様　式　Ｃ－１９、Ｆ－１９、Ｚ－１９　（共通）

科学研究費助成事業　　研究成果報告書

平成 ２７ 年　 ５ 月 ２０ 日現在

機関番号：１２５０１
研究種目：基盤研究(B)
研究期間：2012～2014
課題番号：２４３０００９８

研究課題名（和文）多施設間の統合退院サマリーデータベースの構築

研究課題名（英文）Construction of the integrated multicentre discharge summary database

研究代表者
鈴木　隆弘（Suzuki, Takahiro）

千葉大学・医学部附属病院・准教授

研究者番号：４０３２３４２２

交付決定額（研究期間全体）：（直接経費）　　12,800,000円

研究成果の概要（和文）：本研究では複数の医療機関から退院サマリーを抽出し、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持つ大型テキストデータベースの構築を進めてきた。これまでに7つの病院から退院サマリーを抽出した。データの収集と基本的処理は各施設内で行い、連結不可能匿名化した情報のみを集約している。DPCを指標としたクロスマッチ自動判定の結果では、自施設のモデルデータでは高い一致率を示し、他施設のモデルデータでは10～20%程度低下するものの、全施設統合データでは自施設と同程度の値を示した。また、これまでに構築したデータベースを用いて類似症例検索を行うWEBアプリケーションを開発して千葉大学に設置した。

研究成果の概要（英文）：We performed the multi-year project to collect discharge summary from multiple hospitals and made the big text database to built a common document vector space, and developed various application. We extracted 243,907 discharge summaries from seven hospitals (Chiba University hospital, Nagasaki University hospital, Kagawa University hospital, Osaka University hospital, Kochi University hospital, Saga University hospital and St. Luke's International Hospital). There was a difference in term structure and number of terms between the hospitals, however the differences by disease were similar. We built the vector space using TF-IDF method. We performed cross-match analysis of DPC selection among seven hospitals. About 80% cases were correctly matched. The use of model data of other hospitals reduced selection rate to around 10%, however, integrated model data from all hospitals restore the selection rate.

研究分野：医療情報学

キーワード：退院時サマリー　テキストマイニング　多施設共同研究

１．研究開始当初の背景

近年の日本は電子カルテシステムの普及期を迎え、医療情報は電子化された形で蓄積されつつある。千葉大学医学部附属病院においても、電子カルテシステムが 2003 年 6 月より本稼動を始め、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基に、統計手法を用いて新たな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類は先送りされてきた。理由としては業務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。また、個人情報保護法の制約のため、複数施設間の情報を集約した研究は進んでいなかった。

我々は早くから診療文書の解析に取り組んできた。しかし、対象が少数の施設に限られていたために普遍性が課題となっていた。我々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、本手法を用いることで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。この成果をさらに発展させて複数の施設の文書を統合した巨大データベースによるテキストマイニングが可能となれば、より一般的な結果を得られ、精度の向上が期待できるだけでなく、応用範囲も広がると期待された。

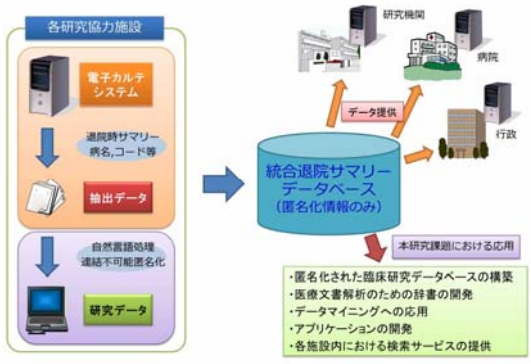

図１　研究の概要

２．研究の目的

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様々な応用を試行し、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

３．研究の方法

(1)対象

千葉大学附属病院、長崎大学附属病院、佐賀大学附属病院、高知大学附属病院、香川大学附属病院、大阪大学附属病院、聖路加国際病院の退院サマリーから、2009-2011 年を対象とし、現病歴、既往歴、家族歴、病名の外部ファイルへの出力を行い、併せて DPC（Diagnosis Procedure Combination）の 14 桁コードを出力した。DPC 自動判定には、14 桁コード毎の症例数が 20 例以上ある DPC を用い、症例をベクトル空間構築用と自動判定の検証用にランダムに振り分けた。

対象のサマリー数は千葉大学 42690 例、大阪大学 42867 例、香川大学 40076 例、高知大学 13976 例、長崎大学 41466 例、佐賀大学 28753 例、聖路加国際病院 37483 例で、総計は 247311 例であった。患者個人情報は本研究には不要なので、出力の際に出来る限り取り除いて連結不可能匿名化を行った。これらのサマリーから重要語を抽出してテキストデータベースを作成した。サマリー本文の情報は１入院に対し１つのテキストファイルを作成し、退院時所見・入院後経過を格納して、ファイル名は入院毎の識別番号とした。属性ファイルは病院毎に１つ作成して、全ての入院分のデータを格納した。各レコードが１入院のデータを表し、病名のように複数のデータが存在する場合は各データを排他的な記号で区切った。

全ての病院で、院外での退院サマリーの使用について利用許可が得られなかったため、データの収集と基本的な処理を各施設内で行い、形態素分解後の元の文章が復元不可能な情報のみを集約した。

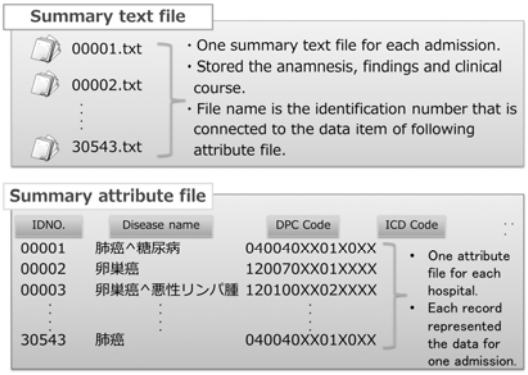

図２　データファイル書式

(2)形態素解析と辞書の整備

本研究では、形態素解析に MeCab を使用した。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。標準的な辞書だけでは医療用語が過剰に小さく分解されてしまうため、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、術式マスターなどから作成した高精度の医療用語辞書を用いて、医療用語が分解されずに索引語として単語認定できるようにした。加えて我々が作成に協力したパラメディカル用医療辞書である ComeJisyo を併用した。

(3)ベクトル空間モデル

抽出したサマリーは症例検索の場合は単独で、DPC 自動判定の場合は各々のコードごとにマージして解析を行った。得られた索引語の中には、疾患・病態と密接に関係したものから、関係の薄いものまで存在する。そこで、その索引語が病態の特徴を表す上でどれだけの重要度を持っているか示す為に、索引語の重み付けの処理が必要となる。本研究では重み付けにシンプルで処理速度が速いことから、索引語の文書集合内での出現頻度と文書集合間の出現頻度を用いる Tf×Idf 法を用いた。得られたデータからベクトル空間が構築され、検索対象のサマリーベクトル空間と入力されたサマリーを比較し、最も類似したベクトルを検索する。類似度の算出はベクトルの内積によって求める。これらの手法により、ＤＰＣの自動判定や類似症例検索などが可能となる。

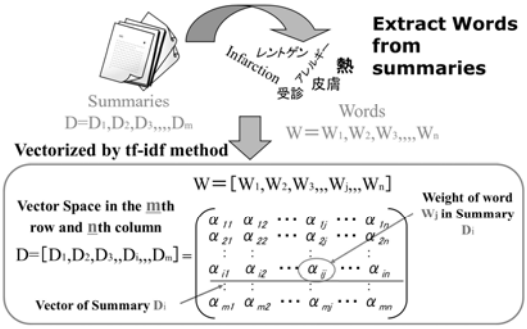

図３　ベクトル空間モデル

(4)施設間の違いの検証

研究参加病院間の診療の差について、サマリーの長さ、使われている単語、診断およびＤＰＣとの関連等について比較を行い、サマリーの違いから検証した。多施設のデータを集積することによって疾患としての全体像を提供し、それと自院との差異を明らかにして自らの特徴を知ることで、医療の品質管理や研修医の指導にも活用できる。

(5)DPC 自動判定

症例数が 20 以上の DPC コードを抽出し、7 対 3 の比率で無作為にモデル作成用と検証用のサマリーに分け、検索対象のサマリーベクトル空間と入力されたサマリーを比較して最も類似したベクトルを検索し、DPC を判定した。類似度の算出はベクトルの内積によって求めた。参加７病院各々の検証用データとモデルデータの組み合わせによる７対８のクロスマッチ判定および、全病院のデータを統合したモデルデータとの間での自動判定を、DPC コードの 14 桁全てと上位６桁のそれぞれで行った。

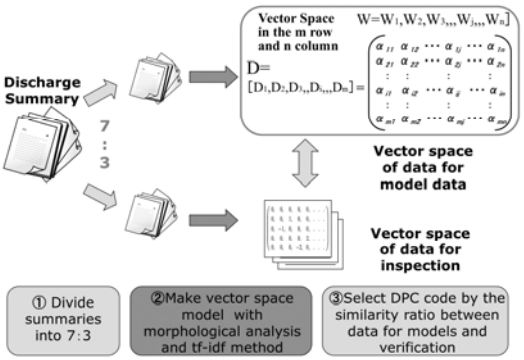

図４　DPC 自動判定の実験手順

(6)応用アプリケーション

本研究で得られたデータベースの応用として、インターネットから類似症例検索が出来る Web アプリケーションと DPC 判定をバッチ処理で行うアプリケーションを作成した。

Web アプリケーションはM言語データベースの Cache に装備されている CSP（Cache Server Page）機能を用いて開発した。検索ボックスに病歴をそのまま入力し、各施設別と全国共通のベクトルを切替えて類似症例を検索することができる機能を備えている。サーバは千葉大学に設置し、インターネットからも利用可能とした。

DPC 判定アプリケーションはリレーショナルデータベースを用いて作成した。

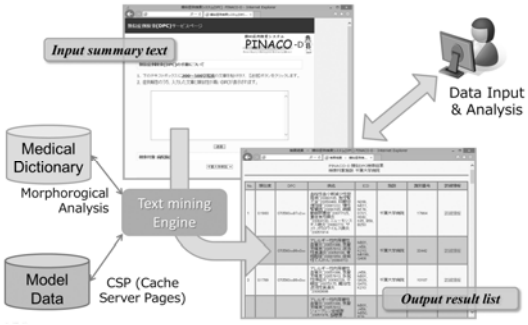

図５　類似症例検索アプリケーション

４．研究成果

(1)施設別比較結果

索引語の総出現数は 27220461 回で、単語数はそれぞれ、千葉大学 63441 語、大阪大学 88750 語、香川大学 75757 語、高知大学 50145 語、長崎大学は 78490 語で全体では 185378 語であった。平均単語数は千葉大学の 160 語から大阪大学の 383 語まで大きな違いが認められた。いずれの大学でも極端に短いサマリーが存在した。図６にサマリーの長さの分布を示す。1 つのサマリーに含まれる索引単語

数による比較では、長崎大学は 200 語台がピークの正規分布を示していたのに対して、千葉大学は 200 語台と 50 語以下の 2 つのピークを持ち、全体に短いものが多かった。大阪大学と香川大学ではピークが 300 語台で 200 語未満のものは殆ど無かった。

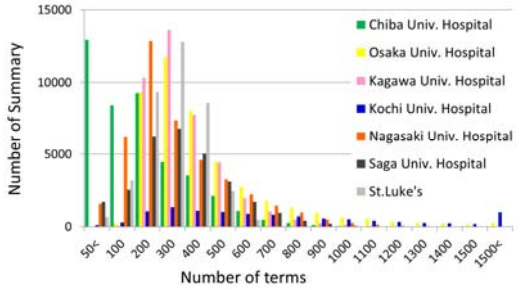

図６　単語数別サマリー数

次に疾患とサマリーの長さとの関係を検討した。サマリー１件あたりの索引語数を MDC(Major Diagnostic Category)毎に平均し、施設毎の平均単語数との比率で比較した。どの病院も眼科疾患、耳鼻科疾患は短く、精神科疾患、呼吸器疾患では長い傾向が認められた。

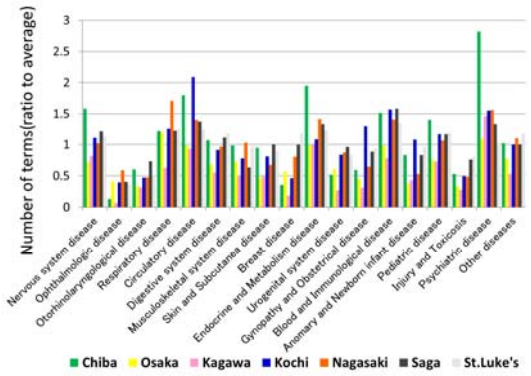

図７　疾患別単語数の平均との比較

(2)DPC 自動判定

図８から１１に、同一施設のモデルデータと検証データの組み合わせでのＤＰＣ判定率を主要疾患群別に示した。

千葉大学では乳房の疾患で 100%正答を示した一方、「その他の疾患」の判定率は低かった。

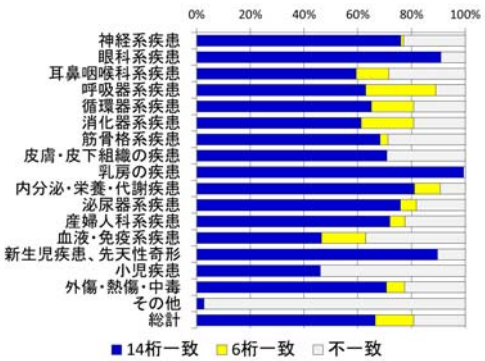

図８　DPC 判定結果−千葉大学

大阪大学では乳房の疾患、泌尿器科疾患、眼科疾患の正答率が高く、小児疾患の正答率が低かった。

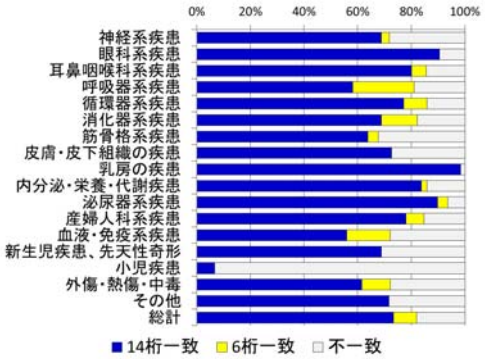

図９　DPC 判定結果−大阪大学

香川大学では眼科疾患が 100%正答を示し、乳房の疾患や新生児疾患の正答率も高く、皮膚科疾患、「その他の疾患」の正答率が低かった。

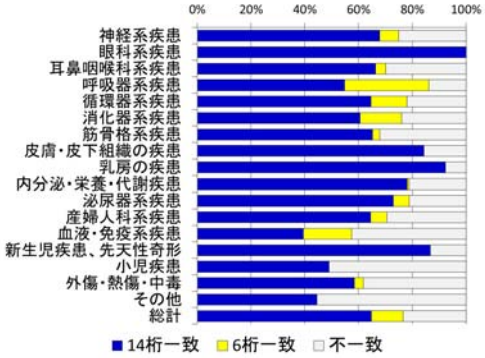

図 10　DPC 判定結果−香川大学

長崎大学では眼科疾患、乳房の疾患、新生児疾患の正答率がほぼ 100%と高く、千葉大と同様に「その他の疾患」の正答率は非常に低かった。

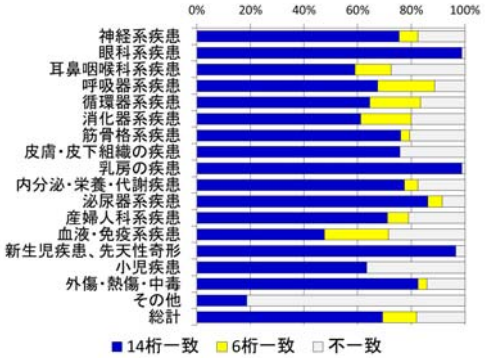

図 11　DPC 判定結果−長崎大学

(3)DPC クロスマッチ判定

疾患を表す DPC6 桁までを用いたクロスマッチ照合の結果を表１に示す。いずれの病院でも自施設のモデルデータと検証データの組み合わせでは、78～85%が一致と高い判定率を示した。異施設との間では 56～82%の判定率に低下するものの、全施設を統合したモ

デルデータとの検証では自施設と同等の判定率を示した。

Model / Verification	Chiba Univ. Hospital	Kagawa Univ. Hospital	Kochi Univ. Hospital	Nagasaki Univ. Hospital	Osaka Univ. Hospital	Saga Univ. Hospital	St. Lukes Hospital	Integrated Data
Chiba Univ. Hospital	83%	75%	69%	77%	73%	77%	74%	83%
Kagawa Univ. Hospital	56%	78%	53%	66%	71%	65%	63%	74%
Kochi Univ. Hospital	75%	75%	80%	80%	75%	77%	73%	81%
Nagasaki Univ. Hospital	77%	78%	72%	85%	79%	82%	78%	85%
Osaka Univ. Hospital	72%	75%	69%	76%	84%	77%	76%	81%
Saga Univ. Hospital	72%	73%	66%	78%	74%	81%	78%	80%
St. Lukes Hospital	73%	72%	64%	77%	73%	80%	81%	77%

表１　DPC６桁によるクロスマッチ結果

表２にDPC14桁全て一致での判定結果を示す。判定率は自施設同士で 58～75%、異施設との間でも 49～66%と 6 桁よりは下がるものの高率を示した。統合モデルデータとの照合では、6 桁の場合と同様に自施設と同等の判定率を示した。

Model / Verification	Chiba Univ. Hospital	Kagawa Univ. Hospital	Kochi Univ. Hospital	Nagasaki Univ. Hospital	Osaka Univ. Hospital	Saga Univ. Hospital	St. Lukes Hospital	Integrated Data
Chiba Univ. Hospital	69%	50%	55%	59%	53%	56%	56%	69%
Kagawa Univ. Hospital	52%	58%	50%	55%	52%	53%	50%	51%
Kochi Univ. Hospital	49%	42%	69%	54%	51%	51%	51%	65%
Nagasaki Univ. Hospital	60%	54%	57%	71%	55%	62%	58%	72%
Osaka Univ. Hospital	56%	53%	56%	60%	75%	58%	60%	72%
Saga Univ. Hospital	56%	50%	54%	61%	52%	67%	61%	66%
St. Lukes Hospital	61%	55%	59%	65%	59%	66%	71%	67%

表２　DPC14 桁によるクロスマッチ結果

(4)応用アプリケーション

図１２に類似症例検索 Web アプリケーションの検索例を示す。レスポンスは入力された文章が 200 文字以下の場合で約 5 秒以内、1000 文字以上でも９秒以内で、平均は 6.7 秒と実用上充分な速度が得られた。精度はオフラインでのクロスマッチテストの結果と同等であった。

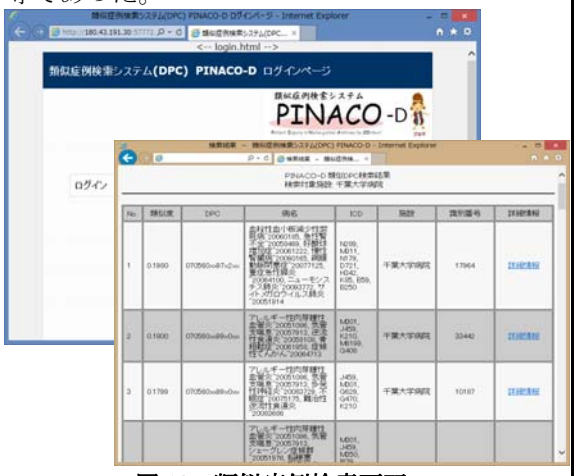

図 12　類似症例検索画面

(5)考察

各施設の退院サマリーの属性、スタイルはかなり異なっており、特に文章の長さでは２倍以上の差が見られた。にもかかわらず 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る。

クロスマッチ DPC 自動判定の結果では、新たな３施設でも以前の報告と同様に自施設のモデルデータでは高い判定率を示した。他施設のモデルデータを用いた検証では 10～20%程度の判定率低下が認められたが、全施設統合モデルデータでは自施設と同程度に復しており、これらからも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る。このデータベースを用いることで、マイニング精度の向上が期待できるだけでなく、応用範囲も広がることが期待できる。応用の一つである Web 版の類似症例検索アプリケーションは十分な速度と精度を示した。

今回使用した TF-IDF 法は、シンプルで計算速度の速い方法であるが、文法的な要素は考慮されておらず、精度を更に向上させる余地がある。もちろん文法は各国の言語に依存しているので、日本語だけに通用する手法となってしまうが、日本語文法を考慮したテキストマイニングの手法がいくつか提案されている。それらを医療文書に応用した研究もいくつかなされており、成果も上がっている。しかし、それらは一施設内の実験で有り、文書の種別も限られている。我々の大型テキストデータベースとそれらの文法的な手法が組み合わされば、更なる発展が期待できる。

今後は今回開発したアプリケーションをパッケージ化して参加施設に配布して、臨床での応用を行っていく予定である。

本研究の成果は日本医療情報学会（2012, 2013, 2014）および国際医療情報学会（2013）にて発表を行った。2015 年開催の国際医療情報学会(Medinfo2015 ブラジル)にも Accept されており、発表を行う予定である。

＜引用文献＞

① Takabayashi K, Doi S, Suzuki T. Japanese EMRs and IT in Medicine: Expansion, Integration, and Reuse of Data. Health Inform Res. 2011; 17:178-83.

② Takabayashi K, Ho TB, Yokoi H, Nguyen TD, Kawasaki S, Le SQ, Suzuki T, Yokosuka O. Temporal abstraction and data mining with visualization of laboratory data. Stud Health Technol Inform. 2007; 129:1304-8.

③ Ono H, Takabayashi K, Suzuki T, Yokoi H, Imiya A, Satomura Y. Extraction of diagnosis related terminological information from discharge summary. Medinfo. 2004; 1786.

④ Suzuki T, Yokoi H, Fujita S, Takabayashi K. Discharge Summaries can be diagnosed from

extracted index terms by text mining. Medinfo. 2007; 2257-2259.
⑤ Suzuki T, Yokoi H, Fujita S, Takabayashi K. DPC Code Selection from Electronic Medical Record -Text Mining Trial of Discharge Summary-. Methods Inf Med 2008; 47:541-548
⑥ Suzuki T, Doi S, Shimada G, Takasaki M, Tamura T, Fujita S, Takabayashi K. Auto-selection of DRG codes from discharge summaries by text mining in several hospitals: analysis of difference of discharge summaries. Stud Health Technol Inform. 2010; 160:1020-4.
⑦ Kudo T, Yamamoto K, Matsumoto Y. Applying conditional random fields to Japanese morphological analysis. Proc. EMNLP, 2004: 230–237

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔雑誌論文〕（計１件）
①Suzuki T, Doi S, Fujita S, Hatakeyama Y, Honda M, Matsumura Y, Shimada G, Takasaki M, Tsumoto S, Yokoi H, Takabayashi K. Construction of the integrated multicentre discharge summary database. Studies in Health Technology and Informatics. 査読有、Vol.192, 2013, pp.1064, http://www.ncbi.nlm.nih.gov/pubmed/23920838

〔学会発表〕（計４件）
①鈴木隆弘，土井俊祐，嶋田元，　高崎光浩，津本周作，畠山豊，本多正幸，松村泰志，横井英人，高林克日己．多施設データを集約した退院サマリー検索システムの構築．第 34 回医療情報学連合大会、2014/11/7、幕張メッセ（千葉県・千葉市）

②Suzuki T, Doi S, Fujita S, Hatakeyama Y, Honda M, Matsumura Y, Shimada G, Takasaki M, Tsumoto S, Yokoi H, Takabayashi K. Construction of the integrated multicentre discharge summary database. Medinfo2013, 2013/8/21, Bella Center (Copenhagen, Denmark)

③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，松村泰史，横井英人，高林克日己．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．第 33 回医療情報学連合大会、2013/11/23、神戸ファッションマート（兵庫県・神戸市）

④鈴木隆弘，土井俊祐，藤田伸輔，本多正幸，津本周作，横井英人，松村泰史，高崎光浩，嶋田元，高林克日己．多施設間の統合退院サマリデータベースの構築.第 32 回医療情報学連合大会、2012/11/15、朱鷺メッセ（新潟県・新潟市）

〔図書〕（計０件）

〔産業財産権〕
○出願状況（計０件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
http://180.43.191.30/csp/chiba/login.html
類似症例検索システム(DPC)　PINACO-D

６．研究組織
(1)研究代表者
鈴木　隆弘（SUZUKI, Takahiro）
千葉大学・医学部附属病院・准教授
研究番号：40323422

(2)研究分担者
なし

(3)連携研究者
土井　俊祐（DOI, Shunsuke）
千葉大学・医学部附属病院・助教
研究番号：90639072

高崎　光浩（TAKASAKI, Mitsuhiro）
佐賀大学・医学部附属病院・准教授
研究番号：70236206

津本　周作（TSUMOTO, Shusaku）
島根大学・医学部附属病院・教授
研究番号：10251555

畠山　豊（HATAKEYAMA, Yutaka）
高知大学・教育研究部医療学系・准教授
研究番号：00376956

本多　正幸(HONDA, Masayuki)
長崎大学・医歯（薬）学総合研究科・教授
研究番号：10143306

松村　泰志（MATSUMURA, Yasushi）
大阪大学・医学部附属病院・教授
研究番号：90252642

横井　英人（YOKOI, Hideto）
香川大学・医学部附属病院・教授
研究番号：50403788

高林　克日己（TAKABAYASHI, Katsuhiko）
千葉大学・医学部附属病院・教授
研究番号：90188079