ポスター発表

[PB] ポスター B

2020年6月5日(金) 09:00 ～ 18:30 ポスター会場(1) (e-poster)

[PB-15] 電子レセプト情報に基づく患者の死亡推定方式の検討 Predicting Patients' Mortality from Medical Insurance Claim Database

*佐藤 淳平[1]、合田 和生[1]、喜連川 優[1]、満武 巨裕[2] （1. 東京大学　生産技術研究所、2. 医療経済研究機構）

*Jumpei Sato[1], Kazuo Goda[1], Masaru Kitsuregawa[1], Naohiro Mitsutake[2] （1. Institute of Industrial Science, The University of Tokyo, 2. Institute for Health Economics and Policy）

電子レセプト情報に基づく患者の死亡推定方式の検討

佐藤 淳平[*1], 合田 和生[*1], 喜連川 優[*1], 満武 巨裕[*2]
[*1] 東京大学 生産技術研究所, [*2] 医療経済研究機構

Predicting Patients' Mortality from Medical Insurance Claim Database

Jumpei Sato[*1], Kazuo Goda[*1], Masaru Kitsuregawa[*1], Naohiro Mitsutake[*2]
[*1] Institute of Industrial Science, The University of Tokyo.
[*2] Institute for Health Economics and Policy

抄録：少子高齢化に伴う労働人口の減少が見込まれる我が国に於いて，医療サービスの質を維持し続けるためには，蓄積されたデータに基づく政策立案・実行が有力なアプローチの一つである．著者らは，厚生労働省や地方自治体等より電子レセプト情報の提供を受け，当該情報を使用した分析を医療・政策立案分野の研究者らと共に行ってきた．しかしながら，当該情報はアウトカム指標である死亡の正確な情報を有しておらず，死亡を考慮する必要がある分析の障壁になっていた．死亡前の患者とその他の患者では疾病や提供される診療行為の傾向が異なる可能性が高く，その傾向の差を特徴量として学習することにより患者の死亡を高い精度で推定可能になるものと期待される．本論文では，電子レセプト情報から患者の死亡を推定する方式を明らかにし，地域自治体の電子レセプト情報と加入者台帳情報を組合せた死亡の推定精度の検証実験を示し，その有効性を定量的に示す．

キーワード 電子レセプト情報，死亡，推定

1. はじめに

少子高齢化に伴う労働人口の減少が見込まれる我が国に於いて，医療サービスの質を維持し続けることは社会課題の一つであり，蓄積されたデータに基づく政策立案・実行は有力なアプローチといえる．著者らはこれまでに，厚生労働省や地方自治体等より電子レセプト情報の提携を受け，医療・政策立案分野の研究者らと共に当該情報の分析[1]や，分析を容易化するフレームワークの開発を行ってきた[2]．電子レセプト情報は診療報酬請求のためのデータであるため，死亡や寛解等により医療の提供がなくなった際に当該情報は発行されなくなる（以下，レセプトの中断）．しかし，当該情報はレセプトの中断の理由を正確に把握可能な情報を有しておらず，例えば，死亡をアウトカム指標とした分析の障壁になっていた．レセプトの中断が生じた患者のうち，死亡患者と他の患者では中断の以前の疾病や提供された診療行為，医療費等の傾向が異なる可能性が高く[3]，これらの傾向を特徴量として学習することにより死亡を高精度で推定可能とし，推定結果に基づく死亡をアウトカム指標とした分析を可能とすることが期待される．本論文では，電子レセプト情報から患者の死亡を推定する方式を明らかにし，地域自治体が保有する電子レセプト情報と加入者台帳情報を組合せた死亡の推定精度の検証実験を示し，その有効性を定量的に示す．

2. 方法

1) 死亡の推定方式

本論文では，各々の患者の最終レセプト発行月から前半年間の電子レセプト情報から作成した特徴量に基づき患者の死亡を推定する方式を提案する．本方式では，各々の患者の最終レセプト発行月を基準月とし，3 つの期間（基準月およびその前月，基準月の 2 か月前から 3 か月前，基準月の 4 か月前から 6 か月前）に於ける傷病名・診療行為・医薬品の各コードの出現回数と総医療費に加え，性別と年齢を特徴量とした特徴量ベクトルを作成する．死亡日の情報を有する加入者台帳情報に基づき，基準月から 1 か月以内に死亡日を有する患者を死亡，その他の患者を生存とすることで正解ラベルを作成する．当該正解ラベルと特徴量ベクトルを組合せ，機械学習に基づく推定モデルの構築を行うことにより死亡の推定を行う．

2) 評価実験

提案方式の有効性を示すため，三重県下の地域保険者および岐阜県下の地域保険者より提供を受けた電子レセプト情報と加入者台帳情報（以下，三重データセット，岐阜データセット）を用い

た推定モデルの構築と評価を行う(Table.1).

推定モデルの構築では, 3 つの学習モデル: Gradient Boosting (GB), Neural Network (NN), Support Vector Machine (SVM) を使用し, 3-fold Cross Validation によるパラメタ調整および学習を行う. 特徴量ベクトルの次元数は, 傷病名コードのみで約 1,000 次元と膨大な数であるため, 適切な学習が行われない可能性がある. そこで, 線形回帰モデルに基づき, 正解ラベルとの相関係数が高い上位 150 個の特徴量を選択し, 学習に使用する. 評価指標には, ROC 曲線下面積:AUC 値, 適合率(Precision)と再現率(Recall)の調和平均:F-score を使用する.

Table.1 データセットの詳細

	三重データセット	岐阜データセット
データ期間	2013/03－2017/11	2014/04－2017/11
レセプト数	7,153 万件	7,435 万件
レコード数	10.68 億件	10.84 億件

3. 結果

三重データセットより 182,173 名(うち死亡 43,956 名), 岐阜データセットより 364,346 名(うち死亡 43,344 名)の特徴量ベクトルと正解ラベルを作成した. まず, 単一県下のデータセットを用いた検証として, 三重・岐阜それぞれのデータセットについて, 無作為に抽出した半数の個人を学習に使用した推定モデルの構築, 残りの半数の個人を用いた評価を行った. 全ての学習モデルに於いて AUC 値 0.98 以上, F-score 0.85 以上と高い値を示した(Table.2 左). 次に, 異なる県下のデータセットを用いた検証として, 単一県下のデータセット内での汎化性の検証に於いて学習した学習済み推定モデルと, 学習に使用した県とは異なる県のデータセットを用いた評価を行った. その結果, 岐阜データセットより学習を行った推定モデルがより高い精度を示し, 学習モデル:SVM, NN に於いて, AUC 値 0.98 以上, F-score 0.88 以上と高い値を示した(Table.2 右).

4. 考察

線形回帰モデルより抽出された特徴量には, 死亡診断加算の有無や緩和ケア病棟入院料の有無等が含まれていた. これらの特徴量はリークと呼ばれる正解ラベルの漏洩の一種である. 電子レセプト情報はその特性上, リーク状態の特徴量が発生し得るため, 比較的簡素な学習モデルでも高い精度で死亡を推定できたと考えられる. 今後, 全国規模の電子レセプト情報を使用した, 更なる性能検証を進めることが課題である.

5. 結語

電子レセプト情報に基づく患者の死亡の推定方式を提案し, 地域自治体の電子レセプト情報と加入者台帳情報を組合せた推定精度の検証実験を行った. その結果, 患者の死亡を AUC 値 0.98 以上, F-score 0.88 以上の精度で推定可能であることを示し, 当該方式の有効性を定量的に示した. 今後, 全国規模の電子レセプト情報を使用した性能検証を進める.

参 考 文 献

[1] Sato J, Yamada H, Goda K, et al: Enabling Patient Traceability Using Anonymized Personal Identifiers in Japanese Universal Health Insurance Claims Database. AMIA Jt Summits Transl Sci Proc, 345-52, 2019.

[2] Sato J, Goda K, Kitsuregawa M, et al: Novel Analytics Framework for Universal Healthcare Insurance Claims Database. Stud Health Technol Inform(264), 1578-79, 2019.

[3] 阿波谷 敏:死亡前一年間の医療および介護費用の検討, 季刊社会保障研究 40 巻(第 3 号) 236-43, 2004.

Table.2 評価結果

	単一県下のデータセットを用いた検証						異なる県下のデータセットを用いた検証					
学習	三重データセット			岐阜データセット			三重データセット			岐阜データセット		
評価	三重データセット			岐阜データセット			岐阜データセット			三重データセット		
モデル	GB	NN	SVM	GB	NN	SVM	GB	NN	SVM	GB	NN	SVM
AUC 値	0.992	0.988	0.987	0.994	0.991	0.988	0.965	0.942	0.982	0.971	0.983	0.984
F-score	0.918	0.907	0.901	0.887	0.876	0.858	0.684	0.563	0.855	0.842	0.891	0.884