

令和 4 年 6 月 16 日現在

機関番号：16401

研究種目：若手研究

研究期間：2019～2021

課題番号：19K19389

研究課題名（和文）バイアスを除いた検査値の自動抽出手法の確立と臨床疫学研究への応用

研究課題名（英文）Establishment of a method to automatically collect laboratory test values without bias and application to clinical epidemiological research

研究代表者

兵頭 勇己（Hyohdoh, Yuki）

高知大学・教育研究部医療学系連携医学部門・助教

研究者番号：50821964

交付決定額（研究期間全体）：（直接経費） 2,400,000円

研究成果の概要（和文）：我々は、電子カルテから臨床的に妥当な値の検査値を抽出し、その分布と基準範囲を推定する新しい手法を開発した。本手法は、次の臨床検査オーダーまでの間隔を医師の臨床判断を含む情報として扱う。次回の検査オーダーまでの検査間隔を重みとしてデータをリサンプリングし、臨床的に許容される検査値の集合を特定した。また、推定されたヘモグロビンの基準範囲の妥当性を検証するために、鉄剤処方と臨床検査値の関係との比較を行った。その結果、本手法は複数の臨床検査値の年齢別基準区間を推定できた。本手法は、医師の臨床的判断を含めた有用な分布と基準範囲を簡便かつ自動的に推定する方法である。

研究成果の学術的意義や社会的意義

診療データは単純に解析すると、誤った結論を導くことが多い。その理由の一つは、診療データには欠損値が多く存在し、そのまま解析に利用すると偏り（バイアス）を伴った結果となる可能性が高くなるためである。さらに医療現場においては「健康な人ほど検査が不要」など、患者の状態に応じてデータの発生確率が変動するためデータが欠損するプロセスが複雑である。本研究では、欠損値への対処方法の一つとして、これらの臨床医の判断を含めて検査値が取りうる値の分布を推定するための手法を確立した。これらの手法や分布を活用することにより、根拠のある欠損値の代入が可能となり、より蓄積された医療データの利活用がなされると期待できる。

研究成果の概要（英文）：We developed a new approach for extracting the laboratory test values that are clinically acceptable and estimating the distributions and reference intervals from the electronic health record. Our approach interprets the intervals until the next laboratory test order as information that includes the physician's clinical decision. We resampled the data using the test interval until the next laboratory test order as a weight, and a set of clinically acceptable laboratory test values was identified. In addition, in order to verify the validity of the estimated reference intervals in hemoglobin, we compared it with the relationship between iron prescription and laboratory test values. Results show that the approach can estimate age-specific reference intervals of multiple clinical laboratory tests. This approach is a simple method of estimating useful distributions and reference intervals including the clinical decision of physicians.

研究分野：医療情報学

キーワード：臨床検査値 検査オーダー 教師無し学習 フェノタイピング

1. 研究開始当初の背景

電子カルテを代表とする診療データの電子化に伴い、医療機関には日々データが蓄積されている。医療ならびに医薬品の諸問題を解決するために、蓄積された大量のデータに対して疫学研究手法の適用やデータマイニング、さらには人工知能(AI)の活用などが期待されている。しかしながら、診療データを単純に解析すると、誤った結論を導くことが少なくない。その理由の一つとして、診療データには欠損値が多く存在し、そのまま解析に利用すると偏り(バイアス)を伴った結果となる可能性が高くなるためである。

最近では、データの欠損値の対処として、多重代入法と呼ばれる統計学的手法による欠損値の補完を行うことがある。しかし、医学研究で多用される検査データなど、医療行為に関するデータの欠損が生じるメカニズムは単純ではない。例えば検査データの場合、臨床医は健康な人ほど検査が不要と判断することが多く、反対に重症な人ほど頻回に検査を実施するという医療プロセス上の特性が存在する。これらのことから、臨床検査データの欠損値の対処としては、欠損がランダムに発生するという単純な仮定での多重代入法のみでは不十分であり、対象者の特性ごとに検査値の取りうる値の頻度分布をあらかじめ推定し、その分布を利用した多重代入法を行う必要がある。

これらのことから、蓄積された診療データに対する医療データマイニングや統計解析における欠損値の対処のために、対象者の特性や状況ごとに、検査値の取りうる値の頻度分布を推定する手法の構築が必要である。

2. 研究の目的

本研究は、診療データのうち医学研究で多用される臨床検査データに限定し、検査結果の取りうる値の頻度分布を推定する手法を構築することを目的とした。臨床検査のデータは「検査値」と検査オーダの「間隔」という2つの独立した情報を有する。近年、これらを利用して臨床医の診療行動に関する情報(推論)を、電子カルテから抽出する試みが存在する。このことは、「検査値」と「検査間隔」の2つの情報のみで、臨床医が「異常」と判断していないとされる検査値を、電子カルテ上から自動的に収集できる可能性を示している。患者特性に応じた臨床検査値の頻度分布を高い精度で推定するために、この医療プロセスの特性に基づいた「検査値」と「検査間隔」の2つの情報を活用する。

3. 研究の方法

まずは、解析対象を外来診療時に血液検査を実施された患者とし、主に頻度分布推定手法の構築を実施した。具体的な手法は、以下の通りである。

【使用データ】

本研究で使用されたデータは、血球数算定や生化学検査項目など40項目の血液検査結果および検査時の年齢や性別、検査日(当該患者誕生日からの相対日付)が含まれる。これらのデータは、高知大学医学部附属病院(以下、当院)の匿名化医療データウェアハウスである、Retrieval sYstem for Open Medical Analysis (RYOMA2) から抽出された。RYOMA2には、当院開院以来36年間の外来および入院患者約31万人のデータが含まれており、小児および成人患者の臨床検査件数約1億6千万件が含まれる。本研究の対象は、1981年から2016年までに当院の外来診療にて血液検査を実施した20歳以上の患者とした。

【推定手法】

本手法は、「臨床医は患者から得られる情報を総合的に解釈および判断し、血液検査をオーダする」という前提に基づいている。すなわち、異常を呈するなど、臨床的に何らかのモニタリングが必要であれば、より頻回に血液検査をオーダすることにより同一検査項目間における検査の間隔が短くなる。反対に、モニタリングが必要でなければ、血液検査の間隔が長くなる。そこで、より「異常」とされる検査値の集合を「検査間隔が短い検査結果値の集合」と仮定した上で、より「問題ない」とされる検査値の集合と仮定した「検査間隔が長い検査結果値の集合」から差し引く。このことにより、病院情報システムに蓄積された検査値の分布から、臨床医が異常と捉えた検査値を取り除くことを実現する。

【データ整形】

対象データは年齢区分ごとに層別化された。すなわち、検査時点の年齢を基準に、20歳以上40歳未満、40歳以上60歳未満、60歳以上80歳未満、全体を代表する区分として20歳以上60歳未満の、合計4つのカテゴリーに分割した。

本手法では、同一検査項目の次回検査日までの日数を「検査間隔」と定義し、検査結果レコードに追加した。なお、同日に複数回検査を行っている場合は、検査間隔を0とした。さらに、1回しか検査を実施していない場合や、検査間隔が365を超える場合は一律検査間隔を365とした。

【再標本化】

以下に示す方法により、対象のデータから、「検査間隔が短い検査結果値の集合」および「検査間隔が長い検査結果値の集合」の2つのデータセットを新たに作成した。

1. パラメータ τ を検査間隔の閾値とし、データを2群に分割する。
2. 標本数を分割前データ件数とした検査間隔を重みとした再標本化を行う。標本数は分割前データ件数に統一する。
 - i. 検査間隔 $\leq \tau$ のデータセットは検査間隔の値の逆数を重みとする。
 - ii. 検査間隔 $> \tau$ のデータセットは検査間隔の値を重みとする。

本研究では、パラメータ τ を30に設定した。

【分布の推定】

以下に示す方法により、2つのデータセットの差分から「臨床医が問題ないと判断した」検査結果の分布を推定する。

1. 各検査項目の最小単位（例えばアルブミンなら0.1 g/dL）を階級幅とした度数分布を算出する。
2. 検査間隔 $> \tau$ のデータセットの度数から検査間隔 $\leq \tau$ のデータセットの度数の差を算出する。算出された度数が負値であれば、0とする。
3. 算出された度数で構成される検査結果値の分布を、「臨床医が問題ないと判断した」検査結果の分布と定義する。

以上の方法により、検査値が取りうる分布を推定した。また、このように推定した分布から確率密度関数を推定し、2.5%点および97.5%点をそれぞれ基準範囲の下限および上限とした。

1. Box-Cox変換にて3.2.3にて推定した分布を正規分布へ変換する。
2. 変換された分布に対し、正規分布のパラメータである平均値および標準偏差を最尤推定法にて推定する。
3. 推定されたパラメータで表現される正規分布の2.5%点および97.5%点を算出する。
4. 算出した2.5%点および97.5%点に対し、Box-Cox逆変換を適用し、元の検査結果値のスケールに戻す。この値をそれぞれ基準範囲の下限および上限と定義する。

また、貧血や糖尿病など一部の病態は、単一の臨床検査項目にて診断や治療方針の決定などの臨床的判断が可能であることから、この特性を利用してさらなる有効性の評価を行った。例えば、ヘモグロビン値が低いほど貧血が重度であり、経口鉄剤を処方する可能性が高まると仮定すると、真の臨床的判断は処方オーダに反映すると定義することが可能であり、貧血における真の臨床的判断を処方オーダから特定することが可能である。

ヘモグロビン (Hb) は一般的な血液検査項目であり、基準範囲に男女差が存在する。さらに、Hb値にて病態の存在が推測できる鉄欠乏性貧血は、日常的に高頻度で遭遇する疾患であり、その初期治療は鉄剤の処方により行われることが多い。そこで、注射用鉄剤および経口鉄剤の処方履歴がある患者は、「臨床医が真に治療の必要性があると考えた鉄欠乏性貧血」が存在すると仮定した。この仮定のもと、Hb検査後3日以内の処方有無および年齢区分ごとに層別化しHb値の中央値を算出、推定した基準範囲と経年傾向を比較することで、推定結果の妥当性を示した。

さらに、本手法の前提条件である、「臨床医は健康な人ほど検査が不要と判断することが多く、反対に重症な人ほど頻回に検査を実施する」という仮定を検証するために、独立したデータセットによる検証を行った。本来、欠損した検査値の本来得られとされる値は不明である。しかし、一部の検査項目においては、蛋白電気泳動波形から検査値を精度よく推定することが可能である。この特性を利用し、欠損した検査値が本来どのような値をとるかを、蛋白電気泳動波形から検査値を予測する予測モデルを構築して検証を行った。使用した蛋白電気泳動波形情報はアルブミン分画から1分画の間のデータを含む、プレアルブミン分画側（陽極側）からグロブリン分画側（陰極側）までの計300区画の易動度を使用した。波形情報の補正および正規化は、アルブミンのピーク位置を易動度75となるよう調整し、さらに、総蛋白値が7.0g/dLの時、波形の曲線下面積が100000となるように波形を補正した。予測する血液検査項目の検査結果値は、年齢および性別に応じて病院内で設定されている基準範囲を閾値とし、「基準範囲内」「範囲外」の二値データに変換した。その上で、予測モデル構築用のデータセットおよび予測モデル検証用のデータセットとデータセットをバリデーションし、機械学習モデルとしてロジスティック回帰モデル、L1正則化付きロジスティック回帰モデル、サポートベクターマシン、一次元畳み込みニューラルネットワーク、ランダムフォレスト、勾配ブースティング決定木の6種類を用いて最良の予測モデルを構築した。

4. 研究成果

上記の手法にて、もともと設定されている基準範囲と比較しても妥当な検査値の分布を得ることができた。例えば、クレアチンホスホキナーゼ (CPK)、クレアチニン (CRE)、Hb など、性別により異なる基準範囲が設定されている項目に関しては、本手法で推定した分布や基準範囲においても性別の違いを反映した結果を示した。さらに特徴的な結果として、アルブミン (Alb)、Hb においては、年齢増加に伴い、推定された検査値の分布および基準範囲の低下が認められた。一方、クレアチニン (CRE)、白血球数 (WBC) においては、年齢区分による大きな変動は認められず、通常参照されている基準範囲と異なる振る舞いをする検査項目が認められた。

鉄剤処方と推定された Hb 値推定分布の関係では、男性は処方の有無に関わらず、年齢増加に伴い Hb 値の分布の中央値の低下が認められた。一方、女性は処方の有無に関わらず、年齢増加に伴う Hb 値の分布の中央値の大きな変動は認められなかった。この傾向は、本手法で推定した Hb の分布の年齢変動に伴う傾向と一致した。

また、蛋白電気泳動波形情報から構築した検査値予測モデルにおいて、使用した 6 つの学習モデルの中で、勾配ブースティング決定木モデルは各検査値において AUC 値の最高値を示すことが多かった。このモデルを使用し、欠損値がどのような値であったかを予測すると、もともとの値の取りうる範囲がランダムであると仮定した場合よりも、もともとの値が基準範囲内であると仮定して発生した欠損と仮定する方が、尤もらしいデータセットであるという結果となった。これは、もともとの前提である、「臨床医は健康な人ほど検査が不要と判断することが多く、反対に重症な人ほど頻回に検査を実施する」が妥当である根拠の 1 つとなり得る。

以上のことから、本研究にて提案した検査値の分布の推定手法により、多くの血液検査項目において、臨床医の判断の情報を内包した有用な分布および基準範囲を推定することができた。特に年齢および性別にて層別化し推定した結果は、それぞれのコホートに対して臨床医が「問題ない」と考える基準範囲が異なる検査項目の存在が示唆された。これらの分布を活用することにより、根拠のある欠損値の代入が可能となり、より蓄積された医療データの利活用がなされると期待できる。

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 1件）

1. 著者名 Hyohdoh Yuki, Hatakeyama Yutaka, Okuhara Yoshiyasu	4. 巻 23
2. 論文標題 A simple method to identify real-world clinical decision intervals of laboratory tests from clinical data	5. 発行年 2021年
3. 雑誌名 Informatics in Medicine Unlocked	6. 最初と最後の頁 100512 ~ 100512
掲載論文のDOI（デジタルオブジェクト識別子） 10.1016/j.imu.2021.100512	査読の有無 有
オープンアクセス オープンアクセスとしている（また、その予定である）	国際共著 -

〔学会発表〕 計4件（うち招待講演 0件 / うち国際学会 0件）

1. 発表者名 兵頭 勇己
2. 発表標題 臨床検査の実施間隔情報を利用した新たな欠損値代入方法の提案
3. 学会等名 第45回中国四国医療情報学研究会
4. 発表年 2020年

1. 発表者名 兵頭 勇己
2. 発表標題 病院の特徴を考慮した検査基準範囲の推定手法
3. 学会等名 第39回医療情報学連合大会
4. 発表年 2019年

1. 発表者名 奥村 健馬
2. 発表標題 検査変動幅は医師の判断にどう影響するか -リアルワールドデータからの集団知の抽出-
3. 学会等名 第39回医療情報学連合大会
4. 発表年 2019年

1. 発表者名 兵頭 勇己
2. 発表標題 蛋白分画検査の波形情報を使用した血液検査の欠損値補間の試み
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
---------------------------	-----------------------	----

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
---------	---------