患者症例レジストリを活用した患者状態モデル構築に関する検討

石井雅通　美代賢吾
NCGM 国立国際医療研究センター
{masaishii, kmiyo} @hosp.ncgm.go.jp

１．はじめに

電子カルテシステムの普及に伴い、医療機関では日々の診療を通して患者の臨床情報が電子的に蓄積されてきているが、蓄積された臨床情報に関して一部の利活用は始まっているが、まだ発展途上である。

電子カルテの臨床情報には構造化データと非構造化データが混在している。構造化データは、厚生労働省標準のデータ交換規格である SS-MIX2 ストレージを介して多施設のデータを収集する事業が複数始まっている。

例えば、当センターでは日本糖尿病学会と連携した「診療録直結型全国糖尿病データベース事業（Japan Diabetes comprehensive database project based on an Advanced electronic Medical record System：以下、Ｊ－ＤＲＥＡＭＳ）」を平成２６年度から開始しており、2017 年 12 月現在、35 施設が参加し、今後 4 年間で 100 施設、10～20 万人規模での登録をめざしている。

Ｊ－ＤＲＥＡＭＳでは電子カルテシステムから「糖尿病標準診療テンプレート」により入力された診療情報に加えて、SS-MIX2 ストレージから患者基本情報、検査結果情報、投薬情報を自動抽出し収集しており、観察する臨床データ項目の標準化による糖尿病診療の質の改善及び、合併症進展リスク因子の検索、合併症進展抑制効果の期待される介入の同定、糖尿病薬の副作用についての発生頻度及びリスク因子の確認、糖尿病の未解決課題の発見を目的としている。こうした多施設臨床データの収集、利活用の事業を継続する中で新たな課題が明らかになってきている。

２．現状の課題

２－１．データクレンジングの負担が大きい

人工知能技術の活用においては必要コストの８，９割はデータクレンジングに投入する必要があるのが実態であるが、多施設臨床データ収集事業においてもデータ活用にあたっては同様である。標準規約に従って収集したデータ項目であっても以下のケースが頻発している。

１）ローカルコードに適切な標準コードがマッピングされていない

・病院内の運用に合わせたローカルコードと標準コードの粒度が異なる場合がある

・ローカルコードと標準コードがＮ：Ｍとなるケースがある

２）データ値の表現が統一されていない

・同一標準コードの医薬品や検査結果値に対して、施設により異なる単位が使用される
・検査機器、試薬等に起因する感度の違いにより測定値が変動し、検査結果に対する正常・異常判定の閾値が変動する

２－２．データ入力の負担が大きい

電子カルテシステムには患者の症状に係る多くの臨床情報が記録されているが、こうした患者状態に関してはほとんどの情報がナラティブに記述された非構造化データである。また、臨床で必要とされる記録と研究目的で必要となる臨床情報に乖離がある。そのため臨床研究を目的としてデータ収集を実施するには、現状では必要とされる表現型データを疾患別の患者レジストリとして、電子カルテシステム上でテンプレート等の機能を活用するなどして、医師自らが判断して新規入力する必要があり、臨床医療の現場への大きな負担となっている。

２－３．既存知識の延長となりやすい

症例別の患者レジストリは専門家によるデータ収集である。収集するデータ項目を決定するフェーズで専門医によるフィルターがかかり、データ項目の選別が実施される。そのため、その時点でのドメイン知識によるバイアスがかかる。

２－４．患者状態の全体像の把握が困難である

レジストリが充実するに従い、データ収集項目が増加し特徴量が多くなるため、患者状態の把握が困難となっている。

３．提案手法

３－１．単位変換の自動化

データクレンジング支援機能として収集データ項目ごとの統計量を知識化し、単位換算量候補を自動算出する

(1) 収集済みのレジストリの各データ項について最頻出の単位表記をもつデータを基準データグループとして抽出する
(2) 基準データグループの統計値（平均値、中央値、標準偏差等）を算出する
(3) ２ＳＤ範囲外の数量をもつデータの数量と単位名を例外データとして抽出する
(4) 抽出された例外データについて、単位名ごとに統計値を算出する
(5) 基準データグループの平均値または中央値より単位換算時の係数を算出する
(6) 単位名について n-gram により単位キーワード抽出する
(7) 上記５，６より単位換算量候補を自動算出する

３－２．患者状態のモデル化

レジストリの収集項目には患者状態を評価して、その領域の専門医が定性的に判断した情

報が含まれる。この情報は、電子カルテには直接記載されていない、もしくは自動的に抽出が困難なためレジストリ登録時に医師が入力した情報であり、専門家による正解ラベルと考えられる。

(1) レジストリ項目より、正解ラベルを選定する

(2) 正解ラベル以外のデータ項目、すなわち電子カルテシステムから抽出された処方薬情報、検査結果情報から特徴量データを導出する
特徴量データとしては、患者ごとの時系列数値データとして統計量を算出しベクトル化する
算出例：患者ごとの平均値、移動平均値、標準偏差、前回からの変化量等

(3) 「糖尿病標準診療テンプレート」で収集した患者の診療録を Mecab により形態素解析を行い、患者症状ワードを抽出する。専門用語辞書には万病辞書を使用する。

(4) 上記(2)に上記(3)で抽出した患者症状ワードを素性として追加して、ベイジアンネットワークによりモデル構築を行う

(5) 上記(4)で構築したモデルを活用して、正解ラベルごとのリスク因子を抽出し、患者状態を表現するモデル要素を同定する

４．今後の展開

現在、プロトタイプにより評価実験を計画している。

提案手法による期待効果としては、専門医によるレジストリデータの入力を正解ラベルとして活用することで効率的に当該レジストリの因果構造モデルを構築できることである。また、構築したベイジアンネットワークを用いることで、感度分析を実施することにより、データ収集事業の以下の目的を達成する一助とすることができると考える。

・合併症進展リスク因子の検索
・合併症進展抑制効果の期待される介入の同定
・糖尿病薬の副作用についてのリスク因子の確認
・糖尿病の未解決課題の発見

また、上記プロセスで患者状態を表現するモデルの要素を同定することで、収集するデータ項目を精査することが可能となり、効率的なデータ収集事業とすることで、データ提供元の臨床医の負担軽減の一助とすることができると考える。

謝辞

この研究の一部は、AMED の課題番号 JP17lk1010010 の支援を受けた。

参考文献

[1]矢野,若宮,荒牧 医療テキスト解析のための事実性判定と融合した病名表現認識器, 言語処理学会 第 23 回年次大会 発表論文集, 2017, p242-245

[2]荒牧,岡久,矢野ほか 大規模医療コーパス開発に向けて, 言語処理学会 第 23 回年次大会 発表論文集, 2017, p1200-1203