厚生労働科学研究費補助金
（政策科学総合研究事業（統計情報総合研究事業））
分担研究報告書
ICD-11 の適用を通じて我が国の死因・疾病統計の向上を目指すための研究
「主疾患を決定付けるパラメータの決定とその予測（DPC/PDPS レコードを「教師」として）」

研究分担者　野口晴子　早稲田大学政治経済学術院

研究要旨

レセプト情報・特定健診等情報データベース（National Database：NDB）は、罹患者数の多い疾患から、患者数が極めて少ない難病に至るまで、治療行為のパターンや詳細な医療費を分析できる潜在力を秘めている。ところが、NDB には、レセプト記載が患者の「主疾病」を必ずしも正確に反映されていないという課題が指摘されており、医学領域では、疾患情報を軸とした分析を行う際の NDB の信頼性は低く、取り扱われる研究課題に限界があった。

そこで、本研究では、NDB サンプリング（トライアル）データを用い、NDB 上に格納されている包括医療費支払制度下での支払い情報に基づき、「教師あり機械学習」によって「主疾患」を分類・特定する手法の開発を試みる。

本研究では、 NDB の中でも、制度上、主傷病が特定されている、つまり、「答え」をもっている DPC/PDPS レコードを「教師」として学習した、主傷病の予測モデルを基に、レセプトデータから主傷病を予測する。第 1 に、本厚生労働研究事業で承認を受けた NDB サンプリングデータを基にして、SQLite（エスキューライト）を用いデータベースを構築し、本番の学習環境では PostgreSQL（ポストグルエスキューエル）を用いる予定である。

第 2 段階では、DPC レコードを基にして、主傷病に寄与する重要なパラメータの選定を行う。例えば、主傷病を説明する説明変数は、年齢、性別、治療行為や処方パターン、副疾患や既往歴等が考えられる。これらの説明変数と、目的変数である主傷病にどれくらいの相関があるかを可視化する。可視化には、クラスタリングや Uniform Manifold Approximation and Projection（UMAP）等の次元削減手法を用いる予定である。

第 3 段階では、予測モデルを構築する。予測モデルでは、因果を単純化し、説明変数を先決変数のみで構成される変数に変形する操作を行う。その後、Random Forest（RF）によって説明変数を選定し、各パラメータの重要性に応じて、重みづけを行って深層学習を行う。

予測モデルとして RF を選定した理由は、解釈性と予測精度との間にはトレードオフがあることが知られているが、 Deep Learning （DL）、Support Vector Machine（SVM）、Linear Regression（LR）、Decision Tree（DT）等と比べ、いずれの点でもバランスがとれている手法だからである。

2023 年度から現在まで、第 1 段階のデータベース構築に当たっているが、2024 年度において、予測モデルの構築を開始する予定である。

A. 研究目的

レセプト情報・特定健診等情報データベース（National Database：NDB）は、罹患者数の多い疾患から、患者数が極めて少ない難病に至るまで、治療行為のパターンや詳細な医療費を分析できる潜在力を秘めている。ところが、NDB には、レセプト記載が患者の「主疾病」を必ずしも正確に反映されていないという課題が指摘されており、医学領域では、疾患情報を軸とした分析を行う際の NDB の信頼性は低く、取り扱われる研究課題に限界があった。

そこで、本研究では、NDB サンプリング（トライアル）データを用い、NDB 上に格納されている包括医療費支払制度下での支払い情報に基づき、「教師あり機械学習」によって「主疾患」を分類・特定する手法の開発を試みる（図 1 参照）。

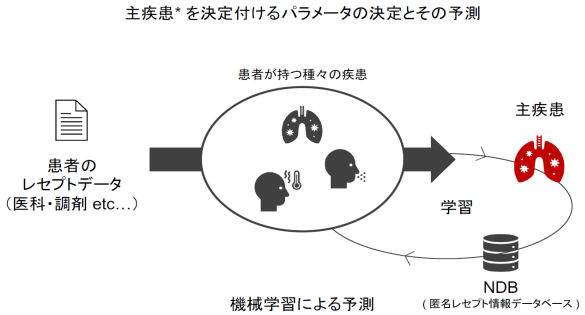

図 1　研究目的の概念図

B. 研究方法

NDB の中でも、制度上、主傷病が特定されている、つまり、「答え」をもっている DPC/PDPS レコードを「教師」として学習した、主傷病の予測モデルを基に、レセプトデータから主傷病を予測する。

現在、本厚生労働研究事業で承認を受けた NDB サンプリングデータを基にして、データベースの構築を行っている段階である。システムは、図 2 に示す通り、実際に機械学習を行う「Processing」と NDB を基に構築した「データベース」に分類される。「Processing」の部分では、プログラム言語の 1 つである Python を用いたデータプロセシングと Sk-learn を援用した Random Forest（RF）および深層学習によって予測モデルを構築する計画である。データベースについては、開発環境として、SQLite（エスキューライト）を用い、本番の学習環境では PostgreSQL（ポストグルエスキューエル）を用いる予定である。

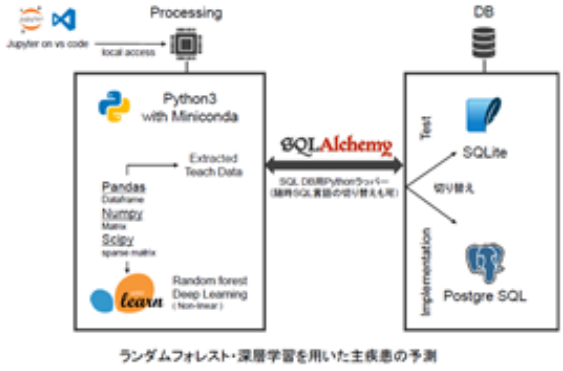

図 2　システムの構成

第 2 段階では、DPC レコードを基にして、主傷病に寄与する重要なパラメータの選定を行う。例えば、主傷病を説明する説明変数は、年齢、性別、治療行為や処方パンターン、副疾患や既往歴等が考えられる。これらの説明変数と、目的変数である主傷病にどれくらいの相関があるかを可視化する。可視化には、クラス

タリングや Uniform Manifold Approximation and Projection（UMAP）等の次元削減手法を用いる予定である。

第3段階では、予測モデルを構築する。予測モデルでは、因果を単純化し、説明変数を先決変数のみで構成される変数に変形する操作を行う。その後、ランダムフォレストによって説明変数を選定し、各パラメータの重要性に応じて、重みづけを行って深層学習を行う。

C. 研究結果

現在、データベース構築中であるため、今年度については、報告すべき研究成果はない。

D. 考察／E. 結論

使用するモデルの選定理由としては、モデルの解釈性と予測精度にはトレードオフがあることが知られている。例えば、Deep Learning （DL）や Support Vector Machine（SVM）等は、予測精度は極めて高いが、モデルの解釈性が低くく、主疾患に影響を与える因子がブラックボックス化する恐れがある。他方で、Linear Regression（LR）や Decision Tree（DT）は、モデルの解釈性は高いが、予測精度が低すぎるため、主疾患の予測という目的を達成出来ない可能性が高い。とりわけ、今回は、NDB という膨大な情報量を処理しなければならないデータを用いるため、LR や DT は不向きだと考えられる。

そこで、本研究では、モデル解釈性と予測精度が中程度である RF を用いる予定である。RF は、説明変数が目的変数に与える影響を定量的にしめすことが可能である同時に、データ自体は、ハイパーパラメータが多く、 最適化しやすいと考えられるためである。

F. 健康危険情報

特に無し。

G. 研究発表

1. 論文発表

特に無し。

2. 学会発表

特に無し。

H. 知的財産権の出願・登録状況(予定を含む)

1. 特許取得

特に無し。

2. 実用新案登録

特に無し。

3. その他

特に無し。

参考文献

Scikit-learn – Machine Learning in Python。 https://scikit-learn.org/stable/（閲覧日：2024 年 5 月 24 日）

厚生労働科学研究費補助金
（政策科学総合研究事業（統計情報総合研究事業））
分担研究報告書
ICD-11 の適用を通じて我が国の死因・疾病統計の向上を目指すための研究
「ICD-11 適用によるがん罹患集計への影響：がん登録状況からみた課題の整理」

研究分担者　奥山　絢子　聖路加国際大学看護学研究科　教授
研究分担者　東　尚弘　東京大学大学院医学系研究科　教授

研究要旨
国際疾病分類第 11 回改訂版（ICD-11）は、2019 年の世界保健総会で採択され、2022 年 1 月に発効となった。本研究では、ICD-11 適用によるがん登録罹患集計への影響を評価するため、がん登録の現状と課題を先行研究から整理することを目的とした。結果、ICD-11 適用によるがん罹患集計への影響については、がん登録の実施状況を鑑みると、がん登録のコーディングや登録対象の範囲など登録のルール、医療機関における病名等を用いた登録対象例の見つけ出しへの影響、そして ICD-O に基づいて登録された情報の ICD-11 へ変換による集計値の変化、それぞれについて検討する必要があることがわかった。

A．研究目的
本研究では、がん登録を用いた ICD-11 適用によるがん罹患集計への影響を評価するために、まずがん登録（全国がん登録と院内がん登録）の状況を分析し、ICD-11 適用による影響の評価における課題を整理することを目的とした。

B．研究方法
がん登録(全国がん登録と院内がん登録)の状況を整理するため医学中央雑誌 Web を用いて、登録開始時から 2023 年 8 月 30 日までに公表された文献を対象に、がん(がん、腫瘍）と ICD（ICD 分類、ICD）のキーワードとシソーラスを組み合わせて文献検索を行った。ICD とがん登録、ICD とがん罹患集計について記述されていた文献を選定した。なお、解説を含め、がん登録と ICD に関する記述がある文献すべてを対象とした。タイトルと抄録による 1 次スクリーニング後、論文全体を読む 2 次スクリーニングを行った。この結果を踏まえ、日本におけるがん登録の状況を鑑み、ICD-11 適用におけるがん罹患集計への影響を評価する際の課題を整理した。

（倫理面への配慮）
本研究に含まれる情報は、すべて過去に発表され、一般に入手可能なものである。そのため、本研究では施設の倫理委員会の承認を得る必要はなかった。

C．研究結果
1）文献検討
検索の結果、540 件がヒットし、そのうち選定基準を満たす可能性のある 32 件に

ついて2次スクリーニングを行い、7文献が基準に合致した。検討されていた課題は次の3つに分類された。

① ICDと国際疾病分類腫瘍学（International Classification of Diseases for Oncology, ICD-O）のコード体系とがん登録

がん登録では、ICD-Oに基づいて登録が行われている。これは、腫瘍の局在分類（部位）と形態学的分類（組織型）の2つを別々にコードし、組み合わせて、がんを表す仕組みとなっている。がんは、同じ臓器であっても様々な組織型のがんが生じること、また逆に同じ組織型のがんが様々な臓器に生じることを考えるとがんの登録の実態に即したものであるとされる。これに対し、ICD-10の分類法では、部位が主な分類の軸であり、腫瘍の詳細なコードはなかった。ICD-11では、より組織型を重視した分類となり、腺癌や扁平上皮癌などの頻出の組織型については、分類項目が設けられた（中山,診療情報管理 2020）。また、拡張コード（Extension code）が準備されており、分類上、腫瘍においても詳細な分類が可能となった。ICD-Oに基づいて登録を行うがん登録では、がんのコードを付ける際に、WHO Classification of Tumoursを参照しつつ登録が行われている。近年改訂されたWHO Classification of TumoursにもICD-11のコードが掲載されており、ある程度統合が可能な仕組みとなった。しかしながら、がん登録を行う側からみると、ICD-11のコードには規則性が弱く、がん登録の実務者にとってはICD-Oの方が使いやすい可能性があることが指摘されていた（東,日本診療情報管理学会学術大会抄録集 2022）。

また、ICD-10では、cervical intraepithelial neoplasia 3（CIN3）のみが、子宮頸部（Cervix uteri）のcarcinoma-in-situ（CIS）として報告されていたが、ICD-11では、CIN2も子宮頸部のCISの中に含まれている（コード：2E66）。そのため、これらCIN2を含めて登録がされるようになると子宮頸部のがんの罹患数が見かけ上増加する可能性が危惧されていた（Lu,日本保険医学会誌,2021）。

② 医療機関におけるがん登録症例の見つけ出し

がん登録は、研修を受け認定を受けたがん登録実務者が中心となって、各医療機関で登録作業が行われている。がん登録を行う医療機関では、届出漏れのないように、施設内の情報を検索し、効率的に症例を見つけだすことが課題となっている。漏れのない症例の登録を行うための方法として、これまでレセプトの病名を用いた症例検索が提案され、実施されている（小原,医療情報学 2019；小原、診療情報管理 2019；小原,日本医療マネジメント学会雑誌 2022）。今後、医療機関で新たにICD-11を用いて病名情報が管理されるようになった場合、ICD-11の病名で効率的にがん登録の登録対象症例を見つけ出す方法を新たに検討する必要がある。

③ がん罹患集計の比較可能性を担保したICD-11準拠の集計分類の検討

WHOは、これまでICDを定期的に改訂してきた。前回、ICD-9からICD-10に改訂された際には、ICD-9とICD-10の変換表が作成され、がん罹患集計や死亡統計による影響について検討がなされた（味木,厚生の指標 1997）。これまで

ICD-10では、項目（章）全体が悪性・良性・上皮内などに大きく分けられていたが、ICD-11ではこれを引き継ぎながらも、中枢神経系、造血器・リンパ組織の新生物はこの枠組みから除かれ、別の分類となった（中山、診療情報管理、2020）。

D．考察

がん罹患集計では、ICD-Oに基づいて登録されたがん情報をICD-10に変換し、集計を行うことで、死亡統計との整合性を担保してきた。がん罹患集計におけるICD-11適用による評価においては、ICD-OからICD-11に変換した場合の評価を行う必要がある。しかし、現時点ではInternational Association of Cancer Registries（IACR）からはICD-OからICD-11への変換表は公表されていない。2023年11月に行われたEuropean Network of Cancer Registries（ENCR）とIACRの合同学術集会において、IACRのがん罹患集計担当者は、がん登録において、現状のICD-O 3版を継続使用するか、ICD-O4版を新規作成し使用するか、あるいはICD-11を用いた登録を行うかについて、2021年にIACR会員に調査したところ、89.9%がICD-O4版の作成を選択したと報告している。これを受け、今後多様化するがんを登録するためにICD-O4版を作成する予定であること、そしてICD-O4版からICD-11への変換ルールを検討する必要があると報告している（Znaor A, ENCR-IACR Scientific Conference in Gradana, Spain 2023）。そのため、がん罹患集計へのICD-11適応による評価においては、今後日本でもICD-O4版に準拠した登録が行われるようになった場合には、ICD-O4版からICD-11への変換におけるがん罹患集計への影響も検討する必要がある。

また、本調査の結果、がん登録情報に基づいたICD-OからICD-11への変換による集計値の変化だけでなく、医療機関における病名登録のICD-11適用によるコーディングやがん登録の対象範囲、医療機関におけるがん登録の対象となる症例の見つけ出し方法等を含めて検討する必要性があることが分かった。日本では1999年より標準病名マスターが公開され、2022年にはレセプトに記載すべき傷病名とされたことから広くICD-10をベースにした標準病名マスターが使用されている。しかし、標準病名マスターとICD-11の対応付けが行われたが、必ずしも一対一に対応していないことが課題とされている（今井,ICD-11と標準病名マスターとのマッピングに関する研究　令和4年度総括・分担研究報告書、2023）。このように、がん登録が、がん登録実務者を中心に各医療機関で登録がされ、それらの情報を用いてがん罹患集計が行われている現状を踏まえると、病名等の変更による登録作業への影響も含めて、がん罹患集計への影響を考慮する必要がある。また、がん登録は標準登録様式や多重がんルール等の一定のルールに基づいて登録がされる。ICD-11を適用した場合に、ICD-11を踏まえて再度がんの登録範囲や登録ルール等の規定を見直す必要があるだろう。

ICD-11では、中枢神経系、造血器・リンパ組織の新生物は別の枠組みに位置付けられた。しかし、ICD-10においても、これらのがんは臓器がんとは別扱いであった点は同じであり、全国がん登録罹患集計の基本分類A表では、これまでも脳・中枢神経系（ICD-10のコード：C70-72）、悪性リンパ腫（同C81-85,C96）、多発性骨髄腫（同

C88、C90）、白血病（C91-96）として集計がなされてきた。これらの対応については、今後詳細に検討する必要がある。がん罹患集計は、経時的にがんの罹患がどのように変化しているのかを捕捉する重要な情報源である。今後、こうしたがん罹患集計をどのように活用していくか、そして従来の集計結果との比較可能性をどのように担保するのかを検討する必要がある。

更に、現在がん診療連携拠点病院等を中心とする院内がん登録では、ICD-O3.2 版、全国がん登録では、ICD-O3.1 版に準拠した登録が行われている。それぞれコード分類異なることから、ICD-O3.1 版と ICD-O3.2 版それぞれにおいて ICD-11 準拠でがん罹患集計を行った際の値の変化について検討する必要がある。また、がん登録実務においては、登録対象かどうかを性状コードによって決められてきたものの、改訂によって性状コードが変更になるものもあり、これを ICD-11 ではどのように定義していくのかと言ったことも考えなければならない。

E．結論

ICD-11 適用によるがん罹患集計への影響については、がん登録の実施状況を鑑みると、がん登録のコーディングや登録対象の範囲など登録のルール、医療機関における病名等を用いた登録対象例の見つけ出しへの影響、そして ICD-O に基づいて登録された情報の ICD-11 へ変換による集計値の変化、それぞれについて検討する必要がある。

G．研究発表

1．論文発表

なし。

2．学会発表

なし。

H．知的財産権の出願・登録状況（予定を含む。）

1. 特許取得

なし。

2. 実用新案登録

なし。

3.その他

特記すべきことなし。