



平成 29 年度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。

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出来るようになった。

## 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手入手できる体制」として、全国の国立病院 55 施設より年間 120 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、NCDA 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、運用を開始した。

## C. 結果

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 29 年度末現在 55 病院が参加、約 65000 床、年間実患者数約 120 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。(参考資料に詳細を添付している)

### データベースについて

#### 【国立病院機構 診療情報集積基盤】

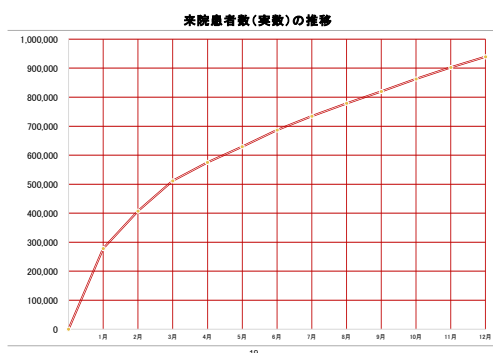
(コクリツビョウインキョウ シンリョウジョウホウシュウセキキバン)

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



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出来るようになった。

## 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

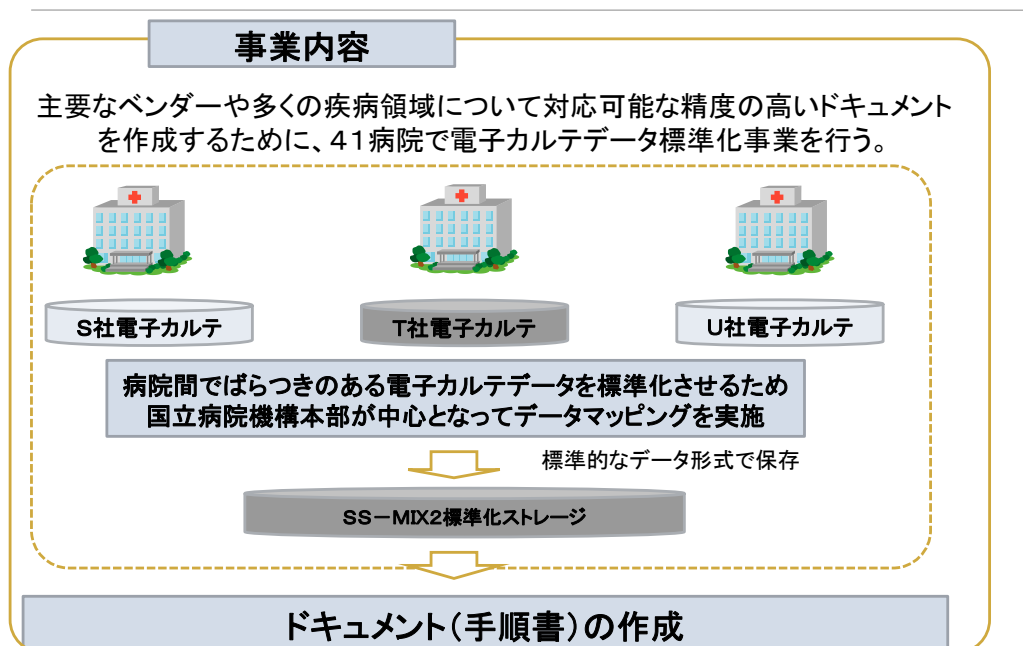
# 国立病院機構診療情報集積基盤について (NCDA:NHO Clinical Date Archives) ～電子カルテデータの標準化～

国立病院機構本部  
情報システム統括部

## NCDA(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。
- 国立病院機構の『電子カルテデータ標準化等のためのIT基盤構築事業(H26補正:13.0億円)』では、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-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

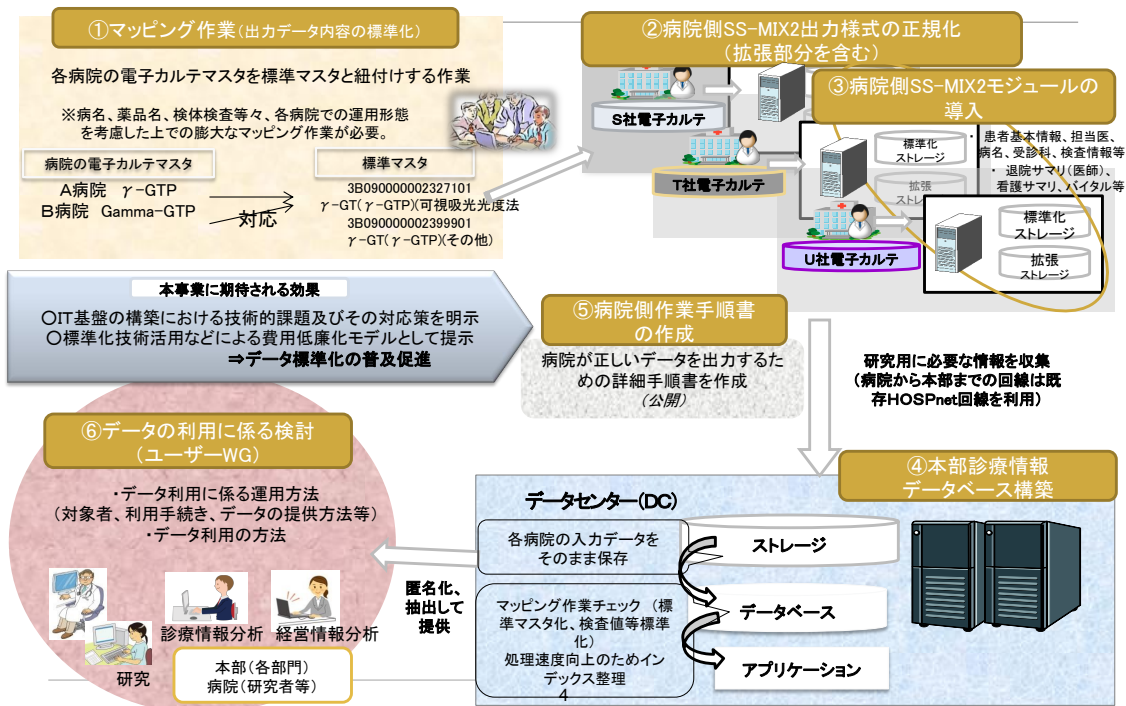
## 国立病院機構のDB事業概要（プロジェクト概要）

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）に対応できるように、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に、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

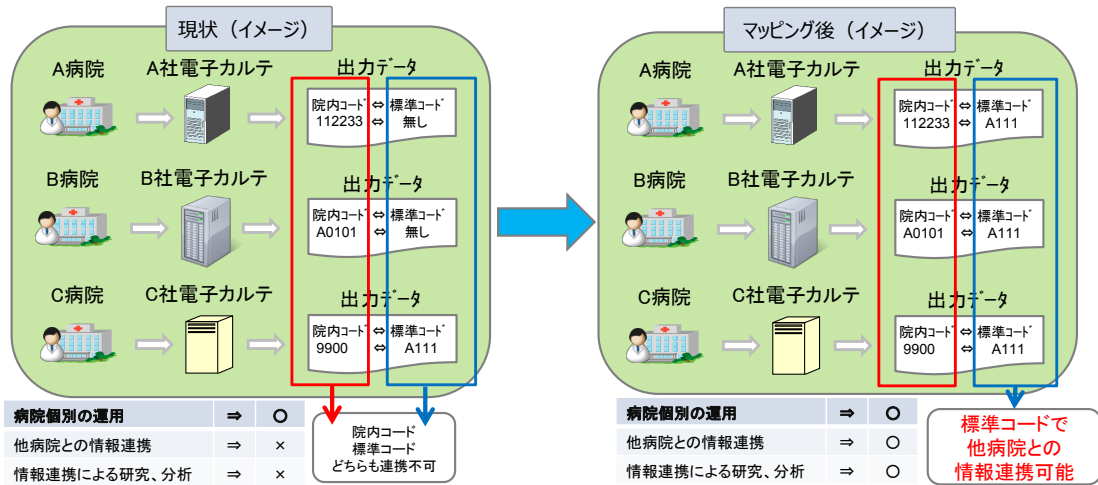
3

SS-MIX2を用いた診療情報データベース構築プロジェクト 作業区分①～⑥

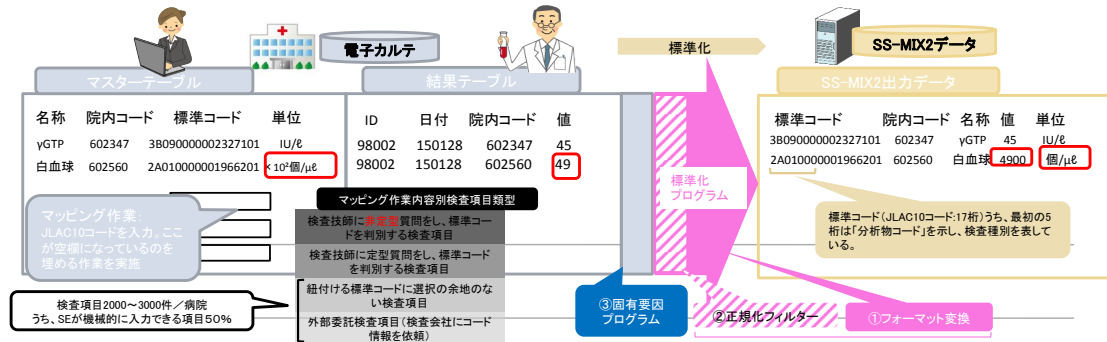


## 病院におけるマッピング作業

- 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。



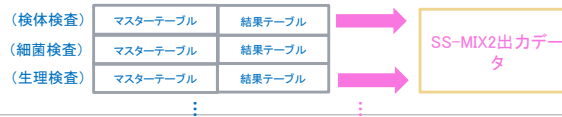
## データ標準化のイメージ(SS-MIX2出力)



- ①フォーマット変換: 電子カルテのフォーマットをSS-MIX2フォーマットに変換する
- ②正規化フィルター: 例えば、同一検査で単位が混在している場合、標準とされる単位に変換する(白血球検査で「個/μℓ」と「×10<sup>6</sup>個/μℓ」が混在している場合、標準の単位が「個/μℓ」であれば、院内電子カルテ上で「×10<sup>6</sup>個/μℓ」の単位にて表されている「値」については、×100してSS-MIX2出力データとする)。他には、使用している文字コードが違う場合、標準とされる方にあわせて変換する、など。
- ③固有要因プログラム: 病院独自の検査表示など実施している場合、出力時に標準に合うように変換する。

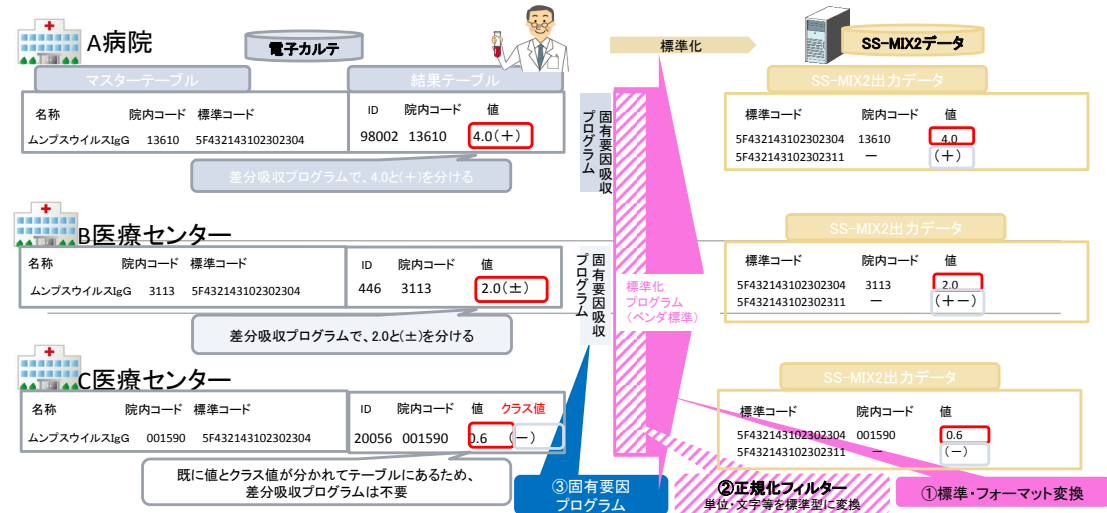
①②は標準化プログラムとして電子カルテ6ベンダーにて開発。  
③は各病院における電子カルテ導入業者が開発。

電子カルテデータからSS-MIX2データへ変換する際、対象となるカテゴリは(検体検査)など数十個存在。  
すべてのカテゴリを変換する標準化プログラムについて、確認・検証する必要がある。



6

## SS-MIX2変換プログラムの構成




- ①②の標準化プログラムは他の施設でも使用する汎用的なもの。
- ③の固有要因プログラムについては病院固有のもの。
- ベンダー側が構築する標準化プログラム①②に固有要因プログラム③の機能が含まれていると、その病院でしか使用できない(汎用化されていない)ことになり、普及促進を図る手順書としての品質は不可。
- 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)を確認する必要がある。
- ※①②③のプログラムの著作権はベンダーにあるため、コード等中身を見ることができない。よってNHOが結果により確認する必要がある。

7

## 事業の成果(標準化の普及促進関係) H28.3時点

---

- 最新のSS-MIX2Ver1.2cに完全準拠しているモジュールを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る。
  - 本モジュールは主要カルテベンダー7社から他の医療機関にも(有償にて)提供可能。
  - 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る。

---

8

## 病院側の負担について

---

- 本事業の病院募集時において試算したところ、病院へのモジュールの導入費用900万~1000万、翌年度以降の(保守・利用料等の)病院負担額は年額100万円以内と設定
- 初期投資についてはベンダー側初期提案額から1~2割の低減を行った。(平均700万円台)
- 本事業内で低廉化について調整したところ全病院年額20万円台で運用できることで調整がついた。
- 電子カルテ更新時のコストについては今後とも交渉を行っていくが、さらなる低廉化がはかれるよう努力していく。

---

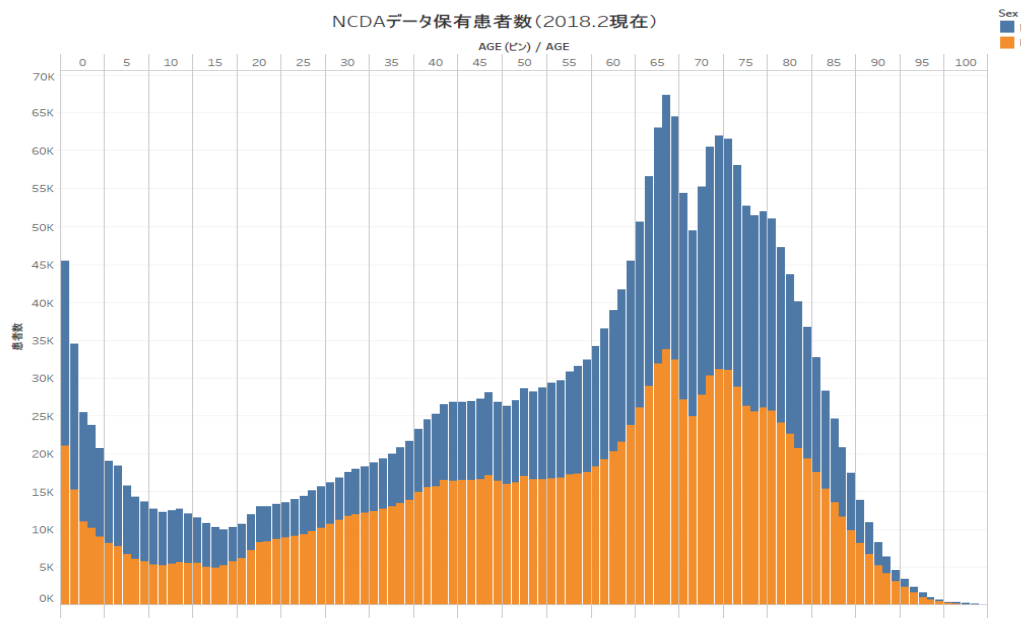
9

## NCDAの現状

- ・平成26年度補助金「電子カルテ標準化等のためのIT基盤構築事業」にて6ベンダー41病院で事業開始(平成28年1月～)
- ・平成29年度末に7ベンダー58病院に拡大
- ・年間患者数約120万人規模のデータ基盤  
平成30年2月末時点で約160万人のデータを蓄積  
(患者数は実患者数、外来データ含む)
- ・一般臨床レベルのデータクオリティはある。  
(例えば、検査データは、全て検査値が統一されアーカイブされている。)
- ・現在安定して稼働中

10

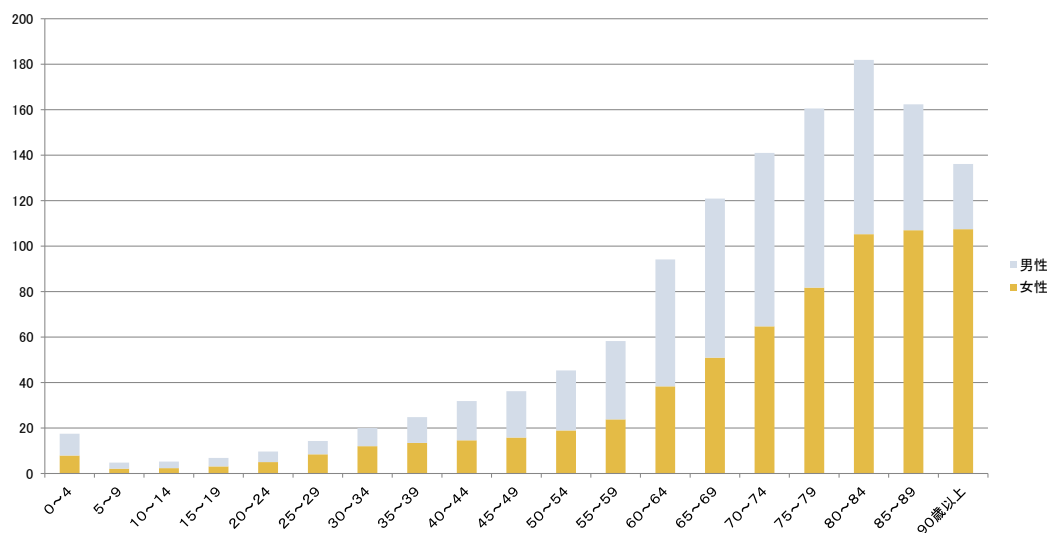
## NCDAに集積している患者数(年齢・性別)の状況



12



## (参考)年齢・性別推計入院患者数(病院)



(単位:千人)

(出典)H26厚生労働省患者調査

13

## NCDAと災害診療記録プロジェクト概要①

### 1 事業名称と予算

◆事業名称:

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」

◆予算総額: 2.5億円(繰越) ※本年度中の執行期限

- ・データセンターにおけるデータ提供モジュール構築に必要な経費
- ・SS-MIX2モジュールのバージョンアップ等に必要な経費 ※病院の電子カルテに関係する部分
- ・手順書作成に必要な経費

◆事業実施主体: 独立行政法人国立病院機構(単独)

### 2 事業の背景

○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。

○この課題に対し、東日本大震災を契機にして「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**」が作成されています(熊本地震で初めて運用開始)。

(※)日本医師会、日本集団災害医学会、日本病院会、日本診療情報管理学会、JICA

## NCDAと災害診療記録プロジェクト概要②

### 3 事業の内容

- 電子カルテデータ標準化等のためのIT 基盤構築事業（NCDA）の機能をバージョンアップし、様々なベンダの電子カルテから、自動的に「災害診療記録」電子フォーマット出力が可能となるように開発及び検証を行い、更に導入手順書を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つもの。

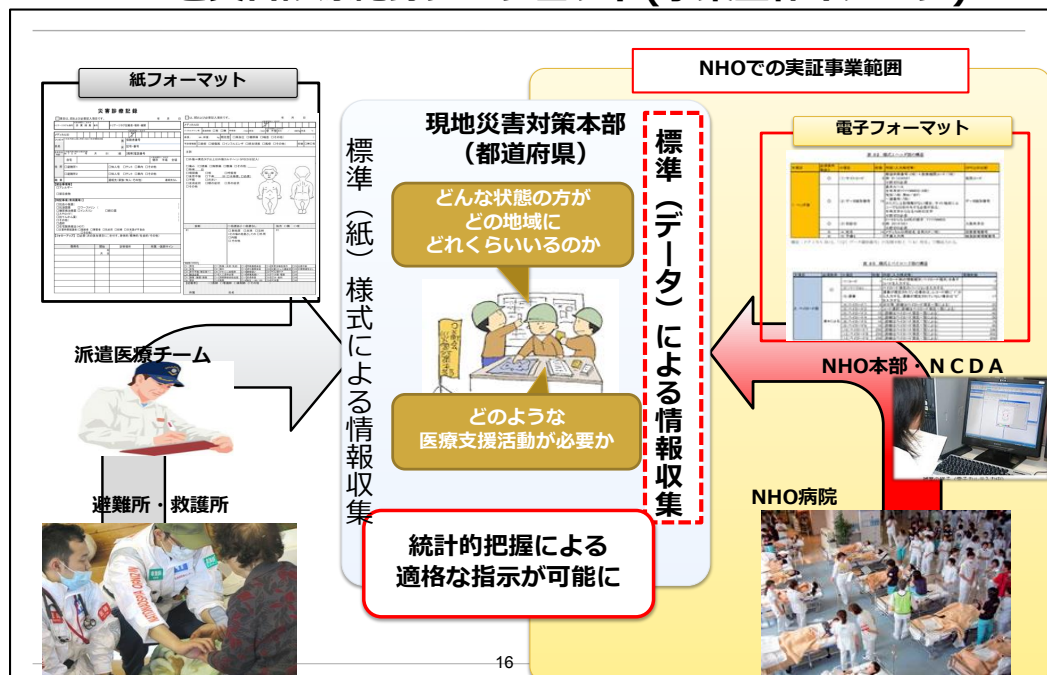
### 4 国立病院機構で実施する理由

国立病院機構は、

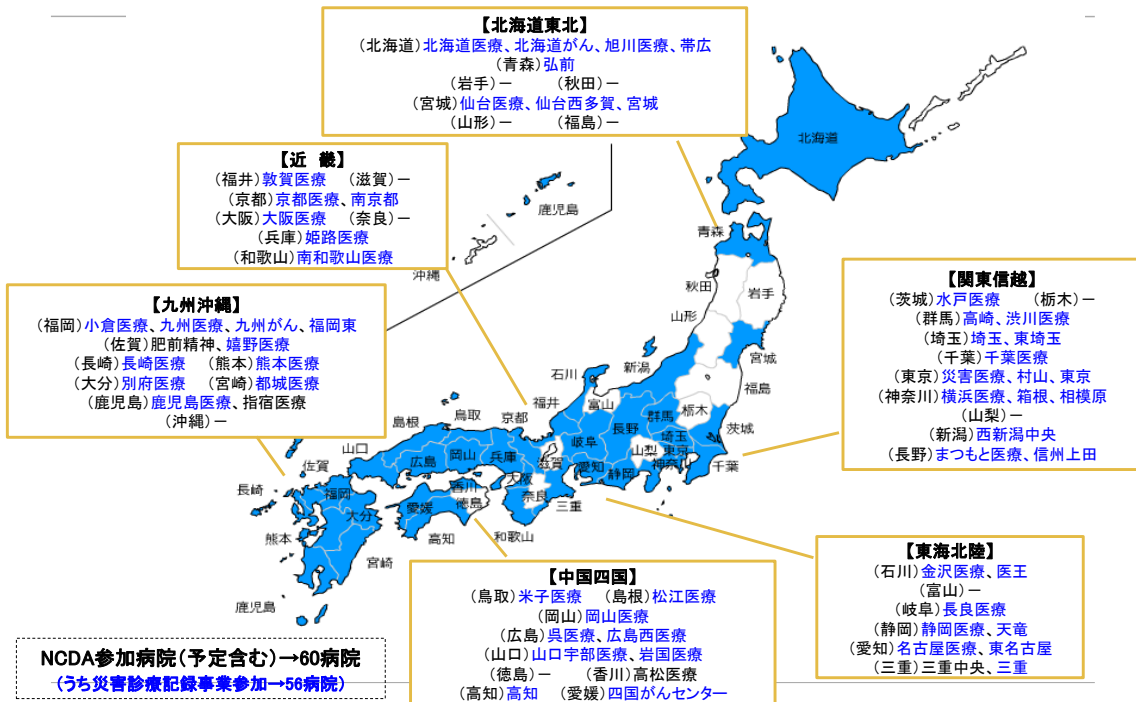
- ✓ 災害対策基本法に基づく指定公共機関
- ✓ NCDAを活用し、効率的に災害診療記録を収集する仕組みの構築が可能
- ✓ 災害診療記録電子フォーマットの普及啓発が可能  
(事業の成果のドキュメント化(手順書の作成等))

15

## NCDAと災害診療記録プロジェクト(事業全体イメージ)



## (参考)NCDA参加病院の状況



## 災害診療記録プロジェクトの成果

✓標準的な情報収集における情報の種類の拡大

ONHOのNCDAにおいて現在収集している標準ストレージ内項目36種類(全て)とバイタル情報に加え、経過記録・退院時サマリについて収集できるようにするための病院向けモジュール改修を実施

○JAHIS標準に準拠した形で定義(HL7 V3. 0 CDAベース)

○本研究費では開発・検証までを実施

○NCDA参加7ベンダーのモジュールの開発及び検証環境下での検証は終了

## NCDA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。
- 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、ガイドライン等に適切に対応
- 研究の倫理指針に適切に対応
- 次世代医療基盤法にも適切に対応していく

---

20

## NCDAの利活用について

---

- 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
  - NHO本部内に、データベース利用審査委員会を設置し、データ利用について審議。
  - 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

---

21



## 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	食事オーダー	OMD^O03
25	OMP-01	処方オーダー	RDE^O11
26	OMP-11	処方実施通知	RAS^O17
27	OMP-02	注射オーダー	RDE^O11
28	OMP-12	注射実施通知	RAS^O17
29	OML-01	検体検査オーダー	OML^O33
30	OML-11	検体検査結果通知	OUL^R22
31	OMG-01	放射線検査オーダー	OMG^O19
32	OMG-11	放射線検査の実施通知	OMI^Z23
33	OMG-02	内視鏡検査オーダー	OMG^O19
34	OMG-12	内視鏡検査の実施通知	OMI^Z23
35	OMG-03	生理検査オーダー	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生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)
- ・ $\mu$  → u (小文字ユー)
- ・語尾に名称 → () で
- ・ $^{\circ}\text{C}$  → cel
- ・‰ → permil
- ・個 → pcs

【上記ルールの適用例】

- ・ mL → mL (ASCII コード)
- ・  $\text{X}10^2/\mu\text{l}$  → .10\*2/uL (かける、乗、 $\mu$ )
- ・ /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
1A0250000001272 01	10.5	mg/l	→	1A025000000127 201	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せら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せら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せられた例

検査名称	院内コード	結果値
ムンプス 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の課題設定である。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ク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ク

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



## 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。まず、実カルテへのアクセスには十分なセキュリティが求められるため、セキュアな処理環境を準備した。

また、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、サマリ生成に必要な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、分析を行った。

### 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. 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。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。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。今後はこれをより大規模なデータに

適用し、実用化を図る。

### 4. 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。

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。

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。結果、ある程度の類似クラスタの形成が見られたが、より安定した分析にはもっと多くのデータが必要と考えられるため、今後はさらにデータ量を増やし分析を進める。

## 5. 次年度以降の課題

次年度以降は、これまでの研究成果を踏まえサマリの生成にむけた実装を行う。それにあたっては、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ものはあるはずで、その傾向を反映した評価の仕組みが必要である。班内の他研究グループの成果を活用し、自動評価を試みる。



## 理想的な退院サマリおよび退院サマリの自動評価に関する調査研究

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、そのために必要な2種類の調査研究を行った。退院サマリを生成するためには、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」の定義が求められる。そこで、理想的な退院サマリについて言及をした国内外の研究の文献を調査した。また、自動生成をした退院サマリが意味のある退院サマリとして成立しているかは評価が必要である。そこで、要約文や自由記載文の自動評価手法に関する国内外の研究の文献を調査した。PubMed や医中誌などを用いて文献を検索し、理想的な退院サマリに関する文献を51報（英文48報，邦文3報），要約文や自由記載文の自動評価手法に関する文献を243報（英文：148報，邦文95報）得た。理想的な退院サマリに関する文献では、記載されることが望ましい項目は何か、各項目が実際の退院サマリに記載されているか、項目ではなく文章としてどのような特徴を有していることが望ましいのか、などが言及されていた。一方、要約文や自由記載文の自動評価手法は、評価の際に正解となる退院サマリが用意できる場合に利用できる手法と、用意できない場合に利用できる手法とに分けることができた。研究で使用する場合には前者が適用できるが、実臨床で使用する場合には後者が必要になると考えられる。次年度は、本研究で得られた知見を整理し、他分担グループと連携をして退院サマリの自動生成および評価に取り組む。

### 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。

退院サマリを自動生成する方法は自明



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。

そこで今年度は、理想とされている退院サマリについて記載された文献を調査すること、および要約文や自由記載文の自動評価に関する先行研究を調査することを目的とした。

## B. 研究方法

理想的な退院サマリの調査のために、英語の文献を PubMed で検索、日本語の文献を医中誌で検索した。退院サマリに関する文献を検索して一覧を作成し、その中から本研究に関係が深いと思われる文献を抽出した。本研究に関係が深いかな否かは、文献のタイトルとアブストラクトから複数の作業員が独立に判断し、判断が一致しなかった場合には第三者（他の作業員）も含めた話し合いによって判断を決定した。抽出されたすべての文献について、全文を入手し、研究の趣旨、研究の結論、退院サマリに書かれるべき（とその文献で述べている）項目、の各点を整理した。この作業中に本研究には適さないものは除外した。

一方、要約文や自由記載文の自動評価手法の調査のために、英語の文献を PubMed, ScienceDirect, Google Scholar, Google で検索、日本語の文献を CiNii Articles で検索した。要約文や自由記載文の自動評価

に関連するキーワード、および先行調査で見つけた文献から拾い出したキーワードを用いて検索をして一覧を作成し、その中から本研究に関係が深いと思われる文献を抽出した。また、文献検索サイトでの検索以外の方法として、検索で抽出した文献から引用されている文献、それらを引用している文献から、関係のありそうな文献を抽出した。

## C. 結果と考察

理想的な退院サマリに関する文献の検索およびその後の精査によって、最終的に 82 報の文献が得られた（内訳は、英語の文献が 73 報、日本語の文献が 9 報）。これらの全文からさらに本研究に適切な 51 報（英語の文献が 48 報、日本語の文献が 3 報）を対象とした。

得られた文献には、退院サマリに書かれている項目を調べた文献、医療従事者などによって項目の評価を行っている文献、項目ではなく退院サマリの文章として望ましい特徴を挙げている文献などがあつた。

具体的な項目として、たとえば、入院の理由 (Reason for admission)、入院中の治療内容 (Treatment in hospital)、退院時の患者の状態 (Patient condition or functional status on discharge)、退院時の診断名の一覧 (List of medications on discharge) などが挙げられている。文献によって、数項目の場合もあれば、数十項目の場合もある。

各文献に記載されている項目は必ずしもその文献で独自に設定されているわけではなく、他の文献やガイドラインの項目を引用している場合もある。たとえば、実際の退院サマリの評価をするためにそれらの項目を利用したり、それらの項目の優先度をアンケート調査によって評価したり、といった研究があつた。一方で、独自に設定さ

れた項目の場合には、それらの項目が書かれるべき根拠が言及されていないこともあった。診療領域によって記載内容や記載量が大きく異なることも示唆されており、これも検討の余地がある。

項目ではなく文章の特徴について言及されている場合には、たとえば、正確性 (Accuracy)、完全性 (Completeness)、適時性 (Timeliness)、文法 (Grammar) などが挙げられている。

今後、51 報の文献のそれぞれに挙げられた退院サマリに求められる項目や特徴、およびそれらに対する評価を整理し、まとめることが必要である。また、ほとんどが海外の研究であるため、日本の医療事情に合わせた検討も必要と思われる。

要約文や自由記載文の自動評価手法に関する文献の検索およびその後の精査によって、243 報の文献が得られた (検索サイトから 148 報、文献の引用・被引用から 95 報)。

要約文や自由記載文の評価手法は、退院サマリへの手法の適用を想定した場合には、2つの分類軸によって分類することができると考えられた。1つの分類軸は、評価の際に「正解文 (もしくは正解要約)」を必要とするか否かであり、もう1つの分類軸は手法開発の際に学習データを必要とするか否か (機械学習を利用するか否か) である。

評価の際に、正解となる要約文を用意することができる場合には、その正解との距離 (類似度) を計算することによって自動生成された要約文を評価することができる。この種類の評価方法としては 2004 年に Chin-Yew Lin が発表した ROUGE が最も有名であるが、それ以外にも様々な方法が提案されている (BLEU, ROUGE, Basic Element, SERA, JSD など)。2016 年に北欧の研究グループが退院サマリ自動生成に関する研究を報告しているが、この中では

生成された退院サマリの評価に ROUGE を用いている [Moen et al, 2016]。ただし、複数の退院サマリの自動生成手法を比較した研究においても一口に要約文と呼んでも、その生成方法や要約の程度などは分野によって多様性があるので、退院サマリの要約に ROUGE などの従来手法をそのまま適用して満足のいく結果が得られるかはわからない。また、実臨床で退院サマリを自動生成し、それが適切な退院サマリになっているかを評価する、といった場面で利用をする場合には、正解要約が必要な手法は利用することができない。しかし、研究においては正解となる退院サマリを用意することが可能な場合があるため、そうした場合には有効な方法となり得る。

正解となる要約文を用意することができない場合には、別の考え方が必要となる。要約文の評価の場合には、原文と比較をする方法がある。たとえば、潜在的意味解析法 (LSA) による要約文の評価が研究されている。それ以外の方法では、主に試験の記述式問題 (短答式も含む) の自動採点において様々な研究開発が行われているようである (Fresa, e-rater, Jess, LSA, JSD など)。アメリカでは以前より大規模な学力試験 (GMAT, GRE, TOEFL など) において用いられているようで、日本においてもセンター試験の後継となる「大学入学共通テスト (仮称)」において記述式問題が採用されることから研究開発が進められている [石岡, 2016]。これらにおいては、修辞、論理構成、内容などの文章の複数の評価基準を設定し、それらを実験的に評価する、といった考え方が採られる。たとえば、GMAT (ビジネススクールの出願者へ課される共通学力試験) で採用されている e-rater では、「Structure」「Organization」「Contents」という 3

つの軸で評価をする。また、PEG (Project Essay Grade) では「Contents」「Organization」「Style」「Mechanics」「Creativity」の5つの軸である。評価者(人)によって評価をした結果が多数得られる場合には、それらを正解として評価器を学習することができる。しかしながら、退院サマリの場合には、「退院サマリを定量的に評価する(採点する)」ことは日常的に行われているわけではなく、その基準が示されているわけでもないため、これは容易ではない。

実臨床で退院サマリを自動生成することを想定した場合には、正解となる退院サマリを必要とせず、また現時点では手法開発の際に人による正解データを大量に必要としないような自動評価手法が望ましい。一方で、研究において正解となる退院サマリを用意できる場合には、正解となる退院サマリとの距離を計算する手法を用いることもできる。

## D. さいごに

今後、理想的な退院サマリの要件をまとめ、自動生成手法の開発グループへ提供することで、生成の方針立案に寄与すものと期待できる。同時に、人手で生成された「理想的な退院サマリ」が、本研究によって抽出された要件を満たしているか否かを比較することは、文献に書かれた理想と現場の感覚との差異を考察したり、また海外と日本の差異を考察したりし、より現実的な理想的な退院サマリの要件の設定につながると考えられる。

また、自動生成された退院サマリの評価を行うために、退院サマリの自動評価手法の実装を行うことを予定している。この際、正解となる退院サマリを必要としない手法の実装には、前述の理想的な退院サマリの要件を参考にする。

## E. 研究発表

### 1. 論文発表

なし

### 2. 学会発表

なし



## 退院サマリの自動生成に向けた研究基盤の構築

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

入院治療を受けた患者が退院するのに際して、医師は、記載していた入院カルテを要約して退院サマリを作成する。入院カルテから退院サマリを自動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、さまざまな波及効果が期待される。本研究班は、昨年度よりこの研究テーマに取り組み、昨年度、退院サマリの自動生成に向けたアプローチを検討した。その結果、退院サマリの自動生成を実現する CASE モデルと称する処理モデルを構築した。

今年度は、このモデルの実証と活用に向けた研究基盤の構築を進めた。まず、研究の推進に際して、自由に研究利用が可能なダミーカルテの整備を行うと共に、アノテーションガイドラインの開発を試みた。また、CASE モデルの実証に求められる「カルテ中センテンスの抽象度分類器」の準備を進めた。さらに、退院サマリの生成システムのユーザーインターフェースを設計し、プロトタイプングを行った。

入院カルテの自動要約に向けた研究は、医療用自然言語処理研究において未開拓な領域であり、まずは研究基盤の構築と合わせた研究方法論の確立が求められる。昨年度、今年度の研究を通じて、研究アプローチの検討と研究基盤の整備が進んだ。

### A. 研究目的

入院治療を受けた患者が退院するのに際して、医師は、記載していた入院カルテを要約して退院サマリを作成する。この作業は、診療上いくつかの意義を有するもので欠かすことができないが、一方で、医師の負担となっている。もし、入院カルテから退院サマリを自動生成することが出来れば、臨床医の診療負担の軽減に資する。また、別掲するように、カルテに含まれる自由記載文の自動解析に向けた技術革新をもたらすことが期待され、医療用人工知能の発展の鍵となる可能性を秘めている。

本研究班は、昨年度よりこの退院サマリの自動生成に取り組み、まず、研究アプローチの検討を行った。そのために、文献調査と医師へのヒアリングを通じて、退院サ

マリに求められる要素の定性的な分析を行った。次に、医療機関への訪問を通じて実際の退院サマリの分析を試みた。また、既存の文書要約技術に加えて、入院カルテの要約研究のサーベイを行った。そのうえで、退院サマリの自動生成に向けたアルゴリズムの検討を行った。

その結果、退院サマリの自動生成を実現する CASE モデルと証する処理モデルを構築することが出来た。このモデルは、退院サマリ中の各センテンスを「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。たとえば、元となる入院カルテに含まれる所見を抜き出して退院サマリに加えた場合、「元カルテに由来」した「抽象度の低い」センテンスとなる。あるいは、入院中の病態について推論し、退

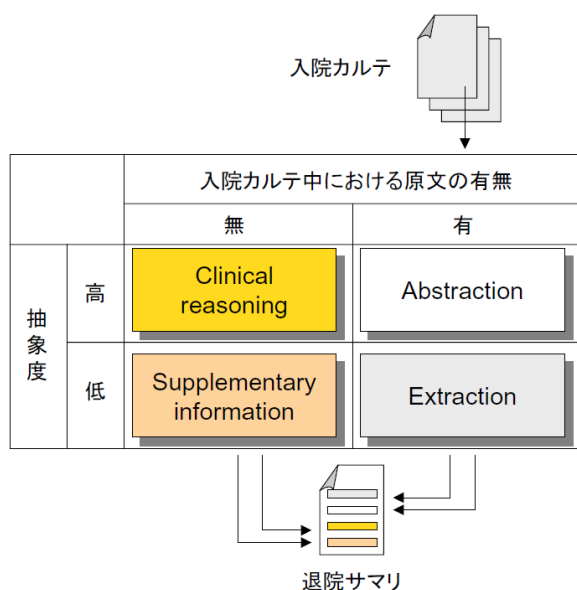


図 1 CASE モデル

院サマリに書き加えたセンテンスは、「元カルテに由来しない」、「事実度の低い」センテンスとなる。このように、CASE モデルは、入院カルテの自動要約に向けて、医師が作成した退院サマリの分析枠組みを提供すると共に、それぞれのセンテンスの生成方法を示唆するモデルとなっている(図 1)。

今年度は、このモデルの実証と活用に向けた研究基盤の構築を進めた。まず、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より成り立つのか、実際の退院サマリの分析を通じて実証する必要がある。しかしながら、膨大な退院サマリを統計処理するためには、自動処理が不可欠であり、退院サマリ中のセンテンスの効果的な分類器が求められる。そこでまず、分類器の研究開発に求められるダミーカルテの整備を行った。また、センテンスの分類に際して、「元カルテに由来するか否か」の検証は技術的に容易であるが、「記載の抽象度」の分類は難度が高い。そこで、この退院サマリ中の記載の抽象度の解析に向けた分類器の構築を試みた。さらに、CASE モデルに基づいて退院サマリを構築するための、退院サマリの自動要約システムの開発を進めた。

## B. 研究方法

退院サマリの自動生成研究には、研究利用できる入院カルテと退院サマリが必須となる。我々の研究班では国立病院機構の診療情報集積基盤(NCDA)に集積される電子カルテデータが活用できるが、高度な個人情報であるため施設外への持ち出しを行うことができない。これは解析に要する各種ツールのチューニングにとって非効率であるため、研究の推進に際して自由に研究利用が可能なダミーカルテが求められる。そこで、本分担において、ダミーカルテの収集・整備を担当した。また、このダミーカルテの特性を解析すると共に、NCDA を解析する方法論の精度管理のため、ダミーカルテ中の各センテンスに対して、上述の 2 軸分類データの付与を試みた。

次に、2 軸分類に求められるカルテ記載の抽象度分類に向けて、自然言語処理研究者への相談を進め、研究アプローチの整理を行った。相談に際しては、遠隔会議に加えて、研究室を訪問してのプレゼンテーションを行い、研究班への協力者拡大を図った。具体的には、奈良先端科学技術大学院大学、首都大学東京システムデザイン学部、産業技術総合研究所人工知能研究センター、北陸先端科学技術大学院大学、理化学研究所革新知能統合研究センターを訪問し、討議を重ねた。

また、上記の検討の後に必要となる入院カルテの自動要約システムの開発を進めた。まず、昨年度の研究において、入院カルテの自動要約に際しては、特性の異なる 4 種類の要約アルゴリズムを組み合わせる必要があることが示唆されていた(図 2)。また、そのアルゴリズムの出力を組み合わせ、順番を整えるとともに、文意が通るよう適切に訂正する必要が想定されていた。そこで、4 種類の要約エンジンからの出力を所与として、その効率的な編集を実現するユーザーインターフェースを設計し、プロトタイピングを行った。

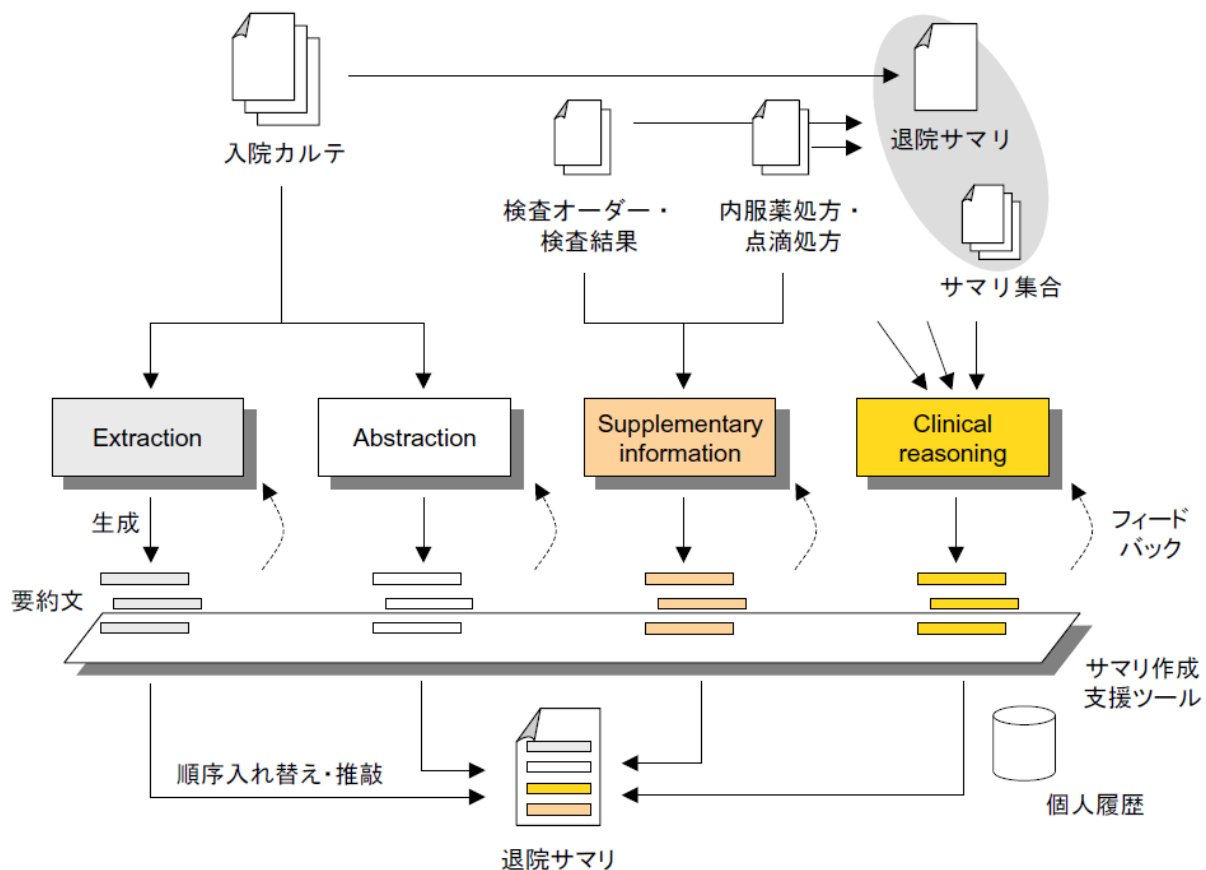


図2 四種類の要約アルゴリズムと要約作成システム

以上の研究に際しては、およそ1ヶ月に1回のペースで研究班会議へと報告すると共に、共同研究者らよりフィードバックを得ながら実施した。

### C. 研究結果

今年度、50組の入院カルテと退院サマリを確保することができた。各入院のデータ量は、最小3.4KByte、平均14.5Kbyte、最大48KByteであった。50件のデータは、統計取得の対象としては少なく、また、原理的にサンプリングにおけるバイアスを排除することが出来ない。しかし、退院サマリに出現するセンテンスの大まかな傾向を明らかとするうえで、自由に研究利用できる貴重なデータ

といえる。そこで、今後の分析と、精度管理用データとするために、暫定的なアノテーションガイドラインを作成し、各センテンスに対するラベル付与を行った。また、今後の研究利用に際しては、文書量の拡大が望ましい。そこで、まずは倍量となる100件の確保を目指して、研究協力者の確保に向けた広報活動と研究協力者の組織化を試みた。

アノテーションに際しては、まず、確かな事実として記載されたセンテンスについて、CT (certain)のラベルを付与した。たとえば、文1はCTに該当する。また、医師が行った推定に対して、PR (probable)のラベルを付与し

た。これらのラベルは、一見、排他的だが、一つの文には複数の節が存在し、接続詞により一文に接続されたり、入れ子等の関係を有している可能性がある。そこで、択一的な分類ではなく、複数ラベルを持ちうる構成とした。

った。医療文書は、利用している語彙や用法が特殊であるため、別分野の一般的な文章によりトレーニングされた分類器が十分な性能を示すかどうかは検討の余地がある。しかし、これら医療文書中のセンテンスにおける抽象度・事実

文 1) 入院中に測定した *IgE* と *RAST* では高値を認め、アレルギー体質が確認された。

文 2) 器質化している可能性が考えられた。

### 図 3 退院サマリに含まれるセンテンス例

次に、これら記載の抽象度・事実度の自動分類に向けたアプローチの構築を試みた。こうした分類器としては、まず、事前に定義されたルール、たとえば、「確認された」や「考えられた」といった事実関係の描写や推論の結果に用いられる述語等を整理しルールとして活用するアプローチが考えられる。また、抽象度や事実度に関するラベルが付与された文章(コーパス)から、出現する要素間の統計的な関係を自動的に学習したモデルを用いた分類器も想定される。その点、日本語における事実度分類に際して、前者にあたるルールベースの分類器が相応の分類性能を示した研究成果が示されていることから<sup>1</sup>、まずは当該研究の処理系の再現を目指す方針とした。なお、その後、後者のアプローチに基づいている開発された「日本語拡張モダリティ解析器 Zunda」が公開されていることが明らかとな

度分類は、相応の作業量を要するタスクとなる。そこで、理化学研究所革新知能統合研究センター知識獲得チームへと相談を行い、同チームにおいてより詳細に検討を進めて頂ける運びとなった。

以上と平行して、入院カルテの自動要約システムの開発を試みた。ただし、昨年度の研究によると、完全な退院サマリを自動生成するのは本研究班の実施年度中は困難であると考えられた。そこで、入院カルテを入力とする4種類の要約エンジンからの出力を元に、その効率的な編集操作を実現するユーザーインターフェースの設計とプロトタイピングを試みた。今年度の成果として開発したプロトタイプを図4に示す。また、システムのバックエンドをコンテナ環境にて構築し、検証を進めた。

## D. 考察

今年度、入院カルテの自動要約に向けた研究基盤の整備を進めた。

<sup>1</sup> 成田和弥, 「日本語事実性解析に関する研究」, 博士論文, 東北大学情報科学研究科 システム情報科学専攻, 2016.1.



その核は、ダミーカルテ、ダミーカルテ中のセンテンスを対象として目視で実施したアノテーション、そして、退院サマリの作成用ユーザーインターフェースのプロトタイプ、となる。そのなかでも、アノテーションにより付与したラベルの特性は、入院カルテの自動要約研究の成否に関わるが、先行研究の少ない分野であるため探索的な作業となった。そこで、今年度の作業を通じて明らかとなった課題に考察を加える。

まず、カルテにおける記載上の誤りをどう取り扱うかという問題が生じた。たとえば、「2011年6月1日頃より体調不良の様であった」という表現は、本来は、「体調不良があった」という事実として記載されるべきと考えられる。しかし、「様であった」との推定的な表現とすることで、言及の事実度が大きく損なわれている。同様に、時間的な制約のなかで記載されるカルテには、誤字や脱字、変換ミスが少なからず混入する。そうした誤りに対する処理が統計に

与える影響を排除するため、カルテにおける誤りを明示するためのラベルを導入するべきかと考えられた。

また、作業を通じて、退院サマリにはいくつかの特徴的なセンテンスが繰り返し含まれることが明らかとなった。まず、事実のバリエーションとして、「治療内容」が具体的に記述されているケースが認められた。また、病院での各種手続きなどに関する「計画」への言及が認められた。これは事実ではないが、確実度の高い未来の事象と考えられる。さらに、事実や推定との間に、診断結果や病状を踏まえ、それらに対する「評価」を行っているセンテンスの存在が認められた。さらに、推定のバリエーションとして、「可能性」について述べたセンテンスが認められた。これは、記載者が確信を有してはいないものの、何らかの推定に基づく有益な情報を補完するケースに当たる。

前の2つは事実の記載に近く、抽象度は低いセンテンスと考えら



図4 退院サマリ作成支援ツールのユーザーインターフェース

れる。また、後ろの2つはそれらの事実に立脚した推論・推測を含むセンテンスと考えられる。さらにそれぞれのセンテンスについて、元の入院カルテに由来するか否かを検討してみると、「治療内容」については入院カルテに含まれるに違いないが、「計画」については入院カルテには含まれない内容が含まれ得ることが分かる。同様に、「評価」は、入院カルテ中の記載に基づくパラフレーズと考えられるが、「可能性」については、入院カルテ中には含まれない事後的な推論の結果が含まれる。これらの知見は、本研究班が提示したCASEモデルに基づいた分類の妥当性を支持するものと考えられた(図5)。

## E. 結論

本研究分担では、今年度、退院サマリの自動生成に向けた研究基盤の構築を進めた。昨年度の研究において提示した入院カルテの自動要約モデル「CASEモデル」に基づいた退院サマリ生成を実現するためには、まず、このモデルの妥当性を実証する必要がある。そこで、国立病院機構の診療情報集積基盤(NCDA)を対象とした退院サマリの網羅的な統計解析を行うために、自動解析に求められる分類器の構築準備と精度管理のためのダミーカルテの整備を進めた。また、退院サマリ生成システムの設計を通じて、ユーザーインターフェースのプロトタイプングを行った。

来年度においては、これら研究基盤のさらなる整備を行うと共に、NCDAに集積される膨大な電子カルテデータの解析に進みたい。そのために、退院サマリに含まれるセンテンスの事実度・抽象度に基づいた分

		入院カルテ中の原文有無	
		有	無
事 実 度	高 (事実)	治療内容	計画
	低 (推定)	検査結果の 評価	可能性 (の示唆)

類器の確保と、精度管理に向けたダミーカルテの整備とアノテーション作業を進める。なお、これらの成果を論文化する際は、ダミーカルテそのものの代表性を担保する必要がある。そこで、今回整備したダミーカルテが真正なカルテと区別できない品質であることを検証する実験に取り組みたい。

もう一つ取り組むべき課題として、電子カルテに保存されている退院サマリの解析と統計取得をする場合、対象とする退院サマリの質を担保することが望ましいという点がある。そこで、他の研究分担において理想的な退院サマリの定義と退院サマリの品質に関する定量化を試みている。これにより良質な退院サマリのみを抽出した研究が可能となるが、もう一つのアプローチとして、入院カルテから理想的な退院サマリを生成したデータセットを持つアプローチが考えられる。そうしたサマリを生成していくことにはコストを要するが、要約エンジンのベンチマークとなりうる点でも貴重なデータとなる。そこで、ダミーカルテ中の退院サマリを医師に提示し、その結果を他の医師に提示するという操作を繰り返すことで、世代を経て「収束」するケースが存在するか否かについて検討を試みたい。

入院カルテの自動要約に向けた研究は、医療用自然言語処理研究において未開拓な領域であり、実際の要約アルゴリズムそのものの検討に先立って、まずは研究基盤の

構築と研究方法論の確立が求められる。今までの研究を通じて、研究アプローチの整備に加えて、研究に求められるカルテや精度管理用データの蓄積が進んだ。来年度においては、これらの基盤の統合を通じて、退院サマリの自動生成研究としての成果発表を進めていきたい。

## 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、産業技術総合研究所人工知能研究センター 高村大也先生、首都大学東京システムデザイン学部 小町守先生、奈良先端科学技術大学院大学 荒牧英治先生には、退院サマリの自動要約アプローチ、とりわけ、センテンス分類に際して貴重なアドバイスを頂戴致しました。一般社団法人情報通信医学研究所 中川晋一所長には、アノテーションガイドラインの開発にご助力を賜りました。また、退院サマリ生成システムの開発に際しては、北陸先端北陸先端科学技術大学院大学の浅井拓也氏に、また、検証に際しては、中部大学河原敏男先生を始め研究室の皆様にご協力を頂きました。この場をお借りして、研究協力者の皆様に感謝を申し上げます。

