平成 28 年度厚生労働科学研究補助金

（臨床研究等 ICT 基盤構築研究事業）分担研究報告書

SS-MIX2 分析用データセットの作成・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
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

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。

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 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。

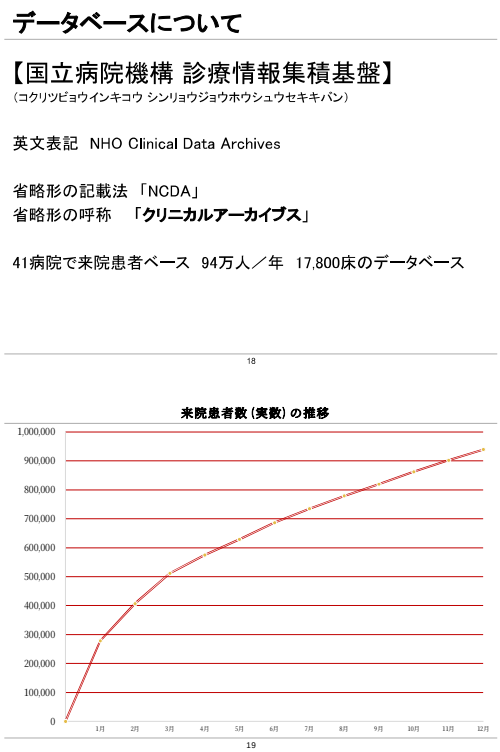
B.方法

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の調整を図ることとした。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うこととした。

C.結果

２, 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、現在 41 病院が参加、約 50000 床、年間実患者数約 90 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、運用開始直後で不安定な状況ではあるが、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。



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。

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。その為、研究期間が 10 月末から開始された後、先ずどのような分析活動を行うかについて数ヶ月にわたり検討を行い、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参考資料 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。

E.結論

本年度、今後研究を実施していくための基礎的な研究基盤の構築に向けた第 1 歩が踏み出せたと考えている。

来年度以降、この分析基盤をきちんと整備するとともに、他の研究分担とともに研究成果を出していきたい。

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
（平成 28 年医療情報学会発表資料抜粋）

国立病院機構における
電子カルテデータ標準化について

国立病院機構本部
IT推進部　医療情報データベース企画室長
堀口　裕正

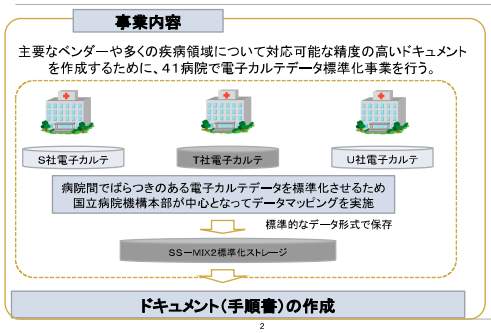
※第36回医療情報学連合大会　COI開示
開示すべきCOIはありません。

1

補助金事業の事業背景

- 平成26年6月24日に閣議決定された「世界最先端IT国家創造宣言」では、地域を越えた国民への医療サービスの提供等を可能とする医療情報利活用基盤の構築を目指し、医療情報連携ネットワークについては、電子カルテを含めたデータやシステム仕様の標準化等を行い、平成30年度までに全国への普及・展開を図ることとされている。
- しかしながら、電子カルテについては、ベンダー毎で開発が行われ、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ーや病院で異なり、標準化の課題となっている。
- 今回の『電子カルテデータ標準化等のためのIT基盤構築事業（13.0億円）』は、このような問題を解消するため、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ー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し、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。

補助金事業の概要　（課題・目的等）



2

事業の成果（標準化の普及促進関係）

- 最新のSS-MIX2Ver1.2cに完全準拠しているモジュールが41病院に導入
  - SS-MIX2 Ver1.2cモジュールの導入
  - SS-MIX2に完全準拠しているモジュール
- HOTコード・JLAC10・ICD10など標準コードを全面的に導入・活用
- 従前のモジュールで課題となっていたベンダー毎の表記ゆれ等の問題が解決され、データ形式の標準化が可能となります
- 本モジュールは6ベンダーから他の医療機関にも（有償にて）提供可能です。
- 他の医療機関が厚生労働省標準規格に準拠（SS-MIX2・標準コード等）したシステムを導入するに当たり、当該事業で作成したドキュメント（手順書）を活用することにより、専門的な知識を要することなく、簡便に導入することが可能となります。

1

事業の成果（標準コード及び標準化団体）

- 標準規格が持つ課題を標準化団体とともに解決
  - HOTコード・・・一般名処方用や持参薬用のコードの整備をMEDISに依頼
  - JLAC10コード・・・体温等の検査コードの採番依頼
  - SS-MIX・・・各種規約の矛盾や、解釈について整理をJAMIに依頼

4

今回のプロジェクトのコンセプト

- 補助金事業として13ヶ月という短納期で仕上げる必要がある。
- 標準化の普及促進に資することを目標とする

以上の条件から以下のコンセプトで事業を実施した

- 検証環境での十分なテスト／検証を行い病院別の開発を極力行わない
- 病院における医療提供に係るユーザーインターフェイスは一切変更しない

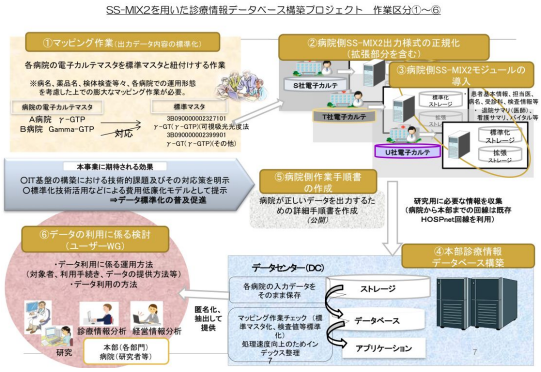
5

国立病院機構のDB事業概要　（プロジェクト概要）

方　針

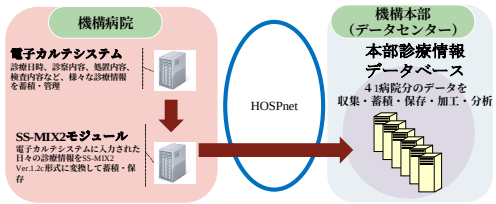
主な作業区分	内容
①マッピング作業（出力データ内容の標準化）	対象４１病院を選定し、データマッピング作業を実施する
②病院側SS-MIX2出力様式の正規化（拡張部分を含む）	全てのSS-MIX機能（メッセージ）に対応できるよう、モジュールを各ベンダで正規化（入力値の正規化・フルセット化等）する。併せて標準仕様以外の拡張データ（バイタル等）が出力できるようにする
③病院側SS-MIX2モジュールの導入	①で選定した対象病院に②で作成したSS-MIX2モジュールを導入する
④本部診療情報データベースシステム構築	データを収集する仕組みを検討し、外部データセンターにデータベースを構築する
⑤作業手順書の作成	本プロジェクト終了後、各病院がSS-MIX2を効率的に導入できるよう、SS-MIX2モジュールを導入するベンダが作業手順書を作成する（手順書は公開予定）
⑥データ利用に係る検討（ユーザーWG）	システム機能とユーザーの要望について調整するデータベースの利用に係る規定（プロセスやルール）や具体的なデータ利用方法を検討する

6



IT基盤構築の仕組み　(標準化、IT基盤とは)

- 本事業では、各病院の電子カルテシステムのデータをSS-MIX2 Ver.1.2c形式に変換（標準化）し、診療情報データベースに収集する仕組み（IT基盤）を構築します。
- この IT基盤は、各病院に導入するSS-MIX2サーバと、機構本部（データセンター）に導入する本部診療情報データベースシステムから構成されます。



8

SS-MIX2を用いたIT基盤構築事業参加病院口覧

ベンダ・病院規模別分布

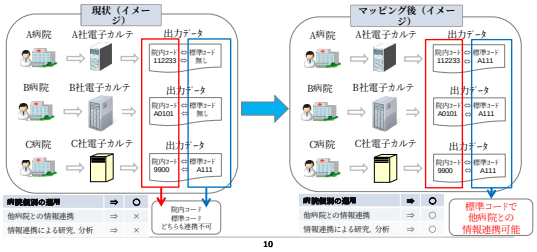
	300床以上	350～499床	349床以下	複合（その他）	障害病床中心	総計
富士通	6病院 金沢医療,名古屋医療,大阪医療,九州医療,長崎医療,熊本医療	5病院 横浜医療，相模原，千葉医療，小倉医療，別府医療	1病院 南和歌山医療	6病院 北海道医療,西群馬,東京,長良医療,村山医療,福岡東医療	4病院 東埼玉,医王,三重,広島西医療	22
日本電気		2病院 北海道がん，埼玉		3病院 旭川医療，帯広，高知	1病院 仙台西多賀	7
ソフトウェア・サービス		5病院 高崎総合，四国がん,九州がん,嬉野医療,鹿児島医療	1病院 米子医療		1病院 高松医療	7
亀田医療情報				2病院 西新潟中央、敦賀医療		2
SBS		1病院 静岡医療			1病院 大竜	2
日本ＩＢＭ	1病院 仙台医療					1
地域毎	7	13	3	11	7	41

北海道東北	関東信越	東海北陸	近畿	中国・四国	九州	総計
6病院	11病院	7病院	3病院	5病院	9病院	41病院

9

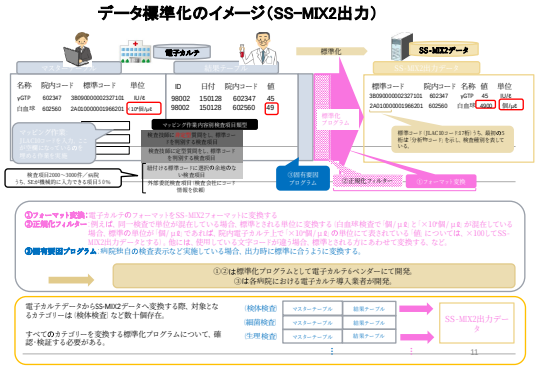
病院におけるマッピング作業

- 院内コードと標準コードを紐付ける対応表を作成します（マッピング作業）。
- 病院毎に異なる院内コードを、標準コードに変換することにより、他病院と連携した診療情報の分析等が可能になります。

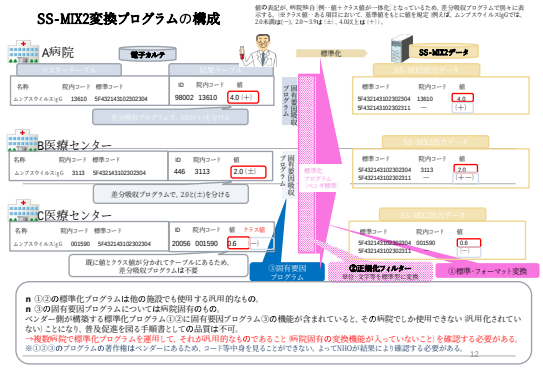


10

データ標準化のイメージ（SS-MIX2出力）



SS-MIX2変換プログラムの構成



12

検査結果の表記揺れについて

病院の部門システム等から発生する検査結果表記に関しては以下のような揺れがあります。今回、病院のモジュールでその揺れを標準的記法に統一しました。

1.検出限界超の値 [> 100(半角の>,スペース,100)]

100<	100 <	100＜	100 ＜	>100	> 100	＞100	＞ 100

2.検出限界以上の値 [>= 200(半角の>,半角の=,スペース,200)]

200≦	200 ≦	200<=	200 <=	200=<	200 =<	200＜＝	200 ＜＝
200＝＜	200 ＝＜	200以上	200 以上	≧200	≧ 200	>=200	>= 200
=>200	=> 200	＞＝200	＞= 200	＝＞200	＝＞ 200	200～	200↑

3.定性値(−)[−(スペース2つ,半角の−)]

−	(−)	フケンシュツ	インセイ	陰性	検出せず	検出セズ
−	(−)	フケンシュツ	ミケンシュツ	ミケンシュツ	ケンシュツセズ	ケンシュツセズ

4. 半定量値 [1 +(スペース,1,スペース,半角の+)]

1+	(2+)	(4＋)	5+	(++)	＋＋＋	(＋＋＋＋)	+++++

13

検査結果の変換について

各JLAC10においてどのような検査値をどのような単位で記載すべきか、どのように表記すべきかについて、マスターを提供・公表しました。

検査項目(参考)	JLAC10	単位
尿量	1A005000000192001	mL
末梢血液一般検査 − 赤血球数	2A990000001930951	,10*4/uL
末梢血液一般検査 − 白血球数	2A990000001930952	/uL

検査	分離前 JLAC10	検査値のパターン	検査結果の例	分離後 JLAC10−1	分離後 JLAC10−2	検査結果 1	検査結果 2
麻疹ウイルス ウイルス抗体IgG	5F4311431023 02304	数値(記号)[文字列]	2.0(−)未満	5F4311431023 02304	5F4311431023 02311	数値[＋文字列]	記号
鼻汁好酸球	2A300000 006360411	(記号)数値	(2+)60%	2A3000000 06360411	2A3000000 06360402	記号	数値

14

検証環境におけるテスト

- 今回の事業を実施するため、本部に6ベンダーの電子カルテシステムをレンタルして各種検証を行った。
- すべての電子カルテに同じ処方オーダーを登録すると、そのデータがどう扱われて、SS-MIXデータに変換あされるのかについて、電子カルテの画面入力からスタートするテストを行って、各社の違いを確認した上で開発調整を行っている。
- これにより、NHO41病院に限らず、汎用的に使えるモジュールになるように開発を行った。

16

データベースについて

【国立病院機構 診療情報集積基盤】

（コクリツビョウインキコウ シンリョウジョウホウシュウセキキバン）

英文表記　NHO Clinical Data Archives

省略形の記載法 「NCDA」
省略形の呼称　**「クリニカルアーカイブス」**

41病院で来院患者ベース　94万人／年　17,800床のデータベース

18

NCDAシステムフロー

各病院は標準化（拡張）ストレージ／標準化（拡張）トランザクションストレージ／インデックスDBの5つを作成

通常時、前日のTRストレージを本部システムが取りに行き、エラーチェックして本部DBに取り込む

不定期に本部のDB内のデータと病院の標準化（拡張）ストレージに齟齬が無いかバリデーションを行う（患者単位で実行可）

TRストレージでのデータ転送により大量の小さいファイルを転送せずに済み、通信コストが大幅に減る

1日～2日遅れでの取り込みとなり、リアルタイム性は望めない

SS-MIX2　標準化ストレージの構造

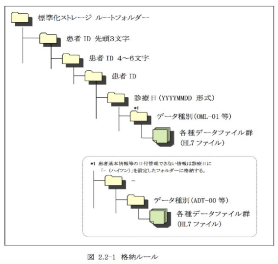


図 2.2-1 格納ルール

（出典）SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン

21

トランザクションストレージ

3.3 3.3.1

トランザクションストレージとは

「標準化ストレージ」は患者(患者 ID)を特定してから、当該患者の診療情報を検索するこ とに特化した物理構造を採用している。しかし、

- 1 何らかの理由で標準化ストレージ再作成しなければならない場合
- 2 災害発生時への対策や地域医療連携の基盤として、外部接続回線を用いてデータセンター等の当該医療施設外に標準化ストレージの複製を作成する場合
- 3 標準化ストレージ以外のシステムにおいて、本ガイドラインで定めた病院情報システムからの伝送データが再利用できると考えられる場合

上記のようなケースでは、診療情報がトランザクションとして標準化ストレージに記録された日時(以下「トランザクション発生日時」という)に着目して診療情報を参照することが必要であると考えられる。したがって、ここでは、病院情報システムから送出される標準化された診療情報そのもの をデータソースとして再利用することによる便宜を考慮して、トランザクション発生日時により診療情報を参照することに特化したストレージとして、トランザクションストレージを規定する。

（出典）SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン

22

トランザクションストレージ

図 3.3-1 格納ルール

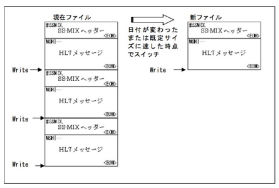


図 3.3-2 トランザクションデータファイルの切り替え

（出典）SS-MIX2 標準化ストレージ構成の説明と構築ガイドライン

23

NCDAにおけるエラーチェック

データ転送後、DB取り込み前に行うエラーチェックは大きく以下の2つ

構文エラー

- SS-MIX2の仕様を満たしているかチェック
  - 型は正しいか、必須項目抜けは無いか、行の順番は合ってるか、項目毎のLengthは守られているか etc
- NHOの要求した仕様を満たしているかのチェック
  - 標準コードは入っているか
  - SN型にあった記載がされているか

毎日チェックした上でエラーを画面に報告、対処を行う

NHOのエラーについては、無視してDBに取り込むことが可能

NCDA本部DBでのデータ修正

基本SS-MIX2の使用に則って、DEL及びINSのSS-MIX(HL7)メッセージを作って登録することでエラー修正を行う。

これにより保存されているファイルからデータベースが再現可能となる

DB設計の際に考慮したこと

HL7要素がすべて格納されること
　→オブジェクト型のDBを利用する
検査結果が抽出可能であること
　→検査結果を単一の値では無く範囲で持つ
いつ抽出作業を行っても同じ結果が出るようにすること
　→すべてのデータにデータ利用可能状態を示す情報に時間範囲を持たせることで実現
抽出処理を行う際にSQL言語が使えること

30

診療情報集積基盤における個人情報取り扱い

1．患者同意
- 病院に掲示されている「個人情報の利用目的」に「国立病院機構診療情報分析基盤での利用」を追加。（平成27年12月中に41病院で実施済み）
- 併せて，ポスター・ちらしでの周知を開始
- 患者の利用不可の申出には対応できるシステムとなっている

2．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法が来年施行見込みであり、今後出てくる政令・ガイドライン等に適切に対応していく
- 研究の倫理指針の見直しがとりまとめられる方向なので、適切に対応していく
- 医療等ID/代理機関等の法令改正が行われた場合にも適切に対応していく

31

参考）postgreSQLにおける範囲型

2013年に組み込みの型としてリリース
(23,34]と記載すると23より大きく34以下という意味となる
中の値としては数値型とtimestamp型が利用可能
‘(3,7)’::int4rangeもしくはnumrange(1.0, 14.0, ‘(]’)と書く

併せて専用の関数が定義されている。

参考）postgreSQLにおける範囲関数

演算子	説明	例	結果
=	等しい	int4range(1,5) = '[1,4]'::int4range	t
<>	等しくない	numrange(1,2,'()') <> numrange(1,1,'[]')	t
>	真実	int8range(1,10) > int8range(2,3)	t
<	より大きい	int8range(1,10) < int8range(1,5)	f
<=	以下	numrange(1,1,'[]') <= numrange(1,1,'[]')	t
>=	以上	numrange(1,1,'[]') >= numrange(1,1,'[]')	t
@>	範囲を包含する	int4range(2,4) @> int4range(2,3)	t
@>	要素を包含する	'[2011-01-01,2011-03-01)'::tsrange @> '2011-01-10'::timestamp	t
<@	により範囲が包含される	int4range(2,4) <@ int4range(1,7)	t
<@	により要素が包含される	42 <@ int4range(1,7)	f
&&	重複する(共通点を持つ)	int8range(3,7) && int8range(4,12)	t
<<	厳密に左に位置する	int8range(1,10) << int8range(100,110)	t
>>	厳密に右に位置する	int8range(50,60) >> int8range(20,30)	t
&<	右端を超えない	int8range(1,20) &< int8range(18,20)	t
&>	左端を超えない	int8range(7,20) &> int8range(5,10)	t
+	隣接	numrange(1.1,2.2) -|- numrange(2.2,3.3)	t
+	結合範囲	numrange(5,15) + numrange(10,20)	[5,20)
*	交差範囲	int8range(5,15) * int8range(10,20)	[10,15)
-	差分範囲	int8range(5,15) - int8range(10,20)	[5,10)

関数	戻り値型	説明	例	結果
lower(anyrange)	範囲の要素の型	範囲の下限	lower(numrange(1.1,2.2))	1.1
upper(anyrange)	範囲の要素の型	範囲の上限	upper(numrange(1.1,2.2))	2.2
isempty(anyrange)	boolean	空の範囲か？	isempty(numrange(1,1,'[)'))	false
lower_inc(anyrange)	boolean	下限は内包されているか？	lower_inc(numrange(1,1,'(]'))	true
upper_inc(anyrange)	boolean	上限は内包されているか？	upper_inc(numrange(1,1,'(]'))	false
lower_inf(anyrange)	boolean	下限は無限大か？	lower_inf('(,)'::daterange)	true
upper_inf(anyrange)	boolean	上限は無限大か？	upper_inf('(,)'::daterange)	true

病院におけるSS-MIX2のデータ精度について

- NHO内の研究チームにおいて本事業の開始「前」から導入されているSS-MIX2モジュールでのデータ精度を調査
- 電子カルテや検査部門システムに残っている検査結果のデータとSS-MIXのストレージ内のデータに齟齬が無いか調査
- 4病院で、それぞれランダムに100人選んでカルテレビュー調査をおこなった。
- 結果、データの一致率は98%を超えた

- NHOとしてのデータ精度の結論
  - データは、間違い無く記載されている。表記の統一がきちんとされているかは別の話。
  - データを受け取るデータベースシステムがきちんと解釈できるかどうかの問題。
  - 上記2点をなるべく汎用的に解決することに取り組むべき

32

診療情報集積基盤における利活用

- 患者に明示した個人情報の利用目的の範囲内で利活用を進める
- 利活用に際しては「利活用要項」を定め、それに従って利用を行う
- 利活用要項の骨子は以下の通り
  - データベース利用審査委員会を設置。データ利用ついて審議。
  - 利活用は匿名化後が原則
  - 研究における利用
    - 本要綱を遵守するとともに、倫理規定等の研究に関連する法令やルールを遵守する

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
3	ADT-01	担当医の変更	ADT^A54

No	データ種別	種別名称	HL7 メッセージ型
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
18	ADT-51	退院予定	ADT^A16
19	ADT-51	退院予定の取消	ADT^A25

No	データ種別	種別名称	HL7 メッセージ型
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
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

No	データ種別	種別名称	HL7 メッセージ型
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

No	データ種別	種別名称	HL7 メッセージ型
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。

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に向けたアプローチの検討

研究分担者　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）

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。

このモデルでは、退院サマリの作成に要する負担を大幅に軽減しつつ、サマリの清書を通じて、自動生成の継続的な精度向上を実現する。さらに、医療用自然言語処理の技術革新に向けた多くの知見をもたらすことが期待される。

A．研究目的

患者が医療機関に入院している際、医師は、回診や検査、処置等の度にカルテ記載を行う。こうしたカルテは入院カルテと呼ばれ、入院の度に一綴りのものが作られる。入院には、経過観察等のために１晩だけ行われるごく短い入院から、１週間や１ヶ月と相応の長さのものがある。それら入院中の経過や、そもそもの入院の理由、退院時の状況等は、医学的に極めて重要な情報となる。そこで、患者が退院した際、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」をまとめる。たとえば、診療所からの依頼により入院をした場合、これらの情報を要約して依頼医に渡すことにより、退院後の治療がよりスムーズに進むであろうことは容易に想像できるであろう。

こうした退院サマリは、最初の外来受診となりうる退院後 2 週間以内に作成されることが望ましいされている。それにも関わらず、評価を受ける病院の公開統計でも 2 週間以内の記載率が 80～90%という数字であることは珍しくない。医師は一般に激務であり、常に病棟や外来の患者に追われている。その結果、患者に直接の影響がないサマリ作成は、後回しにされざるを得ない。もし、入院カルテから退院サマリを自動的に作ることが出来れば、こうした医療現場の負担軽減に繋がる。また、良質なサ

マリの充実は医療の質に貢献すると共に、多くの波及効果を生むことになる。

しかしながら、退院サマリの自動生成は技術的に難度が高い。文章を自動的に要約するための技術や研究は以前よりあるが、それらの多くは、文章の中からポイントとなる文を見つけ繋ぎ合わせる仕組みで動いている。一方、入院カルテの生成に際しては、たとえば、手術をした患者が大きな問題なく経過し術後 1 週間で退院したとすると、入院中のカルテに細かな術後経過が記載されていたとしても、医師は「術後は良好な経過を辿り退院した」と要約しうる。このように、医療における文章の解釈や要約は、単なる文の抜粋ではなく、医学知識を用いての抽象化や一般化が加わる。退院サマリを自動的に作成するに至るまでには、こうした困難な課題を解決していくための効果的な研究アプローチが欠かせないと考えられる。

そこで、本研究分担では、退院サマリの自動生成に向けた研究アプローチの検討に取り組んだ。この分野は、英語圏において多少の関連研究があるものの、自動要約までは実現していない。日本語圏においては、そもそも先行研究自体がほとんど存在しない状態にある。そこで、カルテ要約に際した技術的な知見を一歩づつ得ながら、探索的に研究アプローチを検討した。

B．研究方法

退院サマリの自動生成手法の研究開発に際しては、主に 4 つの観点からアプローチの検討を進めた。まず、文献調査と医師へのヒアリングを通じて、そもそも退院サマリとはいかなるものか、良質なサマリに求められる要件とは何かについて検討を行った。このテーマには、国内外を中心に数十件に上る参考文献が得られた。アンケート調査を元にしたものや診療科を区切らない一般論等、様々な観点からの意見があるため、これら先行研究の整理を通じた研究アプローチの検討を試みた。

平行して、実際の退院サマリの分析を行った。理想的なサマリに求められる要件や必要不可欠な情報を定義したとしても、たとえば、1 泊だけの経過観察入院の場合、定義された情報が得られようがないケースは想起しうる。退院サマリの良し悪しを要件との合致度だけで評価してしまうと、正当なサマリが不当に低く評価されかねない。また、入院には、変化の少ない自動要約に適した入院と、変化が多く自動要約の難易度が高いものがあることが想像された。実際の医師がこれら様々な入院においてどういうサマリを書いているのか、その内容にどういう記述がどれくらい書かれているのかを分析することによって、退院サマリに対するより深い洞察が得られることが期待された。

3 点目として、既存の文書要約アルゴリズム、及び、入院カルテの要約研究をサーベイした。文書を要約するアルゴリズムには、大きく分けて、単一文書要約(single-document summarization)と複数文書要約(multiple-document summarization)がある。退院サマリは、本来的に複数文書要約であることに加えて、入院カルテの自由記載文だけでなく、各種の検査結果に関する複数種類の文書からの要約タスクである点で、独自性が高い。また、文書要約には、文書の中から鍵となるセンテンスを見つけて繋ぎ合わせる抜粋的(extractive)な要約と、複数の記述を高次の概念にて表現する抽象的(abstractive)な要約がありうる。退院サマリの作成に際しては、たとえば、入院までの経過を整理するために入院直接の理由部分を入院カルテから抜粋することも、入院後の経過を医学知識を元に要約することも想定され、これらを組み合わせた総合的なアプローチの必要性が示唆される。しかし、英語圏にて研究されてきたカルテの要約研究は、単一的なアプローチのものが主体であることに加えて、記載された情報を概説する処理(indicative summarization)に主眼があり、情報をコ

ンパクトに要約する試み(informative summarization)は数が限られていること

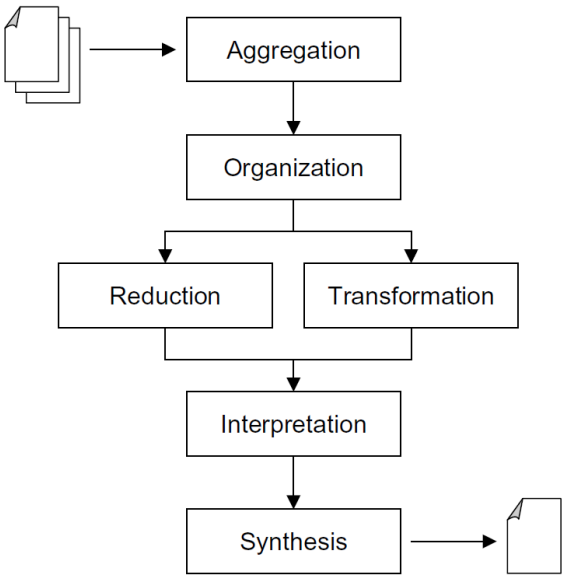


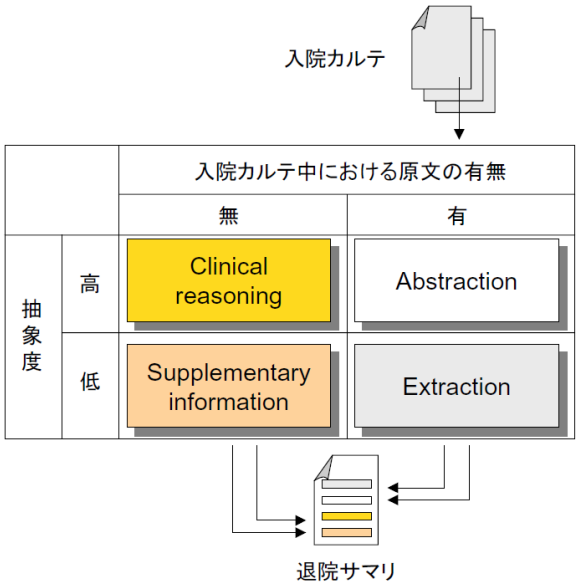
図 1　AORTIS モデル

が明らかとなった。

そこで最後に、上記の知見を総合し、退院サマリの自動生成に向けた独自モデルの検討を進めた。その際、入院カルテの要約過程をモデル化した AORTIS モデルに着目した。これは、(1) Aggregation、(2) Organization 、 (3) Reduction and/or Transformation、(4) Interpretation and (5) Synthesis の頭文字であり、このプロセスを経ることにより退院サマリをシステマティックに作成することが出来るとされる。しかしながら、自動処理に向けたパイプラインとして検討した場合、それぞれの過程における技術的難度が高く、各ステージの性能の低さが全体性能に大きく影響する。さらに、退院サマリにおいては、入院カルテに明示的に記載されていない追加情報が含まれている可能性がある。そこで、AORTIS のモデルの各処理を参考としつつ、より自動要約に適したモデルの構築を目指した。

C．研究結果

今年度の研究の結果、入院カルテの自動要約に向けて、医師が作成する退院サマリの分析枠組みと自動的な退院サマリ生成モデルの双方を兼ね備えた、CASE モデルと証するモデルを構築した。これは、「そもそも退院サマリには何が書かれているのか」という観点より構築されたモデルであり、サマリ中の各文を「カルテに由来するかどうか」という軸と「抽象度が高いか低いか」という軸によって 4 つのクラスに分類する(図 2)。これらのクラスは、退院サマリの分類モデルであると同時に、それぞれ生成に際して固有の処理が求められることから、退院サマリの自動生成に向けた処理モデルとしての性質も有する。以下では、これらクラスの分類軸となる「カルテ記載の有無」、「言及の抽象度・事実度」それぞれを概観した上で、退院サマリの生成モデルとしての展望を記す。

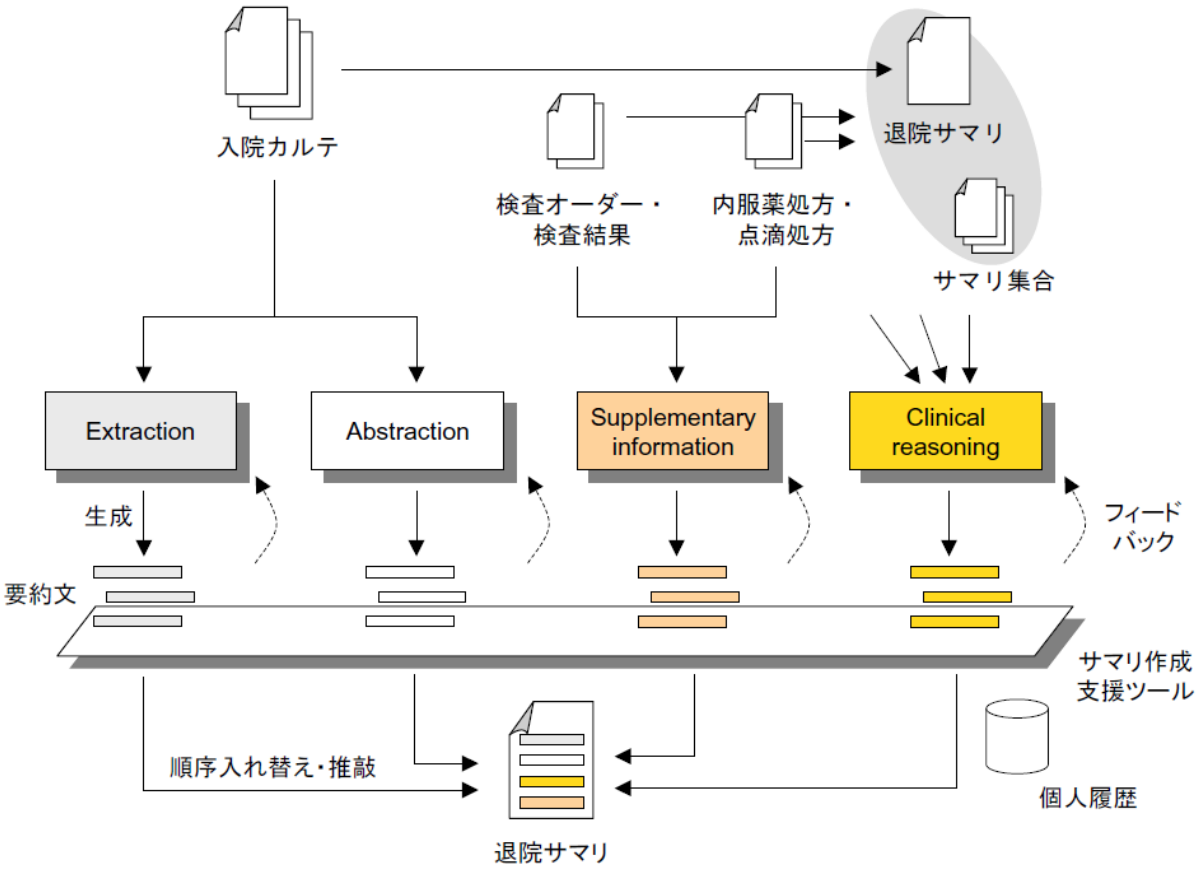


カルテ記載の有無　　退院サマリは、入院カルテに記載された情報を元に作成される。したがって、退院サマリに記載されている情報も、入院カルテに書かれた情報の抜粋ではないかと思われるかも知れない。しかし、退院サマリには入院カルテに直接由来しない文が一定数含まれる。たとえば、

ちょうど退院後のタイミングで検査結果が出た場合や、退院後の療養計画等について退院までにカルテに書き損ねた場合、サマリ中に由来のない文が含まれることになる。あるいは、退院後、入院カルテに書かれた記載を臨床的に評価し、その内容をサマリに記載することもあるかも知れない。いずれにせよ、結果として、入院カルテに直接記されていない文が退院サマリに収載されることになる。今年度の研究を通じて、この「入院カルテ中の原文の有無」について分析を進めたところ、「有」か「無」かではなく、ある程度の濃淡があることが分かった。まず、文の一致として、文レベルで一字一句一致しているのか、あるいは一部変更があるのか。句レベルで一致しているのか、あるいは、類似しているのか。これらのうちどこまでを「有」と取るかによって、区分が変わることになる。また、サマリ文中に複数の句がある場合、それらが入れ子になっている場合等、複雑な事例が少なからず含まれることが明らかとなった。

言及の抽象度・事実度 次に、文における言及の抽象度の分類について記す。退院サマリにおいては、具体的な事実の他に、具体的な事実を抽象化して記載されることが少なくない。たとえば、手術での入院の場合、退院サマリには、入院に至った原因の疾患についての記載に加えて、手術での術式等の情報がまず記載されるであろう。これらは事実に関する言及であり、抽象度が低いと言える。一方、手術後の経過について、日々のカルテに具体的な記載がなされていても、サマリ上では「術後経過は順調であった」と医学的な評価が記載されるケースがある。あるいは、△△処置を何回か行ったとすると、「△△処置を行ったが改善は見られなかった」等と整理して記載するかも知れない。これらは抽象度が高い記述と言える。こうした記載における抽象度の区別については、自然言語処理における事実度の分類技術(factuality classification)が役に立つ可能性がある。ただし、技術的にいくつかの課題があることに留意する必要がある。まず、「骨折を認めた」という記載は明らかに事実度が高いが、「○○の可能性を伝えた」とした場合、伝えたことは事実でも○○についての事実度は低い。同様に、「改善を認めた」といった言及も、医学的な解釈や価値判断を伴っており、骨折のようなケースと比して事実度は低いであろう。このように、事実度を自動的に判定するには、文法と意味の双方を解さなければならない。それでも、退院サマリにおける事実に関する記載の割合や出現位置が分かることで、次に示す退院サマリの自動作成に向けた基礎的な統計が得られることが期待された。

退院サマリの自動生成 最後に、これら退院サマリの分析モデルを用いて、退院サマリの自動生成手法について検討を行った。分析モデルにおける 4 つの区分は、それぞれサマリ作成における戦略の違いに対応している。カルテ中の情報に着目した場合、“Extraction”はそこから重要文を抽出する「抜粋」に相当し、“Abstraction”は、本来の意味での「要約」に該当する。重要文抽出(extraction)は既に多くの先行研究がある分野であり、比較的取り組みやすい。一方、要約操作(abstraction)については、「言い換え」や「重複削除」、「解釈」等、特性の異なるいくつもの手法の総体であると考えられ、技術的難易度が高い。“Clinical reasoning”は、そうしたカルテ記載から医学知識を持って推論や判断を行った結果の記載となる。医学知識に基づいた推論もまた、技術的難易度が高い。しかし、同じ医師や同じ診療科のカルテには往々にして同じような表現が繰り返されることから、大量のデータを機械学習することにより、あ



る程度の有用性が得られる可能性がある。

図 3 退院サマリ作成支援ツール構成図

最後の“Supplementary information”は、入院カルテ中からは直接得ることが出来ない情報であり、検査結果や読影レポート等を対象としたさらなる検討が望まれる。

退院サマリは、これら特性の異なる 4 つの「要約処理」の結果の集合体であり、入院カルテを単一の要約アルゴリズムにより 1 度に高精度な形で生成することは困難と考えられた。そこで、今後の研究の方向性として、高精度な要約アルゴリズムの研究開発ではなく、特性の異なる 4 つの要約モジュールが出力した候補文集合を「退院サマリの下書き」としてユーザーに提示する「退院サマリ作成支援ツール」のプロトタイピングを提案する(図 3)。このツール上では、ユーザーが行う清書作業を、ユーザーからのフィードバックとしてすべて記録することが望ましい。そのデータを用いて各モジュールをチューニングすることで、生成する要約を継続的に改善し、ユーザー嗜好に合致させる処理系の実現が期待される。また、システムを用いて退院サマリを作成する際、利用する専門用語を標準語彙へと半自動的に対応付けながら作成する工夫が望ましい。カルテ中の専門用語は、医療機関やユーザーによって用い方が異なることから、カルテの自動解析処理における障害となっていた。提案手法により、医師に過度の負荷を掛けない形で精度の高い個人辞書の作成が可能となる。とりわけ、各医師の個人履歴には、単語の言い換えに加えて、句から単語への言い換え等、カルテ記載の正規化に向けた豊富な言い換え情報が含ま

れることになる。このデータは、退院サマリ作成支援により効率的に収集が可能であり、医療用自然言語処理に技術革新をもたらすことが期待される。

D．考察

これまでに記したように、退院サマリの自動生成研究は、様々な要素技術が必要であるものの、医療用自然言語処理技術にブレークスルーをもたらしうる研究テーマであると考えられた。以下では、今年度の検討を通じて見出された論点と有用な要素技術について、概要を整理する。

1）退院サマリの基本的な統計の取得

本研究班では、国立病院機構の電子カルテ集積基盤、NCDA に集積する入院カルテと退院サマリデータを網羅的に解析することが出来る。そこでまず、「そもそも退院サマリには何がどれだけ書かれているか」を明らかにしたい。現状では、たとえば、入院カルテから退院サマリが作成されるのに際して文字数がどれだけ圧縮されるのかという単純な統計すら、多施設での検討は進んでいない。退院サマリのセンテンスにおける元の入院カルテに由来する文の割合が明らかとなれば、今後の研究開発に向けた効率的な戦略を検討することが可能となる。CASE モデルが示す記載の事実度の統計も有益であろう。事実記載が多いのか、抽象的な記載が多いのか。これらの比率は、入院カルテと退院サマリで大きく変わるのか否か。CASE のそれぞれのクラスの文は、退院サマリ中、前半や後半のどのタイミングで出現するのか。こうした統計により、医師のサマリ作成タスクにおける特性の解明が期待される。

2）退院サマリ間の類似度評価手法の開発

次に、サマリとサマリの間の類似度についての評価手法の確立が望まれる。文章間の類似度を測る上では、まず、文書毎に利用している単語の利用頻度統計を取得し、その文書毎の単語利用上の類似性を持って文書間の類似度とする手法がある。これは単純な手法ではあるが、ある程度の性能が得られることが経験的に知られている。一方で、こうした形式的な類似度の比較では、同じ分野の退院サマリが類似サマリとして選ばれるものの、入院経過という時間的な軸における類似度の評価が出来ない。また、言及されているトピックや質に関わる類似性を評価することも困難である。これらの類似性指標を確立することが出来れば、以下に示すように「自動要約に適した入院」の特徴を明らかとすると共に、「退院サマリの質の評価」に役立つことが期待される。

3）自動要約に適した入院についての検討

退院サマリ間の類似度を評価することが可能となれば、「退院サマリの自動生成に適した入院」を明らかにすることが出来る。たとえば、退院サマリと入院カルテ中の文重複が高い入院が明らかとなれば、重要文の選択だけで実用性のある退院サマリを生成できるかも知れない。具体的には、産科入院や基礎疾患の無い小児の呼吸器感染症などは、医学的な臨床推論を要する余地が少なく、重要文選択だけである程度の性能が得られる可能性がある。また、糖尿病の教育入院等、退院サマリ同士の類似度が高い入院を発見することが出来れば、それらの退院サマリ集合から高い精度で退院サマリの記載を予測できる可能性がある。一方で、本来は単純な経過を辿るはずのインフルエンザ入院に絞ってみても、実際のカルテを確認してみると、多様な記載が含まれていることが明らかとなった。こうしたカルテのなかから自動要約に適したカルテの種類を発見するためには、カルテそのものの類似度評価以外にも、多くの試行錯誤が求められるかも知れない。たとえば、入院

費の高額な入院はさまざまな処置を行っており経過が複雑な可能性が高いため、自動化に適さず、患者数が多く保険点数が低廉化された入院のサマリほどカルテ記載も薄く自動化が容易であるのかも知れない。DPC 病名毎に平均入院費と入院件数のリストを作ることでも、類似度の高い対象入院をリストアップ出来る可能性がある。

4）退院サマリの質の評価手法

次に、生成された退院サマリを評価する手法が求められる。一つの方向性として、退院サマリに関する文献調査により「理想的なサマリ」に記載されている項目を定性的に整理し、与えられた退院サマリ中にその各項目がどれだけ満たされているかを判定することで質を判断する手法が考えられる。もう一つの方向性として、数多く集めた退院サマリから評価の高いサマリを選び出し、これら優良サマリと各サマリの類似度を定量的に判定することで質を判断する手法も考えうる。なお、これら二つの手法はいずれも退院サマリそのものを評価しているが、そもそものデータ源である入院カルテの質や量が低い場合や、もともとの入院自体が 1 泊だけの経過観察目的でありサマリにもそこまで多くの情報が求められていない場合もある。こうした場合、理想的なサマリとの合致度だけで退院サマリを評価してしまうと正当なサマリが不当に低く評価されかねない。したがって、退院サマリの評価に際しては、サマリそのものの完成度が求められることに加えて、元データからの要約の巧拙や適否に依存する側面もある点に注意を要する。

5）各種記載の正規化手法の検討

カルテに「精神遅滞」と記載されている場合、この語を「精神遅滞」として解釈することはコンピュータにとって容易である。しかし、医師はカルテにおいて様々な表現を行う。「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必要がある。さらに、「学業成績が芳しくない」と記載されたとき、文脈より mild MR と解する柔軟性がなければ、カルテ記載の自動処理を行うことが出来ない。こうした「正規化」処理は技術的難易度が高く、英語圏を中心に研究が進められているものの、日本語圏においては大きく立ち遅れているのが現状である。本研究により、この問題に大きな進展が生じる可能性がある。まず、「表記揺れ」に関する問題について、同一個人や同じ診療科の医師、同じ医療機関の医師では同一の語彙を用いるため、サマリ作成時に個人辞書を整備するアプローチによって解消する可能性が高い。残った問題は、標準語彙とのマッピングと曖昧性解消(word-sense disambiguation)だが、前者については、退院サマリを標準語彙で記載するルールとしたうえで、退院サマリ作成時の訂正履歴を統計処理することにより解決しうる。後者についても、大量の電子カルテデータを用いた共起統計により、ある程度解決するであろう。最後に、より高度な処理が求められる「言い換え」がある。前述のように、カルテにおいては、「学業成績が芳しくない」というように医学的な所見が「用言化」して表記されている場合があり、標準語彙への正規化が困難となっていた。入院カルテから退院サマリを作成する際に、こうした正規化が手作業でなされているとすると、大量の電子カルテを処理することにより自動処理に必要な統計が取得できる可能性がある。また、退院サマリ作成支援ツール上に、この言い換えデータが蓄積していく可能性がある。

E．結論

カルテの電子化により、様々な臨床的情報を自動的に収集し医学研究の発展や医療の質向上に役立てる試みが期待されて来た。しかし、カルテの本体は自由記載にあり、その処理技術が成熟していないことは、電子カルテの効率的な二次利用に向けた大きな制約となってきた。医師の記載には省略や揺れがあることに加えて、解釈に医学知識を要する。そのために、研究に際したコストが大きく､研究が停滞することにより、実用性のあるアプリケーションの開発も制約されて来た。結果として、研究開発投資も伸びない悪循環にあると言える。

そこで本研究では、臨床現場における医師負担の軽減に繋がる退院サマリの自動生成という具体的なアプリケーションを目標に定めた。現在の文書要約技術は、出現する単語の統計的な特徴をうまく捉えることで大まかな要約を行うことに成功している。入院カルテの要約に際しても、これらと同様に、意味の解釈には踏み込まず、統計的な特徴を捉えることで実用性を備えた要約手法を確立することを期待した。

しかし、研究を通じて、カルテの要約が既存の文書要約タスクとは大きく異なる特性を有していることが明らかとなった。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。結果として、CASE モデルと称する退院サマリ記載の分類に基づいた退院サマリの生成モデルを構築することが出来た。このモデルでは、退院サマリに記載されている文をシステマティックに分類し、それぞれのクラスに対して独立して要約アルゴリズムを検討することが出来る。

本手法の発展によって、退院サマリの作成に要する負担を大幅に軽減しつつ、サマリの清書を通じて、カルテ記載内容の正規化と曖昧性解消という自動生成の精度向上に資するデータの収集を行うことが出来る。さらに、継続的な精度向上と生産性の向上を図りつつ、構造化された退院サマリが自動的に蓄積していく環境を実現することが出来る。こうした手法は、現場の負担を下げるとともに、電子カルテデータの価値を高めていくことが出来る妙手であり、医療現場の支持を得やすい。また、最終的な目標であるカルテの自動解析に向けた医療用自然言語処理技術の研究開発に対し、持続的な投資を呼び込むことが期待される。来年度以降、本モデルに基づき、SS-MIX データを対象とした退院サマリ作成支援システムを構築したい。

F．研究発表

1．論文発表

なし

2．学会発表

なし

G．謝辞

本研究に際し、国立病院機構 三重病院 谷口清州 臨床研究部長に、大変お世話となりました。研究班を代表し、深謝致します。

平成 28 年度厚生労働科学研究補助金

（臨床研究等 ICT 基盤構築研究事業）分担研究報告書

退院サマリの自由記載文の特徴解析

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，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。次年度には，他分担が進めた自動要約手法の調査検討結果を踏まえ，退院サマリの自動生成とその評価に取り組む。

A.目的

近年, レセプトや DPC などの大規模な医療データを用いた分析が各所で取り組まれている一方で，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。本研究では，カルテの文章から様々な分析を可能にすることを念頭に置き, 退院サマリ（退院時要約）作成の自動化による記載内容の標準化を目指している。

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。診療行為を大きく外来と入院に分けると，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。なお，退院サマリは 2014 年の診療報酬改定により診療録管理体制加算の条件の１つとして作成が義務づけられており，本加算を算定するためには 9 割以上の退院サマリが退院日翌日から 14 日以内に作成されている必要がある。

退院サマリをいかに自動生成できるかは自明ではない。たとえば，自動要約（automatic text summarization）の技術の応用によって実現できるのか，もしくは自然言語生成（natural language generation）なのか，それすらもわかっていない状況である。そこで今年度は，どのように退院サマリを自動生成しうるかを検討するために，退院サマリがどのように書かれているかを明らかにすることを目標とした。

B.方法

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。

C.結果

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。

D.考察

退院サマリの生成は自動要約なのか自然言語生成なのかという疑問に対して，入院までの経過は前者の傾向が強く，入院中の経過は後者の傾向が強いという結果となった。入院までの経過は，すべての文が入院カルテに由来する（しかも特定の箇所から丸ごと写している）退院サマリが約半数（6/13）あった。入院までの経過は，入院を判断する際に本人や家族，かかりつけ医などから聴取した内容がほとんどであり，また，入院中にすべて書くことができるものであるため，こうした傾向にあると考えられる。ただし，残りの約半数の退院サマリの入院までの経過では，タイプ１以外の文の多くはタイプ４か５であり，これらは自動生成の難易度が高いと考えられる。

入院中の経過は，治療の試行錯誤の過程が書かれており，退院サマリを記載する際にそれらをまとめる必要が生じる。この際に，医学的な知識をもって解釈をする必要がある（タイプ４）と判定されることが多

かった。たとえば,「呼吸器の状態は悪化なく経過した」のように書かれることがたびたびあり，これは退院サマリでは重要な情報である。しかし，入院カルテの日々の記録には, 悪化があればその旨が書かれるが, 悪化があると書かれていないときには，入院カルテ全体を読んで医学的な知識をもって解釈をすることでようやく悪化がなかったと判断することができる（医学的な知識がなくとも推測することができる場合はあるが，確信をもってそう判断することは難しい)。こうした文がどのように自動生成できるかは，今後の大きな課題となり得る。

入院カルテには書かれていない記載（タイプ５）が全体で 2 割強に上ったことは, 退院サマリの性質から考えると好ましいことではないかもしれないが，多忙な診療現場の事情を勘案すると避けられないと考えられる。ただし，医師によってこの事情は異なるかもしれないとも考えられ，今後，様々な医師が記載した退院サマリを同様に分析することでこの傾向についてさらに多くのことが明らかになると期待される。

E.結論

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。

今年度の調査で得られた結果および他の研究分担者が進めた自動要約手法の調査結果を踏まえ，次年度以降は，どのような方法で退院サマリを自動生成するかを検討し, 退院サマリの自動生成および生成された退院サマリの評価に取り組む。

表 1

Doc No	Cycle	入院までの経過／現病歴 (割合)				
		Type 1	Type 2	Type 3	Type 4	Type 5
0004	1	0.333	0.000	0.167	0.333	0.167
0005	1	0.200	0.000	0.000	0.200	0.600
0010	1	1.000	0.000	0.000	0.000	0.000
0011	1	1.000	0.000	0.000	0.000	0.000
0014	1	0.200	0.200	0.400	0.200	0.000
0001	2	1.000	0.000	0.000	0.000	0.000
0003	2	1.000	0.000	0.000	0.000	0.000
0008	2	1.000	0.000	0.000	0.000	0.000
0012	2	0.778	0.000	0.000	0.000	0.222
0015	2	0.714	0.000	0.000	0.143	0.143
0016	2	0.375	0.000	0.000	0.125	0.500
0018	2	1.000	0.000	0.000	0.000	0.000
0019	2	0.714	0.000	0.000	0.286	0.000
Avg		0.720	0.012	0.037	0.098	0.134
Avg (total)		0.428	0.034	0.087	0.240	0.212

表 2

Doc No	Cycle	入院中の経過 (割合)				
		Type 1	Type 2	Type 3	Type 4	Type 5
0004	1	0.222	0.000	0.000	0.222	0.556
0005	1	0.143	0.143	0.143	0.429	0.143
0010	1	0.000	0.111	0.333	0.333	0.222
0011	1	0.143	0.143	0.286	0.000	0.429
0014	1	0.300	0.000	0.300	0.200	0.200
0001	2	0.000	0.000	0.000	0.800	0.200
0003	2	0.000	0.000	0.250	0.500	0.250
0008	2	0.571	0.071	0.000	0.357	0.000
0012	2	0.200	0.000	0.067	0.600	0.133
0015	2	0.500	0.000	0.500	0.000	0.000
0016	2	0.231	0.077	0.077	0.385	0.231
0018	2	0.167	0.083	0.000	0.083	0.667
0019	2	0.294	0.000	0.059	0.353	0.294

Avg	0.238	0.048	0.119	0.333	0.262

表 3

Doc No	入院までの経過／現病歴（カウント）					入院中の経過（カウント）				
	Type 1	Type 2	Type 3	Type 4	Type 5	Type 1	Type2	Type3	Type 4	Type 5
0004	2	0	1	2	1	2	0	0	2	5
0005	1	0	0	1	3	1	1	1	3	1
0010	4	0	0	0	0	0	1	3	3	2
0011	5	0	0	0	0	1	1	2	0	3
0014	1	1	2	1	0	3	0	3	2	2
0001	8	0	0	0	0	0	0	0	4	1
0003	2	0	0	0	0	0	0	1	2	1
0008	10	0	0	0	0	8	1	0	5	0
0012	7	0	0	0	2	3	0	1	9	2
0015	5	0	0	1	1	2	0	2	0	0
0016	3	0	0	1	4	3	1	1	5	3
0018	6	0	0	0	0	2	1	0	1	8
0019	5	0	0	2	0	5	0	1	6	5

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アプローチの検討

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。

このモデルでは、退院サマリの作成に要する負担を大幅に軽減しつつ、サマリの清書を通じて、自動生成の継続的な精度向上を実現する。さらに、医療用自然言語処理の技術革新に向けた多くの知見をもたらすことが期待される。

１．はじめに

本研究の開始にあたり、必要な研究環境の整備と、予備的な調査実験を行った。初年度は年度後半からの開始であり、倫理委員会の承認待ちもあったことから、限られた期間での研究となった。

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。

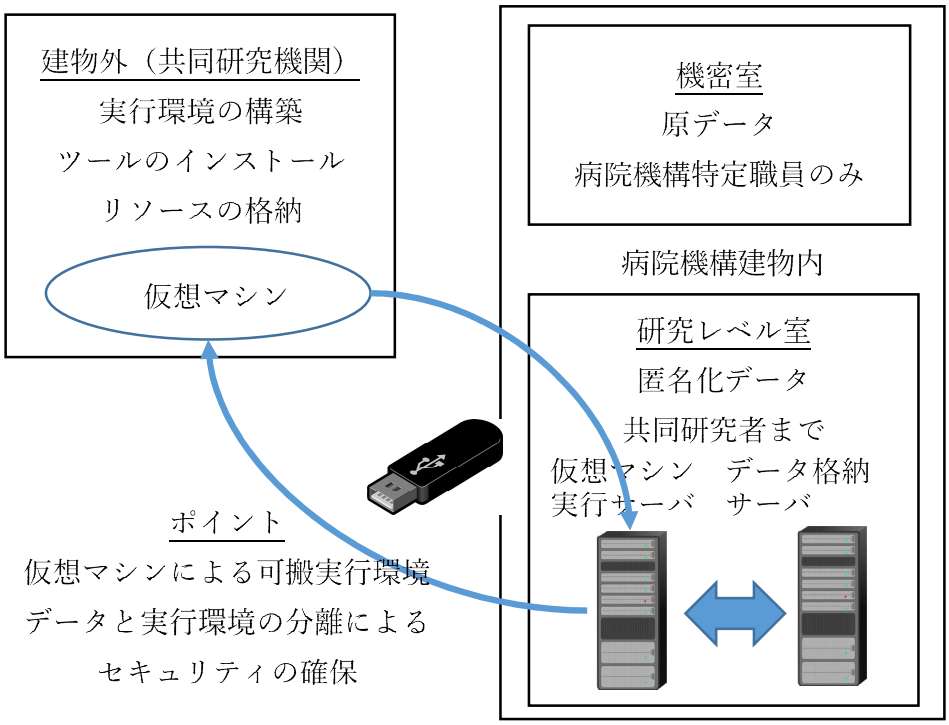
まず、電子カルテという個人情報を扱うことから、厳重なセキュリティ環境が必要である一方、現実的に研究が遂行可能な環境の構築が必要である。

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。他の研究分担者で進められている研究を受け、自動的な処理という観点から、今後の目標とシステム設計について分析を起こった。

さらに、実際の電子カルテに対して、サマリと履歴の間にどの程度の共通性があるのか、予備的な解析を行った。

２．セキュアかつ効率的な研究環境の整備

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。電子カル



テデータを物理的に外部に出さないのが、最も容易な解決策であろう。しかし一方で、他機関の研究者が物理的に訪問滞在可能な期間はごくわずかであり、研究の遂行は現実的には不可能である。本研究分担ではソフトウェアの構築が主となるため、ソフトウェアを遠隔転送して現地で実行をしてもらい、個人情報の含まれない結果のみをフィードバックするという方法も考えられる。しかし、開発中のソフトウェアの実行環境は必ずしも第三者が容易に復元できるものではない。そこで、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する。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する（図）。

本年度は、まずこのような環境整備と実行検証を行った。

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．次年度以降の課題

次年度以降は、既存の手法でどの程度文の選択がカバーできるかを実験し、ベースラインとする。ただし語彙や表現のレベルでは、診療科や病院、医師個人に依存した違いがあるため、その吸収をしなければベースラインとしても妥当な結果を得難いかもしれない。

より根本的な問題として、どのような退院サマリを目指すべきか、という課題がある。サマリの内容やスタイルは、医師、診療科、病院によってさまざまであり、必ずしも唯一の正解があるとも考え難い。ただ、相対的によいサマリというのはあるはずで、その傾向を反映した評価の仕組みが必要である。