Administrative Supplements for P30 Cancer Centers Support Grants (CCSG) to Stimulate Research Using Large Language Models to Assist in Cancer treatment Data Extraction From Clinical Reports or Diagnostic Data From Pathology Reports

Background:

A hallmark of NCI-designated cancer centers is that the science conducted at these institutions meets rigorous standards for transdisciplinary science and state-of-the-art research. The resources and computing environments at these institutions encourage multidisciplinary teams to creatively solve hard problems and ensure solutions are implemented at a relatively fast pace. The NCI would like to leverage these scientifically rich environments and invites teams of biomedical informaticians, data scientists, clinical researchers, and others to use Large (or Medium) Language Models (LLM) to assist in cancer diagnosis, treatment, and other relevant data extraction from unstructured clinical reports.

Large/Medium Language Models or LLMs are generative language models made up of billions of parameters. They are trained on large quantities of unlabeled text using either self-supervised learning or a semi-supervised learning model. The use of LLMs and other generative technologies has made it possible for retrospective treatment and dosage data to contribute to models that aid in modern treatment decisions. Specifically, Generative Pre-trained Transformer (GPT) technology can extract treatment and dosage data from unstructured clinical records. Alternatively, LLMs could be used to extract cancer-relevant information from pathology reports. This extracted information can be analyzed alongside diagnosis, survival, outcome, and quality of life data to contribute to a cancer patient's journey. Solutions that focus on training medium language models for these tasks would be accepted. Through the NCI thesaurus, the NCI maintains a list of common cancer treatments.

1- Large Language Models for Cancer Treatment

Investigators will develop an LLM tool to accurately extract the treatment regimens (i.e. chemo, radiation, targeted therapy, combination therapy, or other medication based) at your Cancer Center and affiliated healthcare facilities. Drug regimen is a required modality. Tools that will map the extracted medications to the NCI thesaurus from clinical records will be given preference. The ideal tool will produce a line of therapies that includes the date, dosage, and drug a cancer patient received. Furthermore, successful approaches will include validation of the extracted components.

2- Large Language Models for Pathology Reports

Diagnostic pathology reports contain vital information for the tumor studied including staging, histology, laterality, anatomical site, behavior, molecular markers, etc. Although some Centers utilize the College of American Pathology provided synoptic reports, many components of the reports are still considered semi/unstructured for programmatic extraction of such useful data elements in scale. NCI is looking for LLM-based approaches to extract at least two commonly used data elements from the diagnostic report (e.g. histology, site, etc.) at your Cancer Center and affiliated healthcare facilities. Preference will be given to those centers with preliminary data on the LLM approaches. Comparison with at least one BERT-

like approach will be expected at the end of the project period. This supplement does not require the participating teams to share clinical or pathology data with NCI.

Cancer Centers can submit their approaches for either of the tasks mentioned. Teams are expected to develop AI-based Natural Language Processing models or repurpose an existing/open-source model that can extract requested information from unstructured notes. Furthermore, investigators are asked to contribute their models to NIH supported/suggested public repositories as well as disseminate the approach via publication(s) and/or webinars coordinated with NCI. At the end of the supplement period, meritorious Teams will be invited to demonstrate their tool sets on a *de novo* cancer data set.

Eligibility and Budget

A. Supplement applications will be due July 1st, 2023.
B. Awards will be made in September 2023.
C. Supplement budget requests may not exceed $300,000 in total costs.
D. Supplements are 1 year in duration.
E. This opportunity is open to all NCI-Designated Cancer Centers.
F. Cancer Centers whose P30 CCSG will be on a merit extension at the time the award is made in September of FY23 are eligible to apply.
G. Cancer Centers whose P30 CCSG will be on a cost extension at the time the award is made in September of FY23 are not eligible.
H. Any proposal that cannot be completed within the 1-year time frame will be viewed as non-responsive.
I. Allowable costs include funding for the PI and his/ her team and the costs for supplies, including compute time. Large pieces of equipment cannot be purchased through this supplement.
J. Teams that have already submitted an NIH grant application like the Projects described above should not resubmit a similar application through this supplement mechanism.
K. Only one supplement request per center will be considered.

Application Submission Format

Applications must be submitted electronically via eRA Commons to the parent award (P30) using PA-20-272 "Administrative Supplements to Existing Grants and Cooperative Agreements (Parent Admin Supplement)" on or before **July 14, 2023**. Your submission should follow the instructions in the funding opportunity announcement, including the following:

Research Plan (5 pages) please include the following elements:

A. The title of the supplement in parenthesis (LLMs for Unstructured Data Extraction)

B. The research proposal should address questions that can be tested by using EHRs collected from patients at the Cancer Center and affiliated healthcare facilities.
C. Proposed research may include computational costs or costs for obtaining records collected from patients at the Cancer Center and affiliated healthcare facilities.
D. Description of the background, preliminary data (if available), relevant cancer center infrastructure, data sources, research teams, etc.
E. Analyses and models that include a diverse population across the spectrum of age, sex, and race are encouraged.
F. Leadership of projects by junior or mid-level investigators is encouraged.
G. Please submit a separate SF424 biosketch form(s) for the Project Leader.

1. Detailed budget and justification

A. Please use the SF424 forms to document your funding request.
B. Appendices and attachments are not allowed.
C. For tracking purposes, please notify NCIClinicalInformatics@nih.gov when you submit your application (but please do not send the application itself).

Evaluation Criteria

Administrative supplements do not receive peer review. Instead, NCI staff will evaluate each supplement request to determine its overall merit. Supplements will be reviewed for quality and responsiveness to application criteria outlined above in the **RESEARCH PLAN SECTION** and **PURPOSE AND GOALS SECTION** of this Funding Announcement.

Awards

The number of awards is contingent upon NIH appropriations and the submission of a sufficient number of meritorious applications.

Reporting Requirements

As part of the annual progress report of the parent NCI Cancer Center Support Grants include information on what has been accomplished via the administrative supplement during the funding period.

Questions

For inquiries about the scientific objectives and goal of LLMs for cancer treatment administrative supplement, please email NCIClinicalInformatics@nih.gov.

Pre-Submission Large Language Models Informational Webinar Material:

https://cbiit.webex.com/cbiit/j.php?MTID=m907d90bb1f91973c596109277ff4355f
Wednesday, June 7, 2023, 10:00 AM | 1 hour | (UTC-04:00) Eastern Time (US & Canada)
Meeting number: 2309 741 9851
Password: PxuAZAV@236

Join by a video system
Dial 23097419851@cbiit.webex.com
You can also dial 173.243.2.68 and enter your meeting number.

Join by phone
1-650-479-3207 Call-in toll number (US/Canada)

Access code: 230 974 19851