別紙 3

令和 4 年厚生労働科学研究費 補助金
政策科学総合研究事業(臨床研究等 ICT 基盤構築・人工知能実装研究事業)
分担研究年度終了報告書

千年カルテの匿名加工医療情報を利用した 入院後敗血症の予測モデルの開発における課題と対策の検討

研究分担者　松田　敦義（株式会社ログビー）

研究要旨

千年カルテの匿名化された医療データを使用して、入院後の敗血症を予測するための機械学習モデルを開発した。本研究グループの 2020 年度の研究と今回の研究で、ランダムフォレストを用いた予測モデルの評価指標を比較したところ、本研究の評価指標（AUC：0.769、F 値：0.528、適合率：0.536、再現率：0.523）がいずれも過去の研究の指標よりも高かった。千年カルテの匿名化された医療データの安全性を担保しながら機械学習モデルを開発するための現状のプロセスの問題点と、その解決策について考察した。問題点として、機械学習モデルの精度と、研究に要する時間の 2 つの観点をあげた。この問題に対し、プロセスの改善策として、テスト環境の準備とデータ集計または機械学習モデル開発のためのテンプレートの提供を提案する。

A. 研究目的

近年、次世代医療基盤法に基づく匿名加工医療情報の活用が注目されている。医薬品の使用による安全性・有効性の研究開発や、患者への個別化医療の提供などを目的とした取り組みが始まっており、期待が高まっている。しかし、その安全性と、AI 活用などにおける有用性についての知見は少ない。

千年カルテプロジェクト（以下、千年カルテと表記）は、2015 年に AMED 研究公募事業に採択された全国共同利用型国際標準化健康・医療情報の収集及び利活用に関する研究である [1]。一般社団法人ライフデータイニシアティブ（以下、LDI と表記）は、次世代医療基盤法の認定匿名加工医療情報作成事業者として国から認可を受けた、千年カルテの EHR（Electronic Health Record）を活用して医療の質・効率性や臨床研究等の研究開発に貢献することを目的とした法人である [2]。

本研究グループは、これまで機械学習と説明可能な AI を用いて、入院後合併症の発症予測と特徴量の寄与度の可視化等の研究を行ってきた [3]。

本研究では、愛媛大学が実施する「認定匿名加工医療情報作成事業者が保有する医療情報を活用した、匿名加工医療情報の作成に依らない AI 研究の実現可能性の検討」研究の一環として、千年カルテの匿名加工医療情報のデータを活用し、LDI の定める安全性を担保するためのプロセスに従いながら、機械学習モデルの開発を行い、その課題と対策の検討を行った。

機械学習による予測モデル開発では、学習に用いる特徴量の探索やデータの前処理を行うため、データを集計、確認しながら方針を定めることが多い。一方で、千年カルテのデータを利用するに当たり、安全性担保のため、開発プロセスにおいては以下の条件を守る必要があった。

1, 個票データを確認することはできないため、統計データのみを抽出する。

2, データベースを直接参照することはできないため、LDI を通して集計結果やエラーなどのヒアリングをする必要がある。

本研究では、上記のような条件に従い千年カルテの安全性を担保しながら、機械学習モデルの開発を行い、そのプロセスの課題の抽出と対策の検討を行うことを目的とした。

機械学習モデルとしては、入院時の患者の状態を元に、入院後の敗血症の発症を予測するモデルを開発した。機械学習モデルの精度評価において、本研究グループが 2020 年度の第 40 回医療情報学連合大会で発表した、宮崎大学医学部附属病院のデータを用いた研究結果との比較を行った。

B. 研究方法

1. 開発プロセス

以下に示す 2 つのステップ A, B で、機械学習モデルの開発を進めた。

A, 宮崎大学医学部附属病院のデータセンター（以下、宮崎大学環境と表記）に、千年カルテの環境と近いデータベース、解析サーバーを構築し、そこでコンテナ型の仮想環境（Docker）を作り、宮崎大学医学部附属病院のデータを用いて前処理、学習と推論を行うプログラムの開発を行った。

宮崎大学環境での機械学習モデル開発の概要を図 1 に示す。

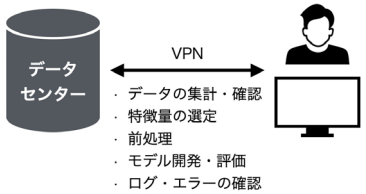

図 1　宮崎大学環境での機械学習モデル開発

B, 千年カルテの本研究用コンピューター（以下、千年カルテ環境と表記）に、A で作成した Docker イメージを持ち込んでコンテナ化し、LDI と連携して試行錯誤しながら開発したプログラムを改修し、機械学習モデルの開発を行った。千年カルテ環境における開発では、以下の手順を繰り返して進めた。

1, プログラムと実行手順を LDI に提供

2, LDI によるプログラム実行とエラー等のフィードバック

3, プログラムを改修し LDI に提供

4, LDI 担当者がデータセンターに入室してプログラムを実行し、データ統計やログを取得

千年カルテ環境での機械学習モデル開発の概要を図 2 に示す。

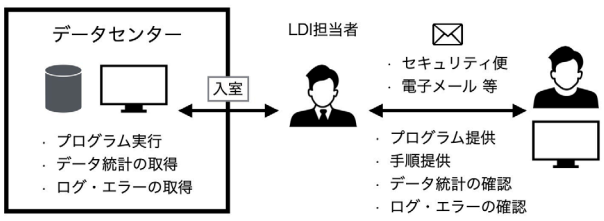

図 2　千年カルテ環境での機械学習モデル開発

最終的に開発した機械学習モデルの評価結果と特徴量重要度を LDI よりセキュリティ便で受領した。

2. 対象データ

千年カルテに登録された 20 の医療機関の 2015～2021 年度の DPC を使用した。また本研究での対象とする患者グループの入院時疾患は ICD10 分類における C00-D48 “新生物＜腫瘍＞” とし、予測対象の入院後合併症を ICD10 分類における分類表記で “敗血症” を含む疾病を対象とした。

予測に用いる特徴量として、DPC の年齢や体重などの基本情報、主傷病や入院時併存症等の診断情報を使用した。使用した特徴量は全部で 166 個であった。

3. 前処理

機械学習に用いるデータは入院単位のデータとなるように整形した。また予測の正当性を守るために、入院時に敗血症を発症している患者は除いた。その他の前処理として、カテゴリデータのフラグ化、欠損のあるデータの削除等を行った。結果として、合計 71,433 件のデータとなった。

以上の処理の後、データ全体を予測モデルの学習に 75%、評価に 25%となるように分割した。学習データについて、データ数が敗血症の発症：非発症 = 1：20 になるように非発症データをダウンサンプリングした。

4. 予測モデル

本研究では、プログラムの実装がしやすく、予測に寄与した特徴量の重要度を得ることができる Random Forests[5]（以下、RF と表記）を機械学習モデルとして用いた。前処理済のデータで学習させ、各ハイパーパラメータは交差検証を行い決定した。

（倫理面への配慮）

本研究は、千年カルテの利用目的等審査委員会[4]。による審査を受け、承認された上で研究を進めた。

C. 研究結果

1. 開発プロセス

千年カルテ環境における開発の手順を進めるにあたり、LDI との電子メールのやり取りは 100 回近くに及んだ。やり取りを通して、時間がかかったプロセスを表 1 に示す。定性的に時間がかかったと考えられる順番で記載した。

各プロセスの中で、「プログラム実行で発生したエラーの確認と対応」に多く時間がかかった。理由としては、エラー時のログについて、LDI にて内容の確認を行い、電子メールにて伝達をする運用であったため、エラーの全体像の把握が難しく、推測してプログラムを直す必要があったことが大きい。また、LDI 担当者もデータセンターのセキュリティルームに入室した上でプログラムを実行し、データ統計やログの取得を行う必要があったため、LDI の運用プロセスにおいても時間がかかった。

表 1　時間がかかった開発プロセス

	開発プロセス
1	プログラム実行で発生したエラーの確認と対応
2	データ授受等、研究の進め方の確認
3	データベース内のテーブルの概要確認と、対象テーブルの修正
4	データベース接続情報の確認と、接続情報の修正
5	Docker イメージでインストールされるライブラリの確認と承認
6	千年カルテから取得するデータ種類の確認と承認

2. 予測モデルの精度評価

予測精度の評価指標として AUC（Area Under the Curve）及び F 値（適合率と再現率の調和平均）、適合率、再現率を用いた。

表 2 に、本研究グループが 2020 年度に発表した RF を用いた敗血症の予測モデルと、本研究における敗血症の予測モデルの評価指標の比較を示す。各指標の値は、敗血症を発症するケースとしないケースの平均をとったものである。いずれの評価指標においても、本研究の精度が高い結果となった。

ただし、2020 年度の研究と本研究では用いた特徴量やデータの母数などの前提条件が異なるので、一概に単純比較はできない。2020 年度の研究では、DPC だけでなく DWH からバイタル、検体検査、実施オーダー情報（薬剤処方、手術、放射線）等も利用し、特徴量としては 358 個を用いたのに対し、本研究では DPC のみから 166 個の特徴量を用いた。また、2020 年度の研究では宮崎大学医学部附属病院のみの前処理済データ 13,527 件を利用したのに対し、本研究では千年カルテの 20 の医療機関の前処理済データ 71,433 件を用いた。

2020 年度の宮崎大学医学部附属病院のみのデータ数に対して、本研究の 20 の医療機関のデータ数が 5 倍程度と、施設数の比率の 20 倍に対して小さかった 1 つの要因として、2020 年度では前処理で欠損値のデータ補完を行ったのに対し、本研究では欠損があるデータはすべて削除したことが考えられる。

表 2　2020 年度の研究と本研究の評価指標の比較

研究	AUC	F 値	適合率	再現率
2020 年度	0.727	0.424	0.374	0.490
本研究	0.769	0.528	0.536	0.523

表 3 に、本研究において、敗血症の発症なし、発症ありで分けた場合の評価指標を示す。各指標において発症なしの値は非常に大きいが、発症ありの各指標、特に再現率の値が小さい事がわかる。このことから、表 2 の本研究の評価指標は発症なしの評価結果により値が高められているといえる。発症ありの精度を高めることが今後の課題である。

表 3　本研究における敗血症の発症別の評価指標

敗血症	F 値	適合率	再現率
発症なし	0.997	0.996	0.998
発症あり	0.058	0.075	0.048

3. 予測モデルの特徴量重要度

本研究の予測モデル（RF）の特徴量重要度について、重要度が大きい順で上位 10 個を表 4 に示す。これを見ると体重、身長、年齢が上位に来ていることがわかる。2020 年度の研究で用いたバイタル、検体検査、実施オーダー情報（薬剤処方、手術、放射線）等の特徴量を追加すると、精度向上に寄与する重要な特徴量が変化するのではないかと期待される。

表 4　本研究における予測モデルの特徴量重要度

	特徴量	重要度
1	体重	0.148074051
2	身長	0.130524826
3	年齢	0.075204027
4	併存症_真菌症	0.036117323
5	併存症_原虫疾患	0.034309911
6	併存症_悪性新生物＜腫瘍＞	0.025880966
7	併存症_高血圧性疾患	0.025158698
8	併存症_糖尿病	0.022879724
9	併存症_食道、胃、および十二指腸の疾患	0.022592592
10	性別	0.021793955

4. 開発プロセスの課題

本研究での千年カルテ環境の開発プロセスにおいて、機械学習モデルの精度の観点と、研究にかかる時間の観点で、それぞれ課題が考えられる。

機械学習モデルの精度の観点では、個票データ

の確認ができないため、データを確認しながら特徴量の内容（例えば検査値の単位やノイズによる影響など）や表記ゆれなどの確認ができないため、データが正しい前提で開発を行う必要があり、実際に何かデータの問題を内包していた場合には、精度が落ちる等の問題が考えられる。

研究にかかる時間の観点では、図 2 に示したように、データの統計量や結果は LDI を通して確認する必要があるため、データを直接確認しながら機械学習モデルの開発を行うよりも時間が多くかかる問題がある。教師あり機械学習モデルを開発する上では、様々なデータを集計し、適切な特徴量を選定するプロセスが必要である。本研究では、宮崎大学環境にて選定した特徴量をそのまま千年カルテ環境でも用いたため、表 1 の「時間がかかった開発プロセス」には含めていないが、より精度の高い機械学習モデルの開発をするためには、特徴量選定に多く時間がかかると考えられる。

上記の問題を軽減するためには、千年カルテ環境で使おうとしているデータに精通した研究者が、手戻りや LDI とのやり取りができるだけ少なくなるように、予め網羅的なデータ統計やログ出力の設計を行い、一度のやり取りで多く情報を取得できるように進めることが必要である。

5. 開発プロセスの課題への対策

開発プロセスの課題への対策として、以下の 2 点が考えられる。

1, 千年カルテ環境と同様の検証環境を用意する。

本研究では、予め宮崎大学環境にて機械学習モデルを開発した上で、千年カルテ環境での開発に臨んだ。このプロセスがなければ、本研究を進めることは難しかったと考えられる。もし千年カルテにて同様の検証環境が用意されれば、まずそこで研究の方針を立て設計することができ、本番環境で手戻りが少なくなると考えられる。

2, データ集計や機械学習モデルを開発するテンプレートを提供する。

千年カルテの利用者は、LDI に集計したい項目や、開発したい機械学習モデルの種類などを提示すれば、結果を得られるという状態が理想的である。網羅的なテンプレートを提供するには、相当の時間がかかると想像されるため、例えば過去の研究で行われたプロセスなどをテンプレート化していくなど、少しずつ進める形が現実的ではないかと考えられる。

6. 多施設の匿名加工医療情報を用いる効果

2020 年度の研究と本研究の評価指標の比較から、多施設のデータを利用することにより、単一の医療機関のデータを用いた機械学習モデルよりも予測精度を向上できる可能性が示唆された。本研究の結果は特徴量の選定や欠損値の補完などがきちんとできていない上で得られたもので、学習前のプロセスをより精密に行うことで、更に精度を高めていく事ができると考えられる。

多施設のデータを用いてデータの母数を増やす事により、これまで難しかった大規模な予測モデルの開発や、多施設間での臨床データの比較、希少疾患の分析への活用などの用途が期待される。

D. 健康危険情報

総括研究年度終了報告書にまとめて記入する。

E. 研究発表

該当なし

F. 知的財産権の出願・登録状況

該当なし

参考文献

1) 千年カルテプロジェクト概要. 日本医療ネットワーク協会. [https://www.gehr.jp/about/index.html (cited 2023-Apr-23)]
2) 認定匿名加工医療情報作成事業者について. 次世代医療基盤法. 一般社団法人ライフデータイニシアティブ [https://www.ldi.or.jp/law (cited 2023-Apr-23)]
3) 松田敦義. et al. 複数の入院後合併症に対する時系列予測モデルの開発と説明可能な AI を用いたリスク要因の比較. 医療情報学会, 2020. [https://jglobal.jst.go.jp/detail?JGLOBAL_ID=202002212479945038 (cited 2023-Apr-23)]
4) 利用目的等審査委員会の公表. 一般社団法人ライフデータイニシアティブ [https://www.ldi.or.jp/committee (cited 2023-Apr-23)]
5) Breiman L. Random Forests. Machine Learning 2001；45(1)：5–32.