一般口演 | 医療データ解析
一般口演16
医療データ解析
2019年11月23日(土) 14:40 〜 16:40 B会場 (国際会議場 2階コンベンションホールB)

[3-B-2-05] 機械学習を用いた医薬品の潜在的リスクの予測

○西川 景太[1]、種村 菜奈枝[1]、矢向 高弘[1]、漆原 尚巳[1] （1. 慶應義塾大学）

キーワード：Machine learning, Classification, Pediatric, Diabetes, Adverse events

【はじめに】本研究は，大規模医療病院事務・ DPCデータを活用し，機械学習アプローチとして k近傍法（以下， k-NN）を用いて投薬後に発現しうる有害事象のカテゴリ予測を行う．開発時に懸念すべき有害事象を明らかにすることで安全性評価の指標とすることを目指す．

【方法】糖尿病治療薬が投与された16歳未満の小児糖尿病患者を対象とした．データソースは，メディカル・データ・ビジョン社が所有する大規模医療病院事務・ DPCデータであり，2016年1月から2017年12月までの期間を分析対象とした．対象患者の糖尿病治療薬，合併症，有害事象データを日本の小児年齢4区分に従って分割し，時系列順にした．合併症は ICD-10コードの3文字目までの分類 (A00, B49, …)，有害事象は ICD-10コードの1文字目までの分類(A, B, …)を用い名義尺度で数値に変換した．作成したデータセットは訓練用(70 %)，評価用(30 %)に分割した．性別・年齢・合併症・糖尿病治療薬の情報を入力パラメータとし， k-NNを用いて有害事象のカテゴリ予測精度の評価を行なった．

【結果と考察】各年齢区分のデータ数は0歳児: 8,018, 1-6歳: 43,634, 6-11歳: 177,821, 12-15歳: 276,864であった．カテゴリ予測精度は，0歳児はおよそ20 %の精度，1-5歳, 6-11歳, 12-15歳の患者はおよそ70 %の予測精度となった．また，12-15歳で最大76 %（ k=39）の精度となった．低精度の原因として入力パラメータである合併症を一次元の名義尺度で数値化したこと，また0歳児のデータ数が少なかったことが考えられる．今後は入力データの多次元化，および年齢区分の最適化により精度向上が期待できる．（本研究は慶應義塾 学事振興資金より助成を受けて実施された．）

©一般社団法人 日本医療情報学会

機械学習を用いた医薬品の潜在的リスクの予測

西川 景太[*1]、種村 菜奈枝[*1]、
矢向 高弘[*1] 、漆原 尚巳[*1]
*1 慶應義塾大学

Prediction of drugs potential risks using machine learning

Keita Nishikawa[*1] , Nanae Tanemura[*1] ,
Takahiro Yakoh[*1] , Hisashi Urushihara[*1]
*1 Keio University

For drug development, the clinical trials of pediatric diabetes, which is characterized by chronic hyperglycemia in childhood, is considered difficult to conduct because the number of pediatric patients with diabetes in Japan is limited and lower than overseas. Therefore, other alternative method to evaluate the safety of drugs is needed. This study aims to reveal the potential risks of drugs by clarifying the relationship between predicted adverse events and inferring adverse events that may lead to serious ones. In this study, we focused on the pediatric patients with diabetes, and predict the category of adverse events that can occur after medication by k-nearest neighbor method (k-NN) as a machine learning approach using a large-scale hospital information database. The analysis dataset, which were derived from the administrative and Diagnosis Procedure Combination (DPC) data, included approximately 2,577 pediatric patients with diagnosis and medications for diabetes, and was split into the datasets for training and test with a ratio of 70:30. The k-NN-based prediction rules for ICD-10 categories of potential adverse events were developed for the age groups by Japanese pharmaceutical regulations, accounting age and concomitant drugs as attributes. The algorithms for 1- to 15-year-old yielded approximately 70% of predictive accuracy. Further elaboration of the rules is warranted by applying other attributes and neural network.

Keywords: Machine learning, Classification, Pediatric, Diabetes, Adverse events

1. 緒論

医薬品の開発では，販売に至るまでに非臨床試験・臨床試験などの様々な試験を行う必要があるため，十年以上の開発期間を要する．有害事象の存在は開発において大きな障害となりうるため，医薬品の候補物質として特定される早期段階から安全性の評価を行うことが求められる [1]．安全性評価は複雑かつ個人の経験や能力に左右され易いという性質も一因となり，本邦において継続的かつ体系化された評価方法は未だ確立されていない [2]． また，小児患者は開発段階では除外され，適応外使用が多いため，承認時またはそれ以降においても安全性の情報が少なく，安全性の評価が難しいという現状がある．特に慢性的な高血糖によって特徴づけられる糖尿病については，小児慢性特定疾患研究事業に登録された小児糖尿病患者数は約 6,200 人であり [3]，海外に比べて発症率が少ないため評価が難しいと考えられる．

そのような中，近年様々な分野において機械学習技術への注目が高まっており，保険医療分野への応用も重要視されている．

そこで本研究は，リアルワールドデータである大規模医療病院事務・DPC データを用いた機械学習分析によって，投薬後に発現しうる有害事象の予測を行った．医薬品開発時に定まる候補物質を含む薬剤群から懸念すべき有害事象を示すことで，医薬品の承認前における安全性評価の指標とすることや市販後安全性対策をより充実させることを目指す．

2. 目的

本研究では機械学習分析によって，投薬後に発現しうる有害事象の予測を行った．予測した有害事象間の関連性を明らかにし，重篤な有害事象に繋がる可能性のある有害事象を推測することで，医薬品が持つ潜在的な有害事象のリスクを顕在化させることを目的とする．

3. 分析方法

本研究では，大規模医療病院事務・DPC データを用いて k-NN による解析を行った．本項では扱ったデータソースおよび解析方法について述べる．

3.1 データソース

メディカル・データ・ビジョン社が所有する大規模医療病院事務・DPC データから得られた患者情報・薬剤情報・疾患情報を用いた．疾患情報は ICD-10 コード，薬剤情報は ATC コードによる分類コードを使用した．また， 2016 年 1 月から 2017 年 12 月までのデータを分析対象とした．

3.2 データクレンジング

本研究では対象疾患（原疾患）を糖尿病(ICD-10 コード: E10～E14 で始まるコード)に絞り，0 歳児から 15 歳までの糖尿病治療薬 (ATC コード: A10 で始まるコード) を投与された小児患者 2,577 人を対象として抽出を行った．選択理由として，小児糖尿病患者は経年変化による人数変動が少ないこと，また合併症や併用薬の数が成人・高齢患者に比べて少ないため，外的要因による影響が小さいことが挙げられる．
データソースから得られた対象患者の薬剤情報・疾患情報を

日本の年齢区分に従って分割し，各年齢区分のデータついて時系列順にソートした(図 1). そして投薬時に有している原疾患以外の疾患(合併症)と投薬後に発症した有害事象を抽出し，以下のようなデータセットを作成した (表 1, 2). ただし糖尿病治療薬間の区別はしないため，薬剤情報はデータセットに含まない. 合併症は ICD-10 コードの 3 文字目までの分類(A00, B49, …), 有害事象は有害事象のカテゴリとして ICD-10 コードの 1 文字目までの分類(A, B, …)を用い，それぞれ名義尺度で数値に変換した. また，作成したデータセットは訓練用(70 %), 評価用(30 %)に分割した.

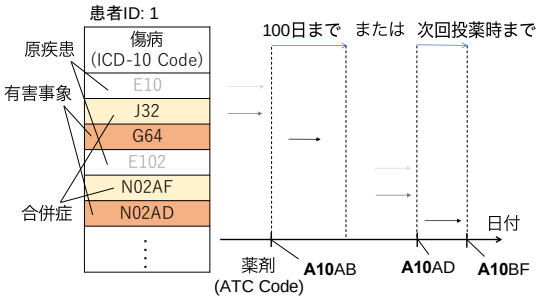

図 1 データクレンジング手法

表 1 データセットの形式 (0 歳児)

患者 ID	性別	年齢	合併症	有害事象
1	1	0	0	0
1	1	0	0	1
⋮	⋮	⋮	⋮	⋮
N	2	0	143	10

表 2 年齢ごとのデータ数

年齢	データ数
0	8,018
1-5	43,634
6-11	177,821
12-15	276,864

3.3 k-NN による分析方法

本研究では性別・年齢・合併症・糖尿病治療薬の情報から，認められた有害事象のカテゴリを予測するために，k-NN を用いて有害事象のカテゴリの分類を行い，その予測精度の評価を年齢区分ごとに行なった.

k-NN は，特徴空間において，入力に近い k 個のデータから多数決によって分類を行う手法である [4]. 解析には機械学習のオープンソースライブラリである Scikit-learn[5]を用いた.

はじめに訓練用データセットを用いて，性別・合併症を入力とし，出力される有害事象が ICD-10 コードのどのカテゴリに分類されるかを予測するモデルを作成した. その後，テスト用データセットを用いて，入力データから予測された有害事象カテゴリがテスト用データセットの有害事象カテゴリと一致するかを比較し，精度の評価を行った.

4. 解析結果

本項では，k-NN による解析結果について述べる. 横軸を k, 縦軸を予測精度とし，k=1~50 について予測精度を年齢別にプロットした(図 2). 0 歳児の患者については，およそ 20 % の精度，1-5 歳，6-11 歳，12-15 歳の患者についてはおよそ 70 %の予測精度となった. また，12-15 歳で k=41 のとき最大 76 %の精度となった.

5. 考察

図 2 に示した通り，1 歳から 15 歳までの患者について，k=2 以上のときおよそ 70 %の予測精度となった. すなわち性別および合併症をもとに未知の有害事象のカテゴリを 70 %の確率で予測することができた. 想定よりも精度が低かった原因として，入力パラメータである合併症を 1 次元の名義尺度で数値化したことが考えられる. したがって多次元的に表現することが必要であると考える. また，0 歳児の予測精度が他の予測精度に比べて低い原因としては，表 2 に示すようにデータセットの数が少なかったことが考えられる. 今後は年齢による精度の揺らぎを減らすため，年齢区分の仕方についての検討が必要と考える.

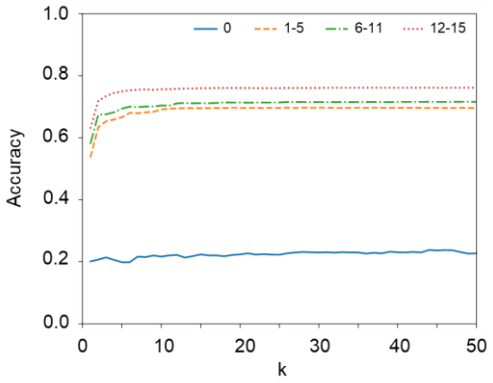

図 2 予測精度

6. 結論

本研究では，大規模医療病院事務・DPC データを活用し，k-NN による有害事象のカテゴリ予測を行なった. 0 歳児のデータについてはおよそ 20 %, 1-15 歳のデータについてはおよそ 70 %と低い精度となった. 今後は入力データの多次元化，および年齢区分の最適化を行うことで，精度向上が期待できる. また，ニューラルネットワークなど，他のアルゴリズムによる精度向上も期待されるため，他の分類手法も視野にいれて比較・検討を行いたい.

7. 倫理的配慮

本研究は慶應義塾大学薬学部 人を対象とする研究倫理委員会の承認（承認番号：承 180418-1），および理工学部生命倫理委員会の承認（承認番号：31-45）を受けて実施した.

参考文献

1) 漆原尚巳. 4. Lifecycle Risk Assess-ment —CIOMS Working Group VI報告書および 米国研究製薬工業団体 SPERT による提案—. 薬剤疫学 2014 ; 19(2) : 123-132.
2) KLEPPER AND COBERT. 薬の安全性を科学する,「くすりの安全性を科学する会」訳 2012 ; 294.
3) 杉原茂孝. 糖尿病の登録・解析・情報提供に関する研究. 「小児慢性特定疾患の登録・管理・解析・情報提供に関する研究」分担研究報告書 2012 ; 127-137.
4) Keller, J. M., Gray, M. R. A fuzzy K-nearest neighbor algorithm. IEEE Transactions on Systems, Man, and Cybernetics 1985 ; SMC-15(4) : 580-585.
5) F. Pedregosa, R. Weiss, M. Brucher. Scikit-learn, Machine Learning in Python. Journal of Machine Learning Research 12 2011 ; 2825-2830.