

<表紙>

臨床試験課題名：「多施設統合退院サマリーデータベースの臨床応用」

研究代表者：

氏名、鈴木隆弘

所属機関・診療科(部)、医学部附属病院企画情報部

住所、千葉市中央区亥鼻1-8-1

電話番号(内線)、043-222-7171 (6473)

FAX 番号、043-226-2373

E-mail アドレス、suzuki@ho.chiba-u.ac.jp

緊急連絡先：043-226-2346

臨床試験実施予定期間：平成28年承認後から平成31年3月

○ 作成日・承認日・改訂日：

2012年10月20日 計画書案 第1版作成

<目次>

1. 研究の背景	34
2. 研究の目的および意義	35
3. 研究対象者の選定	45
4. 研究の方法および研究の科学的合理性の根拠	45
5. 研究実施期間	56
6. 予期される利益と不利益	57
7. インフォームド・コンセントを受ける手続き	57
8. 研究対象者に緊急かつ明白な危機が生じている状況における研究の取り扱い	57
9. 健康被害発生時の対処方法	57
10. 個人情報の保護方法	57
11. 研究資金および利益相反	67
12. 試料・情報の2次利用および他研究機関への提供の可能性	68
13. 研究組織	68
14. 記録等の保管	69
15. 研究成果の発表方法	79
16. 研究対象者等の経済的負担又は謝礼	79
17. 参考資料・文献リスト	79

1. 研究の背景

近年の電子カルテシステムの普及によって、医療情報は電子化された形で蓄積されつつある。当院においても、電子カルテシステムが2003年6月より本稼働を始め、既に大量のデータが集積されている。また、電子化された医療情報をもとに、統計手法を用いて新たな規則や知識を発見する研究も行われてきた。しかし、それらの多くは医療情報の中でも、入力・保存や統計処理が容易な数値データや画像データが対象であり、医師のカルテ記録や退院時サマリーなどの診療文書は見過ごされてきた。それは、システムが複雑で電子化される時期が遅かったこと、従来の自然言語処理手法が診療文書の特殊性に対応できなかったことなどがあげられる。

主任研究者らは、早くから診療文書の解析に取り組み、自然言語処理の手法には、高精度のテキストマイニングを行うことができる形態素解析と tf-idf 法 (図2) によるベクトル空間モデル (図3) を採用し、様々な知識の抽出を行ってきた。

形態素解析では、対象となる文書を形態素に分解して解析を行うが、一般用の辞書では医療用語が分解されてしまうため、高精度の医療用語辞書を用い、医療用語を分解せず単語認定できるようにした。方法としては、一般に利用されている医療用語集の他に、千葉大学医学部附属病院で使用されている処方マスター、検査マスター、術式マスターなど実際に医療機関で使用されている用語集を利用した。また、退院時サマリーに出現する単語を調べ、複数の文書で使われている語に関しては随時追加登録することで、研究を重ねるたびに辞書精度を向上させ、30万語を超す医療用語が既に辞書へ登録されている。抽出された単語の出現頻度に基づいて、tf×idf 法により単語の重要度を算出する。得られたデータからベクトル空間が構築され、ベクトル同士の内積値が文書同士の類似度となる。これらの手法により、診断やDPCの自動判定、類似症例検索などが可能となることを証明してきた。[1-3]。また、内科学会の保有している症例データベースからの類似症例検索システムを構築[4]した際に、他の学会からも同様のシステム構築の検討を依頼されている。

テキストマイニングに関する先行研究では、いずれも1施設、1種類のみの方書を対象としているのに対して、主任研究申請者らは聖路加国際病院及び佐賀大学医学部附属病院との共同研究において、本手法を用いて施設に関わらず高精度のテキストマイニングが行うことができることを示した[4]。平成24年から26年には、これをさらに発展させた複数施設の統合退院サマリーデータベースの構築をおこなった。

前研究である統合退院サマリーデータベースの構築では、各施設の退院サマリートのスタイルはかなり異なっており、特に長さでは4倍以上の違いが見られた。しかしながらMDC毎の違いは似通っており、疾患に共通する要素の存在を示唆していた。施設間クロスマッチDPC自動判定の結果からは多施設のデータを統合することで普遍性のある共通の文書ベクトル空間を持つ大型テキストデータベースを構築できると考えられた。また、他施設のデータを一括検索できる類似症例検索システムを千葉大学に設置しての使用経験からは、原文閲覧への要望が強く寄せられた。しかし形態素解析をしていないサマリートの原文は、ほとんどの施設で外部への持ち出しの許可が得られなかった。今後もサマリー文書全体の持ち出し許可を得ることは困難な見込みだが、個々の症例については匿名化してあれば可能と考えられるので、指定されたサマリー原文を入手できる仕組みを整備する必要性を痛感した。

2. 研究の目的および意義

本研究「多施設統合退院サマリーデータベースの臨床応用」は、複数の医療機関から退院サマリーや症例報告を電子的に抽出し、テキストマイニング技術によって共通の文書ベクトル空間を構築し

た大型文書データベースを用い、類似症例の検索やDPCの自動判定を初めとした応用を試行し、技術的および法律的な課題を整理・解決していくことを目的とする。将来的には全国の主要病院における症例検索を可能にするなど大きく臨床医学に貢献できることが期待できる。(図1)

3. 研究対象者の選定

(1) 選択基準:

2012年1月から2015年12月までの全入院患者を対象とする。症例数は総計で約28万例を見込む。そのうち当院における症例数は約6万例である。

(2) 除外基準:

患者から除外の申し出があった場合は以降の研究から除外する。但し、既に抽出と連結不可能匿名化がなされていた場合には、研究者自身も識別が出来ないために除外は行えない。

4. 研究の方法および研究の科学的合理性の根拠

本研究では連携研究者の所属する病院から退院サマリーを収集し、千葉大学で開発したテキストマイニング技法を用いて解析することで、共通の文書ベクトル空間を構築して自動疾患判定や類似症例の検索を可能とすると共に精度の検証を行い、疾患の自動判定や症例検索を可能とする。

本研究は3カ年計画で実施する。平成28年度は、各施設へのアプリケーションの配布とサマリーの追加収集及び類似症例検索プログラムの改良を主として行う。平成29年度はアプリケーションの操作記録を収集し、分析を行う。また、サマリー原文へのアクセス方法についての検討と試行を行う。平成30年度は、研究の総括として研究成果の評価を行う。

・研究体制

鈴木(主任研究者)は統括とデータ解析を主に担当する。本多(連携研究者)、松村(連携研究者)、高崎(連携研究者)、津本(連携研究者)、横井(連携研究者)、畠山(連携研究者)、嶋田(研究協力者:聖路加国際病院)は各々の施設において退院サマリーの収集と基本的な解析および一般ユーザーへの情報提供を行う。土井(連携研究者)はシステム構築と整備を主に担当する。

・病院情報システムとの接続

前回の研究で整備した機能を用いて病院情報システム内に保管されている退院サマリーの外部への一括出力を行う。

・形態素解析と辞書の整備

本研究では形態素解析に「MeCab(和布蕪)」を使用する。退院サマリーは一般の文書に比べて用語の専門性が非常に高いだけでなく、略語や施設に特有の単語などが多く含まれている。そのために標準的な辞書だけでは電子カルテから十分に医学用語を索引語として反映させることがで

きない。我々はこれまでに約23万語の医学用語を抽出した辞書を準備しており、鈴木、高崎、嶋田が作成に協力したパラメディカル用医療辞書であるComeJisyoも併用する。施設固有の単語などは参加施設の拡大に伴い不足が予想されるので、未知語として抽出された語句の中から重要な単語を選び出して辞書を改良して行く。

・ベクトル空間モデル

抽出したサマリーは索引語に分解して解析を行う。索引語の重み付けに本研究では計算の速いtf×idf法を用いてサマリーベクトル空間を構築する。類似度の算出はベクトルの内積によって求める。

・類似症例検索アプリケーションの配布

全施設のベクトルデータと自施設のサマリー原文を収容した類似症例検索プログラムを協力医療機関内に設置し、院内での利用と分析のための環境を構築する。

平成29－30年度の研究計画・方法

臨床での利用を通じて、操作性の改善点や検索結果の評価、検索に使われるキーワードや操作ログの収集を行って、アプリケーションの評価と改良を行う。また、研究成果を国内外に発表する。

・サマリー原文の閲覧

他院の退院サマリーを閲覧するには個人情報の確実な除去が必要である。原文から個人情報を除去して提供するために必要な要件の整理をおこない、病院の情報資産公開において問題がおきないための技術的・運用的対策を講ずる。

・成果報告

研究成果は国内の医療情報関連学会にて報告する。また、国際医療情報学会(MEDINFO2017)、欧州医療情報学会などの国際学会においても発表を行う。

5. 研究実施期間

平成28年承認後より平成31年3月年まで

6. 予期される利益と不利益

利益：

本研究により被験者が直接受けることができる利益はない。

不利益：

本研究は、既に記載されている診療データを二次利用するものであり、臨床的な不利益及び危険性は存在しない。

解析においては、データ抽出の際に個人識別情報を除去して匿名化した後に単語単位に分解・集計するため、元の文章を復元することは不可能であり、個人情報は存在しない。

7. インフォームド・コンセントを受ける手続き

本研究は過去に遡って当院の一定期間における全ての入院患者を対象にしていることから、インフォームド・コンセントのための手続きを全ての患者に求めることが不可能であり、また診療の結果として得られたカルテ情報から個人識別情報を完全に除去して匿名化したデータの解析であることから、黙示の同意をとることとする。説明は、千葉大学医学部附属病院の「患者の皆様の個人情報取扱いに関するお知らせ」にある、「症例に基づく研究及び学会・研究会・論文等での発表」に相当すると考える。研究についての情報は当院の企画情報部ホームページ上で公開する。

8. 研究対象者に緊急かつ明白な危機が生じている状況における研究の取り扱い

本研究は、既に記載されている診療データを二次利用するものであり、臨床的な不利益及び危険性は存在し無い。

9. 健康被害発生時の対処方法

本研究は、通常の診療の範囲内で得られた情報を研究対象とするので、被験者に対する直接的な研究目的の侵襲性を伴う行為は一切行わない。そのため、本研究に起因する健康被害が発生することはない。

10. 個人情報の保護方法

本研究において必要な情報は退院サマリー本文と診断情報及びDPCコードのみで、個人識別情報は不要であり、収集対象としない。特に、形態素解析後にベクトル化される過程でデータからは個人

情報は完全に削除される。

電子カルテなどの横断的検索にあたってどのように個人情報削除し秘匿化しなければならないか、さらに秘匿化された状態においても電子カルテを用いて広域的研究をするにあたっては個々の患者から何らかの同意が必要であるか、についての規範は十分にできあがっていない。本研究では実施にあたって各施設の倫理委員会に申請すると共に、これらについても十分配慮し、今後行われる同種の研究の先駆けとしてもふさわしい人権の保護と法令の遵守を検討する。

1 1. 研究資金および利益相反

本研究は科学研究費で行われ、利益相反は存在しない。

1 2. 試料・情報の2次利用および他研究機関への提供の可能性

無し

1 3. 研究組織

研究代表者

千葉大学医学部附属病院企画情報部

准教授 鈴木 隆弘 研究試験全般の策定・実施・報告上の調整、および施設間の調整

共同研究者

千葉大学医学部附属病院地域医療連携部

助教 土井 俊祐 システム開発

千葉大学医学部附属病院企画情報部

特任助教 島井 健一郎 データ整理

香川大学医学部附属病院医療情報部

教授 横井 英人 データ収集

大阪大学医学部附属病院医療情報部

教授 松村 泰志 データ収集

長崎大学医学部附属病院医療情報部

教授 本多 正幸 データ収集

佐賀大学医学部附属病院医療情報部

准教授 高崎 光浩 データ収集

島根大学医学部附属病院医療情報部

教授 津本 周作 データ収集

聖路加国際病院医療情報センター

センター長 嶋田 元 データ収集

高知大学医学部附属病院企画情報部

准教授 畠山 豊 データ収集

1 4. 記録等の保管

本研究に係わる必須文書およびデータは各施設の研究協力者所属部門（千葉大学では企画情報部）内に保存し、研究発表後に廃棄する。なお、データは全て抽出時に匿名化されている。

15. 研究成果の発表方法

本研究で得られた知見は、国内外の関連する学会等で公表を行うと共に、得られたベクトル空間を用いて各種検索を行えるアプリケーションを開発し、患者個人情報保護、サマリー公開において問題が起きないための技術的対策を講じた上で、連携医療機関の一般ユーザーからも利用可能とする。

16. 研究対象者等の経済的負担又は謝礼

本研究では参加者の費用負担は皆無であり、謝礼の支払いも無い。

17. 参考資料・文献リスト

1. 鈴木隆弘、小野大樹、横井英人、井宮淳、高林克日己（2006）退院サマリーのテキストマイニングにおけるエンтроピー法と tf×idf 法の比較. 医療情報学, 25:173
2. 鈴木隆弘、藤田伸輔、石塚琳、高林克日己（2005）テキストマイニングによる DPC 自動判定の試み. 医療情報学連合大会論文集 25:686
3. 鈴木隆弘、土井俊祐、藤田伸輔、高林克日己（2007）テキストマイニングによる退院サマリーからの類似症例検索の試み. 医療情報学連合大会論文集 27:635
4. 土井俊祐、鈴木隆弘、藤田伸輔、高林克日己、田村俊世（2009）退院時サマリーのテキストマイニングによる類似症例検索. 第 48 回日本生体医工学会論文集:24am0-14-1
5. 土井俊祐、嶋田元、鈴木隆弘、田村俊世、高林克日己（2008）複数病院間でのテキストデータを用いた DPC 判定の試み. 医療情報学連合大会論文集 28:3-I-2-I
6. 鈴木隆弘、土井俊祐、本多正幸、嶋田元、高崎光浩、津本周作、畠山豊、松村泰志、横井英人、高林克日己（2013）テキストマイニングによる退院サマリーの多施設間クロスマッチ比較. 医療情報学連合大会論文集 33:940-943