The Computing Research Association (CRA)'s Computing Community Consortium (CCC) Response to the National Institutes of Health (NIH)'s Request for Information (RFI): Inviting Comments on NIH's Strategic Plan for Data Science 2023-2028

March 14, 2024

Written by: Tony Capra (University of California-San Francisco), David Danks (University of California San Diego, CCC Council Member), Haley Griffin (CCC), Carl Kingsford (Carnegie Mellon University), Rittika Shamsuddin (Oklahoma State), Katie A. Siek (Indiana University, CCC Council Member), Mona Singh (Princeton University, CCC Council Member), Donna Slonim (Tufts University), and Tammy Toscos (Parkview Health, CRA-I Council Member)

This response is from Computing Research Association (CRA)'s Computing Community Consortium (CCC). CRA is an association of nearly 250 North American computing research organizations, both academic and industrial, and partners from six professional computing societies. The mission of the CCC, a subcommittee of CRA, is to enable the pursuit of innovative, high-impact computing research that aligns with pressing national and global challenges.

Please note any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the authors' affiliations.

Introduction

We were very impressed with this Strategic Plan as a list of priorities for elevating computing in healthcare. NIH policy influences how potential fundees respond to requirements and shape their institutions. Far more of the aspects of this plan, especially data sharing, should be required as a contingency of funding (at the lab and institution level) rather than merely encouraged. We recognize that these requirements could be a challenge for some research groups, so NIH should provide facilitation mechanisms in order to ensure that researchers understand the required processes. We also suggest a gradual phase-in of these requirements.

While many of these objectives are very exciting, they are not feasible if there is not considerable funding allocated to basic research, including basic computing research and infrastructure. Many of these priorities are dependent on continued funding of aspects of

computing like data management and foundational models. Those are often viewed as “outside the scope” of NIH efforts, even though fundamental biomedical advances depend on them. These essential aspects of computing need continuous funding in order to prevent gaps in the research needed for higher-level initiatives (with periodic reassessments to ensure they are still necessary). We should support development of a funding strategy for such essential infrastructure - right now, it is not even clear how such resources can be funded unless they are continuously doing ground-breaking novel research. As just one practical example, there should be increased emphasis on maintenance of critical software or data resources.

Below we provide suggestions for improving the goals and objectives of the plan, and potential research opportunities. We have organized our comments based on the Strategic Plan Goal that they are most relevant to. We also include a list of potential partners for NIH, and a brief conclusion.

Goals 1 & 2

The strategic plan should consider how to capture qualitative and media-rich data that can be used in future data science analysis. Qualitative data provides important triangulation to better understand the context of system use.

Goal 2

The strategic plan should encourage the definition and maintenance of metadata that capture the context and history of data collected. Technology evolves quickly and current investigators need to understand the measurement and data functionalities that were available for past data collection efforts. More generally, context and history play an important part when comparing past data to current data (e.g., when were all citizens able to access mobile internet from their mobile devices to receive personalized recommendations?).

In Objective 2-2, adopting health IT standards for research seems to be missing a critical group of professionals who collect and manage data for local and state departments of health. Given the value of public health datasets and the need to modernize access (based on weakness exposed during the COVID19 pandemic response), we suggest a targeted inclusion of IT leaders from state and local departments of health. Inclusion of public health professionals seems to be foundational for standards that will cut across health care systems and community level data, including social needs data.

In Objective 2-3, there are important strategies discussed for enhancing our capabilities around the collection and use of Social Determinants of Health (SDOH) data. One pragmatic problem with the collection and use of these data is the frequent lack of responses that actually address

individuals' needs. For example, recent studies have shown that despite having effective data collection methods (e.g., electronic health record tools), these data are often not collected and/or reviewed consistently by providers. It is ethically problematic to collect these data with no plan for supporting the identified needs in communities or individuals. Thus, one additional implementation tactic could center on supporting the design of strategic ways to address the social needs of individuals/communities in order to ensure that the data that are collected are representative, ethically sourced, and meaningfully impactful. There also need to be guidelines on what data to collect and when, and support for research and systems that suggest or prioritize which data to collect to maximize its usefulness in model building.

The strategic plan should also define strategies to address miscommunication and lack of awareness among the general public about health data use for research, as transparency does not automatically lead to community understanding.

We appreciate the emphasis on interdisciplinary research; however, we encourage NIH to require higher education institutions to document how they support interdisciplinary research. For example, a document recognizing various contribution types [analogous to the CRA Best Practices Memo on Evaluating Computer Scientists and Engineers For Promotion and Tenure[1]] for all researchers independent of their department/silo for promotion and tenure. Another example would be documentation showing researchers are supported to co-train for interdisciplinary classes and degrees (instead of taking one course in data science taught by a data scientist and another course in biology taught by a biologist; the institution fully supports with pay and resources training that is co-taught by faculty in each to learn from each other and promote collaboration). An additional example is a document that shows pathways for researchers embedded in industry settings to collaborate in an interdisciplinary manner with academic or industry partners. This should highlight ways that the NIH could support such collaborations where the power distribution is uneven but acceleration of substantive research could be gained (e.g., pharmaceutical industry developing vaccination promotion materials and social scientists at a university situated in a community with substantive health inequity/ poor health literacy).

Training requires senior experts to conduct the training. In computing, retaining computer science/technology faculty that have been trained in data science/AI has been challenging due to the resource rich, higher salary industry positions. The NIH Strategic Plan to increase training would increase competition for hiring faculty in a limited pool. There are many pros to having the option for academics to have dual appointments in industry settings[2], but it can also be

[1] http://archive2.cra.org/uploads/documents/resources/bpmemos/tenure_review.pdf
[2] https://cra.org/crn/2019/08/evolving-academia-industry-relations-in-computing-research/

difficult for departments, especially at universities with more limited resources, to hire enough faculty to compensate for the loss in teaching capacity. More initiatives to adequately support dual appointment positions and interdisciplinary positions are needed.

Goal 3

In Objective 3-2 there is an exciting vision for developing new software technologies, including empowering trainees and citizen scientists to develop functional applications with software development platforms. It goes without saying that all software development should begin with clear ends and goals, lest you build a tool that is not useful or does harm. In this regard, we recommend including a citation to the NIH pragmatic clinical trial collaborative[3] and/or other national resources for implementation science. Implementation science remains an under-recognized component of successfully deploying a technology for research and should be combined with any software design initiative. Thus, one implementation tactic could be support for implementation science training or a call to adapt implementation science frameworks in the development of new software technologies.

In Objective 3-3 the speed of new technology innovation is well articulated, but a key stakeholder may be missing. The rapid advances of AI are creating a tidal wave of uncertainty and (probably) uninformed decision making on the part of healthcare executives who need a reliable source of consultation outside of the companies that are trying to get their technologies into the market. While the exploration of "innovative models for public-private partnerships" in the implementation tactics for goal 3 is inspiring, the real-world pressure on our health systems to adapt new technologies may be missed if these types of partnerships are not clearly defined and supported. If scientists are not well attuned to the pain points in healthcare delivery systems, they may miss the mark of this fast-moving market of technology tools. Furthermore, protecting patient privacy and empowering individuals with a better understanding of how their contributed data may be used in the future is important in the consent process. Rapidly changing technology environments may contribute to patient participant mistrust if data use is not explained well. ONC is working on a nutritional label type description for health AI and may be a good partner to prepare suggestions or templates for scientists who want to empower their participants with better understanding of how the data collected in their studies may be used.

We also suggest an operational definition of public-private partnerships and an emphasis on supporting health system leaders to translate quickly changing computational research (e.g., generative AI). This type of emphasis could lead to partnerships (researcher + tech company +

[3] https://rethinkingclinicaltrials.org/

health system leadership) that not only help accelerate the benefit of new computational methods, but also improve the safety and quality of care to patients.

Moreover, the increasing use of deep neural network models in biomedicine requires that computational researchers have access to large numbers of powerful graphics processing units (GPUs) to train models. However, given the expense of GPUs, many researchers do not have access to the necessary computational resources to perform state-of-the-art research. The NIH must support access to such compute resources via both funding for new hardware at diverse institutions, and access to shared cloud resources at rates that are affordable given current NIH grant budget levels.

Building on the above paragraph, study sections and review criteria should support pure computational research that has application to biological data rather than only applied biomedical research. Many biomedical research efforts require advances in fundamental computer science research, including in areas such as programming languages, algorithms, and systems.

Additionally, there needs to be support for systems research at scale. It is essential to prioritize investment in big computer systems and algorithmic challenges to deal with data. Data interoperability, reproducible and distributed processing, low latency data availability, compression, search, and storage of data, etc. are systems challenges that require core computer science research to solve. This is especially true at the scale that is mentioned in the Strategic Plan. Wearables, imaging, genomics, etc. contain large amounts of diverse data that are going to require complex systems integration.

When using AI/ML to enhance biomedical research, there should also be consideration of the issues and opportunities of synthetic data generated by AI/ML systems. Data scientists are having to address artificially generated data today, and this factor is only going to increase over time.

While the bias that can plague data is considered through the Strategic Plan, there needs to be a plan for when incorrect data is integrated. Researchers should never assume that their data is correct, and should have a plan for checking for accuracy and efficacy. The development and/or access to AI/ML tools for identifying and correcting errors should be supported.

Goal 4

The Strategic Plan contains a robust proposal for requiring researchers to make data plans, but it is very distributed. For data scientists to make use of the data, it is essential that it be in a

standardized format. These formats should include requirements on data content (required fields, standardized terminology). This kind of infrastructure would help ensure that data is ready to be inserted into AI systems and analyzed.

Goal 5

We applaud the emphasis on broadening community participation in data science. Based on our experiences, we would emphasize that institutions should be required to have checks and balances to ensure people from historically excluded groups are provided with real research experiences and treated ethically. This could be done with comparative pre-, mid-, and post-research experience surveys with one data group to make comparison reports. NIH does have guidelines for reviewing trainees, but it would be great to have a repository so there are some comparisons between groups (for instance, the two different REU comparisons that CRA's Center for Evaluating the Research Pipeline (CERP) provides in annual reports[4]). In addition, qualitative interview data is needed to hear about participants' experiences in this research training so that the institution and NIH can identify best practices for training diverse, interdisciplinary scholars.

In addition, to help the pipeline of future data science researchers, NIH should fund summer research opportunities for MS students - especially those who complete "intensive" 1-1.5 year programs. Typically, these students do not have the time in their training to get research experiences, and need funded experiences to continue in their training.

We ask NIH to use mechanisms, documentation, and reporting as necessary to show how funded institutions have worked to decrease the need to teach diverse groups about "resilience." We acknowledge that a certain amount of resilience is needed to persist within interdisciplinary, STEM fields; however, typically the resilience referred to in regards to diversity is how to deal with hostile or toxic environments. In this case, NIH has an opportunity to require funded institutions to provide documentation quantifying their culture (again, perhaps by a standardized instrument - such as the CRA Data Buddies program that provides comparable institutional culture data from survey of students[5]) and goals with timelines for improvement. Ideally, there should be other funding mechanisms to help with interventions to improve community culture scores and provide resources to support scholars. This would be a huge step in ensuring that the burden is not on historically excluded groups to persist through an unwelcoming culture and instead, would shift the burden to the institutional leaders to improve their culture for the benefit of all stakeholders.

[4] https://cra.org/cerp/wp-content/uploads/sites/4/2023/09/REU_Site_Report_Sample.pdf
[5] https://cra.org/cerp/data-buddies/#methodology

We appreciate the emphasis on recruitment, training, and mentoring of historically excluded groups. Research shows that this type of mentoring is typically done by fellow researchers from historically excluded groups which, although rewarding, adversely impacts their scholarly publication and grant production. We encourage NIH to provide funding to mentors to not only mentor, but to also keep their research going with low overhead research funding proposals. In addition, we would encourage NIH to require documentation from institutions on how research mentoring of historically excluded groups is valued in their promotion and tenure in service, teaching, and research.

Based on our experiences mentoring various groups from historically excluded groups, we encourage NIH to have funding mechanisms that help trainees stay in the training pipeline. Some trainees from historically excluded groups are affected by social determinants of health and experience (either personally or within their families) negative health outcomes. In these instances, it becomes increasingly difficult to balance training and caring for themselves or loved ones. NIH has the opportunity to provide funding mechanisms to help address these needs by allowing trainees to take a funded break to address health needs and then come back (e.g., modeled after NIH's Family Friendly Initiatives[6] and NSF's Career-Life Balance Initiative[7]).

Many communities do not have regular access to health care systems, including individuals who might not be in the US legally. As a result, there can be significant gaps in data, including those generated and used by disparate government agencies. We urge NIH to include closure of these data gaps as a major goal or subgoal in the plan.

Access to data is also incredibly crucial for research. Well-funded, established institutions have much easier access to data and greater compute abilities. These opportunity gaps should be considered in grant budgets in order to make funding accessible to all health organizations. Investigators from these well-funded institutions also have more opportunities to take part in data generating/sharing consortia, and earlier access to these data–before landmark papers–and/or broadening of existing consortia should be considered.

Furthermore, accessibility of data and infrastructure needs to be improved. Using large, heterogeneous data resources requires not only computational sophistication but a sizable investment of time and expertise. Developing tools to help users easily contribute to, access data within, and interpret information derived from these resources (like the NIH's website) would expand access and ease of leveraging data.

[6] https://grants.nih.gov/grants/policy/nih-family-friendly-initiative.htm
[7] https://www.nsf.gov/career-life-balance/

Opportunities for NIH to Partner

A partnership with local nonprofits/community organizations would help NIH reach under-resourced communities, provide funding where it is needed most, and communicate with impacted populations. There are many organizations adjacent to, and embedded in, the healthcare system that provide supportive care that would be great to reach out to. Partnering with local communities organization would enhance Goal 2 in the following ways:

- More work must be done to create "uber consensual" (as a non-profit recently explained to a respondent here) ethical consents where there is a balance between limited burden and understanding how one's data may be used in the future.
- In addition to traditional IRB mechanisms, participants should be solicited for feedback on interactions with researchers so that researchers can improve their methods and treatment of participants and their data - again with limited burden. In cases of secondary use, this would be critical to raise awareness of researcher and data fair-citizen use.[8]

Below is a list of other stakeholder groups that NIH should consider collaborating with:

- Federal institutions that support data and/or systems research, including FFRDCs that have a major emphasis on data science and data management (e.g., the Software Engineering Institute)
- Public health experts, as it is essential to understand the public health network and the way patient care fits in. Public health professionals often do not have the latest EHR, nor the funding required to integrate with computing technologies.
- Pharmaceutical companies, as even though they are very unlikely to share data, they use a lot of public data and address public health needs, so working with them would be beneficial.

In addition, the following federal agencies could be worthwhile collaborators:

- NSF (especially supercomputing centers), including NSF AI Institutes with a focus on biomedical challenges (e.g., AI-CARING) as well as divisions within the CISE directorate that focus on systems, programming languages, computational biology, and algorithms.
- Department of Energy (DOE)
- Military research systems
- Veteran Affairs (VA) - The VA hospitals and associated care systems collect large amounts of patient data representing both common (e.g. cardiovascular) and unique (e.g. combat-related PTSD) health challenges. Partnering with them might provide unique data resources and highlight very different patient and provider perspectives.

[8] https://cra.org/ccc/wp-content/uploads/sites/2/2024/03/CDARTS-Workshop-Report_Final.pdf

Conclusion

The authors of this Strategic Plan have put together an impressive set of aspirational priorities, but we have concerns that there is not sufficient acknowledgement of the significant challenges that researchers and funders will face while implementing them. It is becoming increasingly difficult for laboratories to be cutting edge due to the increasing amount of funding it takes to reach a state of the art capacity. There will also need to be a shift in the workflow of research, and NIH needs to create a multi-tiered plan to address these changes. Ultimately, researchers need to broaden participation in their research, as well as have access to large datasets and significant compute power in order to make progress on the big systems challenges that are raised throughout the Strategic Plan. This should be reflected in the mechanisms for supporting research funded by NIH.