別添4

令和6年度厚生労働行政推進調査事業費補助金（厚生労働科学特別研究事業）
分担研究報告書2

診療行為の構造化と生成 AI 等を活用した標準化されたレセプト作成機能開発の為の基礎的調査研究（24CA2031）

研究代表者　八木　摂子　株式会社　FIXER・エンタープライズ部門

研究要旨：総括報告書参照

C．結果-2

B. CANBEモデル検討の為の基礎的調査・実験
1.生成AI活用シーン探索の為の基礎実験: 検体検査における診療報酬請求算定ルール（＝診療報酬点数表）をつかって算定項目を自動生成する試み

【A. 研究目的】
診療報酬請求算定ルールを予め学習させた生成AIが、診療報酬算定ルールに基づき、診療行為等データの標準化(表記ゆれの補正)が可能か、そして、電子カルテ検体検査オーダーから請求点数が計算可能か、実験した。

【B. 研究方法】
A. 対象データ
1.電子カルテ情報: 令和４年６月分
株式会社富士フィルム:統合診療支援プラットフォーム: Clinical Intelligence Technology & Architecture (CITA Clinical Finder: 電子カルテと連携しているシステム)より抽出
2.電子レセプト情報: 上記期間令和４年６月に対応するデータを準備
3. 検証症例の選定(以下の条件を同時に満たす症例)
・診療報酬算定を踏まえた「病院における一般的な検査」を定義し、これに合致するもの
・レセプト結果(正解データ)と電子カルテオーダ項目がマッチング可能な症例
・診療報酬の算定ルールのひとつである、包括規定の影響を受けないもの
・検査項目が検査実施料として算定されるもの
・トークン制約に抵触しないもの
4.算定ルール
診療報酬の算定ルールは厚生労働省による「令和４年度診療報酬改定について」にて公開されている、次の告示および通知を基にした。
•令和4年厚生労働省告示第54号「診療報酬の算定方法の一部を改正
•令和4年3月4日保医発0304第1号「診療報酬の算定方法の一部改正に伴う実施上の留意事項について（通知）」
株式会社ケアネットによる運営Webサービス「しろぼんねっと」は以上の公開情報を簡便に参照できるサービスであり、「しろぼんねっと」内の各種資料を参考に、プロンプト入力可能な構造に加工した。
5.レセプト電算処理システム医科診療行為マスター
社会保険診療報酬支払基金ホームページにて公開されている、2024年5月17日版レセプト電算処理システム医科診療行為マスター 全件マスターファイルを使用した。
B. 生成AI: 使用モデル: Claude Sonnet 3.5 V1
C. プロンプト設計
Retrieval-Augmented Generation(RAG)を 使用すると生成AIのブラックボックス部分にふれる可能性が懸念されている。加えて、生成結果に影響する変数としてプロンプトにRAGを加えることで、生成精度に対する評価や判断、対策が難しくなることが指摘されている。
ファインチューニングは、安定して精度上昇等の効果を見込めるかどうかは不透明とされる。また、ファインチューニングの使用は、コスト等との兼ね合いから、モデルの選択肢が限定されてしまい、採用し得るモデルは現行の主流または最新のモデルと比べ、性能が低い傾向がある。
AIによる生成に必要な全ての情報は、プロンプトに直接入力する事で精度上昇が期待されることもあり、本実験では現状のLLMの精度を測ることを目指し、本実験ではプロンプト設計のみを実行する事で検証を実施した。

D. 解析・正答判定
正答判定は以下を同時に満たす場合に、「正答」と判定した。
•算定された診療行為がすべて正解データと一致すること
•正解データに存在しない診療行為が算定されていないこと
•各包括計算グループを構成する診療行為が正解データと一致すること
•各グループの点数が、出来高計算グループであれば構成診療行為の点数の合計に、包括計算グループであれば構成する診療行為の数から導かれる所定点数に一致すること
•合計点数が正解データと一致すること
正答解の評価条件に含まないもの: 出来高計算グループはグルーピングに明確な基準はないため、次の点は不問とした。
•算定されているグループの数が正解データと一致すること
•各出来高計算グループを構成する診療行為が正解データと一致すること

【C. 結果】
A.　実験対象
実験対象は、補足資料2. 対象選定フローに則って、以下の3症例が選択された(詳細は202406031A-sonota6参照)。
糖尿病ケース: 項目数59・検体種類2
潰瘍性大腸炎ケース: 項目数51・検体種類1
術前検査ケース: 項目数49・検体種類1
また、プロンプト実行の予備調査の為、サンプルケースを設定した。その為、サンプルケースに使用された検体種類は1つとし、検体項目も、外来診療で測定される事が多い項目を20個選択して設定した。サンプルとして選択・使用した項目は以下に記載する:
総蛋白定量・アルブミン・総ビリルビン・GOT・GPT・LDH・ALP・γ-GTP・Na・K・Cl・尿素窒素・クレアチニン・CRP・白血球数・赤血球数・血色素量・ヘマトクリット・血液像・血小板数。

B. プロンプト設計
プロンプト設計は、ヒトによる検査オーダーを診療報酬として算定するまでの作業工程を複数のステップに分解し、ステップ単位に具体的な指示を実施するプロンプトを組み立てた(詳細は様式A8参照)。
1巡目（ステップA～ステップG）: 算定ルールに基づき、診療行為および請求点数を算出する。以下に詳細を示す。
◆ステップA: 指示項目⇒診療行為マッチング
検査オーダーの各指示項目を、AIによる推論を加えながら点数表上の該当する診療行為に変換する。
◆ステップB: 診療行為重複項目の除去
変換した診療行為に重複がある場合、重複をとりのぞく。
◆ステップC: 診療行為グルーピング
点数を算出するため、各診療行為をグルーピングする。
◆ステップD: 請求点数計算
各診療行為のグループについて、点数を決定する。
項目数によって算定する検査グループについては、項目数に対応する所定点数を算出する。 それ以外のグループについては、構成される診療行為の点数を合計する。
◆ステップE: 算定ルールに照らして、追加算定できる診療行為を算定する。
検査判断料や検体採取料、加算など、オーダーでは直接指示されない診療行為の追加算定をおこなう。
点数表記載の算定要件を参照し、要件を満たしているものを算定する。
◆ステップF: 算定ルールに照らして、算定できない診療行為を消去する。
ルール上、算定要件を満たせていない診療行為や、互いに同時算定ができない診療行為の組(=背反の関係)がある場合、点数の低い一方を消去する。
◆ステップG: 各グループの明細と点数を示し、最後に合計点数を示す。
合計点数を算出することで、一連の処理が完了する。

2巡目（ステップH～ステップI）：電子フォーマットによる算定内容の構造化をおこなう
(電子レセプトとして診療報酬を審査支払機関に請求するために、電子レセプトファイル(拡張子UKE)のフォーマットで電送をおこなう必要がある為。)
◆ステップH:
ステップGを経た算定結果と医科診療行為マスター、診療日といった背景情報を入力とし、算定した診療行為をレセプト電算処理コードによって符号化する。
◆ステップI:
診療行為の算定日や回数とともに、電子レセプトの構造を反映したYAML形式として出力する。

C. 正答率
総合結果を以下に示す。
1巡目（ステップA～ステップG）正答率
ケース1. 潰瘍性大腸炎: 79.9%(正答数195/試行回数244)
ケース2. 術前検査: 6.5%(正答数55/試行回数852)
ケース3. 糖尿病: 0.6%(正答数2/試行回数314)
参考: サンプル: 63.2%(正答数156/試行回数247)
2巡目（ステップH～ステップI）正答率
ケース1. 潰瘍性大腸炎58.2%(正答数110/試行回数189)
(1巡目で正答が得られた195件から2巡目のプロンプトに不備のあった6件を省いた数)
ケース2. 術前検査 74.5%(正答数41/試行回数55)
ケース3. 糖尿病: 1巡目の正答率が低く未実施
参考: サンプル: 73.0%(正答数114/試行回数156)
各ステップ正答率を以下に示す。
◆ステップA.検査オーダー項目から診療行為への変換結果
ケース1. 潰瘍性大腸炎: 73.7%(正答検査項目数7933/回答検査項目数10761)
ケース2. 術前検査: 79.9%(正答検査項目数7592/回答検査項目数9506)
ケース3. 糖尿病: 94.5%(正答検査項目数11681/回答検査項目数12355)
参考: サンプル: 97.1%(正答検査項目数4180/回答検査項目数4060)
◆ステップC: 診療行為グルーピングと項目数の集計結果
•CRPのグルーピング誤り: 出来高算定となるべき「CRP（C反応性蛋白）」を、少なくない頻度で項目数算定グループに含むという誤判断をした。
ケース1. 潰瘍性大腸炎: 98.0%(正答数239/1巡目試行回数244)
ケース2. 術前検査: 98.6%(正答数710/1巡目試行回数720) (1巡目の試行回数852件からトークン溢れにより出力が途切れた132件を除く)
ケース3. 糖尿病(CRP指示無): 正解できず。
(複数検体にまたがるグルーピング: 糖尿病症例における「クレアチニン」と「クレアチニン（尿）」を同一グループに含めるべきケースで、ほとんどのケースで誤って別グループに分類した)
参考: サンプル: 68.4%(正答数169/1巡目試行回数247)
◆ステップD点数計算結果
ケース1. 潰瘍性大腸炎: 項目数算定: 100%(正答数233/検証回数233)
ケース1. 潰瘍性大腸炎: 出来高算定: 99.6%(正答数2278/検証回数2288)
潰瘍性大腸炎では、項目数算定であれば項目数から点数を算定ルールに基づいて導出し、出来高算定であれば単に点数を合計するという作業は、高い精度で正答が可能であった。
◆ステップE. 追加算定結果
検証した検体検査管理加算の算定・検体採取料の算定・外来迅速検体検査加算・判断料の算定では、プロンプト最適化後は概ね正確に追加算定された。
◆ステップF. 消去の結果
不規則抗体検査の消去: 術前検査症例における「不規則抗体検査」は、一般的な算定要件チェック指示ではほぼすべてのケースで誤って算定された。不規則抗体検査の算定要件を明示的に確認するよう具体的な指示を追加した後は、改善が見られ、多くのケースで適切に消去された。

【D. 考察】
・実験結果要約
本実験では、ケース間で正答率79.9-0.6%とばらつきがあった。
これは、LLMの計算対応能力が症例の複雑性に強く影響を受けていることを示唆していると考えられた。正答率が低かった症例は、より複雑な手続を要すると考えられた。具体的には、算定ルール上の変数が多い症例であり、各ルールについての個別正答は見られても、総合的にすべてを正答できたケースは極めて少なかった。
本実験結果は、LLMによるルール判定は、単一の算定ルールに対しては一定の精度が保たれている一方で、複合的・統合的なルールの判定は精度が落ちる可能性を示唆している。
・実験結果解釈
本実験におけるLLMによるルール判定では、検査項目から診療行為への基本的変換における高い柔軟性を示していた。これは、ベースとなるLLMに既に一般的な医学情報を計算可能な状態で有しており、プロンプトからインプットされた診療報酬の基本体系を基に適切な計算・判定を導き出せる可能性を示唆している。一方で、「クレアチニン」と「クレアチニン（尿）」の区別など、項目名だけでは決定できない状況では精度が低下した。これは、LLMによる暗黙的な文脈の推測には限界があることを示している。また、「血液化学検査グループ」と「血液学的検査判断料」に代表される、類似した専門用語の文脈依存的な区別についてもLLMの限界が示されたと考えられた。加えて、LLMを診療報酬選定過程全体に適応する事に対して、幾つかの課題が示された。具体的な課題例として、糖尿病ケースにおいて、LLMは血液・尿といった検体横断の判断を正しく解釈できなかった。横断的なカウントの根拠となる算定ルールは、点数表上、複数箇所において断片的に記載されており、適切な関連づけ・統合しての理解が必要となる事が、その理由と考えられる。

二つ目の課題例として、術前検査ケースにおける「不規則抗体検査」の算定要件判断の失敗が挙げられる。検査の臨床的背景（術前検査）と算定要件（特定条件下での手術当日のみ算定可）の関連性判断の失敗は、LLMの臨床文脈理解の限界を示していると考えられた。

・実験の限界と今後の展望

1. 検証範囲の限定: 単一医療機関における3ケースのみの検証であり、一般化には多様な症例での検証が必要。

2. 実験環境制約:Claude Sonnet 3.5 V1のみを使用し、他LLMとの比較が未実施。プロンプト最適化や代替手法（RAG・ファインチューニング）の検証も検討の余地がある。

3. 限定された適用範囲: 実験範囲が検体検査算定に限定され、診療報酬全般の適用可能性は未検証。加えてトークン制約やコスト・処理速度の評価も不十分。

以上を踏まえ、LLMによる判定・計算精度の改善の為、次のような検討点が挙げられる。

1. オーダー情報の標準化・構造化など入力情報の質向上

2. 現在の行政文書形式の 診療報酬算定ルールを、LLMによる解析が容易な体系的なデータ構造化する。

3. 従来のプログラムによる機械的な構造化データの処理と、LLMによる非構造化データの自然言語処理を組み合わせた ハイブリッドアプローチの検討

【実験結語】

1. LLMは検査オーダーからの診療行為導出は基本的に可能だが、追加コンテキスト情報がないと精度が低下し、特に複数解釈可能な検査項目では補足情報が不可欠と考えられた。算定ルールに基づく点数計算は単純なケースで可能だが、複雑なグルーピングや特殊ルールでは困難で、複雑な条件では明示的な指示が必要。

2. 今年度実験を基にした、レントゲン、手術等診療行為における、生成AI活用の机上検証

・画像診断

画像診断は、原則として「撮影料+診断料+部位コメント」の構成で算定する。

撮影枚数のように実施時に定まる情報もあるため、指示オーダーに加えて実施結果が算定のためのインプットとして必要となる。

部位コメントについては、診療報酬では「選択的コメント」と呼ばれる仕組みで実現されるが、本研究で検証していないものの、生成AIへの入力に選択的コメントの一覧やマスターを加えることで、生成AIによる適切なコメントの選択および必要事項の追記は十分可能であると想像される。以上より画像診断は、今回の検体検査の検証結果が概ね適用できると考える。

・手術等(手術、輸血)

必要な情報としては、術式や薬剤、材料、監視内容といった実施記録に加え、麻酔記録や輸血記録も必要になり多岐に渡る。加えて、診療報酬上は手術として算定するが、臨床的には手術そのものではない行為があり、それらの判定やインプットは検討が必要と思われる。最後に、手術料は概して高単価であり、更に高額な薬剤や特定保険医療材料の算定が多い。その為、高額な請求漏れや査定減にも繋がりかねないため、他分野に比して高い精度が求められる。以上より、手術等(手術、輸血) は、算定における難度が比較的高い領域であると考えられる。

3. 表記ゆれ等の補正の精度検証

ステップA: 検査オーダー指示項目⇒点数表収載診療行為変換結果を再掲

ケース1. 潰瘍性大腸炎: 73.7%(正答検査項目数7933/回答検査項目数10761)

ケース2. 術前検査: 79.9%(正答検査項目数7592/回答検査項目数9506)

ケース3. 糖尿病: 94.5%(正答検査項目数11681/回答検査項目数12355)

参考: サンプル: 97.1%(正答検査項目数4180/回答検査項目数4060)

LLMによる表記ゆれの補正については、一定の精度を示していると考えられた。

4. 実験に使用された計算資源・通信資源量の計算

Claude 3.5 Sonnet V1は、入力トークン量の上限は20万トークンで、出力トークン量の上限は8,132トークンである。検証に使用した実行環境では、出力トークン上限が4,096トークンであった。

1. サンプルケース: 合計入力トークン量: 159,971・合計出力トークン量: 3,240

2. 潰瘍性大腸炎ケース: 合計入力トークン量: 162,141・合計出力トークン量: 5,269

3. 術前検査ケース: 合計入力トークン量: 165,669・合計出力トークン量: 5,404

4. 糖尿病ケース: 入力トークン量: 37,903・出力トークン量: 2,836

潰瘍性大腸炎ケースで、トークン当たり文字数を算出した。

トークン当たり文字数は、入力0.92字-0.75字・出力0.97字-1.34字

・入力: 65,168字 / 71,086トークン = 0.92字

・出力: 3,472字 / 3,683トークン = 0.97字

であった。

次に医科における算定ルール(診療報酬点数表)の全文をプロンプトとして掲載するために必要なトークン数を概算する。

令和4年度 診療報酬点数表は全4章15部で構成されており、1巡目のプロンプトには第2章第 3 部「検査」から、主に検体検査に関する算定ルール(注と通知を含む)を部分的に抜粋した。検査の部は点数表コードが264あり、プロンプトに掲載したものはうち11であった。プロンプトの消費トークン量71,086に対し、検査の部全体の所要トークン量は、71,086トークン / プロンプト抜粋点数表コード数11 × 部全体点数表コード数264 = 1,706,064と試算できる。続いて、令和4年度 医科診療報酬点数表において、点数表全体930ページ・検査の部146ページ(全体の15.6%)であるので、点数表全体のトークン量は 1,706,064 / 0.156 = 10,936,308 と試算された。