ポスター
# ポスター2
# 情報セキュリティ・プライバシー
2017年11月21日(火) 14:15 〜 15:15 L会場（ポスター会場2） (12F ホワイエ)

[2-L-1-PP2-1] 日本語自然文で記載された電子カルテ記述の匿名化

渡辺 淳, 仲野 俊成 （関西医科大学 大学情報センター）

【背景・目的】改正個人情報保護法施行と次世代診療情報基盤整備法成立に伴い、完全匿名化診療データの利活用が期待されている。本研究では、診療記録の自然文における個人を特定し得る情報（以下、個人情報と称する）の完全匿名化の問題点について検討した。【材料と方法】患者プロファイル等を含む構造化データを除去した電子カルテ記載から抽出したアセスメント項およびプラン項の日本語自然文（約2万文）における個人情報記述の状況を調べて匿名化にあたっての問題点を洗い出し、処理方略を検討した。【結果と考察】自然文1,000文あたり、患者姓名（姓のみの場合も含む）は1~2箇所観察された。姓・名共に記載されていたのは逆紹介のための診療情報提供書やコンサルテーション依頼文書(一部)で、他は姓のみであった患者名に必ず敬称(「様」）が付されており、人名と強い共起関係がみられた敬称（様）を抽出することで、患者氏名を遺漏なく抽出できた。しかしながら、敬称がついた名前のほとんどは「患者様」「奥様」など固有の人名を指さないもの（1000文あたり10箇所程度）であり、さらに病態等を示す「〜様」等の敬称以外の用法が1,000文あたり約300箇所含まれていた。姓名とも記載された患者氏名の少数については、匿名化されなかった場合、近傍に記載された医療機関名を匿名化しない場合には、一意に特定される可能性が示唆された。他方、診療スタッフ等の氏名の記載は1000文あたり20〜30箇所存在し、敬称・職位等を指標に抽出できないものが1,000文あたり20箇所程度存在した。それらの少数(4%)は、近傍記載の診療科名等の記述から、 Webに公開された情報の検索で個人の特定に至った。また、医師についてはk-匿名化でkが3以下となるケースが10%を超えた。匿名加工の自動化による匿名化率は人名辞書の収載語数に大きく依存し、目視確認なしでは姓名の検出漏れによる残存によって適切な匿名化が困難となるケースが生ずる可能性が示された。

日本語自然文で記載された電子カルテ記述の匿名化

渡辺 淳[*1]、仲野俊成[*2]、
*1 *2 関西医科大学 大学情報センター

Anonymization of descriptions in the electronic medical record written in Japanese natural language.

Jun Watanabe[*1] , Toshiaki Nakano [*1]
*1 University Information Center, Kansai Medical University

An effective utilization of anonymized medical records has been expected by the enforcement of the Amended Personal Information Protection Law and the establishment of the Next Generation Medical Information Infrastructure Improvement Law. The present study was conducted to examine the problems of anonymization of information that allowed identification of individuals written in natural Japanese sentences in medical records. After removal of descriptions concerning patient profiles described as the structured data in the electronic medical records, about 20,000 Japanese natural sentences written in the assessment and plan sheets in the records were examined using text mining and visual examination. Of the 20,000 natural sentences, about 30 patient names (including the last name only) were observed. Full name of individual patient frequently appeared in the records if doctors referred the patients to other departments, clinics or rehabilitation hospitals. Except for this case, patient names were written as their surname followed by the honorific title "sama". A strong co-occurrence was found between patient's name and "sama", and thus the patient name could be identified using the honorific title as a marker. However, the title could not be used for the marker of individual patient name. Many individual staff names were found in the records, and most of them were extracted using honorific and job/position titles. A small number (4%) of them reached identification of individuals by web search. In addition, not a few physicians' names were narrowed down using the web search (k < 3 for 10% or more of the detected physicians' names), using the same approach. These findings suggest that an automatic anonymization is practically difficult due to detection failure or detection error of the personal information.

Keywords: Anonymization, Personal information, Natural sentences, Medical Records

1. はじめに

1.1 背景

改正個人情報保護法（個人情報の保護に関する法律）[1]の施行に続いて、次世代医療基盤法（医療分野の研究開発に資するための匿名加工医療情報に関する法律）[2]が可決・成立したことによって、匿名化された診療データを、医療機関以外の大学・研究機関や製薬企業等の第 3 者機関がビッグデータとして分析することに道が拓かれつつある。当初は、医療における喫緊の諸課題に対して、おもに構造化診療データを解析するためのデータ収集のルール・基盤の整備が優先されると予測される。一方、診療情報は要配慮情報であり、データの利活用に際しては、個人情報（本稿では、氏名、生年月日その他の記述等により特定の個人を識別することができるもの、および他情報との照合によって、特定の個人が識別可能となるものを指す）を適切に処理し、個人を特定可能な情報の漏出を防ぐことが必須となる。構造化された診療データでは、所見や検査値に紐付けられている個人情報の所在は明確であり、データと個人情報との結びつき方も、多くの場合、シンプルかつ明確であると推定される。そこで、構造化された電子化診療データの利活用に際してデータに紐付いた個人情報を処理する場合、それらのデータが紐付いている個人情報を「どの程度まで、どういった方策によって匿名化するか」という点は検討すべき課題ではあるが、その個人情報がどのようなかたちで電子化された診療システム（電子カルテシステム等）のどこに格納されているかは明確であり、個人情報をシステマティックに検出・特定することが可能である。

電子的診療記録には日本語自然文として記載された非構造化ディジタルデータも大量に蓄積されつつある。それらの非構造化データには患者の病状に対する診療スタッフによる所見の評価や治療方針の根拠などが含まれている。近年の自然言語処理技術および AI の導入によって、日本語自然文として記載された非構造化診療データの活用についても道が拓かれつつある。そこで、電子的診療記録に自然文で記述された情報の利活用への期待も大きい。

日本語自然文として記載された非構造化診療データの利活用には自然文で記載された記述の計算機処理が可能な必要がある。我々は、日本語自然文を、誰が読んでもその意味がわかり、かつ、計算機を用いた処理が可能な文構造を持つ文（正規化文）に変換するためのアルゴリズムの構築と検証を進めてきた [3-7]。その過程で、診療録に記載された日本語自然文に、個人の識別・特定を可能とする情報が大量に含まれることを見出した。それらの情報は、診療スタッフの記載するシートに前後の文脈と密接に関係して記載されており、診療情報システムが構造化データとして管理している個人情報とは独立している。そこで、これらの非構造化データを構造化データと同様の方略を用いてシステマティックに匿名化することは困難である。他方、自然文では個人の姓名の前後に、姓名と組み合わされることで個人識別を可能・容易とする情報が存在する可能性があることから、それらを適切に処理し、非構造化診療データの匿名性を確保する手法の確立は必須と考えられる。そのためには、まず、電子化された非構造化診療データにおける個人情報を検出・特定するための方略の策定とその検証が必要になると考えられる。また、今後の非構

造化診療データの効果的な利活用に際しては、非構造化診療データにおける個人情報の検出・特定を計算機で処理可能かどうかが、ひとつのポイントになると予測される。

1.2 目的

診療記録に自然文で記載・蓄積されている大量の非構造化データを、医療資源として利活用するためには、それらの非構造化データに含まれている個人情報を完全に匿名化し、意図しない個人情報の流出。拡散を防ぐ必要がある。さらに、匿名化処理ができるだけ人手をかけずに実施できることが、実際のデータの利活用に際しては重要になると考えられる。本研究の目的は、1) 電子化された非構造化診療データにおける個人情報を検出・特定するための方略の策定、2) その方略を用いた際の匿名化の検証および匿名化処理の自動化に向けての問題点の洗い出しにある。そのために、まず、非構造化診療データにどのような個人情報が、どのようなかたち(文脈)で記載されているかについて、テキストマイニングと目視・用手法を併用して検討した。その検討結果にもとづいて匿名化プロセスを自動化するためのアルゴリズムを設計し、コードを実装して検証するとともに、個人特定に至る情報の漏出を防ぎつつ、データを有効活用するための要件・課題の洗い出しを試みた。

2. 材料と方法

2.1 材料

約 750 床の特定機能病院に 2015 年 1 月から 4 月に入院していた患者の電子カルテ記述のうち，アセスメント項およびプラン項の自然文(約 2 万文)を試料とした。

試料の採取にあたっては、まず、電子カルテ DB から解析対象期間の SOAP の記述(xml)を患者毎に時系列順に記載が並ぶように抽出した。抽出直後のデータには患者基本情報や記載者情報は含まれていないが、構造化データとして、入退院日、シートの記載日時(タイムスタンプ)、患者 IO が含まれている。抽出後、直ちに患者 ID、サブジェクトシートとオブジェクトシートの記載、およびアセスメントおよびプラン記述以外の要素(タグ、他項目からの引用情報、患者 ID)を、GNU sed, grep を中心としたシェルスクリプト、Perl で記述した小プログラム群とエディタ (Vim, Mi) を組み合わせて除去した。上述の処理後、時系列に並べられた患者ごとのアセスメントシートおよびプランシートの自然文の集合に対して、仮 ID を紐付け、入退院日の代わりに入院日の月と入院日数、記載日の代わりに入院日からの経過日数を付したデータを作成した(以下、ID 置換データと称する)。また、仮 ID および日時のデータを除去し、アセスメントシートおよびプランシートの自然文だけから成る文例集(コーパス:以下、自然文データと称する)も用意した。

2.2 方法

まず、ID 置換データを用いて、文中に出現する個人情報(同一患者に関する記述において複数の情報-語句を組み合わせた場合を含めて特定の個人を識別できる可能性を有する名詞、名詞句)を、当該個人の立場・種別(患者、患者家族、診療スタッフ、連携先医療機関のスタッフ、司法・行政関係者など)を付して、目視・用手法で抽出した。

次に、ID 置換データから抽出した情報(名詞、名詞句)を自然文データ上にマーキングしなおし、マーキングされた語句に隣接するか、または近傍に位置する語群との共起関係を、テキストマイニングを用いて解析した。また、個人情報を含む文については、必要に応じて、形態素分析の後、係り受け解析に供した。さらに、個人情報が出現する文脈の特徴について、目視・用手法を用いて検討した。なお、テキストマイニングに際しては、匿名化処理前のデータに加えて、文中に出現する氏名、診療機関名等の機関。組織名称を、換字法を用いて匿名化したデータも解析した。匿名化に際しては、氏名・機関名の先頭2文字の読みを一定の規則を用いて換字して平仮名表記とし、続いて自然文データでその語が出現した行番号を 6 桁の数字で記した。なお、ID 置換データについては、データシート(1回に記載された記述)ごとに通し番号を振り、通し番号(5 桁)を行番号の前に置いた。この方法によって、匿名化されたデータは見かけ上一意となるため、テキストマイニングへの影響が最小化される。さらに、正規表現を用いてそれらの情報が書かれた場所を容易に特定、抽出可能とした。

上述の抽出・解析を実施後、個人情報と共起関係または係り受け関係にあった語をプローブとした場合に個人情報をどの程度検出・抽出できるかどうかを当該個人の立場・種別ごとに調べ、実際に記載された情報のどの程度を検出できたか:感度に相当)、および検出された情報のうちどの程度が個人情報であったか(特異度に相当)を算出した。

続いて、匿名化前に抽出した個人情報を用いて、個人の特定が可能かどうかを、Web 検索を用いて調べた。検索エンジンは google を用い、調査対象となった医療機関のサイト(病院ホームページ、講座・診療科のホームページ、研究者データベース)を除外して検索した結果、およびそれらを含めて検索した結果から、患者(家族を含む)、診療スタッフ、その他関係者がどの程度まで特定できるかを調べた。

テキストマイニングの実装には KH coder [8] を用い、前後に出現する単語の出現頻度と位置を KWIC コンコーダンス (Key Word in Context Concordance) およびコロケーション統計を用いて調べた。形態素解析には MeCab [9]、係り受け解析には CaboCha [10] を用い、結果の描画および統計解析には GNU R を用いた。

3. 結果

3.1 自然文における個人情報の出現頻度

目視・用手法による個人情報の検出に際し、複数の情報-語句を組み合わせた場合を含めて特定の個人を識別できる可能性を有する語として、患者の情報にあっては患者氏名(姓のみの場合を含む)、電話番号、住所、親族の氏名と続柄、診療スタッフについては氏名(姓のみの場合を含む)。職種、連携先医療機関のスタッフおよび司法・行政関係者などの氏名(姓のみの場合を含む)を選定した。

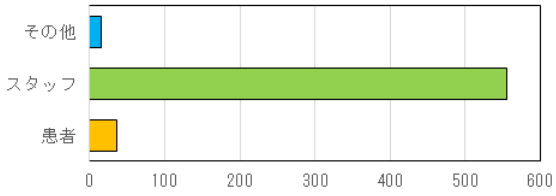

図 1　自然文における姓名（姓のみを含む）の出現数

診療記録の自然文約2万文から検出された数。橙色は患者、緑色は医師を含む診療スタッフ、青は司法・行政関係者等。

目視・用手法で抽出した患者姓名(姓のみの場合も含む)は自然文 1,000 文あたり 1~2 箇所、全体で 36 箇所が検出さ

れた。診療スタッフ等の氏名の記載は 1000 文あたり 20〜30 箇所（全体で556箇所）存在し、1ヶ月間の記載に百箇所を越える診療スタッフの氏名（性のみの場合含む）が記述されていることが明らかとなった。その他、連携先医療機関のスタッフおよび司法・行政関係者などの氏名と属性の組が1ヶ月あたり数か所（全体で 15 箇所）観察された。また、診療施設等の機関固有名称の記載は1ヶ月あたり 10~20 箇所（全体で 63 箇所）存在した (図1)。

3.2 自然文匿名化前データからの個人特定

【患者】 ID 置換データを用いた場合には、氏名（姓のみを含む）記載のあったほとんどの患者の入院診療科を前後の記述内容によって推定可能であり、少なくとも 3 割 (12 名)の患者については入院病棟（転棟情報を含む）の特定または推定が可能であり、当該患者が SNS 等に情報を公開していた場合等には個人の特定に至る可能性が危惧された。なお、3例については、患者氏名に加えて連絡先電話番号、紹介予定先施設名とその理由等の複数の情報の記載があり、電話帳と詳細な住宅地図を併用することで、個人特定が可能なことが判明した。一方、順不同に配置した自然文データでは、文例周から抜き出した1文だけで患者の特定に至る確率は ID 置換データに較べて低いと推測された。しかしながら、個人の特定に至った上述3例中1例については、順不同の自然文データから抽出した1文と詳細な住宅地図の併用によって個人を特定できる可能性が示された。これらのことから、自然文記事際における匿名化処理が適切でないか遺漏があった場合には、患者を特定できるケースがあり得ることが判明した。

【患者家族】 患者の家族に関する記述だけから、個人を特定し得る可能性が推測されるケースは、氏名と連絡先携帯電話番号が記載されていた1例を除いて認められなかった。しかしながら、患者本人の情報が加味された場合には、特定に至るケースがあり得る可能性が示唆された。

【診療スタッフ】

ID 置換データを用いた場合、氏名（姓のみを含む）の記載があったほぼ全てのスタッフについて、所属診療科・部署および職種の特定または推定が可能であり、一部については職位、アルバイト先の診療機関名の記載もあった。これらのことから、自然文記事際における匿名化処理が適切でないか遺漏があった場合には、診療スタッフの個人特定に至る可能性が高いことが推定された。

【その他】 連携先医療機関のスタッフおよび司法・行政関係者については、すべてのケースで氏名に加えて勤務場所、職種、職位、連絡先の2項目以上が組み合わされて記述されており、15 例中 11 例で個人の確実な特定に至った。

3.3 テキストマイニングを用いた共起関係の解析

自然文データにおける個人情報に隣接または近傍に位置する語群との共起関係を解析した。個人の姓名（姓のみの場合も含む）とそれに続く敬称の間に強い共起関係が観察された。目視・用手法による確認作業の結果、今回、検索対象とした自然文中の患者および患者家族の姓・姓名の後には敬称として「様（さま）」が付与されていた。医療スタッフについては、記載差以外の医師には「先生、それ以外のスタッフには「さん」が付けられているケースが多かった。連携先医療機関や司法・行政機関のスタッフには「氏」、「さん」、「殿（どの）」が用いられていた。

次に、個人名と関連性の高かった敬称を指標として、敬称に用いられている語の直前に位置する語が人名であるかどうかを、KWIC コンコーダンスの結果を目視評価することで調べた。その結果、「様」をキーワードとして抽出した場合、キーワードの直前に位置する語の約 6 割は個人名であったが、残りの大半（全体の3割強）は状態等を示す語に付与された「様（よう）」であり、残りは「奥様（おくさま）、ご子息様、患者様等の非固有名詞であった。「さん」の約3割は個人名または個人名＋職種・職位を表す語の連続であったが、大部分は個人名を欠く職種・職位に敬称を付した記述（たとえば「師長さん」、「主任さん」、「PT さん」など）であった。「先生」の直前の語の半数弱は人名（多くが姓のみ）で、残りは「先生」または「担当医の先生」など、個人特定に結びつく固有名詞を欠いていた。敬称として用いられた「殿（どの）」の直前には個人名が位置していたが、「殿」を抽出すると「殿部」「殿筋」等、人名と関係しない語も抽出された。

患者氏名を含む固有名詞の記述が最も多かったのは医師で、メディカルソーシャルワーカー(MSW)が僅差で続いた。これは、連携先病院との病床の調整報告や地域医療パスの受け渡し先情報の提供等で連携先施設の医師、担当者の姓名、連絡先、連携先施設名称等の記載が多かったことと、MSW が照会・連絡先として自己の姓、職種、連絡先を、必ず文中に記載していたことによる。

引き続いて、敬称の直後に続く語について同様の手法を用いて検討した。「様」については、「は」「を」「へ」等の特定の助詞が接続している場合には、様の前に人名または「患者あるいは患者親族の続柄を表す語」が非常に高い確率で存在することが明らかとなった。他方、「様」に続く助詞が「の」、「に」等であった場合には、様の直前語が個人名を含めた人を指すのか、状態等を指すかを区別することが困難なことが判明した。

抽出・解析を実施後、個人情報と共起関係または係り受け関係にあった語をプローブとして個人情報をどの程度検出・抽出できるかどうかを当該個人の立場・種別ごとに調べ、実際に記載された情報のどの程度を検出できたか：感度に相当）、および検出された情報のうちどの程度が個人情報であったか（特異度に相当）を算出した。「様」をプローブとした場合、患者氏名（姓名または姓のみ）をほとんどすべて（今回の解析ではすべて）、検出することができた（感度=100%）。しかしながら、上述のように敬称以外の用法や一般名詞＋敬称の型として用いられていたケースが 40%程度存在しており、特異度は約 60%であった。敬称に続く助詞を加えることで、特異度を 80%強とすることができたが、その際の感度は 80%台に低下し、多数の検出漏れが生じることが明らかとなった。

一連の解析の最後に、敬称の前置語が人名かどうかの判定支援のために、形態素解析に用いた MeCab に人名辞書の導入を試みた。辞書の導入によって、辞書に収載した語の多くについては、固有名詞であることの判定を有効に支援できる可能性が示された。しかしながら、辞書が効果を発揮できないケースが少なくないことも判明した。たとえば今、回の検討対象の記述に比較的多く出てきた「奥様」という語については、それが一般名詞としての女性配偶者を指すのか、患者の姓が「奥」氏であるのかを判定できなかった。

上述の結果は、自然文データ、ID 置換データで変わらなかった。なお、匿名化後のデータ（加工済み）については、匿名化された部位が患者特定に至る固有名詞であって容易かつ正確に抽出可能となった。これらのことから、匿名化処理が適切に実施できれば、その後の解析精度を維持できる可能性が示唆されたが、匿名化処理を施行すべき語の判定を、目

視・用手法を用いずに適切に施行することが容易でないことも、また、示されたと考えられる。

3.4 個人情報が出現する文脈の特徴

個人情報が出現する文脈の特徴について、目視・用手法を用いて検討した。ID 変換データを解析した場合、患者個人名は患者毎に時系列でまとめられた記述の先頭に近い場所に出現する傾向があった。ただし、姓名が記載されているか、それとも「患者様は」と、一般名詞で記述されているかどうかについて、前後の文脈や係り受け関係から推定することはできなかった。他方、医師を含む診療スタッフの姓名については、その出現場所に特徴が見られるケースがいくつか、明らかとなった。今回の検索対象とした記載では、少数の診療科の手術記録が「手術レポート」としてアセスメントシートに記載されていた。その場合、手術記録の決まった場所に、術者（執刀医）、第一助手等の手術スタッフや麻酔医氏名、主治医氏名等が記載されており、それらは【手術記録】という文中表記を指標として抽出することができた。また、アセスメントシート記載の他科コンサルテーション依頼文書等の他診療科や他医療施設向け文書の本文部分等において、その前段にコンサルテーション依頼先等の医師の氏名、紹介患者の氏名、末尾に依頼元医師の氏名が書かれていることが多い傾向が観察された。そこで、コンサルテーション依頼記述をたとえば「侍史」等の敬称群で特定することで、個人名が高確率で出現し得る場所を特定できる可能性が示された。また、ID 変換データでは、シート毎の記載末尾に、記載者氏名が記されていることが少なくなく、これは医師の他にも多くの職種で共通する傾向があった。さらに、他の医療スタッフに何かを依頼する記述があった場合、「・・・お願いします。」「・・しておいて下さい。」等の依頼文の後に記載者の姓または姓名が記載される傾向も認められた。しかしながら、これらの結果をもとに、逆の操作、すなわち個人情報記載場所を機械的・効率的に遺漏なく選び出すことは、現時点では困難であることが判明した。加えて、患者や診療スタッフ等の個人名が想定した場所に出現するか否かについて、文脈から推定することもまた、困難であった。

3.5 Web 検索による個人情報の特定

匿名化前に抽出した個人情報を用いて、個人の特定が可能かどうかを、検索エンジンに google を用いて調べた。調査対象とした医療機関のサイト（病院ホームページ、講座・診療科のホームページ、研究者データベース）を除外して検索した場合、姓名等の個人情報が含まれていても、web 検索だけからでは患者の特定に至ることは、今回の検討では困難であった。医師に関しては姓名（姓のみの場合もあり）と前後の記述から抽出または推定した診療科（他に職位等の情報があればそれを加味）の組、他の診療スタッフについては氏名と職種（職位等の情報があればそれモ加味）の組を検索語とした。調査対象とした医療機関のサイト情報を除外した場合、指名記載のあった医師の約 4%が一意（k=1）で検出され、1%弱の誤検出（他人が一意でヒット）が観察された。検索にヒットした個人が複数でその中に該当する医師が含まれているケースは少なくなく、4 名以下の候補者に含まれたケース (k<5) は 10%を超えた。調査対象とした医療機関のサイトを含めた場合、姓名または姓の記載があった医師全体の 20%以上が診療科と組み合わせられることで特定された（k=1）、4 名以下の候補者(k<5)に含まれたケースは 30%を超えた。また、人名以外に、医師をはじめとするスタッフに関係する診療機関を特定または推定できる情報が残っていた場合、その多くを一意に特定できることも判明した。

4 考察

本研究によって、構造化データとして保持された個人情報を除去しても、まだ、大量の個人情報が自由記述された自然文（非構造化）データが残存していることが明らかとなった。また、患者個人情報については、本研究で対象としたデータにおいては、特定の敬称「様（さま）」をプローブとすることによって高感度（遺漏の少ない）の検出が可能である可能性が示された。しかしながら、敬称に用いられる語・文字は固有名詞のみならず一般名詞にも付けられており、さらに、他の用語にも多く含まれていた。「様」の場合、一部については人名辞書の利用や後置助詞によって患者氏名かどうか、敬称なのか容態を示す接尾語なのかを判定できたが、判定困難な場合の方がはるかに多かった。また、検出感度が高い場合でも、特異度をあまり高くすることができないこと、そして、特異度が高くないため、検出された匿名化が必要な場所に対して匿名化処理を機械的に実施すると、現時点では多くの箇所で誤った処理が行われてしまう危険性が高いことが判明した。これらのことから、プローブに用いた語・文字を指標として姓名等の個人情報を高精度で特定することは、単純な機械的手法では困難と考えられる。また、規則を用いた網羅的抽出が困難出会った診療スタッフの個人情報の記載を検出することは、一層、困難であると推測される。加えて、診療スタッフに関しては、職種によって、自然文に記述された情報から、多数の個人特定が可能となることが示された。特に、医師の姓名と診療科を組み合わせた場合、母集団が極めて限定されてしまうことに加えてインターネットに個人名・診療科情報が豊富なケースが少なくないことによって、個人特定に至るケースが多いことが明らかとなった。

診療記録に自然文で記載された診療情報の利活用には、解析の前に、それらの記述に含まれている個人特定が可能な情報を遺漏なく匿名化する必要がある。本研究によって、診療記録から構造化データとして保持されている個人情報を除去しても、非構造化データである自然文の記述には大量の個人情報が残存していることが示された。また、自然文の解析によって明らかとなった個人情報に関係する語の特徴は、目視・用手法での個人情報に関する記述検出の効率化に役立つことが期待できると考えられる。

このように、目視・用手法を効率化するための方略は明らかとなったが、個人情報の遺漏のない検出・匿名化処理を目視・用手法を用いずに行うための方略については明らかにできなかった。ビッグデータとして存在する非構造化診療データの匿名化を、従来の目視・用手法のみで行うための労力、時間、コストは軽視できないものではあるが、自然文にける個人情報の遺検出・匿名化処理の自動化は、現時点では容易でないことが示されたと考えられる。今後、検討すべき方略としては、人工知能・機械学習を試料の前処理のステップである匿名化処理についても個人情報の検出・判定のために導入する手法、あるいは、構造化データの匿名処理前に自然文を構造化データに記載されている個人情報と対比（紐付け）して特定した後に自然文を匿名化する方法などが考えられる。ただし、少なくともしばらくの間は機械学習の結果の検証に目視・用手法が必要となることが予測される、また、構造化データとの対比による匿名化では患者氏名とシート記載者の氏名については精度の高い匿名化が期待できるものの、本研究によって、それら以外の個人情報が自然文に含まれているこ

とが判明している。これらのことから、匿名化に際しての目視・用手法の完全な排除は、今後、しばらくの間については困難と予測される。

5 結語

電子カルテのデータから構造化された個人情報を除去しても、自然文の記述に個人特定に結びつく情報が多数残存していた。本研究によって、自然文における個人情報の目視。用手法を用いた検出の効率化に、個人情報に関する語・記述の特徴を利用可能なことが明らかとなったが、同時に、現時点では、自然文における個人情報を自動的に遺漏なく検出して匿名化処理することは、は極めて困難なことが明らかとなった。また、匿名化に遺漏があった場合、患者情報だけでなく、特に診療スタッフの個人情報が守られない状態でデータが取り扱われてしまう危険性について留意すべきであることが判明した。

6 謝辞

本研究の一部はJSPS科研費 26330337の助成を受けた。

参考文献

1) 個人情報の保護に関する法律. 最終改正:平成二八年五月二七日法律第五一号 2016 [http://law.e-gov.go.jp/htmldata/H15/H15HO057.html (Cited 2017-Sep-07)]
2) 医療分野の研究開発に資するための匿名加工医療情報に関する法律. 平成二九年五月一二日 法律第二十八号 2017 [http://www.sangiin.go.jp/japanese/joho1/kousei/gian/193/pdf/s031930531930.pdf (Cited 2017-Sep-07)]
3) 渡辺 淳、仲野俊成、北村 臣. 電子カルテに記載された少量情報を用いた意思決定の道筋解析と展開予測に及ぼす暗黙知の影響. 医療情報学 2012 ; 32s :1458-1460.
4) 渡辺 淳、仲野俊成. 日本語自然文で記載された診療記録の構文解析の試み. 医療情報学 2013 ; 33s :832-835.
5) 渡辺 淳、仲野俊成. 日本語自然文で記載された診療記録記述のパラレルコーパスを用いた正規化. 医療情報学 2014;34s:786-789.
6) 渡辺 淳, 仲野俊成, 石原久美子, 夜野敏明. 日本語自然文で記述された診療記録を正規化変換するためのパラレルコーパス構造の検討. 医療情報学 2015; 35s: 260-263.
7) 渡辺 淳, 仲野俊成, 夜野敏明. 石原久美子. 対比文例集を用いた非構造化診療データ記述の正規化に向けたテキストマイニングによる日本語自然文の解析. 医療情報学 2015; 35s: 906-909.
8) 樋口耕一. テキスト型データの計量的分析 -2 つのアプローチの峻別と統合-. 理論と方法 2004; 19: 101-115.
9) MeCab: Yet Another Part-of-Speech and Morphological Analyzer 2006-2013 [http://taku910.github.io/mecab/ (Cited 2017-Sep-07)]
10) 工藤 拓、松本裕治. チャンキングの段階適用による日本語係り受け解析. 情報処理学会論文誌 2002 ; 43:1834-1842.