National Advisory Council for Human Genome Research

September 18, 2023

Concept Clearance for RFA

ML/AI Tools to Advance Genomic Translational Research

Purpose:
NHGRI proposes a new initiative to spur the development of novel Machine Learning/Artificial Intelligence (ML/AI)[a] tools to explore their potential for advancing genomic translational[b] research. Specifically, this initiative aims to uncover novel relationships between genotypes and phenotypes through the learning and classification capabilities of ML/AI to model pleiotropy and variable penetrance in pathogenic variants. The ultimate goal is to create a strong foundation for both the development and validation of ML/AI tools and a systematic approach for assessing the Ethical, Legal, and Social Implications (ELSI) of incorporating and utilizing these tools in clinical decision making. This initiative includes two RFAs: one RFA to establish a consortium of sites charged with the development of generalizable and robust ML/AI tools, and their validation in genomic translational research settings that mimic realistic clinical decision-making scenarios, in accordance with a shared and agreed upon ELSI framework; and a second RFA to establish a Coordinating Center (CC) for the consortium.

Background:
ML/AI techniques have successfully identified novel patterns in non-genomic clinical datasets in fields such as radiology, cardiovascular medicine etc., and have received FDA approval[3]. Yet, the application of ML/AI in translating genomic research findings into clinical applications remains largely untapped[4]. The need for increased research in this area was highlighted in NHGRI-convened workshops (Genomic Medicine XIII[5], NHGRI Machine Learning In Genomics Workshop: Tools, Resources, Clinical Applications and Ethics[6]). These research communities and the NHGRI 2020 Strategic Vision[7] also emphasized the need to establish new collaborations to enable a concerted effort to develop and validate such tools.

The potential for using ML/AI tools in genomics is facilitated by the vast amounts of multimodal data made increasingly available due to advances in interoperability standards and policies to encourage sharing data in a FAIR[8] manner, which is necessary for effectively training ML/AI tools. Multimodal datasets include genomics, multi-omics, phenotypic, Social Determinants of Health (SDOH), and other ancillary data from various sources like EHRs, clinical sequencing and genetic testing, and Internet of Things (IoT) sensors, in addition to published literature. These tools harness the power of high-performance or cloud computing and are adept at processing large multimodal datasets. The potential to utilize datasets not traditionally associated with health and disease can pave the way to shed new light on the interplay between genomic and non-genomic factors to help identify novel genotype-phenotype associations. Furthermore, the capability of such tools to learn from new data,

---

[a] Artificial intelligence is the capability of a computer system to mimic human cognitive functions such as learning and problem-solving. Machine learning (ML) is an application of Artificial Intelligence (AI) where mathematical models of data are used to help a computer learn without direct instruction[1].

[b] Translational research fosters multidirectional integration of basic research, patient-oriented research, and population-based research, with the long-term aim of improving the health of the public[2].

knowledge, and methods can increase applicability for potential clinical decision making in the long term.

While there is growing understanding of the potential for these ML/AI approaches to significantly transform healthcare, there is increasing awareness that the ethical challenges posed by their use will need to be addressed concomitant with their development[9]. There are concerns about incorrect diagnoses due to biases in data and algorithms, use of data without explicit consent, risk of patient confidentiality, and unwarranted reliance on ML/AI tools, all of which can have detrimental societal impacts. The community is recognizing the need to develop resources to address the ELSI of integrating ML/AI approaches into patient care. Examples of such resources include Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis[10] (TRIPOD), MI-CLAIM[11], recommendations for adaptive CDS[12], and the recent Coalition for Health AI (CHAI) blueprint[13].

The time is opportune to assess whether ML/AI tools developed using existing datasets can be used to advance genomic translational research. The intent of the effort is to demonstrate the feasibility and impact that these tools may have on clinical decision making, with an initial focus on improving our understanding of pleiotropy and variable penetrance. Importantly, we can build on existing ELSI-relevant resources to catalyze the development of a framework for ML/AI tool development.

Proposed Scope and Objectives:
This initiative includes two RFAs:

i. The first RFA will establish a consortium of tool development sites charged with the development and cross-validation of ML/AI tools to model pleiotropy and variable penetrance of well-characterized pathogenic variants of genes (such as those in the American College of Medical Genetics and Genomics (ACMG) gene list[14]) using multimodal datasets and in accordance with a consortium-agreed-upon ELSI framework. Multimodal datasets could include genomics and multi-omics, phenotypic, SDOH, and other ancillary data from various sources like EHRs, clinical sequencing and genetic testing facilities, and Internet of Things (IoT) sensors, in addition to published literature.

ii. The second RFA will support a Coordinating Center (CC) for the consortium.

The first RFA is planned to include two phases described below.

<u>Design Phase (Years 1-2; UG3 funding mechanism)</u>
During the Design Phase, the tool development sites (with the help of the CC) will work collaboratively to select and agree upon the following aspects: the disease/s, genes, and associated pathogenic variants that the consortium will focus on; the end-points and outputs of the ML/AI tools and a consortium-wide validation plan to test accuracy, performance, and generalizability of the tools by clinical research physicians at each site. In this Phase, the sites will also prepare the relevant datasets representative of the patient populations affected by the disease/s of interest for tool training and make the datasets accessible to each other as appropriate for cross validation purposes. The consortium will also begin to formulate a draft ELSI framework including an initial set of best practices to address the ELSI of ML/AI tool integration into clinical decision making, and identify assessment criteria for trustworthiness, interpretability, and usability of the tools.

Towards the end of the Design Phase, NHGRI will convene an evaluation panel to assess the progress and accomplishments made by the consortium. The assessment will be used by

NHGRI to decide whether the sites and the CC can proceed to the ML/AI Tool Development and Validation Phase.

ML/AI Tool Development and Validation Phase (Years 3-5; UH3 funding mechanism)
During the ML/AI Tool Development and Validation Phase, the tool development sites (with the help of the CC) will develop the ML/AI tools in accordance with the plans developed in the Design Phase and the ELSI framework, implement the agreed-upon cross-validation plan, and collaboratively refine the ELSI framework. It is anticipated that at the end of this Phase, the generated resources, including the cross-validated ML/AI tools and models, the ML/AI-ready datasets, the ELSI framework and best practices generated by the consortium, and lessons-learned will be broadly disseminated according to the FAIR principles and NIH Data Management and Sharing Policy[15].

ML/AI Tool Development Sites
The sites are anticipated to include multi-disciplinary teams of experts including ML/AI tool developers, software engineers, clinical research physicians for tool development and testing, and ethicists/social scientists to develop the ELSI-framework. A demonstrated ability of the sites' teams to collaborate with other groups in large research consortium projects will be required.

Coordination center (2+3 years; U01 funding mechanism)
The CC, supported by a second RFA, will be responsible for coordinating and facilitating the consortium activities and logistics. The U01 awards to the CC for years 3-5 will be contingent upon the UH3 awards being made to the tool development sites. If the UH3 awards are not made to the sites, the CC U01 award will end concurrently with the UG3 award.

Relationship to Ongoing Activities:
This is NHGRI's first initiative designed to establish the foundational technical and methodological framework for the creation and validation of machine learning and artificial intelligence tools for genomic translational research within an Ethical, Legal, and Social Implications (ELSI) context. This proposed initiative is complementary to and could potentially leverage resources such as harmonized datasets generated by other NHGRI initiatives for example, the eMERGE[16], PRIMED[17], and GREGoR[18] for training the ML/AI models. Other publicly accessible multimodal datasets such as TOPMed[19], AllofUS[20], the UK BioBank[21] and Bridge2AI[22] can similarly be leveraged. This initiative can also leverage ELSI-relevant resources resulting from the Artificial Intelligence/Machine Learning Consortium to Advance Health Equity and Researcher Diversity (AIM-AHEAD) program established to enable mutually beneficial, coordinated, and trusted partnerships with the goal of equity and fairness.

Mechanism of Support/Funds Anticipated:
NHGRI will commit approximately $6M/year total cost for 5 years, beginning in FY2025 for a total of $30M. This includes:
- 3-4 ML/AI Tool Development Sites (UG3/UH3; $4.8M/year total cost)
- 1 CC (U01: $1.2M/year total cost)

Co-funding will be sought from other NIH Institutes and Offices.