PHASE I PROPOSAL CHECKLIST

- ☐ Technical Element
  - ☐ Proposal Cover Sheet (Appendix A)
  - ☐ Table of Contents
  - ☐ Abstract of the Research Plan (Appendix B)
  - ☐ Content of the Technical Element
    - ☐ Identification and Significance of the Problem or Opportunity
    - ☐ Technical Objectives
    - ☐ Detailed Approach and Methodology
    - ☐ Related Research or R&D
    - ☐ Relationship with Future R&D
    - ☐ Innovation
    - ☐ Potential Commercial Applications
    - ☐ Senior/Key Personnel and Bibliography of Directly Related Work
    - ☐ Subcontractors/Consultants
    - ☐ Multiple PI/PD Leadership Plan (NIH Only)
    - ☐ Facilities and Equipment
    - ☐ Resource Sharing Plan(s)
  - ☐ Enhancing Reproducibility through Rigor and Transparency
  - ☐ Research Involving Vertebrate Animals
  - ☐ Dual Use Research of Concern
  - ☐ Human Subjects and Clinical Trials Information Form
  - ☐ Draft Statement of Work (Appendix E)
- ☐ Human Subjects and Clinical Trials Information Form and Attachments (Appendix H.2 and, if applicable, H.3)

Appendix A

DEPARTMENT OF HEALTH AND HUMAN SERVICES

SMALL BUSINESS INNOVATION RESEARCH PROGRAM

PHASE I PROPOSAL COVER SHEET

TOPIC NO.:

PROJECT TITLE:

FAST TRACK PROPOSAL: ☐ YES ☐ NO

SUBMITTED BY (*Firm name, address, and telephone number*):

YEAR FIRM FOUNDED:

NO. OF EMPLOYEES (*Include all affiliations*):

NOTICE TO OFFERORS

The offeror organization and the principal investigator are jointly responsible for the accuracy and validity of all the administrative, fiscal, and scientific information in the proposal. Deliberate withholding, falsification, or misrepresentation of information could result in a determination of non-responsibility [see Federal Acquisition Regulation (FAR) 9.104] which would preclude an award to the offeror. In addition, sanctions such as suspension, debarment, and criminal penalties could apply.

YES	NO	CERTIFICATIONS
☐	☐	1. The above organization certifies that it is a small business concern as defined in this Solicitation.
☐	☐	2. The above organization also certifies that it is one or more of the following small business concerns as defined in FAR 2.101: ☐ 8(a) ☐ HubZone ☐ Service-Disabled Veteran-Owned ☐ Small Disadvantaged Business ☐ Woman-Owned ** Note: Capture of this information is strictly for statistical purposes.*
☐	☐	3. The above organization certifies that, if this proposal results in a contract award, more than one-half of the principal investigator's time will be spent in the employ of the firm.
☐	☐	4. The above organization and / or principal investigator(s) have submitted contract proposals or grant applications for essentially equivalent work (as defined in this Solicitation) under other federal programs, or have received other federal awards containing a significant amount of essentially equivalent work. (If YES, include information required for "**Prior, Current, or Pending Support of Similar Proposals or Awards"** in Appendix C – Pricing Proposal, as described in the solicitation.)
☐	☐	5. If this proposal does not result in an award, is the Government permitted to disclose the title and abstract of your research project, and the name, address and telephone number of the corporate official of your firm, to organizations that may be interested in contacting you for further information or possible investment?
☐	☐	6. This proposed project involves human subjects. (See instructions in Solicitation.) Clinical Trial? ☐ Yes ☐ No Agency-Defined Phase III Clinical Trial? ☐ Yes ☐ No
☐	☐	7. This proposed project involves vertebrate animals. (See instructions in Solicitation.) If YES, identify by common names and circle primates.

NOTICE OF PROPRIETARY INFORMATION

The information identified by asterisks (*) on pages of this proposal constitutes trade secrets or information that is commercial or financial and confidential or privileged. It is furnished to the Government in confidence with the understanding that such information shall be used or disclosed only for evaluation of this proposal; provided that, if a contract is awarded as a result of or in connection with the submission of this proposal, the Government shall have the right to use or disclose information herein to the extent provided by law. This restriction does not limit the Government's right to use the information if it is obtained without restriction from another source.

PRINCIPAL INVESTIGATOR/PROJECT MANAGER	CORPORATE OFFICIAL
NAME:	NAME:
SIGNATURE:	SIGNATURE:
DATE:	DATE:
TITLE:	TITLE:
PHONE:	PHONE:
E-MAIL ADDRESS :	E-MAIL ADDRESS:

Table of Contents

ABSTRACT OF RESEARCH PLAN

NAME, ADDRESS, AND TELEPHONE NUMBER OF OFFEROR ORGANIZATION	AGENCY NAME: SOLICITATION NUMBER: TOPIC NUMBER:

TITLE OF PROJECT

KEY PERSONNEL ENGAGED ON PROJECT

Name (First, Middle, Last)	POSITION TITLE	ORGANIZATION

ABSTRACT OR RESEARCH PLAN: State the proposal's long-term objectives and specific aims, making reference to the health-relatedness of the project. Describe concisely the research design and methods for achieving these goals, and discuss the potential of the research for technological innovation. Summarize the results that are expected. Avoid summaries of past accomplishments and the use of the first person. This abstract is meant to serve as a succinct and accurate description of the proposed work when separated from the application. If the proposal is funded, this description, as is, will become public information. **Therefore, do not include proprietary/confidential information. DO NOT EXCEED 200 WORDS.**

Provide key words (8 maximum) to identify the research or technology.

Provide a brief summary of the potential commercial applications of the research.

Advancing Clinical Trial Engagement in Cancer Prevention and Treatment: Harnessing Large Language Models for Data Augmentation to Develop Multi-Faceted Tools, Enhancing Recruitment, Participant Understanding, and Retention in Diverse Populations.

Significance of the Problem

The landscape of clinical trials is rapidly evolving, with an impressive 19.1% annualized growth rate from 2017 to 2022, resulting in over 462,000 registered studies, as reported by the US National Library of Medicine [1, 2, 3]. Those participating in these trials gain potential access to cutting-edge treatments, often unavailable in standard medical practice, and play a pivotal role in advancing medical science [4, 5, 6]. Clinical trials, while serving as essential conduits for medical advancements, come with significant financial implications as well. With an expenditure of US$19 million per trial or US$41,000 per patient [7], the stakes are high to ensure each participant is appropriately matched to a given trial; however, finding an ideal candidate for a clinical trial can be a complex and time-consuming task [8]. By 2022, records from the clinicaltrials.gov database indicate that 14.5% of interventional trials faced premature termination [9]. Moreover, only 7.9% of new drug applications resulted in an actual formal registration [10]. The financial impact is particularly more severe in later-phase trials, such as Phase 2 and 3, due to their larger sample sizes and prior investments [11, 12]. Thus, optimizing trial design is key to the study's success, efficient resource use, and ethical participant involvement.

Current advances in artificial intelligence mostly focus on predicting trial outcomes [13, 14, 15, 16] rather than on clinical trial design. Research studies centered on forecasting the success of clinical trials use machine learning techniques, drawing from structured and unstructured data derived from clinical studies. Such research predominantly aims at predicting the premature termination of clinical trials by integrating data about trial attributes with unstructured information. On the other hand, the limited exploration of a collaborative-focused design of clinical trials [17, 18] still faces three crucial challenges: (1) the complexities of ensuring ethical practices in both patient involvement and data handling, (2) limited access to comprehensive data, and (3) the lack of uniform standards in data collection and interpretation [19, 20, 21]. As seen in Fig. 1, a significant challenge is found in the mismatch of two key data components (i) the diversity of terminologies and ontologies found in patients' Electronic Health Records (EHR) and (ii) the intricate inclusion and exclusion Elegibility Criteria (EC) of clinical trials.

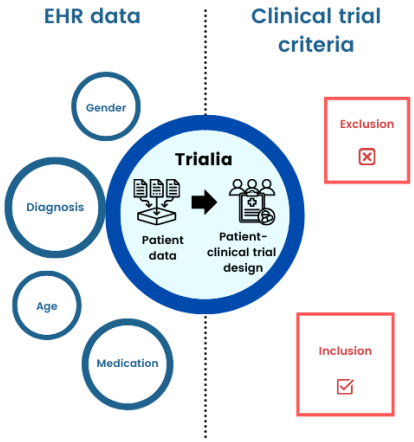

Figure 1: lAIrt aims to explore, understand, and close the current gap in clinical trial design while enhancing patient recruitment and retention.

The evolution of large language models (LLMs) [22] has ushered in a wave of applications in healthcare. These models have showcased unmatched aptitude in diverse scenarios, from answering complex patient queries [23] to composing comprehensive medical notes [24]. Yet, when steered toward the intricate process of clinical trial design, their exact potential becomes a subject of investigation. Moreover, initial observations indicate that there might be a better solution than a straightforward implementation of LLMs, revealing performance deficits and triggering concerns over patient data privacy when interfacing with LLM APIs [25]. Despite the evident promise and transformative power of LLMs across different sectors, their application in clinical trial design requires a nuanced approach. Solutions tailored for clinical trial design should facilitate cross-institutional collaboration, efficiently navigate extensive medical lexicons, and uphold ethical standards, particularly when handling sensitive patient data. To truly harness LLM's potential, a balance must be struck between optimizing its language processing capabilities while ensuring strict data protection standards. This complex task underscores the need for a collaborative effort among data scientists, medical professionals, and ethical committees to create solutions where LLMs can function efficiently while

upholding patient trust and the overarching goals of medical research.

To address the challenges of deploying LLMs in clinical trial design, the work of our team at AI POW LLC prioritizes three methods to address the challenges mentioned above: (1) proper data de-identification and the need for only a limited amount of human-labeled examples as seeds, (2) the creation appropriate prompts to guide LLMs data augmentation task and verifying that it generates reliable datasets, and (3) the standardization between EHR data and clinical trial inclusion and exclusion criteria. We want to highlight that our Phase I proposal centers on using desensitized, rather than raw, patient data to guide LLMs to enhance their reliability in zero-shot tasks while minimizing potential privacy breaches, ensuring a privacy-aware cross-institutional collaboration. At the same time, our preliminary work has shown promising results that give us an early indication that our approach can be successfully implemented in clinical trial design. Particularly, our prior work excelled in extracting structured insights from unstructured healthcare records in biological named entity recognition (NER) and relation extraction (RE). By preserving the intrinsic value of the data, we have a high level of confidence that LLMs can be properly prompt-guided to produce high patient data quality.

Finally, in the broader landscape of medical research, the effective matching of patients to clinical trials is a matter of scientific precision and resource optimization, cost-efficiency, and enhancing public trust. Utilizing LLMs for privacy-conscious data augmentation provides the comprehensive computational tools necessary to revolutionize the design of clinical trials for better patient-trial matching. Beyond ensuring the success of individual trials, our approach can significantly streamline resources and reduce unnecessary financial expenditures. In an era marked by continuous concerns over data security, this approach presents an innovative solution that maintains the integrity of patient data while harnessing the unmatched capabilities of LLMs. By striking this delicate balance, we not only enhance the efficiency of medical research but also fortify public trust, setting the stage for a more collaborative and inclusive future in clinical trials.

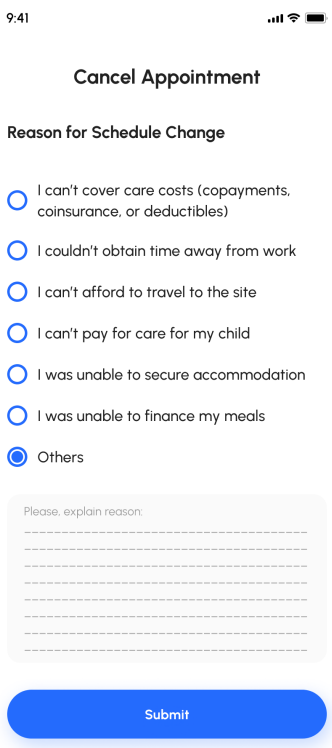

Figure 2: lAIrt, a tool to Enhance Recruitment, Participant Understanding, and Retention in Diverse Populations

Technical Objectives

In Phase I, we propose to further leverage the unique capabilities of LLMs for privacy-aware data augmentation and apply this to enhance cross-institutional clinical trial design through the following aims: In Aim 1 Privacy-Aware Augmentation Techniques (Months 0-4), our team will leverage LLMs and limited seed data samples labeled by experts to generate a comprehensive synthetic dataset and ensure the data is reliable and adherent to stringent privacy standards, allowing for cross-institutional collaboration.

In Aim 2 Patient-Clinical Trial Matching Precision (Months 4-8), we will focus on the creation of a privacy-focused classification framework based on a Memory and Highway network approach that matches patients' EHR to clinical trials' inclusion and exclusion criteria effectively and ensures the matching process is finely tuned to optimize their potential for enrollment.

In Aim 3 - Enhancing Recruitment, Participant Understanding, and Retention in Diverse Populations (Months 8-12), our focus will be on developing strategies and tools to improve the recruitment and retention of participants from diverse backgrounds in clinical trials. This will involve the creation of customized, culturally sensitive informational materials and communication methods that enhance understanding and engagement among diverse populations. Additionally, we will implement advanced analytics to identify barriers to participation and retention, thereby enabling targeted interventions. Our approach will ensure inclusivity and equitable access, enhancing the generalizability and applicability of clinical trial results across varied demographic groups. This aim not only addresses the crucial aspect of diversity in clinical trials but also contributes to the overall success and validity of the research by ensuring a representative participant pool.

At the onset of Phase I, our team will develop a model prototype that provides a hands-on tool to critically assess and validate the efficiency, accuracy, and reliability of our patient-trial matching system using six different stroke clinical trials,

namely NCT03263117, NCT03496883, NCT03545607, NCT03735979, NCT03805308, NCT03876457, using ClinicalTrials.gov as our source. Our proposed Phase I work is also tailored to give us a better understanding of a diverse population of patients when undergoing a clinical trial, setting a robust foundation for our Phase II endeavors.

Team and long-term goal. This project brings together investigators with complementary expertise in Data Mining (Costilla), Machine Learning Automation (Hu), and collaborator for model validation and clinical trial implementation (Jiang). The successful completion of our phase I work will lay the foundation for a tool that sets new standards in clinical trial design to facilitate collaboration among medical professionals to enhance their ability to connect patients to potentially life-altering treatments and drive transformative outcomes in patient care and clinical research.

Detailed Approach and Methodology

Phase I Overview. This proposal will allow our team to perform a feasibility study of patient-clinical trial matching leveraging LLMs. In this phase, we will leverage our long-term partnerships, experience, and expertise gained from developing a framework for synthetic data generation using LLMs to help in clinical text mining and clinical trial design.

Problem setting overview: Patient records (input 1) are extracted from raw clinical documentation and converted into structured tables highlighting diagnoses, medications, and procedures as character strings. Alongside this, clinical trials (input 2) are considered, specifically focusing on eligibility criteria derived from raw trial documents. The first task, Patient-Criteria Matching, is approached as a multi-class classification with three categories: "Match" where patient records align with trial criteria, "Mismatch" indicating no alignment, and "Unknown" for ambiguous alignments. The second task, Patient-Trial Matching, determines a match if a patient satisfies all inclusion criteria and does not meet any exclusion criteria.

Aim 1: Privacy-aware augmentation techniques

Acquiring comprehensive and high-quality data will present significant challenges, especially with concerns about costs and potential privacy infringements. To address this, our team will introduce a data augmentation technique that harnesses the capabilities of LLMs. This approach will aim to generate supplementary data points while preserving the semantic coherence of the original trial's inclusion and exclusion criteria. Initially, we will employ the Chain-of-Thought method to guide the LLMs in gradually generating prompts. These prompts will be specifically crafted to ensure that machine learning models will easily interpret the output data without losing its semantic essence. After establishing these prompts, we will use desensitized patient data combined with clinical trial data to execute a data augmentation process, thus guaranteeing the safeguarding of privacy. To illustrate this, Fig. 3 provides examples of the augmentation component of our proposed Large Language Model for Patient Trial Matching (LLM-PTM). On the technical front, given the criteria of a clinical trial represented as T=[i1 ,i2 ,...,ini ,e1 ,e2 ,...,ene], (where i and e are the inclusion and exclusion criteria respectively) we will task the LLM with producing augmented data points T that conform to specific constraints. For every criterion ik and el in T, we will formulate input strings i'kand e'l using the equation i'k=oik, e'l el , where o is the designed prompt and signify the concatenation operation. The LLM will then process these strings to generate augmented data points Aik and Ael, respectively. By viewing the LLM as a function, the expressions Aik=LLM(i'k) and Ael=LLM(e'l) are derived, leading to the creation of the complete augmented trial dataset represented as: =k=1n Aik l=1mAel

Patient and Criteria Embedding. Once the text data has been preprocessed, it will undergo embedding through the taxonomy-guided deep learning method. We anticipate that latent representations of both a patient's visit record and trial criteria will be obtainable via LLMs. For the primary text encoding, our team will adopt the pretrained BERT [26]. When focusing on patient embedding, we will leverage a memory network [27], denoted as Mem(·). This approach will be instrumental in effectively preserving the sequential nature of visit data within the embedding space. Formally, the patient record embedding, denoted here as xP, will be derived from the encoding function fP(), and can be expressed as xP=fP(P)=Mem(BERT(a1), BERT(a2),...,BERT(an)), where P represents the patient's records. Regarding the embedding of criteria, to capture essential features within the embedding space, we propose a method that intertwines a convolutional neural network (CNN) with a highway layer [28]. This combination is envisioned to distill patterns across various levels for the semantic matching task [29]. Formally, the encoding function fc() will be employed to produce an EC embedding xc, defined as xi/e=fc(c)=Highway(BERT (c)), where c T. Conclusively, the outputs from the highway network will be determined by the equation Highway()=Sigmoid()Conv()+Conv()(1-Sigmoid(Conv())).

Potential Problems and Alternatives. We anticipate that GPT will demonstrate its capabilities in generating high-quality synthetic data, a capacity evidenced by our previous preliminary work. However, an emerging concern will be GPT's training on publicly available datasets. This raises the concern that the model might have already been exposed to the datasets we will use in our Phase I feasibility study, potentially leading to unintentional disclosures from the primary dataset. To counteract this potential pitfall, we will employ the sentence transformer to secure embeddings for both the original and the synthetic data. Subsequently, these embeddings will be projected using t-distributed stochastic neighbor embedding. Identifying distinct patterns between the synthetic and original data will be crucial, serving as indicators to ensure that GPT hasn't merely reproduced the dataset from its internal memory.

Aim 2: Patient-clinical Trial Matching

Throughout the model optimization process, our primary objective is to enhance patient-trial matching. A pivotal part of this endeavor involves addressing the distinct differences between inclusion and exclusion criteria. To reach our goal, we will introduce a composite loss function. This function will incorporate several loss terms, starting with the classification loss. As part of our strategy to optimize the classification performance, we intend to use a cross-entropy loss term Lcla=-yTlog(ŷ)-(1-y)Tlog(1-ŷ) which will measure the discrepancy between the predicted outcome ŷ and the ground truth y.

Our optimization process will also involve the creation of an inclusion/exclusion contrastive loss term. By constructing this term, we aim to directly address the alignment between the patient embedding and the EC embedding, for both inclusion and exclusion stipulations. The essence of this loss term lies in its ability to enable the model to pinpoint specific features, such as negation words, within the inclusion/exclusion criteria. This, in turn, facilitates the decision-making process regarding whether to include or exclude a patient. From a mathematical perspective, our goal will be to heighten the similarity between the retrieved patient memory and the inclusion criteria embedding, represented as (xP, xi), while concurrently reducing the similarity with the exclusion criteria embedding, denoted as (xP, xi). The formulation of this contrastive loss term will hinge on the following pairwise distance loss: Lcon=a=1,...,ni(1-s(xia,xP))b=1,...,nemax(0,s(xeb, xP)-)

Figure 4: Overall model framework

where s(·, ·) represents the similarity function between two vectors. Our team plans to opt for the cosine similarity function, anticipating its efficacy in measuring the distance between two data modalities. The hyperparameter will be introduced to define the minimum allowed similarity between the exclusion criteria embedding and the patient's memory. When a patient's data aligns with an inclusion criterion, the model's design will aim to amplify the cosine similarity between the associated embeddings, making the value of 1-s(xi, xP) converge to zero. Conversely, if a patient's profile leads to an exclusion based on a specific criterion, the model will strive to diminish the similarity between the corresponding embeddings. In this scenario, the value of max(0, s(xe, xP,)- will be driven downwards, but it will always remain above , ensuring that xi and xe maintaining distinguishable distances relative to xP in the latent space. Ultimately, the strategy will encompass the joint minimization of the loss functions via backpropagation, executed in an end-to-end manner as depicted by L = · Lcla+ (1) · Lcon, where will serve as a crucial parameter adjusting the weight of classification loss. A detailed representation of this model's architecture and workflow will be available in Fig. 4.

Proposed Work: Experimental. We will collect data from six different stroke clinical trials, namely NCT03735979, NCT03805308, NCT03263117, NCT03496883, NCT03876457, and NCT03545607, using ClinicalTrials.gov as our source. Our focus will be on both the inclusion and exclusion criteria, which will result in at least 150 sentence-level statements extracted.

Potential Problems and Alternatives: There is a potential negative impact of the number of synthetic sentences on the effectiveness of our proposed method. To address this issue, we will conduct experiments with varying numbers of synthetic sentences and ratios of seed examples. We propose a series of tests that uses a range of synthetic data to train our local model and to vary the pool size of our seed examples. The results of these examinations could indicate, for example, that increasing the number of synthetic sentences can improve model performance up to a certain point, beyond which the improvement becomes marginal; or that using a larger number of seed examples can increase the quality and diversity of the generated data.

Aim 3: Enhancing Recruitment, Participant Understanding, and Retention in Diverse Populations

The focus of Aim 3 is to improve the recruitment and retention of participants in clinical trials, specifically targeting diverse populations that are often underrepresented in clinical research. This objective is crucial as it ensures the trial results are generalizable and applicable across different demographic groups. To achieve this, we will integrate innovative strategies for participant engagement and retention, leveraging the capabilities of LLMs and AI-driven communication tools.

Firstly, we will develop culturally sensitive and linguistically appropriate educational materials using LLMs. These materials will be tailored to address common misconceptions and barriers to participation often encountered by diverse populations. For example, LLMs will be used to generate explanatory texts and FAQs in multiple languages, ensuring comprehensibility and cultural relevance.

Secondly, we will employ AI-driven communication tools to maintain engagement with participants throughout the trial. These tools will facilitate personalized interactions, regular updates, and reminders, thereby enhancing the overall trial experience for participants. The communication will be adapted to individual preferences, whether it's through text, voice messages, or virtual meetings, ensuring participants feel connected and informed.

Thirdly, we will utilize advanced analytics to identify patterns in dropout rates and factors affecting participant retention. By analyzing these patterns, we can proactively address potential issues, improving retention rates. For instance, if certain demographic groups show higher dropout rates, targeted interventions can be developed to address the specific needs or concerns of these groups.

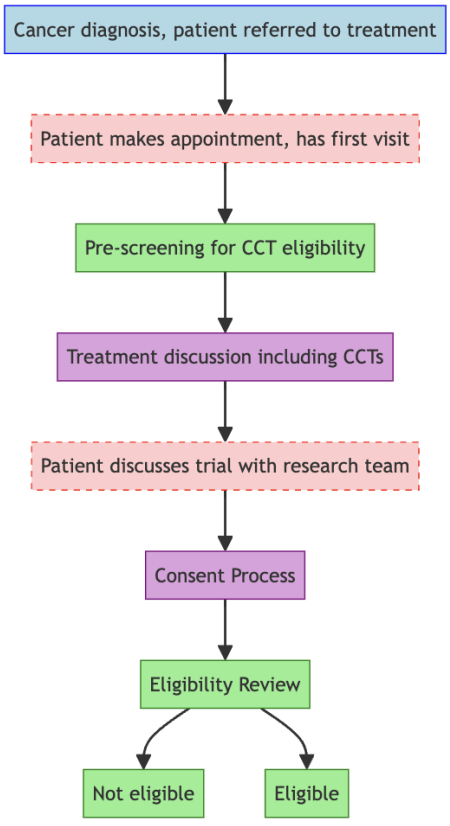

Figure 3: Navigating the Cancer Clinical Trial Process: A Patient's Journey

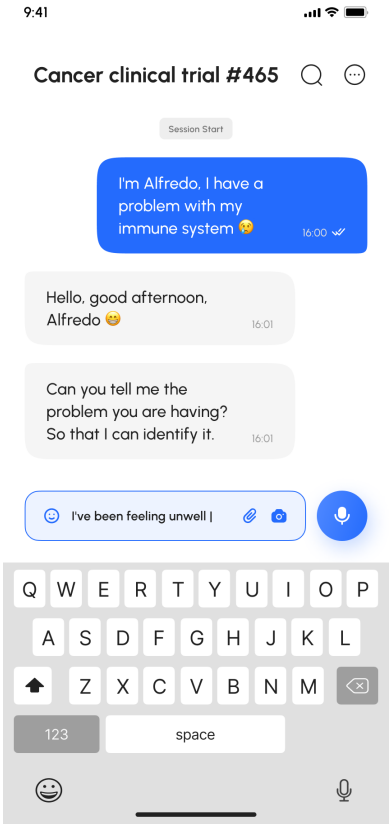

Figure 4: lAIrt leverages LLM to understand patients' participation at the individual level to help them navigate the clinical trials to increase retention in diverse populations

In addition, we plan to collaborate with community organizations and leaders to build trust and facilitate the recruitment process. These partnerships will be vital in reaching out to underrepresented communities, providing them with the information needed to make informed decisions about participating in clinical trials.

To evaluate the effectiveness of these strategies, we will track recruitment and retention metrics across different demographic groups. We will compare these metrics with those from trials that did not employ these strategies, to assess the impact of our approach on enhancing diversity in clinical trial participation.

Potential Problems and Alternatives: One potential challenge is the varying levels of technology access and literacy among different populations, which may affect the effectiveness of our digital engagement strategies. To mitigate this, we will offer alternative means of communication and support, such as phone calls and in-person meetings, to ensure inclusivity. Additionally, we will continuously monitor and adapt our strategies based on participant feedback and engagement levels, ensuring that our approach remains effective and responsive to the needs of diverse populations.

Figure 5: Flowchart of the recruitment and retention strategies in diverse populations.

Related Research or R&D

Preliminary Work:

Considering privacy and confidentiality of patient information are of utmost importance, AI POW LLC has developed a mechanism to ensure robust privacy protections that prevent unauthorized access to sensitive information. In our preliminary work, we developed a framework for measuring and improving the reliability of LLM for zero-shot tasks aimed at Biomedical NER and Biomedical RE while mitigating the privacy risk.

First, to assess the zero-shot performance of current LLM models for healthcare tasks, we conducted experiments on Generative Pre-trained Transformer (GPT) to investigate its ability to extract structured information from unstructured healthcare texts, specifically for biological NER and RE tasks. Our findings suggest that GPT directly only yields poor performance compared to SOTA models trained on the dataset for precision (P), Recall (R), and F-1 score (F). This result highlights that while GPT has demonstrated impressive inference and reasoning abilities in various classic natural language understanding (NLU) tasks, it is not adequate to apply GPT alone to healthcare tasks since it doesn't ensure both the required performance and the privacy requirements needed for cross-institutional collaboration for this domain [30, 31].

Design: Our preliminary study was geared towards assessing the efficacy of LLMs in creating a large volume of superior synthetic data with labels, using LLMs, and fine-tuning a local model for Biomedical NER and RE. As illustrated in Fig. 5, our primary work involved the creation of an innovative training paradigm to address the challenges of using LLMs for healthcare tasks.

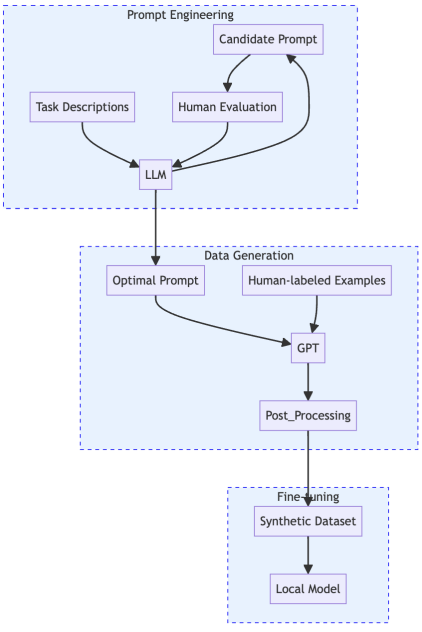

Figure 5: An overview of the workflow for synthetic data generation using GPT

Contrary to the direct application of LLMs in a zero-shot setting, we generated a substantial amount of synthetic data with labels using LLMs. To enhance the quality and diversity of the synthetic data, we utilized a limited number of human-labeled examples as seeds and developed suitable prompts to guide LLMs in creating a range of examples with diverse sentence structures and linguistic patterns. A post-processing step was employed to eliminate low-quality or duplicated samples generated by LLMs. Finally, we used synthetic data to fine-tune a local pre-trained language model. This design is particularly beneficial for multi-center clinical trials, where synthetic data can be easily shared among different local sites. Each of these sites can then use this data to fine-tune their individual models. Given that online LLMs like GPT4 are not sharable, this approach ensures that the performance of all sites can be significantly improved. As demonstrated in the Results section below, our experiments on four representative datasets showed that our proposed pipeline substantially improved the performance of the local model compared to the zero-shot performance of LLMs. Moreover, the approach effectively addressed data privacy concerns by minimizing the need for uploading patient data to an LLM API.

Results. Our work used LLMs to produce high-quality synthetic data, subsequently serving as the bedrock for fine-tuning dedicated models. Our approach yielded a significant performance boost, with F1-scores increasing from 23.37% to 63.99% in NER (Table 1) and from 75.86% to 83.59% in RE (Table 2), but also

minimized data acquisition timelines and expenses. Furthermore, generating data using LLMs can significantly reduce the time and effort required for data collection and labeling and mitigate data privacy concerns. Our proposed framework sets a strong precedent to present this promising solution to enhance the applicability of LLM models to solve the critical clinical text mining challenge for efficient clinical trial design. Experimental Approach to Leverage LLMs for Clinical Trial Design. As stated earlier, patient-trial matching involves finding appropriate patients for a specific clinical trial using their EHRs. These records reside in patients' medical databases. On the other hand, clinical trials are characterized by their descriptions, eligibility requirements, and other relevant details. Below, we elaborate on our problem setting and specific aims.

Table 1: Test results in biomedical named entity recognition. Precision (P), Recall (R), and F1 (F) scores on each dataset are reported. All the numbers are in percentage and computed based on 3 trials.

		Zero-shot	Fine-Tuned on Synthetic Data		
	Metrics	GPT	BERT	RoBERTa	BioBERT
NCBI Disease	P	32.84	39.41 ± 0.11	42.83 ± 0.48	43.14 ± 0.18
	R	44.86	59.15 ± 0.53	62.78 ± 2.37	63.92 ± 0.41
	F	37.92	47.30 ± 0.09	50.91 ± 1.10	51.51 ± 0.22
BC5CDR Disease	P	17.03	62.51 ± 0.40	64.47 ± 0.59	63.08 ± 0.68
	R	43.56	61.85 ± 0.08	62.95 ± 0.18	64.63 ± 0.59
	F	24.48	62.18 ± 0.16	63.70 ± 0.19	63.84 ± 0.31
BC5CDR Chemical	P	5.76	62.45 ± 2.42	67.56 ± 0.84	68.88 ± 0.83
	R	11.69	81.96 ± 1.89	83.36 ± 1.06	86.36 ± 0.76
	F	7.72	70.84 ± 0.95	74.63 ± 0.81	76.64 ± 0.78
Average	P	18.54	54.79	58.28	58.36
	R	33.37	67.65	69.69	71.63
	F	23.37	60.10	63.08	63.99

Table 2: Test results in biomedical relation extraction. Precision (P), Recall (R), and F1 (F) scores on each dataset are reported. All the numbers are in percentage and computed based on 3 trials.

		Zero-shot	Fine-Tuned on Synthetic Data		
	Metrics	GPT	BERT	RoBERTa	BioBERT
GAD	P	76.32	82.39 ± 0.93	83.59 ± 1.01	84.28 ± 1.03
	R	79.82	90.21 ± 0.15	92.57 ± 0.47	94.21 ± 1.35
	F	78.03	86.12 ± 0.72	87.85 ± 0.68	88.96 ± 1.01
EU-ADR	P	72.01	72.05 ± 1.02	73.44 ± 1.07	75.81 ± 1.43
	R	75.43	78.13 ± 0.50	79.22 ± 0.22	81.20 ± 1.00
	F	73.68	74.96 ± 0.81	76.22 ± 0.55	78.41 ± 0.77
Average	P	74.16	77.22	78.52	80.05
	R	77.62	84.17	85.90	87.70
	F	75.86	80.54	82.04	83.69

Relationship with Future R&D

Long-term goal. Starting with Phase Zero, our team utilized design thinking principles to empathize with our users and accurately define the problem space. This led us into Phase 1, where lean methodology guided us in the creation and testing of a prototype; this will allow us to learn and pivot as necessary. Also in Phase I, our team will showcase the effectiveness of our LLM-driven platform in allowing cross-organizational collaboration for clinical trial design. Our solution will ensure

data de-identification with minimal human-labeled data, guide LLMs in reliable data augmentation through precise prompts, and align EHR data with clinical trial criteria. This next stage, Phase II, harnesses Agile development methodologies to mature and optimize the technology, further de-risking our proposed platform for patient trial matching. Specifically, in Phase II, AI POW LLC aims to refine and scale the platform, integrating feedback from Phase I to enhance accuracy and user experience. We'll also expand partnerships with leading healthcare institutions to ultimately drive forward the vision of AI-powered, collaborative clinical trial design. Finally, Phase 2+ focuses on broader implementation and scaling of the solution while actively seeking external investors to fuel the growth and reach of this proposed work.

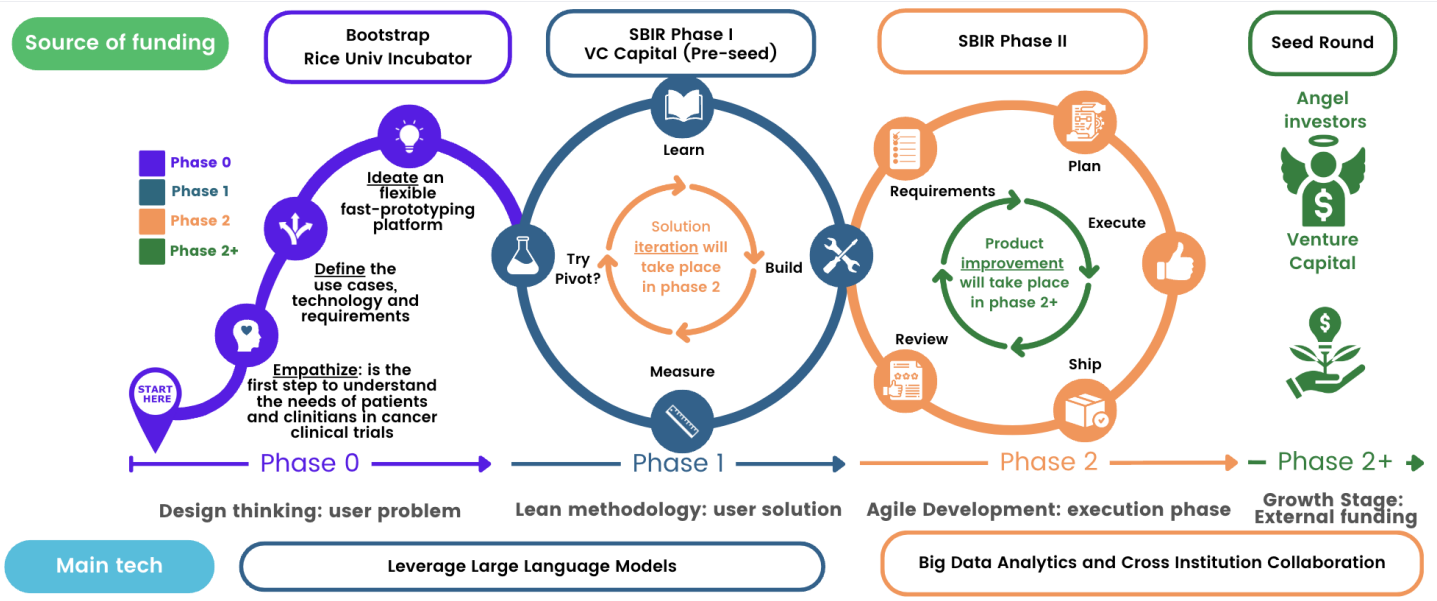

Figure 6: Relationship of current preliminary work Phase Zero, our proposed Phase 1 work and future R&D work of Phase II, Phase 2+, and beyond

Innovation

Clinical trials, which are pivotal in the advancement of innovative disease treatments, frequently encounter setbacks, with many being unable to start due to overwhelming recruitment obstacles [8]. The emergence of automated patient-trial matching introduces a groundbreaking and promising solution, poised to revolutionize the intricacies of cross-institutional collaboration for clinical trial design. The essence of this proposed approach lies in identifying eligible patients for clinical trials based on their EHR data and trial EC, which encompass both inclusion and exclusion criteria. If data limitations are solved, the problem can be framed as a classification problem with a given input consisting of a patient's complete EHR data and a single trial's EC, the output can be classified as either match, mismatch, or unknown.

The specific innovations of our proposed approach include the following:

Privacy-aware Data Augmentation: Data augmentation in natural language processing (NLP) traditionally involves diverse transformations of text data to bolster model training. In patient-trial matching, the challenge lies in the limited richness of the available training data. To address this, we introduce an effective data augmentation method designed to generate a more varied dataset, allowing machine learning models to better capture the intricacies of patient and eligibility criteria information. However, the vast potential of big data in healthcare is accompanied by significant privacy concerns. These concerns range from the ethical implications of data usage to the technical challenges of data de-identification and linking patient data from different sources. Our proposal's innovation lies in introducing privacy-conscious augmentation techniques using open-source components such as BERT [26]. The primary goal of these techniques is to enhance the accuracy of patient-trial match identification while ensuring no leakage of private-source data, making possible collaboration

from multiple institutions to safely share data and models for comprehensive clinical trial design.

Leveraging LLMs for EHRs and Clinical Trial Descriptions: Our second innovation centers around the potential of LLMs to enhance the standardization and interoperability between EHRs and clinical trial descriptions. By utilizing the advanced natural language generation capabilities of LLMs, the approach aims to improve patient-trial matching by homogenizing the diverse terminologies and ontologies present in both EHRs and clinical trial criteria. The proposed approach not only enhances the matching process but also ensures the security and confidentiality of sensitive patient data.

Holistic Framework for Enhanced Medical Research: Finally, our proposed combined work presents a holistic framework where LLMs address multiple challenges in medical research – from text mining and data extraction to collaborative work in clinical trial design. The dual focus ensures efficient information extraction from clinical texts and optimizes patient-trial matching, all while maintaining strict adherence to data privacy.

References

[1] Trends, Charts, and Maps. https://classic.clinicaltrials.gov/ct2/resources/trends, 2023. Accessed: 17-Aug-2023.

[2] A. T. McCray and N. C. Ide. Design and implementation of a national clinical trials registry. *J Am Med Inform Assoc*, 7:313–323, 2000. doi: 10.1136/jamia.2000.0070313.

[3] C. Laine, R. Horton, C. D. DeAngelis, J. M. Drazen, F. A. Frizelle, F. Godlee, et al. Clinical trial registration: looking back and moving ahead. *Croat Med J*, 48:289–291, 2007. URL https://www.ncbi.nlm.nih.gov/pubmed/17589970.

[4] T. P. Bardyn, E. F. Patridge, M. T. Moore, and J. J. Koh. Health sciences libraries advancing collaborative clinical research data management in universities. *J Escience Librariansh*, 7, 2018. doi: 10.7191/jeslib.2018.1130.

[5] L. M. Friedman, C. D. Furberg, D. L. DeMets, D. M. Reboussin, and C. B. Granger. *Fundamentals of Clinical Trials*. Springer International Publishing, 2015. doi: 10.1007/978-3-319-18539-2.

[6] O. T. Inan, P. Tenaerts, S. A. Prindiville, H. R. Reynolds, D. S. Dizon, K. Cooper-Arnold, et al. Digitizing clinical trials. *NPJ Digit Med*, 3:101, 2020. doi: 10.1038/s41746-020-0302-y.

[7] T. J. Moore, J. Heyward, G. Anderson, and G. C. Alexander. Variation in the estimated costs of pivotal clinical benefit trials supporting the us approval of new therapeutic agents, 2015-2017: a cross-sectional study. *BMJ Open*, 10:e038863, 2020. doi: 10.1136/bmjopen-2020-038863.

[8] B. Hargreaves. Clinical trials and their patients: the rising costs and how to stem the loss. Pharmafile, 2023. URL http://www.pharmafile.com/news/511225/clinical-trials-and-their-patients-rising-costs-and-how-stem-loss. Accessed: Aug-2023.

[9] CTG labs - NCBI. https://clinicaltrials.gov/ct2/about-site/background, 2023. Accessed: 18-Aug-2023.

[10] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal. Clinical development success rates for investigational drugs. *Nat Biotechnol*, 32:40–51, 2014. doi: 10.1038/nbt.2786.

[11] H. A. Glick, J. A. Doshi, S. S. Sonnad, and D. Polsky. *Economic Evaluation in Clinical Trials*. OUP Oxford, 2014. URL https://play.google.com/store/books/details?id=Xqi1BAAAQBAJ.

[12] M. F. Drummond and G. L. Stoddart. Economic analysis and clinical trials. *Control Clin Trials*, 5:115–128, 1984. doi: 10.1016/0197-2456(84)90118-1.

[13] F. D. Beacher, L. R. Mujica-Parodi, S. Gupta, and L. A. Ancora. Machine learning predicts outcomes of phase iii clinical trials for prostate cancer. *Algorithms*, 14:147, 2021. doi: 10.3390/a14050147.

[14] K. M. Gayvert, N. S. Madhukar, and O. Elemento. A data-driven approach to predicting successes and failures of clinical trials. *Cell Chem Biol*, 23:1294–1301, 2016. doi: 10.1016/j.chembiol.2016.07.023.

[15] L. Follett, S. Geletta, and M. Laugerman. Quantifying risk associated with clinical trial termination: A text mining approach. *Inf Process Manag*, 56:516–525, 2019. doi: 10.1016/j.ipm.2018.11.009.

[16] M. E. Elkin and X. Zhu. Predictive modeling of clinical trial terminations using feature engineering and embedding learning. *Sci Rep*, 11:3446, 2021. doi: 10.1038/s41598-021-82840-x.

[17] H. Hassanzadeh, S. Karimi, and A. Nguyen. Matching patients to clinical trials using semantically enriched document representation. *J Biomed Inform*, 105:103406, 2020. doi: 10.1016/j.jbi.2020.103406.

[18] M. Alexander, B. Solomon, D. L. Ball, M. Sheerin, I. Dankwa-Mullan, A. M. Preininger, et al. Evaluation of an artificial intelligence clinical trial matching system in australian lung cancer patients. *JAMIA Open*, 3:209–215, 2020. doi: 10.1093/jamiaopen/ooaa002.

[19] A. Bhatt. Artificial intelligence in managing clinical trial design and conduct: Man and machine still on the learning curve? *Perspect Clin Res*, 12:1–3, 2021. doi: 10.4103/picr.PICR_312_20.

[20] J Gao, C Xiao, LM Glass, and J Sun. Compose: Cross-modal pseudo-siamese network for patient trial matching. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery Data Mining*, pages 803–812, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3394486.3403123.

[21] X Zhang, C Xiao, LM Glass, and J Sun. Deepenroll: Patient-trial matching with deep embedding and entailment prediction. In *Proceedings of The Web Conference 2020*, pages 1029–1037, New York, NY, USA, 2020. Association for Computing Machinery. doi: 10.1145/3366423.3380181.

[22] K Sanderson. Gpt-4 is here: what scientists think. *Nature*, 615:773, 2023. doi: 10.1038/d41586-023-00816-5.

[23] K Singhal, T Tu, J Gottweis, R Sayres, E Wulczyn, L Hou, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023. URL `http://arxiv.org/abs/2305.09617`.

[24] M Hosseini, CA Gao, DM Liebovitz, AM Carvalho, FS Ahmad, Y Luo, et al. An exploratory survey about using chatgpt in education, healthcare, and research. *medRxiv*, 2023. doi: 10.1101/2023.03.31.23287979.

[25] HA Aziz. A review of the role of public health informatics in healthcare. *Journal of Taibah University Medical Sciences*, 12:78–81, 2017. doi: 10.1016/j.jtumed.2016.08.011.

[26] J Devlin, M-W Chang, K Lee, and K Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. URL `http://arxiv.org/abs/1810.04805`.

[27] J Weston, S Chopra, and A Bordes. Memory networks. *arXiv preprint arXiv:1410.3916v11*, 2014. URL `http://arxiv.org/abs/1410.3916v11`.

[28] RK Srivastava, K Greff, and J Schmidhuber. Highway networks. *arXiv preprint arXiv:1505.00387*, 2015. URL `http://arxiv.org/abs/1505.00387`.

[29] Q You, J Luo, and Z Zhang. End-to-end convolutional semantic embeddings. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5735–5744. IEEE, 2018. doi: 10.1109/cvpr.2018.00601.

[30] T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

[31] C Qin, A Zhang, Z Zhang, J Chen, M Yasunaga, and D Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint arXiv:2302.06476*, 2023. URL http://arxiv.org/abs/2302.06476.

Potential Commercial Applications

Detail why the proposed project has potential commercial applications and discuss market dynamics.

Senior/Key Personnel and Bibliography of Directly Related Work

Identify key personnel and provide summaries of their most relevant experience or publications.

Subcontractors/Consultants

No contractors will be requested for this project.

Facilities and Equipment

Detail where the research will be conducted, describe facilities, and list important equipment.

Data & Safety Monitoring

The project team is fully committed to supporting the security of the data collected during all testing procedures as well as maintaining the confidentiality of all registered users of the system. Secure data management policies will be followed to ensure data safety and confidentiality is maintained.

Data monitoring will be extensive as each data class and element will be characterized and approved, then denoted for data completeness. Data will be monitored through edit checks and data quality control designed for the study and documented in the study data management plan and associated guidelines. Data transfers between locations (if any) will be conducted in a secure manner, such as using SFTP or SSL webservices.

Research Involving Vertebrate Animals

This project does not involve vertebrate animals.

Dual Use Research of Concern

This project is not subject to the DURC policy since we will not use any agent or toxin during the course of the research work.

Human Subjects and Clinical Trials Information Form

Source of data

We will use simulated and de-identified data to conduct our proposed research. We expect this project to be eligible for NIH Exempt Human Subjects Research under §CFR 46.104(d)(4) because the collection/study of data are recorded such that subjects cannot be identified. Our research teams are highly educated, trained, and experienced in legal, regulatory, and ethical requirements and best practices of research involving human data. We will maintain robust confidentiality protections, including standard technical safeguards to remove identifiers from our study.

Interaction with patients

The proposed study is focused on methodology development and will use only existing genomics databases. Any form of interaction with the patients or intervention will not take place. Based on these characteristics of the study, our self-assessment is that the study presents no more than minimal risk to human subjects. Nevertheless, we will seek IRB guidance as to whether the study is classified as having a minimal risk and whether a waiver of consent would be warranted given the impracticality of seeking consent from the targeted patient group.

Potential risks

The data sets that are used for method development are from existing data. The primary potential risk to subjects is loss of confidentiality with potential linkage to external data to reveal identity. Given our team's previous experience of handling such sensitive data and a strong background in health privacy, this is unlikely to happen. Our project will not reveal any patient-level information.

Adequacy of Protection Against Risk

This study will use retrospectively collected data to study fairness enhancing methods for AI models in the healthcare context. Although the investigation has little risk, we consider the privacy risk as being yet unknown, so we will still maintain a high standard of protection in our study • The servers will be located behind firewalls in our institutions. • Access to the development servers will be limited to the approved study investigators and their collaborators.

Data and Safety Monitoring Plan

In addition to the security layers deployed by IT Security team at UTHealth, the School of Biomedical Informatics has a strong and dedicated IT team that maintains the privacy and security of relevant research. All SBMI servers are hosted in a private HIPAA-compliant environment to provide privacy-preserving storage and high-performance computing capabilities to biomedical and behavioral researchers. The data collected and generated in this project will be stored and processed in this secure SBMI computing environment. The research team will review any issues with data safety in a regular meeting and modify or stop the research protocol if necessary. We will abide by all the rules and plans developed for human subject protection purposes that apply to this proposed project.

Potential benefits

Our study has the potential to enhance data sharing and inform better decision making for transplantation. These results are valuable in promoting and generating new scientific evidence for biomedical research. The discoveries from this project may not directly benefit individuals whose data will be used for the analyses. Overall, the risks to subjects are reasonable in relation to the potential benefit to research participants and others. We will abide by all the rules and plans developed for human subject protection purposes that apply to this proposed project.

Appendix E

SAMPLE FORMAT- STATEMENT OF WORK
TITLE: (Normally carries the same title as Phase I, but may be changed if the previous title no longer accurately describes the project.)

Background Information

Prove a general, brief summary of what needs the research on this contract will be addressing. Information from the Topic description may be useful to include for context.

Scope

Provide a one to two sentence statement of what is expected to be accomplished under the contract.

Objectives

State the specific objectives and anticipated end results of the proposed Phase II effort

Services to be Performed.

Begin this section with the following language: "The contractor shall independently perform all work and furnish all labor, materials, supplies, equipment, and services (except as otherwise specified in the contract) to perform the following services:"

Specific Requirements – List all tasks in a logical sequence, in an outline format, to precisely describe what is expected of the contractor in performance of the work. Tasks should contain enough detail to establish parameters for the project and keep the effort focused on meeting the objectives. Do not include any proprietary information.

If the research plan includes the use of human subjects or animals, briefly describe the parameters of this use. Refer to Sections 4 and 8 of the Solicitation for further guidance.

Reports & Deliverables - Describe end products and deliverables, and describe periodic/final reports required to monitor work progress under the contract.