リアルワールドデータを用いた予測モデルの開発と評価

瀬戸　ひろえ

リアルワールドデータとは，医療現場の日常診療で発生する患者情報や，社会生活者が日常生活を送る中で生成される健康関連のデジタルデータの総称である．患者情報の電子化と利活用が促進されたことや，スマートフォンアプリやウェアラブルデバイスが普及したことから，リアルワールドデータを活用して，日々の意思決定や行動変容のサポートなど医療分野にとどまらない多様なソリューションが社会に実装されること期待されている．リアルワールドデータを用いてリアルワールドエビデンスを生成することに注目が集まっている一方で，将来の疾病発症のリスクを予測することにも大きな期待が寄せられている．しかし，リアルワールドデータは，変数数やレコード数が膨大であること，研究目的で収集されたデータと異なり不均一で，様々なバイアスを含むことなど，多岐にわたるデータ解析上の課題を有する．そのためリアルワールドデータの解析に対しては，ビッグデータの効率的な利用が期待できる機械学習を用いることや，これまで医療分野では用いられてこなかった解析手法を用いること，データやタスクに合わせ必要な新しい統計手法が開発されることなどが必要となってくる．本論文にはこのような背景の下で，リアルワールドデータを用いた疾病予測モデルの開発と評価における課題の解消に向け行った研究結果を収録する．

第 1 章では，リアルワールドデータを用いた予測モデルの開発と評価の概要を説明する．ここではまず，リアルワールドデータの定義と特徴，予測モデルの開発と評価を行う一般的な手順を紹介する．また，予測モデルの開発に用いられる代表的な学習アルゴリズムや，予測モデルの評価に用いられている代表的な概念と評価手法を紹介する．そして，本論文に収録された研究の位置づけを整理する．

第 2 章では，糖尿病予測モデル開発における機械学習の有用性を検討した研究を取り上げる．治験データなど小規模なデータを用いることが多かった医療分野においては，リアルワールドデータのようなビッグデータを用いた研究がまだ少なく，ビッグデータを用いた場合に機械学習がどれほど良好な予測精度を示すのか明らかにされていなかった．こうした背景に対し，本研究では国民健康保険データを用いて，糖尿病予測モデルを開発する際に機械学習を用いることの有用性を検討した．特に，これまで機械学習を用いた疾病予測モデル開発に関する研究がほとんど評価してこなかったキャリブレーションの評価に焦点をあて，機械学習モデルと古典的モデルの比較を行った．

第 3 章では，確率予測モデルに対する変数ベースのキャリブレーション評価手法を開発した研究を取り上げる．スマートフォンなどを介してモデルが算出する予測値を個人に対し個別に提示する状況などが期待されている現代においては，個別の集団に対して適切な予測値が算出できるのかを評価することの重要性が高まっている．こうした背景の中で本研究では，どの変数のどのような値において予測モデルのキャリブレーションが良好かを判断できる新しい評価手法を開発した．そして，理論的証明と数値実験によって，開発した手法がキャリブレーション評価手法として妥当であることを確認するとともに，その特性や従来手法との比較を示した．さらに，国民健康保険データを用いた実データ解析を実施し，提案手法の実運用上の有用性を示した．

第 4 章では，安定性に対する公平性を評価できる新しい評価手法を開発した研究を取り上げる．リアルワールドデータのように，対象者の特性がコントロールされずに収集されたデータでは，国籍や人種など，差別に対する配慮が必要な属性について，外国籍や少数民族など，差別から守られるべき集団のデータが少数となってしまうことが多い．学習アルゴリズムの多くが少数の集団に対し適合の悪いモデルを開発する可能性が高いことから，少数の集団に対し不公平なモデルが開発される可能性があり，倫理的・

法的に重大な問題が生じうる．こうした背景の中で，予測モデルの公平性を評価する手法が様々に開発されてきたが，安定性に対する公平性を評価できる手法が存在しなかった．そこで，本研究では，安定性に対する公平性を評価可能な新しい評価手法を開発した．さらに，提案手法を用いて，既存の学習アルゴリズムの安定性に対する公平性を評価する数値実験と実データ解析を行い，学習アルゴリズムごとの安定性に対する公平性の特性を明らかにした．

第 5 章では，メタボリックシンドローム（metabolic syndrome; MetS）の罹患率とその診断マーカー（MetS マーカー）の季節変動を調査した研究を取り上げる．リアルワールドデータのように，長期間継続的に収集されるデータには，季節変動が含まれており，これが予測モデルを開発するうえで解析結果にバイアスを生じさせる可能性がある．よって，開発に用いる変数に対する季節変動を正確に把握することが必要となる．特に，MetS の診断マーカーである血圧や血糖などの検査値は，季節によって大きく変動することが示されており，多くの先行研究が MetS の罹患率や MetS マーカーの季節変動を調査してきた．しかし，先行研究が用いている解析手法には，複雑な変動を抽出できない，長期トレンドを考慮できないなどの問題があった．こうした背景に対し本研究では，長期トレンドや複雑な変動を考慮可能な季節性解析手法である，seasonal trend decomposition procedure based on loess を用いて，大阪府の国民健康保険データベースから取得した特定健康診査データの MetS の罹患率と MetS マーカーの季節変動を調査した．

第 6 章では，まず本論文で取り上げた研究を総括し，さらに本論文の限界を整理した．また，本論文で取り上げた研究の拡張性について検討した．さらに，本論文では取り上げられなかったリアルワールドデータを用いた予測モデルの開発・評価における更なる課題を紹介した．（行動統計科学）