Report of the ACD AI WG

December 6, 2019

TABLE OF CONTENTS

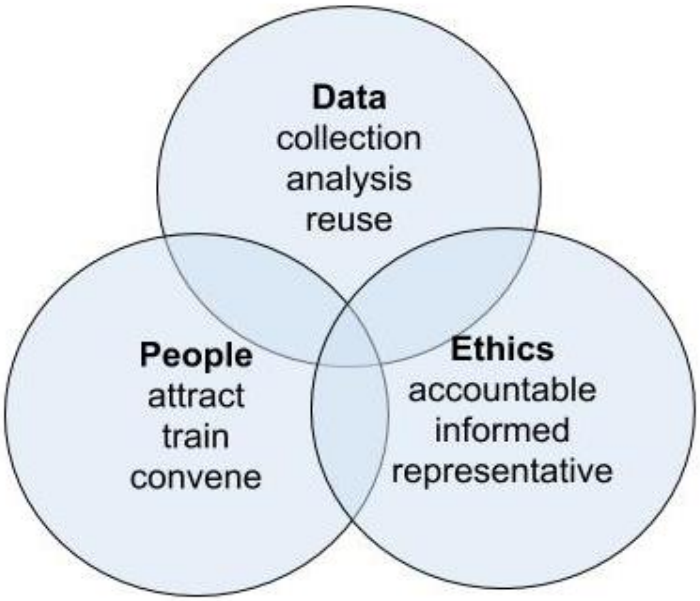
Fusing Biomedicine and Machine Learning

We live at the threshold of a new world of biomedicine, fueled by parallel revolutions in data generation and data analysis. In the life sciences, technologies such as DNA sequencing, high-resolution imaging, longitudinal electronic health records, and wearable and contactless sensors are providing more data about the biology and health of more individuals than ever before. In the computational sciences, advances in machine learning (ML) and other forms of artificial intelligence are transforming consumer technology, transportation, energy, and agriculture. The confluence of these revolutions is opening the door to a new world of ML-BioMed, with biomedical experiments that are designed for ML, and ML that's designed for biomedical experiments. In this report we make a set of recommendations on how the NIH can best ensure the use of machine learning to advance biomedical research and global health, responsibly.

Machine learning does not replace traditional biomedical data practices; instead, it adds new complementary tools to the toolbox that offer new insights and reward new kinds of data collection. These new tools embrace open-ended questions, can discover unexpected patterns, and enable the generation of new hypotheses -- see Opportunities below. However, they also pose specific risks and potential harms that require careful consideration and protections for individuals and communities -- see Challenges below.

In the sections below we expand on the opportunity we see for fusing machine learning and biomedicine, describe the challenges we see in order to realize that opportunity, and provide specific recommendations for moving forward (see Recommendations below), organized into:

(i) initiatives to create more data designed for ML-BioMed,
(ii) initiatives to advance ethics and consent practices for ML-BioMed, and
(iii) initiatives to create more people skilled in this new world of ML-BioMed.



Opportunities

The new world of ML-BioMed has the potential to drive improvements ranging from biomedical science and enhancement of clinical care, to improvements in health at the community level. As biomedicine and machine learning fuse, computational talent will be drawn to biomedicine by the mission and potential impact of working in the field. As they collaborate with biomedical experts to create and analyze ML-friendly datasets, we anticipate a world of vibrant cross-domain collaborations that lead to advances in ML and high impact discoveries in biomedicine.

A few examples illustrate the wide range of these novel applications.

Biomedical science: Understanding mechanisms. Biomedical science has made and is making enormous progress in studying fundamental components of human biology. Examples include the characterization of the human genome (including the DNA sequence, the human genes, most common genetic variants in human populations, large numbers of regulatory elements, genes underlying most rare Mendelian diseases, and 70,000 genetic variants associated with risk of common diseases) and the human cell types (through technologies for single-cell analysis of isolated cells and in tissues, and international efforts to create a Human Cell Atlas). Tools such as CRISPR have provided powerful ways to study genes, including by altering genome sequence and gene expression in living cells. Advances in large-scale screening have also made it possible to characterize the effect of perturbing cells with drugs and genetic changes.

The challenge ahead will increasingly be to systematically learn how the components *work together* to give rise to biological mechanisms underlying health and disease (such as regulatory networks, cellular programs and tissue-level interactions), to be able to understand the role of the components in these large context, and to be able to reliably predict the effect of altering them.

Given the huge number of components, inference will necessarily play a large role — and thus machine learning will be an essential tool. We will need to learn how to 'fill in' pictures, at many levels, from partial information. Examples include: imputing large-scale gene expression into histological images, based on partial data; predicting the effects of both coding and non-coding mutations, based on large but incomplete data; and inferring the effects of combinatorial perturbations too vast to ever interrogate experimentally. These studies will undergird fundamental biological research, disease studies, and drug development.

Biomedicine will push the boundaries of machine learning, and vice versa. For many ML-applications, it is enough to **make predictions**; for biomedicine, we must **infer mechanisms** in order to advance science and to develop treatments. For many ML-applications, it is only possible to observe the system; for biomedicine, the extraordinary range of tools for **experimental perturbations will drive advances in active learning and inference of causality**. Active learning uses ML to search through high-dimensional spaces, by creating intelligent feedback loops where the results of existing measurements are used to prioritize the gathering of new measurements. This technique can be fruitfully applied to many vast parameter spaces

in biomedicine -- for example, in protein design, drug regimens matched to the genomic state of cells, and using whole genome CRISPR multiplexed screens to dissect cellular circuitry. Conversely, machine learning will drive the development of new experimental approaches, using highly multiplexed readouts or automated robotics to run experiments, collect data, and use ML analysis of the data to select the next experiment to run.

Clinical care. The increasing ability to measure, in amazing detail, both the "inputs" to a person's health (from the molecular activity in individual cells, to the functional activity of their metabolism, to their physical activity as they live their life) and the health "outputs" (by extracting interventions and outcomes from their medical records) creates opportunities to train ML models to identify signals that predict particular outcomes.

While the concept of creating predictors is not new (for example, the $CHADS_2$ score[1] is used to predict stroke risk based on a handful of measurements), most such tools were developed in the era of limited data and limited computation, often restricted to what a clinician could do by hand. In the new era of ML-BioMed, much larger datasets will allow development and widespread use of more accurate markers for more conditions.

Imagine if a patient's primary physician could provide, as part of regular physicals, not just population-level advice (e.g. "lose weight", "eat healthy", "exercise more") but rich personalized information derived from new ML-powered markers:

- risk of developing metabolic syndrome and recommended interventions (specific diet, exercise, and medication) based on genome and microbiome information;
- early warnings about autoimmune disease risk and flare-ups, based on blood tests that use single-cell analysis and cytokine levels to infer the status of the immune system;
- early signs of neurological diseases, based on wearable motion sensors that feed data to your phone;
- early detection of heart disease and sleep disorders based on wearable cardiac and respiration monitors and new blood-based marker signatures;
- early cancer detection and cancer vaccine effectiveness monitoring via immune cell, cell-free DNA, and exosome analysis; and
- cancer recurrence risk via initial tumor genome analysis followed by regular blood tests.

Imagine if hospitals could monitor in-patients to know 24 hours in advance which patients were at high risk of complications and remotely monitor recently discharged patients to know days in advance which patients were at high risk of readmission.

These aspirations might be realized with sufficient data collection and ML-aided analysis. With continued advances in inexpensive and comprehensive detection technology, and continued computational advances in optimizing models to run on ubiquitous devices such as phones, the resulting learned models could be suitable for deployment at scale and equitably to large and diverse populations.

[1] https://en.wikipedia.org/wiki/CHA2DS2%E2%80%93VASc_score

There are already important early successes, including interesting examples of surprising new hypothesis generation. For example, a project to use deep learning (a type of ML) to improve and scale the detection of eye disease in retinal fundus images[2] succeeded, as hoped, at detecting referable diabetic retinopathy with high sensitivity and specificity. That same research also led to unexpected results, where the authors were able to predict cardiovascular risk factors[3] from the same types of images. That kind of unexpected result, derived by ML analysis of biomedical data, "suggests avenues of future research into the source of these associations, and whether they can be used to better understand and prevent cardiovascular disease"[4]. As the right datasets become available in other fields, similar results, pointing to similar new promising areas of research, are likely to materialize.

Social understanding of health. As biomedicine and machine learning fuse, social scientists and humanists can be included from the beginning of projects, to complement technological and biomedical knowledge. Their expertise can improve our understanding of systemic health inequities, organizational practices of healthcare, and diverse cultural approaches to health. (Relevant disciplines include Science and Technology Studies, medical anthropology and sociology, law and policy, and Human-Computer Interaction). This inclusive approach can help to ensure that ML contributes to a future that reduces structural injustices and social harms, rather than amplifying and reinforcing them.
Benefits of this more inclusive future could include:

- earlier detection and resolution of data bias issues before systems are widely deployed;
- consideration of non-technical determinants of health, such as the political and economic structures within which health technologies are deployed, and misaligned incentives (e.g. between insurers and healthcare providers);
- mechanisms for measuring and modulating patient-doctor communication as an important variable when studying technological interventions that affect clinical workflows; and
- data-driven health programs that improve patient health without compromising patient privacy.

Ethics and data sharing. The new discipline of ML-BioMed has an opportunity to drive the creation of best practices for data sharing, patient consent, and responsible data use that will promote research insights and clinical interventions, as well as inform and shape trustworthy and accountable data practices. Combining the ethos of experimentation from the ML field with the traditions of responsible data practices in the biomedical space will allow for research

---

[2] Gulshan V, Peng L, Coram M, et al. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. JAMA. 2016;316(22):2402–2410. doi:https://doi.org/10.1001/jama.2016.17216

[3] Poplin, R., Varadarajan, A.V., Blumer, K. et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng 2, 158–164 (2018) doi:10.1038/s41551-018-0195-0

[4] Ibid.

breakthroughs that improve the lives of all people, while preserving the privacy, agency, and respect of patients and their data.

These new best practices can lead to the creation and widespread adoption of tools such as:

- novel ways to build informed consent throughout the data lifecycle of a clinical study
- accountability and auditing mechanisms to record what health data is used and where
- methods to assess the efficacy of models trained on one population for use on others
- processes to include patient groups in the design of health data collection and use

Challenges

Building a bridge from today's world to the vision described above requires addressing several major challenges, including avoiding the risks of harms.

Data Challenges

ML-powered biomedical discovery and care is fueled by data. The awareness that data holds tremendous value, particularly when combined with data science and machine learning, has spurred many efforts to aggregate existing biomedical datasets. Unfortunately, these datasets can fall drastically short of their potential if they are not well-suited for the application of machine learning methods and can even present serious risks.

Raw "exhaust" datasets collected in the course of healthcare delivery may not be very useful if they lack crucial metadata — about demographic parameters, collection methods, systematic inaccuracies (e.g., due to the desire to achieve reimbursement) and many other types of information. In many cases, they can contain biases that would severely compromise machine learning models trained on the data — including biases that would harm specific groups.

Similarly, experimental data may not be useful if it lacks crucial metadata and quality controls (to avoid results dominated by artifacts such as batch effects and other data collection noise); lacks sufficiently rich information (across modalities, time points, and interventions) to include key causal factors; has a sample size that is too small; or has access policies that are too restrictive.

While attempts are sometimes made to try to "fix" datasets that were not properly designed for ML, these efforts are typically slow, expensive and yield inadequate results. Achieving the effective convergence of biomedical data and machine learning requires datasets to be thoughtfully designed from the outset to be valuable for machine learning-based analysis.

The data challenge ahead includes (i) developing experimental approaches that are designed to efficiently capture information optimized for machine learning; (ii) applying these methods to collect massive data designed to address key biomedical needs; and (iii) ensuring that the datasets are carefully described and made available in ways that maximize its value for intended purposes, while minimizing the risks of adverse unintended use.

Consent Challenges

We lack appropriate guidelines and tools for participant-facing consent and researcher-facing data access consistent with the opportunity for ML-BioMed. Many important insights will come from wide re-use of data -- not limited to a single disease state, pooled and combined with other datasets, and used to train models for academic and commercial purposes. Therefore, we need clear guidelines on when such re-use is allowed, and on how to inform participants about the many potential uses of data they contribute. Similarly, many important insights will come from a wide pool of researchers, with diverse backgrounds beyond a single institution or

research area. Therefore, we need clear guidelines on who are "bona fide researchers" and what data they are allowed to use. And we need streamlined data access mechanisms, potentially including a "data passport" model where one approval process grants access to multiple global datasets.

There is a substantial gap between consent standards typically required in biomedical research and consent standards typically applied in ML, where it is common practice for ML engineers and scientists to create training sets for ML models by scraping the internet of public text, images, and videos without explicit additional consent for such reuse. Widespread scraping of public data has raised serious privacy and ethical considerations (such as in this article on Facial recognition's 'dirty little secret[5]) and is already regulated in Illinois under the Biometric Privacy Act (BIPA).

These practices are of special concern for data focused on historically marginalized groups, for data requiring consultation with sovereign American Indian and Alaska Native Tribal nations, and for data that might involve individuals living in other countries where laws governing data are weak or essentially non-existent. These are also populations who are typically under-represented in biomedical research, so it is important to have standards that protect all populations, while ensuring that the benefits of research will flow to all populations.

Ethics Challenges

ML technologies are rapidly changing the landscape of healthcare systems, from assisting and identifying new directions in biomedical research, to aiding in diagnosis, and informing decision-making in health. Oversight and investigation into the use of ML tools has not caught up to the proliferation of use, leading to a number of unanswered questions.

- **Fairness and equity:** How can we build ML systems to identify and mitigate harm to disadvantaged and marginalized groups in an adequate and timely manner? How can we ensure that ML systems do not replicate historical bias and discrimination in biomedicine and public health research? What measures will be taken to account for existing health disparities as well as data gaps and inequalities where disadvantaged communities are under-represented, mis-represented, or entirely missing in existing datasets? How can ML assist in identifying opportunities to shed light on and potentially mitigate health inequities? How can we ensure inclusion throughout the research, development, and deployment pipeline of ML systems, and what mechanisms will be built for soliciting and incorporating feedback from affected groups?
- **Privacy and consent:** What measures will be put in place to ensure the privacy and informed consent of patients and communities, and to address cross-use of data that is gathered for one purpose but used for many others? Specifically, what steps should researchers follow for large Web and social media data, which is actively used in a

[5] Erik Carter, NBC News, Facial recognition's 'dirty little secret': Millions of online photos scraped without consent; https://www.nbcnews.com/tech/internet/facial-recognition-s-dirty-little-secret-millions-online-photos-scraped-n981921

variety of contexts, including diagnosing mental health and disabilities? How will informed consent be defined in these varied contexts? What data sharing practices ought to be put in place?

- **Reliability, safety, and security:** ML tools are used in a variety of situations, including many life-and-death ones. What guidelines will be put in place for testing for reliability, safety, and security of these tools? I.e., that they are functioning as per their intended use? What metrics will we use and how will this be done in harmony with the FDA's role? How will we ensure that these measures are multi-dimensional, accounting for the varied short- and long-term harms that they might cause? How will we ensure that these tools comply with the law, especially in multi-national contexts where the data used or communities affected might come from many nations?
- **Accountability and governance:** What roles and regulations govern the use of ML in biomedicine and health? How do we define the intended function of these tools, both initially and as their use evolves? Who is accountable when ML tools are used in health contexts throughout the end-to-end process? How do we test for the transparency and traceability of the use of ML in biomedicine? How do we validate the output of ML systems? What evaluation metrics will be put in place?
- **Education:** What education around ethical concerns should be provided for researchers in ML and biomedicine, and for practitioners who use ML in their day-to-day decision-making? What skills and training should we provide patients and doctors around human-computer interaction with ML-powered systems? How do we educate the general public about the use of ML in this context?

Without new coordinated efforts, machine learning can reinforce existing blind spots and biases in medicine, adding a new unjustified veneer of technical credibility. Obermeyer et al.'s recent paper, Dissecting racial bias in an algorithm used to manage the health of populations[6], shows how a widely used algorithm that attempts to predict who will most benefit from health interventions systematically underrepresents Black patients. The reason is that the algorithm's predictions are based on historical healthcare *utilization*. However, healthcare utilization tends to be much lower for Black patients than White patients with similar medical conditions. In this case, the authors were able to pick apart the contributing inputs to the algorithm and understand where bias was introduced. As future ML-based algorithms are introduced, that level of retrospective explainability could be lost, unless careful attention is paid to these issues.

People Challenges

Experts in computational fields are often interested in using their skills to advance biomedicine, but do not have the vocabulary or context to do so efficiently. Experts in biomedical fields are often interested in taking advantage of new computational tools, but do not have the vocabulary or context to know how to do so, or even to know which tools make sense for which

[6] Obermeyer et al, Science 2019; https://science.sciencemag.org/content/366/6464/447

problems. And neither group alone has the skills and background needed to evaluate whether the results are true signal.

Realizing the full value of ML-BioMed requires collaborative teams, with experts from multiple domains. Until those experts understand enough about each others' disciplines to collaborate effectively, though, they will find it challenging to:

- make biological/clinical data ready for ML (data wrangling, while dealing with batch effects, confounders, and artifacts)
- formulate machine learning modeling tasks that are suited to available biomedical data and address important biomedical questions
- articulate biologically and clinically relevant metrics of success, and translate these into model evaluation metrics
- troubleshoot models for both technical and biomedical issues
- interpret models, and not overinterpret models
- look out for blind spots, caveats, and biases of datasets and models

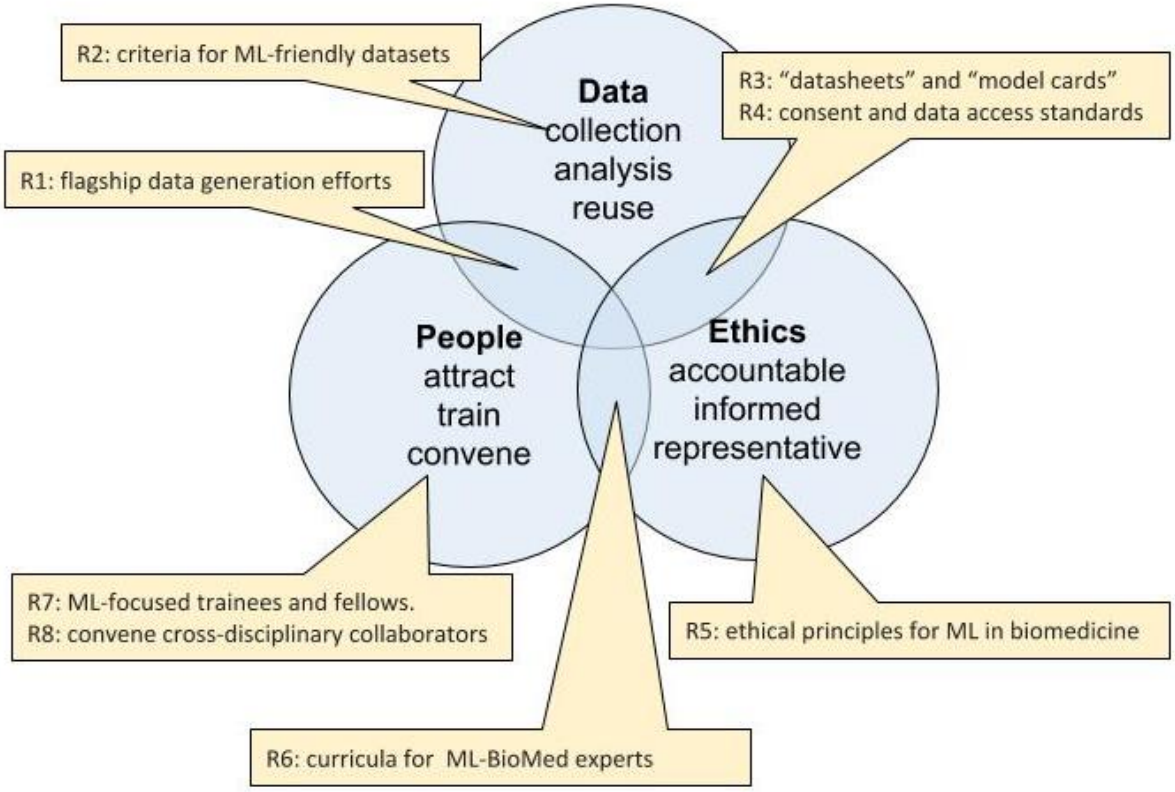
The current educational system, from high school through college, grad school, and ongoing professional education, treats machine learning and biomedicine as largely disjoint fields, with little opportunity for cross-training or even cross-awareness. The current professional conference landscape, which can provide a fruitful environment for cross-disciplinary cross-pollination, is also largely disjoint, with attendees from either biomedicine or computation, but rarely both.

Recommendations

The Artificial Intelligence Working Group was assembled in the closing months of 2018 and was charged[7] with addressing how NIH could take advantage of new opportunities in AI to advance biomedicine, including workforce and ethical considerations. Over the course of the past year, we have met twice in person, conducted several teleconferences, and presented a mid-year report[8] at the ACD meeting in June 2019.

We have eight specific recommendations that collectively address the needs for investment in data, ethics, and people as we move into the world of fused biomedicine and ML.

For clarity, we intend the term 'biomedicine' below to include a broad range of activities — including studies of the molecular and cellular basis of biological and disease mechanisms; development of methods for disease prevention, diagnosis and treatment; clinical health care; and relevant social, behavior, and humanistic sciences.



[7] AI Working Group Background, https://acd.od.nih.gov/working-groups/ai.html

[8] AI Working Group Update, https://acd.od.nih.gov/documents/presentations/06132019AI.pdf

Recommendation 1: Support flagship data generation efforts to propel progress by the scientific community.

The NIH should support a program of ambitious flagship projects to propel a deep fusion of biomedicine and machine learning by supporting centers and networks aimed at important topics at the cutting edge of these fields.

These efforts would generate large-scale experimental data, with billions of data points designed to:

- **(i) be well-suited for ML analysis and inference,**
- **(ii) address key biomedical challenges, and**
- **(iii) stimulate new approaches in machine learning.**

And implement processes designed to:

- **(iv) develop improved criteria and technical mechanisms for data access, and**
- **(v) strengthen ethical criteria for dataset use, with a focus on consent, privacy and accountability.**

While the efforts should include appropriate analysis capabilities, the primary purpose should be to empower the entire biomedical, computational, and social scientific communities while respecting patients and participants. Accordingly, the efforts should:

- **(i) be required to make the datasets generated available to the scientific community immediately (subject only to necessary quality control) and freely (subject only to any access controls required to protect patient privacy and honor informed consent);**
- **(ii) include mechanisms that promote rapid exchange of ideas and progress, such as challenge problems, easily computed model performance metrics, and jamborees for the community; and**
- **(iii) to the extent feasible, include mechanisms to test hypotheses produced by the community through machine learning analyses.**

Given the value of fusing biomedicine and machine learning to a wide range of areas of importance to the NIH and the importance of broadly engaging the NIH community in this work, the activities would ideally be supported through both NIH-wide mechanisms (e.g., Common Fund) and mechanisms involving one or several ICs.

These efforts would aid in the development and real-world validation of the technical and ethical criteria discussed below. Therefore, as part of this program, the NIH should also support a cross-cutting activity to develop standards as described in several of the recommendations below. This effort could be led by an advisory panel and should involve and draw on the experiences of the flagship projects.

The topics should primarily be determined based on the quality of proposals received, with the recognition that some NIH ICs may wish to support particular areas. For the sake of illustration, potential topics might include:

- Cellular pathways: Inferring cellular pathways based on large-scale gene perturbation data.
- Genetic variants: Inferring the role of non-coding variants based on observational and perturbational data.
- Disparities in healthcare: Predicting minority patients at risk for death or complications after surgery based on data from disparities databases and clinical records.
- Histology: Automatic annotation of cellular structures and/or gene expression in histological images.
- Microbiology: Inferring interactions in microbiome among bacterial species and with their human hosts.
- Chemistry: In silico drug-like molecule creation; retrosynthetic planning for drug-like molecules; physical property prediction; toxicology predictions.
- Medical images: Detecting the presence of abnormalities in radiographic images; real-time simultaneous spatial and temporal cellular images.
- Clinical data: Predicting patient outcomes from longitudinal electronic health record data.
- Sensors: Inferring health attributes from wearable, contactless, and other types of digital health sensors.

These flagship projects will aim to address key biomedical challenges using machine learning methods, and to advance machine learning methods for future use in biomedicine. As such, they should involve much more than straightforward application of existing machine learning methods. Rather, they will press the boundaries by propelling new ways to gather massive data in biomedicine and develop new methods in machine learning. Getting this right will often involve interplay between the capabilities of experimental and computational science (for example, in deciding what is best measured and what can be imputed using machine learning methods). The scientific teams in these projects should involve a strong engagement of high-caliber investigators from both disciplines, including leading researchers in machine learning (not just in biostatistics, bioinformatics, or computational biology).

The projects should be selected based on a combination of their relevance to important biomedical needs (including understudied questions), their ability to produce transformative data sets, their value in advancing ML analysis methods for biomedical data, and the timeliness with which they can be created. Project review should incorporate expertise in machine learning as well as traditional biomedical domains. It is critical that new data sets be generated with machine learning in mind, including compliance with the best practices being specified as part of Recommendation 2.

Recommendation 2: Develop and publish criteria for ML-friendly datasets.

The NIH should develop (possibly as part of the cross-cutting activity described in Recommendation 1 above) and publish criteria for evaluating datasets based on their value for ML-based analysis.

The criteria should include illustrative evaluations of existing datasets against the criteria, serving as examples of where current practices are well-aligned with the opportunity for ML-powered discovery, and where there are gaps that need to be addressed.

We suggest the criteria initially be published as guidelines, to foster community input and real-world feedback. We further suggest that, within two years of initial publication:

(i) the criteria be updated;
(ii) a review process be established to assess significant data releases against them; and
(iii) a subset of the criteria be recommended as requirements for future NIH-funded datasets.

Possible criteria include:

- clear provenance -- as much metadata as possible, to detect and correct for batch effects
- well-described data -- what does each variable mean? what's the distribution of values?
- accessible data -- flexible data access policy, reasonable data access process
- large sample size -- to allow training (and evaluation) without overfitting
- multimodal data -- to study complex systems from multiple perspectives
- perturbation data -- includes outcomes ("outputs") as well as measurements ("inputs")
- longitudinal data -- to allow modeling and prediction of progression
- active learning -- data grows over time, incorporates new data-gathering techniques, and uses ML-based analysis of existing data to inform future data generation

The UK Biobank[9] is a recent example of a dataset that addresses many of these criteria. Provenance is clear, data is well-described and accessible, 500,000 people is the beginning of a sufficiently large sample size, and there are data points from multiple modalities and time points. Furthermore, data in the UK Biobank are collected using rigorous protocols that minimize batch effects and artifacts.

The NIH *All of Us* Research Program[10] is a new dataset that intends to also address many of these criteria, including provenance, data description, and data access. The targeted sample

[9] UK Biobank; https://www.ukbiobank.ac.uk/

[10] NIH All of Us Research Program: https://allofus.nih.gov/

size (over a million people) and data modalities will be useful for ML, and the data collection protocol aims to minimize batch effects.

Recommendation 3: Design and apply "datasheets" and "model cards" for biomedical ML.

The NIH should develop (possibly as part of the cross-cutting activity described in Recommendation 1 above) and publish best practices for "datasheets" that describe and evaluate training datasets, and "model cards" that do the same for generated models. The best practices should include examples created after the fact for existing biomedical datasets, ideally with the participation of the original dataset creators.

We suggest that, within two years of initial publication:

- **(i) the best practices be updated based on feedback from applying them in the real world;**
- **(ii) the NIH require that all extramural NIH grant applications and all intramural NIH projects that involve ML research must include datasheets and model cards; and**
- **(iii) the NIH encourage journals to require the submission of such datasheets and model cards along with submission for publication of any paper involving ML research.**

Datasheets for datasets[11] used in ML-BioMed could include: where the content was sourced; the relevant demographics and "under-represented in biomedical research" (UBR) characteristics of the data (e.g. using the participant characteristics reported on by the *All of Us* Research Program[12]); and any potential legal and ethical issues including privacy, consent and copyright. The datasheet should also have a section discussing any potential harms that the set could cause, so that future users can be aware of those risks. One critical intent of datasheets is to be explicit about known blind spots in the data, which could otherwise create hidden biases in derived models, leading to problematic downstream effects.

Model cards for generated models (as described here[13] and here[14]) would include: what training data was used (including datasheets where possible), how training and validation were done, intended use of the model, known trade-offs and limitations on applicability, and estimates of performance in various circumstances. The model card should also have a section discussion ethical considerations, including potential harms of inappropriate use of the model.

[11] Gebru et al, arXiv 2019; https://arxiv.org/pdf/1803.09010.pdf

[12] N Engl J Med. 2019 Aug 15;381(7):668-676. doi: 10.1056/NEJMsr1809937

[13] Mitchell et al, arXiv 2019; https://arxiv.org/pdf/1810.03993.pdf

[14] Model Cards, Google; https://modelcards.withgoogle.com/

Recommendation 4: Develop and publish consent and data access standards for biomedical ML.

The NIH should charge a (new or existing) working group to address the substantial gap between consent standards typically required in biomedical research and consent standards typically applied in ML, where it is common practice for ML engineers and scientists to create training sets for ML models by scraping the internet for content.

Standards should be developed that ensure appropriate consent for biomedical ML, by reconciling common ML practices, existing biomedical best practices, and ongoing efforts in the global biomedical community to harmonize consent and data use standards to facilitate the widest responsible use of data, while ensuring protections against potential harms.

Once draft standards are developed, the NIH should establish a process to review significant new projects against the standards, in order to test theory against real-world practice and refine the standards. After the standards are finalized, the NIH should implement appropriate mechanisms to require adherence.

Recommendation 5: Publish ethical principles for the use of ML in biomedicine.

The NIH should charge a (new or existing) working group to move rapidly, within the next year, to develop a set of ethical principles for the use of ML in biomedicine, including guidelines for ensuring fairness, equity, governance, respect, accountability and transparency. This working group will be tasked with grappling with the unique set of ethical challenges in this space, that add to existing challenges for the use of ML in other public and private sector settings.

We recommend the working group include researchers and practitioners in ML, biomedicine, law and public policy, Science and Technology Studies (STS), and related disciplines, including representation of communities that will potentially be negatively impacted by ML technologies in biomedicine.

The working group should formalize these principles and create short- and long-term strategies for the development and use of ML techniques that can enhance biomedical research and the delivery of health-care while ensuring that all those at risk from harm are protected.

Once draft principles are agreed on, the NIH should establish a review process to assess significant new publications against the principles, thus testing theory against real-world practice, and using the gaps to refine the principles. After refinement, the NIH should implement appropriate mechanisms to require adherence with the principles.

Topics of focus for this working group include the issues discussed in the Challenges above:

- **Fairness and equity:** ensuring that ML systems not only avoid reinforcing existing biases, but also do not contribute to future health disparities and inequities.
- **Privacy and consent:** coordinating with the Recommendation 4 work on consent, opt-out mechanisms, and data access standards
- **Reliability, safety, and security:** extending existing biomedical oversight mechanisms as needed to ensure reliability, safety, and security of ML-powered tools
- **Accountability and governance:** extending existing biomedical accountability mechanisms and governance procedures as needed to account for the unique attributes of ML-powered tools
- **Education:** coordinating with the Recommendation 6 work on including ethics content in new curricula for researchers, and going further to inform broader audiences

Recommendation 6: Develop curricula to attract and train ML-BioMed experts.

The NIH should fund the development of curricula, at multiple levels from high school through professional education, designed to

(i) entice upcoming and established data experts into the field of biomedicine, educate them on the opportunities and challenges, and help them to successfully collaborate with those in the biomedical field and experts in the social and humanistic sciences;

(ii) inform upcoming and established biomedical experts about modern ML techniques, including the techniques' strengths and limitations[15], and help them to successfully collaborate with experts across multiple fields;

(iii) Invite social scientists and humanists with a focus on data and its wider social implications to collaborate on studies and inform on best practices; and

(iv) raise the awareness of biomedical policymakers and decision makers about the opportunities and risks for applying modern ML techniques, and help them know what questions to ask to aid in their decision making.

These curricula should incorporate elements on the risks of mis-applying ML in biomedicine, including wider social and ethical considerations, such as problems of hidden bias from non-representative training sets.

For experts in biomedicine and experts in machine learning to become effective multi-lingual researchers (ML-BioMed experts), it is important to develop curricula tailored to their respective backgrounds. Possible elements of such curricula include:

Bio4ML: train upcoming and established data experts to successfully collaborate with biomedical experts	**ML4Bio:** train upcoming and established biomedical experts to successfully collaborate with data experts
● Machine learning courses that include biology and clinical medicine applications as motivating examples ● Biology and clinical medicine courses that emphasize fundamental concepts and problems amenable to machine learning analysis, rather than detailed memorization of facts	● Machine learning overview courses that emphasize fundamental concepts (e.g. how to assess the suitability of a problem for ML, how to design datasets for ML, and how to assess applicability and limitations of developed ML models), rather than minute technical details of ML algorithms

[15] https://xkcd.com/1425/

<table>
<tr><td>

- Hands-on exercises to build models from real-world biomedical datasets, focused on understanding the unique access policies and analysis characteristics of such data

</td><td>

- Hands-on exercises to use real-world ML tools to train and evaluate biomedical models, focused on understanding the train - evaluate - deploy process

</td></tr>
<tr><td colspan="2">

- Opportunities to take part in curated challenges, ideally in collaborative teams with students from the complementary discipline. Ideal challenges should:
  - be of wide appeal and real-world importance
  - be tractable but not require extensive biological background
  - utilize biomedical data with student-friendly data access policies
  - use datasets with real-world idiosyncrasies (e.g. confounders), that are properly annotated so that participants learn to take them into account in modeling efforts

</td></tr>
<tr><td colspan="2">

- Education on the principles of ML and healthcare ethics, including consent, fairness, and privacy

</td></tr>
<tr><td colspan="2">

- Fellowships and other programmatic opportunities to gain in-depth experience in a complementary (biomedical or data science) environment

</td></tr>
</table>

Opportunities such as workshops and other events for data experts and biomedical experts to mingle, learn from each other, and form new collaborations will also be mutually beneficial.

Recommendation 7: Expand the pilot for ML-focused trainees and fellows.

The NIH should continue and expand the inclusion of ML-focused projects in its existing trainee and fellow programs. This approach was successfully piloted in summer 2019 with three ML projects in the Civic Digital Fellowship[16] and two in the Graduate Data Science Summer Program[17]. The NIH should make ML a major ongoing focus of these and similar programs going forward.

For example, future trainee and fellow projects could include efforts similar to the ones done in the summer of 2019:

- use topic modeling to categorize grants by subject to inform portfolio distribution based on best fit with program officer subject matter expertise
- develop machine learning models to predict migration paths and shape changes of fibroblast cells in dishes
- develop algorithms to extract, validate, and load missing data to the EYEgene genotype/phenotype database using optical character recognition/computer vision
- use topic modeling in genomics for gut microbiome taxonomy
- use machine learning methods to augment data for drug development pipelines

And other efforts such as:

- develop mechanisms to assess training data for potential biases

[16] https://www.codingitforward.com/fellowship

[17] https://www.training.nih.gov/data_science_summer

Recommendation 8: Convene cross-disciplinary collaborators.

The NIH should continue and expand support for biomedical tracks and workshops at leading computational conferences.

This approach was piloted at NeurIPS in December 2019, and should be expanded to other conferences and other opportunities for convening experts from different fields.

We suggest considering targeting the following computationally-focused conferences:

- AAAI (Association for the Advancement of Artificial Intelligence)
- ICML (International Conference on Machine Learning)
- CSCW (ACM Conference on Computer-Supported Cooperative Work and Social Computing)
- FAT* (ACM Conference on Fairness, Accountability, and Transparency)
- NeurIPS[18] (Neural Information Processing Systems)
- CVPR[19] (Computer Vision and Pattern Recognition)
- ACL[20]/NAACL/EMNLP (Meeting of the Association for Computational Linguistics)
- CHI (ACM CHI Conference on Human Factors in Computing Systems)
- RECOMB[21]
- ISMB (MLCSB)[22] (International Society for Computational Biology)
- MLCB[23] (Machine Learning in Computational Biology)

And more general biomedicine and scientific conferences, such as these:

- 4S[24] (Society for the Social Studies of Science)
- ACS[25] (American Chemical Society)
- FASEB[26] (Federation of Societies for Experimental Biology)
- ASBMB[27] (American Society for Biochemistry and Molecular Biology)
- SfN[28] (Society for Neuroscience)

[18] https://nips.cc/
[19] http://cvpr2019.thecvf.com/
[20] https://acl2020.org/
[21] https://www.recomb2020.org/
[22] https://www.iscb.org/about-ismb
[23] https://sites.google.com/cs.washington.edu/mlcb/
[24] https://www.4sonline.org/meeting
[25] https://www.acs.org/
[26] https://faseb.org/Science-Research-Conferences.aspx
[27] https://www.asbmb.org/annualmeeting/
[28] https://www.sfn.org/

Conclusion

Recent advances in data generation and data analysis have brought biomedicine to the cusp of a new world of ML-BioMed. The computational and biomedical communities are poised to jointly drive transformative progress in biomedical research -- leading to new insights into how all living systems work -- and in care delivery -- leading to improvements in the health of all humans and all communities.

The NIH is well positioned to accelerate that progress, by supporting the three complementary areas of **data** to fuel the analysis engines, **ethics** to always be steering in accordance with our highest values, and **people** to select and drive projects forward. The eight recommendations in this report suggest specific ways to propel progress. We look forward to seeing the results unfold.

Acknowledgements

The co-chairs wish to acknowledge, with thanks, the outstanding contributions of each of the committee members. Very special thanks go to Jessica Mazerik, Ph.D., for her extraordinary effort in coordinating this activity and for her contributions to preparing the report. The group is thankful for George Santangelo, Ph.D, and Rebecca Meseroll, Ph.D., from the Office of Portfolio Analysis, NIH, and our colleagues at the U.S. Food and Drug Administration, Matthew Diamond, M.D., Ph.D., Donna Mendrick, Ph.D., Robert Ochs, Ph.D., Nicholas Petrick, Ph.D., Berkman Sahiner, Ph.D. for sharing their time and expertise to brief our group. Thanks also go to Nicole Garbarini, Ph.D. for her help during the face-to-face meetings. We also thank Courtney Coombes, Ph.D., Susan Gregurick, Ph.D. and Tara Schwetz, Ph.D., for their participation and support during the many conference calls.