H23年度 統合化推進プログラム進捗報告会

大規模ゲノム疫学研究の統合情報基盤の構築

京都大学医学研究科附属ゲノム医学センター
松　田　文　彦

JSTバイオサイエンスデータベースセンター
「基盤技術開発プログラム」および「統合化推進プログラム」
平成23年度　進捗報告会
2012年2月24日　TKP東京ビジネスセンター

研究開発の目標・ねらい

- **ゲノム疫学研究の情報基盤の構築と公開**

  「ながはま0次コホート研究」の一万人の生活習慣・環境情報、臨床情報、ゲノム・オミックス情報を標準化し、データベースを構築する。

  集積した情報を、個人情報保護のもと、医学・生命科学研究者に提供する。

- **データベースの枠組みの提供と情報の連結**

  これをモデルケースとして、同様の研究をおこなう際に即時活用可能なかたちで、分子疫学研究者にデータベースの枠組みを提供する。

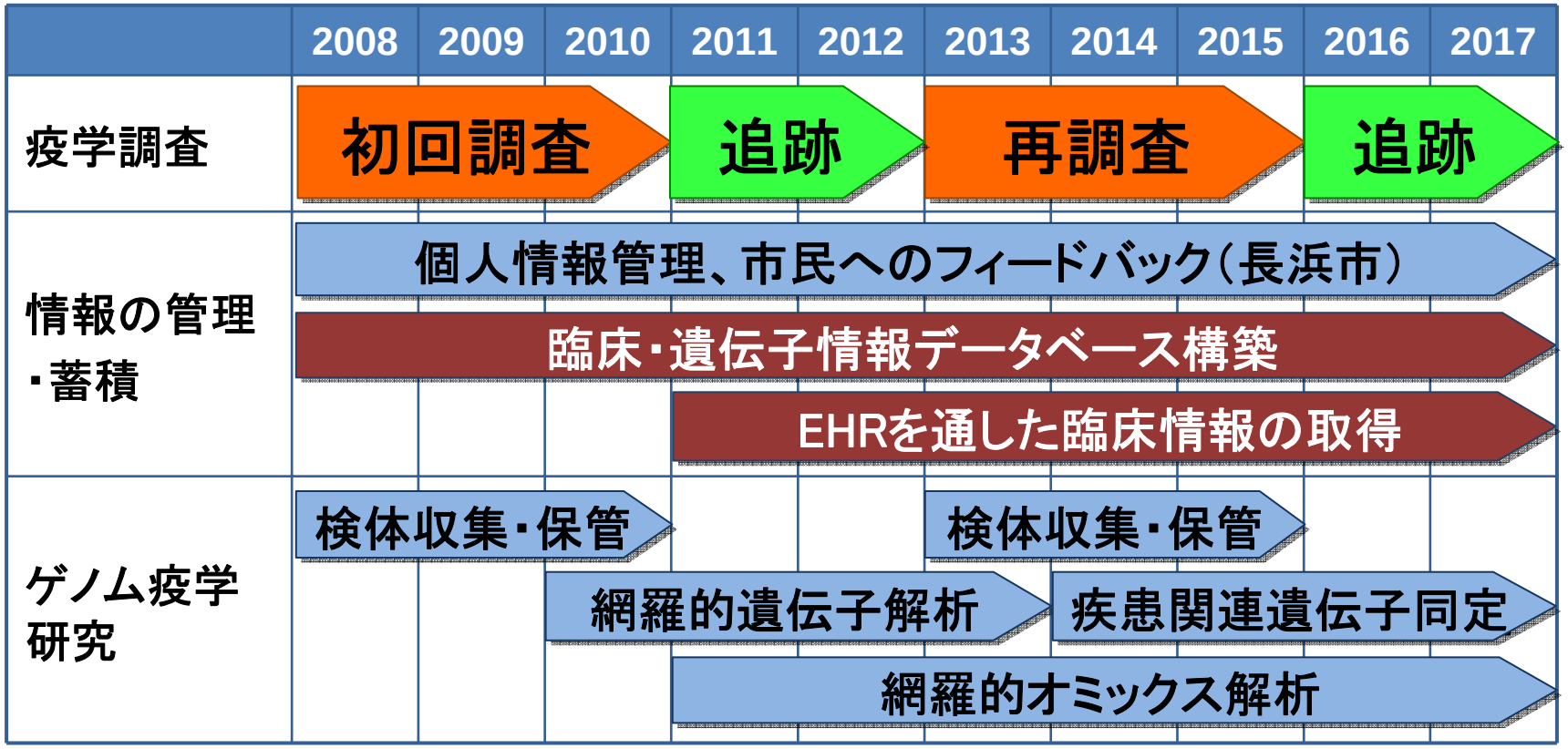
  他の研究で蓄積された遺伝型・表現型データを連結、共有することで、個別の研究で得られた情報の一元化によるそれらの再利用を促す。

- **ゲノム情報科学の若手研究者の育成**

  バイオインフォマティクス、遺伝統計学の若手研究者に教育訓練（OJT）をおこない、これらの分野の将来における中心的研究者の育成をはかる。

「パーソナルヘルスレコード」の情報提供先として機能できる汎用性の高い健康情報管理システムを提案

ながはま0次コホート事業のロードマップ



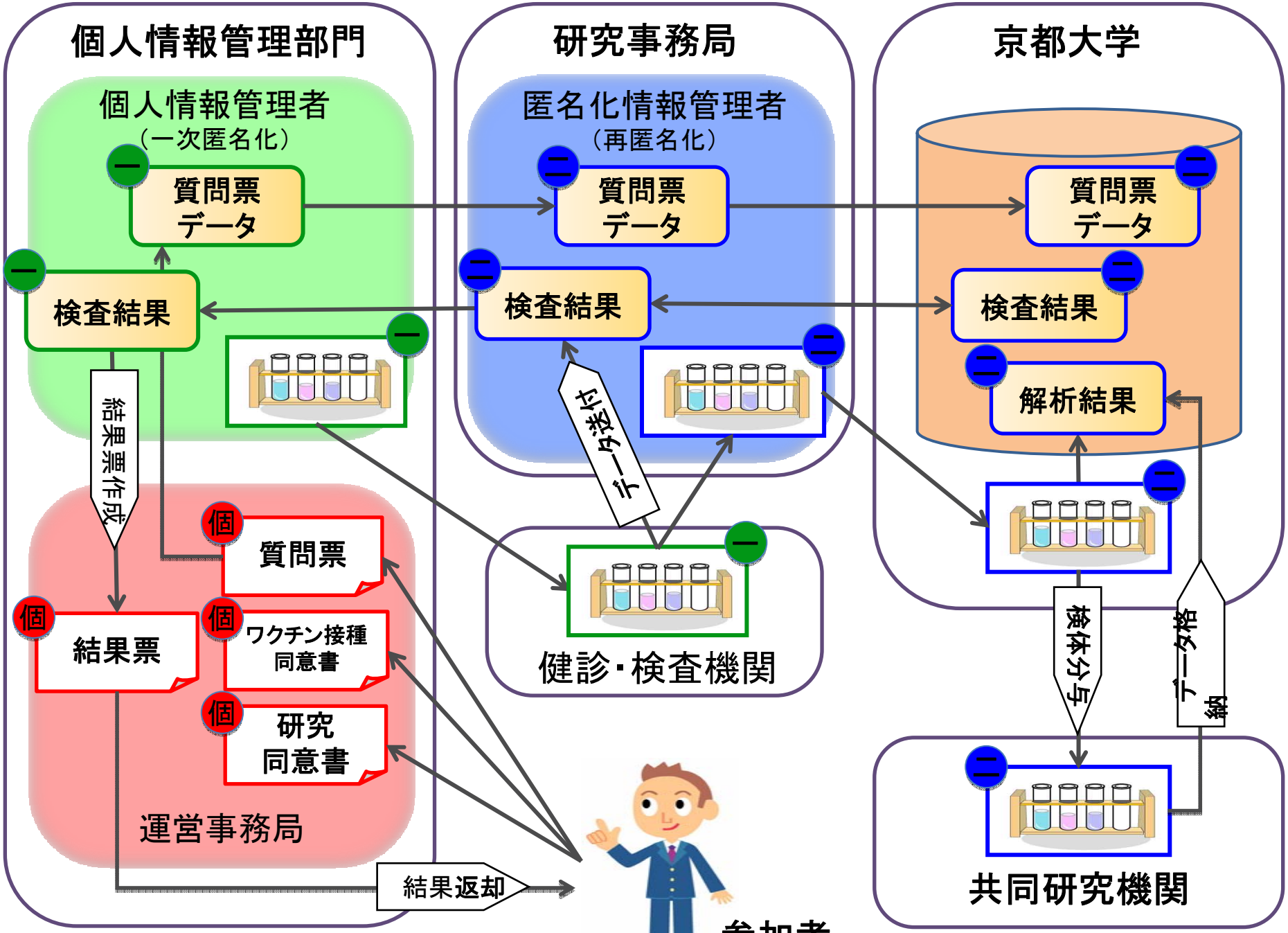
参加者: 10,080人 （男性 3,333人、 女性 6,747人）

取得情報: 環境・生活習慣情報 （質問票） 742項目 合計748万件
生理学、血液学、生化学測定値 145項目 合計146万件

ゲノム解析: ゲノムスキャン（Illumina） 3,713検体 合計55億ジェノタイプ
エクソームシークエンシング 64検体

本事業でNBDCに蓄積するデータ項目

- 質問票による環境・生活習慣情報
  - 742項目　約748万件
- 生化学・血液学・生理学的測定値
  - 145項目　約146万件
- 参加者のゲノムスキャン情報
  - Illumina610K（約1,800検体）、2.5Mアレイ（約3,200検体）
- エクソームシークエンス情報
  - 500検体を目標とする
- 網羅的メタボローム解析情報
  - 島津製作所GC-MSを利用　全検体解析　200～250ピーク分
- EHRによる疾患罹患情報
  - ITネットワークが整備されることが前提



データベースシステム概念図

京　都　大　学

統合DBセンター

SNPアレイ
全ゲノム塩基配列

質量分析・
その他測定機器

採血・診察・問診

ゲノム
DB

中間形質
DB

臨床情報
DB

分子疫学
DB

管理者

ミラーリング
（VPN）

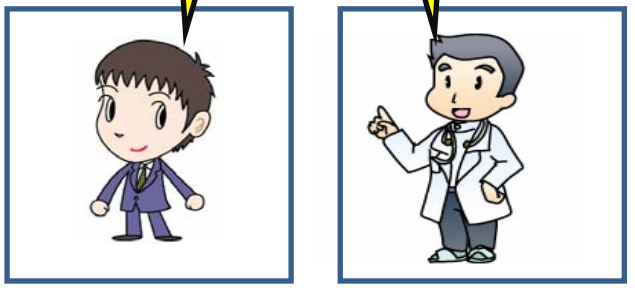
権限承認
（SSL）

共同研究用

一般公開用

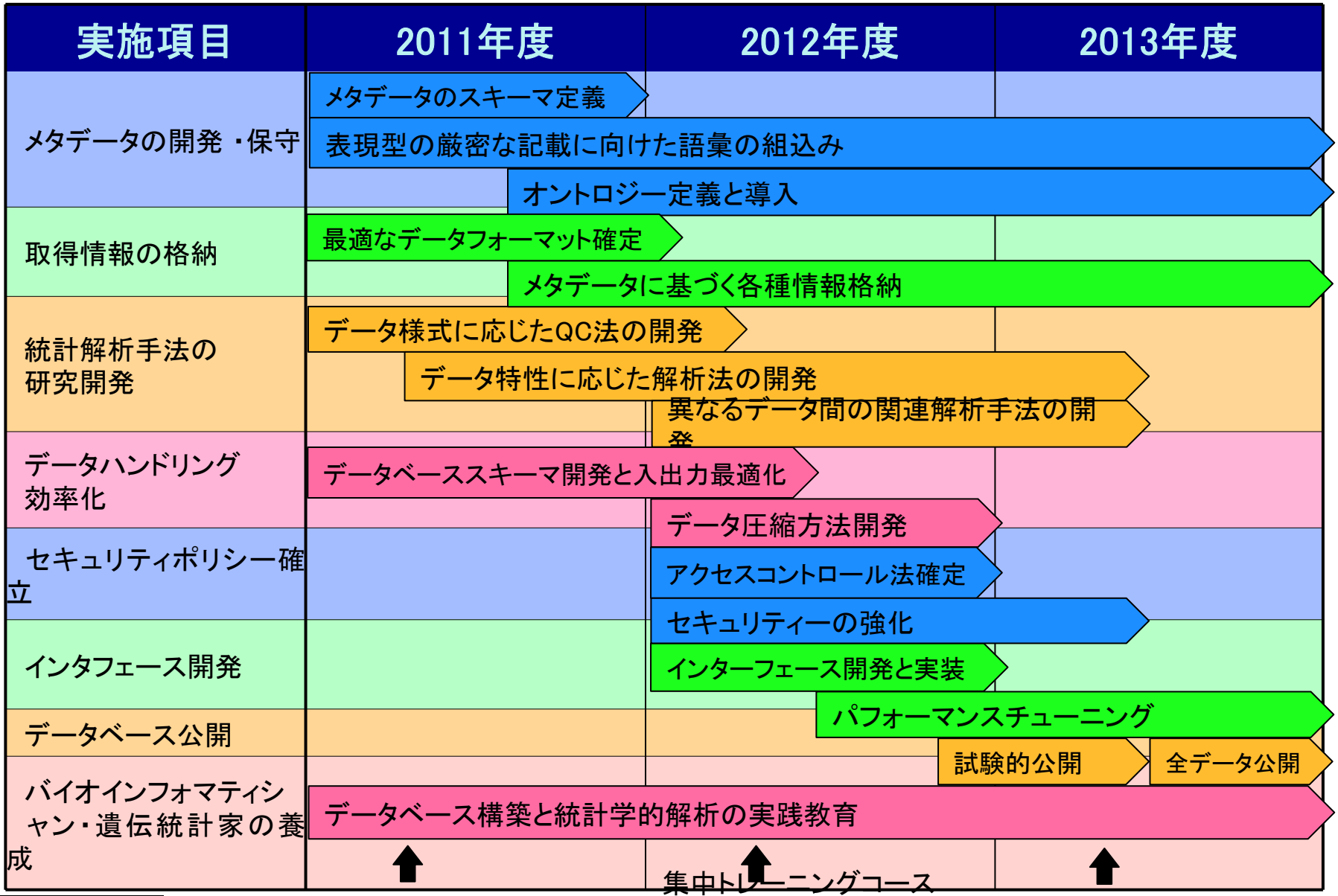
アクセスコントローラー

閲覧・登録・解析
（SSL/VPN）



データベース利用者

本研究開発のロードマップ



H23年度の開発項目

項目１　メタデータの開発・保守

項目２　取得情報の格納

項目３　データハンドリングの効率化

項目４　統計解析手法の研究開発

項目５　EHRの導入による疾患関連情報の取得
（新規開発項目）

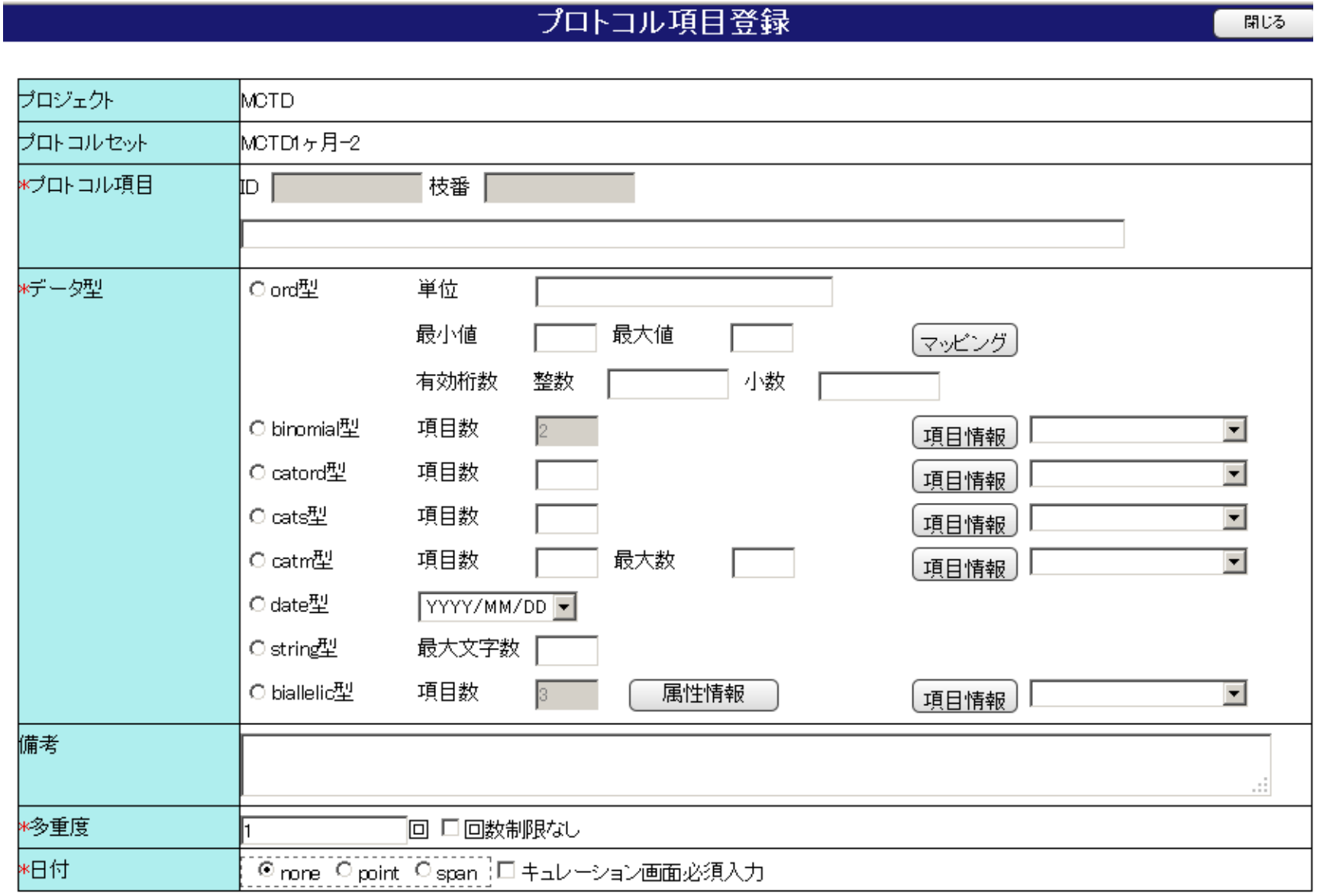
各データ項目にはデータ型と、その型に応じた制約を定義する

データ型	制約	例
連続値	最大最小、打ち切りの有無	バイオマーカー
順序ありカテゴリ	カテゴリの値とコード	質問票
順序なしカテゴリ	カテゴリの値とコード	質問票
SNP情報	ゲノム上の位置とアレル	SNP
その他の多型情報	ゲノム上の始終点	CNV
波形	ピークの位置と強度	中間形質情報
文字列・日付		自由記載情報

ながはまO次コホート事業　データ項目数

種別	連続値	カテゴリー値		文字列・日付	総数
		順序あり	順序なし		
質問票	290	187	485	52	1014
尿・血液検査	85	24	2	0	111
中心血圧	24	0	0	9	33
CAVI	111	0	9	7	127
呼吸機能	32	0	0	5	37
眼軸長	29	0	16	7	52
総計	571	211	512	80	1374

データ項目定義インターフェイス



データの標準化

- 類似する項目の統合・関連付け
- オントロジー化がなされた外部の標準データと関連付け
- オントロジー化が不十分な場合は、独自にオントロジーを追加
- 加えて外部標準データ等を組みいれ、疫学標準データ項目セットを定義
- キュレーターによる認証

複数のプロジェクト統合に必須である
外部利用者にとってデータの突合が容易になる

テキストキュレーション
文字列型→カテゴリ型への変換とオントロジー（ICD10）付与

変換前	変換後 （ICD10）
かんぞう・肝・肝臓	肝癌 （20057051）
すいぞう・すい臓	膵癌 （20078952）
だいちょう・大腸	大腸癌 （20069739）
のど・喉	喉癌 （ ）
コウトウ・こうとう・喉頭	喉頭癌 （20061154）
リンパ腫	リンパ腫 （20053644）
胃	胃癌 （20054178）
胃肺	胃癌 （20054178） &肺癌 （20073570）

取得情報の格納とハンドリングの効率化

- 実験をともなう分析・解析結果

  原則として実験機器から得られる生データを最小限の加工でデータベース登録可能なパイプラインを構築

- 臨床情報

情報登録用統一フォーマットを設計

網羅的ゲノム解析ツールの Webインターフェイス構築

- 検体・マーカーセットによるジェノタイプ抽出
- 実験結果の品質管理
- 検体間の遺伝的近縁度 （PCA/近縁）
- ゲノムワイド関連解析
  - 2x3分割表
  - 線形回帰系（logistic, linear, Tobit, ordlog)
  - 詳細物理地図作成
- 形質間の関連
  - Linear, logistic, Johnckeere-Terpstra, trend, ..
- ハプロタイピング、ジェノタイプ欠損補完

いかにして疾患罹患情報を取得するか

- 質問票による調査

  自己申告情報の確度に問題

- 特定健診の情報

  受診率の低さに問題

- 地域の疾病登録制度の利用 （がん登録など）

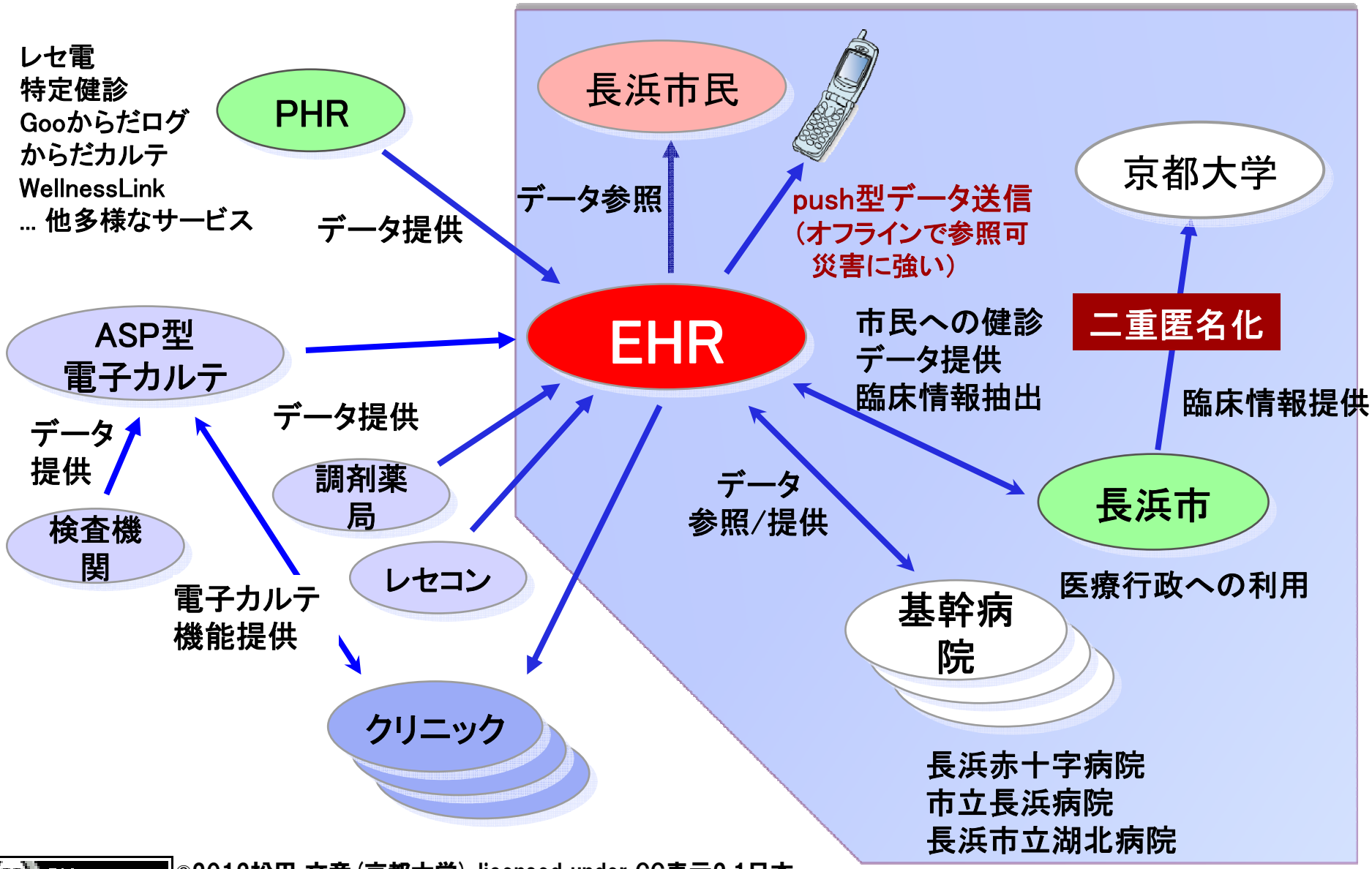
  情報が一部の疾患に限定される

- 移動・死亡情報 （住民基本台帳、死亡小票の閲覧）

  死亡者の情報のみしか集まらない

EHRを利用した疾患関連情報取得の検討を開始

まいこネットの疫学研究利用



疾患関連情報取得システムの段階的開発

１） 一疾患・一施設

希少難治性疾患　京都大学医学部附属病院

患者数百人における133項目の臨床情報
4データポイントの追跡

２） 一疾患・多施設

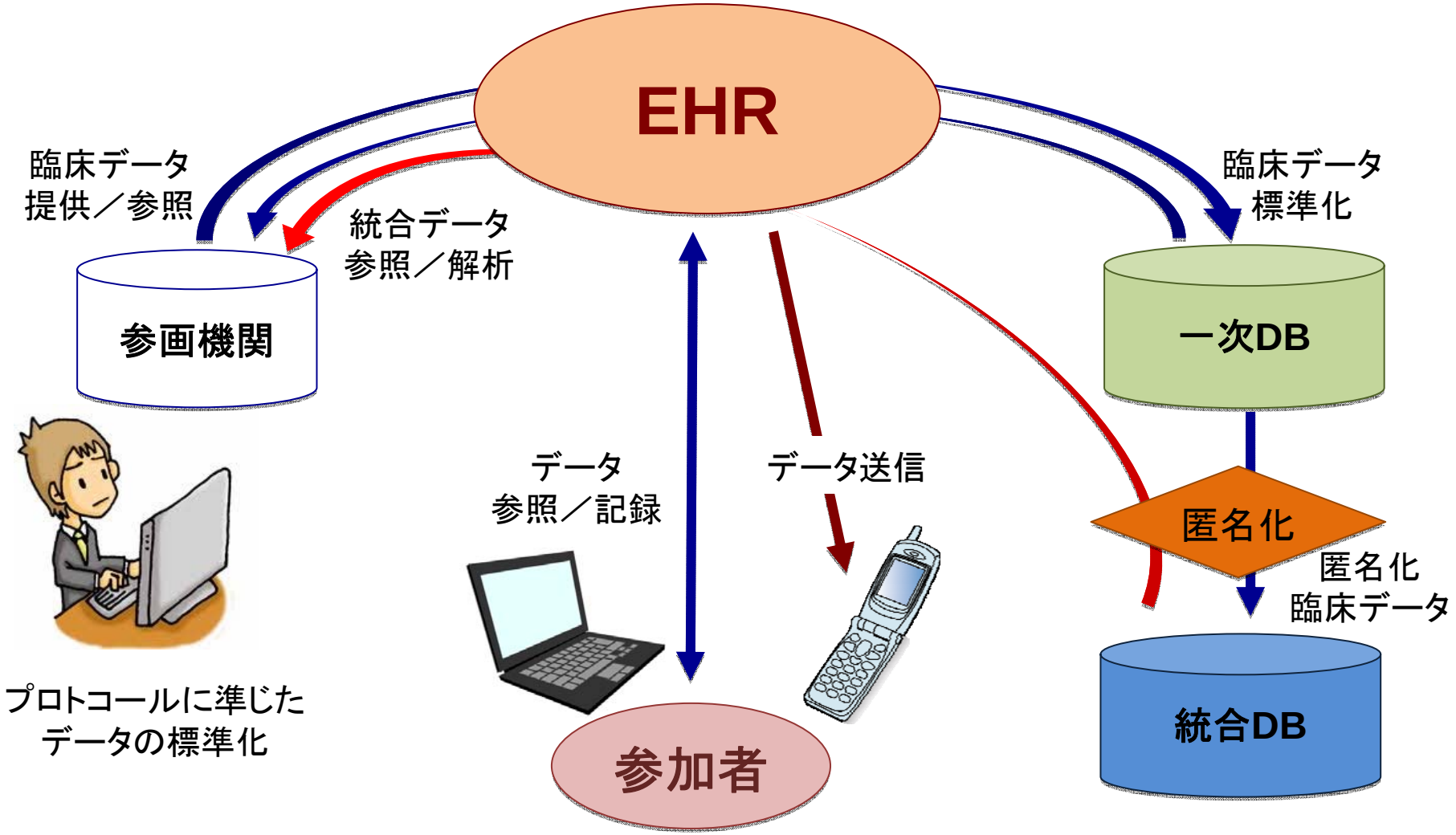
希少難治性疾患　共同研究機関

３） 多疾患・多施設

ながはまコホート　長浜市の三基幹病院

参加者10,080人
参加登録前からの疾患罹患情報

大規模コホートとEHRの統合による統合疫学データベース構築



MML (Medical Markup Language)形式で出力した疾患関連情報

```
          <claim:bundle claim:classCode="211" claim:classCodeId="Claim007">
            <claim:className>内服薬剤(院内処方)</claim:className>
            <claim:admMemo>1 × 朝食後</claim:admMemo>
            <claim:bundleNumber>7</claim:bundleNumber>
            <claim:item claim:subclassCode="0" claim:subclassCodeId="Claim003" claim:code="210">
              <claim:name>内服薬剤</claim:name>
            </claim:item>
            <claim:item claim:subclassCode="2" claim:subclassCodeId="Claim003" claim:code="621375001">
              <claim:name>バファリン配合錠A81　81mg</claim:name>
              <claim:number claim:numberCode="12" claim:numberCodeId="Claim004"
claim:unit="T">1.0</claim:number>
            </claim:item>
            <claim:item claim:subclassCode="2" claim:subclassCodeId="Claim003" claim:code="612170710">
              <claim:name>ノルバスク錠(5mg)　★Ca拮抗薬★</claim:name>
              <claim:number claim:numberCode="12" claim:numberCodeId="Claim004"
claim:unit="T">1.0</claim:number>
            </claim:item>
          </claim:bundle>
          <claim:bundle claim:classCode="211" claim:classCodeId="Claim007">
            <claim:className>内服薬剤(院内処方)</claim:className>
            <claim:admMemo>1 × 夕食後</claim:admMemo>
            <claim:bundleNumber>7</claim:bundleNumber>
            <claim:item claim:subclassCode="0" claim:subclassCodeId="Claim003" claim:code="210">
              <claim:name>内服薬剤</claim:name>
            </claim:item>
            <claim:item claim:subclassCode="2" claim:subclassCodeId="Claim003" claim:code="610412202">
              <claim:name>パリエット錠(10mg)</claim:name>
              <claim:number claim:numberCode="12" claim:numberCodeId="Claim004"
claim:unit="T">1.0</claim:number>
            </claim:item>
```

希少難治性疾患のデータ取得状況

評価ランク		割合
◎	1対1で対応付けが可能	16% (21/131)
○	1対1では対応付けが困難であるが、参考情報をして参照が可能	16% (21/131)
■	コホートシステム側で算出が可能	18% (24/131)
▲	調査が必要であるが、結果により対応付けできる可能性がある	5% (7/131)
△	条件により1対1で対応付けが可能	1% (1/131)
×	EHR上に移行されていない情報	32% (42/131)
—	電子カルテ上に保存されていない情報	12% (15/131)

キュレーションインターフェイス

一次ID： 123456789012

再調査に向けたWebQ&Aを活用した情報取得システムの構築

- Webを通した情報の直接収集で、収集およびキュレーションにかかる労力を削減
- 項目定義、表示の順序、回答権の制約を定義→WebQ&A画面の自動生成

ID	項目名	データ型	制約	設問	選択肢	回答権制約
2	性別	binomial	0=male 1=female	性別をお答えください	0=男 1=女	
3	身長(cm)	numeric	<220	身長をお答えください		
15	子宮がん	binomial	0=no 1=yes	子宮がんと診断されたことはありますか	0=いいえ1=はい	2=1

H24年度の研究実施内容

- 疫学システム全体の洗練・多機能化・セキュリティの強化
  - 個人情報保護・匿名化の枠組みの洗練
  - データ共有の枠組み
  - EHR連携
- データ項目の標準化　特にメタボローム データ
- インタフェイスの実装
  - データキュレーション用インタフェイス
  - 質問票のWeb化
- 解析手法の開発・実装