リアルワールドデータの多施設統合解析を実現する

JPP : Japan Precision-medicine Platform

新医療リアルワールドデータ研究機構株式会社 岡田 昌史

第9回クリニカルバイオバンク学会シンポジウム

2024.8.2

リアルワールドデータ

リアルワールドデータとは何か？

- **臨床試験の**世界と**現実の世界**
- 臨床試験の世界
  - 対象集団は、プロトコルで定義された包含/除外基準により選択される。
  - 受診はプロトコルによって決定される
  - 薬剤投与はプロトコルによって管理される
  - アウトカムは、プロトコルで定められた方法で測定される。
- 実世界
  - 母集団は一般集団、または病院の所在地や専門分野による
  - 患者は症状があるときや薬が切れたときだけ病院を訪れるようになる
  - それぞれの患者の問題を解決するために薬が処方される。
  - 結果は測定されない

典型的なリアルワールドデータ

- 病院の電子カルテ（EMR）データベース
- レセプトデータベース（日本ではDPC(Diagnosis Procedure Combination)を含む）
- (日本の）健康診断データ
- 学会等主導によるレジストリデータ

- 活用時には一般には匿名化されているが、各病院や各保険会社内での受診は縦断的に追跡調査可能となっている場合が多い
- 後ろ向きコホート研究、もしくはコホート内ケースコントロールデザインで解析される場合が多い
- 電子化されているため、臨床試験と比べてデータサイズが大きく、また多くのエラー、欠損、表記ゆれ等がそのまま残されているため、扱うには多くのテクニックが必要となる。

電子カルテ(EMR)データの取得

レセプトの限界

- レセプトデータの病名は、実際の問題を表すだけではなく、処方と矛盾しないように病名がつけられていることがある。重篤度等の情報はなく、すでに解決した病名がそのまま記録され続ける場合も多くみられる。

- EMRの場合、病理/放射線レポート、臨床病期、詳細診断、遺伝子変異情報等はフリーテキストで書かれている。フリーテキストはその中に個人情報が記載されている危険性があるため、研究用にそのまま提供されることはほぼない。

- 特に、悪性腫瘍を対象としたデータベース研究を行う場合には、病理診断情報や病期、治療方針（化学療法レジメン等）といったテキストで記載された情報が必要になることが多い。

- とくに病理診断レポートには、パネル検査やシーケンサーによって得られる遺伝子変異情報のなかから、治療方針決定に寄与する情報が病理医によって選択されて書かれており、重要な情報となる

- また、放射線レポートには放射線診断医による治療効果判定が書かれており、アウトカム情報として重要である

- これらのテキストに含まれる情報が**構造化**された形でデータベースに入力されれば、研究に使うことができる。

EMRから構造化データを取得する

- 戦略1: 自然言語処理技術を使って、フリーテキストから知識を得る
  - 医療スタッフに余計な仕事をさせない
  - 精度はおおむね良好であり、LLMの発達によりさらなる精度向上が期待されるが、100%ではない
    - そもそも人間が読んでもわからない / 必要な情報が記載されていない場合もあり
- 戦略2: 医療スタッフが構造化された情報を入力するために、EDCや「テンプレート」を使用する
  - データ入力は医療スタッフに余分な仕事を強いる
  - 暗黙の意図と解釈を捉えることができる
    - 例えば、薬剤や処置の組み合わせの意図（治療方針）、処方が中断された理由、ある症候が「ない」という観測事実や、陰性であった検査結果などである。このような情報は、診断に寄与しているにもかかわらず、EMRのテキストでも欠落していることがある。

戦略1: 自然言語処理技術を使って、フリーテキストから知識を得る

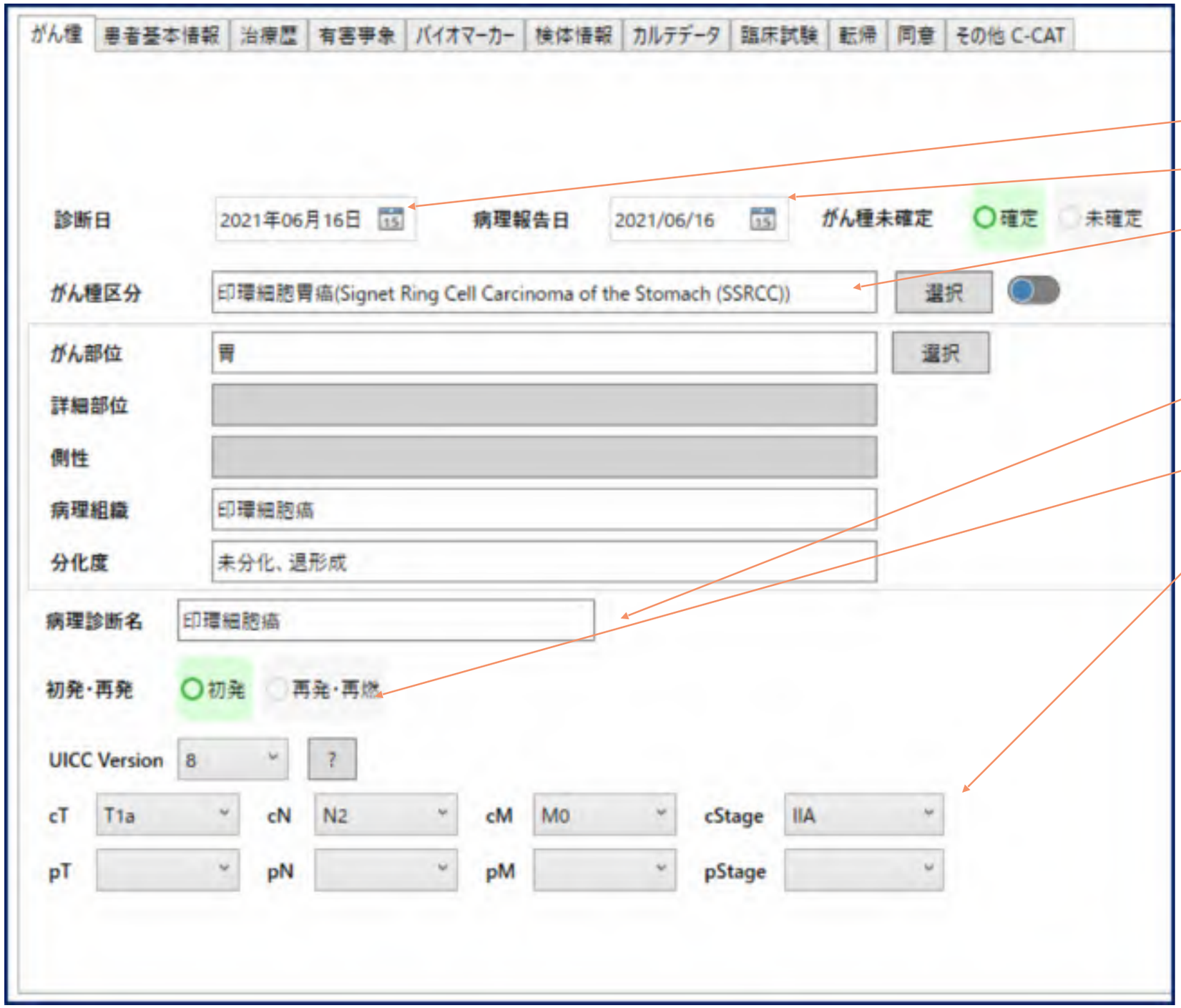
- たとえば

  Ge, J., Li, M., Delk, M. B. & Lai, J. C. A Comparison of a Large Language Model vs Manual Chart Review for the Extraction of Data Elements From the Electronic Health Record. *Gastroenterology* **166,** 707-709.e3 (2024).

では、

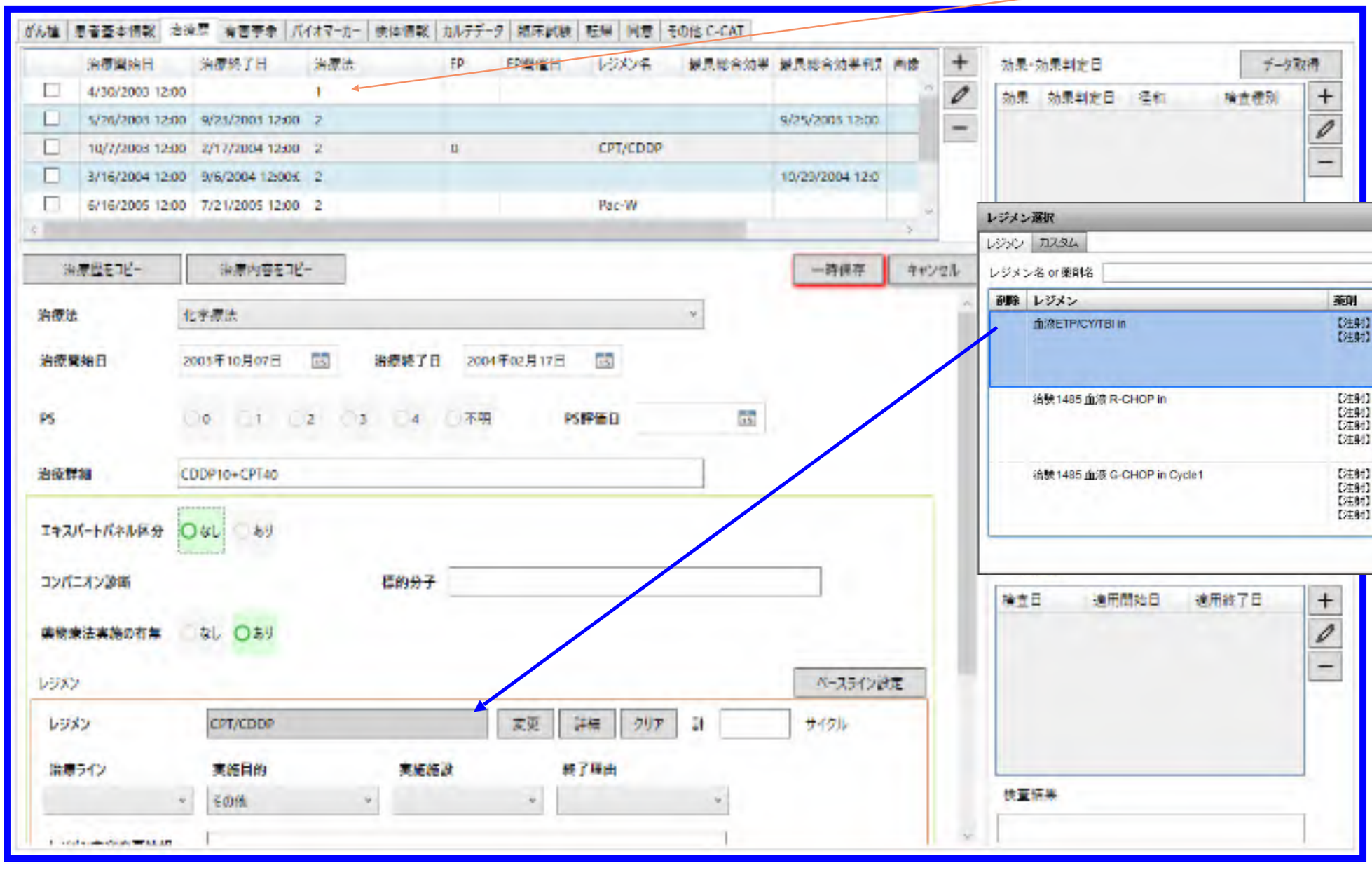
- GPT-4相当のLLMに肝臓画像診断レポートを読み込ませて、Overall accuracy 0.934 という結果を出している
- PrecisionとRecallにはそれほどスコアに差がないので、False positive とFalse negativeのどちらが出やすい、ということでもないようではあり、明確なハルシネーションの影響は出ていないように読める
- とはいえ100%にはいたらない

戦略2: 医療スタッフが構造化された情報を入力するために「テンプレート」を使用する



Cyber Oncology®のスクリーンショット

化学療法の詳細
(治療ライン
レジメン、
対応評価、
有害事象)

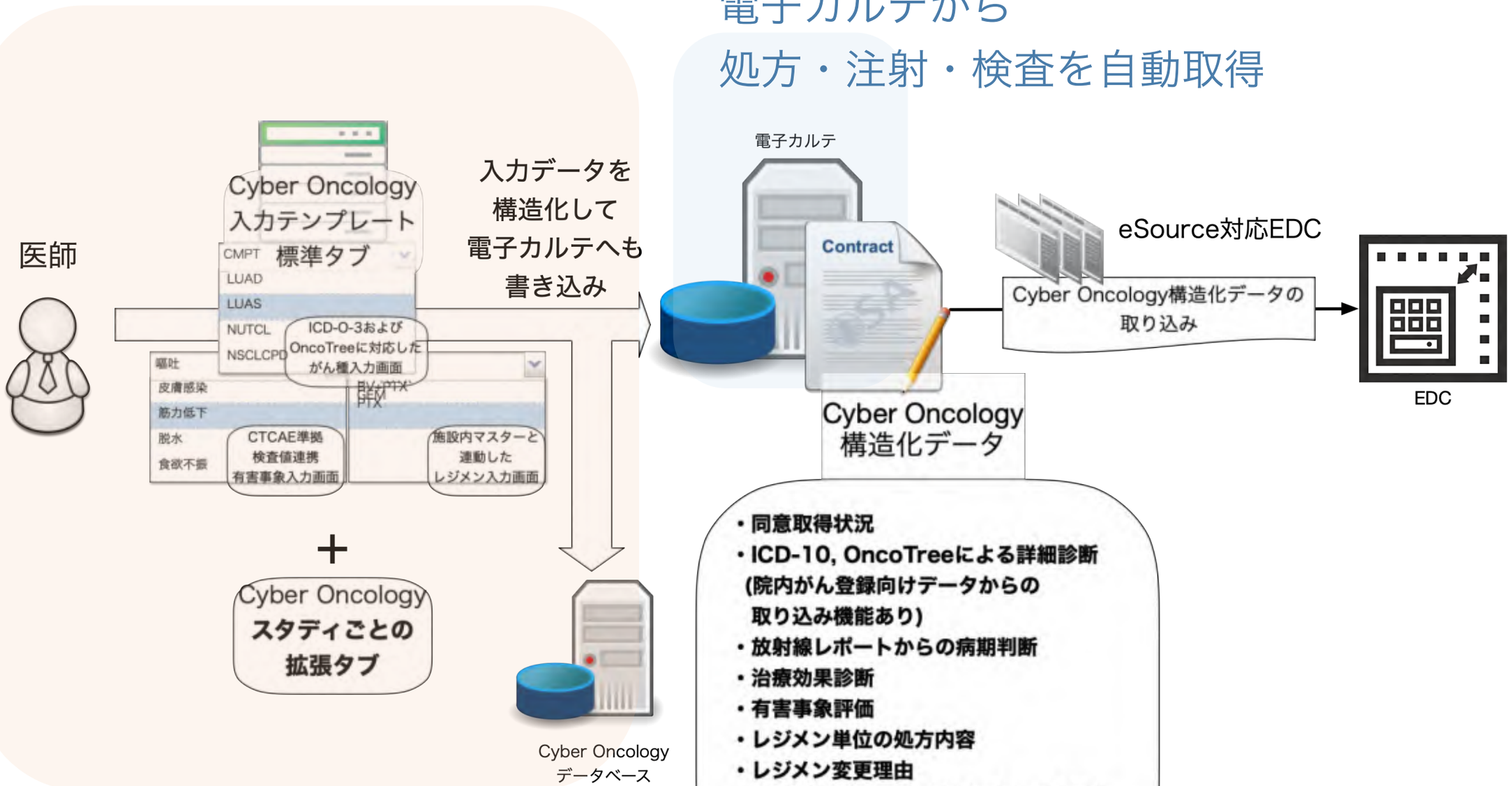


レジメンはEMRのマスターから選択

Cyber Oncology®のスクリーンショット

Cyber Oncology® によるEMRテンプレート

電子カルテから
処方・注射・検査を自動取得



- 同意取得状況
- ICD-10, OncoTreeによる詳細診断
  (院内がん登録向けデータからの
  取り込み機能あり)
- 放射線レポートからの病期判断
- 治療効果診断
- 有害事象評価
- レジメン単位の処方内容
- レジメン変更理由
- バイオマーカー(免疫染色, シーケンサ)
- 臨床検査結果(カルテから自動取り込み)
- 処方/記録(カルテから自動取り込み)

医師がテンプレートを使用して、
構造化してカルテを入力

EMRテンプレートの標準化

- EMRテンプレートを導入するには、EMRごとにカスタマイズが必要である。
- HL7 FHIR（Fast Healthcare Interoperability Resources）標準の一つである "FHIR Questionnaire "はテンプレートの標準的な記述を定義している。
- FHIR Questionnaire の日本語化は、JASPEHRプロジェクト（Japan Standard Platform for Electronic Health Records）として進行中である。
- JASPEHRに対応した電子カルテを使用することで、研究者はEMRベンダーに関係なく、多くの病院にEMRテンプレートを配布することができる。
- 弊社ではCyber Oncology®の機能を移植した JASPEHR対応テンプレートシステムである CyberJASPEHR (仮称) もリリース予定

電子カルテデータの活用

電子カルテデータの多施設データ統合分析

- リアルワールドデータを用いた臨床研究
  - ランダム化していないため、**交絡因子**の影響を受ける。PS Matching 等で調整をするためには多数のサンプルが必要
  - プロトコル通りに収集、測定されたデータではないため、**測定バイアス**の影響を受けやすい
  - 相対的に低コストでデータ収集が可能であるため、**サンプルサイズ**を拡大しやすい
  - サンプルサイズを拡大した結果として、**一般化可能性**が高い結果を得られる

- サンプルサイズを拡大するには？
  - 単一施設で長期間リクルートする　→　数は増えるが一般化可能性はそれほど上がらない
  - **多施設で収集する → 一般化可能性が上がる**ことが期待できるが、**測定バイアスの影響**を受けやすく、**コスト**もかかる

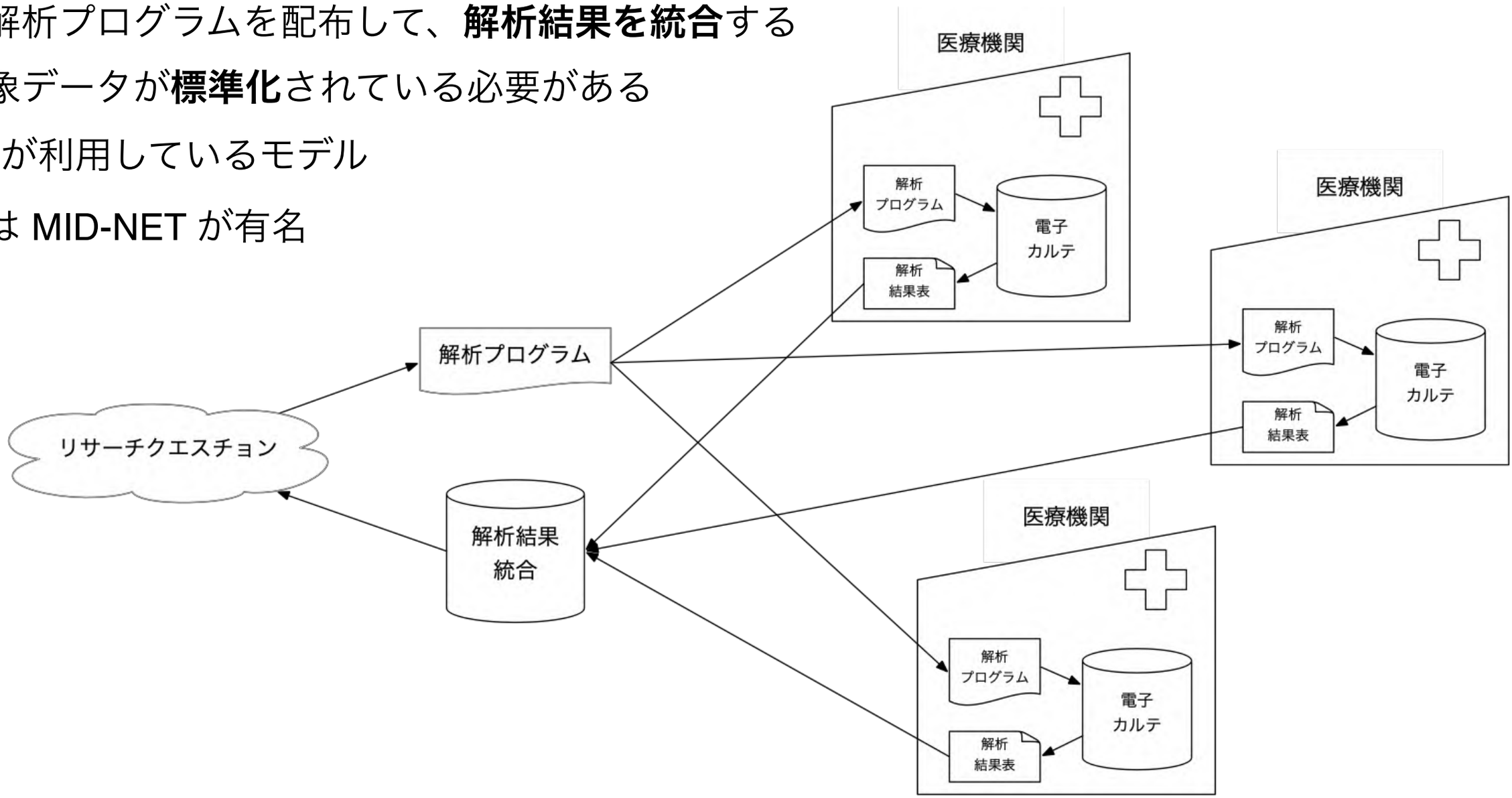
- 測定バイアスを少なく、コストを抑えて多施設データの統合分析ができないか？

多施設データ統合分析の問題点

- 個人情報保護
    - 診療目的を超えて個人情報を収集するには、**同意取得、倫理審査**が必要になる
    - 中央倫理審査も可能となっているが、施設別の審査もまだ必要とされる場合がある
    - 同意取得やデータ収集業務に関する**契約締結**も案件別・施設別に実施すると大変な量となる
- 測定バイアスの軽減
    - Inclusion Criteria / Exclusion Criteriaを**厳密に定義**する、測定**項目の定義**を厳密にすることで測定バイアスを避けたいところ
    - そのためには、収集項目(診断・検査・処置・処方)の**定義の標準化**が必要
- 分析のコスト
    - 各施設から提供されるデータのフォーマット・変数名・欠測の表現法などが少しでも異なると、そのたびに**解析プログラムの修正**が必要
    - 実施検査の変更、**電子カルテの更改**などがあると作り直し
    - 作り直す必要があることに長いこと気付かない場合も

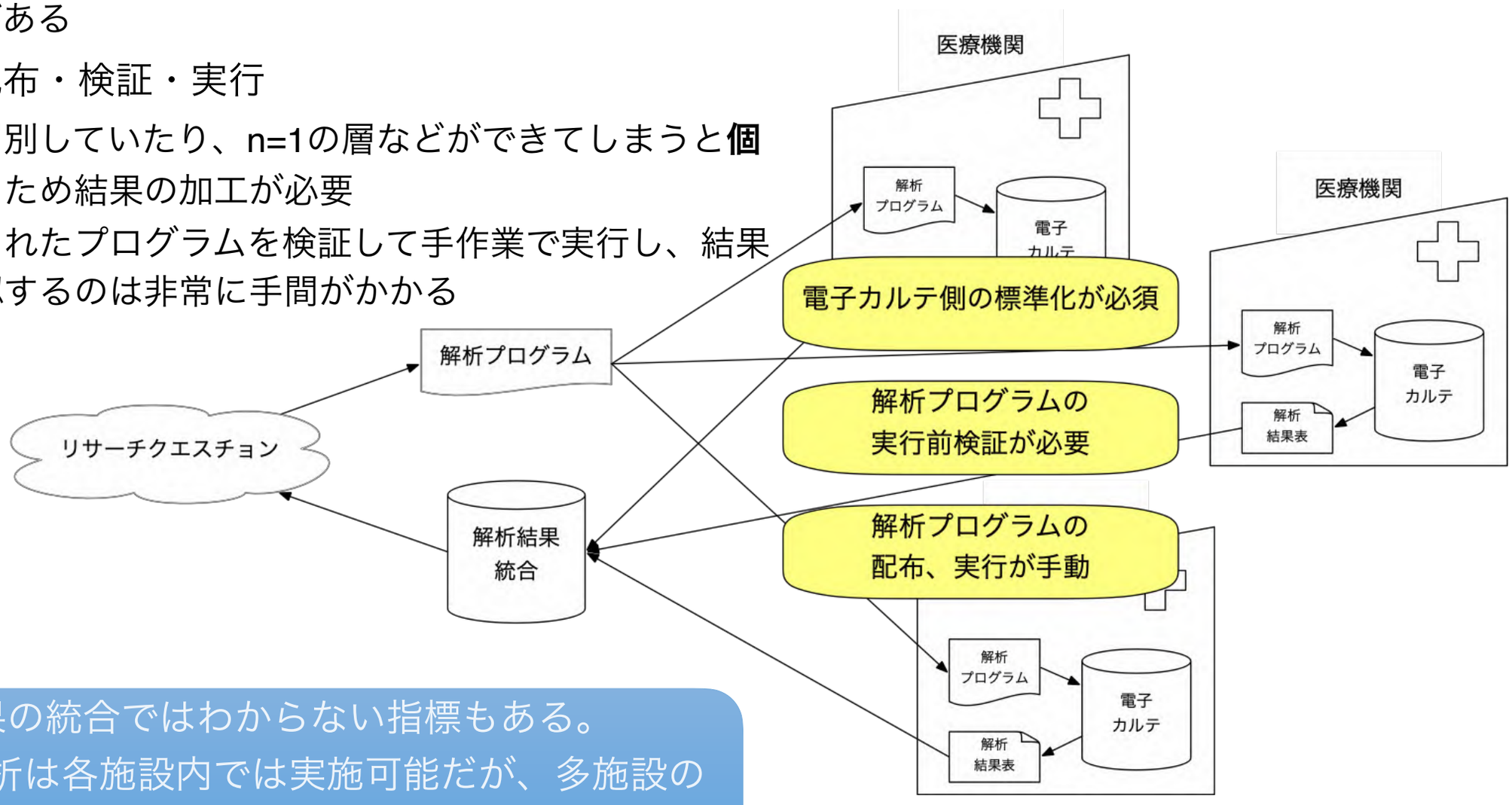
Federation Model

- 個人情報を**医療機関の外に出さず**、施設内で**解析済みの表**になってから出す
- 同一の解析プログラムを配布して、**解析結果を統合**する
- 集計対象データが**標準化**されている必要がある
- OHDSI が利用しているモデル
- 日本では MID-NET が有名



Federation Modelの問題点

- 各施設内データの標準化が必須
  - OMOP CDMのように、**変数名や用語辞書のレベルまで標準化**されている必要がある
- 解析プログラムの配布・検証・実行
  - 患者IDごとに層別していたり、n=1の層などができてしまうと**個人情報にあたる**ため結果の加工が必要
  - しかし、配布されたプログラムを検証して手作業で実行し、結果を人の手で確認するのは非常に手間がかかる



各施設での解析結果の統合ではわからない指標もある。
たとえば生存時間分析は各施設内では実施可能だが、多施設の情報を統合したものは難しいが、層別解析、患者背景の集計、治療パターンの分析、医療費の分析などは可能。

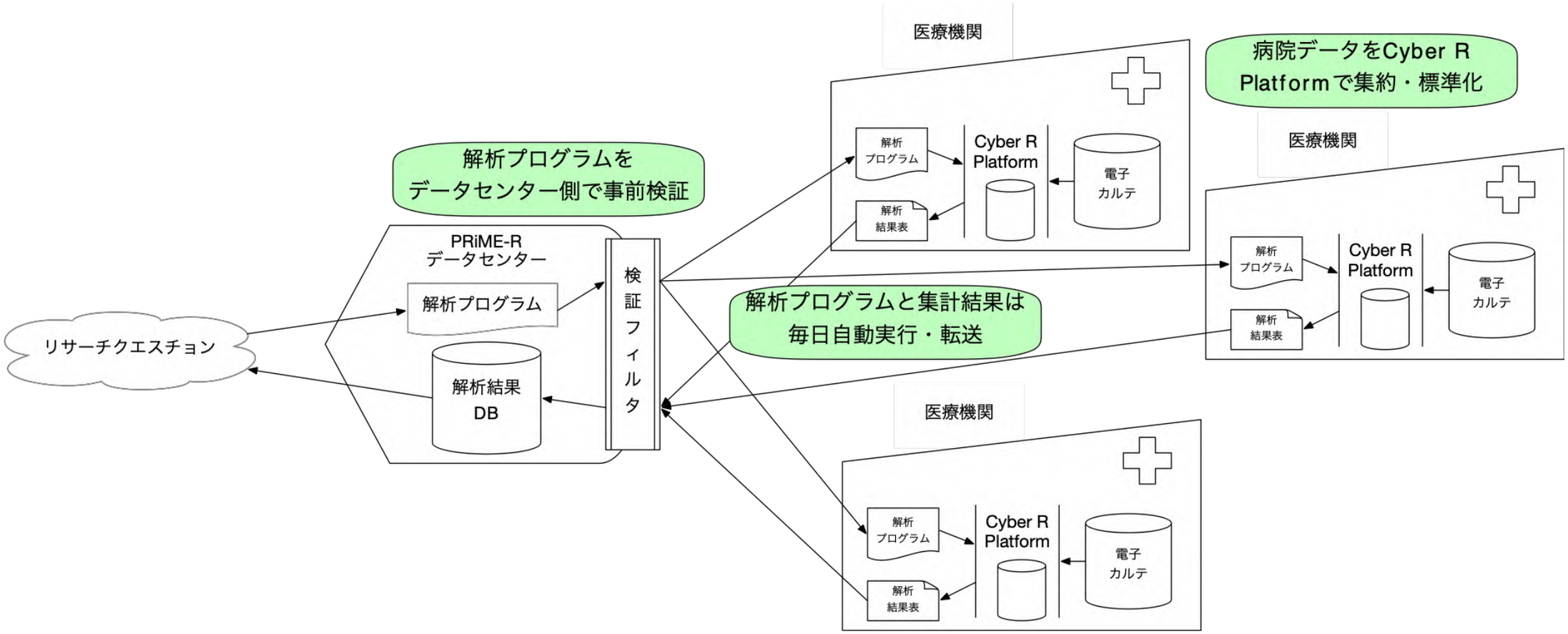
Cyber Oncology ® 解析プラットフォーム

- 院内に設置した**Cyber Oncology**が、**テンプレート入力データ**や電子カルテから取得した**検査・処方・注射情報**、がん登録向けの**病理診断・初回治療内容** などを含む各種病院内データを標準化して解析向けに格納

- 解析プログラムはデータセンターで検証の上、問題ないことが確認されたものが**オンラインで配信**され、毎日**自動実行**

- 解析結果は**k-匿名化**により、個人情報にあたる可能性があるデータを除いて、統計情報として検索を可能にする

- 後ろ向き観察研究の適格基準該当症例がどの病院に何症例あるのかを素早く知ることができる

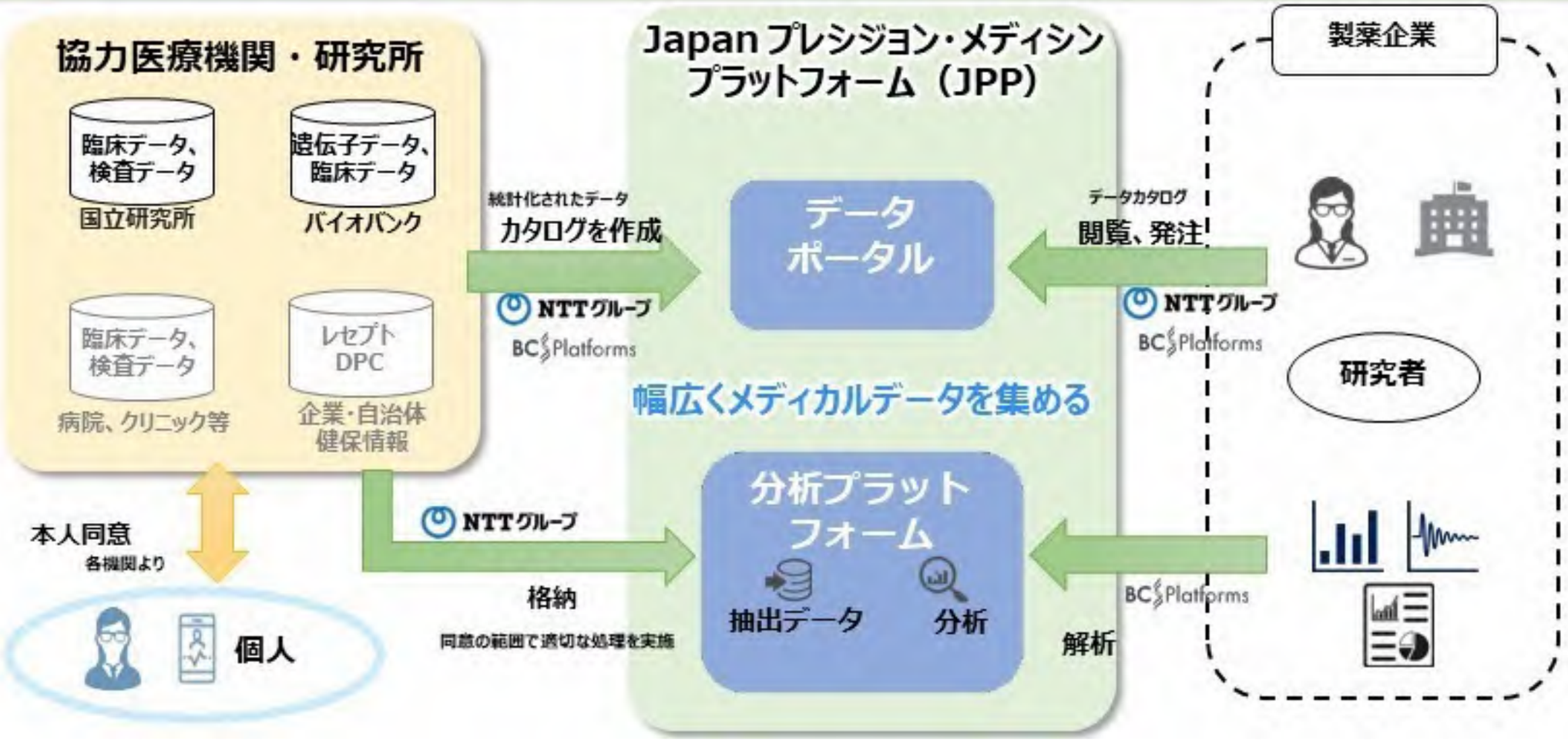


Japan Precision Medicine Platform(JPP)

PRiME-R

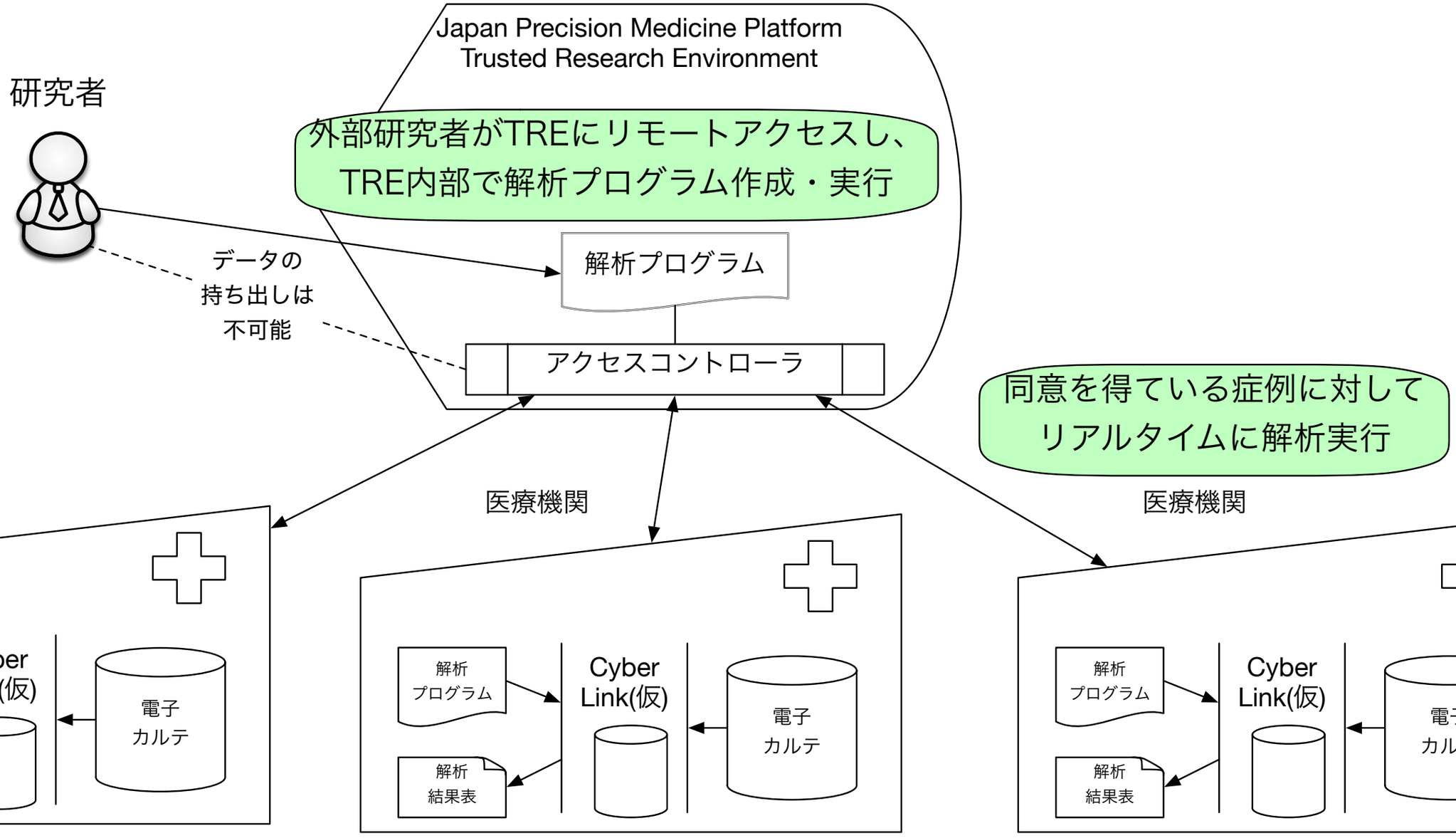
Japan プレシジョン・メディシン プラットフォーム - 将来への安心 -

めざす姿：日本の研究者が国内の次世代医療のための研究ができる仕組み
NTTグループがBC Platforms AGのシステム/ノウハウを活用し、日本の医療機関のメディカルデータを整備



個別症例分析のためのTrusted Research Environment

- Federation Modelで適格症例がどの医療機関に何例ほど受診しているかがわかったら、倫理審査を通過したプロトコルによる臨床研究に移行する
- 同意取得された症例の個別症例データを 1 箇所に集約する環境 (Trusted Research Environment) を準備し、研究者がこの環境に外部からアクセスして、解析結果のみを持ち出すことが可能
- 研究者は個別データ自体を持ち出すことがないため、データがコピーされ管理不可能になるリスクを回避



まとめ

- リアルワールドデータは臨床試験とは異なる集団で、実際の患者集団における治療実態や治療効果を知ることができるため、近年活用が進んでいる。
- これまでは医科レセプト, DPCレセプトを中心としたデータが主流であったが、病理レポートに記載されている遺伝子変異情報や放射線診断レポートにある治療効果判定情報を加えることができれば、検証できる研究テーマが大幅に広がる
- 自由記載文書に記載されている情報を構造化するためには自然言語処理技術が有用だが、完全ではないため、電子カルテの入力テンプレートによる手動入力もまだ必要である。Cyber Oncology® では、カルテに記載される情報を**記載と同時に取り込む**ことで二重入力の負担を軽減している。
- リアルワールドデータを用いた臨床研究では、**多施設でデータ収集を行いサンプルサイズを増やす**ことで**一般化可能性が高い**結果を得られることが期待されるが、施設間での違いによる**測定バイアス**や**コスト**が問題となる
- OHDSIやMID-NETで使われている**Federation Model**により**個人情報の取得を避け**つつ多施設データの**統合解析**が可能となる
- Japan Precision Medicine Platform では、Cyber Oncology® と統合されたBC Platforms のソリューションを用いて、Federation Model により後ろ向き観察研究の適格基準該当症例がどの病院に何症例あるのかを素早く知る機能を提供する。さらに、個別症例データを用いた臨床研究の実行において病院データと連動した安全な解析環境を提供することで、電子カルテデータを用いたリアルワールドデータ研究の迅速な実行を目指していく