, F., Nissenbaum, H., Bach, R.L., Kreuter, F., & Zins, S. (2021). Individual acceptance of using health data for private and public benefit: Changes during the COVID-19 pandemic. *Harvard Data Science Review*. <https://doi.org/10.1162/99608f92.edf2fc97>; Ghafur, S., van Dael, J., Leis, M., Darzi, A., & Sheikh, A. (2020). Public perceptions on data sharing: key insights from the UK and the USA. *Lancet Digital Health*, 2, e444 - e446. [https://doi.org/10.1016/S2589-7500\(20\)30161-8](https://doi.org/10.1016/S2589-7500(20)30161-8); Grande, D., Mitra, N., Iyengar, R., Merchant, R.M., Asch, D.A., Sharma, M., & Cannuscio, C.C. (2022). Consumer willingness to share personal digital information for health-related uses. *JAMA Network Open*, 5, e2144787. <https://doi:10.1001/jamanetworkopen.2021.44787>; Jagsi, R., Griffith, K.A., Jones, R.D., Krenz, C.D., Gornick, M., Spence, R.A., de Vries, R.G., Hawley, S.T., Zon, R.T., Bolte, S., Sadeghi, N., Schilsky, R.L., & Bradbury, A.R. (2019). Effect of public deliberation on patient attitudes regarding consent and data use in a learning health care system for oncology. *Journal of Clinical Oncology*, 37, 3203-3211. <https://doi.org/10.1200/JCO.19.01693>.

There is no one set of established public values and, especially for contentious topics such as the appropriate use of personal data in biomedical research, the values of different communities may conflict. Nonetheless, feedback from the community conversations as well as scholarly literature converge on certain ethical issues (e.g., interests in trustworthiness, transparency, respect, and privacy).⁴⁵ When a clear mandate emerges in public preferences and expectations, NIH should evaluate and update agency policies and structures accordingly. To ease administrative burdens, NIH should attempt to harmonize any changes and communication on these topics with executive agencies across the United States Government.

Finding 4. It is important to consult with research participants, individuals whose data are used in secondary research, families, and communities throughout the biomedical research process when novel types, uses, and analyses of data are involved.

Recommendation 4: NIH should help establish long-term relationships between members of the public and research communities to facilitate sustained public engagement on evolving issues regarding the use, analysis, management, and reuse of personal health data in biomedical research.

Commitment to long term relationships is central to trustworthiness and can lay the foundation for responsible use and reuse of personal data. Opportunities and guidance for meaningful engagement in communities around these topics will sustain partnerships beyond an individual study.

Recommendation 5: NIH should encourage investigators to prospectively address bioethical considerations in emerging areas of data science that are particularly important to participant communities, as reflected in public engagement, in their research programs and proposals.

As indicated in the previous recommendations, investigators should engage with communities early and often, including when developing and prioritizing areas of inquiry. Doing so will not only produce better research but will also help identify ethical issues raised by the collection and use of personal data. NIH should encourage investigators to embed bioethics and community engagement experts in research teams and to leverage existing sources of such expertise at their home or partner institutions.⁴⁶

Topic II: Informed Consent for and Control of Data Sharing and Linkage

Observations from community conversations: Some community members noted frustration with feeling like their options for making decisions about their data were limited in certain contexts. For example, some indicated frustration that consent forms are often long and challenging to understand. Moreover, since they are often eager to use a specific technology or participate in a given study, their consent to data sharing and/or linkage felt like acquiescence. Many community members desired to know more about how and when their data may be used after the data have been collected. While many understood that they may consent for use of their data beyond the initial research, many also desired to have more control over those uses.

Observations from public workshop: Panelists noted that current approaches to informed consent for research make it difficult for people to understand their commitments or remain engaged with a project after agreements are signed. This is particularly true in the context of consent for research beyond the initial study that may be complex

⁴⁵ Kim, J., Kim, H., Bell, E.A., Bath, T., Paul, P., Pham, A., Jiang, X., Zheng, K., & Ohno-Machado, L. (2019). Patient perspectives about decisions to share medical data and biospecimens for research. *JAMA Network Open*, 2, e199550; Kraft, S.A., Cho, M.K., Gillespie, K.A., Halley, M.C., Varsava, N., Ormond, K.E., Luft, H.S., Wilfond, B.S., & Soo-Jin Lee, S. (2018). Beyond consent: Building trusting relationships with diverse populations in precision medicine research. *American Journal of Bioethics*, 18, 20-23; Mello, M.M., Lieou, V., & Goodman, S.N. (2018). Clinical trial participants' views of the risks and benefits of data sharing. *New England Journal of Medicine*, 378, 2202–2211; Wilkins, C.H., Mapes, B.M., Jerome, R.N., Villalta-Gil, V., Pulley, J.M., & Harris, P.A. (2019). Understanding what information is valued by research participants, and why. *Health Affairs*, 38, 399-407.

⁴⁶ Shah, S., Gross, M., & Nebeker, C. (2022). Optimizing ethics engagement in research: Learning from the ethical complexities of studying opioid use in pregnancy. *Journal of Law, Medicine & Ethics*, 50, 339-347. doi:10.1017/jme.2022.61.

or challenging to understand, such as when using PPRL. They acknowledged that research participants would like more control over how their data are used and reused and that this may be at odds with the open science movement in biomedical research. One suggestion was that government infrastructure be developed to give participants greater access to more control and security in sharing their data. Such a platform could be used for improving transparency around how personal data are used, though the use of such infrastructure might also be confounded by potential lack of trust in government oversight of personal data. Several panelists suggested seeking input from communities and participants as a first step during research design with an emphasis on understanding the plurality of approaches for obtaining consent and cultivating processes tailored to communities.

Other relevant considerations: There are a number of constraints on improving research participant understanding and awareness through informed consent processes and documents. Research on comprehension of consent forms has demonstrated that even concise and easy-to-read consent forms often do not lead to high levels of understanding.⁴⁷ It is also extremely difficult to predict future uses of personal data, which limits the ability to get meaningful informed consent for those uses. Re-consenting participants and offering options for granular and/or individual control of their data may necessitate costly infrastructure that must be regularly updated. Participants may also feel burdened by frequent requests for re-consent. The legal requirements for consent are also narrowly defined in certain contexts and are codified by regulations overseen by agencies beyond NIH. Regulatory bodies may also be reticent to support increased volume of documents, notifications, and infrastructure around the consent process, as it could be challenging to harmonize information and ensure that everything conforms to legal standards.

Finding 5. More transparency is needed around the research process, how data will be collected, analyzed, and used, and the risks and benefits to research participants.

Finding 6. Current regulations and procedures for the use of biomedical research data provide ways for data to be reused in research without consent or adequate oversight.

Finding 7. Linkage of data in the absence of robust informed consent raises concerns about data privacy, transparency, and respect for research participants.

Recommendation 6 (encompassing Findings 5-7): NIH, in consultation with other federal agencies as appropriate, should develop and disseminate best practices on informing participant communities about how their personal health data are being managed, reused, and linked to other data, especially when these best practices extend beyond regulatory requirements.

Communicating information about research solely through an informed consent document is not sufficient to build trustworthiness and may in some cases even impair it (e.g., when consent forms are too long or burdened with jargon). Innovative methods are needed to improve communication both within and outside of the informed consent process. Modifications to communicate more effectively might include interactive consent processes, illustrations and graphics that convey the most important information about how personal data are shared and protected, layered consent that provides multiple opportunities to view information about a study and change one's level of participation, and signage in hospitals and research settings to indicate how personal data might be used in research and how value is returned to research participants and communities.⁴⁸

⁴⁷ Beskow, L.M., Lin, L., Dombeck, C., Gao, E., & Weinfurt, K.P. (2016). Improving biobank consent comprehension: a national randomized survey to assess the effect of a simplified form and review/retest intervention. *Genetics in Medicine*, 19, 505 - 512. <https://doi.org/10.1038/gim.2016.157>; Grady, C., Touloumi, G., Walker, A.S., Smolskis, M.C., Sharma, S., Babiker, A.G., Pantazis, N., Tavel, J.A., Florence, E., Sánchez, A., Hudson, F., Papadopoulos, A., Emanuel, E.J., Clewett, M., Munroe, D., & Denning, E.T. (2017). A randomized trial comparing concise and standard consent forms in the START trial. *PLoS ONE*, 12. <https://doi.org/10.1371/journal.pone.0172607>.

⁴⁸ Wilbanks, J. (2018). Design issues in e-consent. *The Journal of Law, Medicine & Ethics*, 46, 110 - 118. <https://doi.org/10.1177/1073110518766025>; Raj, M., Ryan, K.A., Amara, P.S., Nong, P., Calhoun, K., Trinidad, M.G., Thiel, D.B., Spector-Bagdady, K., de Vries, R.G., Kardia, S., & Platt, J.E. (2023). Policy preferences regarding health data sharing among patients with cancer: Public deliberations. *JMIR Cancer*, 9, e39631. <https://doi.org/10.2196/39631>.

Topic III: Return of Value for Participation in Research and Data Sharing and Linkage

Observations from community conversations: Community members' support for research was dependent on whether there would be a significant return of value to them, their communities, and the public. Some community members noted that communities have not historically received information about the results of their participation in research, nor how they or others may have benefitted, and that this was particularly true in marginalized communities. Though context mattered, many expressed a desire to know more about personal health outcomes from research. Additionally, some noted that when investigators (especially in industry) profited from this research, research participants should share in these profits.

Observations from public workshop: Panelists agreed that sharing data for research should translate into a return of value for research participants and communities. Some panelists noted that with industry development of AI, it has become even more evident that companies are profiting from personal data while those who contributed their data receive very little in return. They also noted that different communities may have varying conceptions of what type of return should be expected. Panelists discussed the importance of providing a return specifically when conducting research with marginalized communities. They also suggested working more closely with communities to understand what they expect to receive in return for sharing their data, as it may not always be financial compensation or rights to intellectual property.

Other relevant considerations: As evidenced by previous efforts, it can be challenging to return research results responsibly and in a way that does not pose other risks to research participants.⁴⁹ It can be expensive and complex to find and securely share private health information with the appropriate liaisons to participants (e.g., returning results to their health care provider with sufficient contextual information to convey the results).

While return of results may be one form of return of value, value could take many different forms depending on the individual participant or community engaged. It remains unclear what "value" would be most meaningful to participants and communities (e.g., financial, healthcare services, return of results) in part because there is a limited empirical base of evidence. Additionally, it can be difficult to assign value to participant contributions to a research product or patent, and determining adequate compensation raises complex ethical issues (e.g., potential for coercion and issues of justice and equity).

Finding 8: Beyond potential promises of precision medicine, research participants deserve greater return of value for themselves and their communities.

Recommendation 7: NIH should leverage prior research on return of results and support further public engagement to inform best practices for return of results from biomedical research using emerging data science and to determine implementation infrastructure needs.

Several guidelines, frameworks, and initiatives to improve the return of results to individuals and communities have been developed.⁵⁰ For a variety of reasons, these previous efforts have not been adequately implemented into biomedical research processes. Better dissemination, use, and implementation of these resources would help investigators meet many community members' preferences.

Finding 9: Providing research participants with financial benefits and/or increased control over personal data may bolster community support for biomedical research conducted by for-profit entities or in public/private partnerships.

⁴⁹ National Academies of Sciences, Engineering, and Medicine (2018). *Returning Individual Research Results to Participants: Guidance for a New Research Paradigm*. <https://nap.nationalacademies.org/catalog/25094/returning-individual-research-results-to-participants-guidance-for-a-new>; Wolf, S. M., Ossorio, P. N., Berry, S. A., Greely, H. T., McGuire, A. L., Penny, M. A., & Terry, S. F. (2020). Integrating rules for genomic research, clinical care, public health screening and DTC testing: Creating translational law for translational genomics. *Journal of Law, Medicine & Ethics*, 48, 69–86.

⁵⁰ Future of Life (2023, March 22). Pause giant AI experiments: An open letter. <https://futureoflife.org/open-letter/pause-giant-ai-experiments/>.

Recommendation 8: NIH should support exploration of methods for providing benefits and return of value beyond returning research results to research participants and their communities, when appropriate and as consistent with current regulations.

Beyond return of research results, more studies are needed to define what form of returns might provide meaningful value to research participants and communities.

The NExTRAC offers these recommendations to inform and guide the development of future policies and programs at NIH. The Committee also underscores the importance of public engagement throughout the research process to address emerging needs and preferences regarding the use of personal health data. Public engagement should be seen as essential to good research and a core component of the biomedical research process, not an ancillary task to be conducted when convenient.

It is also essential for NIH, as well as investigators and institutions, to think about *how* these recommendations will be implemented. More work is needed to define what successful public engagement looks like, how it can be measured, and generally how to evaluate the role of engagement in biomedical research. The NExTRAC encourages the development of systems and processes to ensure that the implementation of these recommendations leads to positive outcomes.

DRAFT

VI. Conclusion

Recent advances in data science promise to accelerate discovery in biomedical research and improve human health. New methods to collect personal data in real time and connect them to lived experience and environmental exposures could enable multidimensional, personalized, precision medicine and spur more rapid responses in public health emergencies. For example, a person with high blood pressure can track and store their heart rate and rhythm, skin conductance, neural recordings, quality of sleep, and blood sugar levels simultaneously on their watch. Their social media data can also be combed through to identify other subtle signs and/or predispositions to heart disease. All these data can be linked and shared with researchers to produce novel insights into human health and identify at risk groups in an emergency.

These emerging technologies also present complex and evolving ethical issues that can shift the balance of benefits and risk for individuals, families, communities, and society. The sophistication of devices and the computational power of algorithms are accelerating. At the same time, widespread applications beyond research and personal health have spurred public debate about the collection, use, and monetization of personal data. For example, the idea of collecting personal data on individuals and populations from social media posts has caused alarm in some communities, raising awareness about real and potential harms. Additionally, some have called for a societal “pause” on the development and use of generative AI (e.g., ChatGPT) until the implications of these technologies are better understood.⁵⁰ The Committee does not wish to weigh in on those debates directly but recognizes that the time appears ripe for the development and implementation of new forms of governance. In line with its framework for evaluating the use of emerging technologies in biomedical research, the NExTRAC supports the concept that public engagement must inform new policy, programs, and processes for managing how personal data are used and analyzed in biomedical research.

Researchers should partner with communities to consider questions like: When should data be shared and linked? How can AI technologies be responsibly harnessed to provide new insights while avoiding bias? How can investigators ensure data privacy while gaining increasingly granular pictures of individuals and their families? How can they limit stigma to vulnerable communities and ensure that those communities directly benefit from research? Some of these questions inspired the Committee’s effort; most are offered to NIH as the foundations for future inquiry through the Committee’s recommendations.

Historically, NIH has often relied on methods to query the public that, while broadly disseminated (e.g., Requests for Information, public workshops), may not be accessible to all. The NExTRAC’s effort in meeting this charge showcases how NIH can enhance public engagement and cultivate long term relationships through richer, more productive dialogue. In taking the time to meet members of the public through trusted liaisons in their communities and tailoring case studies to maximize discussion, the Committee garnered myriad perspectives on how people feel about the use and analysis of their (and their community’s) data in biomedical research. Three themes emerged: 1) NIH should help ensure that transparent, sustained engagement with members of the public undergirds policy development and buttresses trustworthiness of biomedical research; 2) as investigators seek to collect and reuse greater volumes of personal data, NIH should develop and support approaches to keep people informed about those uses; and 3) NIH should guide researchers to find ways to share results and return value to participants and their communities.

The Committee recognizes that the depth and breadth of public engagement should be tailored to different types of research. When this research incorporates new uses and analyses of personal health data, these practices show respect for and inclusion of participants as partners and can strengthen enthusiasm for NIH-supported studies and the biomedical research ecosystem more broadly. NIH should use the community feedback, findings, and recommendations in this report as a roadmap to build trustworthiness and maximize the utility of tax dollars towards its mission of seeking fundamental knowledge about the nature and behavior of living systems and the application of that knowledge to enhance health, lengthen life, and reduce illness and disability.

VII. Appendix 1: Further Description of Phase 2 Efforts

Appendix 1 includes further context and details on Phase 2, including the locations of the community conversations, the case studies used at each location, the text of those case studies and discussion questions, and the red-yellow-green analysis for case studies under Topics #2 and #3. This mirrors the analysis for Topic 1 in Table 5 from Section IV.

A. Schedule of Community Conversations and Webinars

Six in-person conversations were held in different geographical locations across the United States. In addition, three virtual webinars were held to gather feedback from rural health advocates, American Indian and Alaska Native populations, and rare disease advocates. Table 6 depicts the locations, dates, and case studies discussed at each site.

Location	Date	Research Topic Area #1: Novel Data Types Used Outside of the Traditional Healthcare System	Research Topic Area #2: New Ways of Analyzing Health Data	Research Topic Area #3: Linking and Aggregation of Different Types of Health Data
Harlem, NY (Northeast USA)	11.09.2022	Case Study 1: Wearable, app, social media data to monitor heart health	Case Study 3: Algorithm to detect lung cancer (imaging data)	Case Study 5: Linkage of data from a dementia study with COVID data
Bronx, NY (Northeast USA, en Español)	11.10.2022	Case Study 1: Wearable, app, social media data to monitor heart health	Case Study 4: Algorithm on social media to identify signs of suicide in teens	Case Study 5: Linkage of data from a dementia study with COVID data
Alamosa, CO (Southwest USA)	12.05.2022	Case Study 1: Wearable, app, social media data to monitor heart health	Case Study 3: Algorithm to detect lung cancer (imaging data)	Case Study 5: Linkage of data from a dementia study with COVID data
Jackson, MS (Southeast USA)	12.14.2022	Case Study 2: In-home sensors, smart devices to monitor an older relative's health	Case Study 4: Algorithm on social media to identify signs of suicide in teens	Case Study 5: Linkage of data from a dementia study with COVID data
Dearborn, MI (Midwest, USA)	01.18.2023	Case Study 2: In-home sensors, smart devices to monitor an older relative's health	Case Study 4: Algorithm on social media to identify signs of suicide in teens	Case Study 5: Linkage of data from a dementia study with COVID data
Webinar (Rural Health Advocates, Rare Disease Advocates)	01.24.2023	Case Study 2: In-home sensors, smart devices to monitor an older relative's health	Case Study 3: Algorithm to detect lung cancer (imaging data)	Case Study 5: Linkage of data from a dementia study with COVID data
Santa Clara, CA (West, USA)	01.31.2023	Case Study 1: Wearable, app, social media data to monitor heart health	Case Study 3: Algorithm to detect lung cancer (imaging data)	Case Study 5: Linkage of data from a dementia study with COVID data
Webinars (American Indian, Alaska Native Advocates)	02.16.2023 and 02.17.2023	Case Study 1: Wearable, app, social media data to monitor heart health	Case Study 4: Algorithm on social media to identify signs of suicide in teens	Case Study 5: Linkage of data from a dementia study with COVID data

B. Case Studies and Analysis of Case Study Conversations in Phase 2

A series of case studies and questions were developed to discuss with each community as part of Phase 2 engagements. The full text of the case studies and discussion questions are provided below, though these were revised slightly over the course of the community conversations in response to questions and feedback from community members. Below each case study are color-coded themes (as described in Section IV) from the conversations about those case studies. Responses to each case study are also described in detail in Section IV.

Case Study 1: Use of a wearable, smartphone app, and social media data to monitor heart health

You have been asked to participate in a research study about heart disease. You are in good health, but heart disease runs in your family. The research study is trying to identify the signs of when or how things start to go wrong. The researchers hope that by tracking several things that you do every day, they can identify “healthy” patterns that people at risk of heart disease should follow and “unhealthy” activities to avoid.

As part of the study, the researchers loan you a watch that can measure your heart rate and blood pressure. They also ask you to download an app on your phone. The app collects information from the watch, as well as other information from your phone (such as how active you were throughout the day). They also ask you to keep a record of the food that you eat and enter that information into the app. The app automatically shares all of this information with the researchers. All you have to do is keep your watch on your wrist and your phone on or near your body (such as in your pocket or a bag that you carry), and the researchers will study the data that have been collected.

To get an even clearer picture of how you're doing, the researchers also ask for permission to monitor the information you post on your social media accounts (if you are active on social media). The researchers hope that by looking at information from all of these places, they will better understand early signs of heart disease.

Case Study 1 Discussion Questions

1. Are there particular personal health or other (e.g., social media) data that you would be more comfortable sharing with researchers than others? Which data, and why?
2. What do you think is the best way for researchers to make use of this data after the study? Who should have permission to access the data? What should people be allowed to do with the data? Who should decide who can access the data?
3. Would you let your children participate in this study? Would you have any concerns about children sharing any of the specific types of data that the researchers are looking to collect?
4. In this case, researchers have outlined for participants the data that they would collect, and they have asked your permission to collect it. However, some of this data may be publicly available or collected and made available for purchase by a company that makes one of the devices or apps or by a "data aggregator" that compiles and packages information from multiple sources. How would you feel about researchers using this publicly available information or purchasing information to create a picture about your health without asking for your permission?

Case Study 2: Use of in-home sensors and smart devices to monitor an older relative's health.

Over the holidays last year your friends gave you a gift—a set of "smart" speakers, similar to a Google Echo and Amazon Alexa. The speakers not only play music and the radio but also listen to you when you talk out loud. The speakers are connected to the internet and have a virtual assistant who you can talk to, ask questions, and ask to control other devices in your home.

Recently, an elderly family member has moved into your home. They are having some difficulties remembering things clearly and also have some trouble maintaining their balance, which makes you nervous about leaving them alone. You often go to doctors' visits with your relative, and at one such visit, their doctor approaches you and your relative about a study helping researchers at the local hospital increase the safety of elderly individuals in the home.

The researchers offer to outfit your home with a network of sensors that connect to the smart speakers. For example, currently the Google Nest can be set to automatically turn the temperature of your home up or down based on your preferences. In this potential study, some of the sensors the researchers bring could detect the temperature and tell the smart speakers to turn on your home's heating and fans when it gets to a certain temperature inside. There are other sensors that can be used to monitor health conditions, such as a sleep tracker mattress, a microchip in the toilet that detects elevated lab markers in urine, a smart scale that measures weight loss, and a digital mirror that notices changes in appearance. Sensors can also analyze movement patterns to detect whether someone might be injured if they fell or were confused. However, the researchers note that none of these functions are foolproof and that the technology sometimes makes mistakes.

The researchers want to collect data on your relative's speech, movement, and vital signs using an app that will help them identify potential causes of safety issues. The speakers that are already in your home can be connected to the sensors to constantly collect data on you and your family members and send the data back to the company and researchers for analysis. This is how the researchers would monitor your relative's health status. For example, the data might be able to tell the difference between when your relative is having a stroke or just has low blood sugar. The

company and researchers claim to keep the data private and protected, but a lot of personal data is shared with the company through the app and the sensors.

Case Study 2 Discussion Questions (Table 5 in Section IV shows color-coded analysis for case studies #1 and #2)

1. How do you feel about the company getting your and your relative's data? How would you feel if other researchers were able to get this data from the company to help study other conditions?
2. Suppose you live with other family members and frequently have family and friends over to your home. Many sensors can't distinguish between different people and just collect data on everyone in the home at any given time. How would this influence (affect) your decision to enroll in the study?
3. What concerns would you have (if any)?
4. Are there factors that would increase your likelihood of participating?
5. How important is it to you to understand the details of your data and how they were analyzed? Do you want/need to understand all of the details? Who would you prefer to help you in analyzing (understanding/interpreting) your data (e.g., your physician, a researcher, or someone else)?

Case Study 3: Use of an algorithm on imaging data to detect early-stage lung cancer.

Lung cancer is a complex disease. There are people who smoke daily throughout their adult lives and remain cancer-free. Others live their entire lives in clean-air environments, avoid smoking, and have no family history with lung disease but for some reason develop lung cancer. There's still a lot to learn, and researchers continue to search for early risk factors to prevent people from getting lung cancer or to diagnose that cancer at an early stage to enable early treatment.

At your yearly checkup with your health care provider, you see a flier seeking healthy volunteers in your age group to participate in a research study looking for early risk factors of lung disease. It's a minimal time commitment. You have a family history of lung cancer, and while your affected relatives were heavy smokers and you've never smoked, you decide to participate in the study.

When you talk to the recruiter for the study, they inform you that the researchers are developing a way to compare the medical records of healthy people with the medical records of people with lung cancer, using artificial intelligence (AI). As part of the study, you will undergo a screening procedure (computerized tomography, or CT, scan). Researchers will compare your scan to scans of both healthy patients who never got cancer and patients with a variety of different types and stages of lung cancer. The researchers inform you that you have the option to receive your results from the study, and they will schedule an appointment with a lung doctor for you to discuss these at your convenience.

A few months after enrolling in the study, you see the lung doctor, who gives you the following results:

- There are a few factors in your medical records that are considered risk factors. While you haven't smoked, your parents did in your home throughout your childhood. Additionally, it was discovered that you work and have lived near some industrial factories that release chemicals that may cause cancer.
- The doctor informs you that the scans were analyzed with an AI algorithm that was specifically designed to detect early-stage lung cancer. However, the doctor cannot tell you much more about how the algorithm works. The doctor only knows the outcome: no tumor detected. The doctor has no further concerns at this time. However, given the risk factors discovered, they recommend more frequent screenings (yearly as opposed to every 5 years).

Case Study 3 Discussion Questions

1. Algorithms (i.e., computerized calculations) consider a lot of information assisting your doctor. Are you more or less likely to trust your doctor's diagnoses and guidance than that of an AI-based tool? What else, if anything, would you want to know about the researchers' process of scanning social media posts? Are there factors that would increase or decrease your trust in their use?
2. Who would you want to hear from to help you understand this sort of predictive or diagnostic health information?
3. What, if anything, would you want to know about the algorithms being used?
4. Are there factors that would increase or decrease your trust in their use?
5. Do you have any concerns about privacy?

6. How do you balance the benefits of knowing this health information with any concerns about privacy that you may have?

Case Study 4: Use of an algorithm on social media data to identify signs of suicide in teens.

Suppose you live in a town where there has been a recent spike in teen deaths. The pandemic has been hard on everyone—people have been stuck in their homes, the economy has taken a big hit, and the pressure on middle and high school students with virtual school is higher than ever. Because these teens haven't been able to interact with each other in person for some time, they're mostly expressing themselves and connecting with each other through social media apps (e.g., Instagram, TikTok, Facebook). Your neighbors have all kinds of theories about what might be causing so many deaths among younger individuals, but it's hard to know for sure. Regardless, everybody agrees that it's a problem and something needs to be done.

Researchers have developed a way to rapidly scan and interpret social media posts, including video, images, sound, and text. Public health officials in your town have contacted these researchers to determine whether this resource could be used to get at the underlying causes of teen deaths, identify signs of suicidality in teens, and develop ways to solve this problem. The researchers agree to test their process in your community and will scan social media posts and use the data to obtain information such as median (average) income and education level of the neighborhood and crime rates.

Your public officials have organized a town hall with these researchers to engage with you and your neighbors so that you can discuss the pros and cons of using this technology.

Case Study 4 discussion questions:

1. What is your first question for the researchers? For the public health officials?
2. What else, if anything, would you want to know about the researchers' process of scanning social media posts? Are there factors that would increase or decrease your trust in their use?
3. If you agree to work with the researchers, they would like to publish their results and share the data for reuse in future studies or by other researchers. Typically, any identifying information would be removed from the data before publication or sharing, though this can limit the ability to use the data to help people. Is it enough to remove identifying information? Are there other conditions that you feel would be necessary before you would consent to publication and sharing of these data about your and your neighbors' children?
4. Would you feel differently about your response if we were talking about adults and not kids?

Table 7: Qualitative Analysis of Community Member Support for Research Topic Area #2

Case Study	Green: increased support	Yellow: uncertainty or mixed support	Red: diminished support
Case Study 3: Algorithm to Detect Lung Cancer	Transparency in how algorithm was created	Privacy and security	Inaccuracy of algorithms
	Transparency in algorithm funding	Returning results	Data unrepresentative of population
	Accuracy or validity of algorithm		Risks from secondary use of data
	Data representative of population		
	Trust in both doctors and algorithms		
	Trusting relationship with personal doctor		
Transparency about uses			
Case Study 4: Social Media Algorithm to Identify Signs of Suicidality	Clear communication about scope of research and future data use	Broad scope of social media data	Causing stigma
	Transparency and honesty about the research	Collection of social media data	Risks from secondary use
	Community input on the research	Keeping minors safe	Lack of privacy and anonymity
	Clear explanations of how algorithms would work	Ensuring anonymity in social media data	
	Reassurances of privacy and security	Relevance of social media data to suicide	
	Explicit and clear informed consent	Stigmatizing information	
	Participant agency over data collection and participation	Inaccurate or unreliable social media data	
		Risk/benefit ratio	
	Unclear benefits		

Case study 5: Linkage of data from a dementia study with COVID and other data, along with use of a linkage database to facilitate recruitment for studies related to dementia.

Suppose that you decided to participate in a study of dementia looking for early risk factors and treatments. You don't have any health concerns and consider yourself a fairly healthy person. However, you've seen friends and family suffer from dementia and you hope to live well as you age, so you agree to participate. All you are asked to do is answer a basic questionnaire about your daily habits. You also agree to share your personal health data to help this and other research projects. You check a box that says, "I agree that my data will be shared for future biomedical research."

Suppose that completely separate from this study, you provided blood and saliva as part of routine testing for COVID-19. You learned that you had COVID-19, but your symptoms were mild. You had a runny nose and sore throat that went away after about one week. You again check a box that says, "I agree that specimens and associated data from my COVID-19 test will be shared for future biomedical research."

Months after the dementia study and COVID-19 testing have finished, you receive a call from a different research team. They ask whether you want to participate in a study of Long COVID. They are trying to determine how previous COVID-19 infection, which can interfere with brain function (from the loss of smell and taste to "brain fog"), may affect early indicators of dementia. But why are they calling you?

The researchers explain that they received information from a database that links data across studies. The database indicated that you had provided blood and saliva as part of routine testing for COVID-19. Although that was a while ago and not for research, you had consented for those specimens to be preserved for future research use. As it turns out, when you participated in the dementia study, the database automatically linked your dementia survey data to your COVID-19 testing data. Those data are now available for other researchers to request for use in other studies. The Long COVID researchers were able to use the database to see that you had previously participated in research relevant to their new study.

Since dementia might be related to Long COVID, the researchers are very interested in your participation. The researchers have already linked the data from the dementia survey to your data from the COVID-19 testing. They let you know that there is still a lot to learn about the long-term impact of COVID infection and that getting additional personal

health data from you will inform new interventions and therapies. Additionally, if the researchers find anything significant in your personal data, you'll have the option to discuss these results with their team physician.

Case study 5 discussion questions

1. How would you feel if you were recruited for the Long COVID study in this way?
2. Are you comfortable with having your data linked from different studies like this? Are you comfortable with researchers linking and using other pieces of your data that you have consented to having collected and stored (e.g., from COVID-19 testing)?
3. Is there any information that you think shouldn't be linked in these ways? (a) For example, what if data were linked from a drugstore rewards card purchase? (b) What if data came from Medicare or Medicaid or some other health database that has your personal information? (c) What if data came from the national census or databases on local crime, pollution, housing, and other factors affecting communities?
4. In this case, you consented to share your information for future biomedical research but not specifically for data linkage. Would you expect that researchers should have to come to you for permission again before getting any linked data about you?
5. What else, if anything, do you think should be done in order to link your data in these ways?

Table 8: Qualitative Analysis of Community Member Support for Research Topic Area #3

Case Study	Green: increased support	Yellow: uncertainty or mixed support	Red: diminished support
Case Study 5: Data Linkage	Explicit informed consent	Transparency about possibility of linkage	Linking to sensitive data
	Renewed consent for new uses	Details about linkage	Linking to non-health data
	Participant agency over new uses	Consistency between linkage and initial consent	
	Clear communication of future uses of data	Anonymity	
	Clear communication of data collection	Security and privacy	
	Ability to withdraw	Accuracy and reliability of linked data	
	Trusting relationship with researchers	Translation to benefits	
	Notification of new data uses	Risk/benefit ratio	
	Transparency about the research	Risks from secondary use	

IX. Appendix 2: Generative AI as a Use Case Study for WG Recommendations

A. Introduction

Generative AI and particularly NLP models were in development and use at the outset of this effort in November of 2021. At that time, large language models (LLMs) were being used to extract information from EHRs and social media posts, but these technologies had yet to burst into the public vernacular and popular media. Accordingly, the Committee noted these technologies and their potential applications in addressing the first part of the charge (Section III: Models and Algorithms).

In early November of 2022, OpenAI released an early demo of ChatGPT.⁵¹ This provided the first broadly available interface for public users to explore the utility of a LLM, expanding discussions of the technology from local networks in industry and academia to popular media. This resulted in provocative discussions on the potential application of LLMs across all facets of society, including the use of LLMs in biomedical research. Because LLMs like ChatGPT had not yet emerged in the public zeitgeist, they were not a primary focus of Phase 2. To supplement the discussion of models and algorithms in case studies #3 and #4, the NExTRAC chose to explore the applicability of their recommendations through the following case study as a test for how recommendations in this report might apply to LLMs.

B. Case Study

Consider a theoretical research proposal submitted as part of a hypothetical NIH-funded program on generative AI and the return of clinical results to patients:

Hypothetical NIH-Funded Program: Applications for funding are requested on the use of HIPAA-compliant LLMs to deliver the results of clinical tests to patients.

Hypothetical Proposed Research Study: A randomized controlled trial to determine how research participants perceive the return of their results from a computerized tomography (CT) scan through their patient portals in one of two forms: a standard report based on current radiology practices and using language intended for clinicians versus a report on the scan that has been translated by a HIPAA-compliant LLM to be more readable and easily interpreted by research participants.

The table below delineates one way that the NExTRAC recommendations (Executive Summary and Section V) could be interpreted for this test research question:

Recommendation	Application to Hypothetical 1) Program and 2) Research Study
1. NIH should catalogue, evaluate, and, as appropriate, coordinate efforts among current initiatives addressing public engagement, ethical, and technological issues in data science.	<p>1) NIH should catalogue existing efforts using LLMs in biomedical research and return of research results. Where possible, these efforts should be evaluated for best practices and generalizable guidance and points to consider should be developed as resources for the research community. These resources should be linked to the funding announcement for the program and investigators should be encouraged to reference them in proposals.</p> <p>2) Relevant efforts that the investigator could cite may include:</p> <ul style="list-style-type: none">• Use of AI generally in medicine• Explaining AI models and outputs• Informed consent in clinical research• Data sourcing and potential biases of LLMs• Community engagement in clinical research

⁵¹ OpenAI (2022, November 30). *Introductiong ChatGPT*. <https://openai.com/blog/chatgpt>.

	<ul style="list-style-type: none"> • Neuroethics projects looking at reporting and explaining imaging results
<p>2. When NIH invests in public engagement, especially around emerging technologies, the agency should be prepared to update processes and infrastructure and to revise its programs and policies in accordance with public expectations and social justice principles, or clearly explain to the public why the agency may not revise its course.</p>	<ol style="list-style-type: none"> 1. NIH should provide resources for how public engagement could be tailored to proposals expected for these type of research engagements and could require a plan for this public engagement as a part of the proposal. 2. Public engagements should be used to inform the design of the trial and how information is provided to participants. Given the novelty of presenting AI-generated explanations to clinical trial participants, revisions may be needed to relevant processes and infrastructure. Public engagement should be used to inform such changes, with clear justifications provided if public feedback ultimately is not incorporated into programs and policies.
<p>3. NIH, in consultation with other federal agencies as appropriate, should align its policies on data and research governance with widely held public values.</p>	<ol style="list-style-type: none"> 1. Public feedback may indicate a mismatch between public values and current policies on the conduct of clinical trials specifically when AI-generated results are involved. For example, the community conversations indicated public support for transparency in the use of models of algorithms. If clinical trials like that mentioned in the case study were found to lack transparency, new policies may be needed specifically on the use of AI in clinical trials (or perhaps the use of AI in all biomedical research).
<p>4. NIH should help establish long-term relationships between members of the public and research communities to facilitate sustained public engagement on evolving issues regarding the use, analysis, management, and reuse of personal health data in biomedical research.</p>	<ol style="list-style-type: none"> 1. As described for recommendation 2, NIH should develop guidance and best practices on public engagement that could be easily tailored for proposed research studies and require a public engagement plan as a term of the award. Where applicable, the NIH should support development and maintenance of community-investigator networks. 2. The investigator and research team should consult best practices for community engagement and build these into their research process. Community engagement on the novel uses of LLMs should occur prior to conceptualization of the study, inform the study design, and continue after the study has been completed. NIH should play an instrumental role in facilitating this type of long-term engagement.
<p>5. NIH should encourage investigators to prospectively address bioethical considerations in emerging areas of data science that are particularly important to participant communities, as reflected in public engagement, in their research programs and proposals.</p>	<ol style="list-style-type: none"> 1. NIH should continue to support bioethics programs, cores, and institutional efforts to embed bioethics in biomedical research. 2. The research study and proposal should include consultations with relevant bioethicists and public engagement experts. Relevant issues should be addressed in the study design, plans for engagement, participant recruitment, and plans for disseminating data and results at the completion of the study.
<p>6. NIH, in consultation with other federal agencies as appropriate, should develop and disseminate best practices on informing participant communities about how their personal health data are being managed, reused, and linked to other data when consent or other</p>	<ol style="list-style-type: none"> 1. The novelty of sharing AI-generated outputs may present unique challenges in explaining to potential research participants how their data may be managed, reused, and linked to other data. The research team should consult existing best practices for explaining data practices both within the consent process and through other channels. As the knowledge base on AI expands, NIH should assist in disseminating updated best practices.

<p>such notification extends beyond regulatory requirements.</p>	
<p>7. NIH should leverage prior research on return of results and support further public engagement to inform best practices for return of results from biomedical research using emerging data science and to determine implementation infrastructure needs.</p>	<ol style="list-style-type: none"> 1. NIH should develop best practices and points to consider for return of results through collecting and evaluating current efforts (Recommendation 1). As the knowledge base on returning AI-generated results expands, NIH should assist in disseminating updated best practices. 2. Because the clinical trial is aiming to inform best practices on return of results, the research team would need to incorporate certain best practices in its assessment of LLM-generated explanations. For example, there may be some results that would be harmful to communicate to participants, and the risks therein would likely need to be highlighted in the consent process, following existing best practices.
<p>8. NIH should support exploration of methods for providing benefits and return of value beyond returning research results to participants and their communities, when appropriate and as consistent with current regulations.</p>	<ol style="list-style-type: none"> 1. The research team should consider alternative ways of returning value to participants, beyond providing results of their CT scan. If the researchers are financially benefiting from the study, or if participants are supplying the data used to generate LLM outputs, there may be a particular need to provide additional compensation. Alternative methods pursued should be informed by empirical evidence from community engagement.

As the table conveys, the NExTRAC’s recommendations to NIH are flexible and can easily be tailored to research proposing the use of novel and emerging data science for biomedical applications, including the use of generative AI LLMs. Additionally, this exercise reinforces how many resources are needed to help investigators embed ethics and public engagement in their research programs. In following these recommendations, NIH would not only change the landscape of how data science is integrated into NIH-supported research but could cultivate a community of standard practice in leading by example.