(別添 1)

平成 28 年度厚生労働科学研究費補助金
政策科学総合研究事業(臨床研究等 ICT 基盤構築研究事業)

カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築, 及び, 自動構造化機能を有した入力機構の開発に関する研究

平成２８年度 総括・分担研究報告書

研究代表者 荒牧　英治

平成２９（２０１７）年3月

(別添 2)

目次

(別添 3)

平成 28 年度厚生労働科学研究費補助金
政策科学総合研究事業(臨床研究等 ICT 基盤構築研究事業)

カルテ情報の自動構造化システムと疾患数理モデルの逐次的構築に関する研究

電子カルテは患者情報が全て記録されているものの，非文法的かつ断片化した表現が多く自然言語処理を応用した利活用は困難であった．これを二次利用するため申請者等（申請者荒牧及び分担者河添が所属する研究室主宰者の大江ら）は，2008年から電子カルテから医療用語の自動抽出及び自動コーディングを行う研究に従事してきた．その成果は，日本内科学会の症例報告検索システムなどとして実用化され，現在も用いられている．本研究は，電子カルテの二次利用のさらなる実用化に向けて問題となる次の２つの課題を解決する．

（課題１）実用化可能な解析精度の達成（マッピング精度80%）

（課題２）電子カルテに組み込み可能な実装の開発

荒牧英治 奈良先端科学技術大学院大学 研究推進機構

若宮翔子（奈良先端科学技術大学院大学 研究推進機構･博士研究員）

河添悦昌（東京大学医学部附属病院 企画情報運営部･講師）

A. 研究目的

本研究の目的は，電子カルテに自由記載された文章を対象に，これを二次利用可能な状態に自動変換する技術を確立することである．これを実現するために，（問題１）現状の解析システムの解析精度を向上させ，これを（問題２）多様かつ複雑な電子カルテシステムに組み込む．これは３つのモジュールから構成される（**図1**）．

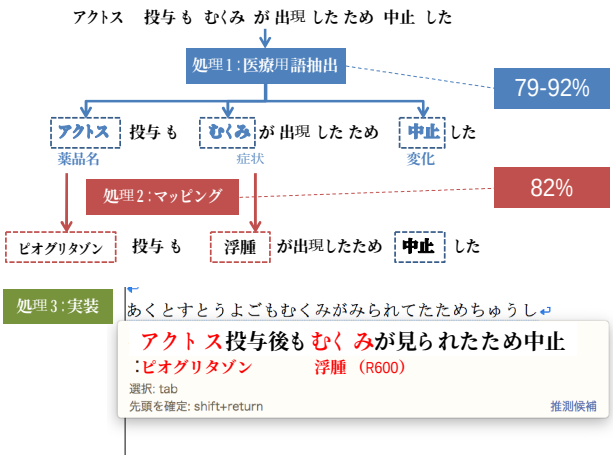


図 1: 本研究にて開発する３つの処理

（２つの言語処理技術とその実装）

（１）【処理１：医療用語抽出】

電子カルテ中の自然文から医療表現（時間表現と疾患／症状表現）を抽出する．

（２）【処理２：標準化変換（マッピング）】

自由記載された病名をICD10コードへマッピングする．これまで，出現頻度が低い（まれな）コードへのマッピングは困難であったが，前段（医療用語抽出モジュール）の結果を用いて，どのような患者がどのようなコードを付与されやすいかという確率モデルを構築する．

（３）【処理３：実装】

処理１と処理２により，既存の電子カルテ情報については後ろ向き解析が可能となるが，それでも一定の誤りが含まれてしまう．そこで，新たに電子カルテに医師等が入力する際に，標準化を行った結果をサジェストするという前向き処理機構を開発する．これにより，現場の医師の負担となることなく，自然と標準的なデータが蓄積されることを目指す．

【特色と独創性】

従来の多くの同様の研究は，ラボレベルの実験にとどまるものが多かった．本研究は新しく入力する際にも利用可能な実装も含めて開発の範疇に入れている．このような出口を持つことで，さらにデータの蓄積が爆発的に増加することが考えられ，それが解析精度を向上させるというプラスの循環を作ることができる．

B．研究方法と成果

以下の３つのシステム，リソース，アプリの開発を行い目的を達成した．

（１） 汎用病名抽出器 MedEX/J の開発／配布／評価

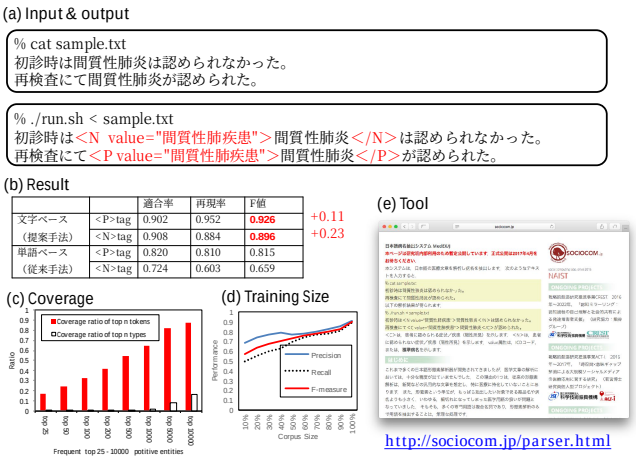


図 1: MedEX/J の概要．（a）MedEX/J への入力（上）と出力（下）．陽性所見が＜P＞，陰性所見が＜N＞としてタグ付けされて出力される．（b）タグ付けの精度．陽性所見は 0.926, 陰性所見 0.896 と，従来の単語ベースの手法を 10 ポイント程度上回っている．（c）コーパスにおける病名の出現統計．頻出する n 病名が対象とするコーパス全体に出現する病名をどの程度カバーしているかを

示したもの（塗りつぶし＝延べ単語数，白色＝種類数）．頻出する 10000 病名が 90%近くをカバーしているが，種類数としては 20%に満たない．すなわち，少数しか出現しない病名が無数にあることを示している．(d) 教師データの量（X 軸）と病名特定精度（Y 軸）．教師データが増えると順調に抽出精度は向上する．来年度増強するコーパスで 95%を達成する試算である．(e) MedEX/J 配布サイト．本ページにてツールを試験公開している．

本システムは，日本語の医療文章を解析し病名を抽出する．例えば，図 1（a）上のテキストを入力すると，図 1（a）下の解析結果が得られる．ここで，＜P＞は，患者に認められる症状／疾患（陽性所見）を示し，＜N＞は，患者に認められない症状／疾患（陰性所見）を示す．value 属性は標準病名を示す．

予備実験の結果，病名抽出においては形態素解析を用いず，いきなり文字そのものを処理する方式の方が高精度であることが分かり（図 1（b）：陽性抽出の F 値 0.926，陰性抽出の F 値 0.896），この結果を受けて，形態素解析部を省くことで，よりコンパクトな解析器を構築できることになった．速度についても汎用機（core-i7-6core 3.4GHz; Memory 32GB；70 万円相当）にて，3000 退院サマリを 120 秒で処理可能など十分実用に耐えうる．

現在は，班内および共同研究者に試験配布を行っている（図 1（e）http://sociocom.jp/parser.html ）．

（2）MedEX/J に利用する辞書「万病辞書」構築

カルテ文章調査の結果，延べ45万症状表現（種類数としては6.2万種類）が得られ，その28.3%（種類数としては87.5%）が，標準病名でカバーされていないことが分かった．このうち高頻度（頻度30回出現の5,600病名）を扱い医療従事者3名によりコーディングを行い，意見が食い違ったものはその曖昧性も残したまま辞書リソース化した（通称「万病辞書」）．この万病辞書により，現在すでにカルテに出現する80%（ただし種類数としては20%）の症状／病名を標準病名に変換可能である．

（3）日本語入力パレットの開発

日本語入力パレット（通常のIMEを用いて入力を行うと標準病名に変換した結果がサジェストされる）を開発した（図2）．

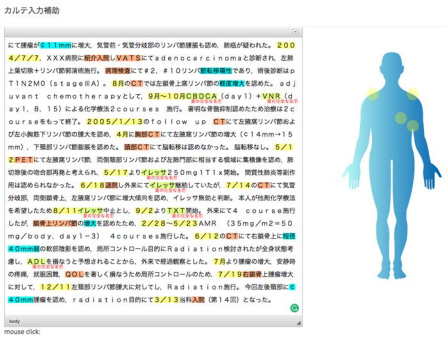


図 2:試作した日本語入力パレット．電子カルテに入力する前に，本パレット上で入力を行い，入力した病名が

標準的かどうかを確認しながら入力ができる．また，クリックにより，サジェストされる標準病名と置換可能である．

D．結論

これまで多くの日本語形態素解析器(mecab, juman など)が開発されてきたが, 医学文章の解析においては, 十分な精度が出ていなかった．この理由の１つは，従来の形態素解析は，新聞などの汎用的な文章を想定し，特に医療に特化していないことにある． また，形態素という単位が，もっぱら抽出したい対象である薬品名や病名よりも小さく，いわゆる，細切れになってしまう問題もある．

このような問題を解決するために，本研究班で開発する MedEX/J は，形態素ではなく，病名用語抽出に特化し，その後処理として，標準病名への標準化，事実性判定など，研究，臨床的に重要な処理も組み込んだ．

E．研究発表

1. 論文発表

- E.Aramaki, K.Yano, S.Wakamiya: MedEx/J: A One-scan Simple and Fast NLP tool for Japanese Clinical Texts, MedInfo, 2017.（採択）

2. 学会発表

- 矢野憲，伊藤薫，若宮翔子，荒牧英治: 深層学習による医療テキストからの固有表現抽出器の開発とその性能評価： 人工知能学会全国大会（JSAI）, 2017.（査読なし）
- 矢野憲，若宮翔子，荒牧英治： 医療テキスト解析のための事実性判定と融合した固有表現認識器，言語処理学会年次大会，2017.（査読なし）

F. 健康危険情報

該当なし

(別添 4)

平成 28 年度厚生労働科学研究費補助金
政策科学総合研究事業(臨床研究等 ICT 基盤構築研究事業)
分担研究報告書
カルテ文章からの病名自動抽出に関する研究

研究分担者：若宮翔子 奈良先端科学技術大学院大学 研究推進機構

A.研究目的

電子カルテを初めとする医療テキストには，患者情報か あまねく記録されていると考えられるものの計算機て 扱いか 困難な非文法的かつ断片化した表現か 多く含まれており，利活用は困難て゛あ った．このため，自然言語処理を用いた医療テキスト処理が注目されているが，日本語テキストを扱う標準的な言語処理ツールは未だ存在しない．本研究では，日本語テキストを処理するために，日本語による医療用自然言語処理ツール MedEx/J を提案する．これは以下の２つの処理を行う．

- **事象認識**：医療テキストにおける病名および疾患名を識別する．これは，一般的な固有表現認識タスクと同等である．以降，本稿では，この処理をER (Entity Recognition) とも呼ぶ．
- **陽性／陰性 (P/N) 分類**：テキストには，予測，疑い，可能性，否定など様々なモダリティの事象が含まれている．これらは大別すると， 現在の疾患（陽性所見）と，それ以外の所見（陰性所見）に分類できる．これらを区別する必要がある。以降、本稿ではこの事象の分類を**P/N分類**と呼ぶ．提案手法は，以下の２つの重要な違いがある．
- **文字ベースの処理**
- **2つの処理（事象認識とP/N分類）の融合**

以下がそれぞれの処理の内容である．

(1) **文字ベースの処理**：従来のNLPの研究では，単語（形態素）を最小単位とみなすものが多い．しかし，医療テキストは，長い複雑な複合名詞（「傍大動脈リンパ節郭清」など）や一般には用いられない専門的なひらがな名詞（例えば，「びまん性」）が多く出現し，しばしば形態素解析の誤りにつなが

る．したがって，本研究では，形態素解析処理を行わない文字ベースの方法を採用する．

(2) **２つの処理の融合**：通常，事象のP/N分類タスクは，最初の固有表現認識ステップの後に適用される．しかし，P/N分類に必要な情報は固有表現認識の情報と重複する部分も多い．例えば，「～が認められる」「～が認められない」は，ともに病名出現の大きな手がかりであるとともに，P/N分類の手がかりにもなる． このため，本研究では，固有表現認識とP/N分類の2タスクを1つに融合する．

B.研究方法

B-1.コーパス

本研究では,NTCIR の共有タスクデータと互換性のある医療テキストデータセットを用いる．これらのデータでは，症状および診断に関係する表現は，“<p> </ p>”（陽性の場合．以降，<p>タグと呼ぶ）または“<n> </ n>”（陰性の場合．以降，<n>タグと呼ぶ）でマークされている．本研究では，468 症例（計 9,286 文）の日本語テキストを用いた．

B-2.文字ベースのラベリング

本研究で提案する MedEx/Jの入力は文であり,出力は<p>タグや<n>タグが付いた文である．この出力は，(1) ERと (2) P/N分類の2つの処理の結果であり，処理(1) の結果は，タグ付けされた範囲として表現され，処理(2) の結果は，タグのタイプ(<p>タグまたは<n>タグ)として表される．

ERは，一般の固有表現認識の手法にならい入力テキストシーケンス上で開始 (B)，内側 (I) および外側 (O) という3種類のラベル付与タスクとして実現する．ただし，ラベルの付与単位は，文字ベースである．図1に提案手法（文字ベース）と既存手法（単語ベース）の系列ラベリングの違いを示す．

通常，単語ベースのアプローチでは，品詞や原型などの単語単位の情報を特徴として使用しているが，文字ベースのアプローチではそれらの特徴を使用できない．そのかわりに，ひらがなやカタカナ，漢字といった文字種の情報を用いる．

図2は，文字ベースのアプローチ（図2(a)）と単語ベースのアプローチ（図2(b)）にて，それぞれ使用される素性関数を生成するためのテンプレートを示している．文字ベースのCRFテンプレ

ートにある「Char」と「Char type」はそれぞれ表

(a) 入力テキスト

腫瘤は肝細胞癌ではなく肝の孤立性形質細胞腫と診断された.

(b) 文字ベース系列ラベリング

腫	瘤	は	肝	細	胞	癌	で	は	な	く	孤	立	性	形	質	細	胞	腫	と	診	断	さ	れ	た
O	O	O	B-N	I-N	I-N	I-N	O	O	O	O	B-P	I-P	I-P	I-P	I-P	I-P	I-P	I-P	O	O	O	O	O	O

(c) 単語ベース系列ラベリング

腫瘤	は	肝	細胞	癌	で	は	なく	孤立	性	形質	細胞腫	と	診断	さ	れ	た
O	O	B-N	I-N	I-N	O	O	O	B-P	I-P	I-P	I-P	O	O	O	O	O

(d) タグ付き出力テキスト

腫瘤は<N>肝細胞癌</N>ではなく肝の<P>孤立性形質細胞腫</P>と診断された.

図1: 系列ラベリングにおけるシーケンス表現

面文字と文字タイプを表す．文字タイプは漢字，ひらがな，カタカナ，英数字である．「N-gram」は，ユニグラム，バイグラムまたはトライグラムのうちのどのnグラムが使用されるか指定する．各設定のウィンドウサイズは，予備実験にて調整した．

（倫理面への配慮）

本研究については以下の課題名で，奈良先端科学大学院大学情報学系の倫理審査に申請し，申請が受理されている．

公開用課題名:

電子的診療録の自動構造化を有した自然言語処理解析装置の研究開発

受付番号　2016-I-30

Feature Type	**N-gram**	**Window Size (unit : character)**
Char	1, 2-gram	-2,-1,0,1,2,3,4,5
Char Type	1, 2-gram	-2,-1,0,1,2,3,4,5

(a) **文字ベース素性**

Feature Type	**N-gram**	**Window Size (unit : word)**
Word	1,2,3-gram	-2,-1,0,1,2
Word[-1]	1,2,3-gram	-2,-1,0,1,2
Word[-2:-1]	1,2,3-gram	-2,-1,0,1,2
Yomi	1,2-gram	-2,-1,0,1,2
POS	1,2-gram	-2,-1,0,1,2

(b) **単語ベース素性**

単語ベース CRF では 5 種類の特徴を使用する．表中の, 'Word', 'Word[-1]' and 'Word[-2:-1]' **はそれぞれ表層単語，前単語，前 2 単語を示す．**'Yomi'**は表層単語のひらがな読みで** 'POS' **は表層単語の形態素を示す．**

図 2: 素性関数テンプレート

C.研究結果

C-1. コーパスマテリアルと設定

構築されたCRFのモデルを評価するために，B. 研究方法で述べたコーパスとは別に．<p>タグと<n>タグを付与した日本語退院サマリ 500症例（計10,266文）を用いた．モデルの学習の際に必要となる パラメータにはデフォルト値を用い

た．単語ベースと文字ベースの2つの手法の性能を調査するため，両手法を実装した

評価方法

評価は，(1) ERおよび (2) ER + P/N分類（融合タスク）の2種類の方法で行った．ERは事象抽出のみを行う処理であり，P/N分類は，陽性または陰性の2つのタイプのモダリティを判定する処理である．

C-2. 評価指標

性能は，既存研究の手法に基づいて，適合率，再現率およびF-尺度（β=1）を用いて評価した．抽出された事象が教師データファイル内の対応する正解の事象と完全に一致する場合にのみ，正しいとみなした．評価方法は，CoNLL-2000共有タスクで使用したものと同じである．評価に使用したPerlスクリプトは，CoNLL-2000のWebサイト[1]から入手できる．

表1は，文字ベースおよび単語ベースの方法でのERの性能評価の結果である．提案した文字ベースの方法は，3つの評価指標のすべてにおいて単語ベースの方法より優れていた．

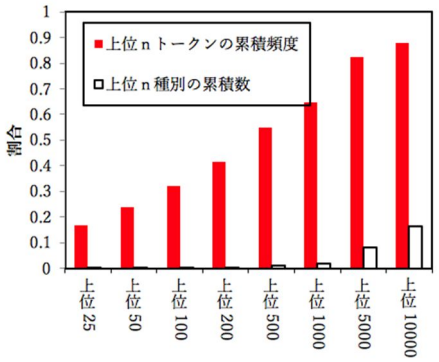
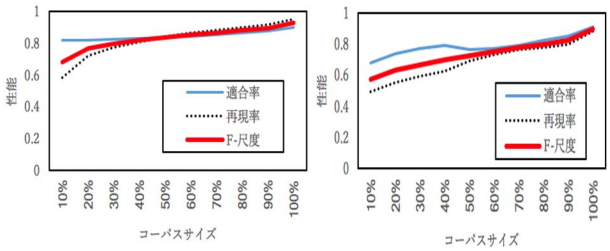


図4: 頻度が上位25から10000位までの陽性事象のトークン（赤）とその種別（白）のカバー率

表2は，ER+P/N分類融合法の性能評価の結果である．この処理においても，文字ベースの方法がすべての指標で単語ベースの方法よりも優れていた．ここで，<N>タグの精度は，<P>タグのそれと比較して低く，P/N分類がむしろ困難な課題であることを示唆している．P/N分類は時に，アノテーションの際にも難しい場合があり，この結果は妥当である．

[1] http://www.cnts.ua.ac.be/conll2000/chunking/



(a) 陽性事象<P>タグ　(b) 陰性事象<N>タグ

図 3: 性能とコーパスサイズの関係

D.考察

表 1: ER 性能

	適合率	再現率	*F*-尺度
文字ベース	0.912	0.954	0.933
単語ベース	0.854	0.827	0.841

表 2: ER + P/N 分類融合

	P/N	適合率	再現率	*F*-尺度
文字ベース	<P>tag	0.902	0.952	0.926
	<N>tag	0.908	0.884	0.896
単語ベース	<P>tag	0.820	0.810	0.815
	<N>tag	0.724	0.603	0.659

本章では，コーパスサイズの性能への効果，事象のカバー率，処理時間および今後の課題について考察する．

D-1.コーパスサイズの効果

学習に必要なデータ数を確認することが重要である．これを調べるために，文字ベースでのER+P/N分類融合法の性能を，コーパスサイズを10%ステップで10%から100%に変更しながら分析した．図3(a)と図3(b)に，<p>タグと<n>タグの抽出性能の変化をそれぞれ示す．コーパスサイズが大きくなるにつれて，それぞれの性能が徐々に向上しているのが分かる．この図より再現率は，適合率よりもコーパスのサイズの影響を受けやすい傾向があると言える．また，コーパス全体を使用しても性能が飽和状態に達しないこと，従って，コーパスのサイズを増やすことによってさらに改善の余地があることを確認できる．

D-2. 事象のカバー率：トークンレベルでの分析

トークンレベルでの分析として，頻度上位n位までの事象の種別およびそのトークンのカバー率を調査した．カバー率とは，コーパス中の全てのトークンまたは事象の種別の頻度の累計における，上位ｎ位までのトークンまたは事象の種別の頻度の累計により求められる割合である．図4より，上位500位までの事象がトークンの約50%を占める．一方，事象の種別に関しては上位10,000位までを対象としても全体の2割弱のカバー率であった．つまり，典型的なロングテールの形を示しており，低頻度の多数の種類の事象が医療テキストに現れることを示している．

D-3. 処理時間

処理速度は実用的なシステムを構築するうえで，重要な指標である．実際に，大学病院などの大規模な病院では毎日約3,000人もの患者が診療を受け，およそ6万もの文書が生成されている．

処理時間は以下のスペックを有する計算機を用いて調査した．

- CPU: コアi7 6800K 6core/12thread 3.4GHz（ターボブースト3.6GHz）
- メモリ：32GB (8GB×4) DDR4-2133クワッドチャネル

結果として，医療テキスト1,000件の処理時間は，モデルの学習処理では2分44秒，分析処理では2秒であった．これは，1日で生成される全ての文書に対する処理時間が120 (= 2×60,000/1,000) 秒あれば十分であることを示している．ほとんどの病院にとって，これは実際に実現可能な処理時間と考えられる．

E.結論

本研究では (1) 事象抽出と (2) P/N分類の2つの処理からなる日本語による医療テキスト解析ツールMedEx/Jを開発した．提案した文字ベースのアプローチは，2つの処理を1つの系列ラベリングとして実行する．このアプローチには2つの重要な利点がある．1) 形態素解析が不要なため，事前処理が簡素化された点と，2) 単語ベースの方法よりもはるかに単純な特徴セットを使用しているが，単語ベースの方法よりも性能が優れている点（ERタスク（$F_{\beta=1}$ = 0.93)，および，ER+P/N分類タスク（$F_{\beta=1}$= 0.91)）である．

得られた結果は，提案したアプローチが医療テキストの解析できわめて有効であることを示唆している．

F.健康危険情報

該当なし

G.研究発表

１．論文発表

- E.Aramaki, K.Yano, S.Wakamiya: MedEx/J: A One-scan Simple and Fast NLP tool for Japanese Clinical Texts, MedInfo, 2017.（採択済）

２．学会発表

- 矢野憲，伊藤薫，若宮翔子，荒牧英治: 深層学習による医療テキストからの固有表現抽出器の開発とその性能評価： 人工知能学会全国大会（JSAI）, 2017. （査読なし）

- 矢野憲，若宮翔子，荒牧英治： 医療テキスト解析のための事実性判定と融合した固有表現認識器，言語処理学会年次大会，2017.（査読なし）

H.知的財産権の出願・登録情報

該当なし

平成 28 年度厚生労働科学研究費補助金
政策科学総合研究事業(臨床研究等 ICT 基盤構築研究事業)
分担研究報告書

カルテ文章からの病名自動抽出に関する研究

研究分担者：河添悦昌 東京大学医学部附属病院 企画情報運営部

A．研究目的

東京大学医学部附属病院の退院サマリから本研究の解析対象となる病名情報を抽出する。

B．研究方法

B-1. 2004 年 1 月 1 日から 2016 年 12 月 31 日の期間を対象として、東京大学医学部附属病院において記載された退院サマリを抽出した。退院サマリの個人識別情報を削除し、「年齢」、「性別」、「担当診療科」、「退院時診断（文字列、ICD-10 コード、複数可）」、「サマリ本文に含まれる症状・所見・疾患の文字列」の項目からなる XML データを作成した。症状・所見・疾患の文字列の抽出に際しては奈良先端大荒牧研究室で開発した病名抽出ツールを用いて行った。

B-2.研究の実施に際しては東京大学大学院医学系研究科の倫理承認を得て行った。

承認番号：11446

URL:http://www.m.u-tokyo.ac.jp/medinfo/wp-content/uploads/2013/08/ethics-20170208.pdf

C．研究結果

対象となった退院サマリは 291,642 件であった。退院サマリからデータが正確に抽出されているかを検証する目的で、抽出項目をサマリが記載された年ごとにまとめた（表１）。2004 年はシステムの移行期にあったため、退院サマリの件数が少なかった。その他、サマリの記載年で抽出した項目に大きな違いはなかった。

D．考察

データ抽出過程のため特になし。

E．結論

データ抽出過程のため特になし。

F．健康危険情報

データ抽出過程のため特になし。

G．研究発表

データ抽出過程のため特になし。

H. 知的財産権の出願・登録情報

該当なし。

表1 抽出した退院サマリの要約

記載年	件数	年齢中央値	男性の割合	診断病名数（重複あり）		本文中病名数（重複あり）	
				全件数	サマリあたり件数	全件数	サマリあたり件数
2004	3402	59	46.7%	8659	2.5	68020	20.0
2005	15927	59	50.9%	40884	2.6	354949	22.3
2006	17141	62	55.3%	41903	2.4	442117	25.8
2007	19193	61	54.6%	45707	2.4	528254	27.5
2008	19139	61	54.0%	43944	2.3	557520	29.1
2009	24371	61	51.0%	55598	2.3	581559	23.9
2010	26889	61	49.4%	61038	2.3	601114	22.4
2011	27043	61	49.3%	64513	2.4	543237	20.1
2012	28190	62	49.1%	69508	2.5	664561	23.6
2013	28287	62	48.9%	69936	2.5	644391	22.8
2014	28299	62	49.2%	67971	2.4	605666	21.4
2015	26716	61	48.9%	61566	2.3	597685	22.4
2016	27044	62	49.5%	61851	2.3	656598	24.3
平均	22434	61	50.5%	53314	2.4	526590	23.5

別添5

研究成果の刊行に関する一覧表

論文

発表者氏名	論文タイトル名	発表誌名	巻号	ページ	出版年
E.Aramaki, K. Yano, S.Wakamiya	MedEx/J: A One-scan Simple and Fast NLP tool for Japanese Clinical Texts	MedInfo	In press	In press	2017

学会発表

発表者氏名	論文タイトル名	発表会議名	発表日時
矢野憲, 伊藤薫, 若宮翔子，荒牧英治	深層学習による医療テキストからの固有表現抽出器の開発とその性能評価	人工知能学会全国大会	2017/05/23
矢野憲，若宮翔子，荒牧英治	医療テキスト解析のための事実性判定と融合した固有表現認識器	言語処理学会年次大会	2017/03/14