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　５ 年　 ６ 月　 ８ 日現在

機関番号：１６４０１
研究種目：基盤研究(C)（一般）
研究期間：2020～2022
課題番号：２０Ｋ１０３４８

研究課題名（和文）電子カルテに基づいた慢性疾患重症化時期の新しい予測手法

研究課題名（英文）A new method for predicting the timing of chronic disease severity based on electronic medical records

研究代表者
畠山　豊（Hatakeyama, Yutaka）

高知大学・教育研究部医療学系連携医学部門・教授

研究者番号：００３７６９５６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：検査値の時系列データに対する予測を行う際に、問診項目などの他の構造化データや経過記録などのテキスト情報から患者状態を推定し、その推定結果を組み合わせて検査値予測を行うアルゴリズムの開発を行った。予測結果はこれらの情報を組み合わせたことにより精度向上が実現したことを確認できた。これらのアルゴリズムは患者状態を非構造化データからでも推定できること、及び患者情報の統合に基づきより詳細な患者状態が把握可能であることを示した。今後病院情報システムなどにおいて様々な種類のデータを取得ができるようになるため、構築アルゴリズムは有用であると考える。

研究成果の学術的意義や社会的意義
手法の新規性としては、経過記録などの非構造化データから患者状態の定量的な指標に変換して検査データなどの構造化データと統合して処理を行った点が挙げられる。電子カルテ情報以外のテキスト情報が電子情報として蓄積され始め膨大なデータとなり、これらの非構造化データを解析する需要が増大すると考えられる。そのため、膨大な医療データに対して統合処理を行う本手法は今後さらに必要とされる手法である。

研究成果の概要（英文）：The prediction algorithms for time series laboratory data were developed based on medical text data such as progress records and medical interview items, which were combed with laboratory test values and these data. The prediction results confirmed that the combination of these information improved the accuracy. These algorithms showed that patient status can be estimated even from unstructured data, and that more detailed patient status can be obtained based on the integration of patient information. These results suggest that the construction algorithms will be useful for hospital information systems, which will be able to acquire various types of data in the future.

研究分野：医療情報学

キーワード：医療データ解析　予測モデル

１．研究開始当初の背景

生活習慣病における重症化予防及び発症予防を行うため、検査値の時系列データに対する予測を行い、発症時期などのイベント予測を行うことは重要である。しかし、イベント直前に検査値が急激に変動する傾向があること[1]が知られており、このような時系列データに対する予測アルゴリズム[2]、[3]は既に提案されている。実際の診療データの特徴である不定期に行われている検査データに対しても、先行研究では十分に対応できることが示されている。

欠損データが多い期間における時系列データに対する予測は、予測される変動パターンが多く、必ずしもイベント予測に有用であるとは限らない。そこで、検査値以外の経過記録などの診療情報から患者の状態変化を推定することで、時系列データに対する予測の精度向上が期待できる。また、同様にアプローチとして、問診情報などの健診時の検体検査以外の情報を用いることで、健診受診時のイベント予測に寄与できることが期待できる。

２．研究の目的

検査値以外の情報を組み合わせて、検査値時系列データに対する予測を行うアルゴリズムの開発を行った。病院情報システムに蓄積されている患者データ及び健診受診者データに対し、それぞれ以下のアルゴリズムの開発を行った。

（１）HbA1c 値と一定期間内の HbA1c 変動量を状態モデルとした状態空間モデルを構築しパーティクルフィルタ[4]による実装を行った。さらに HbA1c 変動量が欠損している場合は深層学習モデルによって観測データを補間し長期間の HbA1 の変動を予測するアルゴリズムの開発を行った。高知大学医学部附属病院の患者データに適用し、長期間の時系列変動に提案手法が追随対応可能であるかどうかを評価した。

（２）本論文では健診受診者の受診時の HbA1c 値と回答した問診項目から HbA1c≥6.5 を超えるまでの日数を推定するアルゴリズムの開発を行った。健診受診は年 1 回の間隔で行われるため時系列変化を扱うことは困難であるため、受診日の HbA1c 値及び問診情報から日数を推定するモデルを構築する。実際に基準値を超えた受診者の日数と HbA1c 値の分布の特徴に基づきモデルパラメータを同定し、実際の受診者に適用し推定した日数の妥当性を評価することで構築した予測モデルを評価した。

３．研究の方法

（１）HbA1c 値と一定期間内の HbA1c 変動量を状態モデルとした状態空間モデルの開発

各時刻における HbA1c 値を予測するモデルを状態空間モデルによって記述する。状態空間モデルは、観測できない内部状態を記述する状態モデルと観測データを記述する観測モデルの 2 つモデルを組みわせた時系列モデルであり、以下の式で定義される。

$$\boldsymbol{x}_t = f_t(\boldsymbol{x}_{t-1}) + \boldsymbol{v}_t, \boldsymbol{y}_t = h_t(\boldsymbol{y}_t) + \boldsymbol{w}_t$$

時刻 t の状態モデル、観測モデルの値を$\boldsymbol{x}_t$, $\boldsymbol{y}_t$と記述し、それぞれのモデル関数をf_t, h_tとする。状態モデルにおけるシステムノイズ、観測モデルにおける観測ノイズを$\boldsymbol{v}_t$, $\boldsymbol{w}_t$として記述する。提案モデルでは、状態モデルが推定する HbA1c 値($HbA1c_{state,t}$)、HbA1c 変動量($\Delta HbA1c_{state,t}$)及び観測モデルにおいて計測された HbA1c 値($HbA1c_{space,t}$)、HbA1c 変動量($\Delta HbA1c_{space,t}$)の 4 変量を記述する。つまり、

$$\boldsymbol{x}_t = \{HbA1c_{state,t}, \Delta HbA1c_{state,t}\}, \boldsymbol{y}_t = \{HbA1c_{space,t}, \Delta HbA1c_{space,t}\}$$

と定義する。

実際の予測アルゴリズムの流れ図 1 にて示す。

状態空間モデルをパーティクルフィルタにて実装をおこなった。状態モデル及び観測モデルの各パーティクルにおいて予測を行い、実測値との尤度に基づきフィルタリング処理を行い各パーティクル分布の修正を行うことで予測結果の修正を行った。これを時刻ごとに繰り返し行うことで時系列データに対する予測処理を行った。

HbA1c 変動量クラス情報を推定する識別モデルを、実際に HbA1c 変動量を取得できる際に記載された経過記録情報に基づき構築した。識別モデルとして深層学習モデルである BERT[5]を利用した。入力データとなる経過記録から BERT の入力データとなるトークンを生成する方法として sentencepiece[6]を利用した。日本語版 Wikipedia の情報に基づき事前学習されたモデルパラメータ[7]を各モデルパラメータの初期値として利用した。

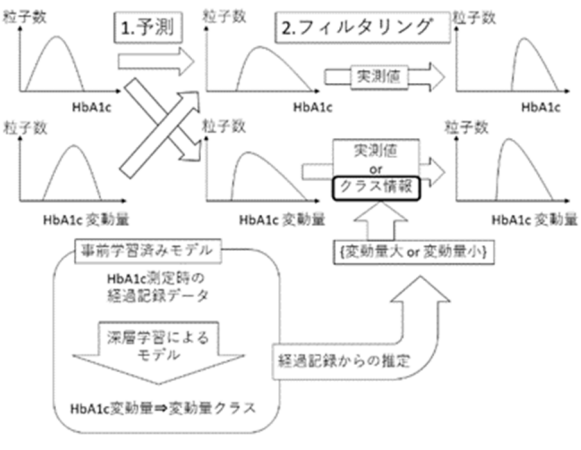

図１　経過記録も利用した予測アルゴリズム

HbA1c 最大値と最小値の差を HbA1c 変動量と定義し、HbA1c 変動量クラスを変動量:大と変動量:小の 2 クラスとし定義した。各モデルパラメータは高知大学医学部附属病院の患者データに基づき学習を行う。2009 年から 2019 年において、120 日間に複数回 HbA1c 検査を実施しており、経過記録の文字数が 300 から 700 である患者データを対象とする。対象期間中の 1940 件のデータを学習データとして利用し、fine tuning によって 2 クラス識別モデルを構築した。別の 277 件のテストデータによって構築した識別モデルの性能を評価した。

（２）健診受診患者に対する発症日予測アルゴリズム
受診日から HbA1c が閾値を超える日までの残り日数を予測対象とするアルゴリズムの開発を行った。（１）と同様に状態空間モデルに基づく予測モデルとして構築を行った。予測対象受診日における健診受診回数を t(≥1)と定義した際の状態空間モデルを以下に定義した。
状態モデル：$Days(t) = Days(t-1) + interval(t) + \omega$
観測モデル：$HbA1c(t) = f(Days(t), \boldsymbol{Questionnaire}(t)) + \omega$
閾値を超えるまでの残り日数を$Days(t)$、前回受診からの経過日数を$interval(t)$とする。対象受診日の年齢、性別、BMI、及びダミー変数化した各問診回答を$\boldsymbol{Questionnaire}(t)$と記述し、残り日数$Days(t)$及び受診日の状況$\boldsymbol{Questionnaire}(t)$から受診日の$HbA1c(t)$を推定する関数を f と記述する。正規ノイズをωとする。
観測モデルにおける関数 f を先行研究[1]及び（１）の結果に従い指数関数として定義する。

$$HbA1c(t) = exp(a_0 * Days(t) + \sum_{i=1}^{29} a_{i*} Questionnaire_i(t)) + b_0 * Days(t) + \sum_{i=1}^{29} b_{i*} Questionnaire_i(t) + c$$

各モデルパラメータを糖尿病群の学習データを用いて最小二乗法によって推定を行った。さらに、AIC による変数選択を行い最適な問診回答項目を決定する。構築した状態空間モデルの状態推定は粒子数 5000 による粒子型フィルタによって行い、粒子数の比率を残り日数となる確率として評価を行った。

４．研究成果
（１）HbA1c 値と一定期間内の HbA1c 変動量を状態モデルとした状態空間モデルの開発
2 クラス識別モデルにおいて 277 件のテストデータに対する識別精度が 0.74 となった。
192 人の HbA1c 測定患者に対し提案予測モデルを適用した。1 患者における適用結果を図 2 に示す。HbA1c 値のフィルタリング後のモデル値（Model HbA1c）と実測値(Observed HbA1c)の時系列変化を図 2 に示す。各時刻の値は各粒子の重心値を示している。直線が引かれていない期間は未測定期間であることを示し、フィルタリングが未実施であることを示す。

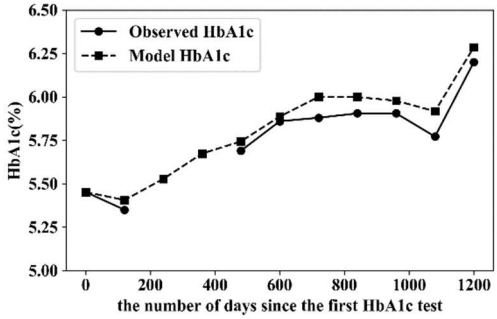

図２　1 患者データにおける HbA1c 提案モデルフィルタリング値
(実測値:実線、フィルタリング値:点線)

最終期間において、HbA1c 値モデル予測分布及びフィルタリング分布の各重心値と実測値との RMSE はそれぞれ 0.30 と 0.14 となった。モデル予測分布の重心値と実測値の相関係数は 0.87 となった。最終期間において HbA1c 変動量を測定できる患者数は 124 であり、HbA1c 変動量予測分布とびフィルタリング分布の各重心値と最終期間における実測値との RMSE は、それぞれ 0.25 と 0.15 となった。
対象データとなる病院情報システムに蓄積されている時系列データは欠損値がほとんどであるため、様々なデータ種類を組み合わせて予測精度を向上する手法が重要になると考える。これらのことから、本構築アルゴリズムにおける経過記録のテキストデータから数値データを推定し、予測モデルの観測データとして利用する手法は病院情報システムデータの解析に有効であると考える。

（２）健診受診患者に対する発症日予測アルゴリズム
観測モデルにおける関数 f においてＡＩＣによる変数選択を行った結果、脂質異常の有無、脳卒中の有無、貧血の有無、１０キロ増加の有無、運動１年の有無、身体活動の有無、体重増減の有無、食べる速度、就寝前夜食の有無、朝食抜きの有無、酒量（4 段階）の項目が選択された。学習データに対する関数 f によって算出された推定 HbA1c 値と実際の HbA1c 値の分布を図 3 に示す。実際の測定値と推定値の相関係数は 0.63 となった。
糖尿病群のテストデータに適用した予測結果の例を図 4 に示す。図 4 は各時刻での粒子数のヒストグラムであり、その比率が残り日数の確率として評価する。つまり、推定残り日数が 3 年以内の粒子数の比率が 3 年以内に閾値を超える確率として評価する。3 年以内での閾値を超えるかどうかについてのＲＯＣを行った。AUC は 0.926 であり 95%信頼区間は(0.937-0.914)となった。

この時のカットオフ値は 0.368 であり、3 年以内の確率が 0.8 の時、陽性的中率が 83.3%陰性的中率 98.1%であることが得られた。

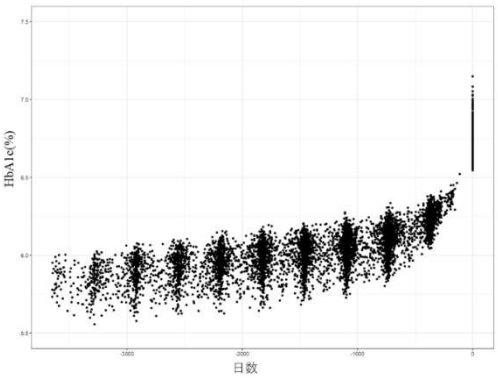

図 3　観測モデルによる学習データ適用結果

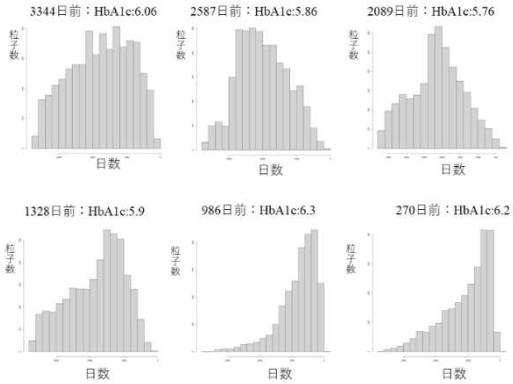

図 4 1 受診者に適用した際の粒子分布推移
HbA1c≥6.5 を初めて満たした受診日を 0 日とし、
それ以前の受診データに対して構築予測モデル
を適用した際の粒子分布

HbA1c 値が閾値をこえる健診受診者の各残り日数における HbA1c 値は閾値直前で急激な上昇を示すなどの一定の傾向を示しており、その傾向を考慮した予測モデルを適用することで適切に受診者の将来発症リスクを提示可能な結果が示されたと考える。

引用文献

[1] Tabak AG, Jokela M, Akbaraly TN, Brunner EJ, Kivimaki M, Witte DR. Trajectories of glycaemia, insulin sensitivity, and insulin secretion before diagnosis of type 2 diabetes: an analysis from the Whitehall II study. Lancet 373: 2215-2221. 2009.
[2] Chen-Ling Huang, Usman Iqbal, Phung-Anh Nguyen, et al.: Using hemoglobin A1C as a predicting model for time interval from pre-diabetes progressing to diabetes. PLoS One. 5;9(8), 2014
[3] Hatakeyama Y, Kataoka H, Nakajima N, et al: Prediction model for glucose metabolism based on lipid metabolism. Methods Inf Med. 53(5):357-63. 2014.
[4] Kitagawa G. Monte Carlo filter and smoother for non-Gaussian nonlinear state space models. J Comp Graph Stat 5: 1-25. 1996.
[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, et al: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. CoRR. abs-1810-04805, 2018
[6] Taku Kudo, John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. CoRR. abs-1808-06226,2018
[7] BERT Pretrained model Trained On Japanese Wikipedia Articles https://github.com/yoheikikuta/bert-japanese (2020/02/01)

5．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　0件）

1．著者名 畠山　豊, 兵頭　勇己,奥原　義保	4．巻 40
2．論文標題 経過記録情報を用いた深層学習による欠損値補間した HbAlc 予測モデルの構築	5．発行年 2021年
3．雑誌名 医療情報学	6．最初と最後の頁 231-237
掲載論文のDOI（デジタルオブジェクト識別子） 10.14948/jami.40.231	査読の有無 有
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計1件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 畠山豊
2．発表標題 問診回答を利用した健診受診者に対する糖尿病発症日数予測モデル
3．学会等名 第41回医療情報学連合大会
4．発表年 2021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	奥原　義保 (Okuhara Yoshiyasu) (40233473)	高知大学・医学部・特任教授 (16401)	
研究分担者	兵頭　勇己 (Hyohdoh Yuki) (50821964)	高知大学・教育研究部医療学系連携医学部門・助教 (16401)	

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