

審査の結果の要旨

氏名 香川璃奈

本研究は、患者が特定の疾患(phenotype)を持つか否かを計算機処理により自動で EHR データから識別する計算機処理技術(e-phenotyping)において、目標精度や利用する評価指標に応じた、用いるべき計算機処理技術の類型化を、特定の医療制度や phenotype に依らずに行うため、医師が疾患の有無を EHR データ要素から判断する際に用いる情報のうち、意味の深い理解が必要となる情報の類型化と、それに基づいた annotation criteria の作成とその e-phenotyping フレームワークへの転用を試みたものであり、下記の結果を得ている。

1. 東京大学医学部附属病院を受診した患者 104,522 人からランダムに抽出した症例の EHR データを実験対象とし、対象とする phenotype は多様な特徴を持つ 10 疾患とした。医師が疾患の有無を EHR データ要素から判断する際に用いる情報について、深い意味の理解が必要か否か、という観点から EHR データ要素を医師 1 名で整理を行った。その結果、医師が EHR データから疾患の有無を推測する際に用いる情報のうち、意味の深い理解が必要となる情報として、病名記載の意味を持たない病名記載、病名記載以外の情報、病名記載の変化、検体検査結果と医師の判断が異なるもの、の 4 種類が存在し、対象疾患全てにこれらの情報のうち少なくとも最初の 3 種類が存在することを示した。この結果に基づき計算機処理における難易度を考慮した annotation criteria を作成した。既存の annotation criteria では判断に迷うような症例を、両軸の正解付けの結果から annotation の根拠を共有しながら言語や施設ごとの患者の特性に依存せずに明示的に annotation できる利便性も示唆された。

2. 作成した annotation criteria を用いて医師が annotation を行った。対象疾患のうち 9 疾患で  $\kappa$  係数  $> 0.81$  と、本 annotation criteria の妥当性が示唆された。また、間接的な情報を用いなければ正しく annotation できない症例が肺塞栓以外のすべての対象疾患で認められ(平均 13.9%)、間接的な情報の利用を明示した本 annotation criteria の有用性が示唆された。さらに、annotation criteria が異なると正例集合の性質が有意に異なり、開発された e-phenotyping 手法の精度の比較も十分に行えないことが示唆された、本研究の annotation criteria の有用性が示唆された。

3. 作成した annotation criteria に基づいて縦軸と横軸それぞれを正しく識別する計算機処理を行うことで、目的とする正例を識別するものを e-phenotyping のフレームワークとし

た。このフレームワークに **annotation** 結果の症例数を適用した計算結果から、**e-phenotyping** 研究において、目標精度が感度または特異度 **95%**以下、または陽性的中率 **80%**以下の場合には病名記載と検体検査値の診断基準を正しく識別するための技術開発を行うべきことが示された。その一方で、**100%**に近い感度や特異度、あるいは先行研究で採用された **95%**以上の陽性的中率を達成するためには間接的な情報も用いた識別が可能な技術開発が必要なことが示唆された。

以上、本論文では、意味の深い理解が必要となる **EHR** データの情報の類型化と、それに基づいた **annotation criteria** の作成を行った。それを **e-phenotyping** フレームワークへ転用したことで、**e-phenotyping** 研究において、目標精度が感度または特異度 **95%**以下または陽性的中率 **80%**以下の場合には病名記載と検体検査値の診断基準を正しく識別するための技術開発を行うべきこと、**100%**に近い感度や特異度、あるいは先行研究で採用された **95%**以上の陽性的中率を達成するためには間接的な情報も用いた識別が可能な技術開発が必要なことが示唆された。医師が疾患の想起に用いる **EHR** データ要素の類型化と、それに基づく **e-phenotyping** フレームワークの作成はこれまでにないものであり、**e-phenotyping** 研究に重要な貢献をなすと考えられ、学位の授与に値するものと考えられる。