ポスター

[PA1～PA19] ポスター

2018年6月22日(金) 15:00 ～ 16:00 ポスター会場 (3階・中会議室302)

[PA17] 病院情報システムの構造化データから病態を抽出する Phenotyping 手法の開発

伊豆倉 理江子 （九州大学病院 メディカル・インフォメーションセンター）

病院情報システムの構造化データから病態を抽出する Phenotyping 手法の開発

伊豆倉 理江子[*1], 野原 康伸[*1], 山下 貴範[*1], 濱田 直樹[*2], 鈴木 邦裕[*2], 福山 聡[*2], 松元 幸一郎[*2], 朴 珍相[*1], 高田 敦史[*1], 若田 好史[*1], 神田橋 忠[*1], 中西 洋一[*2], 宇山 佳明[*3], 中島 直樹[*1]

[*1] 九州大学病院 メディカル・インフォメーションセンター, [*2] 九州大学病院 呼吸器科, [*3] 独立行政法人 医薬品医療機器総合機構

Establishment of phenotyping to detect the diseases using the structured data on hospital information system

Rieko Izukura[*1], Yasunobu Nohara[*1], Takanori Yamashita[*1], Naoki Hamada[*2], Kunihiro Suzuki[*2], Satoshi Fukuyama[*2], Koichiro Matsumoto[*2], Jinsang Park[*1], Atsushi Takada[*1], Yoshifumi Wakata[*1], Tadashi Kandabashi[*1], Yoichi Nakanishi[*2], Yoshiaki Uyama[*3], Naoki Nakashima[*1]

[*1] Medical Information Center, Kyushu University Hospital

[*2] Dep. Respiratory Medicine, Kyushu University Hospital

[*3]Pharmaceuticals and Medical Devices Agency

【目的】病院情報システム（HIS）の構造化データから疾患/病態を抽出するための新たな phenotyping 手法を開発し、MID-NET への適用可能性を含め検討した。【方法】間質性肺炎（IP）を対象とし、phenotyping 開発に以下の手法を試みた。1). 専門医の見解やガイドライン等から作成した構造化データを用いた抽出ルールで症例を抽出。2). 専門医レビューで真偽判定。3). 2)の真症例を目的変数とし機械学習を用いて陽性的中度（PPV）やケース内感度をみながらルールを修正。4). 初期ルールでは抽出されない真の症例を CT レポートのキーワードで検索。5). 専門医レビューで真偽判定。6). 機械学習を用いて有効な抽出ルールを模索。7). 3)6)から最終抽出ルールの作成。8). 2)5)での真症例を期間中の全ての真症例と仮定し、各ステップの PPV、ケース内感度を再算出。【結果】機械学習を用いて初期ルールを修正することで、PPV は 36.5%から 80.0%へ上昇した。一方、6).において AUC が 0.58 と低く構造化データを用いた有効な抽出ルールの作成はできなかったが、少なくともその時点で可能なかぎり推定される真症例数がわかり、感度の推定が精緻化された。【考察・結論】1)～9)のうち 6)以外は達成した。6)の達成は疾患・病態に依存する。IP 以外の疾患・病態においても同様のステップによる Phenotyping が可能と思われた。

キーワード MID-NET, Phenotyping, 機械学習, ケース感度, 間質性肺炎

1. はじめに

医療情報データベース事業（MID-NET）では、厚生労働省と PMDA を中心に医薬品の副作用検知のためのデータベース構築を進めてきた。平成 30 年度からの本格稼働に際して、副作用調査対象となる病態を発症した症例を、病院情報システム（HIS）のデータから一定のルールで正確に把握される精度により抽出すること、つまり「Phenotyping」の確立は重要かつ急務である。

本研究では、医薬品リスク管理計画の中から、報告例の多い有害事象である間質性肺炎（以下 IP）を対象とし、発症例を抽出するための HIS の構造化データを用いた phenotyping 手法を開発し手法の MID-NET への適用可能性を検討した。

2. データ・方法

1) データ属性

2014 年と 2015 年の九州大学病院（以下、本院）の全入外来患者 117,401 名を対象とした。構造化データは、患者属性、傷病名（HIS/DPC/退院サマリ）、検体検査値、放射線検査有無、注射実施、処方依頼、生理検査実施を用い、非構造化テキストデータの CT レポートを利用した。

2) Phenotyping 開発工程の構築

以下 3 工程（9 段階）を構想した（Fig.1）。

工程 1: メインの抽出ルール作成

(1) 専門医の見解やガイドライン等をもとに構造化データ（e.g. 病名、処方、検査結果など）を用いた初期抽出ルールを作成

(2) 初期抽出ルールによる症例を抽出

(3) 専門医のレビューによる真偽判定

(4) (3)の真症例を目的変数として機械学習を行い、陽性的中率（PPV）やその時点で把握された真症例から算出した感度（ケース内感度）をみながら PPV が改善するルールに修正（抽出ルール 1）

工程 2: 工程1で非抽出の真症例の抽出

(5) 初期抽出ルールで抽出されない真症例を、構造化/非構造化データを用いて抽出

(6) 専門医レビューによる真偽判定

(7) 機械学習を用いて構造化データで抽出できる有効な抽出ルールの模索（抽出ルール 2）

(8) (4)と(7)から最終ルールの作成

工程 3: 把握された真症例数（all possible case）による感度の精緻化

(9) (3)と(6)の真症例を期間中における当該機関での全真症例と仮定し、各工程の感度を再計算

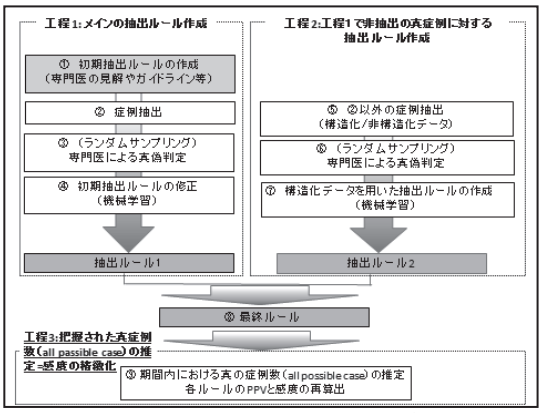

Fig.1 phenotyping の開発工程

3) 倫理的配慮

本研究は九州大学医学研究院倫理審査委員会により承認された（承認番号 29－167 番）。

3. 結果

2.-2)の工程（3 工程/ 9 段階）を IP の Phenotyping 開発へ適用した。

工程 1: メインの抽出ルール作成

(1) 初期抽出ルールは、高い感度かつ単純で明確なルールを想定し、HIS 傷病名と KL-6 を用いて以下のように作成した。

初期抽出ルール: a or b
a. HIS傷病名にIPに関連した確定病名がある
b. KL-6値が施設基準値以上（≧430 U/mL）

(2) (1)から 1424 名が抽出された。

(3) (2)からランダムサンプリングして 200 名を選定し、専門医レビューを実施した。専門医判定は、症例につき 2 名の専門医が独立して 5 段階評価し、評価スコアに差がある症例については、間質性肺炎症例の診療経験年数の多い専門医の評価を採用し、kappa 係数を算出して評価の信頼性を確認した（k=0.74）。本研究では 5 段階のうち「間違いなく確実に IP：スコア 5」を真 IP 症例とし、73 名が真と判定された（PPV36.5%）。その結果、対象期間内に本院では初期抽出ルールで真 IP 症例が 520 名と推定された。

(4) 73 名の真 IP 推定症例を目的変数とし、構造化データの 1406 変数を説明変数として勾配ブースティング（GBDT）手法による AUC から抽出ルール修正の実現可能性を評価した。AUC は 0.798 で、少なくとも 50%以上のケース内感度を確保しつつ、PPV を最大化することを目標とし、関数 rpart の分割基準の中のエントロピーに従い、PPV 80.0%, ケース内感度 60.3%の抽出ルール 1 を策定した。

抽出ルール1: a and (b or c)
a. 初期ルール
b. SP-D値が145μg/mL以上である
c. "間質性肺炎"を含むDPC傷病名がある

工程 2: 工程1で非抽出の真症例の抽出

(5) 初期抽出ルール以外の真 IP 症例の抽出に CT レポートのテキスト検索を用いた。専門医の見解や文献を基に選定した 8 個のキーワードが一つでも含む症例を対象とし、5141 名を抽出した。

(6) (5)からランダムサンプリングして 84 名を選定し、専門医レビューで真の症例は 39 名であった（kappa 係数: k=0.61）。この検索から対象期間内に本院では 2,387 名の真 IP 症例が推定された

(7) 39 名の真症例を予測するため、(4)と同様のデータを用いた GBDT では、AUC が 0.582 と低く、構造化データを用いた有効な抽出ルール 2 は作成できないと判断した。

(8) (7)の結果により最終ルールは(4)とした。

工程 3: 把握された真症例数（all possible case）による感度の精緻化

(9) (3)および(6) から、対象期間内における本院の推定される真 IP 症例数（all possible case）は 2,907 名となった。これを基に各抽出ルールの感度を再算出すると、初期抽出ルールでの感度は 17.9%で、最終ルールでは 10.8%と推定された。

4. 考察

本研究では、MID-NET に適用可能な構造化データを用いて疾患を抽出するルール（Phenotyping）の精度を正確に評価しつつ、かつできる限り精度を向上する手法として 2 つの試みを行った。一つは、機械学習を用いた抽出ルールの精度の精緻化であり、もう一つは初期ルールで抽出されない真の症例を非構造化テキストデータを用いて抽出し、その予測に構造化データによる別の抽出ルールの作成と感度の正確な把握、であった。本研究では前者は達成でき、後者においては予測モデルの AUC が低く、適したルール策定は困難であった。しかし、CT レポート検索から多くの真症例の存在が判明し、All possible case の推定から感度が精緻化された。最終の感度算出においては低い値を得たが正確な感度に近づけたことは一つの成果である。IP のように非構造化テキストデータ上に真の症例が隠れている病態・疾患は多いと推察され、これらを抽出しようという本研究の試みは臨床・疫学研究において大変重要であると考える。

5. 結語

MID-NET で適応可能な構造化されたデータを用いた phenotyping の開発手法を提案し、IP において実現した。

6. 謝辞

本研究は、AMED 事業「アウトカム定義のバリデーション等に関する研究（番号：17mk0101088h0001）（代表：宇山佳明）」によって行った。関係各者に深謝する。

参考文献

[1] 島井良重他：第 34 回医療情報学連合大会論文集 34 558-591, 2014 .