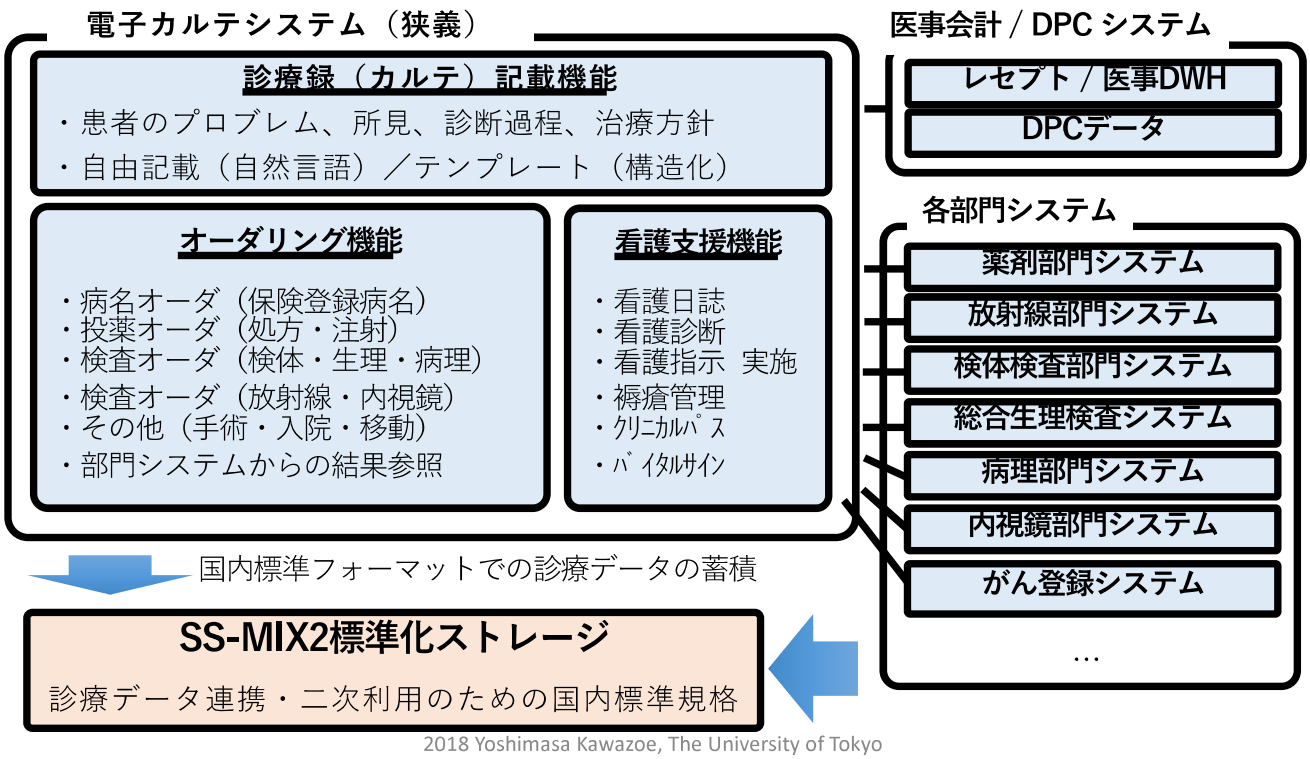
医療におけるビッグデータ
― 医療情報学はどう関わってゆくのか、何をなすべきか ―

ビッグデータの経緯と現状・課題③

電子カルテデータの活用とe-Phenotyping

東京大学医学部附属病院　企画情報運営部
JSTさきがけ
河添 悦昌

病院情報システム



2018 Yoshimasa Kawazoe, The University of Tokyo

診療データの種類・可用性

画像データ	レントゲン、CT、MRI、病理、眼底	構造化 ◎、標準化 ○
波形データ	心電図、脳波、光トポ、バイタル	
ゲノム・オミクス	FASTQ、SAM/BAM、VCF、パネル	
保険病名	標準病名、ICD-10、登録日、転機	構造化 ◎、標準化 ◎
処方情報	薬品種類、用法、用量、処方日	
検査結果	検体種類、検査項目、結果、検査日	
診療報酬請求	レセプト・特定健診、DPC	
診療記録	医師記録、看護記録、退院サマリ	構造化 △、標準化 X
検査レポート	放射線、病理、内視鏡、生理、ゲノム	

診療データを活用した研究

1. **診療報酬請求データを活用した研究**
   - レセプト・特定健診情報データ
   - DPC（Diagnosis Procedure Combination）データ

   => 保険請求目的で収集されたデータを活用
2. **症例レジストリデータを活用した研究**
   - 放射線治療DB（日本放射線腫瘍学会）
   - 慢性腎臓病DB（日本腎臓病学会）
   - 救急症例DB（日本救急医学会）
   - J-DREAMS（日本糖尿病学会）、J-DOME（日本医師会）
   - がんゲノム医療中核拠点

   => 項目をあらかじめ決め人手で集める
3. **電子カルテデータを活用した研究**
   - 特定の疾患を有する症例を抽出したい。既存薬の新たな効能を発見したい。
   - **e-Phenotypingの技術が必要**

1. 診療報酬請求データを活用した研究

- 主に傷病名、診療行為、管理料などの保険請求に付随する情報から、臨床的・医療経済的分析を行う研究。
- レセプトの第三者提供により行われた研究例[1]
  - 併用禁止医薬品、重複投与等の処方実態研究
  - 向精神薬の処方パターンの探索的分析
  - メトホルミン及びブホルミンの処方実態の分析

- **メリット：**
  - 極めて多量のレコードを利用可能
  - レセ：10億超件／年、健診：3000万件／年、DPC：1000万件／年
- **デメリット：**
  - **患者の状態に関する情報が限られる**（傷病名のみ）
  - 傷病名は真の罹患を反映していないのではないか
  - オーダ時レセプトチェッカーの普及

2018 Yoshimasa Kawazoe, The University of Tokyo

1. 第16回レセプト情報等の提供に関する有識者会議資料より

2. 症例レジストリデータを活用した研究

- 特定の疾患で特定の基準を満たす症例について、収集したいデータ項目をあらかじめ決め、人手により情報を入力。
- 電子カルテのテンプレート機能を使うことで、診療業務と研究用途の入力を一貫して行える。また、SS-MIX2と連携する臨床データ登録システムもある(http://mcdrs.jp/)

- **メリット：**
  - 人手で入力するため、臨床研究に直結する程度に質の高い、必要な情報粒度でのデータ収集が可能。
- **デメリット：**
  - 学会等の大きな組織による統制力が必要。また、あれもこれもで雪ダルマとなりがちな膨大な項目を正確に入力する負担が大きい。
  - 多施設で集める場合、複数ベンダーの電子カルテ上に、同じ仕様のテンプレートを作成する必要がある。

2018 Yoshimasa Kawazoe, The University of Tokyo

電子カルテ入力テンプレート（J-DREAMS）

3. 電子カルテデータを活用した研究

- 日常診療の記録を使い、関心のある臨床的特徴（表現型）を持つ症例を抽出し研究対象としたい。
- ゲノム研究：遺伝子の変異に対応する臨床的特徴もつ症例
- 臨床研究：適格・除外基準を満たす候補症例
- 薬剤疫学研究：既存薬の新たな効能を発見

- **メリット：**
  - 診療記録・投薬情報・検査結果を組み合わせることで、レセプトより詳細な患者の状態が抽出できるのではないか。
  - 症例レジストリのようにインテンシブな入力が不要。
- **デメリット：**
  - 日常診療で発生する”Exhaust Data”から、目的とする情報を抽出する必要がある。

e-Phenotyping：カルテからの表現型抽出

- どの程度の粒度での表現型（疾患）の抽出を目指せそうか

構造化データ

- 登録病名：保険請求を目的とするため真の病名を反映しない
- 検査結果：検査値だけで診断できない疾患も多い。疎な時系列データ。
- 処方注射：同一薬が複数疾患に使われることや、投薬治療されない疾患も多い。

これらを組み合わせることで、**本態性高血圧、２型糖尿病、気管支喘息**など**通常の疾患粒度**で罹患の有無を推定することはできそう。

自然言語文書

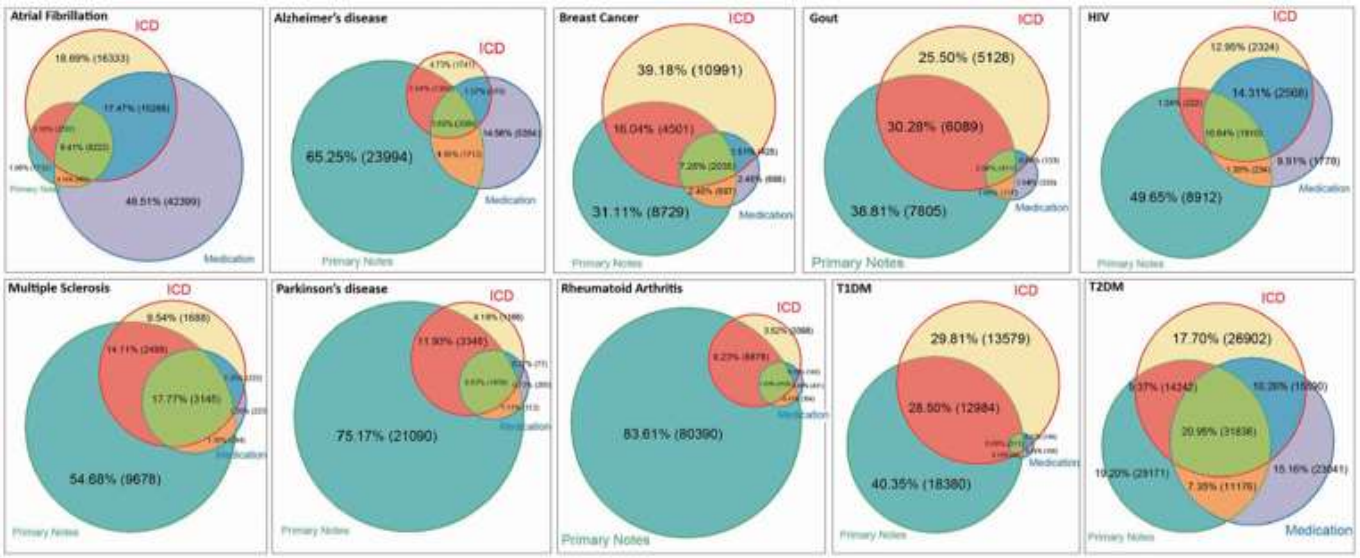
- 診療記録：記録が断片的、表記ゆれ、略語の多用、時期不明、事実不明
- 退院サマリ：外来患者には記録されない。

更に、**自然言語文章の解析を加える**ことで、精度の向上のほか、稀な表現型の抽出に役立つかもしれない。

ただし、いずれにおいても、罹患時期を同定するタスクを加えると、更に困難。

どの情報カテゴリを使うと疾患有無が判断できるか

病名コード（ICD-9）、処方情報、診療記録のどのカテゴリに、疾患有無を判断できる情報があるか



対象疾患の選び方にもよるだろうが、診療記録の役割が大きい。

Wei-Qi Wei, et.al. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc. 2016 Apr;23(e1):e20-7

ご参考

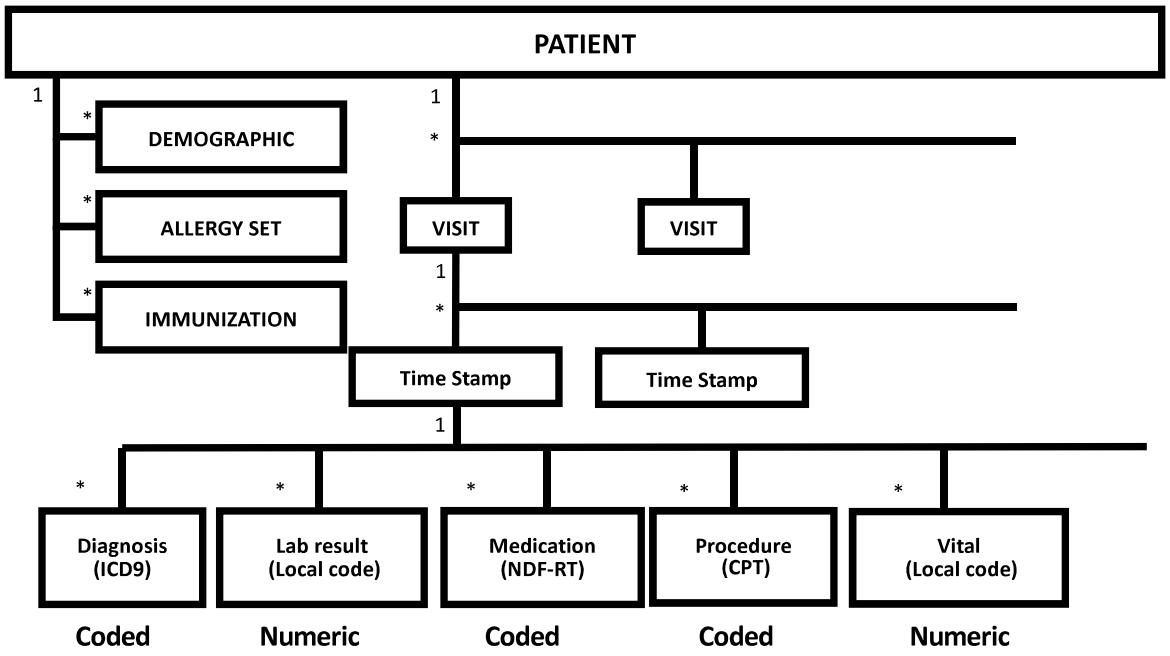
i2b2: Informatics for Integrating Biology & the Bedside

https://www.i2b2.org/

2018 Yoshimasa Kawazoe, The University of Tokyo

ご参考

Structure of I2B2 Data



2018 Yoshimasa Kawazoe, The University of Tokyo

Phenotypingアルゴリズムの開発例

- 本邦でのアルゴリズム開発の必要性
  - 疾患診断基準の違い、自然言語文の解析精度の違い、診療情報の種類の違いなどから、海外で開発されたアルゴリズムを利用する場合に、報告される精度を外挿することはできない。

2型糖尿病 [1]	陽性的中率	感度	正解率
機械学習	89.77%	67.86%	**95.77%**
ルール形式	93.94%	46.97%	87.54%

本態性高血圧 [2]	陽性的中率	感度	正解率
機械学習	79.64%	66.86%	**89.80%**
ルール形式	75.50%	75.50%	71.56%

1. 香川 瑞奈 他. SS-MIX2標準化ストレージを用いた2型糖尿病症例のEHR Phenotyping手法の開発と評価. 医療情報学35(Suppl.), pp.672-675, 2015.
2. 香川 瑞奈 他. 高血圧のphenotyping手法の開発および他疾患との比較検討. 医療情報学36(Suppl.), pp.770-773, 2016.

2018 Yoshimasa Kawazoe, The University of Tokyo

アルゴリズム開発における問題（1）

正解ラベルを付けるのに労力を要する

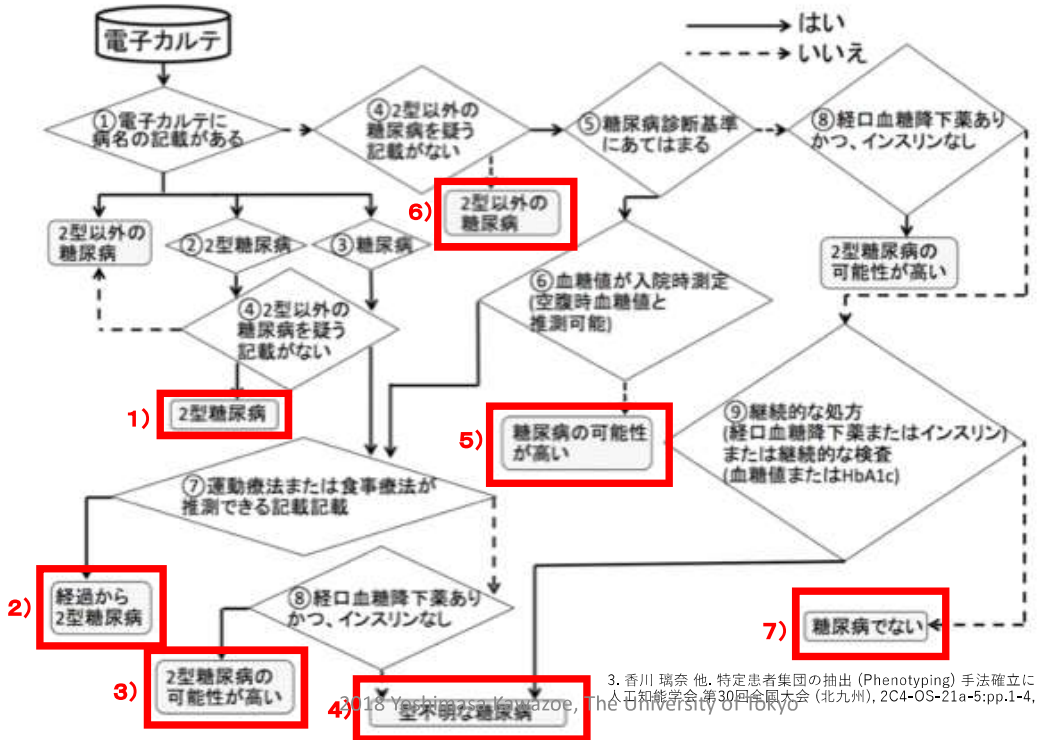
- 陽性的中率（PPV）だけを評価するだけであれば..
  - 陽性と判定した症例のカルテを開き結果が正しいかどうか判定
- PPVだけでは十分でない理由
  - 確実にケース例といえる症例を1件だけ見つけるアルゴリズム
  - 有病率の異なる施設にアルゴリズムを適用する場合は、感度と特異度が指標としてより適切
- 正解ラベルを付きデータセットが必要
  - 感度、特異度の算出に必要
  - 開発例では650人の正解ラベル付きセットを作成した
  - データセット作成に時間を要することがアルゴリズム開発の律速

2018 Yoshimasa Kawazoe, The University of Tokyo

アルゴリズム開発における問題（2）

正解ラベルの明確な基準が必要

人手による2型糖尿病の正解ラベル付けのフローチャートを作成[*3]



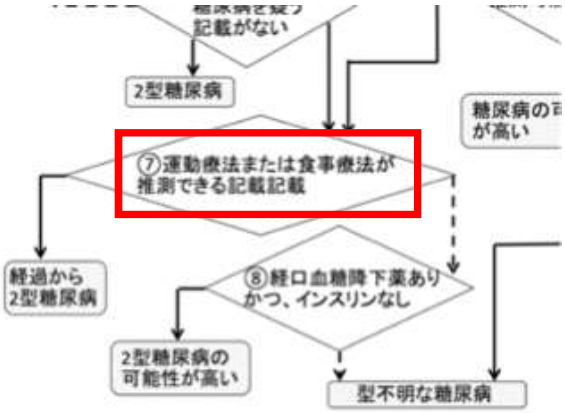
3. 香川 瑛奈 他. 特定患者集団の抽出 (Phenotyping) 手法確立に向けた技術的課題に関する考察. 人工知能学会 第30回全国大会 (北九州), 2C4-OS-21a-5:pp.1-4, 2016.

2018 Yoshimasa Kawazoe, The University of Tokyo

計算機処理が困難である理由（１）

- 前述のフローチャートを計算機処理するとして、その難しさを検討

- **⑦「運動療法または食事療法が推測できる記載」**
  - 診療録に栄養指導を行ったことの記載はあるものの、2型糖尿病に対する指導であることが明示されないものがある。
  - 医療者であれば、栄養指導の対象となる疾患はいくつかに限定されることと、各疾患に対して行われるべき指導内容が想定されることから、2型糖尿病を対象としたものかどうかの判断は比較的容易。
  - そのような**知識がない計算機**にとって、記載される内容だけで判断することは難しい。

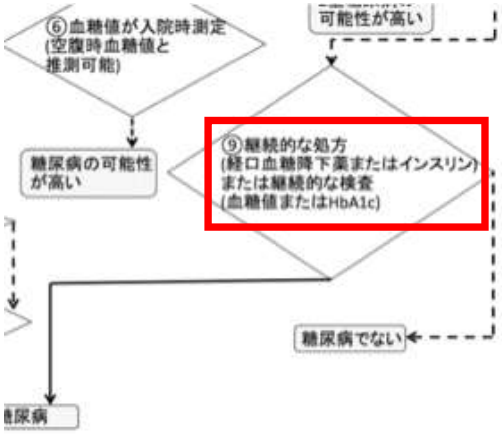


医療知識が必要。データドリブンのみでは難しい。

2018 Yoshimasa Kawazoe, The University of Tokyo

計算機処理が困難である理由（２）

- **⑨「経口血糖降下薬またはインスリンの継続的な処方」**
  - 一時的な血糖上昇に対して血糖を下げる薬剤が使われることがあるため、そのような症例を除外する目的。
  - 医療者であれば一時的かどうかは、血糖上昇の前にどのような医療イベントが生じていたのかをみることで比較的明らかに判断できる。
  - 計算機でこれを行うためには、血糖の上昇を伴う医療イベントの種類が何であるかということの**知識が必要**。
  - また、一時的という表現を単純な期間に置き換えられないことから、計算機処理で判断することが難しい。



医療知識に加えて、一時的・長期的といった期間に関する解釈が必要

e-Phenotypingの現状

- e-Phenotypingアルゴリズムの必要性
  - 現状の電子カルテの記録方式では、何の病気で診療を受けているか、ということすら計算機判断が難しいため。
  - 診療録の記載に対する単純な文字列のマッチングやレセプト病名だけは、対象症例が十分に抽出できない。
  - 複数カテゴリの診療情報が必要。特に診療記録の役割が大きい。
- e-Phenotypingアルゴリズム開発における問題
  - アルゴリズムの性能を知るために正解データセット作成、判定基準を明確にする必要があるなど、開発に時間を要する。
- 計算機処理がむずかしい理由
  - 自然言語文章からの疾患有無の事実判定が難しい。
  - カルテは人間が読むこと（文脈を読む、知識で行間を補うこと）を前提とするため。

これからの展望 ..とにかくデータの質が重要

短期的には

- 信頼性の高い情報源を使う（DPC、がん登録）[1]
  - DPC自体が解析を意識した制度であり「資源病名」は信頼性が高い。
  - がん登録に含まれるICD-Oコードは、訓練を受けたスタッフよりコーディングされており、信頼性が高い。

中・長期的には

- 診療記録を活用するための整備
  - 医療分野の自然言語処理技術の向上のためにできること[2]
  - 退院時サマリなど、確実に記録される文章のミニマル標準化[3]
  - 日常診療で医療者がPhenotype情報をスムーズに正しく入力されるテンプレート機能の開発。医療者へのフィードバックを通し診療記録の重要性を認識してもらう。

1. 中島直樹. 日本におけるPhenotypingの必要性と可能性. 医療情報学連合大会プログラム2017・抄録集 巻：37th ページ：202.
2. 森田　瑞樹, 荒牧　英治. 医療分野における言語処理研究の環境整備に向けての提案. テキストアノテーションワークショップ 2012.
3. 木村通男. 退院時サマリの電子化と標準化. http://www.hl7.jp/docs/65seminar_1_HL7.pdf