目　　次

目　　次

厚生労働科学研究補助金

（臨床研究等 ICT 基盤構築・人工知能研究事業）総括研究報告書

電子カルテ情報をセマンティクス（意味・内容）の標準化により分析可能なデータに変換するための研究

堀口　裕正 国立病院機構本部総合研究センター　診療情報分析部　副部長
岡田 千春　国立病院機構本部　　企画役
狩野 芳伸　静岡大学情報学部行動情報学科　准教授
森田 瑞樹　岡山大学大学院医歯薬学総合研究科　准教授
奥村 貴史　国立保健医療科学院研究情報支援研究センター　特命上席主任研究官

平成 28-29 年度においては、我々が日本語における医療用自然言語処理の研究コミュニティを形成し研究に取り組んで来た標準化技術を実カルテへと適用することで、カルテからの情報抽出の自動化に向けた予備的な検証を行うことを計画した。
研究代表者堀口及び分担研究者岡田は、国立病院機構本部との調整を中心とした基盤構築を行った。まず、NDCA データの研究利用に向け、倫理審査申請に加えて、内部規定にて定められている内部委員会の調整を図った。また、閲覧・解析に特化した自然言語処理用の研究基盤の構築を行った。研究基盤の概念図は図に示したとおりで、セキュリティを維持しつつ、空間的制約をなるべく少なく研究が進められるようなものになっている。
研究分担奥村は、臨床的なニーズを自然言語処理における個別技術へと橋渡しする役割を担った。具体的には、臨床医側より退院サマリの自動生成に求められる要件定義を進めるとともに、先行研究の整理を行い今後の研究アプローチの策定を行った。さらに、今後の研究に役立てられる入院カルテ・退院サマリの高品質な個人情報を含まない模擬のデータセットを構築した。
研究分担狩野・森田は、上記の実データ・テスト用データ双方を活用し、自然言語処理の医療テキストへの適用を進めた。狩野は、時系列で蓄積してく入院カルテデータを対象として、既存の自然言語処理ツールによる処理性能と今後の改良に向けた課題抽出を図った。森田は、医師が要約した退院サマリデータを対象として、医師の記載する退院サマリの定量的・定性的な特徴の把握を図った。この知見は、今後、入院カルテの自動要約技術の研究に際した精度管理に役立てられる。
平成３０年度については研究の最終年度として以下のテーマに取り組んだ。
研究代表者堀口及び分担研究者岡田は、他のため研究者の分析に資するため、国立病院機構本部との調整を中心としたセキュリティを維持しつつ、空間的制約をなるべく少なく研究が進められるような基盤構築・運用を行った。
また研究代表者堀口は、本研究の当初の目的である、汎用的にどの電子カルテからで

も利用できるシステムの構築という課題に対して、JAHIS の HL７CDA 規約に基づくデータを SS―MIX 拡張ストレージに保管するという現在日本における標準的なデータ交換フォーマットを起点として、特定の患者の医師記録を抽出し、そこからサマリを作成するフローを実現するモジュールの開発を行った。SS-MIX ストレージを活用したデータ保管システムはすでに国内の大手ベンダー８社においては NCDA 上で実現されており、横展開可能なシステムの構築という本研究に求められたミッションの１つは達成できたと考えている。

研究分担者奥村は、入院カルテの自動要約に向けた研究基盤の整備に取り組んできた。臨床医は、入院治療をしていた患者が退院する際、それまでに記載していた入院カルテから退院サマリを作成する。この退院サマリの作成を効率化することができれば、医師の診療負担を直接軽減することが出来ることに加えて、様々な副次的な効果が期待される。

今年度は、昨年度までの研究をさらに発展させ、4 つの課題に取り組んだ。まず、一連の研究には自由に研究利用できるカルテの存在が必要となる。そこで、ダミーカルテの収集を進め、100 件の整備を目指して活動を進めた。また、収集したダミーカルテを対象として、機械的な処理を可能とするためのアノテーション作業に取り組んだ。その際、初年度に行ったアノテーション、昨年度に試みたアノテーションを踏まえ、さらなる改善を図った。次に、このアノテーション作業と平行して、退院サマリの分析モデルであり生成モデルである「CASE モデル」の改善に取り組んだ。また、退院サマリの自動生成処理を精度管理するうえで必要となる「理想の退院サマリ」の確保に向けた検討を行った。

研究活動の結果、ダミーカルテは、目標を超える 108 件を集めることが出来た。また、これらのカルテを対象としたアノテーションを進めると共に、アノテーションガイドラインを高品質化することが出来た。さらに、このアノテーション済みカルテを用いて実現する退院サマリの分析に向けたモデルとして、修正 CASE モデルを提示することが出来た。理想のサマリに向けた検討では、高品質な退院サマリを低コストに実現するための作業仮説を整理すると共に、実証に向けた研究デザインを策定することが出来た。

研究分担者森田は、退院サマリの自動生成技術の実現を目指し，昨年度までの調査結果の整理およびそれを踏まえた退院サマリの自動評価のコンセプト検証を実施した。退院サマリを生成および評価するためには，どのような退院サマリを生成しなくてはならないかを示す「理想的な退院サマリ」の定義が必要となる。昨年度に理想的な退院サマリについて言及をした国内外の文献調査を実施し，今年度はその結果を整理した。退院サマリの記載に関して「記載すべき項目」および「定性的な要件事項」を抽出し，前者として 38 項目を得た。このうちの 14 項目はカルテの構造化データより抽出できるものの，残りの 24 項目はカルテの自由記載より抽出してサマリとして文章を作成や要約をする必要があるものと考えられた。また，後者として 9 項目を得た。このうちの 5 項目は医学的な知識・経験がないと採点が難しいものの，残りの 4 項目は形式的に判断をすることが可能と考えられた。この 4 項目のうち医学用語辞書を必要としない 3 項目について文章の特徴を用いて自動評価することを試み，人による評価との相関および点数の分布を踏まえて自動評価の可能性を考察した。退院サマリの自動評価手法の確立に向けた今後の課題が明らかとなった。

研究分担者狩野は、退院サマリの自動生成に向けたテキストの分析についての研究を

行った。入院患者の退院に際し、医師は入院中に記載したカルテ等の情報から退院サマリを作成する必要がある。この退院サマリを自動的に生成することが出来れば、臨床現場の負担を下げることが出来ると共に、医療の質に貢献することが期待される。
カルテの処理にあたっては、事前に匿名化が必要となる。匿名化作業を自動化するための匿名化ツールの実装と性能向上に取り組んだ。そのために、既存の正解付き模擬カルテデータに加え、別のダミーカルテデータセットに対し匿名化のためのアノテーション付与を行い、これらを用いてルールベースおよび機械学習による匿名化ツールの実装と性能検証を行った。
サマリ生成にあたっては、対象とするカルテやサマリのドメイン、すなわち診療科や疾患により、サマリ生成に必要となる情報が異なると考えられる。サマリと対応する電子カルテの履歴データについてクラスタリングを行い、どのようなタイプのサマリやカルテがどう類似しうるかの分析を行った。

A.目的

本研究は、電子カルテに実装可能な医療用語の標準化を行うシステムの開発に向け、そのステップとして、臨床現場からのニーズが極めて大きい退院サマリの自動生成技術の実現を目標と定める。電子カルテの自動解析は技術的な難易度が高く、実用的な精度を実現するためには多額の研究開発投資が求められる。そこで、本研究提案では、医療現場に直接的なメリットが生じる研究課題に取り組むことによって、現場の協力と今後の追加的な研究開発投資を呼び込み、その過程を通じて実用性の高い電子カルテの自動解析技術を実現する戦略を採る。

初年度、我々が今まで模擬カルテを用いて研究開発を進めてきた標準化技術を、国立病院機構の有する広域電子カルテ網(NCDA)上の実カルテへと適用し、技術的な課題を抽出する。2 年目には、NCDA を用いて集積した電子カルテに加えて、退院サマリ情報を用いることで、電子カルテの自動要約技術の検討を行う。3 年目においては、両技術の統合により、継続的な精度向上の体制を実現するとともに、研究成果を既存の各社電子カルテへと組み込む枠組みを構築する。

本研究により、退院サマリの自動要約技術や紹介状の作成支援技術等、医療用の自然言語処理に関連する多彩な応用技術が実現する。これは、医療現場における負担軽減策として極めて効果が期待される。また、こうした応用の発展により、要素技術である電子カルテ上の記載からの自動情報抽出において、継続的な精度向上が実現する。この手法は、電子カルテにおける用語の標準化技術単独に研究開発投資を行うことと比して、投資効率が極めて高いと考えられる。さらに、こうして医療用自然言語処理技術が発展することにより、大量の電子カルテからの効率的な情報抽出が実現する。これは健康医療政策に資する統計データの収集コストを劇的に低廉化し、今後、政策に求められる様々なエビデンスを継続的に生み出していく基盤となることが期待される。

B.方法

本研究の今年度申請時に記載した研究計画および方法は以下のとおりであった。

【取り組む課題】

本研究では、電子カルテの記載より標準化した情報の抽出を行う技術の研究開発を行う。たとえば、カルテに「精神遅滞」と記載されている場合、この語を解釈することは機械にとって容易である。しかし、カルテの記載は単純ではない。「MR」と記載されていた場合、その語が精神遅滞(Mental Retardation)を指すのか、僧帽弁閉鎖不全症(Mitral Regurgitation)か、製薬会社の営業担当(Medical Representative)かを正確に判別することは機械にとって容易ではない。さらに、「学業成績が芳しくない」と記載されたとき、医学的には文脈よりmild MRと解する柔軟性も求められる。カルテ記載は、往々にして略記法や省略が多用される。「腹部：圧痛、反跳痛なし」という

記載から、圧痛の有無を判別することは、高度な処理を要する。医療用自然言語処理はこのように難度の高い処理であり、現時点では情報抽出の精度に限界があり、研究開発体制におけるブレークスルーが求められている。

【研究体制】

研究代表者 堀口は、「大規模に電子カルテデータを入手できる体制」として、全国の国立病院 41 施設より年間 80 万患者の電子カルテ情報を自動収集する診療情報集積基盤（NCDA）を構築し、運用している。この研究基盤を用いることにより、大学病院等のように患者の偏りが生じない多彩な施設からカルテ情報を収集することが出来る。

研究分担者 狩野・森田は、日本語カルテに記載された所見や病名を自動的に標準化するコンテスト(NTCIR MedNLP)を主催する、医療用自然言語処理における我が国の代表的な研究者である。このコンテストは、日本語カルテからの情報抽出精度を競う唯一の学術集会であり、森田は参加者としても首位の成績を収めて来た。狩野は、コンピュータによる大学入試問題の自動解答に向けた人工知能研究「ロボットは東大に入れるか」プロジェクトにおいて、社会科分野の科目担当リーダーを務めている。両名は、公募研究課題の求める「工学系人工知能研究者・自然言語処理研究者」の役割を果たす。

分担研究者 奥村は、我が国においては数少ない計算機科学分野において学位を取得した医師であり、自然言語処理の医療応用分野で継続的な貢献を果たしてきた。本研究において奥村は、「病名・治療の標準化に詳しい臨床医学系研究者」として、基礎技術研究と応用研究とを橋渡しするコーディネータ役を担う。

分担研究者 岡田は、本研究の対象となる各国立病院カルテを記載する臨床医を取りまとめ、研究開発成果の臨床的有用性の担保に取り組む。

【H28 年度実施状況】

初年度においては、上述の NTCIR MedNLP コンテストを中心に研究開発を進めてきた標準化(正規化)技術を、NCDA 上の実カルテへと適用し、予備的な検証を行うこととしていた。実際には

１，実データは利用場所が限られている等、研究利用に対して制限が多いことから、実際の診療を行っている医師の協力を得て、実際の診療経過に沿ったダミーデータを作成し、各種研究開発を加速する為の基盤を作成した。

２，実際の分析を効率的に行うため、セキュリティを考慮した分析環境の整備/構築を行った。

３，退院サマリと、電子カルテの医師記録についての関係について，現場医師から情報提供を受け予備的な検討を行った。

【H29 年度計画】

2 年目においては、NCDA を用いて集積した電子カルテに及び、退院サマリデータを対象とした電子カルテの自動要約技術の検証を行う。平成２８年度に構築した分析環境基盤を活用して、まずは、医師記録や検査

データ等入院期間内に生成された情報と、退院サマリデータの差分の検証を行い、医師がどのような思考経路を経て情報処理を行い、文書を作成しているかについての検討を行う。その際、一般的な文章を対象とした自動要約技術の応用に加えて、医療文書を対象とする正規化技術を活用した疾患や病期毎のテンプレートを用いた情報抽出を試みる。この過程は、医療従事者による詳細な手本(教師データ)を用いることで精度向上が望める。その際、平成２８年度に作成したダミーデータを活用し、本番データの分析環境と平行して各々の施設内での作業を並行して行っていくことで研究活動の効率化を図る。

【H30 年度計画】

最終年度においては、それまでに培った知見を取りまとめるとともに、上記の過程を組織化し、継続的な精度向上の体制を実現する。また、研究成果の電子カルテへの組み込みもしくは SS-MIX ストレージを活用した１部門システム化を目指し、各種ベンダーのシステムへの退院サマリ自動作成支援システムの適用可能なモジュール化を試みる。

この研究計画を踏まえ我々が日本語における医療用自然言語処理の研究コミュニティを形成し研究に取り組んで来た標準化技術を実カルテへと適用することで、カルテからの情報抽出の自動化に向けた予備的な検証を行うことを計画した。

その計画の実現に向け、採択後研究者全員で組織する「総括・企画調整班」を作り、平成２８−２９年度においては月 2 回程度のミーティングを行い、研究の方向性の整理及び各分担班の作成及び役割の決定、進捗の管理及び調整を行う方法で研究を遂行することとした。また、「総括・企画調整班」以外の分担班についてはそれぞれ責任研究者を決め、その裁量で研究を進める方法をとった。

平成３０年度は、昨年度まで実施していた研究者全員で組織する「総括・企画調整班」の機能を縮小し、それぞれの分担班においてそれぞれ責任研究者を決め、その裁量で研究を進める方法を中心として運営を選択した。

（尚、総括・企画調整班以外の分担班については総括・企画調整班の活動結果から生まれたものであり、それぞれの班の目的・方法についても C. 結果セクションで記載することとする。）

C.結果

１．総括・企画調整班

まず、本研究班はその応募要項の段階からデータ収集に掛かる部分については研究の中に組み込まないことを求められており、データの収集基盤の構築・運営については本研究のカバーする範囲ではない。しかしながら、本研究の前提となる国立病院機構が作成・維持運営する NCDA について本報告書でその概要や意義について記述を行わないとするならば、本報告書の内容の理解に大きな妨げになると考えここに報告を行うこととする。

NCDA の概要については参考資料１にその概要を資料を添付した。また、実際の病院における SS-MIX データ作成に掛かるシステム仕様についても参考資料２に示した。これらの仕様等のドキュメントについては

その改版履歴も含め、github 上で管理、公開している。
https://github.com/nhoHQ/SSMIX2_support_documents

次に、実際の本分担研究班の活動についての報告を行う。本研究班においては、まずは研究者が独立して研究活動を進めるのではなく、１０回の研究班会議（うち８回は Web 音声会議）を行い、１つの有機的な研究班として活動が行える環境で運営してきた。

各班会議での調整事項は以下の通りである。

第 1 回
- 研究管理面の話題
- 各人の状況 update
- 仮説構築作業
- 「退院サマリとは何か？」

第 2 回
- 作業仮説構築

ツールドリブン/リソースドリブン/臨床ニーズ/ 病院管理ニーズからの整理
- 倫理審査に向けた論点整理

病院訪問に向けた調整

第 3 回
- 三重病院にて退院サマリの記載内容について臨床家とともにディスカッション

第 4 回
- 今後のスケジュール確認
- 三重病院訪問での成果確認
- 倫理審査に向けた調整
- 研究分担の整理
- 「良質な退院時サマリとは？」問題の整理
- ダミーカルテ作成の是非

第 5 回
- 継続申請書類の作成について
- 倫理審査の申請書確定について
- 医師アンケート企画について
- テストデータについて
- 解析のアプローチについて

第 6 回
- H29 継続申請について
- H28 倫理審査の状況報告
- 年度内達成目標の再確認
- 各分担研究状況報告
- テストデータについて

第 7 回
- 各分担研究状況報告
- 研究基盤の整理
- 倫理委員会・利活用審査委員会の報告

第 8 回
-各分担研究状況報告
-退院サマリの関する文献サーベイについて
- カルテ要約の要素技術についての議論
- スケジュール確認

第 9 回
-各分担研究状況報告
-退院サマリの関する文献サーベイについて
- 報告書作成について
- 分析環境の整備について報告

第 10 回
-各分担研究状況報告
- 報告書作成について

第 11 回
- 平成 29 年度計画の確認

第 12 回

- 各自状況報告
- 研究成果報告書
- 記録の保管について

第 13 回

- 各自状況報告

研究成果報告書

年間スケジュール

第 14 回

- CLEF シェアードタスク
- 荒牧班との共同研究班会議について
- 情報共有／相談
- 研究分担の整理
- 「良質な退院時サマリとは？」問題の整理
- ダミーカルテ作成の是非

第 15 回

- 匿名化について
- 荒牧班との共同研究班会議
- 各自状況報告

第 16 回

- 匿名化について
- 各自状況報告
- 年度内達成目標の再確認
- 各分担研究状況報告
- テストデータについて

第 17 回

- 各分担研究状況報告
- 研究基盤の整理

第 18 回

-各分担研究状況報告

-研究発表の場について

- スケジュール確認

第 19 回

-各分担研究状況報告

- 報告書作成について
- 分析環境の整備について報告

第 20 回

-各分担研究状況報告

- 報告書作成について

なお、NCDA データは国立病院機構が契約するデータセンター内で厳重に管理されている。研究に際しては、このデータベースから研究テーマごとに匿名化したサブセットを切り出し、国立病院機構本部内のオンサイト利用に限っている。以上により、データセットの利用対象と利用目的を厳しく制限することにより、患者個人情報の保護を行っている。それに対応する分析基盤の作成に関して、分担研究班を組織し、堀口・岡田が責任者として活動を行うこととした。(分担研究の結果は後述)。また、最終年度においてはこの分担班で「部門システム化を目指したモジュール開発」のタスクを担当することとした。

また、「退院サマリの自動生成に向けたアプローチの検討」というテーマの分担研究を奥村が、「退院サマリの自由記載文の特徴解析」というテーマの分担研究を森田が、「退院サマリの自動生成に向けたアプローチの検討」というテーマの分担研究を狩野が担当することとした。

２．SS-MIX2 分析用データセットの作成・開発班

研究代表者堀口及び分担研究者岡田は、国立病院機構本部との調整を中心とした基盤構築を行った。初年度 NDCA データの研究利用に向け、倫理審査申請に加えて、内部規定にて定められている内部委員会の調整を図りともに承認を得た。また、閲覧・解

析に特化した自然言語処理用の研究基盤の構築を行った。研究基盤の概念図は図１に示したとおりで、セキュリティを維持しつつ、空間的制約をなるべく少なく研究が進められるようなものになっている。
また、本研究で中心的に使われる医師記録等（経過記録・退院サマリ）については、SS-MIX2 の標準仕様に含まれていないが、JAHIS の提供している仕様を参考に、資料１で示した仕様で NCDA 内に実装することとした。
汎用的にどの電子カルテからでも利用できるシステムの構築という課題に対して、JAHIS の HL７CDA 規約に基づくデータを SS―MIX 拡張ストレージに保管するという現在日本における標準的なデータ交換フォーマットを起点として、特定の患者の医師記録を抽出し、そこからサマリを作成するフローを実現するモジュールの開発を行った。SS-MIX ストレージを活用したデータ保管システムはすでに国内の大手ベンダー８社においては NCDA 上で実現されており、横展開可能なシステムの構築という本研究に求められたミッションの１つは達成できたと考えている。

３．退院サマリの自動生成に向けたアプローチの検討班

入院カルテの自動要約に向けた研究基盤の整備に取り組んできた。臨床医は、入院治療をしていた患者が退院する際、それまでに記載していた入院カルテから退院サマリを作成する。この退院サマリの作成を効率化することができれば、医師の診療負担を直接軽減することが出来ることに加えて、様々な副次的な効果が期待される。
今年度は、昨年度までの研究をさらに発展させ、4 つの課題に取り組んだ。まず、一連の研究には自由に研究利用できるカルテの存在が必要となる。そこで、ダミーカルテの収集を進め、100 件の整備を目指して活動を進めた。また、収集したダミーカルテを対象として、機械的な処理を可能とするためのアノテーション作業に取り組んだ。その際、初年度に行ったアノテーション、２年度に試みたアノテーションを踏まえ、３年目にさらなる改善を図った。次に、このアノテーション作業と平行して、退院サマリの分析モデルであり生成モデルである「CASE モデル」の改善に取り組んだ。また、退院サマリの自動生成処理を精度管理するうえで必要となる「理想の退院サマリ」の確保に向けた検討を行った。
研究活動の結果、ダミーカルテは、目標を超える 108 件を集めることが出来た。また、これらのカルテを対象としたアノテーションを進めると共に、アノテーションガイドラインを高品質化することが出来た。さらに、このアノテーション済みカルテを用いて実現する退院サマリの分析に向けたモデルとして、修正 CASE モデルを提示することが出来た。理想のサマリに向けた検討では、高品質な退院サマリを低コストに実現するための作業仮説を整理すると共に、実証に向けた研究デザインを策定することが出来た。

４．退院サマリの自由記載文の特徴解析班

退院サマリの自動生成技術の実現を目指し，昨年度までの調査結果の整理およびそれを踏まえた退院サマリの自動評価のコンセプ

ト検証を実施した。退院サマリを生成および評価するためには，どのような退院サマリを生成しなくてはならないかを示す「理想的な退院サマリ」の定義が必要となる。昨年度に理想的な退院サマリについて言及をした国内外の文献調査を実施し，今年度はその結果を整理した。退院サマリの記載に関して「記載すべき項目」および「定性的な要件事項」を抽出し，前者として 38 項目を得た。このうちの 14 項目はカルテの構造化データより抽出できるものの，残りの 24 項目はカルテの自由記載より抽出してサマリとして文章を作成や要約をする必要があるものと考えられた。また，後者として 9 項目を得た。このうちの 5 項目は医学的な知識・経験がないと採点が難しいものの，残りの 4 項目は形式的に判断をすることが可能と考えられた。この 4 項目のうち医学用語辞書を必要としない 3 項目について文章の特徴を用いて自動評価することを試み，人による評価との相関および点数の分布を踏まえて自動評価の可能性を考察した。退院サマリの自動評価手法の確立に向けた今後の課題が明らかとなった。

５．退院サマリの自動生成に向けたアプローチの検討班

退院サマリの自動生成に向けたテキストの分析についての研究を行った。入院患者の退院に際し、医師は入院中に記載したカルテ等の情報から退院サマリを作成する必要がある。この退院サマリを自動的に生成することが出来れば、臨床現場の負担を下げることが出来ると共に、医療の質に貢献することが期待される。

カルテの処理にあたっては、事前に匿名化が必要となる。匿名化作業を自動化するための匿名化ツールの実装と性能向上に取り組んだ。そのために、既存の正解付き模擬カルテデータに加え、別のダミーカルテデータセットに対し匿名化のためのアノテーション付与を行い、これらを用いてルールベースおよび機械学習による匿名化ツールの実装と性能検証を行った。

サマリ生成にあたっては、対象とするカルテやサマリのドメイン、すなわち診療科や疾患により、サマリ生成に必要となる情報が異なると考えられる。サマリと対応する電子カルテの履歴データについてクラスタリングを行い、どのようなタイプのサマリやカルテがどう類似しうるかの分析を行った。

退院サマリの自動生成にあたっては、extractive な処理を行うこととし、電子カルテ内の各文についてサマリに含めるべきか否かの判断を行った。判定には、文末表現に着目する手法、文ベクトルを生成して文間の類似度で決定する方法などいくつかの手法を行い、文一致率、単語一致率、ROUGE など異なる指標での評価を行った。結果、いずれの手法でもある程度の一致率を達成しうることがわかった。今後、より実用的な性能の達成のためには、表記揺れの吸収、さらに大規模なデータの利用による学習性能の向上などの研究が考えられる。

D・E.考察及び結論

医療用情報システムの研究開発においては、医療現場に直接の恩恵が及ばないゴー

ルが設定されることで、研究開発が現場のニーズから乖離するとともに、継続した開発投資に繋がらない悪循環が往々にして生じてきた。本研究提案は、医療現場における負担軽減策として期待が大きい退院サマリの自動要約技術の開発を目指す。これにより、紹介状の自動作成技術等、電子カルテの自動解析技術に関連する継続的な研究開発投資の実現が期待される。この研究開発サイクルを確立することにより、要素技術である電子カルテ上の記載から自動情報抽出における継続的な精度向上が期待される。

こうして確立する医療用自然言語処理技術は、大量の電子カルテからの効率的な情報抽出を実現し、健康医療政策に資する統計データの収集コストを劇的に低廉化することが期待される。とりわけ、様々な傷病や治療に関して､既存の DPC やレセプトには表れてこない深遠な実態を明らかとし、医療の質向上・均てん化・各種医療技術の臨床開発に必要なエビデンスを生み出してくことが期待される。

また、国立病院機構の有する広域電子カルテ網は、各病院が独自に調達した電子カルテベンダー主要８社を網羅している。本研究によって、大口顧客としての交渉力を背景としたこれら主要ベンダーへの研究開発成果の技術移転が期待される。これらは、医療の情報化を進める厚生労働行政にとって、新たな政策手段の実現をもたらす

加えて、本研究で作成した入院カルテ・退院サマリの高品質な個人情報を含まない模擬のデータセットは、研究成果として公表することにしており、電子カルテの現物を持たない情報系研究者が、容易に本領域の発展に貢献できる環境を提供出来る点も大きな成果になるものと考えている。

今年度の研究においては、昨年度に引き続きサマリ作成についての各種技術の研究を着実に進めたとともに、本研究の当初の目的である、汎用的にどの電子カルテからでも利用できるシステムの構築という課題に対して、JAHIS の HL７CDA 規約に基づくデータを SS―MIX 拡張ストレージに保管するという現在日本における標準的なデータ交換フォーマットを起点として、特定の患者の医師記録を抽出し、そこからサマリを作成するフローを実現するモジュールの開発を行った。SS-MIX ストレージを活用したデータ保管システムはすでに国内の大手ベンダー８社においては NCDA 上で実現されており、横展開可能なシステムの構築という本研究に求められたミッションの１つは達成できたと考えている

Ｆ．研究発表

１．論文発表

古崎晃司,堀口裕正,奥村貴史,津本周作: OS-27　人工知能の医療応用 人工知能 33(6):843-848 2018

堀口裕正: 国立病院機構のデータベースを用いた臨床研究 Progress in Medicine 38(2):17-20 2018

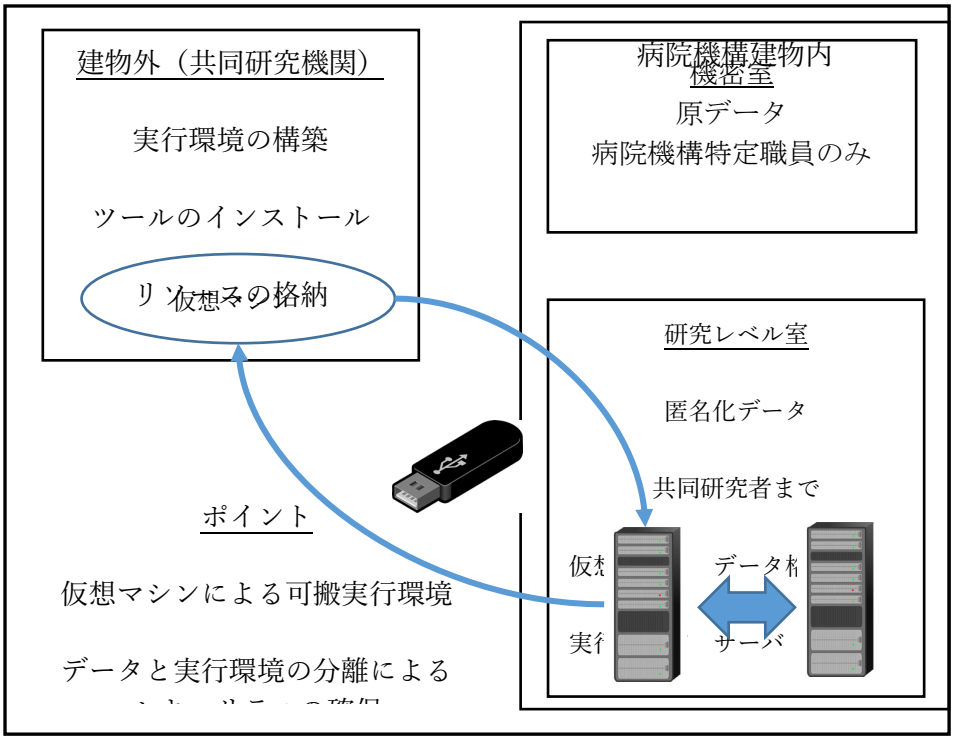
堀口裕正: NCDA の現況と成果および今後の展望 月刊新医療 2018 年 2 月号

２．学会発表

森田瑞樹, 奥村貴史, 狩野芳伸, 堀口裕正. 退院サマリの自由記載は何を書くことが望ましいのか：文献レビュー, 第38回医療情報学連合大会, 2018年11月22～25日, 福岡.

[Kajiyama 18] Kajiyama, K. Horiguchi, H. Okumura, T. Morita, M. Kano, Y., 2018. De-identifying Free Text of Japanese Dummy Electronic Health Records. The Ninth International Workshop on Health Text Mining and Information Analysis(LOUHI2018) (p. 65).

図 1



資料１　NCDA における医師記録等の仕様書

趣旨

本事業では、各社の SS-MIX2 モジュールの拡張ストレージへの出力機能を利用し、以下の情報を出力することを求めている。その際、SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d（以下、ガイドライン）に記載している仕様に対応していること。また、トランザクションストレージ、インデックスデータベースも同時に生成すること。

- 経過記録
- 退院時サマリー
- 診療情報提供書

以下に仕様を示す。

ドキュメントデータ 物理構造

```
|-- 拡張ストレージ ルートフォルダ
    |-- 患者 ID 先頭 3 文字
        |-- 患者 ID 4〜6 文字
            |-- 患者 ID
                |-- 診療日
                    |-- データ種別
                        |-- コンテンツフォルダ
                            |--主文書ファイル
```

診療日

特に指定しない。

データ種別

ガイドライン P4 (4)「データ種別フォルダ」について に則ること。

```
[ローカル文書コード]^ローカル文書名称^[ローカルコード体系コード]^標準文書コード^標準文書名称^標準コード体系コード
```

以下のように標準コードに対しローカルコードが複数あることは許容される。

```
L12345^入院診療録^99ZZZ^11506-3^経過記録^LN
L12346^外来診療録^99ZZZ^11506-3^経過記録^LN
```

コンテンツフォルダ

ガイドライン Ver.1.2d P5 (5)「コンテンツフォルダ」について に則ること。

```
患者 ID_診療日_データ種別コード_特定キー_発生日時_診療科コード_コンディションフラグ
```

いずれの文書も削除は想定していないが、電子カルテシステムによっては修正はあり得ると考える。その場合、ガイドライン P6 ④修正が発生する場合 に則り改版すること。

主文書ファイル

XML CDA R2 で出力すること。XML ファイル以外に画像ファイルや CSS ファイル等を出力してもかまわない。

HEADER 部

いずれの文書も JAHIS　診療文書構造化記述規約 共通編 Ver.1.0 に則ること。
P27 6.3.11.検査・診療等行為 "documentationOf/ServiceEvent" によると、documentationOf の制約・多重度は 0..1 となっているが、 経過記録、退院時サマリについてはこれを 1..1 と読み替えること。
経過記録は serviceEvent classCode(サービスイベントクラスコード)を ENC(診察)とし、effectiveTime(実施日)は low value、high value ともに記録タイミングを出力すること。
退院時サマリは serviceEvent classCode(サービスイベントクラスコード)を ACCM(入院、滞在)とし、effectiveTime(実施日)は low value に入院タイミング、high value に退院タイミングを出力すること。
タイミングの粒度は日以上であれば良い。

BODY 部

診療情報提供書は、日本 HL7 協会 患者診療情報提供書 規格 Ver.1.00 に則ること。
診療情報提供書以外は、XML の文法に則ること。

参考資料

１．NCDA データベースの説明資料

国立病院機構
診療情報集積基盤について

～電子カルテデータの標準化から利活用へ～

NCDA

NHO Clinical Data Archives

独立行政法人
国立病院機構
National Hospital Organization

平成31年3月現在

Contents

1．NCDAとは

国立病院機構のデータベースの特徴

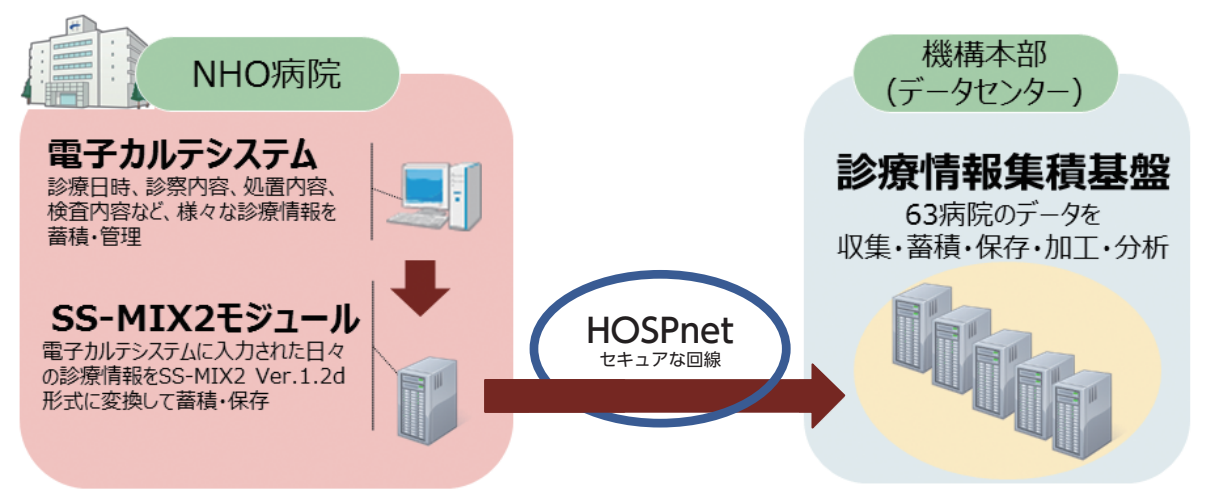
国立病院機構では、厚生労働省より平成26年度地域診療情報連携推進費補助金をうけて『電子カルテデータ標準化等のためのIT基盤構築事業』を実施しています。当事業では、SS-MIX2規格を利用し、データの標準化に最大限取り組んだ２次利用可能なデータベースとして国立病院機構診療情報集積基盤（NCDA：NHO Clinical Data Archives、平成30年度末で63病院が参加）を構築すると共に、SS-MIX2規格を用いて電子カルテデータの標準化を進め、その工程を示したドキュメト（手順書）を作成・公開しています。

NCDAの特徴は、以下のとおり３つ挙げられます。

①平成28年１月以降のSS-MIX2規格（標準化ストレージ機能）に含まれる全データ種別（病名・入退院・転棟・外来来院・食事・処方・投薬・検査）のデータを保有。
②入院患者のバイタルサインデータ（血圧・体温・心拍数）を保有。
③国立病院機構では単一の法人格によって複数の病院が運営されていることから、利活用におけるデータ提供の際には匿名化が行われているが、収集・データベース化の段階では匿名化せずに患者番号や生年月日・住所といった個人識別子を含んだ状態で運用。

例えばNCDAとＤＰＣデータやレセプトデータとのデータ結合や病院における追加調査の実施等において技術的な制約がない点が特徴としてあげられます。加えて、IT基盤構築事業の結果、標準化されて各病院で同じマスターを利用し、数値が数値として認識でき、数値の単位も標準化されたデータベースとなっている点も大きな特徴です。

本パンフレットは、平成30年度末現在のNCDAの特徴と現況について説明を行っています。この資料が皆様のNCDAへの理解と利活用に関する興味を持っていただくことの一助になることを期待しております。



2．NCDAの経緯

年表

平成26年	SS-MIX2データの収集に向けて、パイロットスタディを4病院で開始
平成26年3月	パイロットスタディ参加病院を12病院まで拡大
平成26年4月	国立病院機構第3期中期計画開始（「電子カルテ情報の収集・分析について具体的な検討を進め、臨床研究等のIT基盤の充実を図る」ことを明記） 電子カルテ情報の収集・分析について具体的な検討を開始
平成27年2月	平成26年度地域診療情報連携推進費補助金「電子カルテデータ標準化等のためのIT基盤構築事業」開始
平成28年3月	IT基盤（NCDA）構築事業完了 6ベンダ41病院でNCDAを運用開始

電子カルテデータ標準化等のためのＩＴ基盤構築事業

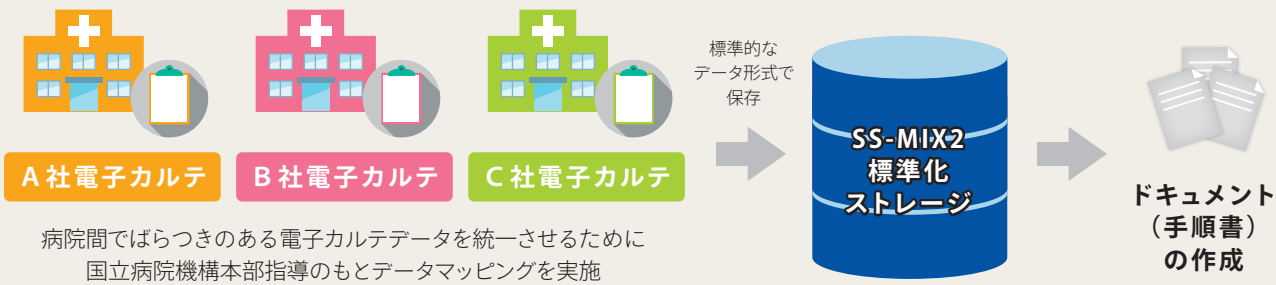
平成２６年度半ば～平成２７年度末

電子カルテについては、ベンダごとで開発が行われ、各病院が使いやすいようにカスタマイズされるなど、電子カルテデータの形式が標準化されないまま普及したことから、電子カルテ上で使用されている病名や医薬品等のコードがベンダや病院で異なり、標準化の課題となっている。

本事業は、上記のような問題を解消するため、厚生労働省から平成26年度地域診療情報連携推進費補助金として事業を付託されたものであり、各病院の電子カルテデータを厚生労働省の定める標準コードに紐付けするデータマッピングを行い、SS-MIX2規格（標準化ストレージ機能）を用いて電子カルテデータの標準化を実施するとともに、その工程を示したドキュメント（手順書）を作成・公開することを目的としている。

事業内容

主要なベンダや多くの疾病領域について対応可能な精度の高いドキュメントを作成するために、41病院で電子カルテデータ標準化事業を行う。



病院間でばらつきのある電子カルテデータを統一させるために
国立病院機構本部指導のもとデータマッピングを実施

平成28年11月	国立病院機構診療情報データベース利活用規程制定 平成28年度地域診療情報連携推進費補助金「電子カルテによる『災害診療記録』電子フォーマット自動出力実証事業」開始
平成29年8月	災害診療記録事業に7ベンダ55病院が参加
平成30年3月	災害診療記録事業完了 NCDAの運用は8ベンダ58病院まで拡大
平成31年3月	NCDAの運用は8ベンダ63病院まで拡大

電子カルテによる『災害診療記録』電子フォーマット自動出力実証事業
平成28年度半ば～平成29年度末

大規模災害時において、災害対策本部（都道府県）が被災地の医療概況を把握し、適確な医療支援活動を展開するうえで、極めて重要な情報は「疾病別症例数」等の集計情報であるが、それを迅速に集計する手法の確立が課題となっている。

この課題に対し、東日本大震災を契機に「災害時の診療録のあり方に関する合同委員会*」が設置され、災害時の標準的記録フォームといえる「災害診療記録」が作成され、平成28年熊本地震で初めて運用開始された。本事業はこの電子フォーマットの電子カルテへの実装と収集に係る実証を行うことを目的とする。

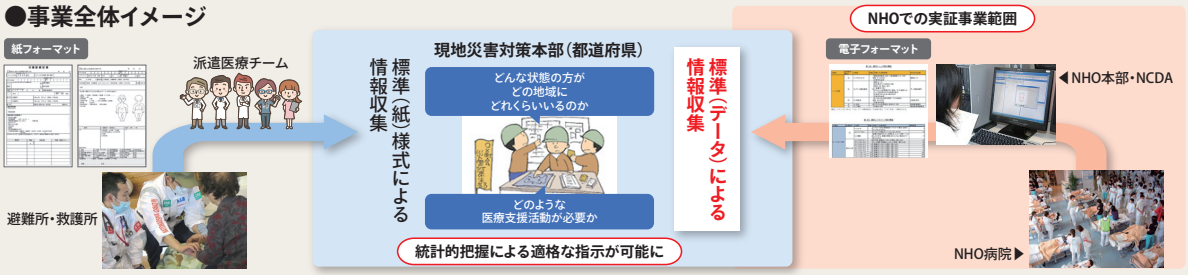
平成29年度末時点で、ＮＣＤＡ参加病院のうち災害拠点病院を中心に55病院で本モジュールを導入済みである。

*日本医師会、日本病院会、日本集団災害医学会、日本救急医学会 などが参加

事業内容

ＮＣＤＡの標準化機能を活かして、様々なベンダの電子カルテから災害診療記録用の電子フォーマットの出力が可能となるよう対応モジュールをバージョンアップし、災害時に必要な診療情報の自動抽出化等の開発及び検証を行い、その結果を導入手順書として公開することを通じて、災害発生時の適確な医療支援活動の展開に役立てるもの。

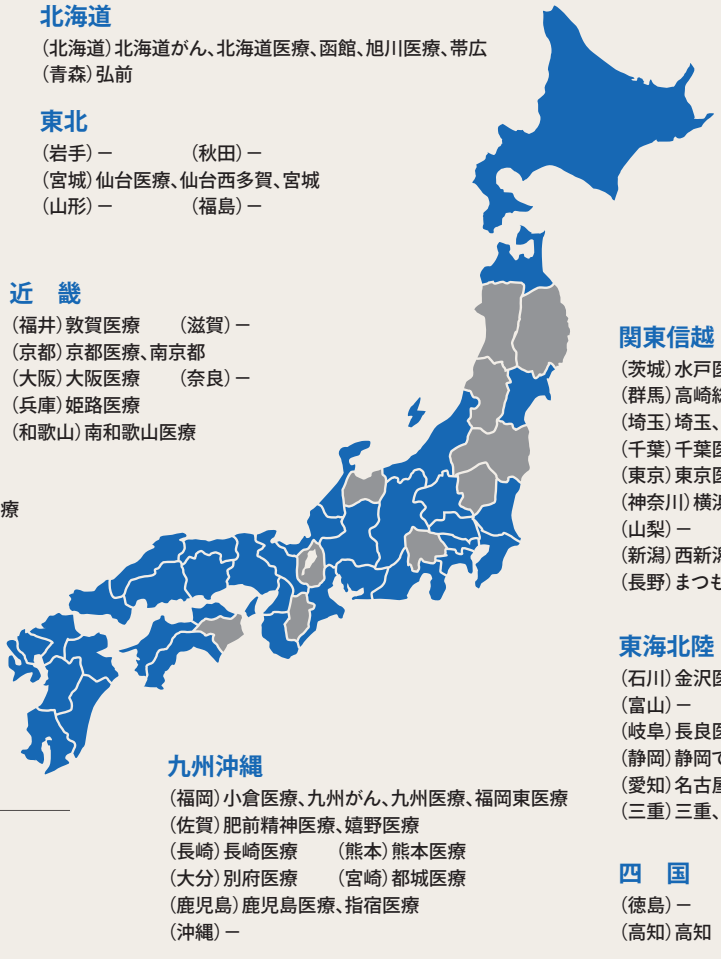
●事業全体イメージ



3．NCDAの現状

NCDA参加病院の状況（平成31年3月現在）

NCDA参加病院
63病院
うち災害診療記録事業参加
60病院



北海道
（北海道）北海道がん、北海道医療、函館、旭川医療、帯広
（青森）弘前

東北
（岩手）－　　（秋田）－
（宮城）仙台医療、仙台西多賀、宮城
（山形）－　　（福島）－

近　畿
（福井）敦賀医療　　（滋賀）－
（京都）京都医療、南京都
（大阪）大阪医療　　（奈良）－
（兵庫）姫路医療
（和歌山）南和歌山医療

中　国
（鳥取）米子医療　　（島根）松江医療
（岡山）岡山医療
（広島）呉医療、広島西医療
（山口）山口宇部医療、岩国医療

関東信越
（茨城）水戸医療　　（栃木）－
（群馬）高崎総合医療、渋川医療
（埼玉）埼玉、東埼玉
（千葉）千葉医療
（東京）東京医療、災害医療、東京、村山医療
（神奈川）横浜医療、箱根、相模原
（山梨）－
（新潟）西新潟中央
（長野）まつもと医療、信州上田医療

東海北陸
（石川）金沢医療、医王
（富山）－
（岐阜）長良医療
（静岡）静岡てんかん、天竜、静岡医療
（愛知）名古屋医療、東名古屋
（三重）三重、三重中央医療

九州沖縄
（福岡）小倉医療、九州がん、九州医療、福岡東医療
（佐賀）肥前精神医療、嬉野医療
（長崎）長崎医療　　（熊本）熊本医療
（大分）別府医療　　（宮崎）都城医療
（鹿児島）鹿児島医療、指宿医療
（沖縄）－

四　国
（徳島）－　　（香川）高松医療
（高知）高知　　（愛媛）四国がん

ベンダ・病院種別分布

	500床以上	350～499床	349床以下	複合（その他）	障害病床中心	精神科病床中心	総計
富士通	**9病院** 金沢医療、名古屋医療、京都医療、大阪医療、呉医療、岩国医療、九州医療、長崎医療、熊本医療	**7病院** 千葉医療、横浜医療、相模原、三重中央医療、姫路医療、小倉医療、別府医療	**2病院** 南和歌山医療、指宿医療*	**6病院** 渋川医療、東京、村山医療、長良医療、山口宇部医療、福岡東医療	**9病院** 宮城、東埼玉、箱根、医王、東名古屋、三重、南京都、松江医療、広島西医療		33
日本電気		**3病院** 北海道がん、埼玉、災害医療	**1病院** 信州上田医療	**6病院** 北海道医療、旭川医療、帯広、まつもと医療、静岡医療、高知	**1病院** 仙台西多賀		11
ソフトウェア・サービス	**1病院** 岡山医療	**6病院** 水戸医療、高崎総合医療、四国がん、九州がん、嬉野医療、鹿児島医療	**1病院** 米子医療		**1病院** 高松医療*		9
亀田医療情報				**1病院** 敦賀医療	**1病院** 西新潟中央		2
SBS					**2病院** 静岡てんかん、天竜		2
日本IBM	**2病院** 仙台医療、東京医療						2
CSI			**3病院** 函館、弘前、都城医療				3
ナイス						**1病院** 肥前精神医療*	1
総計	12	16	7	13	14	1	63

* 災害診療記録事業には参加していない病院

病院数・病床数・患者数

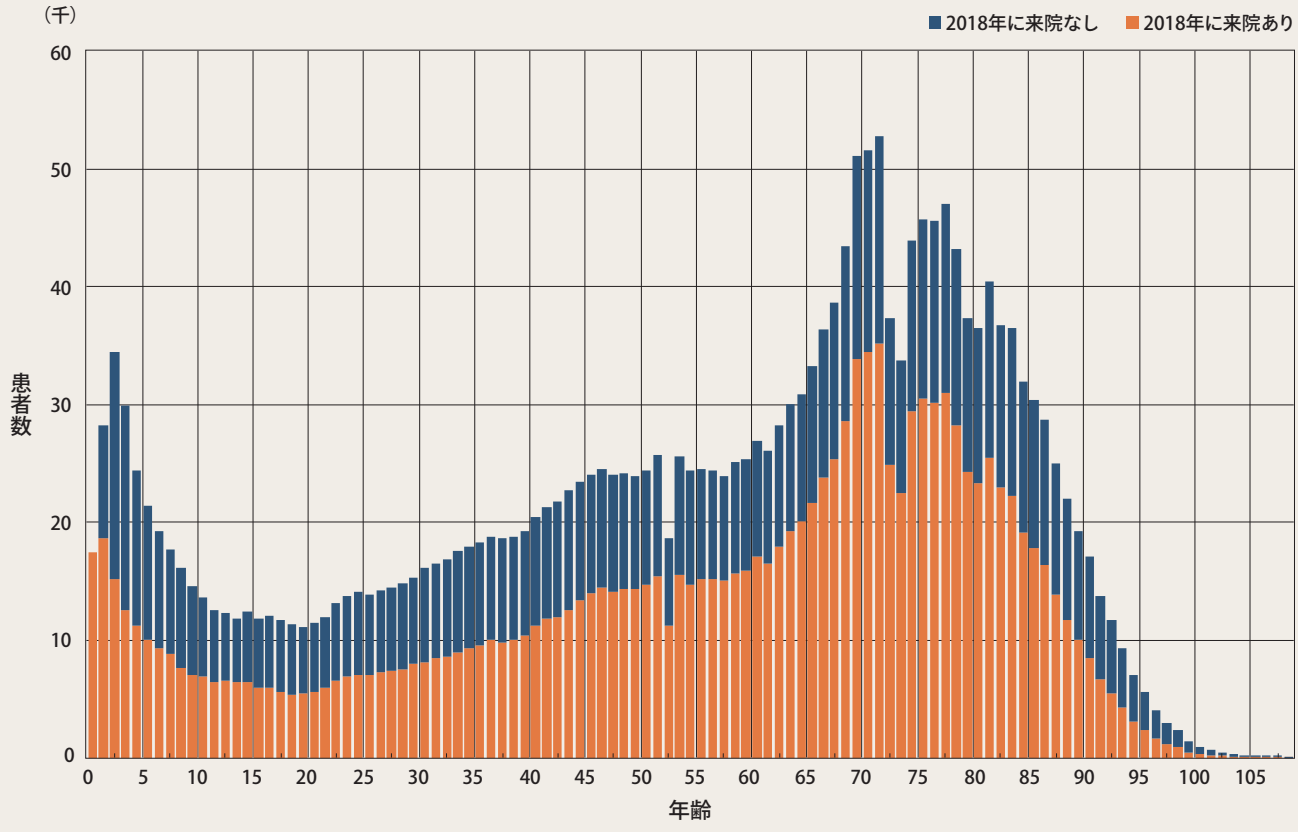
平成28年度末時点
病院数 41病院
病床数 1.8万床
患者数 約100万人

平成29年度末時点
病院数 58病院
病床数 2.4万床
患者数 約160万人

平成30年度末時点
病院数 63病院
病床数 2.7万床
患者数 約190万人

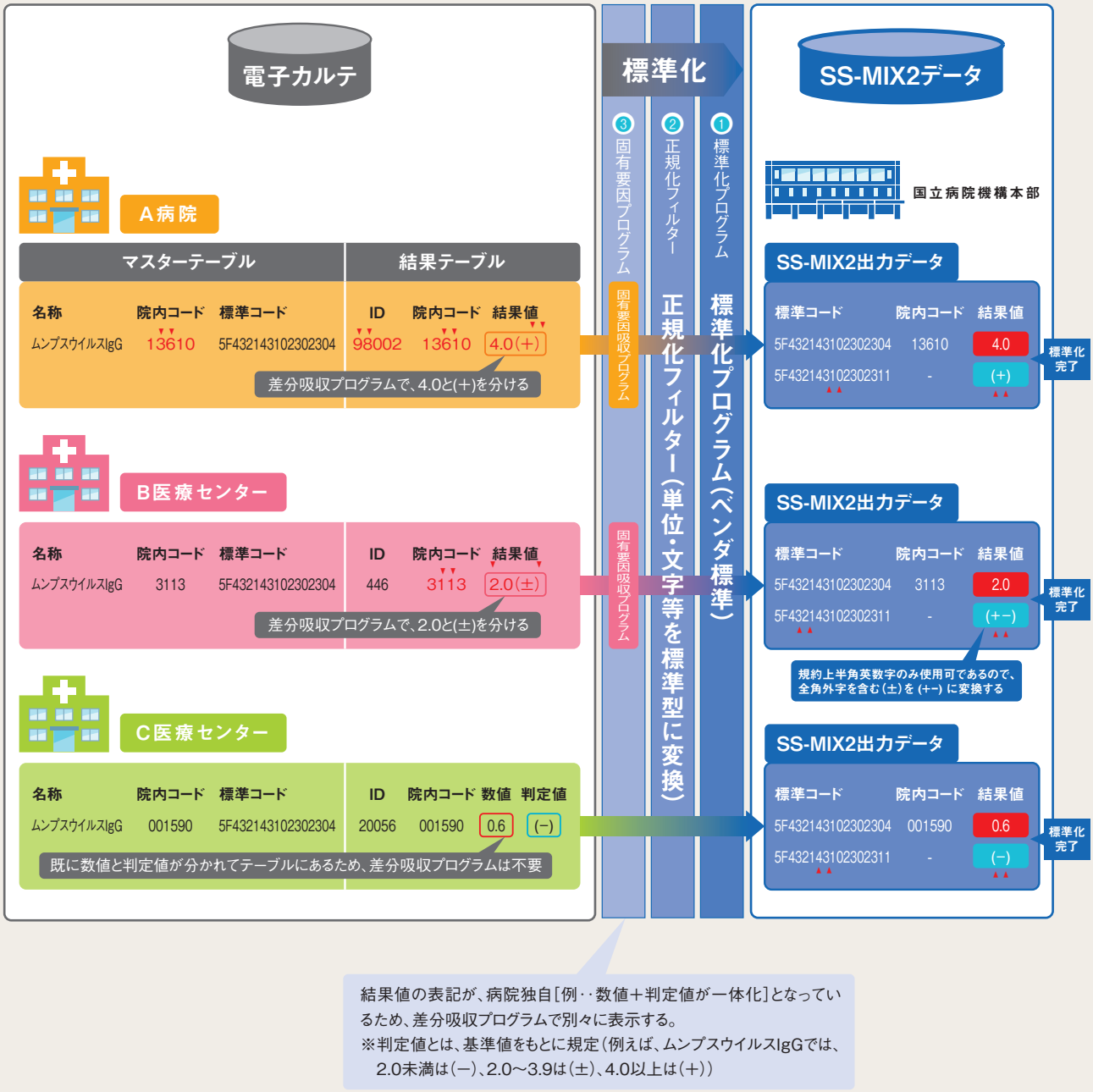
NCDAに集積している患者数（年齢）の状況

NCDAデータ保有患者数（平成30年12月現在）



4．SS-MIX2変換プログラムの構成

SS-MIX2変換プログラムの構成



結果値の表記が、病院独自[例‥数値＋判定値が一体化]となっているため、差分吸収プログラムで別々に表示する。
※判定値とは、基準値をもとに規定（例えば、ムンプスウイルスIgGでは、2.0未満は(−)、2.0〜3.9は(±)、4.0以上は(+)）

■①②の標準化プログラムは他の医療機関でも使用可能な汎用的なもの。
■③の固有要因プログラムについては病院固有のもの。
各ベンダーが構築する標準化プログラムに固有要因プログラム③の機能が含まれていると、病院独自の仕様となり汎用化できないため、普及促進を図る手順書としての品質は不可。
→今回の事業では、複数病院で標準化プログラムを運用して、それが汎用的なものであること（病院固有の変換機能が入っていないこと）を確認する。
※①②③のプログラムの著作権はベンダーにあるため国立病院機構はコード等中身を見ることは不可。
国立病院機構は出力結果により汎用性を確認する。

5．保有するデータ種別

データ種別	データ内容	使用している標準コード
ADT-00 ADT-01 ADT-12 ADT-21 ADT-22 ADT-31 ADT-32 ADT-41 ADT-42 ADT-51 ADT-52 ADT-61	患者基本情報 担当医 外来診察受付 入院 外出泊 転科・転棟(転室・転床) 退院 アレルギー情報	
PPR-01	病名（歴）情報	ICD-10
OMP-01 OMP-02 OMP-11 OMP-12	処方・注射	HOT9
OML-01 OML-11	検体検査	JLAC-10
OMG-01 OMG-02 OMG-03 OMG-11 OMG-12 OMG-13	放射線・内視鏡・生理検査	
L-OBSERVATIONS^OBSERVATIONS^99ZL01	バイタル検査結果	
^（ローカル名称）^^11506-3^経過記録^LN	診療録（外来/入院含む）	
^（ローカル名称）^^34108-1^外来診療録^LN	診療録（外来）（入院・外来が別の場合）	
^（ローカル名称）^^34112-3^入院診療録^LN	診療録（入院）（入院・外来が別の場合）	
^（ローカル名称）^^18842-5^退院時サマリー ^LN	退院時サマリー	
^（ローカル名称）^^57133-1^紹介状^LN	診療情報提供書	
L-JSPEED^災害時JSPEED記録^99ZL01^74465-6^災害時JSPEED記録^LN	災害時JSPEED記録	

6．構築・運用費用について

各病院における構築費用については、ベンダ側からの当初提示額から1～2割の低減を行い、1病院あたり平均約700万円となりました。また、本事業を安定的に継続していくため、運用コストの低廉化について検討した結果、1病院あたり年間約20万円での運用が可能となりました。

各病院での電子カルテ更新の際に発生するコストについては、今後も交渉を行い、さらなる低廉化が図れるよう努力していきます。

7．NCDAの利活用について

平成28年11月に「国立病院機構診療情報データベース利活用規程」を制定し、研究における利活用については本規程を遵守するとともに、倫理規程等の研究に関連する法令やルールを遵守することを定めました。利活用の際は、匿名化を原則としています。

また、国立病院機構本部内に「診療情報データベース利活用審査委員会」を設置し、データ利活用及び成果公表の適切性について審議を行っています。

利活用申請件数（データソースがNCDAのもの）

	平成28年度	平成29年度	平成30年度
業務	1	0	1
研究	2	0	4

8．NCDAにおける個人情報の取扱い

NCDAにおいては、以下の方針で個人情報を取り扱っています。

●患者同意

各病院において掲示している個人情報の利活用目的の範囲内で実施することを原則とします。そのために、ポスターを院内に掲示するとともに、患者からの利用不可の申出に対応できる体制を整備しています。

●法令対応

個人情報保護法、独立行政法人における個人情報保護に関する規程、ガイドラインおよび研究の倫理指針等に適切に対応します。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にも適切に対応していきます。

9．ＮＣＤＡを活用した研究例

【研究名】

「腹部悪性腫瘍手術患者における術前血糖管理と術後創部感染症の関連に関する研究」

（研究目的）

腹部悪性腫瘍患者において術前の血糖管理状態と術後の創部感染症の関連を検証し適切な術前血糖管理目標を特定する

【研究名】

「電子カルテ情報をセマンティクス（意味・内容）の標準化により分析可能なデータに変換するための研究」

（研究目的）

電子カルテに実装可能な医療用語の標準化を行うシステムの開発に向け、そのステップとして、臨床現場からのニーズが極めて大きい退院サマリの自動生成技術を実現する。

10．MIAについて

～電子レセプトデータとDPCデータの利活用～

国立病院機構では、NCDAとは別に、診療情報データバンク（MIA:Medical Information Analysis databank）を運営しています。

MIAでは、国立病院機構の全病院から診療情報の利活用を目的としてDPCデータ及びレセプトデータを収集しており、これらの既存データを二次利活用することで、疫学研究の推進に役立てています。

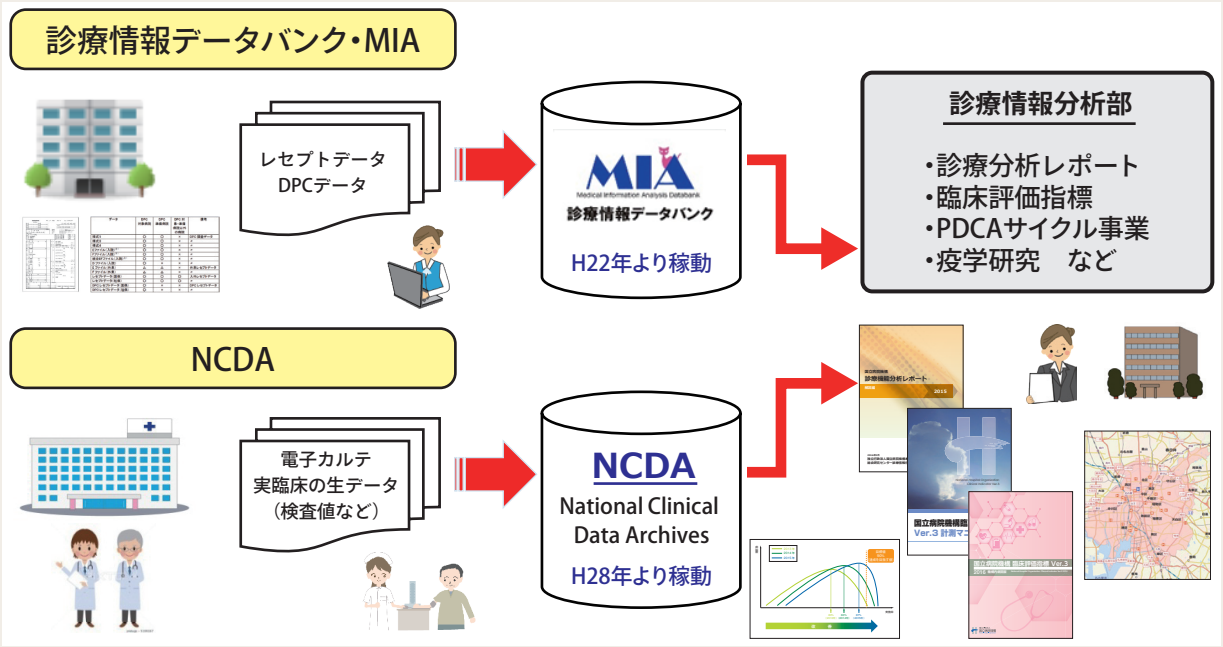
MIAの特徴

NCDAの特徴として挙げた、「③国立病院機構では単一の法人格によって複数の病院が運営されていることから、利活用におけるデータ提供の際には匿名化が行われているが、番号や生年月日・住所といった個人識別子を含んだ状態で運用。」について、MIAも同様の運用を行っています。そのため、NCDAとのデータ結合や病院における追加調査の実施等において技術的な制約がありません。

病院種別の変遷とMIAに蓄積されているデータ

	H22	H23	H24	H25	H26	H27	H28	H29
電子レセプト	144	144	144	143	143	143	143	143
DPC参加病院	45	49	53	52	54	54	64	64
DPC準備病院※	9	4	4	4	15	28	25	43
電子レセプトのみ	89	90	86	86	73	60	54	36

※データ提出加算病院を含む



MIA利活用の経緯

平成22年 4月	電子レセプト、DPCデータの収集開始
平成22年10月	臨床評価指標の開発 診療情報分析レポートの編纂
平成26年 4月	国立病院機構第3期中期計画開始（「電子カルテ情報の収集・分析について具体的な検討を進め、臨床研究等のIT基盤の充実を図る」ことを明記） 電子カルテ情報の収集・分析について具体的な検討を開始
平成28年 4月	医療の質改善事業（PDCA事業）開始に伴い、臨床評価指標の集計サイクルを四半期ごととする
平成28年11月	国立病院機構診療情報データベース利活用規程制定
平成29年 4月	NCDAとのデータ相互利活用の開始
平成30年 3月	データの品質を管理するためのバリデーション研究開始

MIAを活用した研究例

【競争的資金の獲得による研究例】

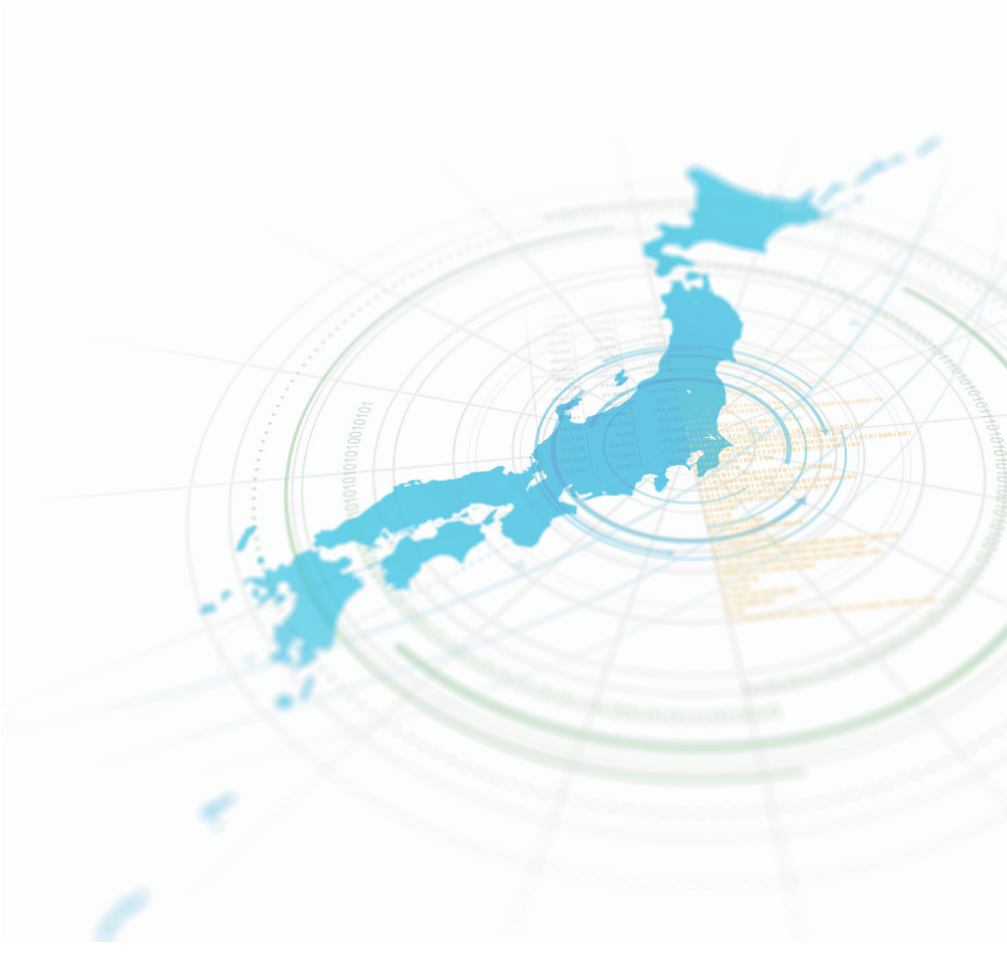
- アレルギー疾患対策に必要とされる疫学調査と疫学データベース作成に関する研究
- 診療情報データベースにおける重症敗血症および播種性血管内凝固症候群の特性に関する研究
- 医薬品使用パターンによる重症度リスク補正を用いたアウトカム評価手法の開発
- B型肝炎ウイルス再活性化に関与するウイルス・宿主要因の解明に基づく予防対策法の確立を目指す研究
- 肝硬変患者の予後を含めた実態を把握するための研究
- 診療情報データベースに記録された情報についての妥当性検証研究
- 新薬へのスイッチの実態が後発医薬品推進政策へ及ぼす影響を評価する研究
- 医療情報データベースを用いた治療効果検証手法の開発：カルテ調査との比較を通して
- 手術患者の術後日常生活活動(ADL)悪化に対する鎮痛薬・鎮静薬・向精神病薬の影響

11. 本事業の今後

これまでの取り組みにより、SS-MIX2規格を用いて電子カルテベンダ毎に異なるデータを標準形式に変換して集積するIT基盤（NCDA）を構築し、その導入手順等の工程を標準作業手順書として作成、公開を行いました。

引き続きNCDAを運用するとともに、今後は、対応ベンダや実施病院の拡大、更には集積されたデータから新たな臨床評価指標の作成・掲示モニタリング、臨床疫学研究の推進、診療機能分析レポートの作成、薬剤副作用調査、被験者データベースによる治験の推進などの利活用を積極的に進めます。

公開した手順書は、他の医療機関・病院グループにおいても、より簡便に変換作業ができるようにすることを目指すものであり、問い合わせに対しても助言等対応を行い、我が国の医療情報の標準化の普及推進に継続的に取り組んでいきます。



国立病院機構本部
情報システム統括部

〒152-8621　東京都目黒区東が丘2-5-21
TEL 03-5712-5130

２．NCDA システム仕様書

SS-MIX2 を用いた診療情報データベース構築の為の SS-MIX2 モジュール技術仕様書

1. システム要件

国立病院機構の各病院にて「国立病院機構診療情報分析基盤(NCDA)」に参加する為に調達する SS-MIX２モジュールの機能は以下の通りである。 但し、本体の電子カルテシステム等の仕様上、作成が不可能であるものについては作成を要しない。その場合、何が不可能かを導入標準作業手順書に記載すること。

1.1 SS-MIX2 Ver.1.2d 機能

SS-MIX2 Ver.1.2d に準拠することとして、以下の機能を有すること。

- 日本医療情報学会発行の「SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン Ver.1.2d」、「SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d」、「SS-MIX2 標準化ストレージ仕様書 Ver.1.2d」、「標準化ストレージ仕様書別紙：コード表 Ver.1.2d」、「SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d 別紙：標準文書コード表」に記載している仕様に対応していること。（尚、当初 Ver.1.2c 準拠としていたが、標準ストレージ部分では Ver.1.2c からの変更点について影響がないため Ver.1.2d 準拠ということとした。）
- 標準化ストレージ、拡張ストレージ、トランザクションストレージ、インデックスデータベースの４つのファイルを生成すること。
- 標準化ストレージにはデータ種別として 36 種のデータを出力すること。

（表 1-1　標準化ストレージ格納データ）

No	データ種別	種別名称	HL7 メッセージ型
1	ADT-00	患者基本情報の更新	ADT^A08
2	ADT-00	患者基本情報の削除	ADT^A23

No	データ種別	種別名称	HL7 メッセージ型
3	ADT-01	担当医の変更	ADT^A54
4	ADT-01	担当医の取消	ADT^A55
5	ADT-12	外来診察の受付	ADT^A04
6	ADT-21	入院予定	ADT^A14
7	ADT-21	入院予定の取消	ADT^A27
8	ADT-22	入院実施	ADT^A01
9	ADT-22	入院実施の取消	ADT^A11
10	ADT-31	外出泊実施	ADT^A21
11	ADT-31	外出泊実施の取消	ADT^A52
12	ADT-32	外出泊帰院実施	ADT^A22
13	ADT-32	外出泊帰院実施の取消	ADT^A53
14	ADT-41	転科・転棟(転室・転床)予定	ADT^A15
15	ADT-41	転科・転棟(転室・転床)予定の取消	ADT^A26
16	ADT-42	転科・転棟(転室・転床)実施	ADT^A02
17	ADT-42	転科・転棟(転室・転床)実施の取消	ADT^A12

No	データ種別	種別名称	HL7 メッセージ型
18	ADT-51	退院予定	ADT^A16
19	ADT-51	退院予定の取消	ADT^A25
20	ADT-52	退院実施	ADT^A03
21	ADT-52	退院実施の取消	ADT^A13
22	ADT-61	アレルギー情報の登録／更新	ADT^A60
23	PPR-01	病名（歴）情報の登録／更新	PPR^ZD1
24	OMD	食事オーダ	OMD^O03
25	OMP-01	処方オーダ	RDE^O11
26	OMP-11	処方実施通知	RAS^O17
27	OMP-02	注射オーダ	RDE^O11
28	OMP-12	注射実施通知	RAS^O17
29	OML-01	検体検査オーダ	OML^O33
30	OML-11	検体検査結果通知	OUL^R22
31	OMG-01	放射線検査オーダ	OMG^O19
32	OMG-11	放射線検査の実施通知	OMI^Z23

No	データ種別	種別名称	HL7 メッセージ型
33	OMG-02	内視鏡検査オーダ	OMG^O19
34	OMG-12	内視鏡検査の実施通知	OMI^Z23
35	OMG-03	生理検査オーダ	OMG^O19
36	OMG-13	生理検査結果通知	ORU^R01

「SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン Ver.1.2d　p11」

1.2 拡張ストレージへの出力機能

現在の SS-MIX2 モジュールでオプションとして既に導入している拡張ストレージへの出力機能は、そのまま提供すること。また、1.3.0 で規定する出力を行うこと。

1.3　NHO 対応としての設定

1.3.0 拡張ストレージへの出力機能

各社の SS-MIX2 モジュールの拡張ストレージへの出力機能を利用し、以下の情報を出力すること。その際、 日本医療情報学会発行の「SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d」に記載している仕様に対応していること。また、トランザクションストレージ、インデックスデータベースも同時に生成すること。

No	データ種別	種別名称	HL7 メッセージ型
1	L-OBSERVATIONS^OBSERVATIONS^99ZL01	バイタル検査結果	HL7 V2.5 ORU^R30
2	^（ローカル名称）^^11506-3^経過記	診療録（外来/入院含	HL7 CDA R2

No	データ種別	種別名称	HL7 メッセージ型
	録^LN	む）	
2.1	^（ローカル名称）^^34108-1^外来診療録^LN	診療録（外来）（入院・外来が別の場合）	HL7 CDA R2
2.2	^（ローカル名称）^^34112-3^入院診療録^LN	診療録（入院）（入院・外来が別の場合）	HL7 CDA R2
3	^（ローカル名称）^^18842-5^退院時サマリー^LN	退院時サマリー	HL7 CDA R2
4	^（ローカル名称）^^57133-1^紹介状^LN	診療情報提供書	HL7 CDA R2

1.3.1　バイタル検査結果通知の出力

（1）バイタル検査結果通知のデータを、別紙の形式で拡張ストレージに出力する。 尚、「診療日」に出力する日付は OBX-14 トランザクション日時（測定した日）とする。

（2）ファイル作成の単位は、データの格納構造として日付の下にあるため、最大でも一日分が１ファイルにまとまっている形とする。一日の中で測定のたびに作成するのでも良い。 一日１ファイルなら、特定キーは測定日を出力する。一日に複数回のデータを出力する場合は、特定キーに測定日の時間まで（YYYYMMDDHH）出力すること。

1.3.2　バイタルデータの項目及び形式等

（1）バイタルデータとして取得する項目は、「拡張期血圧、収縮期血圧、脈拍数、呼吸数、体温」の５項目とする。

（2）OBX-3 検査項目に出力するコードは JLAC10 コードとする。バイタルデータを参考に適切な JLAC10 を選択すること。

（３）上記以外の項目を SS-MIX2 に出力することは問題ないが、今回の対応では扱わない。但し、今後の検討で仕様として扱うことになる場合は、JLAC10 コードを基準とした標準コードを必須とすることを想定している。この今後想定される検査項目は別表として提供する。

1.3.3　標準コード変換機能

SS-MIX2 データの出力に際しては、コードのマッピング表などに従って、院内のローカルコードを厚労省が定める標準コードに変換する機能を有すること。またマッピング表については、容易にその内容を変更できるマスターメンテナンスプログラム等の機能を有すること。

JLAC10 コード、JANIS コード、HOT コードについては、機構病院が NCDA 事業に参加する場合においては機構から提供する。

1.3.4　標準化ストレージにおける文字コードについて

メッセージの文字コードについては、「標準化ストレージガイドライン」で示されているとおり、1 バイト系文字は ISO IR-6（ASCII）、2 バイト系文字は ISO IR87 （JIS X 0208 第一水準、第二水準）とする。 ただし現実には上記以外の文字コードが電子カルテシステムに登録されている可能性があるため、以下のように対応することとする。

1 半角カナ文字　→　全角カナ文字に置き換えて SS-MIX2 に出力する。
2 外字　→　■で置き換えて SS-MIX2 に出力する。
3 環境依存文字については変換表を機構より提供するのでそれにより変換して SS-MIX2 に出力する。

1.3.5　単位の文字表記の統一

SS-MIX2 データの出力に際して、臨床検査データの OBX セグメントの 6 フィールド目の単位の文字表記を統一すること。

【単位の文字表記の統一ルール例】ASCII コードで表記すること

・かける　→　.（ドット）
・乗 →　*（アスタリスク）
・μ →　u（小文字ユー）
・語尾に名称 →　（）で

・℃ → cel
・‰ → permil
・個 → pcs

【上記ルールの適用例】
・ｍＬ → mL （ASCII コード）
・X10＾2/μl → .10*2/uL （かける、乗、μ）
・/HPF → /(hpf) （語尾に名称）

1.3.6 単位変換機能

SS-MIX2 データの出力に際して臨床検査データの単位に関しては､JLAC10 コードごとに、機構が定める単位に変換を行った上で SS-MIX2 データを生成すること。 尚、JLAC10 コード別の単位表は別途機構から提供する。単位表は「SS-MIX2 標準化ストレージ仕様書 Ver.1.2」にも別表として添付する。

【単位変換例】

JLAC10 コード	数値	単位	→	JLAC10 コード	数値	単位
1A025000000127201	10.5	mg/l	→	1A025000000127201	1.05	mg/dL

1.3.7 計測値等の表記方法について

（1）定性値・検出限界以下・検出限界以上の表記

- OBX（検体検査結果）セグメントの５フィールド目（検査値）に検査結果を記述する場合、現在そのデータ形式は OBX-2 フィールドの説明にあるように NM 型、ST 型、CWE 型のうちいずれかの形式で記述することとなっている。
- 今回の仕様では、定性値・検出限界以下・検出限界以上のデータについては、SN 型の表現方法を用いて SN 型の”^”を” “（スペース）に置き換える。
- この件の説明は、「SS-MIX2 標準化ストレージ仕様書 Ver.1.2」 P104 表 3-77 検査結果セグメント（OBX）定義 の OBX-2 の項目説明にも記述する。

（2）複数の要素が一つの値で表現されている場合の表記

複数の要素が組み合わされ一つの結果値として表記されている場合は、それぞれの要素に分離して表記すること。例えば定量値とクラス値が組み合わされた結果値については、定量値とクラス値に分離する。

【定量値とクラス値の分離の例】

定量値とクラス値が組み合わされた例

検査名称	院内コード	結果値
ﾑﾝﾌﾟｽ Virus IgG	001591	2.3(±)
↓		
定量値とクラス値を分離した例		

SS-MIX2 標準コード	院内コード	結果値	備考
5F432143102302304	001591	2.3	
5F432143102302311	001591	+-	（半角スペース 2 つプラスマイナス）

1.3.8　トランザクションストレージのデータ保持期間

トランザクションストレージのデータ保持期間は、現在の標準化ストレージ及び拡張ストレージを作っているデータの再現に必要な分だけ保持しておくこと。

1.3.9 ST 型の長さ

- RXE-23(与薬速度)は ST 型で長さが 6 であるが、正負の記号と小数点を考慮し（例: +266.865）、本事業では 8 桁まで許容するものとする。

- CX 型は先頭成分が ST 型で長さが 15 であるが、IN1-10(被保険者グループ雇用者 ID)に長い名称の保険者が出力される場合などを考慮し、本事業では CX 型の先頭成分は 30 桁まで許容するものとする。
- XAD 型は第 8 成分(その他地理表示)が ST 型で長さが 50 であるが、全角 50 文字(100 バイト)と解釈しているシステムがあり半角文字で 100 文字登録出来るため、本事業では XAD 型の第 8 成分は 100 桁まで許容するものとする。

1.3.10 トランザクションストレージのファイル切り替え機能

SS-MIX2 の仕様上、トランザクョンストレージはカレントの日付が変わった時点、もしくは記録中のトランザクションデータファイルのファイルサイズが一定量を超えた時点で、新たなファイルを作成して記録先を切り替えるものとするとなっているが、同一日付内において一定時刻（例えば 17:00）を経過した時点で記録先を切り替える機能を追加する。

厚生労働科学研究補助金

（臨床研究等 ICT 基盤構築・人工知能研究事業）分担研究報告書

SS-MIX2 分析用データセットおよびシステムモジュールの作成・開発について

堀口　裕正 国立病院機構本部総合研究センター　診療情報分析部　副部長
岡田　千春　国立病院機構本部総合研究センター　企画役

研究要旨
本分担研究において、国立病院機構本部との調整を中心とした基盤構築を行った。まず、NDCA データの研究利用に向け、倫理審査申請に加えて、内部規定にて定められている内部委員会の調整を図った。また、閲覧・解析に特化した自然言語処理用の研究基盤の構築を行った。
また、NCDA データセットから、そのデータ仕様に基づいた匿名化モジュールの開発を行った。本年度、基本４情報を含む単独で個人情報とみなされる情報を削除するモジュールを開発し、そのモジュールを通過させた後に研究者に提供をおこなった。
また、実運用を想定し、SS―MIX の拡張ストレージから医師記録データを読み込み、それをサマリフォーマットに変換して次の処理に渡すことのできるモジュールの開発を行った。

A.目的

本研究は、電子カルテに実装可能な医療用語の標準化を行うシステムの開発に向け、そのステップとして、臨床現場からのニーズが極めて大きい退院サマリの自動生成技術の実現を目標と定める。電子カルテの自動解析は技術的な難易度が高く、実用的な精度を実現するためには多額の研究開発投資が求められる。そこで、本研究提案では、医療現場に直接的なメリットが生じる研究課題に取り組むことによって、現場の協力と今後の追加的な研究開発投資を呼び込み、その過程を通じて実用性の高い電子カルテの自動解析技術を実現する戦略を採る。

初年度、我々が今まで模擬カルテを用いて研究開発を進めてきた標準化技術を、国立病院機構の有する広域電子カルテ網(NCDA)上の実カルテへと適用し、技術的な課題を抽出する。2 年目には、NCDA を用いて集積した電子カルテに加えて、退院サマリ情報を用いることで、電子カルテの自動要約技術の検討を行う。3 年目においては、両技術の統合により、継続的な精度向上の体制を実現するとともに、研究成果を既存の各社電子カルテへと組み込む枠組みを構築する。

本研究により、退院サマリの自動要約技術や紹介状の作成支援技術等、医療用の自然言語処理に関連する多彩な応用技術が実現する。これは、医療現場における負担軽減策として極めて効果が期待される。また、こうした応用の発展により、要素技術である電子カルテ上の記載からの自動情報抽出において、継続的な精度向上が実現する。この手法は、電子カルテにおける用語の標準化技術単独に研究開発投資を行うことと比して、投資効率が極めて高いと考えられる。さらに、こうして医療用自然言語処理技術が発展することにより、大量の電子カルテからの効率的な情報抽出が実現する。これは健康医療政策に資する統計データの収集コストを劇的に低廉化し、今後、政策に求められる様々なエビデンスを継続的に生み出していく基盤となることが期待される。

なお本分担研究では、本研究における「大規模に電子カルテデータを入手できる体制」として、全国の国立病院 63 施設より年間 190 万患者の電子カルテ情報を自動収集する診療情報集積基盤（NCDA）を構築し、運用している基盤を用い、本研究目的のためのデータの収集・分析活動を行うためのシステム構築及び運用を行うことを１つ目の目的とする。また、各社電子カルテへと組み込む枠組みを構築するためのモジュール開発を行うことを２つ目の目的とする。

B.方法

国立病院機構本部との調整を中心とした基盤構築を行った。まず、NDCA データの研究利用に向け、倫理審査申請に加えて、内部規定にて定められている内部委員会での調整結果を踏まえてデータ抽出機能の調整を行い、閲覧・解析に特化した自然言語処理用の研究基盤の構築を行うこととした。その上で、本研究を実施するのに不可欠な NCDA データセットから、そのデータ仕様に基づいた匿名化モジュールの開発を行い、運用を行うこととする。

さらに、各社電子カルテへと組み込む枠組みを構築するためのモジュール開発として

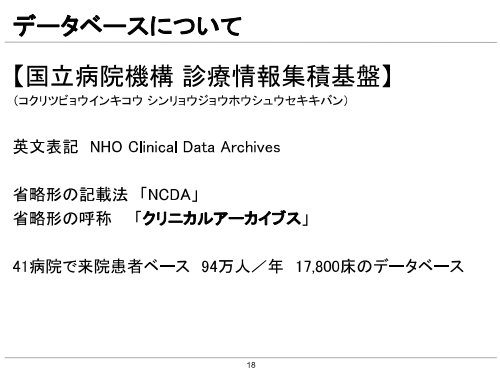
NCDA が各電子カルテベンダーに要求している SS-MIX 拡張ストレージへの医師記録の書き出しを起点として情報を取得し、処理後同じく NCDA が各電子カルテベンダーに要求している SS-MIX 拡張ストレージへの退院サマリの格納場所に要求するフォーマットで書き戻すシステムを想定し、それらに必要な病院内で必要なモジュールの開発を行った。

C.結果

Ⅰ研究用データセットの作成・運用について

国立病院機構が平成 27 年度に構築した NCDA データベースは、平成 30 年度末現在 63 病院が参加、約 27000 床、年間実患者数約 190 万人のデータベースであり、診療日翌日には本部のデータベースに検査値や投薬の情報を含む診療データが届くことになっている。

まだ、本年度運用３年目を迎え、今後、MIA のデータベースで今まで実践してきた分析調査を代替できるポテンシャルを持っている。（参考資料に詳細を添付している）



また、本研究で中心的に使われる医師記録等（経過記録・退院サマリ）については、SS-MIX2 の標準仕様に含まれていないが、JAHIS の提供している仕様を参考に、資料 1 及び 2 で示した仕様で NCDA 内に実装することとした。

本研究はカルテの非定型の記載欄に記入されたデータを使うという研究であり、患者の不利益等を防止するために倫理的な配慮をした上で、倫理審査を受けなければならない。平成 29 年 1 月に国立病院機構中央倫理委員会に侵襲・介入なしの観察研究として倫理審査の申請を行い、3 月に承認された。倫理審査の承認後、データ利用に際して必要な国立病院機構内のデータベース利活用審査委員会への利活用申請を行い、3 月にその承認も受けた。倫理審査申請書については資料 3 に示す。

なお、NCDA データは国立病院機構が契約するデータセンター内で厳重に管理されている。研究に際しては、このデータベースから研究テーマごとに匿名化したサブセットを切り出し、国立病院機構本部内のオンサイト利用に限っている。以上により、データセットの利用対象と利用目的を厳しく制限することにより、患者個人情報の保護を行っている。

また、NCDA データセットから、そのデータ仕様に基づいた匿名化モジュールの開発を行った。本年度、基本４情報を含む単独で個人情報とみなされる情報を削除するモジュールを開発し、そのモジュールを通過させた後に他分担研究者に提供出来るようになった。

Ⅱ病院での運用をイメージしたシステムモジュールの開発

本研究課題は研究の公募段階から全国の病院であまねく運用可能なシステムの構築を求められていた。この分担研究においては上記の条件を以下の方法で実現することとし、その実現に必要なモジュールを開発した。

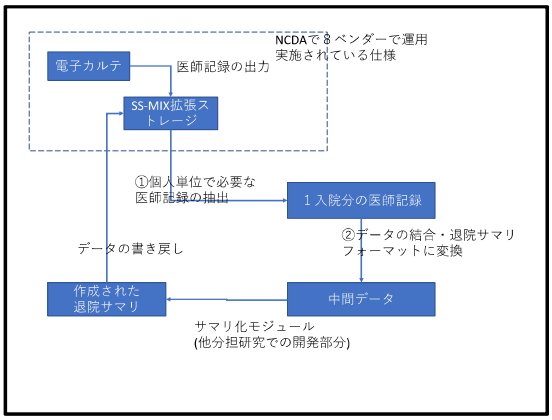
(1) 全国の病院のシステムの接続方式について

本研究では全病院で運用可能な接続として、SS-MIX の拡張ストレージをキーとした運用を採用することとした。

前節で説明をした NCDA における医師記録等の拡張ストレージに対する出力は JAHIS の規約通りである・すでに８ベンダー63 病院で実装済みで、現時点でもその病院数を NCDA 内で増加させることに大きな問題を起こしていない。よってこの SS-MIX を利用する方式は研究の公募段階で要求されていた仕様を満たすものと考えられる。

(2) 本研究で開発を行ったモジュールについて

本研究にて想定したモジュール群の概念図を図に示す。この図の形でシステムを実現するものとし、その中の①SS-MIX 拡張ストレージから必要な医師記録データを抽出するモジュール、②抽出したデータを退院サマリフォーマットに変換・結合するモジュールについて開発を行った。



実際に開発するシステム内で組み込みが容易なように実装は Windows 環境下においてコマンドで動作するものとし、go にて開発を行っている。

それぞれのコマンドの Usage は以下の通りである

① のモジュール

ssmix2extf pcopy

Usage:

ssmix2extf pcopy [flags] nhoid

Flags:

-u, --duration string duration for search condition

-e, --encoding string encoding of logging format (json | console) (default "json")

-h, --help help for pcopy

-d, --logfiledirectory string the logfile directory

-l, --logfilename string the

logfile name (default "pcopy.log")
  -c, --lsoutput string csv file which contain copied files (default "lsoutput.csv")
  -o, --outputdir string output directory (default "pcopyout")
  -p, --patientidlistfile string the file which contain a list of patient id
  -v, --verbose console verbose mode

② のモジュール

ssmix2extf summary

Usage:
  ssmix2extf summary [flags] input.csv

Flags:
  -h, --help help for summary
  -o, --outputdir string output directory (default "summaryout")

モジュールのバイナリおよびソースコードについては https://github.com/nhoHQ **内で公表予定である。**

E.結論

本年度、本研究の他の分担研究を支える分析基盤をきちんと整備・安定した運用を実施することができた。また、本研究課題は研究の公募段階から全国の病院であまねく運用可能なシステムの構築を求められてい他部分について、その実現に必要なモジュールを開発した。

資料１　NCDA における医師記録等の仕様書

趣旨

本事業では、各社の SS-MIX2 モジュールの拡張ストレージへの出力機能を利用し、以下の情報を出力することを求めている。その際、SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d（以下、ガイドライン）に記載している仕様に対応していること。また、トランザクションストレージ、インデックスデータベースも同時に生成すること。

- 経過記録
- 退院時サマリー
- 診療情報提供書

以下に仕様を示す。

ドキュメントデータ 物理構造

```
|-- 拡張ストレージ ルートフォルダ
    |-- 患者 ID 先頭 3 文字
        |-- 患者 ID 4〜6 文字
            |-- 患者 ID
                |-- 診療日
                    |-- データ種別
                        |-- コンテンツフォルダ
                            |--主文書ファイル
```

診療日

特に指定しない。

データ種別

ガイドライン P4 (4)「データ種別フォルダ」について に則ること。

```
[ローカル文書コード]^ローカル文書名称^[ローカルコード体系コード]^標準文書コード^標準文書名称^標準コード体系コード
```

以下のように標準コードに対しローカルコードが複数あることは許容される。

```
L12345^ 入院診療録^99ZZZ^11506 -3^経過記録^LN
L12346^ 外来診療録^99ZZZ^11506 -3^経過記録^LN
```

コンテンツフォルダ

ガイドライン Ver.1.2d P5 (5)「コンテンツフォルダ」について に則ること。

患者 ID_診療日_データ種別コード_特定キー_発生日時_診療科コード_コンディションフラグ
いずれの文書も削除は想定していないが、電子カルテシステムによっては修正はあり得ると考える。その場合、ガイドライン P6 ④修正が発生する場合 に則り改版すること。

主文書ファイル

XML CDA R2 で出力すること。XML ファイル以外に画像ファイルや CSS ファイル等を出力してもかまわない。

HEADER 部

いずれの文書も JAHIS　診療文書構造化記述規約 共通編 Ver.1.0 に則ること。
P27 6.3.11.検査・診療等行為 "documentationOf/ServiceEvent" によると、
documentationOf の制約・多重度は 0..1 となっているが、 経過記録、退院時サマリについてはこれを 1..1 と読み替えること。
経過記録は serviceEvent classCode(サービスイベントクラスコード)を ENC(診察)とし、effectiveTime(実施日)は low value、high value ともに記録タイミングを出力すること。
退院時サマリは serviceEvent classCode(サービスイベントクラスコード)を ACCM(入院、滞在)とし、effectiveTime(実施日)は low value に入院タイミング、high value に退院タイミングを出力すること。
タイミングの粒度は日以上であれば良い。

BODY 部

診療情報提供書は、日本 HL7 協会 患者診療情報提供書 規格 Ver.1.00 に則ること。
診療情報提供書以外は、XML の文法に則ること

２．NCDA システム仕様書

SS-MIX2 を用いた診療情報データベース構築の為の SS-MIX2 モジュール技術仕様書

1. システム要件

国立病院機構の各病院にて「国立病院機構診療情報分析基盤(NCDA)」に参加する為に調達する SS-MIX２モジュールの機能は以下の通りである。 但し、本体の電子カルテシステム等の仕様上、作成が不可能であるものについては作成を要しない。その場合、何が不可能かを導入標準作業手順書に記載すること。

1.1 SS-MIX2 Ver.1.2d 機能

SS-MIX2 Ver.1.2d に準拠することとして、以下の機能を有すること。

- 日本医療情報学会発行の「SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン Ver.1.2d」、「SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d」、「SS-MIX2 標準化ストレージ仕様書 Ver.1.2d」、「標準化ストレージ仕様書別紙：コード表 Ver.1.2d」、「SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d 別紙：標準文書コード表」に記載している仕様に対応していること。（尚、当初 Ver.1.2c 準拠としていたが、標準ストレージ部分では Ver.1.2c からの変更点について影響がないため Ver.1.2d 準拠ということとした。）
- 標準化ストレージ、拡張ストレージ、トランザクションストレージ、インデックスデータベースの４つのファイルを生成すること。
- 標準化ストレージにはデータ種別として 36 種のデータを出力すること。

（表 1-1　標準化ストレージ格納データ）

No	データ種別	種別名称	HL7 メッセージ型
1	ADT-00	患者基本情報の更新	ADT^A08

No	データ種別	種別名称	HL7 メッセージ型
2	ADT-00	患者基本情報の削除	ADT^A23
3	ADT-01	担当医の変更	ADT^A54
4	ADT-01	担当医の取消	ADT^A55
5	ADT-12	外来診察の受付	ADT^A04
6	ADT-21	入院予定	ADT^A14
7	ADT-21	入院予定の取消	ADT^A27
8	ADT-22	入院実施	ADT^A01
9	ADT-22	入院実施の取消	ADT^A11
10	ADT-31	外出泊実施	ADT^A21
11	ADT-31	外出泊実施の取消	ADT^A52
12	ADT-32	外出泊帰院実施	ADT^A22
13	ADT-32	外出泊帰院実施の取消	ADT^A53
14	ADT-41	転科・転棟(転室・転床)予定	ADT^A15
15	ADT-41	転科・転棟(転室・転床)予定の取消	ADT^A26

No	データ種別	種別名称	HL7 メッセージ型
16	ADT-42	転科・転棟(転室・転床)実施	ADT^A02
17	ADT-42	転科・転棟(転室・転床)実施の取消	ADT^A12
18	ADT-51	退院予定	ADT^A16
19	ADT-51	退院予定の取消	ADT^A25
20	ADT-52	退院実施	ADT^A03
21	ADT-52	退院実施の取消	ADT^A13
22	ADT-61	アレルギー情報の登録／更新	ADT^A60
23	PPR-01	病名（歴）情報の登録／更新	PPR^ZD1
24	OMD	食事オーダ	OMD^O03
25	OMP-01	処方オーダ	RDE^O11
26	OMP-11	処方実施通知	RAS^O17
27	OMP-02	注射オーダ	RDE^O11
28	OMP-12	注射実施通知	RAS^O17
29	OML-01	検体検査オーダ	OML^O33

No	データ種別	種別名称	HL7 メッセージ型
30	OML-11	検体検査結果通知	OUL^R22
31	OMG-01	放射線検査オーダ	OMG^O19
32	OMG-11	放射線検査の実施通知	OMI^Z23
33	OMG-02	内視鏡検査オーダ	OMG^O19
34	OMG-12	内視鏡検査の実施通知	OMI^Z23
35	OMG-03	生理検査オーダ	OMG^O19
36	OMG-13	生理検査結果通知	ORU^R01

「SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン Ver.1.2d　p11」

1.2 拡張ストレージへの出力機能

現在の SS-MIX2 モジュールでオプションとして既に導入している拡張ストレージへの出力機能は、そのまま提供すること。また、1.3.0 で規定する出力を行うこと。

1.3　NHO 対応としての設定

1.3.0 拡張ストレージへの出力機能

各社の SS-MIX2 モジュールの拡張ストレージへの出力機能を利用し、以下の情報を出力すること。その際、 日本医療情報学会発行の「SS-MIX2 拡張ストレージ構成の説明と構築ガイドライン Ver.1.2d」に記載している仕様に対応していること。また、トランザクションストレージ、インデックスデータベースも同時に生成すること。

No	データ種別	種別名称	HL7 メッセージ型
1	L-OBSERVATIONS^OBSERVATIONS^99ZL01	バイタル検査結果	HL7 V2.5 ORU^R30
2	^（ローカル名称）^^11506-3^経過記録^LN	診療録（外来/入院含む）	HL7 CDA R2
2.1	^（ローカル名称）^^34108-1^外来診療録^LN	診療録（外来）（入院・外来が別の場合）	HL7 CDA R2
2.2	^（ローカル名称）^^34112-3^入院診療録^LN	診療録（入院）（入院・外来が別の場合）	HL7 CDA R2
3	^（ローカル名称）^^18842-5^退院時サマリー^LN	退院時サマリー	HL7 CDA R2
4	^（ローカル名称）^^57133-1^紹介状^LN	診療情報提供書	HL7 CDA R2

1.3.1　バイタル検査結果通知の出力

（１）バイタル検査結果通知のデータを、別紙の形式で拡張ストレージに出力する。
尚、「診療日」に出力する日付は OBX-14 トランザクション日時（測定した日）とする。

（２）ファイル作成の単位は、データの格納構造として日付の下にあるため、最大でも一日分が１ファイルにまとまっている形とする。一日の中で測定のたびに作成するのでも良い。 一日１ファイルなら、特定キーは測定日を出力する。一日に複数回のデータを出力する場合は、特定キーに測定日の時間まで（YYYYMMDDHH）出力すること。

1.3.2　バイタルデータの項目及び形式等

（1）バイタルデータとして取得する項目は、「拡張期血圧、収縮期血圧、脈拍数、呼吸数、体温」の５項目とする。

（2）OBX-3 検査項目に出力するコードは JLAC10 コードとする。バイタルデータを参考に適切な JLAC10 を選択すること。

（3）上記以外の項目を SS-MIX2 に出力することは問題ないが、今回の対応では扱わない。但し、今後の検討で仕様として扱うことになる場合は、JLAC10 コードを基準とした標準コードを必須とすることを想定している。この今後想定される検査項目は別表として提供する。

1.3.3　標準コード変換機能

SS-MIX2 データの出力に際しては、コードのマッピング表などに従って、院内のローカルコードを厚労省が定める標準コードに変換する機能を有すること。またマッピング表については、容易にその内容を変更できるマスターメンテナンスプログラム等の機能を有すること。

JLAC10 コード、JANIS コード、HOT コードについては、機構病院が NCDA 事業に参加する場合においては機構から提供する。

1.3.4　標準化ストレージにおける文字コードについて

メッセージの文字コードについては、「標準化ストレージガイドライン」で示されているとおり、1 バイト系文字は ISO IR-6（ASCII）、2 バイト系文字は ISO IR87 （JIS X 0208 第一水準、第二水準）とする。 ただし現実には上記以外の文字コードが電子カルテシステムに登録されている可能性があるため、以下のように対応することとする。

1 半角カナ文字　→　全角カナ文字に置き換えて SS-MIX2 に出力する。
2 外字　→　■で置き換えて SS-MIX2 に出力する。
3 環境依存文字については変換表を機構より提供するのでそれにより変換して SS-MIX2 に出力する。

1.3.5　単位の文字表記の統一

SS-MIX2 データの出力に際して、臨床検査データの OBX セグメントの 6 フィールド目の単位の文字表記を統一すること。

【単位の文字表記の統一ルール例】ASCII コードで表記すること

・かける　→　.（ドット）
・乗 →　*（アスタリスク）
・μ →　u（小文字ユー）
・語尾に名称 →　（）で
・℃ →　cel
・‰ →　permil
・個 →　pcs

【上記ルールの適用例】

・mL　→ mL　（ASCII コード）
・X10＾2/μl → .10*2/uL　（かける、乗、μ）
・/HPF　→ /(hpf)　（語尾に名称）

1.3.6 単位変換機能

SS-MIX2 データの出力に際して臨床検査データの単位に関しては、JLAC10 コードごとに、機構が定める単位に変換を行った上で SS-MIX2 データを生成すること。 尚、JLAC10 コード別の単位表は別途機構から提供する。単位表は「SS-MIX2 標準化ストレージ仕様書 Ver.1.2」にも別表として添付する。

【単位変換例】

JLAC10 コード	数値	単位	→	JLAC10 コード	数値	単位
1A025000000127201	10.5	mg/l	→	1A025000000127201	1.05	mg/dL

1.3.7　計測値等の表記方法について

（1）定性値・検出限界以下・検出限界以上の表記

- OBX（検体検査結果）セグメントの５フィールド目（検査値）に検査結果を記述する場合、現在そのデータ形式は OBX-2 フィールドの説明にあるように NM 型、ST 型、CWE 型のうちいずれかの形式で記述することとなっている。
- 今回の仕様では、定性値・検出限界以下・検出限界以上のデータについては、SN 型の表現方法を用いて SN 型の”^”を” “（スペース）に置き換える。
- この件の説明は、「SS-MIX2 標準化ストレージ仕様書 Ver.1.2」　P104 表 3-77 検査結果セグメント（OBX）定義　の OBX-2 の項目説明にも記述する。

（2）複数の要素が一つの値で表現されている場合の表記

複数の要素が組み合わされ一つの結果値として表記されている場合は、それぞれの要素に分離して表記すること。例えば定量値とクラス値が組み合わされた結果値については、定量値とクラス値に分離する。

【定量値とクラス値の分離の例】

定量値とクラス値が組み合わされた例

検査名称	院内コード	結果値
ﾑﾝﾌﾟｽ Virus IgG	001591	2.3(±)
↓		
定量値とクラス値を分離した例		

SS-MIX2 標準コード	院内コード	結果値	備考
5F432143102302304	001591	2.3	
5F432143102302311	001591	+-	（半角スペース２つプラスマイナス）

1.3.8　トランザクションストレージのデータ保持期間

トランザクションストレージのデータ保持期間は、現在の標準化ストレージ及び拡張ストレージを作っているデータの再現に必要な分だけ保持しておくこと。

1.3.9 ST 型の長さ

- RXE-23(与薬速度)は ST 型で長さが 6 であるが、正負の記号と小数点を考慮し（例: +266.865)、本事業では 8 桁まで許容するものとする。
- CX 型は先頭成分が ST 型で長さが 15 であるが、IN1-10(被保険者グループ雇用者 ID)に長い名称の保険者が出力される場合などを考慮し、本事業では CX 型の先頭成分は 30 桁まで許容するものとする。
- XAD 型は第 8 成分(その他地理表示)が ST 型で長さが 50 であるが、全角 50 文字(100 バイト)と解釈しているシステムがあり半角文字で 100 文字登録出来るため、本事業では XAD 型の第 8 成分は 100 桁まで許容するものとする。

1.3.10 トランザクションストレージのファイル切り替え機能

SS-MIX2 の仕様上、トランザクションストレージはカレントの日付が変わった時点、もしくは記録中のトランザクションデータファイルのファイルサイズが一定量を超えた時点で、新たなファイルを作成して記録先を切り替えるものとするとなっているが、同一日付内において一定時刻（例えば 17:00）を経過した時点で記録先を切り替える機能を追加する。

３．倫理審査における計画書

電子カルテ情報をセマンティクス（意味・内容）の標準化により分析可能なデータに変換するための研究

研究責任者：堀口　裕正
独立行政法人国立病院機構本部　総合研究センター
診療情報分析部　副部長

事務局/研究主催
独立行政法人国立病院機構本部　総合研究センター
診療情報分析部
堀口　水本
〒152-8621　目黒区東が丘２‐５‐21
TEL: 03-5712-5133

FAX: 03-5712-5134
E-Mail：horiguchi-hiromasa@hosp.go.jp

第 1.0 版：2017 年 1 月 18 日

１．背景

本研究では、電子カルテに実装可能な医療用語の標準化を行うシステムの開発に向け、そのステップとして、臨床現場からのニーズが極めて大きい退院サマリの自動生成技術の実現を目的とする。これは用語の標準化を目的とする研究として遠回りな課題設定である。しかし、電子カルテの自動解析は技術的な難易度が高く、実用的な精度を実現するためには多額の研究開発投資が求められる。そこで、本研究提案では、医療現場に直接的なメリットが生じる研究課題に取り組むことによって、現場の協力と今後の追加的な研究開発投資を呼び込み、その過程を通じて実用性の高い電子カルテの自動解析技術を実現する。

２．目的

本研究は、電子カルテに実装可能な医療用語の標準化を行うシステムの開発に向け、そのステップとして、臨床現場からのニーズが極めて大きい**退院サマリの自動生成技術の実現**を目的とする

３．研究方法

３－１．研究実施場所

研究実施場所は、国立病院機構本部総合研究センター診療情報分析部（以下、診療情報分析部）研究室及び本部内分析室並びに静岡大学情報学部行動情報学科狩野研究室、岡山大学大学院医歯薬学総合研究科クリニカルバイオバンクネットワーキング事業化研究講座研究室、国立保健医療科学院研究情報支援研究センター研究室とする。

３－２．研究実施期間

研究実施期間は、倫理審査委員会承認後より 2020 年 3 月 31 日までとする。

３－３．研究対象医療機関と対象患者

研究対象医療機関は、国立病院機構病院に所属する DPC 病院のうち、診療情報集積基盤（以下、NCDA）を運用しデータ提供を行う医療機関とする。

対象患者は 2016 年 1 月 1 日から 2019 年 12 月 31 日までに入院し，退院時サマリを作成した全患者とする。

３－４．対象データ

研究に用いるデータは、研究対象医療機関より診療情報分析部に提供された DPC データおよびレセプトデータ、ならびに SS-MIX2 ストレージに格納された情報から抽出した医師記録、退院サマリおよび入院中の検査結果、食事内容および処方内容である。

３－５．分析方法

(1) 対象

退院サマリを作成した全患者

(2) アウトカム

入院中に記載/記録された情報から退院サマリを自動生成する技術を開発すること

(3) 抽出する項目

入院中の医師記録・退院サマリ・入院中の検査結果、食事内容および処方内容

(4) 解析方法

入院中に記載/記録された情報を元データに、機械学習により自動的に情報収集を行い、退院サマリを自動で作成する。その作成結果と、実際の医師の書いた退院時サマリを比較/検討を行い、自動作成技術の能力評価を行い、またその能力の改善を行っていく。

４．倫理的配慮

本研究は、ヘルシンキ宣言、人を対象とする医学系研究に関する倫理指針

（以下、倫理指針）に基づいて実施する。

４－１．インフォームド・コンセント

本研究は既存試料・情報を用いて実施し、人体から取得された試料は用いない。研究対象者等からインフォームド・コンセントは受けないが、倫理指針「第 12 の１ (2) イ」に則り、本計画書の４－３に記す通り、利用目的を含む本研究についての情報を研究対象者等に公開し、研究が実施されることについて研究対象者が拒否できる機会を保障する。なお、NCDA 運用による診療情報の蓄積・利活用についての説明及び同意は、各施設での掲示で既に行われている。

４－２．データ管理、個人情報等の取り扱いに関する配慮

研究の実施並びに種々のデータの収集及び取り扱いにおいては、国立病院機構診療情報データベース利活用規程に従うとともに、患者情報の機密保持に充分留意する。

本研究で用いるデータは、研究対象医療機関に 2016 年 1 月 1 日から 2019 年 12 月 31 日までに退院サマリを作成した全患者のデータであり、個人情報等を取り扱う。倫理指針「第 15 の２ (1)」及び国立病院機構診療情報データベース利活用規程に則り、保有する個人情報等について、漏えい、滅失又はき損の防止その他の安全管理のため、下記の措置を講じる。

データは研究対象医療機関で収集され、本部 IT 推進部に提出される。データが保管されるサーバーを国立病院機構本部２階のセキュリティルームに設置し、セキュリティルーム内で IT 推進部システム開発専門職が匿名化処理を行う。研究者は匿名化後のデータを用いて本部内分析室において分析を実施する。

保有する個人情報に関する事項の公表等については、倫理指針「第 12 の１ (2) イ」、「第 16 の１ (1)」及び国立病院機構診療情報データベース利活用規程第 6 条第 3 項に則り、個人情報の取扱いを含む研究の実施についての情報を研究対象者等に公開する。

４－３．本研究における情報公開

本研究では、倫理審査委員会承認後、倫理指針「第 12 の１ (2) イ」、「第 16

の１ (1)及び国立病院機構診療情報データベース利活用規程第 6 条第 3 項に則り、本部ホームページにおいて、本研究の意義、目的及び方法、研究機関、保有する個人情報に関して利用目的の通知、開示、訂正等又は利用停止の求めに応じる手続き並びに保有する個人情報に関する問い合わせや苦情等の窓口の連絡先に関する情報を公開する（公表する情報については別添資料を参照）。

４－４．研究成果の公表

本研究の成果は、報告書で公表するとともに、学会・論文で発表する。また、本研究結果を内包したソフトウェアの公表を実施する。データの集計・分析結果については、集団を記述した数値データもしくは機械学習の学習結果データとし、個人が同定されるデータの公表は行わない。

５．研究経費

本研究は、厚生労働科学研究費補助金（臨床研究等ＩＣＴ基盤構築研究事業）「電子カルテ情報をセマンティクス（意味・内容）の標準化により分析可能なデータに変換するための研究」（代表　堀口裕正）を用いて研究を実施する

６．研究組織

総合研究センター診療情報分析部が主体となり、本部医療部、保険医療科学院、静岡大学、岡山大学等から協力を得て、研究を行う。

【研究代表者】
国立病院機構本部総合研究センター診療情報分析部　　副部長 堀口 裕正

【共同研究者】

国立病院機構本部　企画役　岡田 千春
静岡大学情報学部行動情報学科　准教授　狩野 芳伸
岡山大学大学院医歯薬学総合研究科
クリニカルバイオバンクネットワーキング
事業化研究講座研究室　准教授　森田 瑞樹
国立保健医療科学院研究情報支援研究センター
特命上席主任研究官　奥村 貴史

別添

「電子カルテ情報をセマンティクス（意味・内容）の標準化により分析可能なデータに変換するための研究」研究実施に関するお知らせ

厚生労働科学研究費補助金（臨床研究等 ICT 基盤構築研究事業）
分担研究報告書

退院サマリの自動生成研究における研究基盤の整備

研究分担者　奥村 貴史
（北見工業大学 工学部・大学院工学研究科 教授）

研究要旨

臨床医は、入院治療をしていた患者が退院する際、それまでに記載していた入院カルテから退院サマリを作成する。この退院サマリの作成を効率化することができれば、医師の診療負担を直接軽減することが出来ることに加えて、様々な副次的な効果が期待される。そこで本研究分担では、この入院カルテの自動要約に向けた研究基盤の整備に取り組んできた。

今年度は、昨年度までの研究をさらに発展させ、4 つの課題に取り組んだ。まず、一連の研究には自由に研究利用できるカルテの存在が必要となる。そこで、ダミーカルテの収集を進め、100 件の整備を目指して活動を進めた。また、収集したダミーカルテを対象として、機械的な処理を可能とするためのアノテーション作業に取り組んだ。その際、初年度に行ったアノテーション、昨年度に試みたアノテーションを踏まえ、さらなる改善を図った。次に、このアノテーション作業と平行して、退院サマリの分析モデルであり生成モデルである「CASE モデル」の改善に取り組んだ。また、退院サマリの自動生成処理を精度管理するうえで必要となる「理想の退院サマリ」の確保に向けた検討を行った。

研究活動の結果、ダミーカルテは、目標を超える 108 件を集めることが出来た。また、これらのカルテを対象としたアノテーションを進めると共に、アノテーションガイドラインを高品質化することが出来た。さらに、このアノテーション済みカルテを用いて実現する退院サマリの分析に向けたモデルとして、修正 CASE モデルを提示することが出来た。理想のサマリに向けた検討では、高品質な退院サマリを低コストに実現するための作業仮説を整理すると共に、実証に向けた研究デザインを策定することが出来た。今後、これらの研究基盤をベースとして、退院サマリの自動生成に向けたさらなる研究の発展を図りたい。

A．研究目的

医師が患者を診察した際には、その情報を遅滞なく診療録に記載する必要がある(医師法第二十四条)。これがいわゆるカルテであるが、カルテには大きく 2 つの種類がある。1 つは「外来カルテ」であり、外来診療において継続的に記載されていく。もう 1 つが「入院カルテ」であり、患者が入院する毎に作成され、入院中の経過が退院に至るまで記載されることになる。入院カルテは、回診毎に記載されることも少なくなく、外来カルテとは記載の密度が大きく異なるうえ、長期入院などにおいては分量も大きなものとなる。そこで医療機関は、入院患者が退院した時点で入院中の経過を要約した文書を作成する。これが、「退院サマリ」である。

この退院サマリの作成は医師にとって負担であるため、退院サマリの作成を自動化することができれば医師の勤務負担を軽減することができる。そこで本研究分担では、この入院カルテの自動要約に向けた研究基盤の整備に取り組んできた。初年度においては、自動要約を実現するための処理モデル(CASE モデル)の提案を行った。CASE モデルは、退院サマリ中に含まれるセンテンス(文)が、「元カルテに由来するか否か」、「抽象度が高いか低いか」という2軸により分類しうると仮定したモデルである(図 1)。その有用性を示すためには、そもそも実際の退院サマリが本当にこの2軸分類により効果的に分類されるセンテンスにより成り立つのかを、実際の退院サマリの分析を通じて実証する必要がある。そこで2年目に、このモデルの実証に向けた研究基盤の構築を進めた。まず、一連の実証に求められる「ダミーカルテ」の収集を行った。また、退院サマリに含まれるセンテンスの自動分類を実現する分類器の構築を目指した。さらに、当該モデルに基づいて退院サマリを生成するための、ユーザーインターフェースの開発を行った。

3年目は、上記研究のさらなる進展に取り組んだ。ダミーカルテは、一定数揃わなければ有効活用することができない。そこで、研究協力者の医師を募り 100 件の突破を目指して活動を進めた。また、このカルテを研究活用していくうえでは、カルテに含まれる文章に対して、機械処理が可能となるようにその解釈情報を付与していく必要がある。この「アノテーション」を、カルテに含まれる全てのセンテンスに対して行うことには多くの手間が掛かるが、今までに付与したアノテーション情報に技術的な課題が生じていることが明らかとなった。そこで、アノテーションの基準を再考したうえで、全データを対象に再度のアノテーションを行った。さらに、これらの検討を通じて、CASE モデルそのものにも課題が明らかとなったことから、CASE モデルの修正を合わせて行った。

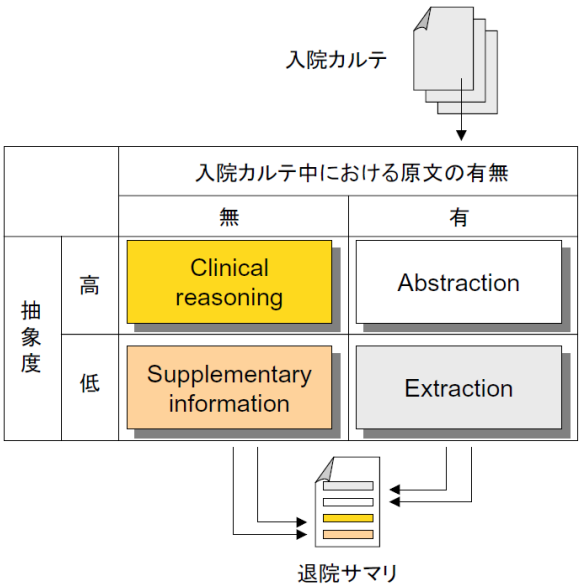


図 1　CASE モデル

本研究班では、上記のように、退院サマリの自動生成研究における研究基盤の整備を進めてきた。これによって、退院サマリの自動生成研究に関する各種の研究が進むことが期待される。一方で、退院サマリを自動生成したとしても、その精度が高まらない限りは有用な技術とすることができない。そのためには、生成した要約を客観的に評価できる必要がある。もし、個々の入院カルテに対して「理想的なサマリ」を用意することができれば、今後、生成したサマリを理想的なサマリとの差異によって評価することが可能となりうる。さらに、理想的な要約を処理系に学習させることで、さらなる精度向上に繋がる可能性もある。そこで、研究基盤整備の一環として、入院カルテから理想的な退院サマリを生成する手法についての検討を行った。

B．研究方法

ダミーカルテ収集

ダミーカルテの収集には、多くの手間が掛かる。カルテを記載するには、医師の協力が必要となるが、多くの医師は、ひたすらデータを生成するような単純作業に興

味を示さない。また、医師の雇用には大きなコストが掛かる。医療用自然言語処理研究の発展が望まれているが、研究分野の停滞の一端は、この、研究利用できるデータの確保に多大な手間を要する点がある。この隘路を突破するためには、本研究の価値を理解し、データ生成に関わって頂ける研究協力医師を見つける必要がある。そこで、医師向けセミナー等にて研究状況を発表すると共に協力を呼びかけ、反応を示して下さった医師と詳細なやり取りを通じて研究協力頂く活動を続けた。医師の負担を少しでも減ずるため、提出データにおける課題の修正や欠損の補填は、別途、看護師の研究協力者へと依頼し、医師にはその作業結果の確認を依頼する形で作業を進めた。提出頂いたダミーカルテは、こうして内容をチェックしたうえで、その後の研究利用に適した状態へとするため、年齢、性別、主病名等の定型情報を整理し、XML と呼ばれるフォーマットへと整理した。

カルテのアノテーション

また、協力医師による尽力によりダミーカルテを数多く用意することができたが、このカルテを研究活用していくうえでは、カルテに含まれる文章に対して、その解釈情報を付与していく「アノテーション」作業が必要となる。昨年までの研究においては、CASE モデルの実証に向けて、カルテ中の全てのセンテンスに対して「事実度」の情報を付与していた。ここで事実度とは、CASE モデルが想定する抽象度と逆の関係にある指標とする。そして、各センテンスに、評価者の判断に基づく「事実度の高低」に関するラベルと、そのセンテンスが肯定か否定かのいずれに関するものかの「極性情報」、その言及が確実なものか不確実なものかに関する「確信度」のラベルを付与していた。しかし、このラベルは、各センテンスを構文解析したうえで節レベルで評価したものでないことから、接続詞によって、1 センテンス中に複数の極性や複数の事実度が混在する問題が生じていた。また、当該データを元にして試行的な分類器の作成を行ってみたものの、高い分類精度を出すことができずにいた。そこで、アノテーションの基準を再考したうえで、全データを対象に再度のアノテーションを試みた。

CASE モデル検討

本研究において当初提案した CASE モデルは、退院サマリ中の各センテンスを、「元カルテに由来するか否か」、「抽象度が高いか低いか」という 2 軸により分類した。このモデルには全センテンスを 2×2 表のいずれかに落とし込むことが出来る分類モデルであることに加えて、4 つに分類されるセンテンスをいかに生成するかという生成モデルを兼ねている。たとえば、Clinical Reasoning 文は、元のカルテに由来せず事実度の低いセンテンスを指す。このカテゴリの文は元カルテからは生成しえないが、カルテ作成者の他のカルテか同一診療科の他の医師のカルテより類似したセンテンスを拾える可能性がある。Assessment 文は、当該センテンスに近い文を元のカルテに見出しうる、事実度の低い文であり、元のカルテに記載されている事実度の高い文、すなわち患者病態等を評価した文であることになる。したがって、要約に際した手法としては、Extractive(抜粋的)な要約ではなく、Abstractive(抽象化的)な要約により生成されることが示唆される。しかし、入院カルテにおいても、SOAP モデル(カルテ記載において Subjective, Objective, Assessment, Plan の各カテゴリ毎に整理する手法)により記載している限り、Assessment 部分や Plan 部分には事実度の低い評価や将来計画が記載されることになる。これらは、退院サマリにおいて、Extractive に抽出され記載される可能性があることから、CASE モデルの前提が崩れかねない状況が生じていた。そこで、CASE モデルを拡張し、入院カルテと退院サマリの関係を整理したうえで、必要な検討を行った。

理想のサマリ研究

個々の入院カルテに対して「理想的なサマリ」を用意するためには、何が理想的なサマリなのかの定義をする必要がある。ただし、言及すべき事項を整理することをもって理想のサマリの定義とすると、疾患によってカルテに記載するべき内容も異なるため、疾患毎に理想の内容を定義する必要が生じ作業量が膨大となる。また、たとえそのような定義を行ったとしても、患者には個体差があることに加えて、それぞれ合併症や同時に罹患している他の疾患がある。さらに、入院日数も異なれば、そもそもの入院カルテの質や量にも違いがある。したがって、理想のサマリを内容によって定義しようとすると、そもそも「理想の入院カルテ」を用意する必要が生じ、医療現場の負担軽減を目指してきた本研究の方針にも反する結果となってしまう。医療現場の負担軽減を図るためには、当面の間は、現場において記載された現状の入院カルテを対象として理想的な退院サマリを生成する必要がある。

そこで、我々は、理想のサマリの定義として、何が書かれているかの内容に関する定義は行わず、代わりに、「入院カルテと退院サマリを第３者の医師に添削させ、その結果をさらに別の医師に添削させるという作業を繰り返し、修正が行われなくなったサマリ」を、理想のカルテであると定義した。これにより、疾患の種別や患者側の既往歴や個人差、入院日数の違いなどに関わらず、理想的なカルテを客観的な形で定義することが実現する。

一方、このような操作的な定義の場合、本当に「安定」した退院サマリが得られるのかという懸念が生じる。たとえば、添削者によって、とある記述が削除された後に、別の添削者によって同じ記述が加筆された場合、退院サマリは「振動」することになる。あるいは、添削者によって加筆や削除するポイントが異なる場合、退院サマリは収束せず、逆に「発散」することになる。糖尿病の教育入院やクリニカルパスに基づく入院の場合には、カルテ記載も定型的になる可能性が高いため、安定サマリが得られやすい一方、診断困難症例の検査目的での入院などでは発散しやすい等、入院の種類によってもこの傾向に差がでることも考えられる。

そこで、この「理想のサマリ」を操作的に定義することにより、本当に理想のサマリを得ることができるかの検証を進めた。具体的には、上記の検証を行うための研究プロトコル策定に加えて、医師を対象として検証実験を行うための検討を進めた。

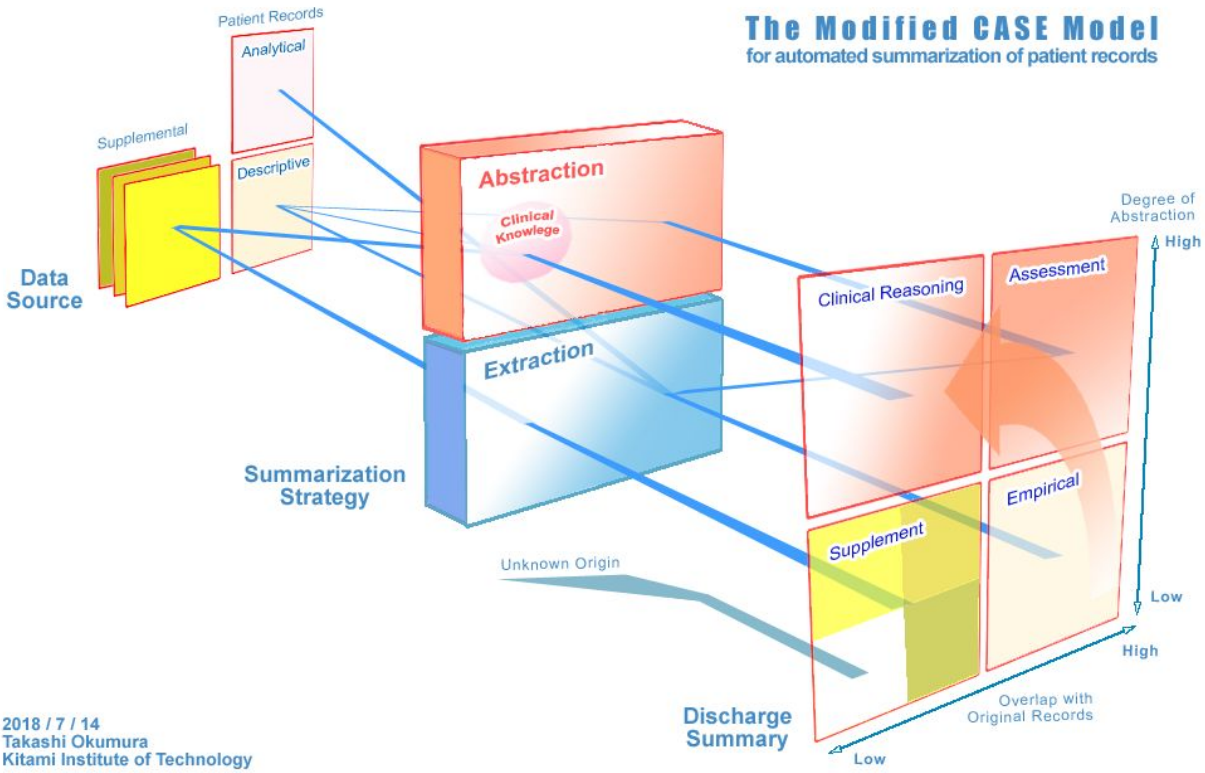


図２　修正 CASE モデル

C．研究結果

ダミーカルテ収集

ダミーカルテの収集と編纂には、全ての過程に医療従事者の関与が求められることに加えて、作業者の錬度が重要となることから、必然的に高コストかつ時間を要することになる。本研究分担では、2 年強に渡る広報活動と収集活動の結果、入院カルテ・退院サマリのペアを含むダミーカルテを 108 件、収集することができた。これらのファイルは、電子カルテより SS-MIX2 形式にて抽出したカルテデータと等価に扱うことができるよう、XML 化したうえで、入院日、退院日、主疾患等のいくつかのメタデータを付与した形となっている。データの品質は高く、一見したところ普通のカルテ情報と遜色がない水準とすることができている。

カルテのアノテーション

新たなアノテーション方式の設計に際しては、旧方式よりも少しでもシステマティックな分類が実現するよう工夫を凝らした。まず、センテンスが言及している内容を、カルテに頻出する話題から整理した 8 つのラベルとして付与した。具体的には、「患者自身に関する(医学的)情報」、「患者以外の何者かに関する情報」、「誰か、何かの行動に関する文」、「誰か、何かの対応に関する文」、「患者の治療に関する言及」、「患者の診断」、「何かの推定」、「可能性に関する言及」について、当てはまるものをマルチラベル式に付与するものとした。これらのうち、前 5 者は事実度が高い内容であり、後 3 者は事実度合いが低い内容と考えられる。そこで、センテンスに付与されるラベルが前 5 者のみの場合には、事実度が高いセンテンスとして扱い、「d: descriptive」とのラベルを付与した。また、後 3 者のいずれかが含まれるセンテンスの場合は、事実度が低い、すなわち、抽象度が高いセンテンスとして、「a: abstractive」のラベルを付与した。ただし、このルールは単純明快ではあるものの、時折例外が生じるため、a、d のラベルについては目視確認のうえ付与した。そのうえで、このデータを用い、汎用性の高い言語表現モデルである BERT

を用いてセンテンスの a/d 分類を試みた。利用したデータセットにおいて、descriptive なセンテンス数は abstractive なセンテンスの 5 倍ほどの分量があるため(d:a = 0.83:0.17)、全ての予測を d としても 83%は正解できてしまうことになる。そこで、a センテンス側の予測性能の向上を目指したところ、再現率(実際に a であるもののうち、a であると正しく予測されたものの割合: recall) 0.70、適合率(正と予測したデータのうち，実際に正であるものの割合: precision) 0.70 の性能を得ることができた。これは、決して高い性能ではないが、単純な手法に基づく予測性能であり、今後のチューニングによってさらなる性能向上が十分に期待できる状況となっている。

修正 CASE モデルの検討

次に、CASE モデルの考察を重ね、アップデート版として修正 CASE モデル(図 2)を策定した。その結果、当初の CASE モデルの弱点であった、退院サマリ中のセンテンスと元の入院カルテ中のセンテンスとの対応付けを行うと共に、そのそれぞれの生成物たるセンテンスの生成メカニズムの検証に道を拓くことができた。たとえば、このモデルでは、退院サマリ中の Assessment 文の由来として、入院カルテに含まれる記述(Descriptive 文)からの Abstractive な要約の他に、もともとのカルテに含まれる分析的な記述(Analytical 文)を Extractive に要約することで生成されうることが示されている。また、センテンスの由来を元の入院カルテに見出すことができない場合、たとえば、Clinical Reasoning 文の由来として、入院カルテから臨床推論を経て記載されるケース、その他の情報源から臨床推論を経て記載されるケースがあることを示している。また、事実度の高い Supplement 文としては、その他の情報源から Extractive に要約されるケースの他に、その他の未知の情報源からの情報が含まれうる点が示されている。これらは、退院サマリ中のセンテンスの分類モデルであると同時に、それぞれのセンテンスをいかに(自動)生成しうるかを示す生成モデルとなっており、その完成度を高めることが出来た。

理想のサマリ研究

自動要約に関する研究分野に、要約研究用のデータセットは存在する。しかし、そもそもそのデータセットに含まれるトレーニング用の要約をいかに高品質に生成するかについては、体系だった研究が見当たらない。エキスパート間のコンセンサスを持って gold standard とするにせよ、誰かが最初の要約作業を行いその要約を修正・添削するのか、複数のエキスパートが平行で要約を行いその結果を用いて討議するのか、品質やコストにとってどれが正しい戦略であるかは自明な問いではない。文献検索を行ってみても、どうも先行研究はみあたらないことから、研究の蓄積が浅い分野であろうことが推察された。近年におけるニューラルネットワークの技術革新によって、高品質なトレーニングデータさえ用意できれば高品質なニューラル要約が可能となる可能性が高い。もし、高品質な要約を低コストに作成できる方法論が確立すれば、退院サマリの要約だけでなく、自動要約研究自体に大きな貢献を果たすことができよう。

そこで、入院カルテを対象とした要約については一旦保留し、より一般的な問題へと条件を緩和して、理想の要約に関する検討を進めた。とりわけ、入院カルテを対象とした検証は被験者(医師)の確保のためのコストが高いことからも、同じ構造を有する他の問題にシフトし基礎的検討を加えることは、研究戦略としても合理的なものと

考えられた。

対象としてまず選択したのは、「就活に関するブログ」の要約タスクである。就活ブログは、ある程度の長さがある時系列なテキストと考えられる。また、複数面接のステップや最終的な結果等、イベントに関する記載が含まれることに加えて、活動を通じた成功要因や失敗要因についての考察が含まれる点で、入院カルテに類似した構成となっている。さらに、就職活動は大学生のほとんどに関わるイベントであり、他人の就活より学べる点も多いことから、大学内での被験者確保に有利なテキストと言える。そこで、今回の理想のサマリ作成の予備実験として、研究協力者に依頼して就活ブログを要約し、その他の被験者に見せ、さらに添削させるという試みを行った。

予想された。そこで、より簡便なタスクとして、「Wikipedia に含まれる会社紹介から適切なサイズの社史部分を抜粋したデータセットを作成し、その要約を試みる」タスクへと修正を加えるものとした。また、この予備実験における挙動を分析することで、「理想のサマリ」を客観的に評価する手法に加えて、要約作業をより低コストに抑える手法について、研究デザインを定めることができた。

D．考察

ダミーカルテ収集

自然言語処理分野の研究論文においては、ダミーデータを用いて研究を行う場合、利

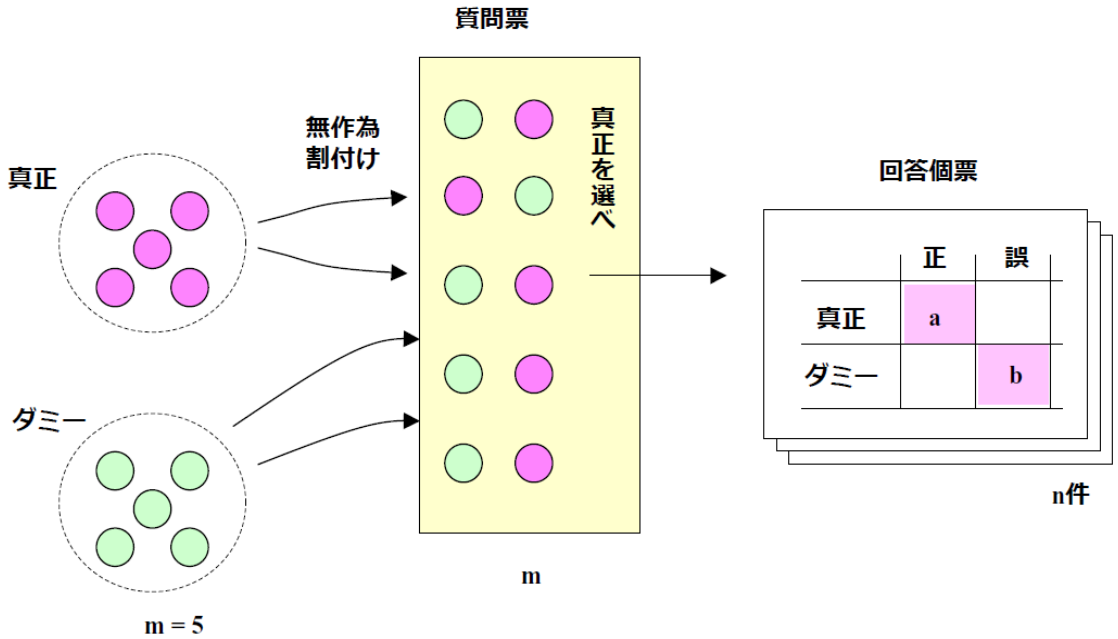


図３ ダミーカルテの真正性検証

試行の結果、まず、最初の要約に時間が掛かり、被験者への負担が予想外に大きいことが明らかとなった。これを第 1 世代の要約と称するとすると、第 2 世代以降の要約は相対的に負担が低くなるが、それでも、今後の研究に際した被験者の確保に困難が

用したテキストが本物と品質等において変わりがないことを証明しなければ評価が下がるケースがある。そこで、この後の研究利用に際しては、まず、「本物であることが証明されているカルテ」とダミーカルテとを医師が区別できないことを示すことが望

ましい。前者については、同意に基づいて、研究協力者より直接提供を受けることで用意が可能と考えられる。後者については、今回収集したダミーカルテからランダムサンプルすることで用意しうるものと考えられる。

研究デザインとしては、無作為に生成した真正データとダミーデータのペアからなる質問票を作成し、これを被験者に提示する手法によって、等質性の検証が実現する(図 3)。もし、真正データとダミーデータが判別可能であれば、真正を正とし、ダミーを誤とした解答が多数となる。一方、この解答がランダムとなるのであれば、両データは目視にて判別不能であり、ダミーデータの品質が真正カルテに等しいことを示すことができる。今後、データと被験者の確保を行い、実際の検証を行いたい。

カルテのアノテーション

上記のように品質管理を行ったダミーカルテについて、今年度、初年度に付与したアノテーションを白紙に戻して、さらに別のアノテーションを行った。その結果、自動分類においても性能向上が認められたが、アノテーションのガイドラインとしてはさらなるブラッシュアップの余地がある。

たとえば、現状では「治療」に言及するセンテンスついて、事実度が高いものと分類している。しかし、治療には、「治療した」という事実の記載だけでなく、今後の治療計画に関するものがあり、現在のラベルはそれらを混在させている可能性がある。また、「誰か、何かの行動に関する文」、「誰か、何かの対応に関する文」という区分についても、「行動」の主体が、患者なのか医療者なのかが不明確となっている。

これらを整理すると、事実度が高いセンテンスについては、主語ないし動作主が患者であるかそれ以外であるかという軸(述語の表す動作の主体)と、「描写」か「行動」かという述語種別の軸によって、2×2 に分類することが、アノテーションガイドラインとしても明確化するのではないかと考えられる。たとえば、「その後空咳も出てきた」という記述において、主語は「空咳」であるものの、動作主は患者となる。ただし、カルテ記載において、この動作主は必ずしも明示的には記載されず、とりわけ患者の場合は少なからず省略されることになる。医療用自然言語処理においては、この動作主の検出というタスク自体に価値がある可能性があるため、今後、再度のアノテーションによりデータ整備を目指したい。

また、事実度の低い、抽象度が相対的に高いセンテンスについても、改善の余地がある。とりわけ、最初のアノテーションにおいて付与した Probable、Certain といったモダリティ情報について、今回のアノテーションでは活用できていない。今後、これらを統合したデータを作成することにより、データセットの価値をさらに高めることが望ましい。

修正 CASE モデルの実証

今回提案した「修正 CASE モデル」は、概念的なモデルであるが、大量の電子カルテがあれば、その妥当性を統計的に検証しうる。

退院サマリ中に含まれるセンテンスの要約手法として、Extractive な要約由来のセンテンスが多いことが分かれば、入院カルテの自動要約に向けたハードルは一気に下がる。また、入院カルテに由来しないセンテンスについて、我々の仮説が示しているように、他の入院カルテに含まれる割合が明らかになることによって、入院カルテの自動要約のハードルはさらに下がることになる。由来が明らかとなれば、元テキストのどういう位置や文脈のテキストが退院サ

マリに出現しやすいか等、多くの統計データを取得していくことが可能となる。

問題は、「テキストに由来しないセンテンスの生成」と、「医学知識に基づいた abstractive な要約の実現」となる。前者については、そもそもそのカテゴリのセンテンスとして、主治医しか知りえない情報なのか、オーダリングシステム中のどこかに存在しうる情報なのか、それぞれがどれくらいの比率で存在するのかという点が重要となる。このいずれにおいても、現在は研究の蓄積が乏しいが、我々の研究によって明らかとなっていくことが期待される。医学知識に基づいた abstractive な要約は技術的な難易度が高いが、こちらも、入院カルテと退院サマリの膨大なペアデータを用いた機械学習により、ニューラル機械翻訳技術と同様にトレーニングデータの物量である程度解決できる可能性がある。

いずれにおいても、今回提案した修正 CASE モデルが基点となって統計取得と研究の発展が見込まれることになる。今後、まずは実データを用いたモデルの実証を目指したい。

<u>理想のサマリ研究</u>

「理想の退院サマリ」の確保に向けては、今年度、客観的な基準に基づく理想的なサマリの作成に向けた仮説の構築と、その仮説の実証に向けた研究デザインの策定を行うことができた。この実証に際しては、退院カルテそのものを用いた研究のコストが高いことから、条件を緩和し、より一般性の高いテキストとして Wikipedia に含まれる企業の社史を対象とした要約タスクとなっている。これにより、文書の要約研究に向けた高品質な要約文書を低コスト、高品質に実現しうることを実証することができれば、次のステップとして、医師を対象とした退院サマリの高品質な整備手法を検証することができる。最終的なゴールである高品質な退院サマリの確保にはまだ道のりが遠いものの、作業としては明確化したことから、実現に漕ぎ着けたい。

E．結論

本研究分担では、退院サマリの自動生成研究の発展に向けて、研究基盤の整備に取り組んだ。そのために、総合に関連した 4 つの課題に取り組み、i) 108 件のダミーカルテ収集、ii) ダミーカルテに対するアノテーションとガイドライン策定、iii) サマリの分析と自動生成に向けたセンテンス分類モデルの構築、iv) 理想の退院サマリの作成に向けた予備実験と研究デザインの策定を行うことができた。

今後、ダミーカルテを対象としたアノテーションの更なるブラッシュアップにより、センテンス分類モデルの実証に向けた分類器の精度管理用データが用意できることになる。この分類器が実用精度となった段階で、実際の入院カルテ・退院サマリを対象とした網羅的な分析を開始することが出来る。この分析は、退院サマリがいかに作成されているかの統計取得を図るものであり、退院サマリの自動生成を実現する研究アプローチを検討するうえで決定的な価値を有することになる。理想のサマリの検討は、この自動生成をニューラル機械翻訳技術の応用により実現する際、生成したサマリを理想的なサマリとの差異によって評価する評価系の実現に大きく貢献する。さらに、高品質なトレーニングデータを大量に用意することができれば、さらなる精度向上に繋がることが期待される。

上記の方向性により、退院カルテの自動生成研究の進展が見込まれる。さらに、自動生成したサマリを提示し、「訂正タスク化」することで、継続的な精度向上が望め

るとともに、医療用自然言語処理の課題の一つである曖昧性解消のためのデータ収集が低コストに実現することになる。また、入院カルテ・退院サマリのペアデータからは、コストの掛かるコーパス整備なしに、様々な副産物が期待される。たとえば、個々の医師やチーム毎に個人辞書を整備することで、「表記ゆれ」の問題を解消することができる。さらに、医療用 NLP の最大の課題であったカルテにおける表記の「用言化」に対し、サマリ訂正タスクを通じた言い換え辞書の整備が実現する可能性がある。

このように、退院サマリの自動作成研究が、技術的な難度が高い医療用自然言語処理技術に対して技術的なブレークスルーをもたらすことが期待される。また、退院サマリの自動生成技術は、実現に向けて医療現場からの期待も高い技術である。今後、コンピュータがカルテ記載を理解できるようになれば、医療用人工知能の発展を通じて、診断や治療において医療従事者を支援する様々な機能が実現する。そのためにも、今後の研究を通じて、さらなる研究の発展を図りたい。

F．研究発表

1．論文発表

なし

2．学会発表

なし

G．謝辞

理化学研究所 革新知能統合研究センター 松本裕治先生、首都大学東京システムデザイン学部 小町守先生、安道健一郎さんには、退院サマリの自動要約アプローチに向けたセンテンス分類に際して多くの助力を頂きました。また、田鎖麻衣さん、大阪大学医学部 宮本紘子さんには、ダミーカルテ整理とアノテーションにご尽力下さいました。北陸先端科学技術大学院大学 浅井拓也さん、北見工業大学 荒田真輝さん、斎藤健斗さん、寺下俊さんは、理想のサマリ研究に貢献して下さいました。また、ダミーカルテの整備に多くの医師の先生方にお力をお借り致しました。この場をお借りして、感謝を申し上げます。

厚生労働科学研究費補助金（臨床研究等 ICT 基盤構築研究事業）
分担研究報告書

退院サマリの品質の自動評価に関するパイロット研究

研究分担者　森田 瑞樹
（岡山大学 大学院医歯薬学総合研究科 准教授）

研究要旨

退院サマリの自動生成技術の実現を目指し，昨年度までの調査結果の整理およびそれを踏まえた退院サマリの自動評価のコンセプト検証を実施した。退院サマリを生成および評価するためには，どのような退院サマリを生成しなくてはならないかを示す「理想的な退院サマリ」の定義が必要となる。初年度は退院サマリの自動生成技術の実現を目指し，予備的な検証を行った。今後どのような方法論で退院サマリの自動生成を行うかを検討するために，その参考になる情報を得ることを目的とし，退院サマリの内容が入院カルテに書かれた文章からどのように生成されているのかを分析した。具体的には，退院サマリの自由記載文は，入院カルテから文，文節，単語を適切に抜き出して組み合わせることで生成されうるのか，それともそれらを単純に組み合わせるだけではなく解釈が必要なのかといった，退院サマリに書かれている文の分析を行った。退院サマリの「入院までの経過」および「入院中の経過」に記載された文章を抽出して文単位に分解し，それぞれの文と入院カルテの記載を比較した。「入院までの経過」は，カルテに書かれた文がそのままか，もしくは文節や単語を組み合わせることで生成できそうな割合が高かった。一方で「入院中の経過」は，カルテに書かれた記述そのままや組み合わせで生成できそうな割合は低く，カルテの記載から解釈が必要なものや，カルテの記載からは作成できない割合が比較的高かった。

２年目に理想的な退院サマリについて言及をした国内外の文献調査を実施し，３年目はその結果を整理した。退院サマリの記載に関して「記載すべき項目」および「定性的な要件事項」を抽出し，前者として 38 項目を得た。このうちの 14 項目はカルテの構造化データより抽出できるものの，残りの 24 項目はカルテの自由記載より抽出してサマリとして文章を作成や要約をする必要があるものと考えられた。また，後者として 9 項目を得た。このうちの 5 項目は医学的な知識・経験がないと採点が難しいものの，残りの 4 項目は形式的に判断をすることが可能と考えられた。この 4 項目のうち医学用語辞書を必要としない 3 項目について文章の特徴を用いて自動評価することを試み，人による評価との相関および点数の分布を踏まえて自動評価の可能性を考察した。退院サマリの自動評価手法の確立に向けた今後の課題が明らかとなった。

A．はじめに

近年，レセプトや DPC などの大規模な医療データ（いわゆる医療ビッグデータ）を用いた分析が研究や病院経営などのために盛んに実施されている一方で，カルテに文章として記載された情報の利活用は進んでいない。カルテの文章の活用を容易

にするためには，記載がある程度は標準化されていることが望ましい。そこで本研究では，退院サマリ（退院時要約）の作成を自動化することにより記載内容を標準化することを目指している。

退院サマリとは，入院していた患者が退院する際に，入院に至った経緯から入院中の経過，および退院後の治療方針などをまとめたものであり，担当医などによって記載される。診療行為を大きく入院と外来に分けると，入院においては外来と比べて短期間に多くの医療行為が実施されるため，カルテの記載量は多くなる。退院して外来に移行する際などに，その内容を効率的に共有するためには入院記録をまとめた退院サマリが効果を発揮すると期待される。現在，医療機関の機能分化が進められており，異なる医療機関や種類の異なる医療施設（病院と介護施設など）でのスムーズな連携を行うために，今後，退院サマリの役割は増していくものと想定される。

退院サマリを自動動成する方法は自明ではない。たとえば，入院カルテからいくつかの文を抽出して組み合わせたり，退院サマリの雛形に必要な情報を入院カルテより抽出もしくは推定して埋めたりするなど，いくつかの方法が考えられる。いずれにしても，どのような退院サマリが望ましいのかを明らかにすることが，自動生成の指標になるものと思われる。また，生成した退院サマリを自動で評価することができれば，その評価結果に基づいてより望ましい退院サマリを選び出すことができるはずである。このため，退院サマリの自動生成のために必要な事項として，理想の退院サマリとは何かを明確に定義すること，生成した退院サマリを自動で評価できること，があると考えた。

昨年度，望ましい退院サマリに関する文献の調査を行い，結果として 51 報の文献を得た（英語：48 報，日本語：3 報）。今年度は，51 報の文献の内容を整理し，退院サマリに記載すべき項目として「記載すべき項目」および「定性的な要件事項」を抽出した。次いで，文献から抽出して整理した要件に基づいて退院サマリを評価するための方法を作成することを試みた。文章の自動評価方法は，正解となる文章との比較に基づいて評価をするものと正解を用いないで評価をするものがあるが，退院サマリの評価においては実臨床で使用することを考慮すると後者の方法しか取り得ない。本研究では，定性的な要求事項を自然言語処理を用いて簡便に評価できることを指向し，退院サマリ 50 報（人工的に作成したダミーデータ）に対して適用した。なお，定性的な要求事項には医学的な知識が必要なものとそうでないものとがあるが，本研究では後者のみを対象とした。

B．研究方法

1）サマリ文の分析

退院サマリの各文について，元になった入院カルテと比較をすることで，その文が入院カルテからそのまま抜き出された文なのか，文や文節などを組み合わせて書かれた文なのか，それとも入院カルテの記載を解釈して新たに生成された文なのか，を決定する。

もし入院カルテから抜き出した文を組み合わせて退院サマリが作成されているのであれば，自動生成のためには入院カルテから適切な文を抜き出して並べることになる。文節や単語を組み合わせて書かれているのであれば，適切な文節や単語を抜き出して文を生成することになる。単語すら書き換えられて入院カルテの記載とは異なる文が書かれているのであれば，入院カルテを入力として文を生成することになる。

退院サマリの各文は次の５つのタイプに分類した：タイプ１．入院カルテの文がそのまま（もしくはほぼそのまま）使われている，タイプ２．入院カルテの文そのままではないが，複数の文や文節を組み合わせることでその文を作ることができる，タイプ３．その文を書くには入院カルテを読ん

で解釈をする必要がある（医療の知識がなくても解釈が可能な範囲である)，タイプ４．その文を書くには入院カルテを読んで解釈をする必要がある（医療の知識がないと解釈ができない)，タイプ５．その文は入院カルテの内容からだけでは書くことができない（情報が不足している)。分類作業は医療の知識がある 4 名で行い，不一致の場合は話し合いによって１つの分類に決定した。13 の退院サマリを使用した。退院サマリは入院までの経過および入院中の経過を使用した。

２）国内外の文献調査

望ましい退院サマリに関して書かれた文献 51 報から，記載すべき項目と定性的な要求事項を抽出して分類整理した。記載すべき項目は，英国 Academy of Medical Royal Colleges（AoMRC）による退院サマリの項目分類（2013 年）に基づいて退院サマリに記載すべき項目として整理した。また，文献に登場した退院サマリの記載における定性的な要求事項を分類した。

定性的な要求事項のうち医学的な知識が必要とされない項目として文や単語の特徴を評価した。評価の対象となったものは，「文法が正しい」，「言葉遣いが適切である」，「簡潔である」の 3 項目（9 項目中）であった。文章校正用のソフトウェアを使用し，そこから得られる解析結果を定性的な要求事項に合うようにスコア化した。文章校正用のソフトウェアとして，Just Right!6 Pro（ジャストシステム)，一太郎 Pro4（ジャストシステム)，PressTerm（NTT データ東北)，Tomarigi（青山学院)，Word（マイクロソフト）を比較し，Just Right!6 Pro を使用した。この方法を用いて評価ができるのは，文法や語句の使用法の適切性（誤字脱字，同音語誤り，同一助詞の連続など)，文章全体や一文あたりの長さ（総文字数，文数，平均文長）である。これらの解析結果を前述の 3 項目に当てはめて評価を行った。

文章の自動評価手法の評価は，人間の評価と相関するかによって確かめられることが多いが，一方で本研究のように主観を排除した文章の評価手法の評価とは必ずしもよい相関を示すとは限らない。そこで，3 項目について人の評価との比較をすることに加えて，よい評価手法ではスコアに適度なばらつきが生じなければならないと仮定してスコアのばらつきによって本研究による方法の妥当性を考察した。人による評価は，3 項目のそれぞれについて 5 段階で評価した（0～4)。ソフトウェアによる「簡潔である」の評価においては，生スコアを平均値で除算することによって正規化した。

C．結果

１）について

全体での各タイプの内訳は，タイプ１：43%，タイプ２：3%，タイプ３：9%，タイプ４：24%，タイプ５：21%，となった。入院までの経過における各タイプの内訳は，タイプ１：72%，タイプ２：1%，タイプ３：4%，タイプ４：10%，タイプ５：13%，となった。13 の退院サマリのうち 6 の退院サマリでは，入院までの経過のすべての文がタイプ１であった。入院中の経過における各タイプの内訳は，タイプ１：24%，タイプ２：5%，タイプ３：12%，タイプ４：33%，タイプ５：26%，となった。入院中の経過ではすべての文がタイプ１の退院サマリはなかった。

入院までの経過は，前半部分に発症からの経過が，後半部分に入院を判断するに至った理由が書かれていることが多かった。入院までの経過は全体的に入院カルテから文をそのまま持って来ていること（タイプ１）が多かったが，すべてがタイプ１ではない場合には，前半部分で特にその傾向が強く，一方で後半部分は医学的な知識がないと解釈ができない文（タイプ４）の割合が若干だが高かった。

入院中の経過は，入院中の症状と治療の経過が書かれ，その最後には退院をした旨と退院後の方針が書かれていることが多かった。退院後の方針は入院カルテの記載だけからでは書くことが難しいこと（タイプ５）が多い傾向にあった。

入院までの経過と入院中の経過を比較すると，入院中の経過はタイプ１の割合が低く，タイプ４と５の割合が高くなっていた。入院までの経過がタイプ１が 72%だったのに対し，入院中の経過は逆にタイプ３～５が計 71%となった。いずれの場合もタイプ２は非常に割合が低かった。

２）について

AoMRC の分類では 82 の退院サマリに記載する項目が 23 に分類されている。記載すべき項目として，AoMRC 以外の文献から分類および項目を追加して 91 項目とし，そのうち 5 報以上の文献に登場した 38 項目を得た（この詳細は第 38 回医療情報学連合大会の抄録参照）。定性的な要求事項として抽出したものを類似の要求事項をカテゴリにまとめて分類し，下記の 4 分類 9 項目となった。

【分類 1：完全性・正確性】内容に不足がない，診断などに関連のないことを記載していない，不適切なコピー＆ペーストをしていない，書かれている情報は正確である
【分類２：見読性・理解容易性】構造化されていて必要な情報にすぐにたどり着ける，文法が正しい
【分類３：言葉の使用】言葉遣いが適切である，診断名などは正確に書き，略語を使っていない
【分類４：分量・文字数】簡潔である

人による定性的な要求事項の評価は次の通りとなった。50 報の退院サマリのいずれもよく書けており，高いスコアになった（0 がよい評価）。

【文法が正しい】平均：0.14，SD：0.40，最低：0，最高：2
【言葉遣いが適切である】平均：0.24, SD：0.48，最低：0，最高：2
【簡潔である】平均：1.02，SD：0.89，最低：0，最高：3

次に，ソフトウェアを用いた評価は次の通りとなった。

【文法が正しい】平均：0.06，SD：0.24，最低：0，最高：1
【言葉遣いが適切である】平均：92.74，SD：19.92，最低：45，最高：135
【簡潔である】平均：5.00，SD：0.58，最低：3.9，最高：6.7

相関係数は，文法が正しい：0.12，言葉遣いが適切である：-0.08, 簡潔である：0.14 となり，ほぼ相関がなかった。

なお，ソフトウェアによる評価の詳細は下記の通りであった。

【文法が正しい】修飾関係：0.06（SD: 0.24，Min: 0，Max: 1），並列関係：0.00（SD: 0.00，Min: 0，Max: 0），ら抜き表現：0（SD: 0, Min: 0, Max: 0），さ入れ表現：0（SD: 0，Min: 0，Max: 0），二重敬語：0（SD: 0，Min: 0，Max: 0），たりの脱落：0（SD: 0，Min: 0，Max: 0），べき止め：0（SD: 0，Min: 0，Max: 0）
【言葉遣いが適切である】誤字脱字：5.18（SD: 2.95，Min: 1，Max: 12），同音語誤り：0.08（SD: 0.27, Min: 0, Max: 1），送り仮名：0.2（SD: 0.49，Min: 0，Max: 2），漢字基準（常用漢字）：18.16（SD:

6.61，Min: 4，Max: 37），公用文：2.46（SD: 1.54，Min: 1，Max: 6），スペルチェック：63.56（SD: 20.80，Min: 15，Max: 105），表記ゆれ：2.68（SD: 1.65，Min: 0，Max: 7），同一助詞の連続：0.20（SD: 0.45，Min: 0，Max: 2）

【簡潔である】総文字数（空白除く）：2246.96（SD: 388.76，Min: 1,420，Max: 3,125），文数：85.44（SD: 33.39，Min: 21，Max: 169），平均文長：30.2（SD: 14.04，Min: 10，Max: 95）

D．考察

退院サマリに記載すべき項目として，投薬された薬剤名，退院後の診療行為，治療中の疾患，検査結果，入院中の治療，入院中の経過などが多くの文献で記載すべき項目として挙げられていた。退院時の担当医のように一部の項目は病院情報システムに登録された情報をそのまま引用できるが，38 項目のうち 24 項目（63%）はカルテの構造化データから自動で抽出することが難しいと思われる項目であった。つまり，文章として記載をすることが求められる項目である。退院サマリの自動生成をする際には，これらの項目は自然言語処理を用いてカルテの自由記述から抽出し，要約をすることが求められる項目である。

記載すべき項目を文献から抽出する方法として，より多くの文献で触れられている項目を採用するという方法をとったが，この方法の課題に，特定の診療科や分野において重要と思われる項目が拾われないことがある。たとえば，褥瘡ケアのために入院前の ADL の把握が必要，終末期の慢性胃炎や長期間の介護を受けていた患者では栄養状態や緩和ケアの状況が必要，といった項目が例として挙げられる。対象領域を限定した研究を実施する際にはこの点を考慮し，当該領域の文献のみで分析を実施することが望ましい。

定性的な要求事項の 9 項目の中には，医療の知識がないと判断が難しいと考えられる事項が一部にあった。内容に不足がない，診断などに関連のないことを記載していない，不適切なコピー＆ペーストをしていないなど，9 項目のうち 5 項目が該当した。残りは知識ではなく形式から判断ができる可能性があるものであり，これらが自動採点の対象となり得ると考えられた。

定性的な要求事項の一部をスコア化することを試みた結果として，スコアを出すことはできるものの，人による評価との相関はなかった。相関なかった要因としてはいくつか考えられる。まず，本研究で用いた退院サマリはダミーとして医師に書き起こしてもらったものであるため，いずれもよく書けており，スコアの分散が小さくなるという点である。人による主観的な評価において退院サマリ間の差がほとんどなく，高いスコアに集中した。この傾向はソフトウェアによる評価において文法のエラーがほとんど指摘されていないことにも表れている。それ以外の指標に関するソフトウェアによる評価は機械的になされるため細かい差がついた。それなりの長さがある文章において文法のエラーを人が 5 段階で点数付けするというタスクは実際に行ってみると安定して実施をすることに困難が感じられ，この点はソフトウェアで評価を実施することの意義を示していると考えられる。

9 項目のうち 1 項目「診断名などは正確に書き，略語を使っていない」は，今回は評価対象に含めなかったが，医療用語辞書（病名，検査名，薬剤名など）を用いて解析をすることで対応が可能と思われる。なお，今回の解析において漢字の使用が不適切である（常用漢字ではない）との判定が頻発しており，その多くは医療用語によるものであった。また，校正ソフトウェアではスペルチェックを行うようになっており，

非常に多くのスペルミスが指摘されているが（退院サマリあたり平均 63.6 箇所），検査名称などがスペルミスとして検出されており，これらも医療用語辞書を用いて対象から外す必要がある。

本研究の方法を実際の評価に使用する際には，9 項目のうち 1 項目「簡潔である」の判断のためにどの程度の数値であれば「簡潔である」であるかを事前に決める必要がある。このために，既存の「わかりやすい日本語」に関する研究の成果を援用することが可能と思われる。たとえば，一文あたりの文字数として 50〜80 文字と言われていたり，また外国人にわかりやすいのは平均 24 文字などと言われたりしている。なお，退院サマリは通常いくつかのブロックに分かれており，文が極端に短い箇所（単語の羅列）と長い箇所（経過に関する記述）とで一文あたりの文字数は極端に異なるので，今回の解析のように全体に１つの点数を付けることは不適切と考えられた。今後はこれらを分けて解析をする必要がある。

文章全体の文字数について参考になる数値として，日本語の論文誌の抄録は 300〜500 文字が上限として設定されていることが多い。この基準を踏まえると，本研究で用いた退院サマリは各文は短くまとまっているが，文の数はいずれも非常に多いと判定された。ただし，こうした適切さの基準はそれぞれ場面を限定して検討されていることもあり，つまり退院サマリとしてどれくらいの値が適切かはこれらとは一致しない可能性もある。よって理想的には，一般的な指標を参考にしつつ退院サマリとして適切な閾値を設定することが望ましい。

本研究の方法では 9 項目のうち 3 項目のみを評価可能と判断した。残りの 6 項目のうち 5 項目「内容に不足がない」，「診断などに関連のないことを記載していない」，「不適切なコピー＆ペーストをしていない」，「書かれている情報は正確である」，「構造化されていて必要な情報にすぐにたどり着ける」を判断するには医療の知識・経験およびそれに基づく内容の理解が必要であり，自動的に判定をするためには本研究とは大きく異なるアプローチが必要である。また，1 項目「構造化されていて必要な情報にすぐにたどり着ける」は医学的な知識は必要ないものの，一文を越えた範囲を解析する必要があるため本研究で使用したソフトウェアでは対応ができなかった。調査をした多くの文献で退院サマリの雛形を使用することの重要性が説かれていたが，雛形において記載箇所が複数に分けられている場合には，それぞれに記載されている内容を判定することで定量化が可能になると考えられる。先の 5 項目とは異なり，医学的な知識や内容の理解がなくとも，使用されている用語の頻度などで一定の判定は可能と想定される。

E．さいごに

本研究では，退院サマリを文ごとに分解し，それぞれの文が対応する入院カルテにどのように書かれていたかを分析した。退院サマリの中の入院までの経過および入院中の経過を調べたところ，入院までの経過の各文は入院カルテをそのまま写していることが多かった一方で，入院中の経過の各文は入院カルテの記述を解釈する必要があったか，もしくは入院カルテの記載からだけでは書くことができないことが多かった。

また、文献から抽出して整理した要件に基づいて退院サマリを評価するための方法を構築することを試みた。この方法で十分な水準で退院サマリを評価し得るものかは本研究によっては結論が出せなかったものの，退院サマリの自動評価方法の確立のために解決すべき様々な課題を明らかにすることができた。

F．研究発表

1．論文発表

なし

2．学会発表

森田瑞樹，奥村貴史，狩野芳伸，堀口裕正．退院サマリの自由記載は何を書くことが望ましいのか：文献レビュー，第38回医療情報学連合大会，2018年11月22～25日，福岡．

厚生労働科学研究費補助金（臨床研究等 ICT 基盤構築研究事業）
分担研究報告書

退院サマリの自動生成に向けた電子カルテの自動分析

研究分担者　狩野 芳伸
（静岡大学 情報学部 行動情報学科 准教授）

研究要旨

入院患者の退院に際し、医師は入院中に記載したカルテ等の情報から退院サマリを作成する必要がある。この退院サマリを自動的に生成することが出来れば、臨床現場の負担を下げることが出来ると共に、医療の質に貢献することが期待される。そこで、本研究分担では、退院サマリの自動生成に向けたテキストの分析についての研究を行った。

そこで、本研究分担では、退院サマリの自動生成に向けた研究アプローチの検討に取り組んだ。まず、文献調査と医師へのヒアリングに基づき、良質な退院サマリに求められる要件について定性的な検討を行った。同時に、実際の退院サマリを対象とした分析を行い、要約過程に関する知見を整理した。さらに、一般的な文書の要約手法と入院カルテの要約手法について文献調査を行った。

その結果、退院サマリの分析枠組みと退院サマリの生成モデルを兼ねた CASE モデルと証するモデルを構築することが出来た。その上で、今後、本モデルが示唆する特性の異なる 4 つの要約処理が出力した候補文集合を退院サマリの下書きとして提示し作成支援するツールのプロトタイピングが望まれることを示した。

カルテの処理にあたっては、事前に匿名化が必要となる。匿名化作業を自動化するための匿名化ツールの実装と性能向上に取り組んだ。そのために、既存の正解付き模擬カルテデータに加え、別のダミーカルテデータセットに対し匿名化のためのアノテーション付与を行い、これらを用いてルールベースおよび機械学習による匿名化ツールの実装と性能検証を行った。

サマリ生成にあたっては、対象とするカルテやサマリのドメイン、すなわち診療科や疾患により、サマリ生成に必要となる情報が異なると考えられる。サマリと対応する電子カルテの履歴データについてクラスタリングを行い、どのようなタイプのサマリやカルテがどう類似しうるかの分析を行った。

１．はじめに

本研究では退院サマリの自動生成を目指している。すなわち、入院中の記録である電子カルテを中心とする患者の履歴を入力とし、その患者の退院時の「まとめ」にあたる退院サマリを出力とするシステムの構築である。

まず、電子カルテという個人情報を扱うことから、厳重なセキュリティ環境が必要である一方、現実的に研究が遂行可能な環境の構築が必要である。

電子カルテの処理にあたっては、自然言語処理によるテキスト処理が必須である。ひとつには、個人情報保護の観点から匿名化処理が必要となる。そのうえで、電子カ

ルテデータにおける個別日時の情報（以下、履歴と呼ぶ）とサマリの間にカルテの種類に応じてどのような関係がありうるか、分析を行い、サマリの自動生成を試みた。

２．セキュアかつ効率的な研究環境の整備と運用

本研究の遂行にあたり、セキュアかつ効率的な研究環境の構築を行った。実行環境を仮想マシンとし、実行環境そのものを遠隔送信し、現地で容易に実行できるようにした。ただし、現地では秘匿すべきデータは別サーバに格納し、ソフトウェアからは一時的なアクセスとして情報を残さないようにすることでセキュリティを確保した。本年度は、このシステムを運用して研究を行った。

３．模擬カルテとアノテーションに基づく退院サマリ考察

他の分担研究により、本年度模擬カルテの提供と、その模擬カルテに基づいた、退院サマリ作成を考慮したアノテーション付与が行われた。

サマリを要約ととらえると、一般的な自動要約技術が適用できそうにも思われる。多くの自動要約は、トピックの統計的な解析を行ったうえで、文書中で重要なトピックが含まれるものを残す、という手法が骨格になっている。しかし退院サマリでは、統計的に重要でない、文書集合中で共通して頻出するトピックであっても、サマリとして残すべきことが多々ある。

また、入力にあたる電子カルテの文章中にない文章や表現が、サマリにどのくらい含まれているかという問題がある。入力のサブセットでよいのであれば、切り貼りの範囲内におさまるが、現実には言い換えに始まり内容的にも新規な文章の挿入がありうる。

分担研究のデータによると、入院までの経過については７割以上のサマリ文がカルテの文章ほぼそのままであった。このことは、医師がサマリを作成する際に文の複製を使用しており、分量的な減少もあまりみられないことから、内容的にもあまり変更を必要としていないことを示唆している。ただし、入院前の記述は他の医師からのお願いの形式をとっており、そのままでは主語や言葉遣い、時制などが不適切なので、そうした部分の変換が必要かもしれない。

入院中の経過については、２～３割程度がそのままの文であった。入院中のカルテの記載は文を完成させず断片的なスタイルのことが多く、一方サマリではきちんとした文にするため多かれ少なかれ文生成の要素が必要と思われる。また、医師本人の記録なので、振り返ることで要素を追加したり、整理したりすることが想像される。

４．カルテとサマリの間の類似性に関する予備的実験

どの文を（一部であっても）サマリとして取り入れるかを自動判定する際、最も基礎的な要素は単語の共通性になる。そこで、予備的実験として、共通する単語の分布を測定した。対象を内容語のみにすると、当然ながら「入院」「退院」「病名」など、入院カルテにおける一般的な単語が上位にみられた。一方で、部位、症状、病名、薬品名、単位などを表す単語も上位にきており、最初の手掛かりになろうと思われる。

５．自動的な匿名化

電子カルテの実データに対し処理を行うには、まず匿名化が必要となる。具体的には、個人氏名、年齢、住所、日付、医院名

など個人を特定しうる情報の自動抽出である。

匿名化の研究は長らく行われているものの、特に日本語医療分野の匿名化は利用可能なリソースが少ないこともあり、研究の数は限られている。利用可能なリソースとしては、NTCIR MedNLP 匿名化タスクで配布された模擬カルテコーパスとそこに付与されたアノテーションが挙げられる。このとき評価に使われた正解アノテーションは期間限定で利用不能となっており、残念ながら直接的な比較を行うことはできない。

そこで、研究班内で作成されたダミーカルテを対象に、新たに匿名化のための正解アノテーションを付与し、学習及び評価に用いた。付与するアノテーションは基本的に MedNLP タスクにおけるものと同種とした。すなわち、年齢・個人名・医院名・性別・時間表現である。

これらのデータを用いて、自動匿名化ツールのプロトタイプ実装を行い、性能を検証した。手法としては、ルールベースのものと機械学習によるもの、およびそれらの混合を試みた[Kajiyama 18]。前述の理由から、MedNLP タスクにおける先行研究との直接比較はできないが、概ね同等程度の性能が得られたと考えられる。また手法としては、文字ベースの LSTM（Long-Short Term Memory）を用いたものが最も安定して高性能であった。ただし、いずれのデータセットも模擬カルテの作成コストが大きく、サンプル数が不足している。そのため、end-to-end の機械学習のみでは性能が不十分であり、ルールベースの併用が必要である。

６．カルテとサマリの間の類似性に関する実験

サマリを要約ととらえると、一般的な自動要約技術が適用できそうにも思われる。多くの自動要約は、トピックの統計的な解析を行ったうえで、文書中で重要なトピックが含まれるものを残す、という手法が骨格になっている。しかし退院サマリでは、統計的に重要でない、文書集合中で共通して頻出するトピックであっても、サマリとして残すべきことが多々ある。

また、入力にあたる電子カルテの文章中にない文章や表現が、サマリにどのくらい含まれているかという問題がある。入力のサブセットでよいのであれば、切り貼りの範囲内におさまるが、現実には言い換えに始まり内容的にも新規な文章の挿入がありうる。

研究班内のデータによると、入院までの経過については 7 割以上のサマリ文がカルテの文章ほぼそのままであった。このことは、医師がサマリを作成する際に文の複製を使用しており、分量的な減少もあまりみられないことから、内容的にもあまり変更を必要としていないことを示唆している。ただし、入院前の記述は他の医師からのお願いの形式をとっており、そのままでは主語や言葉遣い、時制などが不適切なので、そうした部分の変換が必要かもしれない。

入院中の経過については、２～３割程度がそのままの文であった。入院中のカルテの記載は文を完成させず断片的なスタイルのことが多く、一方サマリではきちんとした文にするため多かれ少なかれ文生成の要素が必要と思われる。また、医師本人の記録なので、振り返ることで要素を追加したり、整理したりすることが想像される。

このようにさまざまな要素があるが、現実には診療科や疾患のタイプによって、類似性の高いものとそうでないものがありうる。そこで、国立病院機構内の実際のカルテを対象にクラスタリングを行い、カルテに記載された ICD や DPC といったラベルを含めて類似性の解析を行った。

７．サマリの自動生成と評価

前節までの結果を踏まえ、退院サマリの自動生成とその評価を行った（論文投稿予定）。退院サマリの自動生成にあたっては、extractive な処理を行うこととし、電子カルテ内の各文についてサマリに含めるべきか否かの判断を行った。判定には、文末表現に着目する手法、文ベクトルを生成して文間の類似度で決定する方法などいくつかの手法を行い、文一致率、単語一致率、ROUGE など異なる指標での評価を行った。結果、いずれの手法でもある程度の一致率を達成しうることがわかった。今後、より実用的な性能の達成のためには、表記揺れの吸収、さらに大規模なデータの利用による学習性能の向上などの研究が考えられる。

[Kajiyama 18] Kajiyama, K. Horiguchi, H. Okumura, T. Morita, M. Kano, Y., 2018. De-identifying Free Text of Japanese Dummy Electronic Health Records. The Ninth International Workshop on Health Text Mining and Information Analysis(LOUHI2018) (p. 65).

研究成果の刊行に関する一覧表

１．論文発表

古崎晃司,堀口裕正,奥村貴史,津本周作: OS-27　人工知能の医療応用 **人工知能** 33(6):843-848 2018

堀口裕正: **国立病院機構のデータベースを用いた臨床研究** Progress in Medicine 38(2):17-20 2018

堀口裕正: NCDA **の現況と成果および今後の展望 月刊新医療** 2018 **年** 2 **月号**

２．学会発表

森田瑞樹, 奥村貴史, 狩野芳伸, 堀口裕正. 退院サマリの自由記載は何を書くことが望ましいのか：文献レビュー, 第38回医療情報学連合大会, 2018年11月22～25日, 福岡.

[Kajiyama 18] Kajiyama, K. Horiguchi, H. Okumura, T. Morita, M. Kano, Y., 2018. De-identifying Free Text of Japanese Dummy Electronic Health Records. The Ninth International Workshop on Health Text Mining and Information Analysis(LOUHI2018) (p. 65).