

機械学習を活用したリアルワールドデータからの継続的なエビデンス創出

(Continuous evidence generation from clinical Real World Data utilizing machine learning)

市川 太祐 サスマド(株) 臨床開発部長 医師・医学博士

1 はじめに

近年、医療リアルワールドデータ（RWD）からのエビデンス創出に注目が集まっている。RWDは研究プロトコルに則って収集されるデータではなく、日々の診療や個人の健康管理等によって生じるデータである¹⁾。典型的なデータソースとしては電子カルテ、看護記録、レセプト、DPC、お薬手帳、疾病管理手帳が挙げられ、最近ではアクティブトラッカー等のウェアラブルデバイスで取得したセンサーデータも含まれる。諸条件がコントロールされた臨床試験のデータではなく、実臨床における治療状況を反映したデータとしてRWDは注目されており、実際に医療費請求データや電子カルテデータを用いて臨床試験の結果を置き換えられるか検証した報告もある²⁾。筆者が所属するサスマド株式会社でも複数の学術機関、製薬企業と共同でRWDを用いたエビデンス創出に取り組んでおり、その一部は学術成果として発表し

ている。本稿ではRWDを用いた継続的なエビデンス創出について弊社の経験を踏まえて考察したい。

2 RWDを用いたエビデンス創出

図1にはエビデンス創出の典型的なフローを示した。ステップ1は示唆抽出が目的である。まずドメイン知識を有する専門家とクリニカルクエスチョンについて議論し、RWDを用いて検証する。クリニカルクエスチョンの性質によって、検証には基礎集計及び機械学習を用いた予測モデルを使い分けている。「併用薬の現状を知りたい」といったシンプルなものであれば基礎集計で対応でき、「ある治療方針の選択に寄与する影響因子を見つけない」というものであれば機械学習を用いた予測モデルを利用する。以上のステップ1を複数回繰り返すことで報告に値する示唆が得られ、ステップ2として論文文化を行う。ここではそれまでの解析結果を見直す意味

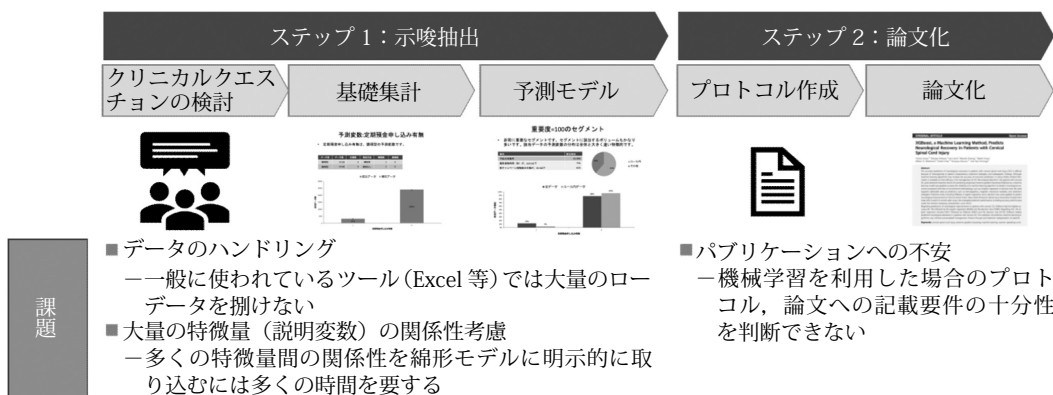


図1 RWDを用いたエビデンス創出フロー

で解析手順をプロトコルの形で清書し、学術論文の形で報告を行う。

実際にこのフローを進めていくにあたってはいくつかの課題がある。まずステップ1においては「データのハンドリング」「大量の特徴量の関係性考慮」という課題が挙げられる。RWDは数十GBのサイズかつ複数のテーブルに分かれた形で提供されることが多く、一般に使われているようなMicrosoft Excel等の表計算ソフトでは基礎集計もままならない。集計についてはリレーショナルデータベースを利用することで対応は可能だが、筆者の経験上多くの企業や学術機関ではこの段階でつまづいているように見える。また、RWDはサンプルサイズだけでなく、各レコードに含まれる検査項目等の特徴量の数も、臨床試験のデータに比べると圧倒的に多い。ここで機械学習を用いることで項目間の関係性を考慮したモデル構築が可能になる。ステップ2の課題としては「パブリケーションへの不安」が挙げられる。これはRWDを用いた研究が始まってからまだ日が浅く、手法の記述や結果の解釈といったテクニカルな部分について十分にノウハウが蓄積されていないことが原因と考えられる。

3 機械学習 PoC の先にある MLops とは

RWDの活用を検討開始した段階ではまだRWDがその企業／学術機関において有効なものか判断がつかない。そのため、先に述べたエビデンス創出フローはPoC（概念実証）として行われることが多い。PoCが成功すると、今度はそれを継続して運用していくための体制づくりが必要となる。先に述べたようにRWDを用いたエ

ビデンス創出については機械学習の利用が必須となるため、体制づくりはRWDを用いた機械学習の継続運用とほぼ同義となる。

機械学習の継続運用については早い段階からビッグデータを扱っていたGoogleやMicrosoftといった情報系企業が先行しており、ベストプラクティスが積み重ねられている。このノウハウについてはMLops (Machine Learning Operations) と名づけられており、その運用について活発に議論されている。図2はGoogleの報告³⁾で示された機械学習システムの構成要素である。各要素の説明については原著を参照してほしいが、ここで重要なのは一般の人が機械学習システムと聞いて思い浮かべるような「ML Code (機械学習のコーディング)」「Analysis Tools (分析ツール)」については全体のごく一部を占めるに過ぎないという点である。PoCの段階では有効性を示すことに注力しているため、システムの継続運用については軽視されやすいためこの点は十分に注意したい。

4 RWDを用いた継続的なエビデンス創出体制を構築するために必要な要素

図2に示したMLopsの構成要素のうち、RWDから継続的にエビデンス創出体制を構築するために必要な要素を図3に示した。なお、MLopsの構成要素には予測モデルを公開した場合に安定稼働させるための要素も含まれている。かつてフラミンガム研究で示されたリスクスコア⁴⁾のように予測モデルを一般に公開し運用する場合はこのような要素も必要となるが、これは稀であるため本稿では割愛する。

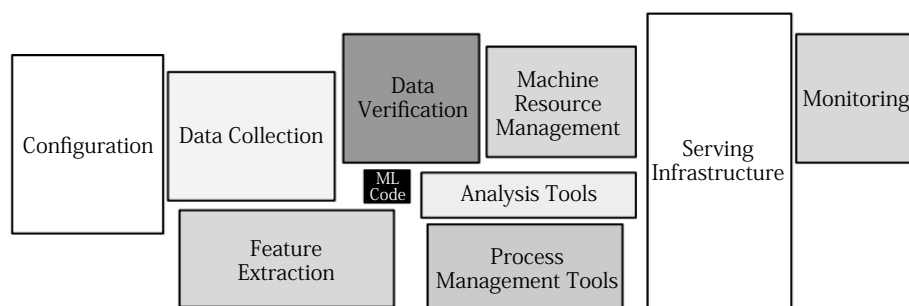


図2 機械学習システムの構成要素と各要素が占める割合 (文献3のFigure 1より引用)

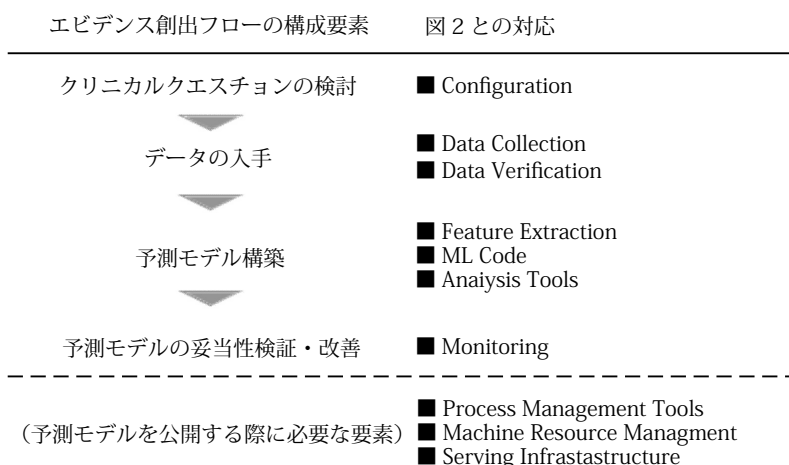


図3 RWD を用いたエビデンス創出フローに求められる MLops の構成要素

4.1 クリニカルクエスチョンの検討

図2の一番左に位置するのが Configuration であり、これは例えば予測モデルにおいて予測対象をどうするかといった一連の設定を指す。RWD を用いたエビデンス創出の場合においてはクリニカルクエスチョンの検討に該当すると言える。他社任せのクリニカルクエスチョンは自社の意思決定につながらないため、これは必ず自社で実施すべきものである。

4.2 データの入手

これは図2におけるデータ収集 (Data Collection)、データ検証 (Data Verification) を含む。RWD を用いたエビデンス創出の場合は多くの場合、外部からデータを入手することになるだろう。国内においてもレセプトデータや電子カルテデータを有償で提供している企業が複数存在する。先にあげたクリニカルクエスチョンの検討を十分に行った上でその検証が可能なデータを入手すべきである。間違っても RWD くらいサンプルサイズの大きいデータであれば何か面白いことが見つかるだろうといった漠然とした期待感のみでデータを入手してはいけない。漠然とした目的で開始した分析は漠然とした結果しか生まない。

4.3 予測モデル構築

ここには特徴量抽出 (Feature Extraction)、機械学習のモデル構築 (ML Code)、分析ツール (Analysis Tool) が含まれる。RWD のローデータから特徴量を抽出し、機械学習の予測モデルに投入し、予測結果を得るという一連の操作で構成される。特徴量とは統計学の用語で言うなら説明変数にほぼ該当する。ただし、医療分野やマーケティング分野で発展してきた古典的統計学における説明変数が人間の立場から見て意味を持つ変数 (例えば血液検査項目) であるのに対して、画像解析やロボティクス分野で発展してきた機械学習で発展してきた特徴量は予測モデルを説明する変数であるという位置づけは変わらないものの必ずしも人間の立場から見て意味を持つ変数とは限らない。その最たるものが近年脚光を浴びた深層学習である。深層学習は特徴量を自動生成して予測モデルを構築する手法と捉えることができるが、そこで得られた特徴量は人間にとって可読性を持たない。

機械学習の予測モデルを構築するための手法については、複数の if-then ルールで構成された決定木、決定木を多数作成し、その結果を合算することで予測結果を高めたランダムフォレスト、データへのオーバーフィッティングを防ぐためにロジスティック回帰に正則化項を加えた L1 ロジスティック回帰等様々なものがある。また先にあげた深層学習も機械学習の一手法である。ここ

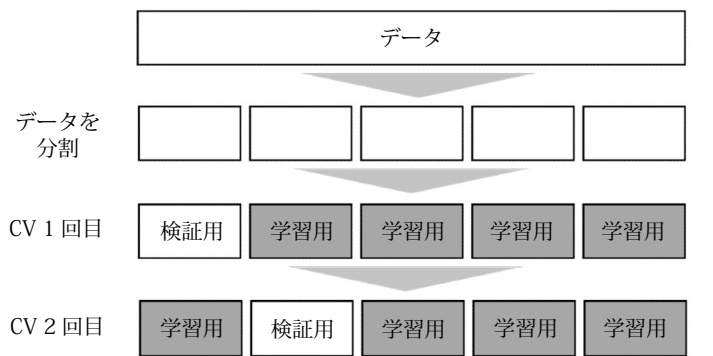
で重要なのはどのようなデータでも最高精度を達成できるような「銀の弾丸」的な手法は存在しないということである。したがって、予測精度の高い予測モデルを構築しようとするなら複数の機械学習手法で結果を比較する必要がある。なお、筆者の経験上、「銀の弾丸」はないものの、比較的高い精度を達成できる手法の候補は絞られているため、その範囲内で結果を比較すれば良いものとする。

上記の機械学習の手法については様々なサービスで実行可能であり、Microsoft Azure Machine Learning や Google AutoML, DataRobot 他多くのツールが市場に出ている。また R や Python 等オープンソースのプログラミング言語を用いて実装することも可能である。ここで MLops の観点からツールの選択の際に重要なポイントは「手法の実行において属人性を排することができる」「メンテナンスが容易である」の2点である。例えば社内にデータサイエンティストを抱えており、追加コストをかけずともデータサイエンティストが機械学習手法を実装し、予測モデルを構築できる場合を考えてみよう。この時データサイエンティストの実装コードは実装者しかその意味がわからない属人的なブラックボックスのコードになっていないだろうか？このデータサイエンティストが辞めても他の人はコードを引き継いで実行できるだろうか。また、少し特徴量を変えて予測モデルを実行したい時に変更・修正は容易に実装できるだろうか？RWDの有効性を確認するPoCであれば属人性やメンテナンスは度外視しても良いかもしれないが、継続的

なエビデンス創出体制を構築していくのであれば無視できない。先に挙げた Microsoft Azure Machine Learning 等のツールは属人性およびメンテナンスについて配慮されており、プログラミングの知識なしに利用することが可能である。十分にデータサイエンティストを雇用する余裕があればオープンソースを利用して独自に機械学習システムを構築するのも良いが、雇用が難しい場合はこのようなツールを利用することを勧めたい。なお、ツールによっては構築した予測モデルの権利がユーザではなくサービス提供者に帰属する場合があるので注意しておきたい。

4.4 予測モデルの妥当性検証・改善

RWDを用いた継続的なエビデンス創出体制を構築するにあたって、最後に必要な要素がモニタリング (Monitoring) である。これはモデルの妥当性検証・改善のプロセスと言い換えられる。先にも述べたが、一般に予測モデルの妥当性を検証する場合は、内的妥当性と外的妥当性という観点から検証する。内的妥当性は学習データの範囲内における予測モデルの妥当性を検証するもので、外的妥当性は学習データ以外のデータで検証を行う。内的妥当性についてはクロスバリデーションという方法で検証することが多い。これは学習データを複数に分割し、その一つを検証用、それ以外を学習用に用いるという操作であり、この操作を分割数の分だけ繰り返す (図4)。予測モデルに利用する学習データが十分に



以降 CV を 5 回目まで実施し 5 回の結果の平均を最終的な結果とする

図4 クロスバリデーション (CV) の実行イメージ

準備できる場合は、学習データの一部を検証用データとして保存しておくとの外的妥当性を検証可能である。しかし、RWDのサンプルサイズが大きいとはいえ、最終的にモデル構築に利用できるデータは少数になることも多く、外的妥当性の検証のために学習に利用しないことももったいないとも言える。したがって、外的妥当性を検証する際は予測モデル構築後にあらためてデータを収集して検証することが多い。

5 サスメド Awesome Intelligence の紹介

筆者が所属するサスメド株式会社では臨床開発における MLops を効率的に実行するシステムとして、機械学習自動分析システム (Awesome Intelligence) を開発した。Awesome Intelligence (AwI) の概要を図5に示す。ユーザは web ブラウザを介して CSV ファイルをアップロードし、表示された列名から予測したい変数 (予測変数) を選択するだけで、複数の機械学習アルゴリズムを用いて予測モデルが構築される。予測モデルの評価結果をまとめた分析レポートは Microsoft PowerPoint の形式でダウンロードでき、分析結果をユーザのレポートにすぐ活用可能である。

分析レポートの内容は各特徴量の分布、予測モデルの評価結果、特徴的なセグメントの抽出結果の3つで構成されている。予測モデルの評価結果は、モデルの内的妥当性の検証結果としてクロスバリデーションを用いた

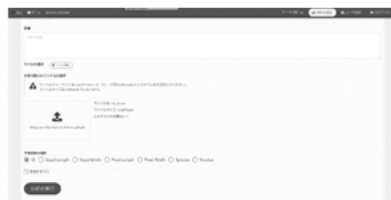
AUROC (Area Under the ROC) 等の評価が表示される。各評価結果には、予測に寄与した特徴量が特徴量重要度という形で優先順位をつけて表示されており、ユーザはどの変数が重要だったのかを確認することができる。特徴的なセグメントの抽出結果については、機械学習の一手法であるスパースモデリングを用いることで、予測変数の分布が全体と比較して大きく異なるセグメントを抽出した結果を表示する。これによって、単一の変数の重要度だけではなく、複数の変数を組み合わせた重要度を確認することができる。

なお構築した予測モデルはユーザに権利が帰属し、ユーザが削除しない限り継続使用できる。予測を実行したい際は、対象データを CSV 形式でアップロードすることで予測結果をダウンロードできるようになる。また予測モデルは Python コードの形でダウンロードも可能であり、機械学習システムとしての継続運用に活用できる。

6 Awesome Intelligence を用いたエビデンス創出の事例

最後に AwI を用いたエビデンス創出の事例を紹介したい。図6は RWD としてレセプトデータおよび診療データを用いた事例である。まず、クリニカルクエスチョンとしてはある処方薬について臨床医がどのような因子に影響を受けて、低用量と通常量のどちらの用量に決定するか検証したいというものであった。入手し

1. データをツールに投入 (csv ファイル)



- ① ファイルをアップロード
- ② 列名から予測変数を選択
- ③ 分析実行

2. アウトプットが PPT 形式で自動で出力

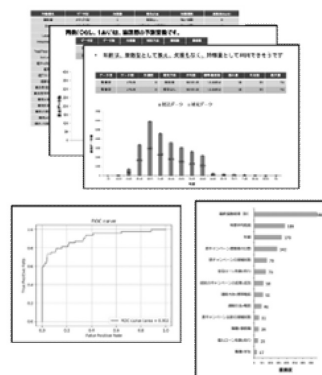


図5 Awesome Intelligence の利用イメージ

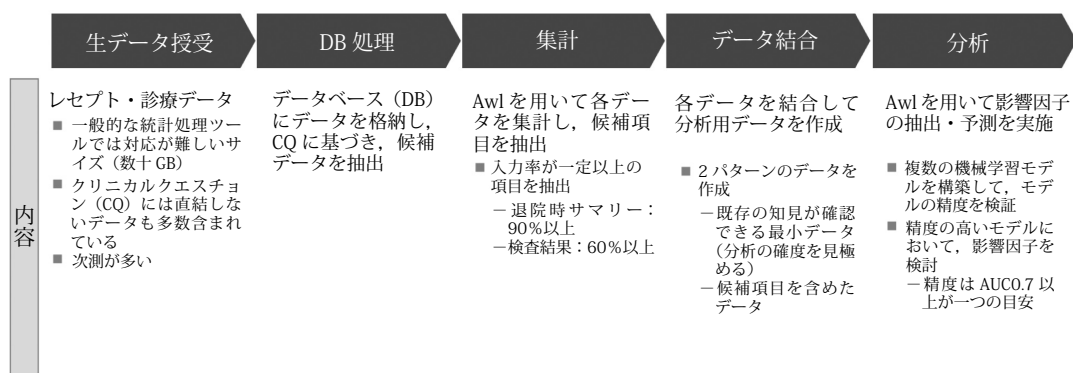


図 6 Awesome Intelligence を用いたエビデンス創出フローの例

た RWD はローデータの段階でデータテーブルが 3 つ、テーブルに含まれているレコード数は最大のもので 4.5 億件と Excel 等の表計算ソフトで扱うには巨大なデータだった。そのため、一旦リレーショナルデータベースに格納した上で、クリニカルクエスションに従って必要なレコードに限定し (DB 処理)、データの妥当性を確認するために各データテーブルにおいて集計を行った (集計)。この際、各項目の入力率を確認した上で一定の入力率を確認した項目のみを分析対象として採用した。集計後、元のデータテーブルを結合して、1 テーブルにまとめた上で AwI に投入した。AwI の実行後得られた予測モデルは AUROC 0.7 を超えており、予測精度の面から内的妥当性については問題ないと判断して、そのモデルにおいて重要度の高い特徴量について検討した。特徴量は治療ガイドラインで留意すべきとされていた腎機能等の検査項目が上位に上がっており、既存知見との合致という観点からみても妥当であることが確認された。さらに特徴的なセグメントの抽出結果を確認すると、先の腎機能について定量化したラインをもって抽出に成功しており、今後臨床医と意見交換する際にケアすべきセグメントが明らかになった。

めの体制づくりについて紹介してきた。製薬業界においても RWD の有効性について検討する PoC のフェーズは終了し、RWD を活用して医学的なエビデンスをどれだけ創出できるかという体制構築のフェーズに移行していると言える。日本国内を見ても RWD の供給については安定化しているものの、体制づくりについては各企業/学術機関によって千差万別である。今後は本稿で紹介した MLops の観点から適切な体制づくりが進められることを期待する。

参考文献

- 1) 岡田美保子, 第 10 回横幹連合コンファレンス (2019)
- 2) Bartlett et al, JAMA Network Open, 2 (10) ,e1912869 (2019)
- 3) D. Sculley et al, Neural Information Processing Systems, 2503-2511 (2015)
- 4) Wilson WF, et al, Circulation, 97 (18), 1837-1847 (1998)

7 最後に

本稿では他業界に先行して機械学習を活用している Google 等の情報系企業が提唱している MLops という観点から RWD を用いて継続的にエビデンス創出をするた