仕様書

電子カルテデータのテキストデータから構造化データを作成するための臨床研究情報入力支援システムの構築

2025 年 5 月

国立研究開発法人 国立循環器病研究センター

1. 委託業務名
電子カルテデータのテキストデータから構造化データを作成するための臨床研究情報入力支援システムの構築

2. 実施業務の概要
国立循環器病研究センター予防医学・疫学情報部は、国立高度専門医療研究センター(国立循環器病研究センター、国立がん研究センター、国立健康危機管理研究機構、国立精神・神経研究センター、国立成育医療研究センター、国立長寿医療研究センターの 6 施設を指す。以下、6NC）での臨床研究管理に伴う人的負担の軽減を目的とした、「医療研究連携推進本部 横断的研究開発費 横断的研究推進費 研究課題番号 JH2024-B-06、研究課題名：電子カルテ情報からの自然言語処理（本仕様書においては、生成 AI による言語処理も含んでいる）によるレジストリーおよび臨床試験情報入力支援システムの基盤構築」研究を行っている（以下、本研究)。
本研究は 3 年計画であり、初年度である昨年は、生成 AI を用いて作成したデモカルテにおいて医学用語を理解できるよう学習し、非構造化データから変数項目を構造化データ（本仕様書においては、SS-MIX や HL7FHIR 等のような標準化ストレージのフォーマットというわけではなく、臨床試験の解析用データセットのような構造化データを指す）として抽出できるよう自然言語処理を用いた事前学習システム(以下、事前学習システム)を作成した。今年度は、実際の電子カルテデータを用いて、事前学習システムのチューニングを行うことを研究マイルストーンとしている。
本委託業務は、本研究における入力支援システム開発を行うためのものである。

3. 納 期　　2026年2月28日

4. 作業対象　事前学習モデルと各NCより提供される既存のテキストデータ

5. 業務の目的および目標
① 臨床研究管理に伴う人的負担軽減を目的に、下記を今年度の目標として定める。各 NC から提供される実際の電子カルテデータ（テキストデータといった非構造化データを含む）を用いて、事前学習システムのチューニングを行い、構造化データの抽出精度を向上させる。ここでのチューニングとは、病名・薬剤名・症状などの一般的な医学用語を理解できるよう学習した事前学習モデルを、6NC 全ての領域(循環器領域、がん領域、感染症領域、糖尿病領域、精神・神経領域、小児・産婦人科領域、高齢者疾患・認知症領域等)において、専門用語や独自の言い回しを理解できるよう、さらなる学習を行うことを意味する。
② そのチューニングされた事前学習システムが、各 NC が定める用件定義を満たすようにすること。
③ 上記を通して、電子カルテデータ（テキストデータといった非構造化データを含む）から、臨床試験情報入力に用いるような構造化データを作成する入力支援システムを構築すること。

尚、昨年度の研究成果を有さない受託者には、必要に応じて秘密保持契約を締結したうえで（本委託業務の受託事業者、国循、そして昨年度の受託事業者との間で)、昨年度の納品物の公開に応じる。ただし、受託契約前に発明されたアルゴリズムやモデルの知財権に関わる内容については、その限りではない。

6. 業務の必要条件

① 当業務の指揮は、臨床経験を有する医師が執ること。

② 本研究全体の目的であるシステム構築に向けて、事前学習モデルを整合性のとれたモデルへチューニングを行うこと。

③ チューニングされたモデルを基に、各NCが期待する入力支援システムを構築し、各NCが利用できる状態で納品すること。

④ 受託者が今年度独自に事前学習システムを作成する場合は、以下条件を満たし、かつ今年度の発注分も含め納期までに完了すること。

- 汎用モデルを利用する場合は、本研究に沿った学習を追加で行うこと(例：fine tuning)。基盤となるようなモデルを既に有している場合は、利用に関して国立循環器病研究センターの研究者と相談して進めること。モデルはローカルな環境で利用可能なものとする。
- 医学用語リスト(病名・薬剤名・症状など)と医学用語を盛り込んだダミーカルテもしくは匿名加工された電子カルテテキストにもとに、3000症例程度のカルテデータに正解ラベルを付与し、正解ラベル付きカルテコーパスとする。これを利用して新たな事前学習モデルの精度検証を実施し、精度が70%を超えていること。ここでのダミーカルテとは、架空の患者の臨床背景・経過・治療・転帰などをNCの医師が記載した退院サマリ等を想定したテキストのことである。可能な限り実物のカルテに近い状態でモデルを学習する必要があるため、医師以外が作成したダミーカルテの使用は禁ずる。
- 正解ラベルは、医師が判断して医学的妥当性を担保すること。特にダミーカルテやカルテコーパスにおいて、カルテ文中の医学用語の表記揺れ、診断病名、疑い病名、除外病名の違いなど、カルテ特有の表現に対応できること。
- 新たな事前学習モデルの精度は、疾患名で85%以上、検査名で75%以上、症状イベントで80%以上、検査結果で75%以上、日付で70%以上であること。
- 6NCが有するデータにアクセスできるようになるまでに時間を要するため、開発初段階においては昨年と同様に受託エリア(6NC)外でモデルを構築すること。

⑤ 本研究に必要な書類(倫理審査委員会への提出書類等)の作成を補助すること。

⑥ 各NCとの実際のカルテデータの受け渡しについては、現地に赴くことも含めて受託者自らが行うこと。

7. 受託者の条件

① 最終的に臨床研究の効率化・迅速化を進めるツールが開発できるよう、臨床医学領域における論文業績があり博士の学位を有すること、またはそれに相当する知見を有すること。

② AIモデルを使った研究実績およびAIソフトウェア/システム開発の知識・経験・コーディングスキルおよびその実績を有し、各NCの様々なフォーマットのデータに柔軟に対応できることに加え、実臨床業務を理解した事前学習モデルのチューニングができること。

③ 臨床医の実務経験あるいはそれと同等の臨床医学用語に対する知見があり、6NC全

ての領域の医学用語やカルテ記載特有の表現方法を正しく理解し、それらをチューニング作業に反映させることができること。

④ 本委託内容の一部は、「人を対象とする生命科学・医学系研究の倫理指針」を下に実施をすすめることになる。そのため、当該指針を熟知し、それを遵守した研究を行った実績があること。そして、当該指針で記載されている研究に参加する機関の要件を満たしていること。

⑤ 機密保持、知的財産等に関して本仕様書が定める責務を受託者が負うよう、必要な処置を実施し、当センターに報告し、承認を得ること。

8. 納品物

・各NCが保有する既存データでファインチューニングを行った事前学習モデルとそのスクリプト（プログラミングコード）
・それをもとに構築された、入力支援システムとそのスクリプト（プログラミングコード）
・そのモデル本体のバイナリーデータ
・そのモデルの使用方法についての説明書
・デモデータ（モデル作成に使用したデータ）
・システム構築の方法と精度評価を記載した報告書とする（要約、introduction, methods, results, discussion, referenceが記載されていること）。
・当センターにあるワークステーションに、入力支援システムプラットフォームを構築し、システムの動作確認や利用方法について説明をすること。加えて、国立循環器病研究センターが用意するテキストデータを使用して、システムのデモンストレーションを実施すること。これらについて、国立循環器病研究センターの研究者の同席のもと、十分な説明をしながら実施すること。

9. 納入場所

国立研究開発法人 国立循環器病研究センター

10. 情報セキュリティ管理

・受託者は、以下を含む情報セキュリティ対策を実施すること。また、その実施内容及び管理体制についてまとめた情報セキュリティ管理計画書を作成し、当センターの承認を受けること。
・当センターから提供する情報を、受託業務を遂行する目的外に利用しないこと。
・本業務の実施に当たり、受注者またはその従業員、本調達の役務の内容の一部を再委託する先、若しくはその他の者による意図せざる変更が加えられないための管理体制が整備されていること。
・受注者の本業務の実施場所について情報提供を行うこと。
・本業務従事者の所属・専門性（情報セキュリティに係る資格・研修実績等）に関する情報提供を行うこと。
・情報セキュリティインシデントへの対処方法を整備していること。
・情報セキュリティ対策に関する履行状況を定期的に確認し、当センターへ報告すること。
・情報セキュリティ対策の履行が不十分であると認められた場合、速やかに改善策を提出し、当センターの承認を受けた上で実施すること。
・当センターが求めた場合に、速やかに情報セキュリティ監査を受け入れること。
・本調達の役務内容を一部再委託する場合は、再委託されることにより生ずる脅威に対して

情報セキュリティが十分に確保されるように情報セキュリティ管理計画書に記載された措置の実施を担保すること。
・当センターから要保護情報を受領する場合は、情報セキュリティに配慮した受領方法にて行うこと。
・当センターから受領した要保護情報が不要になった場合は、これを確実に返却、または抹消し、書面にて報告すること。
・本業務において、情報セキュリティインシデントの発生または情報の目的外利用等を認知した場合は、速やかに当センターに報告すること。
・クラウドサービス（EDC を含む）の利用については、政府情報システムのためのセキュリティ評価制度（ISMAP）クラウドサービスリストに登録されていること。又は、その取得が進められていること。どちらにも該当しない場合は、ISMAP 管理基準についての自己評価を提出し、情報統括部の判断をあおぐこと。
・リモートメンテナンス回線は、センターが提供する VPN 環境で接続すること。
・独自のネットワーク（無線 LAN も含む）を構築しないこと。

11. その他の必要条件
・本業務により作成された事前学習モデルにより生成された構造化データは、すべて当センターに帰属する。なお、受託者が委託契約日までに発明したアルゴリズムやモデルの知財権については、受託者に帰属する。その後、当仕様書をもとに構造化データ作成ツールの商品開発となった場合は、知的財産権の割合等を別途協議する。
・本業務により作成された事前学習モデルは、契約期間終了後も当センターが利用できるものとする。
・当センターは受託者に対し、委託業務の実況等に関し、随時に書面または口頭による報告を求めることができる。
・業務の特質及び秘匿性上、受託者が業務の実施に当たって知り得た情報等は、本作業の目的以外に利用してはならない。また、他に情報を漏らしてはならないものとする。
・解析にかかる費用や送料にかかる費用等、本業務実施に要する全ての費用を含むものとする。

12. 国立循環器病研究センターにおいては、6. ～11. に記載の諸条件が、本件業務を外部委託先で実施するためには必要不可欠と考えられ、これらの諸条件を満たすことができる外部委託先を選定することが必要である。

以上