Data Collaboration Center
Critical Path Institute
1730 E River Road
Tucson, AZ 85718-5893
www.c-path.org

Feb 15, 2021

REQUEST FOR PROPOSAL

Request for Creation of non-identifiable, synthetic patient-level data from electronic health records with strong statistical fidelity to raw data sources and proof of validation of statistical fidelity

Data Collaboration Center's (DCC's) Data Science team is seeking a proposal, including timelines and budget, to provide a synthetic version of patient-level data from electronic health records (EHR's) that will represent a population of patients with rare diseases. The tasks included in the scope of this proposal include:

- Access one or many sources of electronic health records that may include PHI and associated security and privacy concerns
- Create a non-identifiable, synthetic version of data from these sources which have statistical fidelity with the sources
- All code and processes to create the synthetic data must be shareable, at minimum, with C-Path and recognized regulatory agencies (e.g. FDA, EMA). Open source code is encouraged.
- Validation testing and results to provide evidence of statistical fidelity with the source data population

This work will support and accelerate rare disease characterization with the goal of advancing therapy development across rare diseases through the creation of synthetic patient-level electronic health records. The synthetic data should be modeled after real source data with a preference for source data from a healthcare system with a diversity of clinical care settings (large hospitals to local community clinics) and a very large patient population. The source data should be representative of patients across diverse backgrounds with an emphasis on rare diseases and social determinants of health. The synthetic data should represent lifetime longitudinal trends within a patient's record. The data should reflect coding practices, treatment trends, clinical practice guidelines, and prescribing practices with an emphasis on recent standard of care.

This Request for Proposal (RFP) describes the intended scope of work and deliverables expected. Also included is a budget template to be completed by your organization to document the required and any optional task costs associated with completing the defined scope of work and associated deliverables.

Information contained in your proposal will be evaluated by the members of the DCC and will be considered confidential.

Clarifying questions must be received no later than **February 19th, 2021;** a conference call may be convened if deemed necessary. Proposals must be received by **February 26, 2021**. Both are to be sent to:

Amanda J. Borens MS
Executive Director of Data Science
aborens@c-path.org

*** Please include "EHR RFP" in the subject line.**

RFP Provisions

The proposal is a firm offer that will be considered valid for 180 calendar days from the submission date. Please provide the contact information of the person responsible for submitting the proposal. C-Path shall not be responsible for any errors or omissions on the part of the Bidder in preparing this proposal. Bidder shall bear all costs associated with preparing this proposal.

Please prepare your proposed strategy to address the objectives and scope of the generation of synthetic patient-level data from electronic health records. You must demonstrate a knowledge base consistent with the objectives and requirements of this RFP, and describe your strategy and rationale for the solution and validation or quality assurance testing (e.g. algorithm design, validation plan and explicit description of test data sets) that will be used to confirm the adequacy of the proposed solution. All of the required elements (i.e., methods, deliverables, milestones, experience, timelines, and costs) should be clearly explained in 20 pages or less.

Expected Statement of Work

Generate and deliver synthetic patient-level electronic health care records

1. Generate synthetic data representative population of healthy and diseased patients with a focus on rare diseases, examples of rare diseases are given Appendix A.

   Additional specifications for the generated data include:
   i. Source Data Store:
      1. Preference will be given for data modeled after source data from a healthcare system with a diversity of clinical care settings (large hospitals to local community clinics)
      2. The source EHR used by the healthcare system should be well adopted and avoid data modeled from recent transitions between EHR vendors
      3. Develop computable phenotypes to identify and model data of rare diseases

ii. Population Diversity and Social Determinants of Health:
    1. Included source data should be representative of patients across backgrounds including sex, race/ethnicity, income and education levels, health care insurance and access, and geographical regions with preference towards datasets with international representation
iii. Included Data Elements:
    1. The minimum included data elements should include patient demographics, longitudinal observations and repeated measures, clinical encounters (inpatient and outpatient), conditions and observations, procedures, medications, laboratory measurements (a preference will be given to datasets with genetic tests), radiographical and neurophysiological tests (e.g. EKG, PSG) findings and reports
    2. Optional but desired data elements would include mother/child linkage for neonatal diseases, family history, provider notes, administrative data (E&M codes, billing data, etc.), patient reported outcomes, genomics, workflow data, specimens
    3. A preference will be given for datasets that include quantitative measures
    4. All data elements should be represented in an industry standard using established medical and coding terminologies
iv. Temporal Representation and Historical Trends
    1. Data should include longitudinal trends from within a patient record with as close to lifetime representation as possible
    2. The data should reflect coding practices, treatment trends, clinical practice guidelines, and prescribing practices since at least 2010 with an emphasis on recent standard of care

Deliverables

1. Documentation of source data, metadata, data provenance, and clinical phenotypes
2. Synthetic EHR data delivered in one or more standard industry format such as tabular delimited files, a common data model (e.g. Observational Medical Outcomes Partnership CDM, PCORnet CDM, etc.), HL7® Fast Healthcare Interoperability Resources, or another comparable data format
3. If applicable, data dictionaries and associated metadata to define the data files should also be provided
4. Optional: Sharing of open source code

Develop and Deliver Data Validation and Statistical Fidelity

1. Develop transparent statistical analysis and validation
2. Validate the generated synthetic data against source data using quantitative models that demonstrate statistical fidelity
3. Detail quality control measures applied when working between source data and creation of synthetic data.
4. Document and deliver statistical testing and reports

Deliverables

1. Statistical Analysis Plans and Statistical Reports
2. Optional: Sharing of open-source code to allow others to generate synthetic data

Timelines and Overall Project Management

Bidder responding to this RFP is required to provide a detailed timeline including anticipated completion dates for the deliverables and milestones described above. The proposal should include details on a project kick off meeting and regular reporting and project updates to to the DCC Data Science Team. Projected timelines for completion of the project will be an element of the proposal evaluation criteria.

Please format the timeline to show duration of activities based on an anticipated start date of April 1, 2021.

Organizational Experience

Please define your organization's qualifications. Describe previous experience with synthetic electronic health record development and statistical validation. Define prior or ongoing involvement in generating synthetic rare disease data, highlighting the use of real-world evidence for regulatory decision making. Share any unique insight or relationships that would facilitate the identification and generation of data for the target population.

Source Data and Synthetic Data Description

Include a description of the health system(s) used to generate the source data and techniques used to identify patients with rare diseases. Details should include the context of the health system, the time since EHR adoption, and any specialty clinics for rare diseases and specifics on the methodology to determine the clinical phenotypes from structured EHR data. Include a description of generated data disease representation, data elements, data formats, and data models.

Key Personnel

Describe the roles and responsibilities of key personnel on this proposed project. Please include brief descriptions (300 words or less) of all key personnel who will be involved in the project. If necessary, CVs of key personnel and a list of their publications relevant to lung cancer and/or COA development should be provided as an appendix.

Bidder Organization

Please provide a brief description (300 words or less) of your overall organization (e.g., size, locations, and primary business units).

Costs

Costs are to be broken out and identified per task and deliverable. Third party expenses (e.g., subcontractor, honoraria, out-of-pocket, and travel) must be identified and totaled separately from your direct service costs in relation to all tasks of this project. Management and contracting with third parties, including consultants and advisory panel members, is the responsibility of the contractor. Please provide your proposed budget using the templates for direct, pass-through, and optional task costs provided below. The budget template is attached in Excel format to assist in creating the table shown below. Also, please provide proposed payment terms.

BUDGET TEMPLATE

Note: *Below entries are required at a minimum; additional details will be appreciated.*

Direct Costs by Deliverable

Task Name	Time to Completion from Kick-off (in weeks)	Total Hours for all Staff	Blended Hourly Rate	Total
Personnel	N/A			$
Synthetic Data Creation				$
Data Validation and Statistical Fidelity Development and Reporting				$
Statistical Analysis Plans and Reporting Development				$
Additional Tasks or Direct costs (Please specify any additional task costs)				$
				$
				$
			Total Direct Costs	$

PAYMENT TERMS:

Proposed payment terms, tied to specific deliverables or defined milestones for direct costs, are to be provided in the proposal submitted to the Data Collaboration Center (DCC).

Appendix A: Examples of Rare Diseases

Disease Category	Example Diseases
Neurodegenerative Diseases	Huntington's Disease, Atypical Parkinsonian Disorders (e.g. Progressive Supranuclear Palsy[+], corticobasal syndrome, dementia with Lewy bodies)
Neuromuscular diseases	Duchenne Muscular Dystrophy[+], Spinal muscular atrophy, ALS, other muscular dystrophies, and genetic myopathies (e.g. Ryr1, GNE myopathy)
Ataxias	Friedreich's Ataxia, Spinocerebellar Ataxias, Other autosomal recessive ataxias (e.g. ataxia telangectasia)
Neurological Dysfunction	Angelman Syndrome, Prader-Willi Syndrome
Neonatal Disorders	Bronchopulmonary Dysplasia
Blood Disorders	Sickle Cell Disease[+]
Metabolic disorders	Mucopolysaccharidosis , Niemann Pick, Phenylketonuria

[+] preference will be given to these diseases