

ポスター

[PB1～PB19] ポスター

2018年6月23日(土) 15:00 ～ 16:00 ポスター会場 (3階・中会議室302)

[PB19] 臨床データから獲得した属性知識の活用基盤構築への取り組み

石井 雅通（国立国際医療研究センター）

臨床データから獲得した属性知識の活用基盤構築への 取り組み

石井 雅通, 美代 賢吾
NCGM (国立国際医療研究センター)

Proposal for data cleansing method using data knowledge obtained from clinical data

Masamichi Ishii, Kengo Miyo
National Center for Global Health and Medicine

抄録: 多施設で臨床データを収集し活用する事業ではデータクレンジング作業の人的負荷が高い。そこでクレンジング済みの臨床データから統計量として獲得したデータ項目知識に基づいて新たな参加施設から収集した臨床データを評価することでデータ検証の効率化を図る手法を検討した。施設ごとに収集した検査結果値データを同一母集団からの抽出標本であると仮定し予備的な実験を実施したところ仮説は棄却されたが、基本統計量を活用した簡易スクリーニング実現の示唆を得た。

キーワード データ検証, データクレンジング, データ駆動

1. はじめに

多施設で臨床データを収集し活用することを目的とした事業が複数始まっている。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)」^[1] 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に実装された「糖尿病標準診療テンプレート」にて入力された臨床情報に加えて 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のうえ収集しており, 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, 利活用事業を継続する過程で新たな課題が明らかになってきている。

同一施設内においてデータウェアハウス等を活用してデータ解析する場合には問題とはならない

が, J-DREAMS のように多施設からの臨床データを統合し活用するプロジェクトでは前段階でのデータクレンジングは不可避である。例えば医薬品・検査項目においては標準コード設定不具合に始まり, 各施設の実運用と標準コード定義の粒度違いによる不備, 各項目の単位が標準化対象外であるための単位不整合, 同一単位に対する単位表現の違い等, データ活用以前に解決すべき課題が多い

参加施設及び収集データ項目の増加に伴い, クレンジング作業がデータ活用の律速となってきた。従前は人的資源の投入にて対応してきたが, 作業の効率化が喫緊の課題である。人工知能開発を始めとするデータサイエンスにおいても実現コストの大部分がクレンジングに投入されているのが現状である。

2. 方法

多施設から収集した臨床データのクレンジング作業の効率化を目的として, 標準コードの設定不備等に起因するデータ不整合を自動検出する以下の手法について予備的な実験を実施する。

各施設から収集されたデータは国内患者全体を母集団とする抽出標本と仮定し, 既存のクレンジング済の基準データと新規取得の評価対象データが同一母集団からの検査項目データであ

ることを母平均の差の検定により検証する。

当センター本院の2006年11月実施のHbA1c(NGSP)検査30,000件及びJ-DREAMS登録対象データのうち本院の14,025件を基準データ、国府台病院の7,099件を評価対象データとして実験に用いる。

3. 結果

1) HbA1c 基本統計量

実験データから算出したHbA1c(NGSP)の基本統計量をFig.1に示す。全体として最小値3.5%、最大値19.1%のレンジで、中央値は5.8~7.2%、第一四分位点5.3~6.6%、第三四分位点6.7~7.9%の範囲となった。

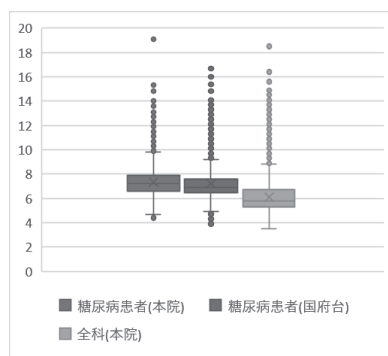


Fig.1 HbA1c 基本統計量

2) 基準データの標本平均基準値分布

基準データの正規確率プロットをFig.2に示す。正規分布は認められない。

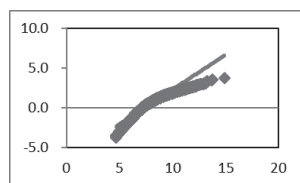


Fig.2 HbA1c 正規確率プロット(基準)

3) 2つの母平均に関する検定統計量分布

基準データ及び評価対象データの2つのデータ群の検定統計量の分布をFig.3に示す。標本サイズは150件、標本調査回数は100回である。

T値平均0.97、標準偏差0.89、歪度-0.11、尖度-0.25であり、t分布が確認された。

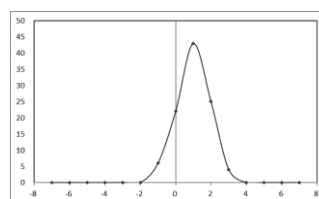


Fig.3 2群の母平均に関するT値分布

4) 2つの母平均の差の検定

基準データ及び評価対象データの2群にウエルチt検定を実施し $P < 0.01$ を得た。同一母集団からの標本データであるとの仮説は棄却された。

4. 考察

基本統計量の比較により、J-DREAMS登録の糖尿病患者については本院を基準データ、国府台を評価データとみなした場合に、評価データの平均値及び中央値は基準データの四分位範囲内であり、データ検証の判断基準に活用できる可能性がある。しかし、本院の全科と比較すると四分位範囲の差異が顕著でありデータ群ごとの患者状態が反映されていると考えられる。

一方、同一母集団からの抽出データとの仮定は母平均の差の検定によって棄却された。選択バイアスによる重症度患者の構成数の違いなど、データ群ごとの特徴が検出された可能性がある。

本結果を踏まえてデータクレンジングに要求される精度に応じた判定手法の検討が必要であり、その選択基準についても今後検証が必要であると考える。

5. 結語

臨床データのデータ検証において統計量による正常性判定実現の可能性が示せた。ただし選択バイアス等の影響を考慮のうえ要求精度に応じた判定手法の検討が課題である。

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] J-DREAMS 診療録直結型全国糖尿病データベース事業 <http://jdreams.jp>