Expressions of Interest: Machine Learning Datasets for Better Healthcare Outcomes

Lacuna Fund: Our Voice on Data

16 June 2021

Table of Contents

1 – Introduction

Purpose and Goals of the Fund

Lacuna Fund supports the creation, expansion, and maintenance of equitable training and evaluation datasets that enable the robust application of machine learning tools of high social value.

The Fund aims to:

- Disburse funds to institutions to create, expand, and/or maintain datasets that fill gaps and reduce bias in labeled data used for machine learning.
- Make it possible for underserved populations to take advantage of advances offered by AI.
- Deepen understanding by the machine learning and philanthropy communities of how to most effectively and efficiently fund the development and maintenance of equitably labeled datasets.

Principles of the Fund

The following principles will guide the governance and operations of Lacuna Fund.

- **Accessibility** – Lacuna Fund is committed to ensuring that the datasets created through its funding are accessible to and benefit underserved communities in service of the goals outlined above. Datasets and related intellectual property will utilize appropriate open data licensing to maximize responsible downstream use. (see IP policy for additional details.)
- **Equity** – Lacuna Fund aims to make AI more equitable by supporting datasets that are created by and responsive to the needs of those with underrepresented identities globally. These datasets should not create or reinforce bias (e.g., they should be gender inclusive and representative of people of color globally), nor should they support systems or technologies that create harm.
- **Ethics –** Lacuna Fund will fund data collection in a manner consistent with ethical labor standards and require recipients to specify steps they will take to protect privacy and prevent harm in the collection, licensing, and use of datasets created with grant funds.
- **Participatory Approach –** Lacuna Fund strives to meet the needs of affected stakeholders by encouraging the leadership or strong engagement of local experts, beneficiaries, and end users in the governance of the Fund and in supported projects. The Fund will consider participation in a manner consistent with our principles on equity and ethics.
- **Quality** – Data generated by Lacuna-funded efforts should be of high quality, enabling beneficial applications in research, communities, and industry.
- **Transformational Impact** – Lacuna Fund aims to unlock the advances offered by AI for poor and underserved communities by funding datasets that address fundamental gaps in AI.

Philosophy of Grantmaking

Lacuna Fund values a collaborative and locally driven approach to data creation, expansion, and maintenance. We recognize that the continued usefulness and maintenance of open data derives from a community invested in that data.

Lacuna Fund hopes to fund datasets that contribute to multiple applications of high social value, whether through research, commercial innovation, or improved public sector services. While "Section 3: Purpose and Need" sets out needs identified by the Technical Advisory Panel (TAP), Lacuna Fund welcomes novel ideas within the domain area that have a clearly articulated benefit aligned with the selection criteria listed below.

Lacuna Fund is supported by The Rockefeller Foundation, Google.org, and Canada's International Development Research Centre. Additionally, this call regarding equitably labeled datasets for better healthcare outcomes is supported by the Wellcome Trust and The Gordon and Betty Moore Foundation.

2- Overview

Organizational Eligibility

Lacuna Fund aims to make its funding accessible to as many organizations as possible in the AI for social good space and cultivate capacity and emerging organizations in the field.

To be eligible for funding, organizations must:

- Be either a non-profit entity, research institution, for-profit social enterprise, or a team of such organizations. Individuals must apply through an institutional sponsor. Partnerships are strongly encouraged; however, only the lead applicant will receive funds.
- Have a mission supporting societal good, broadly defined.
- Are headquartered in or have a substantial partnership in the country or region where data will be collected.
- Have all necessary national or other approvals to conduct proposed research, as well as data use agreements or plans to secure them. The approval process may be conducted in parallel with grant application, if necessary. Approval costs, if any, are the responsibility of the applicant.
- Have the technical capacity – or the ability to build this capacity through a partnership described in the EOI - to conduct dataset labeling, creation, aggregation, expansion, and/or maintenance, including the ability to apply best practice and established standards in the specific domain (e.g. healthcare outcomes) to allow high quality AI/ML analytics to be performed by multiple entities.

Selection Process and Evaluation Criteria

Lacuna Fund seeks to hear from organizations that are interested in responding to a call for proposals to unlock, create, aggregate, and/or improve labeled datasets that can support more equitable healthcare outcomes. Lacuna Fund and its partners will perform an initial screen of Expressions of Interest (EOI) for organizational eligibility and feasibility of the original concept presented in the EOI. Following the initial screen, a Technical Advisory Panel of domain experts, data users, and stakeholders will evaluate the EOIs based on the selection criteria outlined below. Lacuna Fund will then invite a small number of organizations to provide a full proposal based on the EOI submission. Technical Advisory Panel members may not submit an EOI or a proposal in response to an RFP for which they are a reviewer (see Lacuna Fund's Conflict of Interest Policy).

The Technical Advisory Panel for this call will assess EOIs to determine a short list of organizations that will then be invited to provide full proposals for funding. Selections for the short list of invited applicants will be based on the degree to which they meet the following criteria:

- **Quality –** The organization or team proposing the project demonstrates interdisciplinary collaboration between qualified experts in health, research, machine learning, and data management. Partnerships between well-resourced healthcare systems with strong research capacity and systems serving marginalized populations are encouraged.
- **Transformational Impact –** Datasets should improve machine learning and enable a demonstratable social benefit to underserved populations. Examples include, but are not limited to: a) labeling, cleaning, collecting validation data for, or pooling existing datasets to unlock additional value or ensure greater accuracy in the existing dataset; b) creating a new, high-value labeled dataset for an underserved population or problem related to healthcare outcomes; c) making an existing dataset more representative and inclusive of race, gender, ethnicity, ability, etc.; or d) linking existing clinical datasets with data on social determinants of health to create more robust, information datasets.
- **Equity –** There is a compelling theory of change demonstrating how the dataset will be applied to help vulnerable and underserved communities.
- **Participatory Approach –** Datasets should center the needs of affected communities and work with partners to identify community benefit. Specifically, projects should engage patients, clinicians, and community members in data governance decisions, (e.g., what data is curated and how it is used). If the dataset has a geographical scope, the team is predominantly located in the respective area and/or sustains close ties to local actors to ensure sustained maintenance and usage of the dataset by the local community.
- **Ethics –** The project is able to pass an ethical screen (e.g., an institutional review board) that probes: a) privacy concerns, b) potential for downstream misuse c) possible discrimination vectors (e.g., gender), and d) fair and equitable working conditions, if paid labelers are involved in the project.
- **Efficiency –** The proponent has considered existing datasets and proposes to use effective data collection and labeling techniques and tools to speed the collection, cleaning, and sharing of data.
- **Feasibility –** The project is feasible in relation to the budget and scope of work proposed. While the EOI is not expected to include a detailed budget, we do expect EOIs to include a broad estimate of the overall budget expected to complete the project.

- **Accessibility -** The dataset will be made widely accessible under open source licensing pursuant to Lacuna Fund's IP Policy. If this is not possible, a compelling case is made for more restrictive licensing in order to protect privacy or prevent harm, along with a mechanism for providing access under the proposed licensing.
- **Sustainability –** The project has a plan to ensure sustainability and future maintenance of the dataset, e.g., by a dedicated community or a pool of interested parties (for-profit and/or not-for-profit) and a robust governance model for the open dataset.

Timeline

EOI Call Posted Publicly on Lacuna Fund Website	**16 June 2021**
Question and Answer Deadline Please submit questions to secretariat@lacunafund.org	**29 June 2021**
Answers Posted	**7 July 2021**
Expressions of Interest Due	**21 July 2021**

Question and Answer Period: All questions related to the EOI should be submitted to secretariat@lacunafund.org with "Equity & Health 2021 Question" in the subject line. Questions submitted by 29 June will be de-identified and answered publicly by 7 July on the Lacuna Fund website in a document posted on the "Apply" page.

3 – Purpose and Need

Purpose

The purpose of this call for EOI is to identify projects to submit full proposals to develop open and accessible training and evaluation datasets for machine learning (ML) applications that address inequities in healthcare outcomes. Greater implementation of ML models has the potential to provide better information to doctors at the point of patient care. By providing more data and information to clinicians, they can make better decisions about patient diagnoses and treatment options, while understanding the possible outcomes and cost for each one. The value of machine learning in healthcare is its ability to process huge datasets beyond the scope of human capability, and then reliably convert

analysis of that data into clinical insights that aid physicians in planning and providing care, ultimately leading to better outcomes, lower costs of care, and increased patient satisfaction.[1]

A major concern with ML implementation has emerged regarding bias in the datasets that inform algorithms on which these systems are based. Healthcare datasets often lack broad representation across demographic and socioeconomic groups. Where this representation is addressed, the data may also not be totally accurate, resulting on incorrect understanding of differences across those groups. Finally, current data may miss important socioeconomic, environmental, and other data that can determine health outcomes. Improvement of these datasets along these lines, thus, may improve machine learning models and health outcomes.[2]

Need

The Lacuna Fund seeks Expressions of Interest (EOIs) from qualified organizations to develop open and accessible training and evaluation datasets for ML applications that address inequities in healthcare outcomes. The scope of the effort is as follows:

Geographic Scope: Within the U.S. and in low- and middle-income countries (LMICs), with 1-2 proposals and between 30-70% of the pool to be funded in each of these geographic areas depending on the number of datasets are funded.

Substantive Scope: Projects should focus on datasets that can be used to address a health disparity in one or more of the following areas: cancer, infectious disease, or chronic disease. Additionally, applicants may make a case for why a dataset in another area could significantly reduce inequities. In either case, applicants are encouraged to consider if the dataset they are proposing could be used to address other problems, in addition to the intended use case.

Diversity Scope: Our goal is to support the creation, augmentation, or aggregation of datasets that are representative of affected populations and are therefore less biased and more likely to lead to equitable health outcomes. Given historic inequities related to race in the U.S., we are interested in datasets that could help reduce racial disparities in healthcare outcomes in the U.S., and datasets that can mitigate inequities in healthcare outcomes related to identity in LMICs (e.g. ethnicity, tribal affiliation, gender, etc.) In addition, we know that lack of diversity in terms of sexuality, age, ability, geographic location, and even type of care setting (e.g., community health system vs a large academic medical system, inpatient vs outpatient), can make a dataset less representative and lead to unfair outcomes for subsets of a population. We encourage applicants to consider taking an intersectional approach by including data for multiple underserved groups, or to explain why a certain type of diversity is important to ensure equitable outcomes for a specific use case.

Additional considerations include the following:

---

[1] Corbett, Ed. “Real-World Benefits of Machine Learning in Healthcare.” Health Catalyst, 19 Nov. 2020, www.healthcatalyst.com/clinical-applications-of-machine-learning-in-healthcare.

[2] Wawira Gichoya J, McCoy LG, Celi LA*, et al*. “Equity in essence: a call for operationalising fairness in machine learning for healthcare”. *BMJ Health & Care Informatics* 2021;**28:**e100289. doi:10.1136/bmjhci-2020-100289

Existing vs. new datasets: We see great value in unlocking, augmenting, or aggregating existing datasets and is also open to proposals to create new datasets such as benchmark datasets to test the generalizability of algorithms and differences in quality of prediction for underserved populations. Most critically, we want to ensure the underlying dataset is not fundamentally flawed so as to avoid perpetuating bias. All of the approaches described below are of interest:

- Pooling existing data from healthcare systems, health insurance companies, or health data intermediaries to make it accessible to researchers to be able to reveal and correct algorithmic bias (e.g., MIMIC, PhysioNet).
- Filling gaps/making existing datasets more representative (e.g., images of skin cancer, lung x-rays to detect COVID, etc.).
- Linking existing clinical datasets with data on social determinants of health to create more robust, informative datasets.
- Cleaning up existing datasets to ensure accuracy in data about race, ethnicity, gender, disability, etc.

Type of Dataset: Successful applicants will propose creation, expansion, or aggregation of a dataset(s) that can be used to address multiple research questions and correct a bias or disparity.

4 – Expression of Interest Information

Expressions of Intent will only be accepted through the application portal available at www.lacunafund.org/apply. A description of application questions is available below for information only. *Please limit your expression of interest to 4 pages not including references, with 2.5 cm margins and a minimum of 11-point font. Appendices or proposal narrative material beyond 4 pages may not be reviewed.*

The following information is required:

- Applicant Information
  - Name of Participating Organizations
  - Descriptions and Qualifications of Participating Organizations – Provide brief background on the mission of participating organizations, services provided, and constituencies served; how they satisfy the eligibility criteria articulated above; and the applicant's unique qualifications to undertake the proposed work, including experience developing and sharing health/healthcare datasets.
- Summary of the EOI Purpose
  - Problem Identification and Proposed Solutions – Describe the health disparity or inequity, who is most affected by this problem, and the proposed solution (dataset labeling, aggregation, creation, augmentation, or maintenance).
  - General Methodology – Provide a brief overview of the proposed steps (and key assumptions) for developing and implementing the solution set. Explain permissions in place or steps you will take to secure data use agreements.
  - Expected Outcomes and Benefits Following Project Implementation - Explain how the proposed project will contribute to achieving the desired impact. If applicable, describe

how the products could motivate multiple and durable paths of research or commercial application.

- General Timeframe and Overall Budget for Project Implementation:
  - Timeframe – Share a broad timeframe for completion of the steps included in the General Methodology above, including total number of months required for completion.
  - Budget – Provide a broad overview of the expected budget for completion of the steps included in the General Methodology above. Budgets should be submitted in US Dollars. The total pool available is approximately $800,000 USD. We are anticipating proposed budgets in the range of $10k – 100k for small to medium-sized projects and up to $500k for large, complex projects. We anticipate being able to fund 1-2 large projects or a larger number of smaller projects.

Thank you for your interest in Lacuna Fund and your efforts to make machine learning applications in healthcare more equitable. We look forward to reviewing your submission.