



OHDSI Japan

OHDSI Japan Face-to Face Conference Sep.29.2025



Discussion presentation

# CyberOncologyのOMOP-CDM変換による 食道癌データセット構築

RWD R&D

京都大学 大学院 医学研究科  
リアルワールドデータ研究開発講座

DEPARTMENT OF REAL WORLD DATA R&D,  
GRADUATE SCHOOL OF MEDICINE, KYOTO UNIVERSITY

松本 繁巳



# RWD extraction from Cancer Patient Journey



病理（免疫染色）  
腫瘍マーカー  
CT等画像データ

遺伝子パネル検査  
バイオマーカー

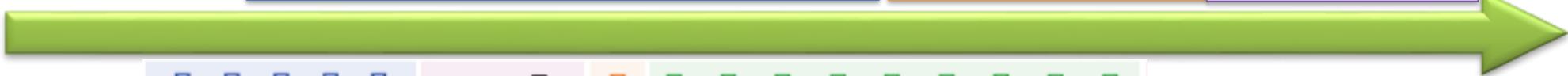
アウトカム  
時系列データ  
有効性と安全性

Omics data  
が重要らしい？



CTCAEによる副作用評価：悪心・しびれ・下痢・・・

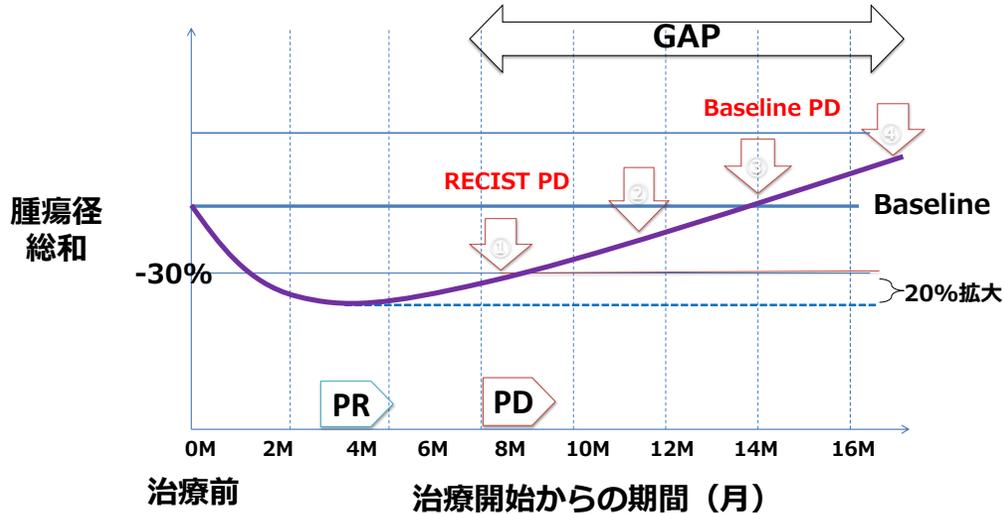
Onset Dx. 1<sup>st</sup> line 2<sup>nd</sup> line 3<sup>rd</sup> line



RECIST PDと実臨床での治療変更タイミング

Gastric Cancer  
tub2, Stage IVb  
  
Claudin IHC (-)  
Her2 IHC (3+)  
PD-1 < 1  
MSS  
  
Liver meta  
(10%)

RECISTによる効果判定





# Agenda

1. **CyberOncologyとAIMGAIN**
2. **OMOP-CDM conversion**
3. **Demonstration**

# RWDに求められる有効性と安全性データ収集・生成→可視化

NTT PRIME-R 副作用入カツール

PRIME-R

有効性

CyberOncology

主観データの  
入力バイアスを  
制御

診療データ入力

CyberOncology®

コピー&ペースト

電子カルテ

連携 連携

オープンDB

同期

電子カルテDB

臨床研究データ

追加入力

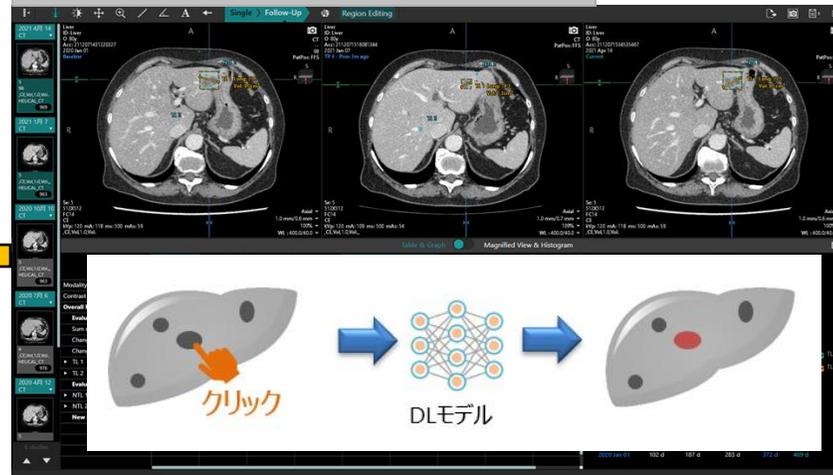
遺伝子パネル検査  
検査会社共通ポータル

H.U.GROUP

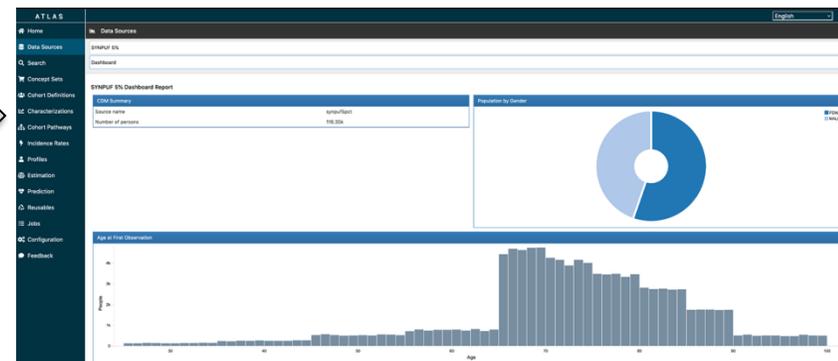
FUJITSU 富士通 Japan

Region Tracking

Canon  
キヤノンメディカルシステムズ株式会社



BIツールで可視化・解析



AMED 革新的医療技術研究開発推進事業（産学官共同型）スタートアップタイプ  
R6年度AIMGAIN採択課題

がん診療の質の向上と研究開発に資するリアルワールドデータプラットフォーム開発

研究代表者 武藤 学



KU:IP 京都大学医学部附属病院



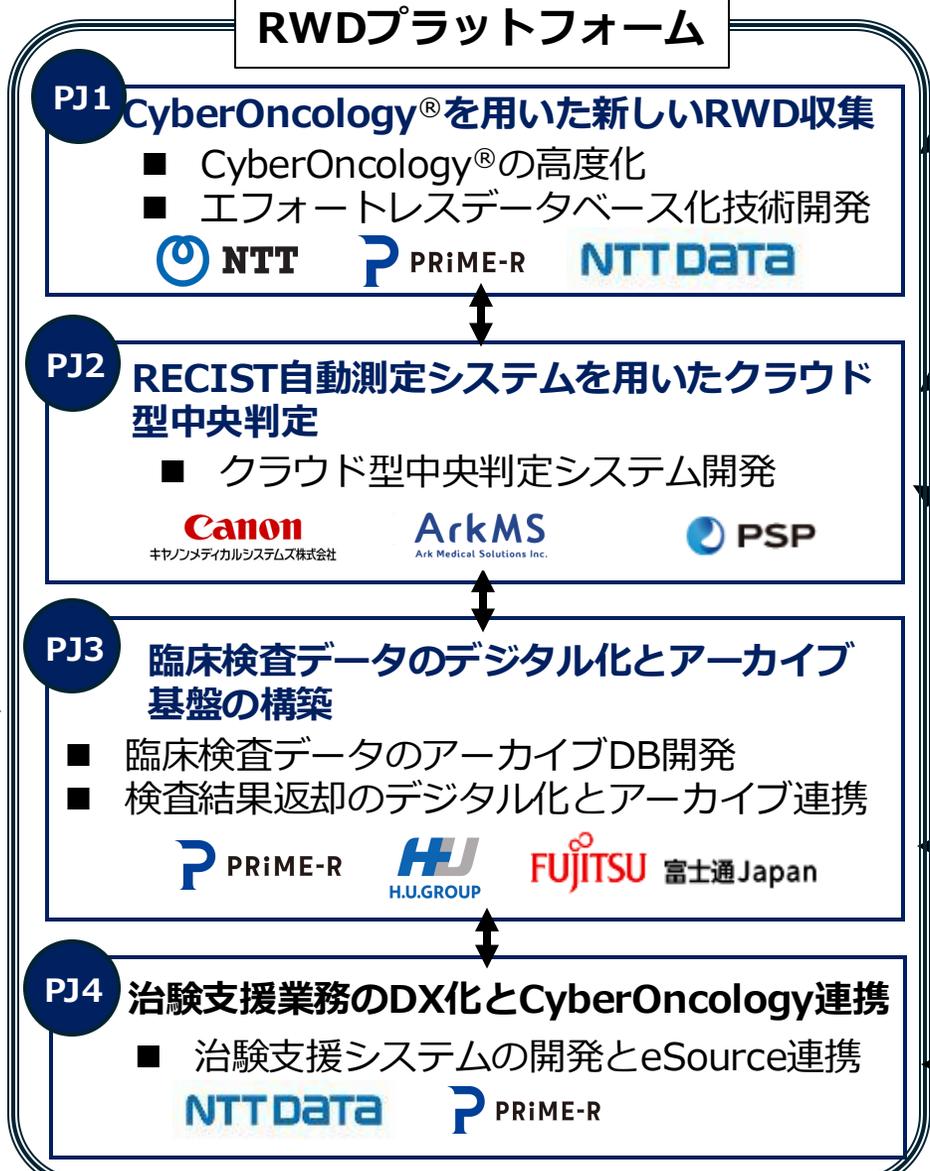
J-CONNECT



# AIMGAIN事業

## 非競争領域（産学官の協調によるRWD収集基盤構築）

## 競争領域



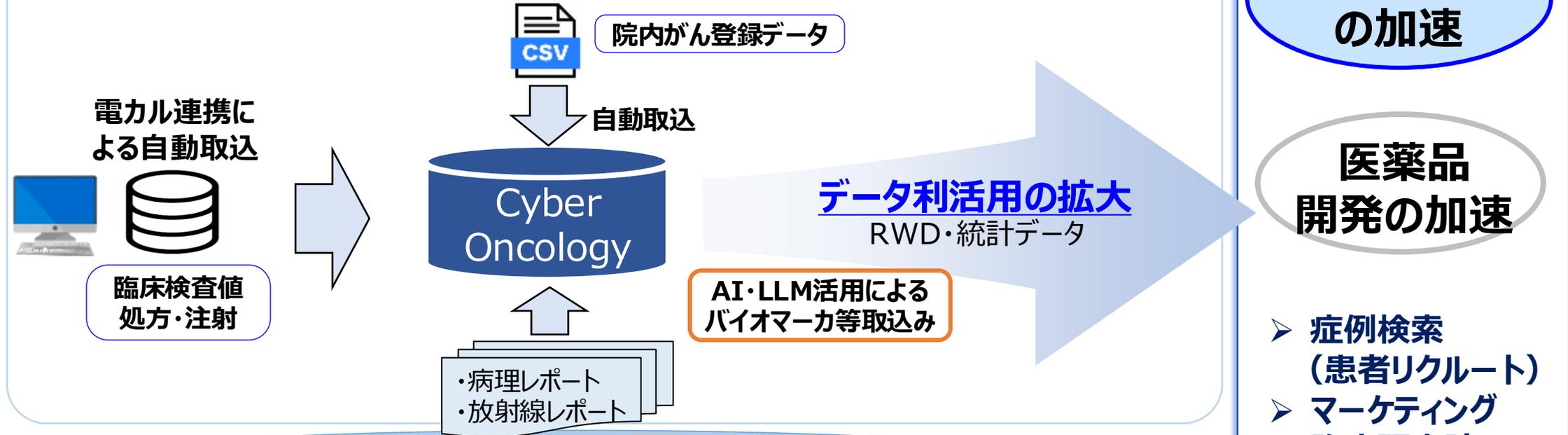
# J-CONNECT コンソーシアム (2023年4月設立)

- ✓ 京大病院を中心とした、がん登録データを活用した多施設共同研究基盤  
電子カルテデータ+院内がん登録+レセプト等をエフォートレスで蓄積
- ✓ 製薬企業の参画による新たな研究基盤として「J-CONNECT」を開始

CONNECT-2:後向きがん登録コホート  
がん診療に関する院内リアルワールドデータの網羅的収集および利活用に関する  
多施設共同研究

## CONNECT-2試験

- 院内がん登録データの取り込み
- ツールを用いて大量のデータを収集



学術研究  
の加速

医薬品  
開発の加速

- 症例検索 (患者リクルート)
- マーケティング
- 臨床研究時のベースデータ

## J-CONNECT 基盤

技術支援



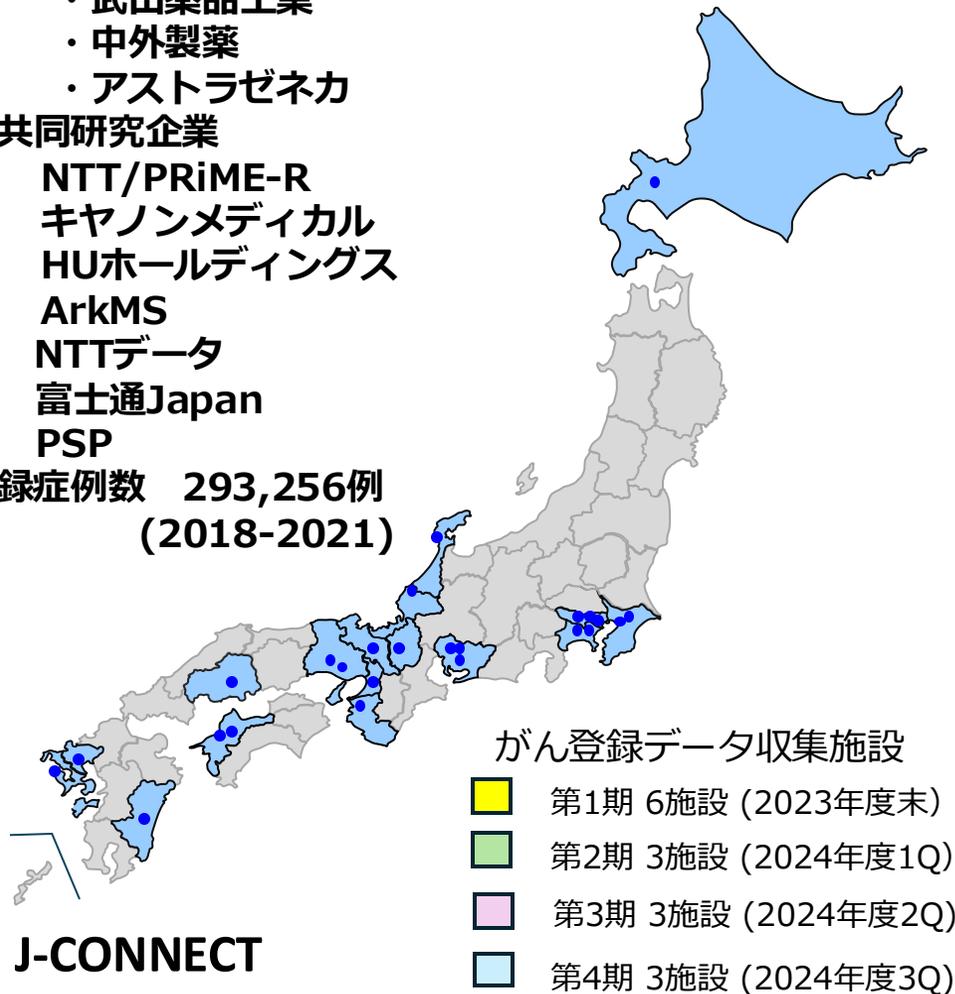
資金的支援

製薬企業等

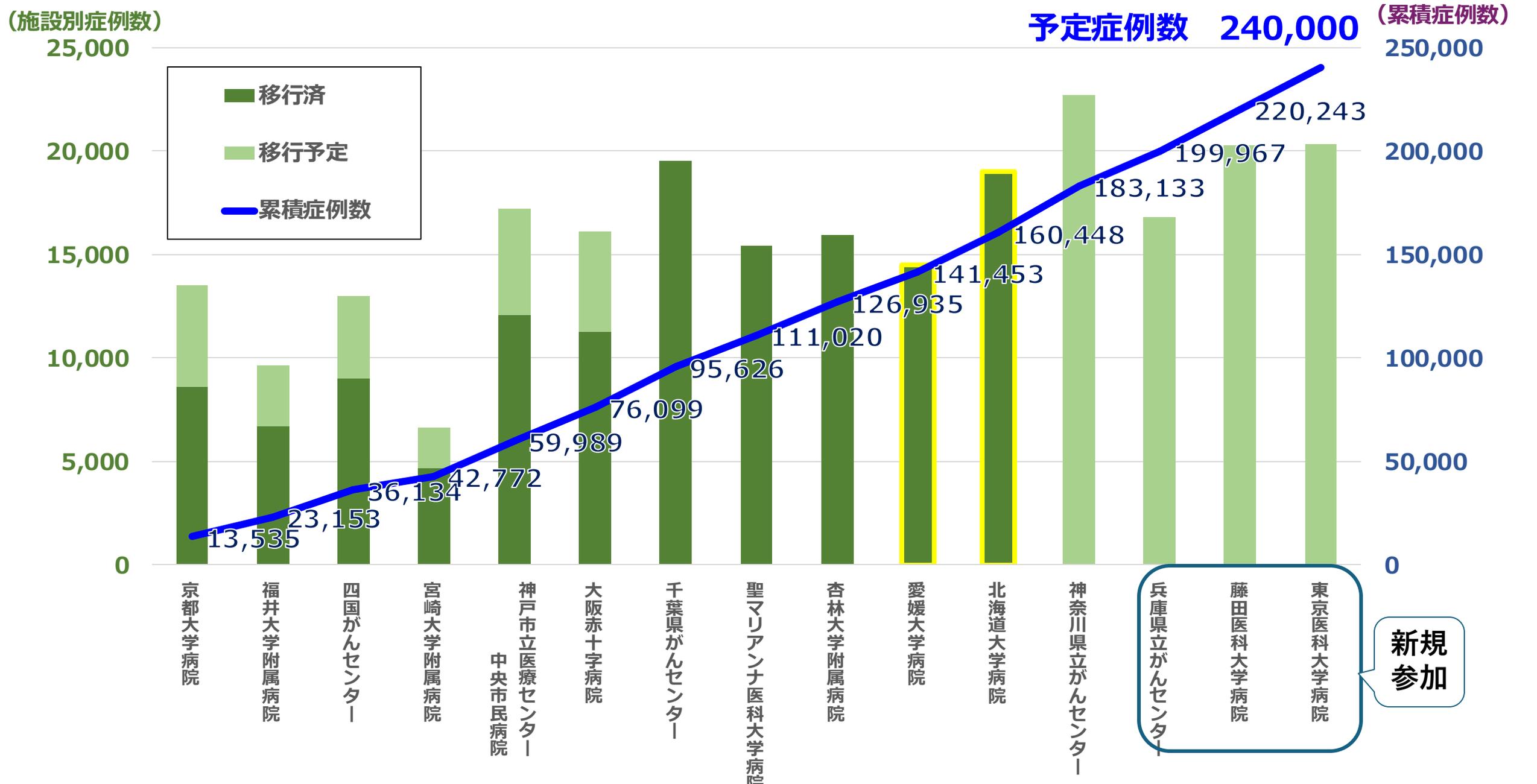
# J-CONNECT 参加25施設

## 参加施設

- ◆ 事務局 京都大学医学部附属病院
- ◆ 医療機関 25施設  
CyberOncology® 導入済み
- ◆ 製薬企業 3社 (共同研究契約済み)
  - ・ 武田薬品工業
  - ・ 中外製薬
  - ・ アストラゼネカ
- ◆ 産官学共同研究企業  
NTT/PRiME-R  
キヤノンメディカル  
HUホールディングス  
ArkMS  
NTTデータ  
富士通Japan  
PSP
- ◆ がん登録症例数 293,256例  
(2018-2021)



	都道府県	医療機関名	主担当診療科名
1	北海道	北海道大学病院	腫瘍内科 (がん遺伝子診療部)
2	東京都	杏林大学医学部付属病院	腫瘍内科
3	東京都	慶應義塾大学病院	腫瘍センター ゲノム医療ユニット
4	東京都	東京医科大学病院	消化器内科
5	千葉県	千葉県がんセンター	診療部
6	千葉県	千葉大学医学部附属病院	食道・胃腸外科
7	神奈川県	神奈川県立がんセンター	血液・腫瘍内科
8	神奈川県	聖マリアンナ医科大学病院	腫瘍内科
9	愛知県	名古屋大学医学部附属病院	消化器外科
10	愛知県	愛知県がんセンター	薬物療法部
11	愛知県	藤田医科大学病院	呼吸器内科学
12	石川県	金沢大学病院	がんセンター
13	福井県	福井大学医学部附属病院	がん診療推進センター
14	大阪府	大阪赤十字病院	腫瘍内科
15	京都府	京都大学医学部附属病院	腫瘍内科
16	兵庫県	神戸市立医療センター中央市民病院	腫瘍内科
17	兵庫県	兵庫県立がんセンター	呼吸器内科
18	滋賀県	滋賀県立総合病院	腫瘍内科
19	和歌山県	和歌山県立医科大学附属病院	第三内科
20	愛媛県	愛媛大学医学部附属病院	臨床腫瘍学
21	愛媛県	四国がんセンター	呼吸器内科
22	広島県	広島大学病院	遺伝子診療科
23	佐賀県	佐賀大学医学部附属病院	がんセンター
24	長崎県	長崎大学病院	臨床腫瘍科
25	宮崎県	宮崎大学医学部附属病院	臨床腫瘍科





# J-CONNECTにおける参照・解析の課題

1. 各施設内で統計・疫学情報にアクセス可能
2. 多施設間・J-CONNECT等のアカデミアでのネットワーク解析を可能
3. RWD解析のための標準的なデータマネジメント



**OMOP-CDM**

# OHDSI APAC : Poster presentation



Marina Bay Sands  
Dec 5-8<sup>th</sup> . 2024



## Challenges in Conducting Federated Analysis in CyberOncology Project in Japan

Shigemi Matsumoto<sup>1</sup>, Kosuke Tanaka<sup>1</sup>, Pei Liying<sup>2</sup>, Masafumi Okada<sup>3</sup>, Manabu Muto<sup>3</sup>

**J-CONNECT**  
PRIME-R

1. Department of Real World Data R & D, Graduate School of Medicine, Kyoto University
2. Department of Medical Oncology, Graduate School of Medicine, Kyoto University
3. Prime Research Institute for Medical RWD, Inc.

### Background

There is a growing demand for the development of frameworks to generate Real World Evidence (RWE) using Real World Data (RWD)<sup>1</sup>. In recent years, in the field of precision medicine, particularly in oncology. Data on patient backgrounds and treatments, and genomic information, including patient-specific biomarkers, have become critical. Additionally, outcomes related to efficacy, safety, and prognosis are indispensable. However, in real-world database research in oncology, there is a strong need to enhance the quality and transparency of RWD sources<sup>2</sup>.

In Japan, there are multiple electronic medical record (EMR) vendors, and customization is common in hospitals. This has posed significant challenges for collecting outcome data, particularly concerning efficacy and safety, which are crucial in oncology. Furthermore, the efficient collection and analysis of RWD from multi-institutional EMRs face various regulatory and ethical challenges, including compliance with the Personal Information Protection Act and ethical guidelines.

To address these challenges, we developed **CyberOncology**, a structured input support tool for EMRs. Through templated input system, this tool allows for standardized data collection across institutions, independent of the EMR vendor, enabling the creation of unified databases and the extraction of statistical insights. Additionally, since April 2023, we have established the **J-CONNECT Consortium**, comprising of 25 institutions nationwide that have adopted CyberOncology. To streamline database creation within CyberOncology, we have integrated cancer registry data from each institution and applied advanced algorithms to ensure comprehensive RWD collection.

**J-CONNECT**  
 ◆ RWD consortium  
 ◆ Established in 2023  
 ◆ 25 Hospitals  
 ◆ 3 Pharma companies

### Method

#### Mapping strategy

Description	CyberOncology Table	Vocabulary used in CO	OMOP CDM v5.4	CDM vocabulary
Patient demographics	episode	-	person	-
Adverse events (begin - end dates)	reaction	-	observation_period	-
Last visit date	episode	-	visit_occurrence	-
Cancer diagnosis	cancer	ICD-O-3	condition_occurrence	ICD-O-3
Prescriptions / Injection	prescription, injection	YJ Code	drug_exposure	ATC
Laboratory data	laboratory	CTCAE	measurement	LOINC
Adverse events	reaction	CTCAE	observation	MedDRA
Biomarker	biomarker	- (Text)	measurement	LOINC
Outcome	outcome	-	death	-

CyberOncology is a structured oncology database integrated with electronic medical records (EMRs), with the definitions of each table outlined in the Mapping Strategy. Additionally, the data structure is designed to be both compatible with the similar to the OMOP Common Data Model (OMOP CDM). The CDM vocabulary is defined as specified in the corresponding documentation.

CDM table	Variable	Value	Concept name	Vocabulary
observation	observation_concept_id	35122651 Anemia	-	MedDRA
observation	observation_date	2021-02-15	-	-
observation	qualifier_concept_id	4309261 Grade 3 on a scale of 0 to 5	-	SNOMED
observation_preiod	observation_period_start_date	2021-02-15	-	-
observation_preiod	observation_period_end_date	2021-02-28	-	-
observation_preiod	period_type_concept_id	32817EHR	-	Type Concept
measurement	measurement_concept_id	3000963 Hemoglobin [Mass/volume] in Blood	-	LOINC
measurement	measurement_date	2021-02-15	-	-
measurement	measurement_type_concept_id	32856Lab	-	Type Concept
measurement	value_as_number	7	-	-
measurement	unit_concept_id	8713gram per deciliter	-	UCUM

The following is an example of a CTCAE transformation recorded in CyberOncology. The variables and vocabulary associated with a case of Grade 3 anemia (Hb 7.0 g/dL) diagnosed on February 15, 2021, and resolved on February 28, 2021, are shown in the table on the right.

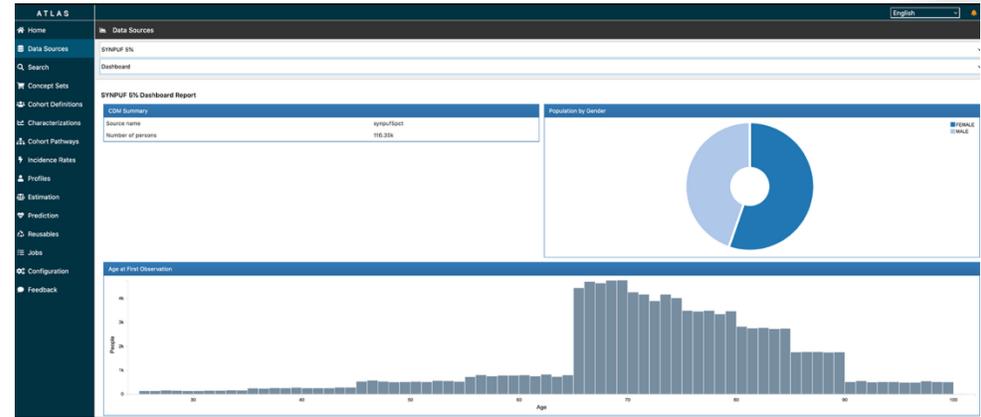
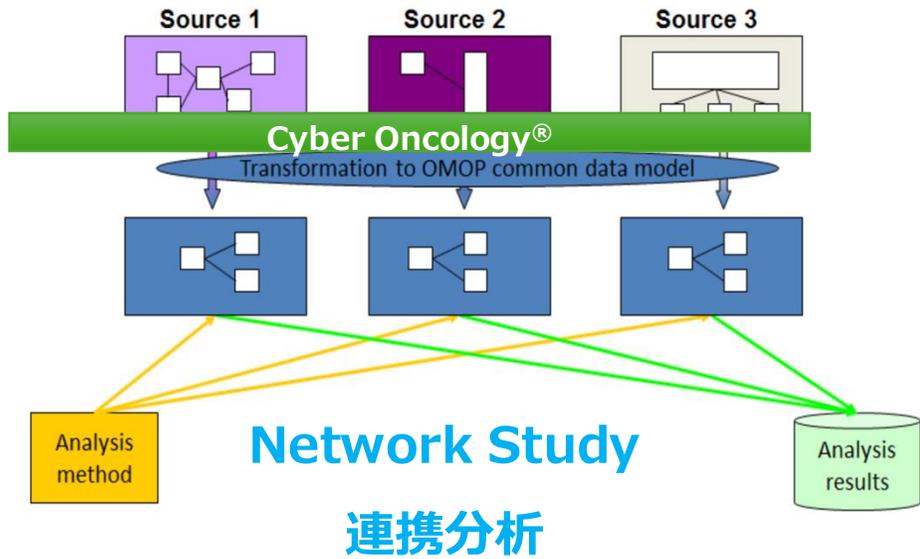
### Summary

We initiated a project to develop a federated analysis platform for the 25 member institutions of the J-CONNECT consortium by transforming CyberOncology, a structured tool integrated with electronic medical records (EMRs), into the OMOP Common Data Model (OMOP CDM). The platform is expected to become fully operational for analysis by April 2025.

### References

1. Concato J, Corrigán-Curay J. Real-World Evidence -Where Are We Now? *New Engl J Med* 2022;386:1680-1682.
2. Ramsey SD, Onar-Thomas A, Wheeler SB. Real-World Database Studies in Oncology: A Call for Standards. *J Clin Oncol* 2024;42(9).

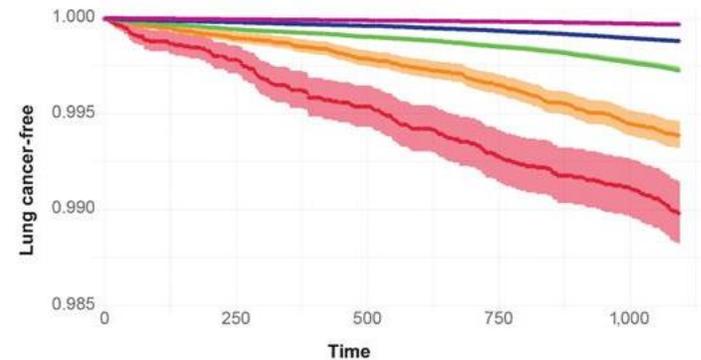
# OMOP化により連携分析が可能→解析環境を整備し勉強会



CANCER EPIDEMIOLOGY, BIOMARKERS & PREVENTION | RESEARCH ARTICLE

## Machine Learning and Real-World Data to Predict Lung Cancer Risk in Routine Care

Urmila Chandran<sup>1,2</sup>, Jenna Reips<sup>1</sup>, Robert Yang<sup>2</sup>, Anil Vachani<sup>3</sup>, Fabien Maldonado<sup>4</sup>, and Iftekhar Kalsekar<sup>2</sup>



- GROUP  $\geq 10X$  ABOVE AVERAGE (N = 15,934)
- GROUP  $\geq 5X$  BUT  $< 10X$  ABOVE AVERAGE (N = 51,344)
- GROUP  $\geq 2X$  BUT  $< 5X$  ABOVE AVERAGE (N = 377,286)
- GROUP  $\geq 1X$  BUT  $< 2X$  ABOVE AVERAGE (N = 1,162,583)
- GROUP BELOW AVERAGE (N = 3,170,459)





# Agenda

1. CyberOncologyとAIMGAIN
2. OMOP-CDM conversion
3. Demonstration

## 1. 京大病院の食道がん症例に対するOMOPデータの生成

- 対象データ：CyberOncology + DWH  
2018年～2023年の院内がん登録に含まれる食道がん患者で、  
2018年～2025年の間に化学療法を実施

## 2. OMOP DBの作成

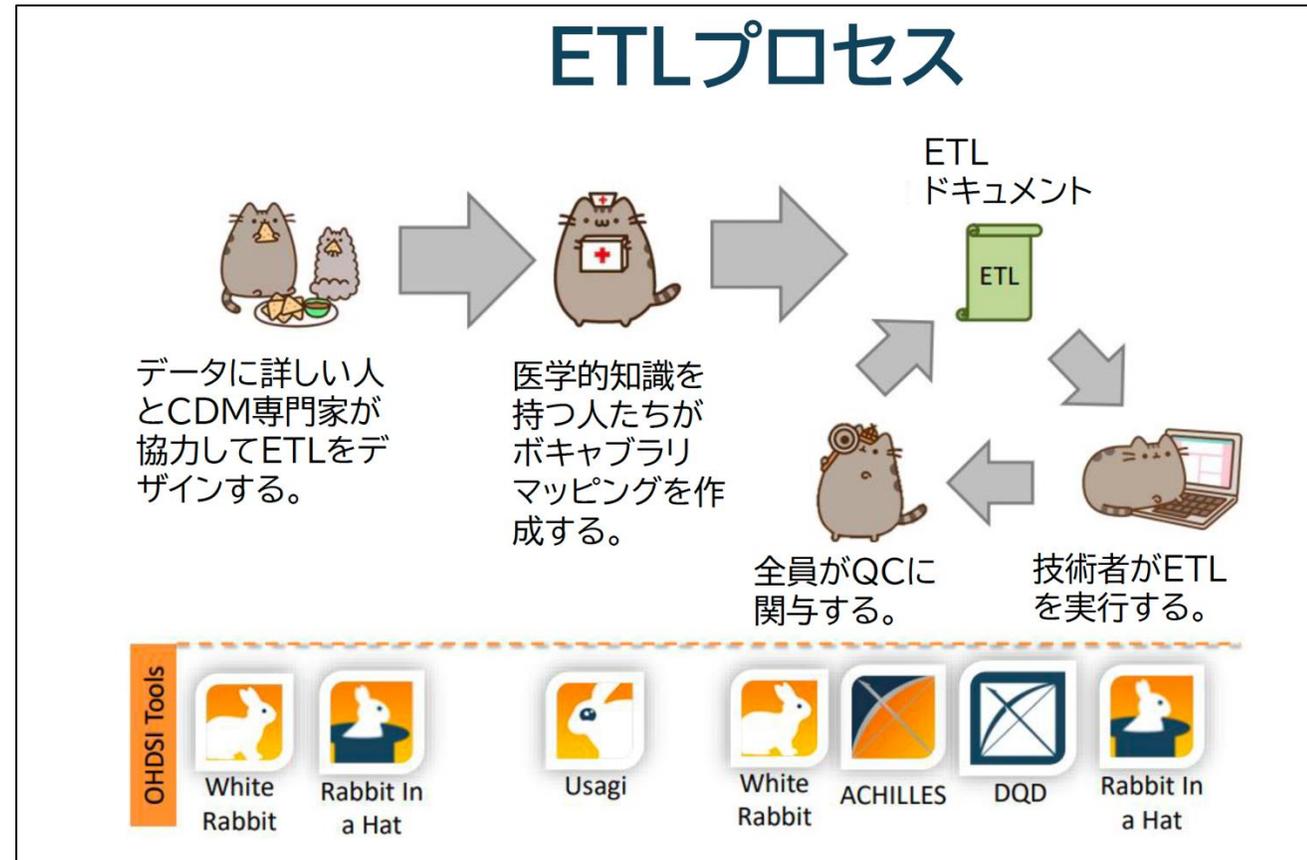
- STEP1: データソース
- STEP2: データマッピング
- STEP3: OMOP変換

## 3. ATLASでの可視化

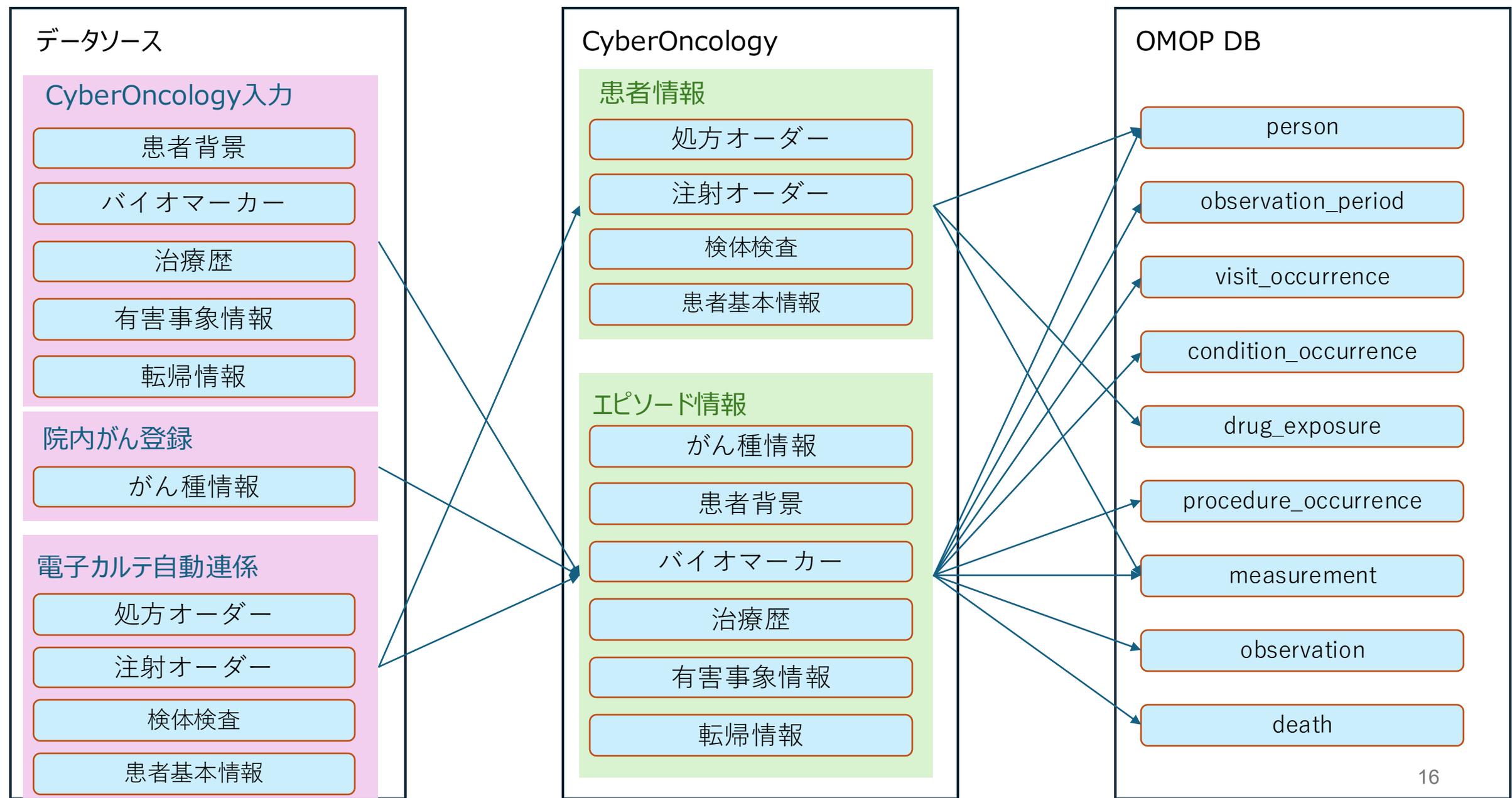
- 可視化結果（デモ）
- 問題点共有

## 4. DQDでのData Quality Check

- チェック結果サマリー
- 問題点と解決策共有

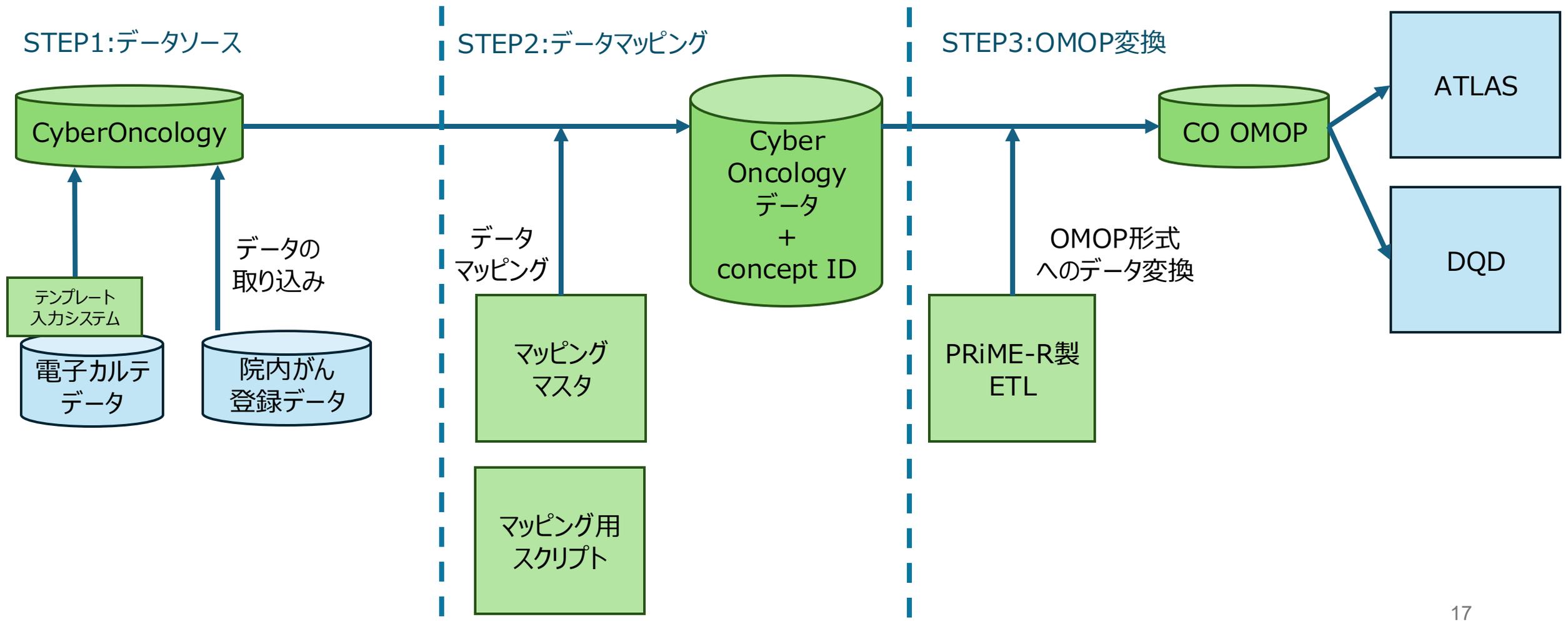


# 2. データソースからOMOP DBのマッピング



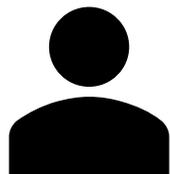
# 2. データ変換フロー

- 今回のOMOP化は以下のフローで実施した

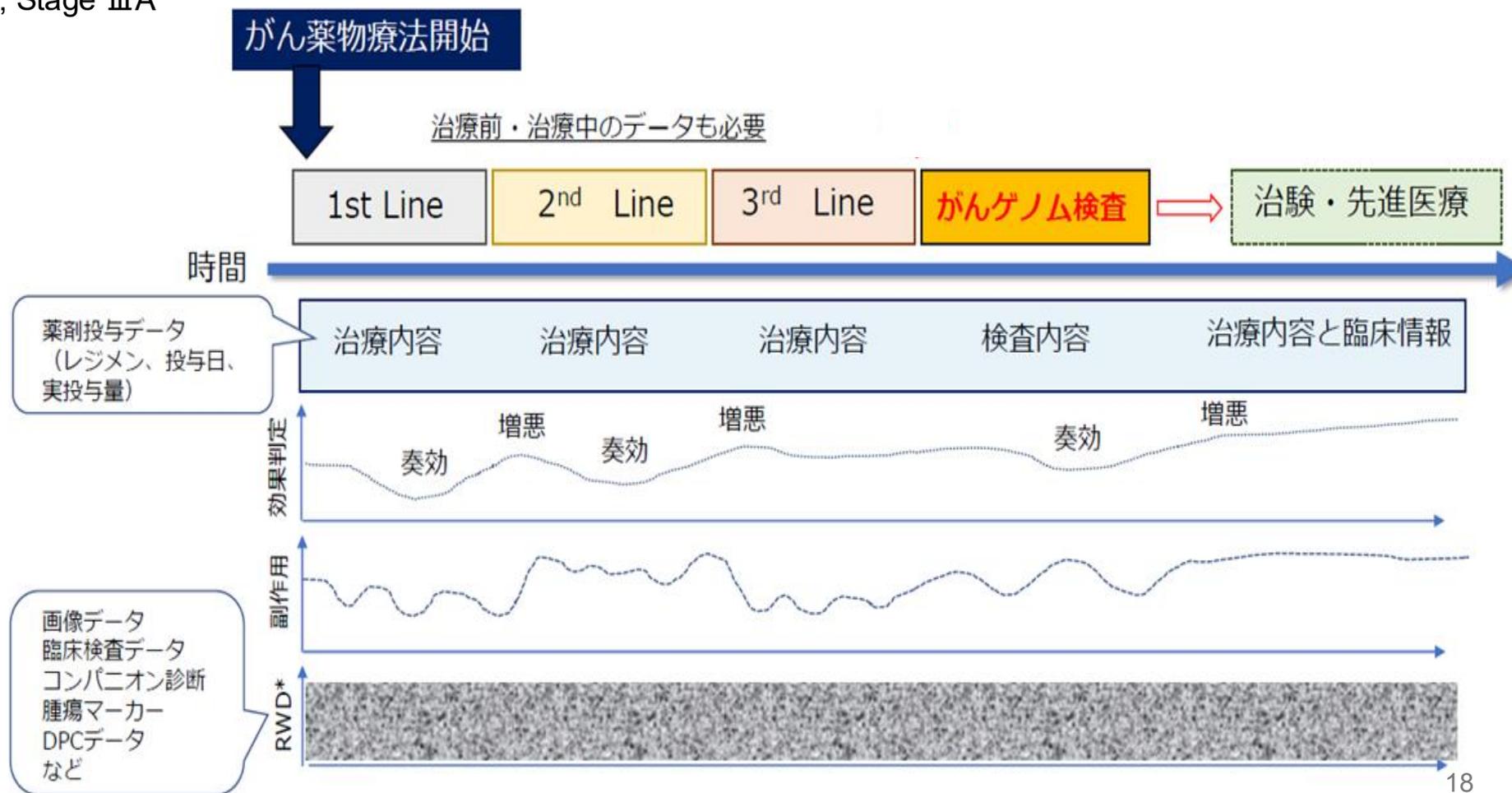


# 2. STEP1: CyberOncology

- 特定の患者様の特定のがん種に対する治療歴をエピソードという管理で構造化して蓄積する。



食道がん  
T2, N3, M1a, Stage IIIA



## cow\_patient (患者基本情報)

- 患者ID
  - 年齢
  - 性別
- など

## cow\_episode (エピソード情報)

- がん種
  - 側性
  - ECOG PS
  - ステージ情報
- など

## cow\_laboratory (検体検査情報)

- 検査日
  - 検査名
  - 検査コード
  - 結果
  - 単位
- など

## cow\_prescription (処方オーダー情報)

- 処方日
  - 薬剤名
  - 薬剤コード
  - YJコード
  - 日数
  - 容量
  - 単位
- など

## cow\_therapeutic (治療歴情報)

- 治療開始日
  - 治療終了日
  - ライン
  - ECOG PS
  - レジメン
  - 効果判定
- など

## cow\_reaction (有害事象情報)

- 発現日
  - アクション日
  - 収束日
  - 有害事象名
  - グレード
- など

- 患者ID
  - 年齢
  - 性別
  - 生年月日
  - 氏名
  - 住所
  - 原発部位
  - 側性
  - 病理診断
  - 診断日
  - 診断施設
  - ステージ
  - TNM分類
  - 外科的治療の施行日
  - 鏡視下治療の施行日
  - 内視鏡治療の施行日
  - 放射線療法の施行日
  - 化学療法の施行日
  - 内分泌療法の施行日
  - TNM分類
  - 生存最終生存確認日
  - 死亡日
  - 生存状況
  - 死因情報
  - 追跡期間
- など

### 現在作成済みのマッピングテーブル

- 医薬品
- 観察期間
- 検体検査
- 死亡日
- 最終生存確認日
- がん種情報 (ICD10, ICD-O-3, OncoTree, TNM分類)
- 効果判定
- 有害事象
- バイオマーカー

### 医薬品:

- drug\_concept\_id : YJコード→ATCコードから変換
- drug\_exposure\_start\_date: オーダー日
- drug\_exposure\_end\_date: オーダー日に投薬予定日数を足した日付

### 観察期間:

- observation\_start\_date: 初回治療開始日
- observation\_end\_date: 最終治療終了日、死亡日、最終生存確認日、のうち最大のもの

### 検体検査:

- measurement\_concept\_id: 施設固有の検査コード
- measurement\_date: 検査日

### 死亡日:

- death\_date: 転記情報の死亡日

### 最終生存確認日:

- visit\_concept\_id: 9202(固定)
- visit\_start\_date: エピソードの最終生存確認日
- visit\_end\_date: visit\_start\_dateと同じ

### がん種:

- condition\_concept\_id: ICD10
- condition\_concept\_id: ICD-O-3 (未対応)
- condition\_concept\_id: OncoTree (未対応)
- measurement\_concept\_id: Clinical TNM, Stage
- measurement\_concept\_id: Pathological TNM, Stage
- condition\_start\_date: 診断日

### 効果判定:

- condition\_start\_date: 効果判定日
- condition\_concept\_id: 効果判定結果

### 有害事象:

- observation\_concept\_id: MedDRAコード
- observation\_date: 発現日
- qualifier\_concept\_id: グレード

### バイオマーカー:

- measurement\_concept\_id: 遺伝子名 (名寄せ後)
- measurement\_date: 検体採取日
- value\_source\_value: 陰性/陽性



# Discussion 1

1. RWE創出のためのマッピングとしては未完了であるが、コメント・ご意見は？
2. Oncology領域では、病名はICD-10では不十分であるが、ICD-O-3やOncotreeを採用していく事に関してご意見？
3. 化学療法のレジメンの標準コード（HemOnc？）
4. バイオマーカー情報（遺伝子検査・免疫染色等）の標準コード？

マッピングルールに基づき、それぞれ割り当てられたconcept\_idやdate情報をcdmスキーマ以下のテーブルに格納する

cdmスキーマの格納対象となるテーブルは以下の通り

- person
- observation\_period
- visit\_occurrence
- condition\_occurrence
- drug\_exposure
- procedure\_occurrence
- measurement
- observation
- death

以下のテーブルは複数のテーブルから値を抽出してくるため、連番を変更している

- condition\_occurrence
  - がん種データ: 0~
  - 効果判定: 1000000000~
  - 有害事象: 2000000000~
- drug\_exposure
  - 処方薬: 0~
  - 注射薬: 1000000000~
- procedure\_occurrence
  - 放射線療法: 0~
  - 手術: 1000000000~
- measurement
  - 検体検査: 0~
  - バイオマーカー: 1000000000~

YJコードからATCコードに変換したが、以下の差分が生まれることを留意する必要がある。

- YJコードしか持たない情報の欠落
- メーカー（商品名）ごとの識別、薬価情報との紐づけ（特にコード単体での容量情報など）など



# Discussion 2

1. 今回のETLの方法に関し、コメント・ご意見は？
2. 複数のテーブルにまたがる変数の場合に、連番の採用は？
3. 薬剤に関し、YJからATCコードに変換しているが、今後の方向性は？

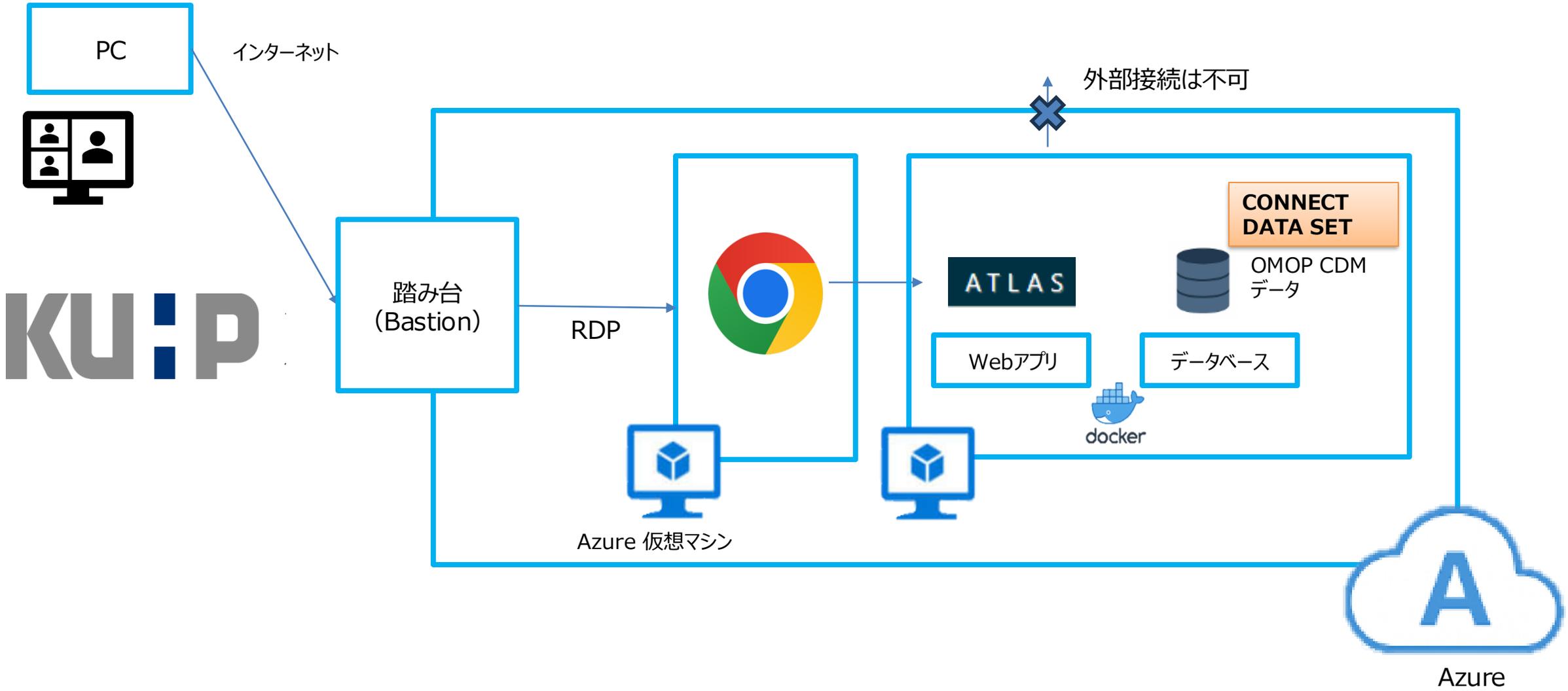


# Agenda

1. **CyberOncologyとAIMGAIN**
2. **OMOP-CDM conversion**
3. **Demonstration**

ATLAS	
 Home	
 Data Sources	
 Search	
 Concept Sets	
 Cohort Definitions	
 Characterizations	
 Cohort Pathways	

メニュー	実施内容
Data Sources	データに対する概要表示
Search	コンセプトの検索
Concept Sets	コホートを定義するためのConcept IDの集合を定義
Cohort Definitions	分析に使うコホートを定義
Characterizations	定義したコホートに対する特徴量を分析
Cohort Pathways	定義したコホートに対する経路を分析 (サンバースト図の描画)



KU:P

## 【対象コホート】

1. 2018-2023年の院内がん登録で食道がん (ICD-10 C15.X) 薬物療法あり

## 【特徴量分析項目】

 Characterizations

1. 年齢分布
2. 性別
3. 病理分類 (ICD)
4. TNMステージ分布

## 【経路分析 (サンバースト)】

 Cohort Pathways

1. 治療法 (手術/放射線/放射線化学療法/化学療法)
2. 化学療法 (oxaliplatin/cisplatin/fluorouracil/docetaxel/paclitaxel)

 Search

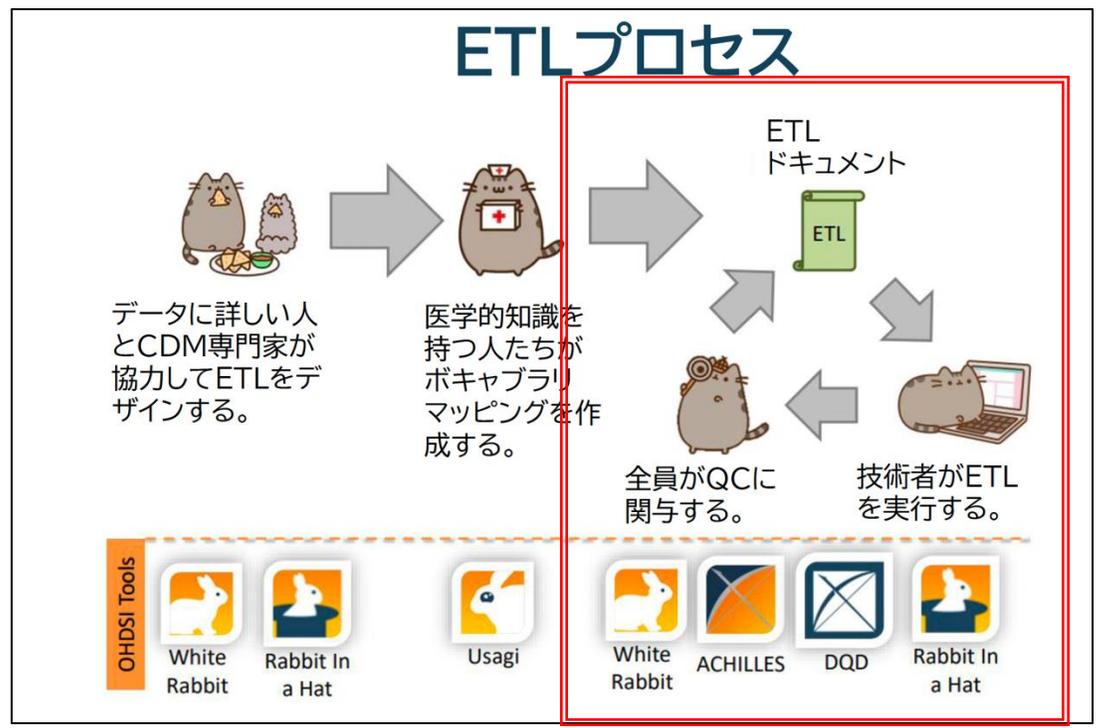
 Concept Sets

 Cohort Definitions



# ATLAS Demo

- 新規薬剤のNivolumabやPembrolizumabのデータが抽出出来ていない  
→データ欠落原因を調査中、マッピングやETLの不具合を確認する。
- 食道がんには通常使われない薬剤が出てきている  
→該当症例のデータに対して元データを確認したところ、すべて重複がん患者えお確認。  
→Observation\_period等の日付の誤値の可能性があり、ETL等の見直しを実施。  
→コホート作成時の設定なども再度見直す。



## Data Quality Dashboard によるチェック

以下の3つの観点からデータのチェックを行い、改善のPDCAにつなげていく予定

- 異常な数値や矛盾するデータがないか
- OMOP CDM データ構造、コンセプトへの定義などが仕様に従っているか
- データの欠損、参照整合性違反がないか



項目	説明	結果値
総チェック数	Data Quality Dashboard の実行したチェックの件数	2,348
NA	該当するデータがなくチェック対象外となった件数	1,362
FAIL	実行したチェックを通過しなかった件数 ( エラーの総数 )	<b>48</b>
SQL ERROR	SQL 実行の戻り値がエラーのため、チェック対象外となった件数 ( データに依存性なし )	22
有効チェック数	実質評価が可能なチェック件数	990
有効チェック範囲の合格数	実質評価可能な範囲でチェックを通過した件数	986
有効チェック範囲の合格率	実質評価可能な範囲でチェックを通過した割合	<b>99.6%</b>

Data Quality Dashboard の結果、データが登録されている有効チェック範囲内において **99.6% の項目が PASS** しながら、**48 の項目が FAIL** ( 下記概要 )

概要	詳細、影響と対策
SQL ERROR	データ依存ではなく、テーブルやフィールドの定義が異なることで発生したエラー 対象のフィールドがないなど、以前のバージョンで定義されているテーブルが混在 テーブル、フィールドの定義、バージョン差異を確認し、OMOP CDM のバージョンの統一化を行う
テーブルにデータがない	今回の構築で対象外としたテーブル ( durg_era, condition_era など ) にデータがない 複数のイベント情報を1つのエピソードしてまとめられていないため、エピソード単位での分析に影響する ETL 上で、生成するためにエピソードをまとめる情報をレジメン等を参考に設定可能か検討する
日時データの不整合	" observation_priod_end_date " が死亡日以降に設定されているレコードが存在 分析データに死亡データが含まれるなどの影響が考えられる ETL 、元データとの整合性を確認し、修正する
concept_id の不整合	一部の concept_id が正しくマッピングできていない ( "0" となっている、domain が異なる など ) 項目 ( condition_concept_id ) によっては集計・分析内容に影響を及ぼす可能性がある マッピング、ETL の確認を行い、修正する



# Discussion 3

1. 今回のDQDに関し、コメント・ご意見は？
2. RWD故に完璧は目指せないが、ETLの必要十分条件？
3. 海外のOHDSIコミュニティと会話するためには AchillesやDQD等のツールをどこまで学習する？



**Thank you!**