様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月 ２５ 日現在

機関番号：１５５０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１０３０３

研究課題名（和文）症例登録事業の根幹をなす精度検証とその向上に資する支援のあり方についての研究

研究課題名（英文）A Study on the Verification of Data Accuracy in Case Registration Projects and Methods for Supporting Its Improvementt

研究代表者
石田　博（Ishida, Haku）

山口大学・大学院医学系研究科・教授

研究者番号：５０１７６１９５

交付決定額（研究期間全体）：（直接経費）　　3,100,000円

研究成果の概要（和文）：症例登録における登録情報の精度向上を目的に、NCD症例登録におけるシステムによる支援なしの登録における精度の状況を胃癌、大腸癌、および乳癌の手術症例について電子カルテ情報を中心とした既存の診療情報と比較し、体重や検体検査、および、病理結果に基づく情報などの登録情報に不一致を確認したことから、書式横断に頻用される項目や情報種別を中心に登録支援ならびに精度確認を可能とするシステムが有用と考えられた。そこで標準、非標準の構造化データに加え、大規模言語モデルにより病理報告等から抽出された固有情報の活用も考慮した登録症例のデータ検証を支援するシステムの機能要件の整理を元にプロトタイプを開発した。

研究成果の学術的意義や社会的意義
エビデンスをもたらす情報源として有用な多施設にわたる様々な症例登録事業が展開されているが、登録データとの本来のデータ間での入力に相違がないかのチェックがデータ品質の確保に必須であり、施設内での電子カルテ情報などの既存データを利用した登録時のデータの検証、あるいは、手入力によらない自動入力による症例登録支援が望まれる。本研究はそのような差異が起こりやすいデータ種別を明らかにし、それらのデータを検証、あるいは、登録支援をするためのシステム開発における要点整理とプロタイプシステムの開発を行ったことに学術的、社会的意義がある。

研究成果の概要（英文）：To improve the data accuracy in case registration, we verified the accuracy of manually entered data in NCD case registration based on existing medical information, primarily electronic medical record information, for gastric cancer, colorectal cancer, and breast cancer surgery cases, and found many discrepancies in registration information such as weight, laboratory values, and information of pathology results. Therefore, a system to support registration and check the accuracy of the information, focusing on frequently adopted items across forms, would be useful, so we organized the functional requirements for a system that supports data verification of registered data, taking into consideration the use of unique information extracted from pathology reports using a large-scale language model, in addition to standard and non-standard structured data, and finally developed a prototype system.

研究分野：医療情報学

キーワード：症例登録　精度管理　構造化データ　固有情報抽出　検証支援システム

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

症例登録は各施設からの臨床情報を収集するため、収集後には、欠損値や通常あり得ないような極端値が登録されている場合の除き、その値が誤りの無いデータであるかの確認が難しく、そのまま登録、利用されることになり、データ品質管理の重要性が従来から指摘されている。また、外部の Web サイトから手入力での症例登録が行われる場合、様々な情報を数多く登録する必要があることから、その転記などにおいて確認誤りやタイプエラーが一定の割合で生じる可能性が高い。元となる情報は、電子カルテなどにすでに登録されている情報であり、別途、手入力を行う際には、登録すべき情報の確認と入力作業自体の負担の中で、誤った認識での判断による情報の登録がなされる場合も想定される。そのため、施設訪問や退院サマリーの送付などによるデータ検証も行われその結果も報告されているが、汎用面、持続面の観点から改善に向けた方策が必要である。

２．研究の目的

自施設の登録データを検証するための機能を有するシステムを構築することを目標に、NCD を中心とした症例登録の項目調査、および実際の登録事例の確認とその結果をもとに、その検証システムとして必要となる要件を整理し、そのプロトタイプシステムの構築を行うこと。

３．研究の方法

以下の(1)～(3)の調査研究とそれらを基にシステム機能要件の整理、プロトタイプシステムの開発を行った。

（1）NCD を中心とする症例登録フォーム（CRF）横断的な共通項目の検討

代表的な症例登録事業である NCD の CRF 中心に登録すべき項目内容の確認を行った。対象は消化器癌関連（6 書式）、呼吸器癌、乳腺癌、血管外科関連（12 書式）、内分泌腺外科関連（3 書式）、小児外科、形成外科、および肝癌登録の 27 書式とし、入退院関連情報、手術・処置関連情報の 2 群に分けて、登録対象情報項目別の CRF 頻度、および頻度の高い項目の検証用の既存データ取得の可能性を確認した。

（2）NCD 症例登録の実例における登録情報の不一致割合の検証

専用 Web サイトへの手入力による NCD 症例登録情報の精度を検証するため、2014～15 年における胃癌・大腸癌手術症例、および 2020 年における乳癌手術症例の登録情報を対象に、電子カルテ等の既存情報との比較による不一致度の検討を行った。胃癌･大腸癌登録情報では、診断、手術、入退院、転帰情報に関連したものを対象とし、乳癌登録情報においては併存症やホルモンレセプターの結果等の病理情報の関連した項目などについて退院サマリ、手術記録、DPC 調査情報などの医事情報、検査情報と比較し、その登録情報の不一致が起こりやすい情報項目を明らかにすることを目的に調査を行った。

（3）大規模言語モデルによる病理報告における固有情報の取得

施設内に構築した大規模言語モデル（LLM：meta-llama/Meta-Llama-3-70B-Instruct）を用い、胃癌、大腸癌、乳癌の術後組織標本の匿名加工後の病理検査報告書を対象として依頼時の臨床所見を含めて、それぞれの癌取り扱い規約に基づいた項目の固有情報抽出を試みた。精度指標は正確度とし、病理報告書にその項目の記載がある場合にその内容を正しく回答、記載がない場合に「なし」と回答した割合の合計を求めた。

（4）登録データ検証機能を有するプロトタイプシステムの構築

電子カルテ、およびその関連システムに蓄積されたデータ（既存データ）を真の情報として、登録された情報の精度を検証することを可能とするプロトタイプシステムの機能要件整理と構

築を行った。

４．研究成果

(1) NCD を中心とする症例登録フォーム（CRF）横断的な共通項目の検討

総登録項目数は 2370 項目で、全体の 7 割を超える項目が一つの書式にしか採用されていない項目であった。書式横断的に頻用されている項目は、性別のように全 27 書式で用いられているものを最頻として 20 書式以上に使われている項目は４、10～19 書式で用いられているのが 39、５～9 書式に用いられている項目は 22 であった。書式横断的に登録頻度の高い項目に対する既存情報からの活用性をみると、DPC 調査情報から患者基本情報（生年月日、性別、郵便番号等）、入退院日、入院時傷病名、身長、体重等を含めた入退院情報、手術術式など、また、SS-MIX2 標準化ストレージより検査結果、処方・注射情報が確実に取得できるもの、また、取得できる可能性のあるものとして併存症、合併症などが上げられたが、術者、時間、出血量、輸液量などの手術時情報は、その実施情報などの非標準情報からの取得が必要であり、氏名イニシャルや各種の治療実施の有無などは蓄積データの集約や値変換等の加工によって活用可能な情報であった。一方、ASA-PS 分類や創感染などの合併症、TNM 分類やリンパ節郭清などは各種レポートのテキスト情報にある場合が多く、構造化データとしての取得は困難な情報であり、それらの情報取得が重要課題と考えられた。

(2) NCD 症例登録実例における情報の不一致割合の検証

①登録症例の完全性の確認

2014～2015 年における消化器外科専門医手術症例、2020 年の乳癌症例については、手術日が前者では 2014 年 1 月 1 日～2015 年 12 月 31 日までの症例、後者では 2020 年 1 月 1 日～12 月 31 日の症例について DPC 調査情報をもとに確認し、いずれも症例登録の漏れがないことが確認された。

②登録情報の不一致の確認

i. 消化器癌（胃癌・大腸癌）手術例：2 年間における対象登録数は 192 症例であり、胃癌 90 例、結腸・直腸癌 101 例、胃および大腸の重複癌 1 例、CRF 別では消火器外科専門医共通項目術式が 98 例、医療水準評価対象術式 94 例であった。術前項目では、喫煙歴(11%)、体重(6%)、術前輸血の有無(5%)、入院時診断(5%)などで比較的多くの不一致を認めた。身長は不一致を認めなかったのに対し、体重は誤入力と考えられるものに加え、異なる日の測定値登録例が見られた。入院時診断は、病変部位の詳細不明としたものが 9 例中 4 例など、局在の記載に不備のあるものを認めた。術中、術後情報で不一致割合が多かった項目は、麻酔種別(10%)、退院日(5%)、TNM 分類の T、N(4%)などであった。麻酔種別では、ほとんどが全麻と硬膜外麻酔の併用の有無による違いであった。退院日は登録が確認できないものを 9 例中 4 例認めた。TNM 分類は手術時の診断に基づいた記載が求められるが、T では術前診断、N では病理診断情報の登録にて不一致となったものを認めた。検査項目では、直近日とは異なる測定日による不一致が最も多く、施設と NCD との基準値範囲の違いが誤判断に影響した可能性のある検査項目が 1～7%、数値の不一致が 1 割を超える項目(HbA1c)が認められた。

ii. 乳癌手術例：両側例が 3 例あり、111 件を対象とした。消化器外科専門医領域と同様に数値型の入力である体重の不一致（12%）、腫瘍径（10%）、浸潤径（7%）、リンパ節転移個数（5%）に認められた。また、組織型(8%)、TNM 分類では M 分類には不一致はなかったが 5～6%で TN 分類の不一致例を認めた。ホルモンレセプター等のエストロゲンレセプター、プロゲステロンレセプターの不一致が 8～10%で、術前の結果と思われる情報の登録によるものがその半数を占めた。一方、併存症については、糖尿病や高血圧で NCD 登録頻度が高かった

が NCD 登録がありながら傷病名に登録されていない例が少なくない状況にあった。また、明確な診断基準が指定されている腎機能障害（クレアチニン≧1.0mg/dl、または、eGFR＜60ml/min/1.73m²）のいずれかを満たすのは 17 例、2 基準ともに満たすものは 2 例で、15 例は eGFR 基準のみ満たす例であったが、腎機能障害として登録された例はなかった。

（3）大規模言語モデルによる病理報告における固有情報の取得

Prompt 内容により LLM より得られた結果が大きく異なり、Prompt の工夫にて更なる精度向上が見込まれた。現状の結果としては、胃癌、大腸癌では、大きさや T 分類、静脈侵襲、リンパ管侵襲のなどの情報抽出の割合（正確度 8 割以上）が高く、組織型、肉眼的分類、浸潤増殖様式、近位・遠位断端の抽出割合が低い結果（正確度 5 割以下）であった。なお、静脈侵襲、リンパ管侵襲については Prompt の中で明確に指示したことで抽出率が高くなったが、断端などの項目についても Prompt で適切に指示することで結果が良くなることが確認された。一方、乳癌手術症例の病理報告では、消化器癌に比し全体の抽出結果は良好で、静脈侵襲、リンパ管侵襲とともに、核異型、組織グレード、各種のレセプターの結果の抽出結果も良好（正確度 9 割以上）であったが、組織型、大きさ、断端などの結果は不十分であった。特に、大きさ（腫瘍径）は浸潤径と報告書内の記載が同じく長さの表現であることから入れ違えや重複した取得により悪化したと考えられた。

（4）登録データ検証機能を有するプロトタイプシステムの構築

① 検証システムにおける機能要件：

i. 検証システムにおける検証項目の範囲

結果（1）のように登録書式（CRF）の項目は多岐に渡る一方、該当書式限定の項目が多く書式横断的な共通項目は限定的な事から、頻度が高く、確実な診療情報として取得できる項目から活用することを基本とした。

ii. システム機能要件

a. システムにおける種々のデータの格納基盤

(ア) 既存データの格納基盤

登録に要求されるデータは、ICD10 コード、術式コードなどを含むテキスト情報、検査結果などの数量情報、また、生年月日や入退院日の日付情報、有り無しの 2 値情報などからなり、さらに項目数の非常に多いデータベース Table 基盤が重要である。データの正規化とレスポンス低下を防ぎ多様な登録項目への対応できる柔軟性とデータ処理速度を保つため、データ縦持ちを拡張した DPC 調査の様式 1 の Table 構造で ID、入院日、手術日をキーとし、登録に関連する情報項目毎に必要なペイロードを有する単一参照型データベースを採用した。

(イ) CRF 別の関連情報の収集と提示のためのマスターの整備と格納

CRF 毎に登録されるデータ内容が異なり、また、性別等の 2 値情報やリスト情報など値変換が必要となる場合が多いことから、CRF 別のマスター設定が必要であり、それらの項目毎に関連情報の整理が必要である。一方で、汎用的なデータ収集後に直接、あるいは加工された既存情報を登録データと比較するが、そのような登録情報と既存情報を対とする紐付けマスター情報が必要である。このようなマスター登録とその格納、および、登録サイトからのダウンロードデータの取込のためのマスタ等の管理が必要となる。

b. データソースからの取得機能の整備

医療機関で取得可能なデータのうち、施設内の二次利用基盤にある DPC 調査情報（様式 1・レセプトデータであるＥＦファイル等）、医療情報の交換・共有基盤用 SS-MIX2 標準ストレージ情報（検査結果等）が現行の主要な一般的な標準化データソースである。しかし、それ以

外の非標準化データである各種オーダ・実施情報等のデータについてもデータウエアハウス（DWH）等により、さらには、退院時サマリや病理レポート等の各種の報告書等のテキスト情報についても施設によっては活用に向けて整備されている。また、今後は HL7 FHIR による標準化された相互運用性のあるデータ活用に向けた整備が進められると考えられるため、現状の取得から REST Api による取得環境への移行を検討する必要があり、今回、その基盤整備を行った。

c. 汎用的なデータ情報収集・加工基盤（汎用インターフェイス）の整備

症例登録においては、標準化データや構造化データそのものが直接には活用ができない場合が少なくない。例えば、患者氏名のイニシャルや検査値の期間最高値、さらに規定の基準に照らし該当するか否かの判断等、データ加工等により比較可能なデータとするため、ロジックを用いたデータ加工や判断を加えることで情報活用範囲の拡大をはかる必要がある。そのような汎用ロジックの登録と実施環境（汎用インターフェイス）の整備が求められる。一方、各種検査や手術、処置等の実施記録、退院時サマリなどのテキスト情報に埋め込まれた情報取得を可能とするため LLM による固有情報の抽出が望まれる。(3) の病理レポートにおける試みのように今後の抽出精度の向上が必須であるが、近い将来での実現可能性が高いと考える。

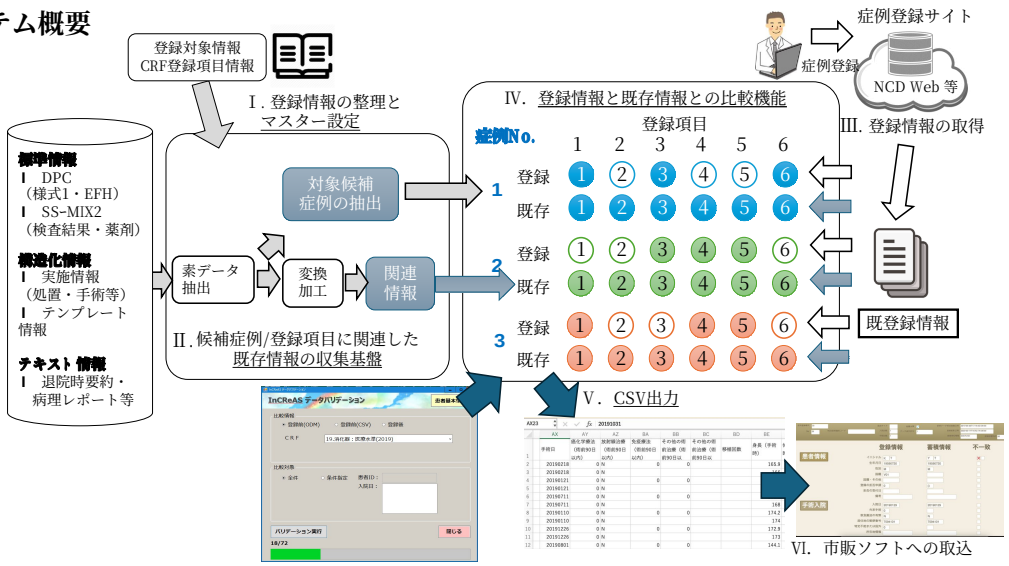
d. データ検証機能環境の構築

検証の対象登録データは各施設での登録データについては登録サイトからのダウンロードした既登録情報の検証、および、支援システム等からのアップロード環境がある場合には事前検証を可能とすることが重要であり、また、後者の場合には、手入力を併用する場合も含めて精度向上につながる登録支援システムの構築を前提としている。一方、ダウンロードデータと既存データを比較する場合、コード化されたデータに対し、比較可能とする変換リストと変換機能、数値の桁数管理なども不可欠である。また、比較可能な既存データが取得できない場合には、その項目自体には自動判断の対象にしないなどの対応が必要となる。

② 登録症例データ検証を有するプロトタイプシステムの概要（図）

上記のシステムの機能要件をもとに、該当 CRF 別の項目マスターにて登録データの検証を行うことを前提に、登録された情報と既存情報を並べた CSV 出力を行うシステムを構築した。また、本システムでは、該当 CRF が対象とする症例について、登録漏れがないかの確認を行うための患者一覧機能を有し、条件（診療科、手術日、入院日等の期間指定）設定での絞り込みと手術名、病名等での並べ替えを可能とし、さらに該当患者毎の基本情報の参照も可能とした。今後の可用性等の検討が課題である。

図：システム概要



５．主な発表論文等

〔雑誌論文〕　計0件

〔学会発表〕　計2件（うち招待講演　0件／うち国際学会　0件）

１．発表者名 石田 博，平野 靖，永野 浩昭
２．発表標題 症例登録における入力精度の検証と正確性向上に資す支援システムのあり方の検討
３．学会等名 第42回医療情報学連合大会
４．発表年 2022年

１．発表者名 石田　博・樫部公一・平野　靖
２．発表標題 症例登録とVerification支援-Web入力型症例登録における精度-
３．学会等名 第48回中国四国医療情報学研究会
４．発表年 2022年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

―

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究分担者	永野　浩昭 (Nagano Hiroaki) (10294050)	山口大学・大学院医学系研究科・教授 (15501)	
研究分担者	平野　靖 (Hirano Yasushi) (90324459)	山口大学・医学部附属病院・准教授 (15501)	

６．研究組織（つづき）

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （機関番号）	備考
研究協力者	中津井　雅彦 (Nakatsui Masahiko)	山口大学・大学院医学系研究科・教授（特命） (15501)	

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

共同研究相手国	相手方研究機関