

一般口演 | 医療データ解析

一般口演25

医療データ解析

2019年11月24日(日) 13:40 ～ 15:10 F会場 (国際会議場 3階中会議室302)

[4-F-3-03] 退院サマリデータ構造化に基づく合併症パターンの分析

○野口 怜¹、鳥飼 幸太¹、齋藤 勇一郎¹（1. 群馬大学医学部附属病院 システム統合センター）

キーワード：Unstructured data, Text mining, Complications, Discharge summary, EMR

【背景・目的】電子カルテのテキストデータは、患者に関する重要かつ一貫したデータを含み、利活用が大きく期待されているが、非構造化データのため取扱いが難しく活用が進んでいない。著者らは、退院サマリを対象として、1症例1レコード形式にデータを構造化した上で、主訴・既往歴の自由テキスト記述から循環器系・腫瘍系の疾患分類に繋がる特徴語抽出に成功している（第23回日本医療情報学会春季学術大会にて口頭発表）。当該研究は「主病名」に絞った分析であったが、本研究では、更なる疾患分類精度向上や、合併症発症予測への応用に向け、退院サマリ内に自由テキストで記述された合併症データから発症パターンを分析することを目的とした。

【方法】本院の電子カルテより、主に循環器内科、消化器・肝臓内科などで用いられている退院サマリ3年半分6,243件（2015/10/1～2019/3/31）を抽出し、正規表現によりデータを1症例1レコード形式に構造化した。合併症がある場合、「診断病名」欄に「1.疾患 A、2.疾患 B」のように自由テキストで併記されているが、病名を分割し、「主病名」=疾患 A、「合併症①」=疾患 Bのように列方向に構造化して分析容易な形式とした。これらデータに対して、項目間の共起性により関係性（ルール）を見つけ出す手法、アソシエーション分析を適用し、合併症パターンの分析を行った。

【結果】主病名が循環器系の疾患では、慢性腎不全→うっ血性心不全、高血圧症→虚血性心疾患などのルールが抽出され、合併症がない場合に比べ、各々1.8倍近くの発症率（リフト値）であった。一方、腫瘍系疾患の場合、C型肝炎→肝細胞癌のルールが抽出され、C型肝炎がない場合に比べ2倍程度の発症率であった。医学知見としても妥当性のある結果が得られ、本手法の有用性が示唆された。今後は主訴や既往歴の特徴語と組み合わせ、疾患分類への活用を行う予定である。

退院サマリデータ構造化に基づく合併症パターンの分析

野口 怜^{*1}、鳥飼 幸太^{*1}、齋藤 勇一郎^{*1}

^{*1} 群馬大学医学部附属病院 システム統合センター

Analysis of Complication Patterns Based on Structured Discharge Summary Data

Rei Noguchi^{*1}, Kota Torikai^{*2}, Yuichiro Saito^{*3}

^{*1} System Integration Center, Gunma University Hospital

Introduction: Text data recorded in the electric medical record system have important and consistent information about patients and are highly expected to be utilized directly for medical examination or diagnosis. We have previously developed a method for processing free text data in discharge summaries into structured data in which each record was unique for each individual case (corresponding to one discharge summary), and for extracting “semantic units”, a kind of keywords, from each record using a novel natural language processing (NLP) engine. This methodology enabled us to extract disease-specific semantic units (namely “feature terms”) for circulatory diseases and tumor diseases from “Chief complaint” and “Past medical history” sections in the discharge summaries, and to distinguish the diseases from each other. In this study, we focused on complications of the main disease extracted from the free text to use the structured data effectively and to predict the complication patterns.

Methods: The data of complications described as free text originally in the discharge summaries were structured and stored in separate columns “Main disease”, “Complication #1”, “Complication #2” ... for each case (e.g. in case of “1. Disease A, 2. Disease B” described in the free text, Disease A and Disease B were extracted separately and stored in “Main disease” and “Complication #1” columns, respectively). We applied association analysis, a kind of pattern mining approaches that can detect relations hidden in large data sets among variables based on co-occurrence, to the structured data and examined what kind of complication patterns were frequently appear.

Results: For circulatory diseases, association rules {chronic atrial fibrillation} $\Rightarrow$ {chronic heart failure} and {chronic kidney disease} $\Rightarrow$ {chronic heart failure} were extracted (A $\Rightarrow$ B means when A occurs, consequent B occurs with a certain probability), and the co-occurrences in both the rules were about 7 and 5 times, respectively, as frequent as single occurrences of the consequent diseases. In contrast, for tumor diseases, a rule {cirrhosis type C} $\Rightarrow$ {hepatocellular carcinoma} was extracted and the co-occurrence was about 15 times as frequent as the single occurrence of hepatocellular carcinoma.

Conclusion: The results were reasonable from a medical perspective, indicating the usefulness of our approach. We will use these results for developing models to distinguish diseases and to predict complication patterns in combination with “Chief complaint” and “Past medical history” data.

Keywords: Unstructured data, Text mining, Complications, Discharge summary, EMR

1. はじめに

電子カルテのテキストデータは、患者に関する重要かつ一貫したデータを含み、診療への利活用が大きく期待されているが、非構造化データのため取扱いが難しく活用が進んでいない。著者らは、退院サマリを対象として、1 症例 1 レコード形式にデータを構造化した上で、主訴・既往歴の自由テキスト記述から循環器系・腫瘍系の疾患分類に繋がる特徴語抽出に成功している。¹⁾ 当該研究では「主病名」として記述された疾患を分析対象としたが、実際には複数の疾患を合併しているケースが多く見られ、これら合併症の発症パターンは主病名や主訴・既往歴と密接に関係している可能性が高い。

合併症は患者の状態を表す重要な情報であり、主訴や既往歴の記述内容と合わせて分析することで、さらなる疾患分類精度の向上や、症例の類似性評価への活用が期待できる。また、新たな合併症パターンの発見や、合併症の発症予測への応用につなげられる可能性もあり、分析の意義は極めて大きい。

そこで本研究では、疾患分類モデルの精度向上や合併症の発症予測などの診療活用に向け、構造化済みの退院サマリ内の病名データを活用して合併症の分析を行い、各症例を特徴付ける合併症パターンを抽出することを目的とした。

2. 方法

2.1 対象データ

群馬大学医学部附属病院の電子カルテに保管されている、主に循環器内科、呼吸器・アレルギー内科、消化器・肝臓内科で用いられている退院サマリ 3 年半分 6,243 件(期間: 2015/10/1~2019/3/31)を抽出した。

2.2 データの構造化

上記の退院サマリ 6,243 件を匿名化した上で python (version 3.6.0) の分析環境(Jupyter Notebook)に読み込み、正規表現(文字列のパターン抽出を行うための表記法)により、診断病名や主訴、既往歴、現病歴といった項目ごとにテキストを分割して 1 症例 1 レコードの形式に構造化した(図 1)。

さらに、合併症がある場合は、「診断病名」欄に「1. 疾患 A、2. 疾患 B」のように自由テキストで併記されているが、病名を分割し、「主病名」= 疾患 A、「合併症①」= 疾患 B のように列方向に構造化して分析容易な形式とした。

2.3 診断病名の標準化

退院サマリに記載された病名は、医師による自由テキスト記述のため、表記揺れが極めて多く、そのままでは分析が困難である。そこで、電子カルテなどから抽出された実際の臨床現場で用いられる病名（出現形）と、国際標準の疾病分類に基づく病名（標準病名）とを対応付けた病名辞書「万病辞書」²⁾を用いて、各病名を標準病名に置き換えることで名寄せを行った。

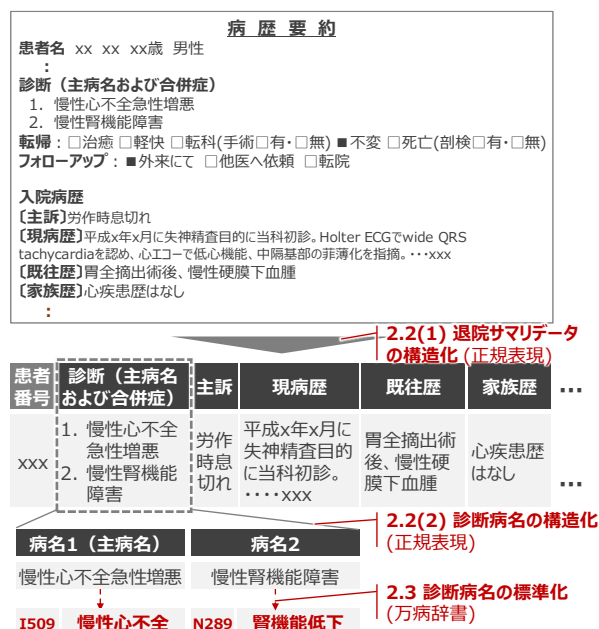


図1 退院サマリデータの構造化と病名標準化

2.4 アソシエーション分析

アソシエーション分析は、主にマーケティング分野において顧客の同時購入商品进行分析する際によく用いられる分析手法である。「教師なし」手法に分類され、膨大なデータの中から、データの共起性に基づいて、発見的に関連性の高いアイテムの組み合わせ（ルール）を見つけることが可能である（例：Xを買う人はYも買う傾向にある）。近年、医療分野でも、医薬品における副作用の関連性分析などで応用され始めており^{3) 4)}、大量データの中での共起ルールを発見したい場合に極めて適合性が高い。

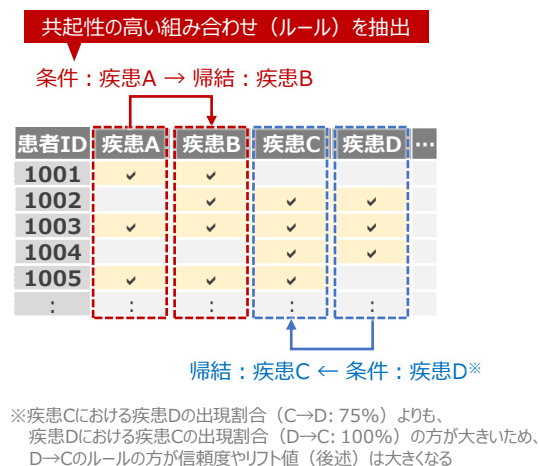


図2 アソシエーション分析適用のイメージ

本研究で分析対象とする退院サマリ内の合併症データは、先述の通り、各病名が列方向に構造化されており、これにアソシエーション分析を適用することで、共起性の高い疾患同士をルールとして抽出した（図2）。なお、分析ツールは JMP Pro v14.1 (SAS 社) を用い、アソシエーション分析のアルゴリズムの1つである Apriori (アプリアリ) を適用した。

抽出されたルールの信頼性や有効性は、「信頼度」「支持度」「リフト値」の3つの指標で評価される（図3）。例えば、「条件：疾患A → 帰結：疾患B」というルールが抽出された場合、「信頼度」が高ければ、このルールの結び付きが強く、疾患Aの罹患患者の中では疾患Bはよく見られる合併症であることを意味する（ルールの信頼性）。「支持度」が高ければ、全般的にもよく起こる合併症であり、全体に対する影響度が大きいことを意味している（ルールの一般性）。「リフト値」が高ければ（1以上）、そのルールの有効性が高く、疾患Bを単独で罹患するよりも、疾患AとBを合併している割合の方が高いことを意味している（ルールの有効性）。

ルール「条件：疾患A → 帰結：疾患B」が抽出された場合

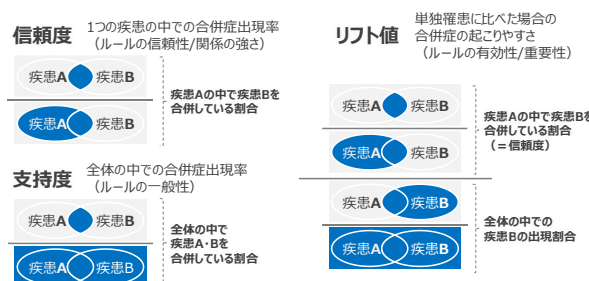


図3 各種指標の定義と考え方

また、上記定義に基づくと、条件と帰結の順序を入れ替えば各指標の算出値が変わりうるということがわかる。つまり、各ルールは「条件→帰結」の方向性を持つことを意味している。例えば、図2で示したように、3例ある疾患Aのうち、すべてのケースで疾患Bに罹患していれば、A→Bのルールは信頼度100%となる。しかし、逆のB→Aのルールでは、合併症割合は4例中3例となり、信頼度は75%に留まる。ただし、この矢印の方向性は必ずしも因果関係を示すものではなく、あくまでデータ上の傾向を反映しているに過ぎない点に注意が必要である。

3. 結果

アソシエーション分析により抽出されたルールを表1に示す。ルールの抽出条件は支持度0.005以上、信頼度20%以上、リフト値1.0以上とした。19件のルールが抽出され、多くが高血圧症や脂質異常症、糖尿病などといった慢性疾患との関連ルールであった。

信頼度50%以上のルールは4件あり、そのうち3件は高血圧症が帰結となるルールであった。信頼度50%以上とは、条件側の疾患の半数以上で帰結側の疾患の合併が見られることを意味し、ルールの信頼性としては比較的高い。また、これらのリフト値も高い値を示し、帰結が高血圧症の3件は4.0以上、残りの{C型肝硬変}→{肝細胞癌}のルールでは15.35という極めて高い値であった。なお、ルール#4は、{脂質異常症、糖尿病}→{高血圧症}という複数疾患が条件となっているルールであるが、脂質異常症と糖尿病を罹患している場合に、高血圧症を同時に合併しているケースが多いことを意味

している。

信頼度 30-40%台では、上位を占める高血圧症・高尿酸血症に関するルールその他、#6{慢性心房細動}→{慢性心不全}、#8{慢性腎臓病}→{慢性心不全}といった心疾患関連のルールがリフト値 5.0 以上という高い値で抽出された。また、#7{睡眠時無呼吸症候群}→{高血圧症}という他のルールの疾患とは傾向の異なるルールも抽出された。

信頼度 20%台では、糖尿病関連のルールや、高血圧症・脂質異常症関連のルールが多く目立ったが、リフト値は 3.0 以下と上位に比べると低い傾向にあった。その中でも、#16{虚血性心疾患}→{慢性腎不全}は、信頼度は 23%ながらも、リフト値は 5.54 と比較的大きい値を示した。

表 1 アソシエーション分析による合併症ルール抽出結果
(支持度 0.005 以上)

#	条件 (疾患A)	帰結 (疾患B)	信頼度	リフト値
1	高尿酸血症	高血圧症	59%	4.78
2	C 型肝硬変	肝細胞癌	57%	15.35
3	脂質異常症	高血圧症	57%	4.56
4	脂質異常症, 糖尿病	高血圧症	53%	4.28
5	高尿酸血症	脂質異常症	43%	8.69
6	慢性心房細動	慢性心不全	42%	6.95
7	睡眠時無呼吸症候群	高血圧症	38%	3.04
8	慢性腎臓病	慢性心不全	31%	5.10
9	高血圧症, 糖尿病	脂質異常症	30%	6.23
10	虚血性心疾患	糖尿病	29%	3.50
11	糖尿病	高血圧症	27%	2.21
12	虚血性心疾患	高血圧症	27%	2.20
13	脂質異常症	糖尿病	27%	3.22
14	慢性腎臓病	高血圧症	27%	2.17
15	高血圧症, 脂質異常症	糖尿病	25%	3.02
16	虚血性心疾患	慢性腎不全	23%	5.54
17	2 型糖尿病	高血圧症	23%	1.86
18	高血圧症	脂質異常症	22%	4.56
19	慢性腎不全	糖尿病	22%	2.59

4 考察

4.1 抽出ルールの医学的な妥当性

表 1 で抽出されたルールは、いずれも医学的にも矛盾しない結果と言える。多くが高血圧症や脂質異常症、糖尿病といった慢性疾患に関連するルールであったが、これらは医学的にも相互に関連性が認められている。

一方、信頼度トップの#1{高尿酸血症}→{高血圧症}は、明確な因果関係があるかはまだ不明とされるが、疫学的研究では関連性が指摘されており、降圧薬服用者における高尿酸血症の割合は男性 40%以上という報告もある。⁵⁾ 同様に、#5{高尿酸血症}→{脂質異常症}においても、明確な因果関係は明らかではないが、高トリグリセリド血症の群で有意に高尿酸血症を合併していたという報告もある。⁶⁾

また、#7{睡眠時無呼吸症候群}→{高血圧症}のルールについても、睡眠時無呼吸症候群による高血圧の合併リスクに関して報告されており⁷⁾、医学的に妥当であると言える。

唯一、腫瘍系疾患に関連するルールとして上がった#2{C 型肝硬変}→{肝細胞癌}においては、C 型肝硬変は肝細胞癌の大きなリスク要因であることが知られており⁸⁾、15.35 という極めて大きいリフト値からも、妥当な結果と言える。

その他、心疾患系のルール、#6{慢性心房細動}→{慢性心不全}、#8{慢性腎臓病}→{慢性心不全}、#10{虚血性心

疾患}→{糖尿病}、#12{虚血性心疾患}→{高血圧症}、#16{虚血性心疾患}→{慢性腎不全}は、いずれにおいても日本循環器学会のガイドラインにて合併症リスクとして挙げられている。^{9) 10)} 以上の観点から、医学的見地からも極めて妥当なルールが抽出されたと言える。

一方、「条件→帰結」というルールの方向性については、医学的な見地とは逆向きになっているルールも存在する。例えば、#12{虚血性心疾患}→{高血圧症}や#19{慢性腎不全}→{糖尿病}は、リスク要因の考え方からは方向性は逆向きであると考えられる。ただし、先に述べたとおり、このルール方向性は必ずしも因果の意味を持たず、あくまでデータ上の発生割合から導出された方向性である。高血圧症や脂質異常症、糖尿病といった慢性疾患は、様々な病気に派生しうる発生件数の高い疾患であるため、データ上もその他疾患と比較して多く存在していると考えられる。このため、図 2 中の注意書きにもある通り、帰結側に発生件数の多い疾患がくる方が、信頼度の高いルールとして抽出されやすくなる。このことを踏まえると、ルールの方向性と医学的な因果関係を直接紐づけるのは難しく、分析結果はあくまで「組み合わせ」として活用し、方向付けは医学的な観点から検討することが望ましい。

また、今回の分析では得られなかったが、もし医学的な妥当性が低いもしくは、不明な結果が得られた場合は、医学的見地からの検証が必要となる。関連診療科の医師と連携し、医学的・疫学的観点からの検証を行うとともに、地域性や偶然性を排除するために、他院との共同研究などを通じて院外データでの分析も併せて行うべきである。

4.2 分析手法の適合性と今後の発展性

アソシエーション分析により、医学的な知見と矛盾しないルールを抽出することができた。本手法は、目的変数を設定しない「教師なし学習」であるため、データドリブンで頻出パターンの探索を行う場合に適している。実際、今回の結果において、医学的に新しい知見こそ見出せていないが、#1{高尿酸血症}→{高血圧症}など、明確な因果関係までは明らかになっていない特徴的な関係性を抽出することができた。

本研究で抽出されたルールと、主訴や既往歴、家族歴といった診療データ・患者背景データを組み合わせ、各ルールを目的変数とした教師あり学習を行うことで、どういった場合にどのような合併症を発症しやすいか分析できると考えられる。さらに、その結果に基づけば、合併症の発症予測にも応用できる可能性がある。

また、各ルールは、疾患分類モデルの説明変数としても活用できる。合併症の発症パターンは、患者の状態を表すデータでもあるため、各主病名を目的変数、主訴・既往歴の特徴語や合併症パターンなどを説明変数とした教師あり学習を行うことで、さらなる分類精度の向上が期待できる。

4.3 分析に用いるべき「病名」

本研究で用いた病名は、退院サマリに記載された診断病名である。自由テキスト記述のため、非構造データであることに加え、表記揺れや略語、記載ミスなどの問題が伴い、本研究においてもまずデータを構造化した上で、万病辞書を活用することで病名を標準化してこの問題にアプローチした。それでもなお、吸収しきれない表記揺れや略語も存在し、さらなる標準化方法の検討が必要である。

一方、DPC データ上にも保険請求に用いるための病名が登録されている。テーブル形式で ICD10 コードと紐付いた形

で DPC 請求単位ごとに病名が記録されており、データ構造化や病名標準化という点では、極めて扱いやすいデータである。しかしながら、保険治療を行わない疾患名は登録されないほか、保険請求を主目的としたいいわゆる「保険病名」が登録されることもある。¹¹⁾ 実際、DPC 導入直後の病院において、多くの医師が退院サマリ上の記載とは異なる傷病名を DPC データに登録しているというアンケート調査もある。¹²⁾ 本研究においても、退院サマリデータと DPC データを結合して、各々の病名記述を比較したところ、完全一致するものはほとんどなかった。特に、退院サマリ記載の病名の方がより詳細に記述されているケースが多く見られ、医師が患者の背景や入院経過を想起しながら記述する退院サマリの方が、より患者の実態を細く表した病名が記録されていると考えられる。そのため、退院サマリ内の病名データを活用した本研究の意義は極めて大きく、より詳細かつ精度の高い分析が可能となる。

5. 結言

構造化した退院サマリ内の病名データに対してアソシエーション分析を適用することで、医学的に妥当性のある合併症の発症パターンが抽出でき、本手法の有用性が示唆された。今後は本分析結果と主訴や既往歴の特徴語と組み合わせ、診療活用に向けた合併症の発症予測モデルの構築や、疾患分類モデルの精度向上につなげたい。

参考文献

- 1) 野口怜, 鳥飼幸太, 齋藤勇一郎. 新しい自然言語処理エンジンを用いた電子カルテデータ構造化と疾患別特徴語抽出. 第 23 回日本医療情報学会春季学術大会, 2019 年 6 月.
- 2) Ito K, Nagai H, Okahisa T, et al. LREC, 2018: 2365-2369.
- 3) 増川直裕. 「アソシエーション分析」に”視覚的なグルーピング手法”をプラスした医薬品副作用の関連性分析. SAS Institute Japan, 2017.
[<https://community.jmp.com/kvoqx44227/attachments/kvoqx44227/discovery-ja-2017-content/7/1/C3-Naohiro-Masukawa.pdf> (cited 2019-May-23)]
- 4) 井上宏高, 杉田かほり, 高松和正, 山田秀. 医薬品副作用の関連性検討手法に関する研究 ～アソシエーション分析とコレスポネンデンス分析を用いた探索的検討～. 薬剤疫学, 2018; 23, s130-s131.
- 5) 土橋卓也. 高血圧患者における尿酸管理の意義. Gout and Nucleic Acid Metabolism, 2016: Vol.40 No.1.
- 6) 大岩加奈, 重見博子, 森田美穂子, 松田安史, 上田孝典, 山内高弘. 健診受診者における高尿酸血症と脂質異常症についての検討. Gout and Nucleic Acid Metabolism, 2016: Vol.40 No.1.
- 7) 荻尾七臣. 睡眠時無呼吸症候群と高血圧. 第 59 回日本心臓病学会学術集会シンポジウム, 2012: Vol. 7, No. 1.
- 8) 肝炎診療ガイドライン作成委員会. C 型肝炎治療ガイドライン. 日本肝臓学会, 2017.
- 9) 日本循環器学会/日本心不全学会 合同研究班. 急性・慢性心不全診療ガイドライン. 日本循環器学会/日本心不全学会, 2017.
- 10) 日本循環器学会 合同研究班. 虚血性心疾患の一次予防ガイドライン. 日本循環器学会/日本心不全学会, 2012.
- 11) 武田理宏, 真鍋史朗, 松村泰志. 電子カルテデータ二次利用の現状と課題. 生体医工学, 2017: 55(4), 151-158.
- 12) 下川忠弘. 傷病名付け・傷病名コーディングについて—診療報酬を請求する観点と診断群分類(DPC)の精緻化を図る観点から—. 診療録管理, 2007 ; 19(2), 52.