

医療ビッグデータ利用の 現状と課題

20190710

東京大学大学院医学系研究科

公共健康医学専攻臨床疫学・経済学

松居 宏樹

本日本話しする内容

この講演の目標

- NDBに含まれる情報（データ構造）を理解する。
- データハンドリングのイメージをつかむ。
- オンサイトセンターで行えることのイメージをつかむ。

本日も話している内容

想定する対象者

- 公的な大規模医療データを利用したことがなく、今後利用を検討している方々

本日も話しする内容

もくじ

- 大規模データベースを利用した研究活動
 - Administrative Claims Database とは？
 - NDBとはなにか
 - NDBのデータに含まれる情報
 - NDBデータのハンドリング
 - NDBデータ構造上の問題
- NDBの利用方法
- NDBオンサイトセンターの紹介

簡単な自己紹介

松居宏樹

東京大学大学院 臨床疫学・経済学教室 助教

主たる研究テーマ

大規模データベースを用いた臨床疫学研究

- DPCデータベース
- NDBデータベース
- 介護レセデータベース

はじめに

NDB とは何か

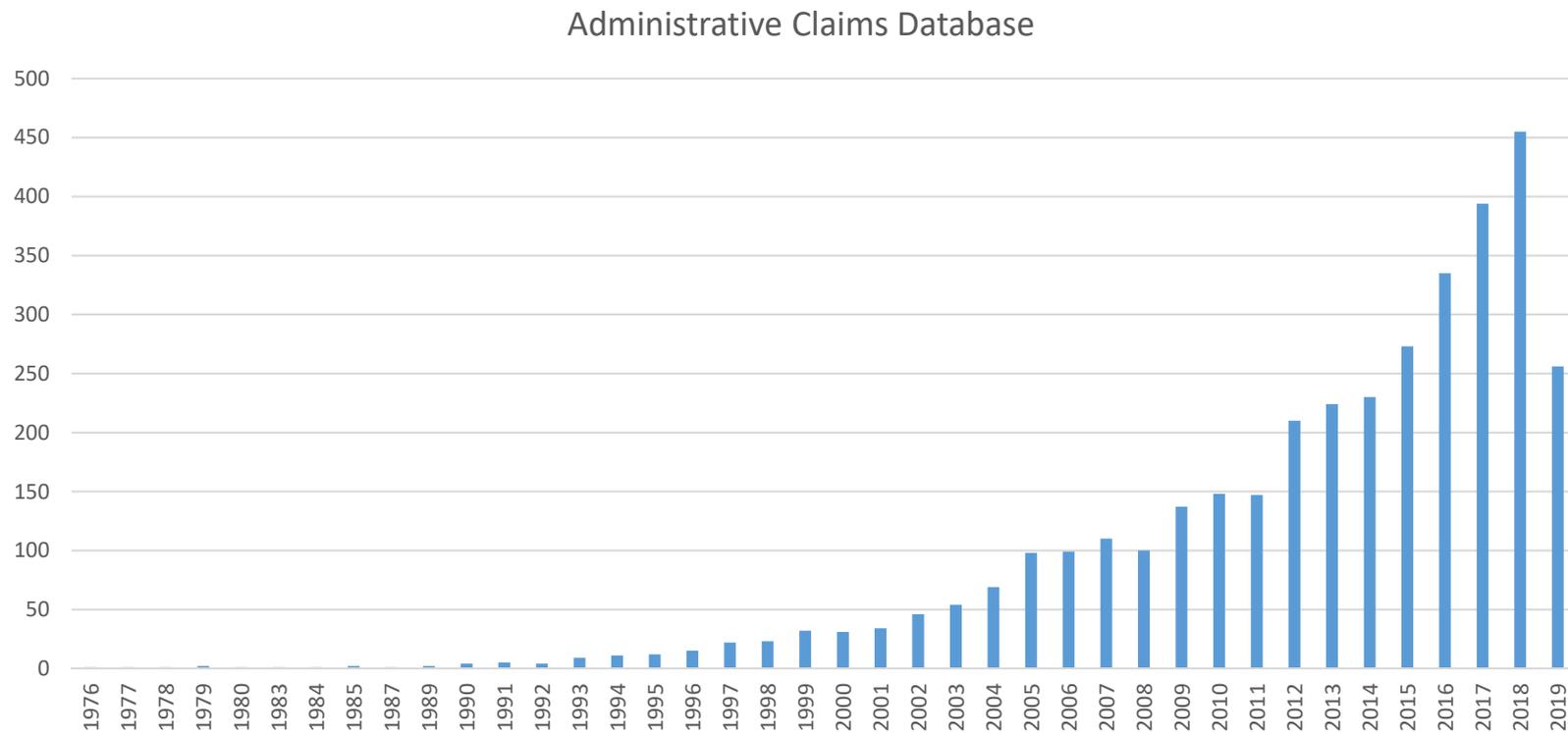
- 正式名称：レセプト情報・特定健診等情報 データベース
- 利用目的は全国医療費適正化計画及び都道府県医療費適正化計画の作成、実施及び評価 に資するため
(高齢者の医療の確保に関する法律 第16条)
- 保有：厚生労働大臣
- 内容
 - レセプトデータ 年間約18億7000万件 (H28年度)
 - 特定健診・保健指導データ 年間約2,730万件 (H28年度)
 - Administrative Claims Database の一種

Administrative Claims Database の研究利用

- 記述統計研究
 - 薬剤の使用状況・疾病の分布など
- 政策評価研究
 - 政策施行前後での患者受療行動の変化など
- Ecological 研究
 - 地域間の記述統計・相関解析など
- 治療やリスクファクターとアウトカムとの関連性（因果関係の検証）
 - 自然実験・操作変数などを用いた治療効果・政策効果の検証など

Administrative Claims Database の研究利用

研究利用する試みは数多く行われている。



検索日: 20190626.12:00

海外の大規模データ

海外の大規模臨床データベース

- 米国のCMS Medicare DataやNISデータ
- 英国のClinical Practice Research Datalink (CPRD)
- 台湾のNational Health Insurance Research Database
- 韓国のHealth Insurance Review Assessment Service

…等々

NDBの研究利用の現状

Administrative Claims Database の比較

	企業が提供する レセプトデータ	NDB		DPC Data	CMS data (米国)
		レセプト	特定健診情報		
収集元	国内の保険者	審査支払機関 社会保険診療報酬支払基金		DPC 病院	Medicare, Medicaid
収集対象	特定の健保加入者	全国民	特定健診対象者	DPC 入院患者	加入者
データ規模	数百万人	1億2000万人	2400万人	1000万人(2017)	5,000万人 (2012)
個人の縦断的観察	可	可	可	可(同一医療機関内)	可
患者の重症度	なし	なし	検査結果あり	あり	なし
医療機関匿名化	匿名化	匿名化	匿名化	匿名化	Facility Number
患者匿名化	匿名化	匿名化	匿名化	匿名化	SSN

NDBの研究利用の現状

Administrative Claims Database の比較

	市販レセプトDB	NDB		DPC Data	CMS data (米国)
		レセプト	特定健診情報		
利用方法	データを研究者に提供	・データを研究者に提供 ・アクセスポイントを提供	・データを研究者に提供 ・アクセスポイントを提供	・ 集計データのみを提供	・データを研究者に提供 ・VPNオンラインアクセスを提供
オプトアウト	可	不可	不可	不可	不可?
利用料金	有料	無料	無料	無料	有料
他DBとのリンク	不可 ※特定健診情報とは突合可	不可 ※特定健診情報とは限定的に突合可	不可 ※レセ情報とは限定的に突合可	不可	可
被保険者台帳	あり	なし	なし	なし	あり
傷病名の正確性担保	なし	なし	なし	なし	なし
データハンドリングサポート	あり	なし	なし	なし	あり

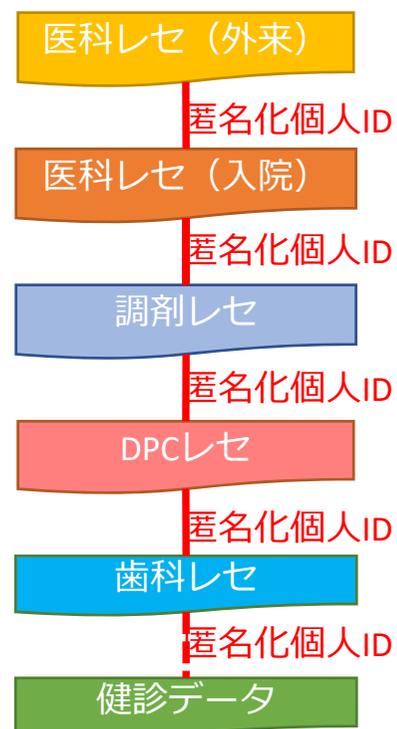
NDBデータの大まかな構造

レセプト部分と健診情報部分がある。

- 医科(MED)
 - DPC(DPC)
 - 調剤(PHA)
 - 歯科(DEN)
 - 特定健診データ
 - 保健指導データ
- レセプト情報
- 健診情報

NDBデータ の大まかな構造

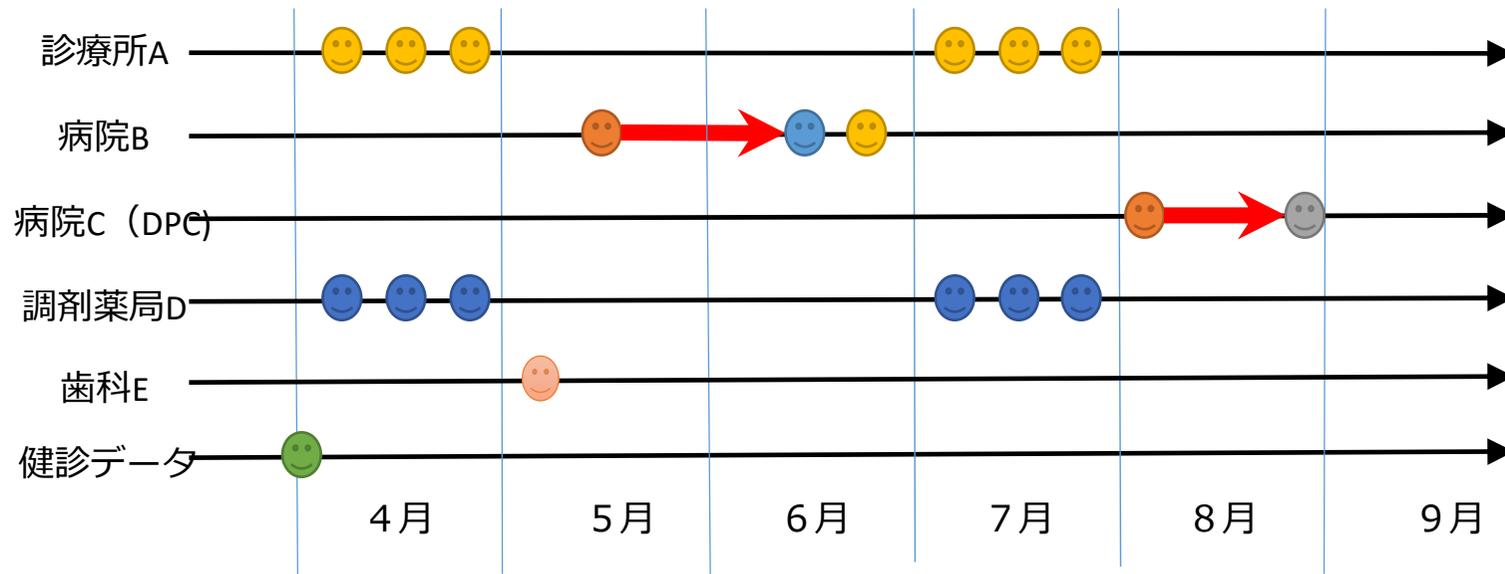
匿名化された個人IDで紐付けを行う事が出来る



- NDB には2つのID が存在する。
- 匿名化ID1
保険者番号、被保険者証 の記号・番号、生年月日、性別を基に作成
- 匿名化ID2
氏名、生年月日、性別を基に作成

NDBデータ の大まかな構造

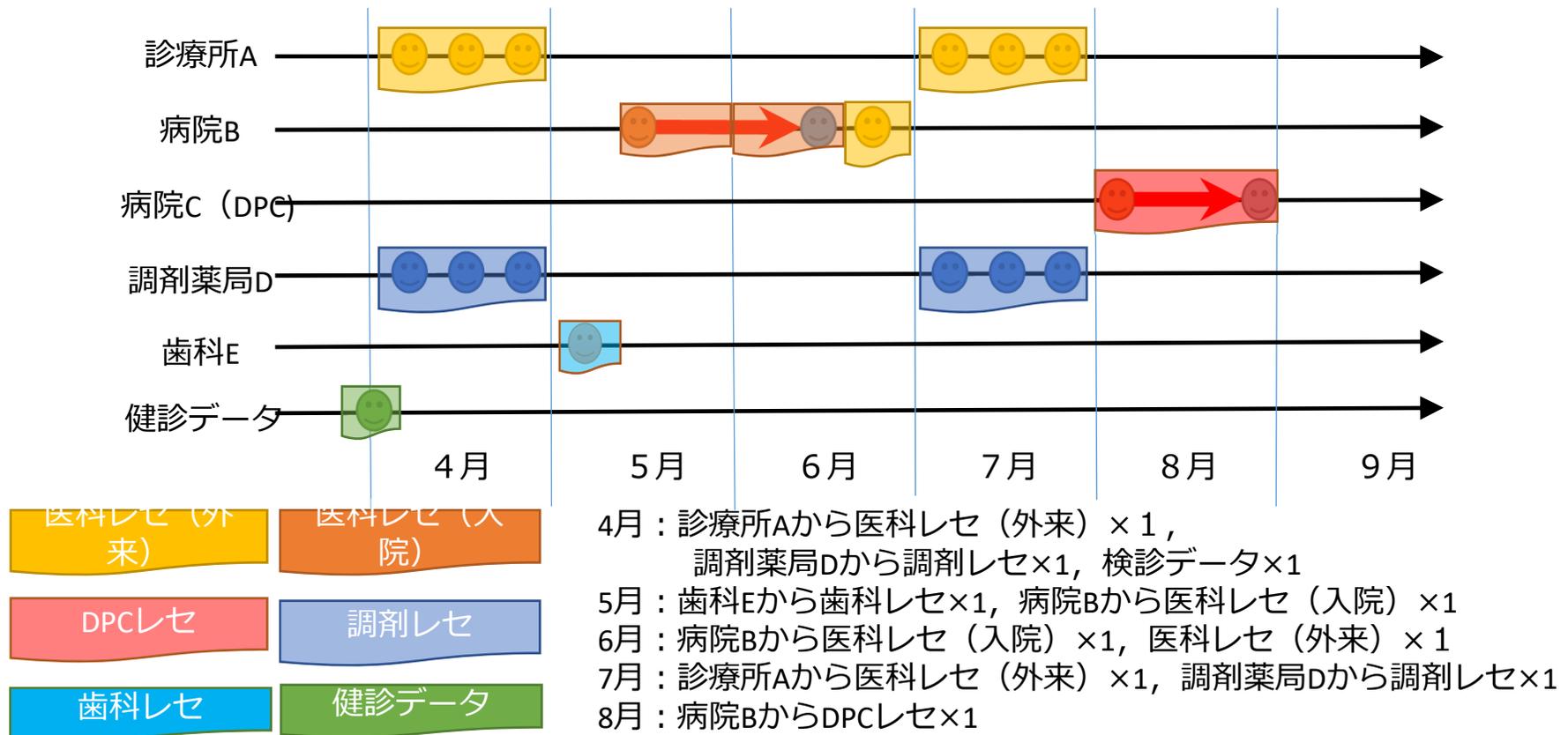
患者のエピソードで考えると以下の様になる。



- 😊 外来受診
- 😄 調剤薬局
- 😊 歯科受診
- 😓 入院開始
- 😁 生存退院
- 😇 死亡退院
- 😌 健診受診

4月：診療所Aに外来受診3回，同日調剤，健診受診
5月：歯科E受診，病院B入院
6月：病院B退院，同病院にて外来フォローアップ
7月：診療所Aに外来受診3回，同日調剤
8月：病院CにDPC入院，同月死亡退院

NDBデータ の大まかな構造

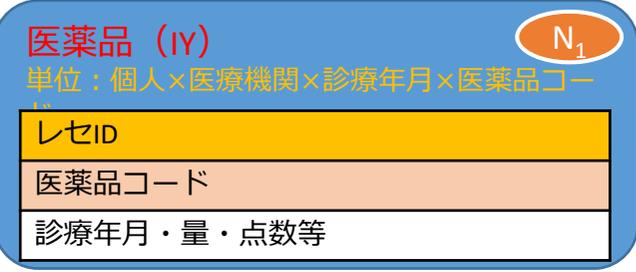


NDBデータの大まかな構造

- 複数のテーブル（右図）で1レセの情報となる。
- **匿名化個人ID** は各REに含まれる、**匿名化ID1**と**匿名化ID2**の2つ
- 他のレセプト（MED, DPC, PHA, DEN）情報（や健診情報）とは**匿名化個人ID**で接続可能
- 月単位で登録されるため、月またぎ入院や月内外来受診の処理が煩雑

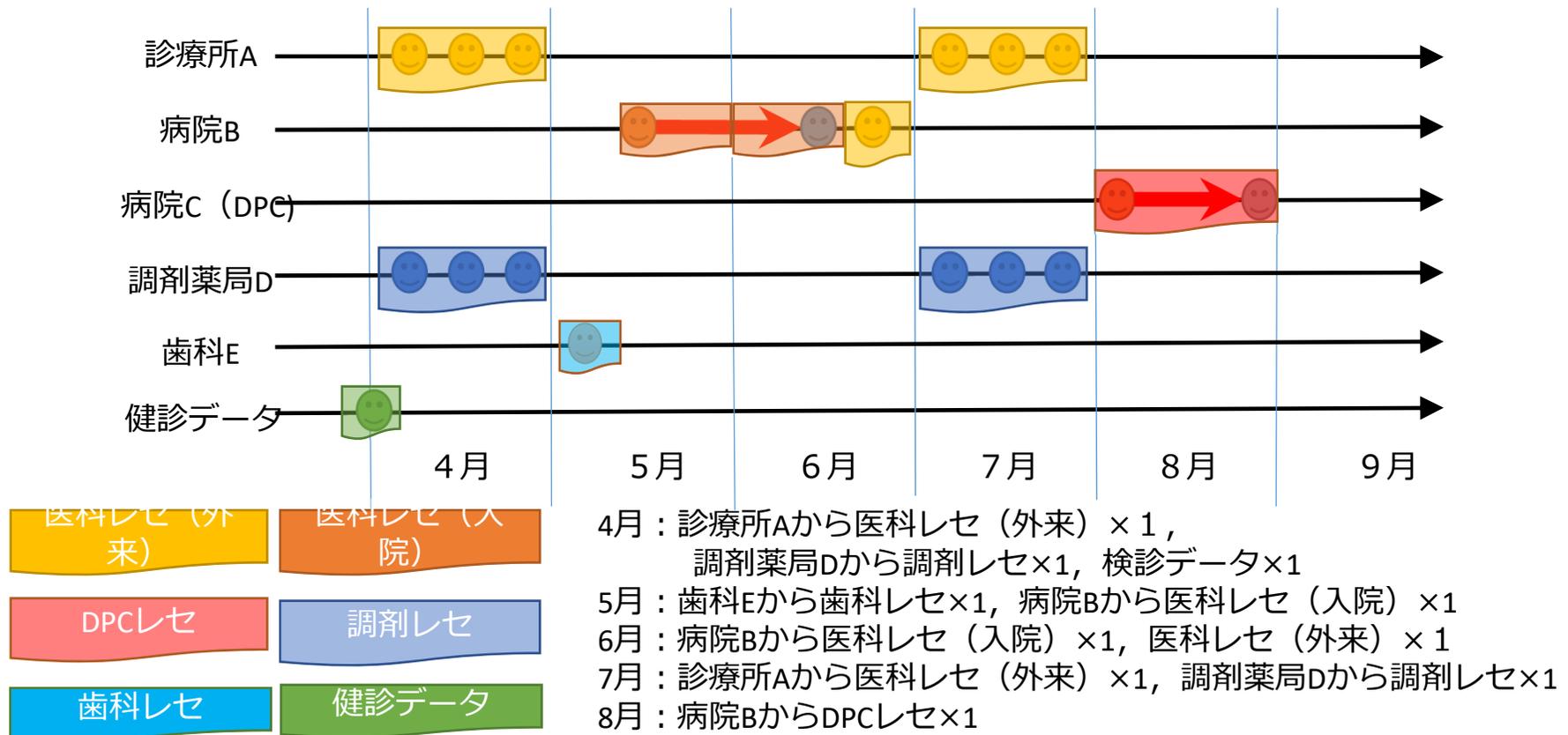


※DPCレセはREが複数個



⋮

NDBデータ の大まかな構造



NDBデータの構造上の問題

- 死亡個票などとの連結は不可能
 - 院内死亡ならある程度追える
 - 院外死亡はルールを決めないと厳しい。
- 他のデータベースとの連結は**禁止されている。**
- 社会保障番号が日本には存在しない。
 - 75歳のタイミングでIDが総入れ替え。
 - 名前などから作成された匿名化IDを元に情報の連結が必須。

ID構造と縦断データの作成

匿名化個人IDリストから、追跡可能な全ての匿名化個人IDリストを抽出する

匿名化ID1	匿名化ID2	ID3_A	ID3_B	ID3_C	ID3_D
A	1	B_1	A_1	C_2	A_1
B	1	B_1	B_2	C_2	B_1
B	2	C_2	B_2	C_2	B_2
C	2	C_2	C_2	C_2	C_2
D	3	D_3	D_3	D_3	D_3
E	4	E_4	E_4	E_4	E_4

ID3という個人IDを作成するとして、

ID3_AはID2を基準に作成（例：名前が変更されると追跡不能だが、保険者変更を追跡可能）

ID3_BはID1を基準に作成（例：保険者変更を追跡不能だが、名前の変更を追跡可能）

ID3_CはID1 or ID2 を基準に作成（例：保険者の変更・名前の変更を追跡可能だが、同姓同名問題がある。）

ID3_Dは両者を基準に作成（例：性別が同じ双子も判別可能だが、追跡率は落ちる。）

縦断ID作成の事例

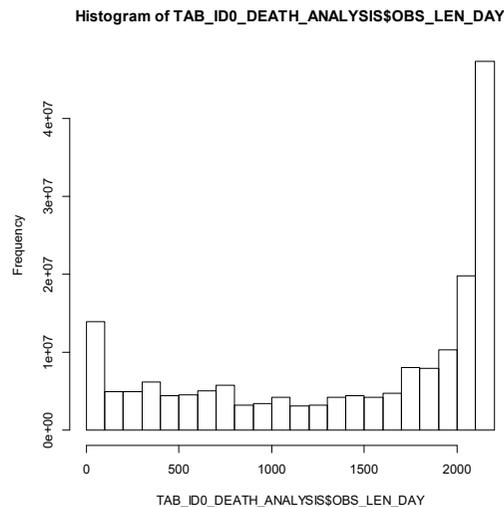
学術論文でOnline になったIDの作成法

National Database of Health Insurance Claims and Specific Health Checkups of Japan (NDB): Outline and Patient-Matching Technique

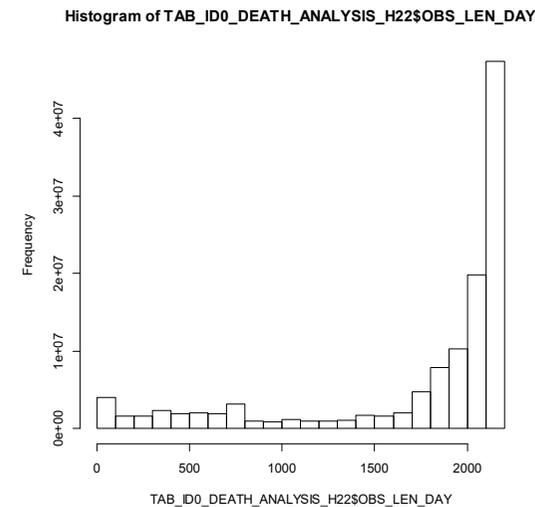
<https://www.biorxiv.org/content/early/2018/03/29/280008.article-metrics>

個人追跡率 (結果)

- H22.1-H27.12の ID0 (独自ID) の総数 : 1.78 億件
- H22.1-H22.12 の間に出現したID0 : 1.20 億件
- 初年度症例の追跡期間は平均 4.73 ± 1.7 年



全ID0



H22.1-H22.12 の間に出現したID0

死亡アウトカムの取得

NDB では一部症例において死亡転帰が取得できる。

- 医科SY, DPC BU, DPC SY, 歯科 RE, 歯科 HS に転帰区分が含まれる。
- コメントレコードには退院先（死亡）情報が含まれる。
- 今まで、NDB を用いて死亡転帰をどの程度補足できているか検証したデータはない。

死亡アウトカムの取得（処理方法）

処理方法

- H22.1-H27.12 の医科SY, DPC BU, DPC SY, 歯科 RE, 歯科 HS を対象とした。
- 転帰区分が“ 3 : 死亡”もしくはDPC 転帰区分が “ 6 : 死亡”、“ 7 : 外死亡”となっているレセプト番号（SEQ2_NO）を抽出した。
- SEQ2_NO に紐付くID1, ID2, 診療年月 を取得した。
- 取得したID1, ID2 にID0 を紐付けた。
- 複数回死亡レセが出現する場合（複数医療機関で死亡が記録される）は最小の診療年月を取得

転帰区分	ID1	ID2	PRAC_YM	ID0	YM_BGN	YM_END	AGE	SEX
3	A	a	HYYMM	****	HYYMM	HYYMM		
3	B	b	HYYMM	****	HYYMM	HYYMM		
6	C	c	HYYMM	****	HYYMM	HYYMM		

経年の男女別死亡者数（結果）

NDBから取得された死亡者数

年	男性	女性	総計（参考値比※）
2010	437538	391884	829422(69.29)
2011	461019	420630	881649(70.36)
2012	473550	435793	909343(72.38)
2013	482391	447899	930290(73.34)
2014	487002	452441	939443(73.8)
2015	495851	464303	960154(73.74)
総計	2837351	2612950	5450301(72.19)

※人口動態統計より

死亡アウトカムの妥当性（結果）

死亡レセの出現後もレセが出現する割合

$$137,374 / 5,450,301 = 2.5\%$$

以下の場合死亡レセの時期がずれる？

- 外来でフォローしていた患者が他院入院中に死亡
- 歯科で入れ歯作成中の患者が他院入院中に死亡

同一ID0で死亡が複数回

- ID突合ミス？
- 転帰情報記載ミス？

NDBデータの利用の方法

申請方法

- 厚労省窓口（NDB+申請で検索）

https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryuu/iryuu_hoken/reseputo/index.html

- 申請できる者

- ① 国の行政機関
- ② 都道府県
- ③ 研究開発独立行政法人（PMDA含む）
- ④ 大学（大学院含む）
- ⑤ 医療保険者の中央団体
- ⑥ 医療サービスの質の向上等をその設立目的の趣旨に含む国所管の公益法人
- ⑦ 提供されるデータを用いた研究の実施に要する費用の全部又は一部を国の行政機関や研究開発独立行政法人等から補助されている者等

NDB利用の手順

NDB の利用には有識者会議の審査が必要

- 有識者会議は年4回
- 開催時期の1か月前くらいに申請書類提出締め切りがある。
- その前に、厚労省担当に事前相談が求められている。
- 申請種別によって、倫理申請を事前に通す必要がある。

NDB の利用形式

NDB にはいくつかの利用形式があり必要な書類等が違う

- サンプルングデータセット
 - ある特定月の横断データデータからレセプトをランダムサンプルリングしたもの
 - 匿名性を担保するため、コードマスキングがかかっている。
 - 外来—調剤間の接続のみなされているが、患者の縦断化は不可能
 - 詳細な公表形式・抽出要件などを求められない。
 - **必要となるセキュリティ要件の緩和**
 - 倫理審査結果の提示は不要

NDB の利用形式

NDB にはいくつかの利用形式があり必要な書類等が違う

- 集計表の取得
 - 集計方法を指定
 - 集計を厚労省に依頼して集計結果のみを取得する。
 - 申請時に詳細な公表形式・抽出要件を指定する。
 - 必要となるセキュリティ要件の緩和（サンプリングデータセットと同程度）
 - 倫理審査結果の提示が必要

NDB の利用形式

NDB にはいくつかの利用形式があり必要な書類等が違う

- 特別抽出

- 特定の条件でレセプト情報・特定検診情報を抽出する。
- 抽出された情報を取得解析する。
- **必要となるセキュリティ要件が高い**
- 申請時に詳細な公表形式・抽出要件を指定する。
- 倫理審査結果の提示が必要

室内環境とルール

レセプト情報等オンサイトリサーチセンター設置場所

東京大学医学部教育研究棟13階

- 公衆衛生学、行動社会医学、臨床疫学・経済学、医療情報学の医学部4講座による合同管理

施設管理者：小林 廉毅

(健康医療政策学分野 教授)



実際の使用感について

BIツール（画面イメージ）

Business Intelligence ツール

汎用的な情報分析ツール

医科・DPC・歯科・調剤・特定健診の定型帳票が約40種類

The screenshot displays the Oracle Business Intelligence interface. On the left, there are four main categories: 【医科】 (Medical), 【DPC】 (DPC), 【歯科】 (Dental), and 【調剤】 (Pharmacy). Each category lists various data sources and reports. Below these are sections for 【特定健診・特定保健指導】 (Specific Health Examination/Specific Health Guidance) and 【レセプトビューワ】 (Receipt Viewer). The main area shows a detailed report for medical data, including a table with columns for '件数' (Number of Cases), '合計金額' (Total Amount), '診療日数' (Number of Treatment Days), and '平均' (Average). The table data is as follows:

診療種別	件数	合計金額	診療日数	平均	診療種別	件数	合計金額	診療日数	平均
合計	21	104,940,421	1,696	18	2,471,288	1,604	2,246,913,356	26	177,777,777
000	1	15	99	1	15	99			
100	1	17	99	1	17	99			
100	1	18	99	1	18	99			
100	1	1234,567	99	1	1234,567	99			
100	1	16	99	1	16	99			
700	1	12,345,678	13						
000	1	999	99	1	999	99			
000	1	10	10	1	10	10			
10-1000	2	2	198	2	2	198			
10-1000	1	1	99	1	1	99			
20-2000	2	1,235,566	198	2	1,235,566	198			
25-2000	1	10	10	1	10	10			
30-3000	1	1	99	1	1	99			
35-3500	1	12,345,678	13						

On the right, there is a bar chart titled '月別集計グラフ (件数)' (Monthly Summary Graph (Number of Cases)) showing a single bar with a value of approximately 20. The interface also includes various filters and navigation options.

実際の使用感について－BIツール

検索画面イメージ

BIツールの標準帳票例

The screenshot shows a search interface for a BI tool. It features a grid of search criteria, each with a dropdown menu for selection. The criteria include:

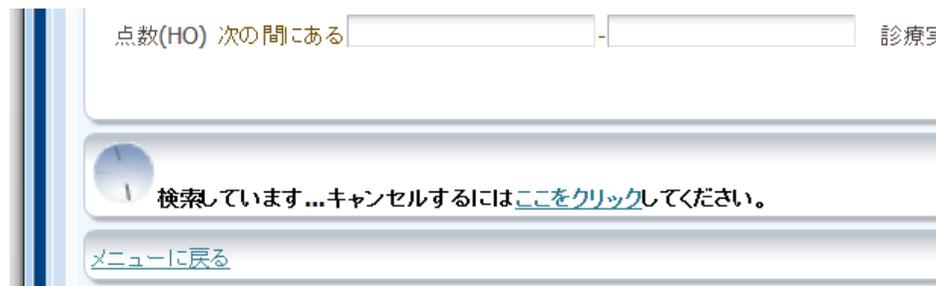
- 請求年月(IR) 次の間にある (Request Date)
- 都道府県(IR) (Prefecture)
- 診療科(IR) (Department)
- 病診区分(IR) (Institution Type)
- * 診療年月(RE) 2015年10月 (Treatment Date)
- 男女区分(RE) (Gender)
- 年齢階層(実)(RE) (Age Group)
- 入院区分(RE) (Admission Type)
- 法別番号(HO) (Legal Number)
- ICD10-1コード(SB) 次で始まる (ICD10-1 Code)
- ICD10-2コード(SB) 次で始まる (ICD10-2 Code)
- 傷病(SB) (Injury/Disease)
- 疑い病名(SB) (Suspected Disease Name)
- 主傷病(システム設定)(SB) (Main Injury/Disease - System Setting)
- 主傷病(医療機関設定)(SY) (Main Injury/Disease - Medical Institution Setting)
- 診療識別(SI) (Treatment Identification)
- コード表用番号(SI) 次で始まる (Code Table Number)
- 診療行為(SI) (Treatment Act)
- 診療識別(IY) (Treatment Identification)
- 薬効分類コード(IY) 次で始まる (Drug Efficacy Classification Code)
- 医薬品(IY) (Medicine)
- 検発品(IY) (Inspected Item)
- 剤形(IY) (Formulation)
- 診療識別(TO) (Treatment Identification)
- 特定器材(TO) (Specific Equipment)
- 点数(HO) 次の間にある (Points)
- 診療実日数(HO) 次の間にある (Actual Treatment Days)

At the bottom of the screen, there is a footer with the following text:

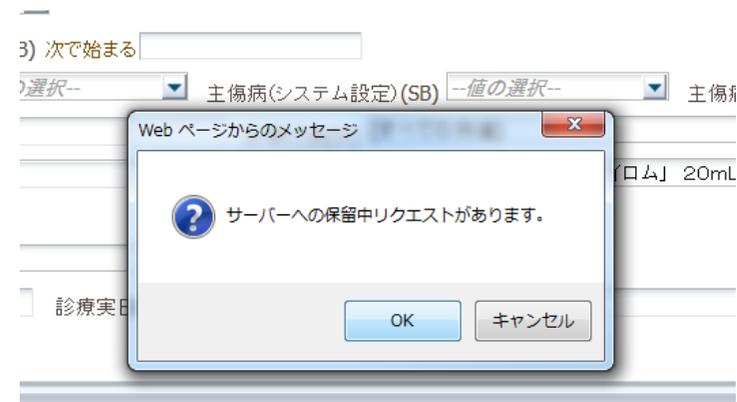
[DPC] 年齢階層別・男女区分別集計
実行時刻: 2016/02/01 16:36:15

レスポンスについて

請求年月、診療年月、診療行為等の抽出が可能・ただし、複数条件を指定した場合の処理時間は延長した。
(詳細はOracleR使用感参照)



<検索作業中のステータス>



<作業途中で検索条件の変更は不可>

Oracle BI の使いどころ

単純なレセプト数カウントなどを行うツール

- 研究者むけのツールではない。
 - 複雑な集計や時系列を追う集計は困難
 - 検索条件を極めて単純にした集計には利用可能
- どちらかといえば、政策担当者向けのツール
 - 単純なレセプトの発生件数を調べることは可能
 - 帳票を自分で作成する自由分析ツールの使用して患者数を調べることも可能
- ただし、適切な使用方法をしないと数値を読み間違えるため、十分なマニュアル整備が必要。

ORE を用いた抽出・集計

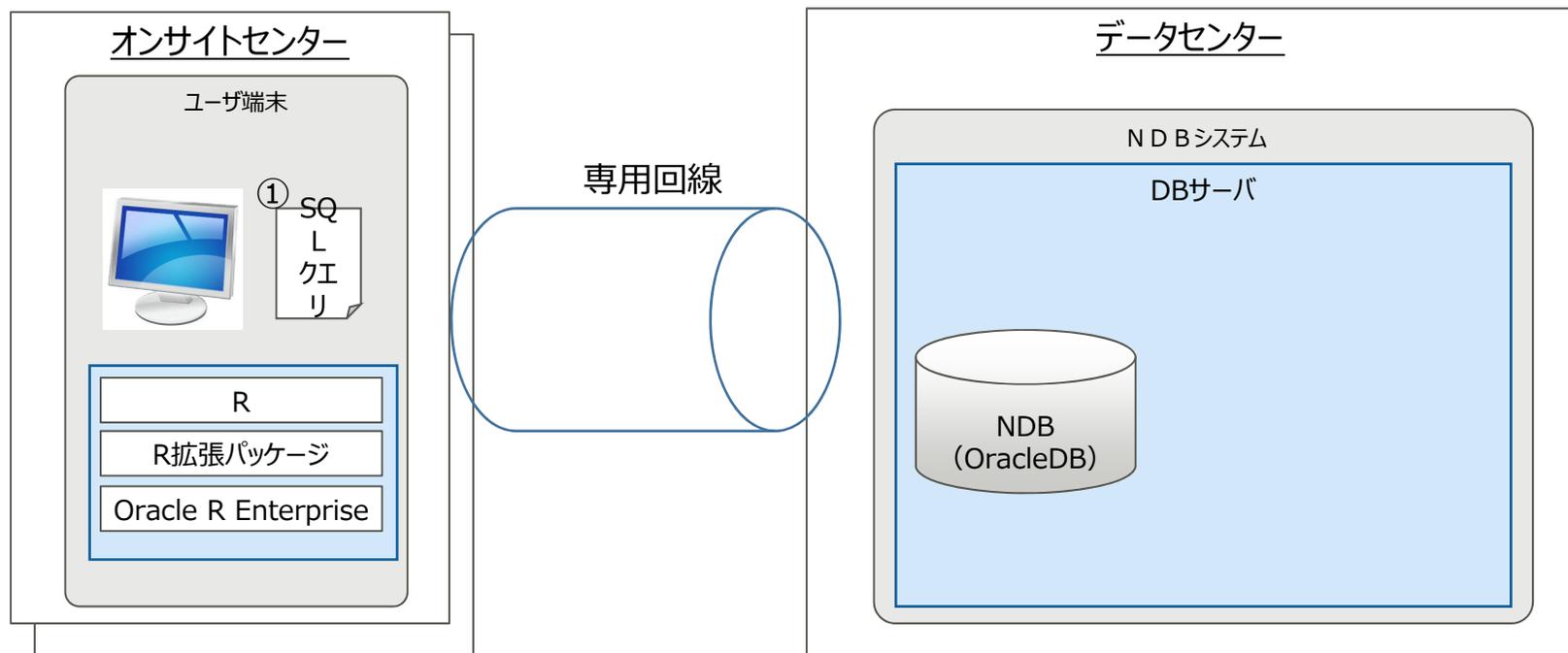
Oracle R Enterprise

- Oracle R Enterprise (以下ORE)はOracle 社が提供している、オープン・ソースの統計プログラミング言語であるRとその環境をエンタープライズ対応およびビッグ・データ対応にする機能を備えたソフトウェアである。

テスト行った処理

- ① オンサイトセンター内にて、SQL のクエリ（データ問い合わせプログラム）を作成
- ② ORE を介してクエリを実行
- ③ ORE を介してデータをオンサイトセンター内に移動

オンサイトリサーチセンターでのデータ処理



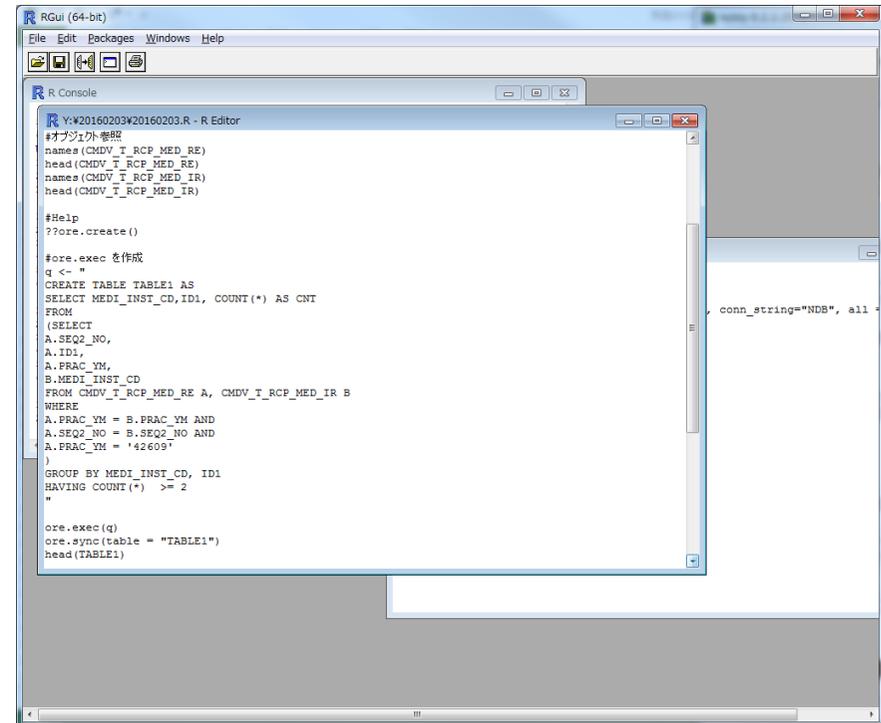
- ・ オンサイトセンター内にて、SQL のクエリを作成した。

ORE (SQL クエリを書く)

シンタックスハイライト可能なエディタを導入

下記の様なSQL文（データの問い合わせプログラム）を組み合わせてながら必要となるデータ形式へデータを抽出・成形するためのクエリを書く

SELECT <カラム名>
FROM <テーブル名>
WHERE <条件>
GROUP BY <集計単位>



実際の開発画面

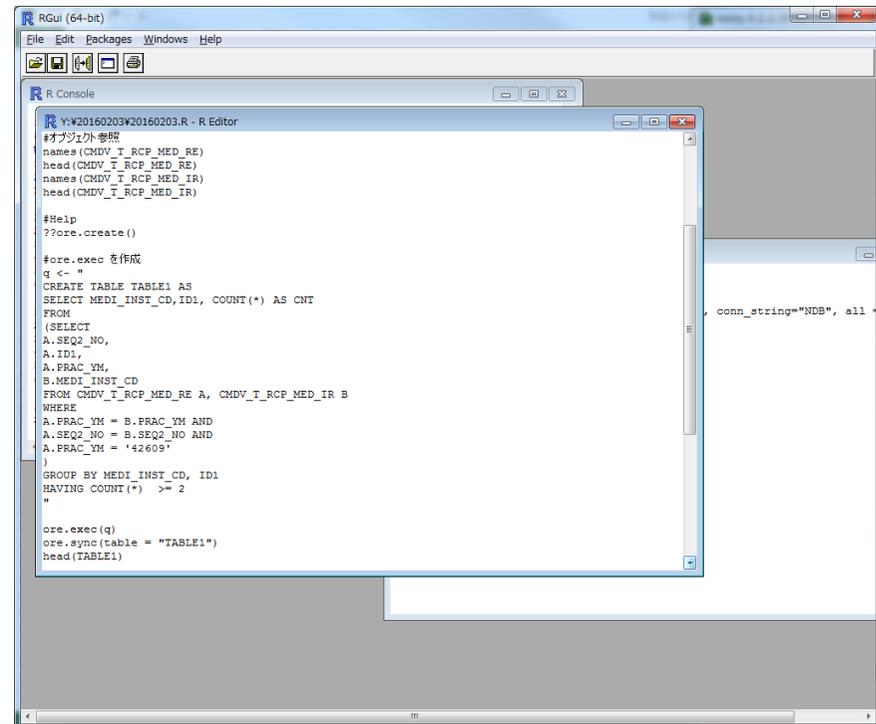
ORE (SQL クエリを書く)

エラーがマスクされるのでバグフィックスが難しかった。

オブジェクト参照機能等がないため、開発に困難が伴う。

実行時エラーなどがマスクされており、デバッグが困難。

→後述のSQL Plus を利用した開発に移行した。



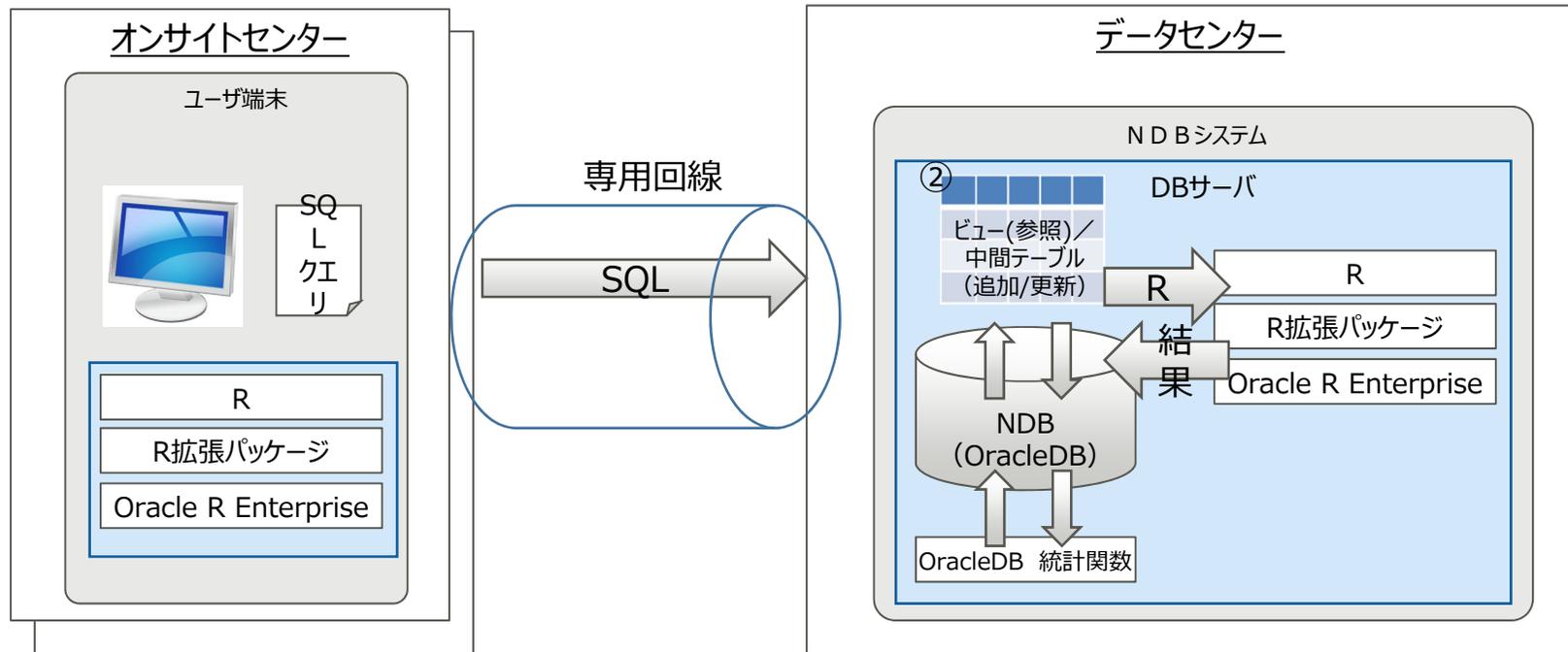
```
Y:\20160203\20160203.R - R Editor
#オブジェクト参照
names(CMDV_T_RCP_MED_RE)
head(CMDV_T_RCP_MED_RE)
names(CMDV_T_RCP_MED_IR)
head(CMDV_T_RCP_MED_IR)

#Help
??ore.create()

#ore.exec を作成
q <- "
CREATE TABLE TABLE1 AS
SELECT MEDI_INST_CD, ID1, COUNT(*) AS CNT
FROM
(SELECT
A.SEQ2_NO,
A.ID1,
A.FRAC_YM,
B.MEDI_INST_CD
FROM CMDV_T_RCP_MED_RE A, CMDV_T_RCP_MED_IR B
WHERE
A.FRAC_YM = B.FRAC_YM AND
A.SEQ2_NO = B.SEQ2_NO AND
A.FRAC_YM = '42609'
)
GROUP BY MEDI_INST_CD, ID1
HAVING COUNT(*) >= 2
"
ore.exec(q)
ore.sync(table = "TABLE1")
head(TABLE1)
```

実際の開発画面

ORE (SQL クエリを実行)



・ ORE を介してクエリを実行した。

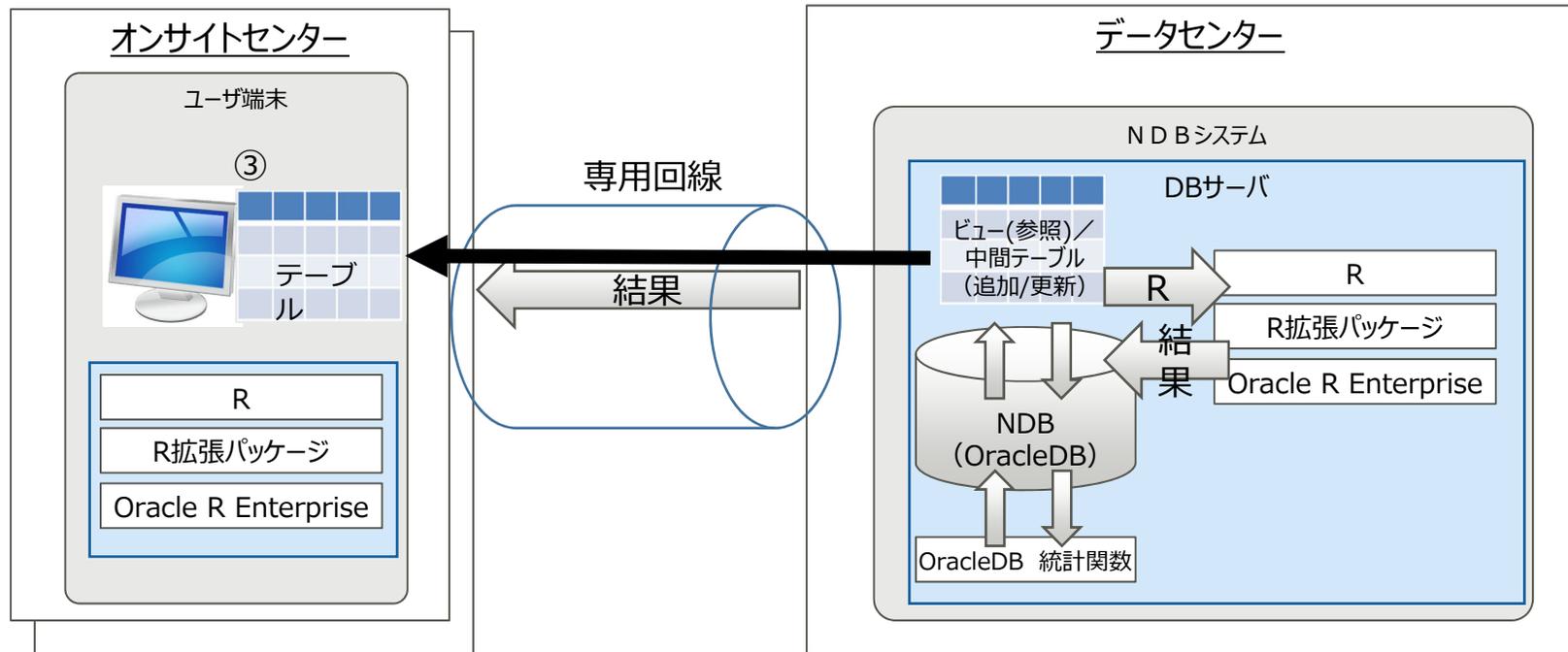
ORE (SQL クエリを実行)

処理	クエリ詳細	処理時間(秒)		
		C34\$ (400名)	I63\$ (50000名)	I50\$ (90000名)
処理 1	H26.10 のDPC SBレコードを対象にICD10 コードがC34\$, I63\$, I50\$ のSEQ2_NO (レセプトID) を抽出する。	0.31	6.77	6.81
処理 2	H26.10 のDPC REレコードを対象に処理 1 で抽出した SEQ2_NO に紐づくID1 (患者ID) を抽出する。	0.41	4.05	2.12
処理 3	H26.10~H27.03 の期間のDPC レセREレコードから、処理 2 で抽出した ID1 に紐づくSEQ2_NOを抽出する。	13.45	13.81	13.67
処理 4	H26.10~H27.03 の期間の医科 レセREレコードから、処理 2 で抽出した ID1 に紐づくSEQ2_NOを抽出する。	26.6	30.94	39.07
処理 5	H26.10~H27.03 の期間のDPC レセSBレコードから、処理 3 で抽出した SEQ2_NO に紐づくレコードを抽出する (Table_SB)。	5.23	17.06	15.36
処理 6	H26.10~H27.03 の期間の医科レセSYレコードから、処理 4 で抽出した SEQ2_NO に紐づくレコードを抽出する (Table_SY)。	79.38	107.74	129.12

- ・クエリが実行可能であることが確認できた。
- ・今回実行したクエリに対するレスポンスは十分であった。

C34\$:肺癌, I63\$:脳梗塞, I50\$:心不全

ORE (データをセンター内にDL)



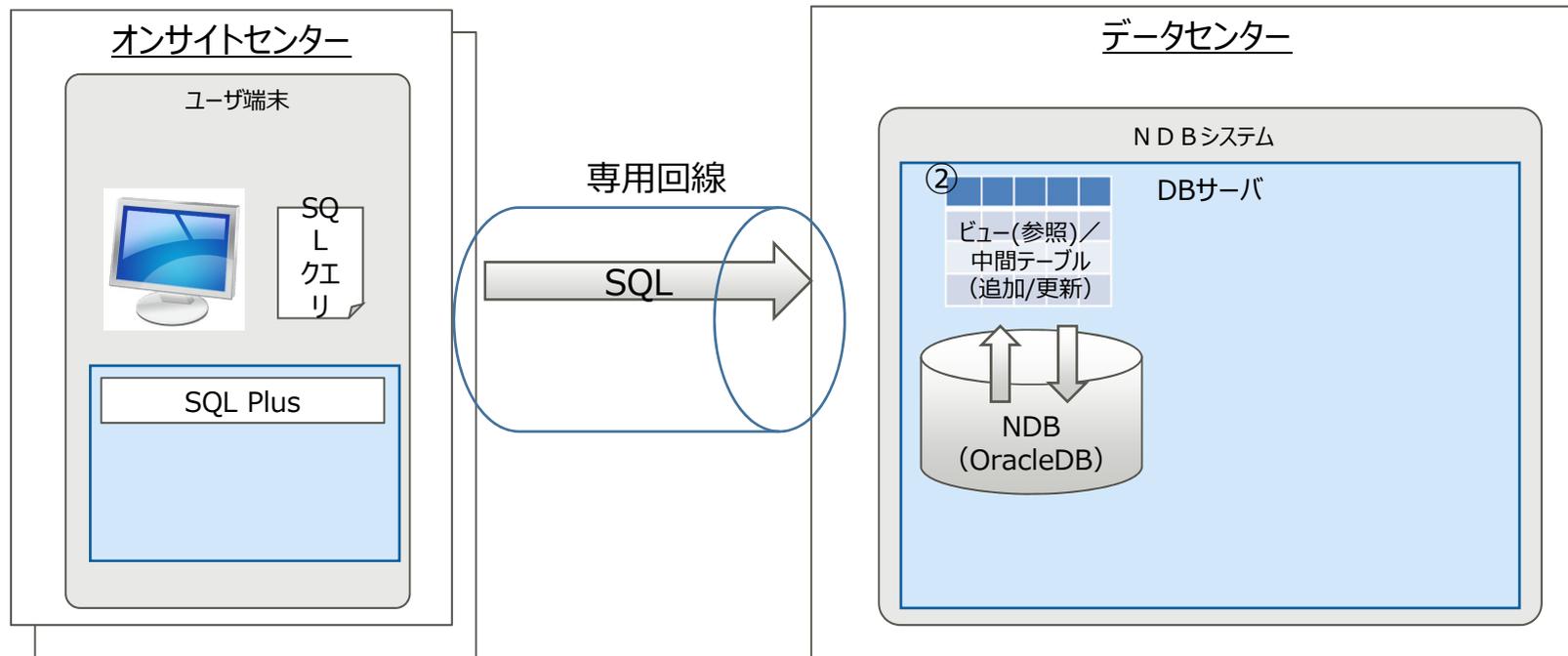
・ ORE を介して得たデータをオンサイトセンター内に移動した。

ORE (データをセンター内にDL)

	Table_SB (19カラム 126桁)			Table_SY (12カラム 128桁)		
	C34\$	I63\$	I50\$	C34\$	I63\$	I50\$
概算行数 (行)	16,000	1,000,000	2,300,000	35,000	3,000,000	7,500,000
ダウンロード時間(秒)	0.64	26.35	71.4	1.50	109.2	247.2

- ・ NDB のサーバーからローカル環境 (オンサイトセンター内) へのデータのダウンロードが可能であった。
- ・ ローカル環境の解析システム (SAS, R等) での解析が可能であることが分かった。
- ・ データサイズが大きくなると、ダウンロードに時間がかかるため、課題が残る。

SQL クエリを直接実行する。

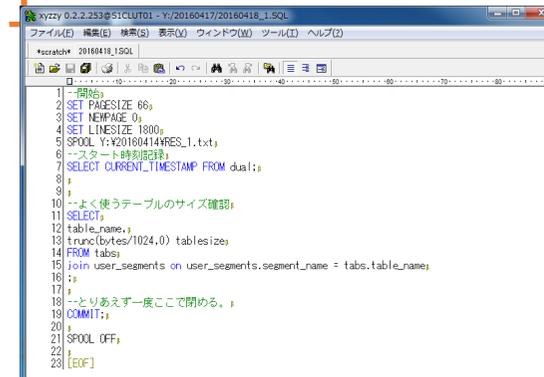


・ SQL Plusを介してクエリを実行した。

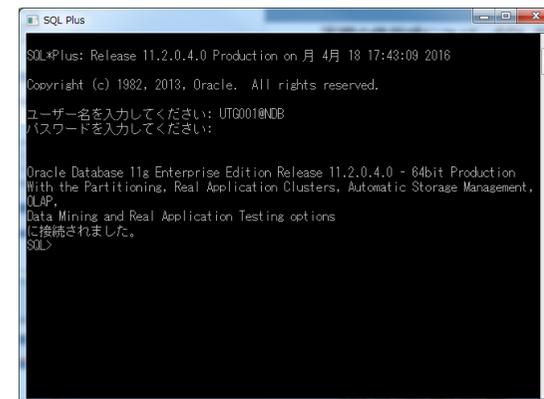
SQL クエリを書く

エディタで書いてSQL Plusで実行

- シンタックスハイライトのつくテキストエディタを用いてクエリを書く。
- SQL Plus にて実行
- 結果をCommit すれば Oracle R Enterprise から参照可能
- エラーの内容が表示される。



```
xyzzy 0.2.2.253@SICLUT01 - Y:/20160417/20160418_1_SQL
*sqlplus* 20160418_1.SQL
1 --開始
2 SET PAGESIZE 66;
3 SET NEWPAGE 0;
4 SET LINESIZE 1800;
5 SPOOL Y:\20160414\RES_1.txt;
6 --スタート時刻記録
7 SELECT CURRENT_TIMESTAMP FROM dual;
8 ;
9 ;
10 --よく使うテーブルのサイズ確認
11 SELECT
12 table_name,
13 trunc(bytes/1024,0) tablesize
14 FROM tabs;
15 join user_segments on user_segments.segment_name = tabs.table_name;
16 ;
17 ;
18 --とりあえず一度ここで閉める。
19 COMMIT;
20 ;
21 SPOOL OFF;
22 ;
23 [EOF]
```



```
SQL*Plus
SQL*Plus: Release 11.2.0.4.0 Production on 月 4月 18 17:43:09 2016
Copyright (c) 1982, 2013, Oracle. All rights reserved.
ユーザー名を入力してください: UT0001@NDB
パスワードを入力してください:
Oracle Database 11g Enterprise Edition Release 11.2.0.4.0 - 64bit Production
With the Partitioning, Real Application Clusters, Automatic Storage Management,
OLAP,
Data Mining and Real Application Testing options
已接続されました。
SQL>
```

テーブルのサイズ

各テーブルの行数をカウント

RE テーブル	行数 (百万行)
医科	6,152
DPC	92
歯科	675
調剤	3,929

IY テーブル	行数 (百万行)
医科	10,122
DPC	530
歯科	208
調剤	15,909

SY テーブル	行数 (百万行)
医科	32,003
DPC (SY/BU)	81/76
歯科 (HS)	1,330

SI テーブル	行数 (百万行)
医科	61,316
DPC	1,206
歯科 (SI/SS)	35/5,705

RE:レセプト共通, SY:傷病名, IY: 医薬品
SI:診療行為

NDB のデータ特性

テーブルのサイズ

各テーブルの行数をカウント

TO テーブル	行数 (百万行)
医科	452
DPC	117
歯科	111
調剤	34

TO:特定器材

データのサイズと抽出に必要な時間

ランダムに医科レセプト (MED) から患者を選んでデータを抽出のに必要な時間とデータサイズ

MED(人)	RE	SY	IY	SI	TO	合計
686 (1:00)	4096 (3:44)	2048 (3:44)	1024 (2:09)	5120 (10:1)	64 (00:08)	12MB (20.7min)
5213 (1:02)	28672 (3:47)	14336 (3:45)	8192 (2:12)	34816 (10:48)	320 (00:09)	84MB (21.7min)
22803 (1:03)	131072 (3:54)	64512 (3:59)	34816 (2:23)	155648 (11:15)	2048 (00:09)	379MB (22.7min)
178106 (1:12)	999424 (4:05)	491520 (5:49)	262144 (3:16)	1179648 (17:28)	9216 (00:15)	2873MB (32.1min)

RE:レセプト共通, SY:傷病名, IY: 医薬品, SI:診療行為, TO:特定器材 ※単位はKB(min:sec)

東大オンサイトセンターの運用

- 厚労省の許可を受け「東大の所属員」を対象に利用を開始
- 現在も利用者を少しずつ増やしながら事例蓄積
- 学外を含めた新規利用者の追加に向けた運用管理規定の見直し

旧	新
1) 施設利用者は、東京大学常勤教員のうち、オンサイトの利用について厚生労働省保険局が了承し、厚生労働省保険局と利用規約に基づく利用契約を結んだ者をいう。	1) 施設利用者は、東京大学常勤教員及び、 施設管理者が認めた者 のうち、オンサイトの利用について厚生労働省保険局が了承し、厚生労働省保険局と利用規約に基づく利用契約を結んだ者をいう。

NDBデータから研究論文を出すまで

研究者のスキルとニーズがずれている。



- NDBで実行可能な研究計画を建てる
- 研究計画に必要なデータを抽出する
- 研究に必要な形式にデータを成型する
- 基準に従いデータをクリーニング
- データの集計と解析
- 論文執筆

研究者はデータを抽出整形する作業のプライオリティを下げている。
その部分のスキルを十分に有する人材が少ない。

データベースを取り扱うことのできるエンジニアや疫学者との協働が必須

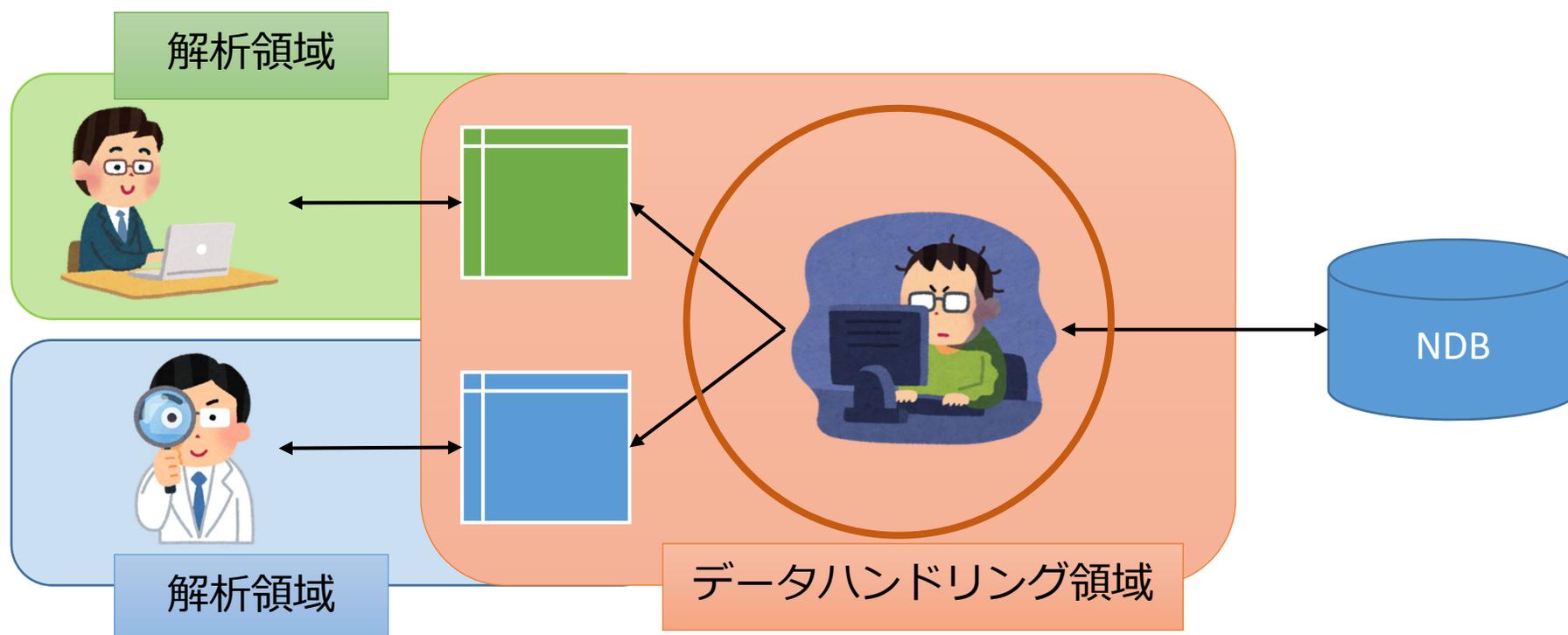
オンサイトセンターにおけるアドバイザーの必要性

オンサイトセンターではこれまでのハンドリングを研究者が行う

- 研究者に高度な技術と知識が要求される。
- しばしば、研究者が誤ったハンドリングを行いエラーを報告している事例がある。
- 東大オンサイトでは学内関係者に向けてデータハンドリングサポートを行っている（学外者に向けては未定）。

東大における運用形態

コンソーシアムを構築し、運用を行うためデータハンドリングプログラムを開発。



レセプトデータ 抽出・整形アプリケーション

アプリケーションの目的

- 研究者が利用しやすい形にレセプトデータを結合・整形する。

研究者の利用しやすいデータ ≈ 分析が行いやすい形式

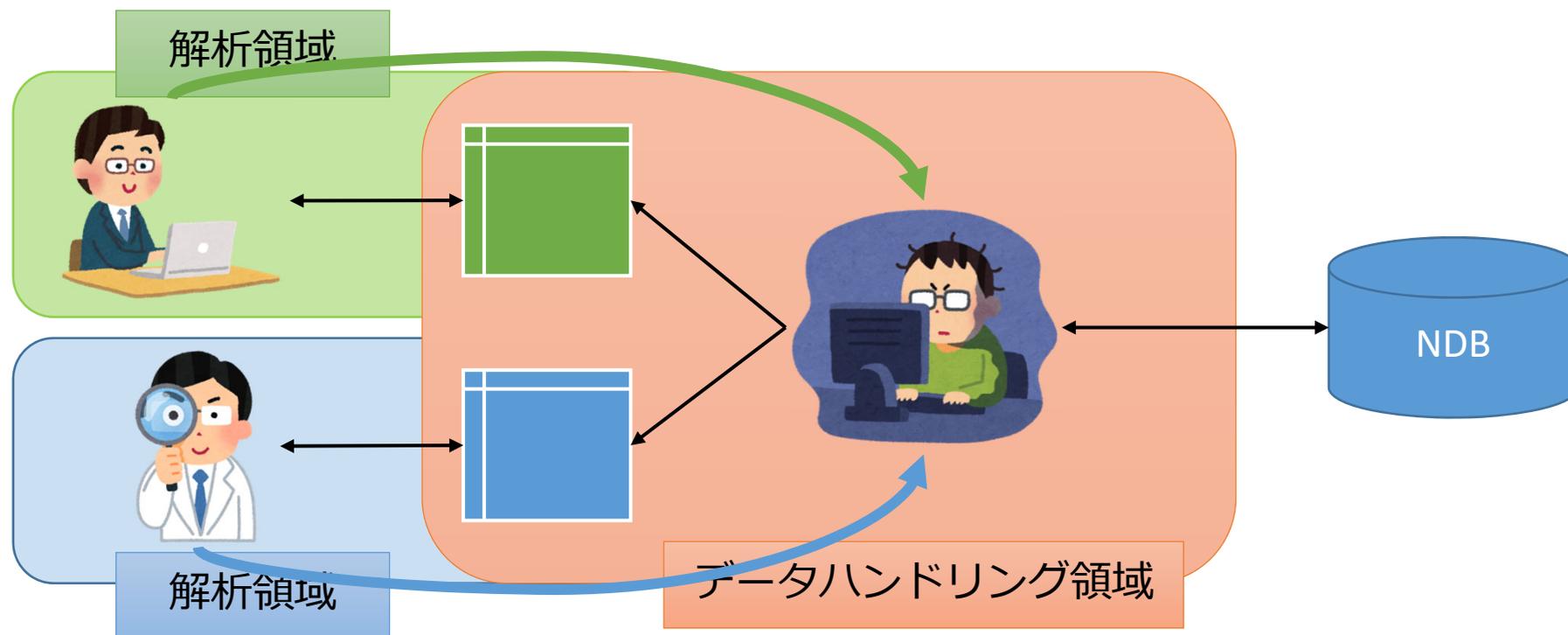
研究者が利用しやすいデータの形式

仕様策定時に決めたこと

- 過去のDPCデータを用いた研究実績をもとに下記を設定
- レセプトから研究者が研究に用いる変数は何か？
 - 年齢性別などの患者背景情報
 - 各観察単位での疾病情報
 - 各観察単位での処方・処置
- 研究者が研究に用いる観察単位は何か？
 - レセプト単位
 - エピソード単位
 - 実施日単位

運用形態

各研究班からデータハンドリングを依頼



現状の課題

- ハンドリング機能を提供せずにオンサイトを研究者に提供できるのか？
 - エラーの頻度増加
 - データベース信用の棄損
 - 研究進捗の遅れ
- データハンドリング機能を提供するマンパワーや行政からのサポートがあるのか？
 - ない。

種々の課題への対応

- 臨床疫学会と協力した教育プログラムの実施
 - 厚生労働省科学研究費（康永班）の援助を受けサマーセミナーを開催
 - http://www.clinicalepi.org/seminar/s_190805.html
- NDB ユーザー会
 - ユーザーの情報交換を促進
 - <http://square.umin.ac.jp/ndb/index.html>

まとめ

- NDBに含まれる情報（データ構造）を学んだ。
- データ利用が可能な形にデータを事前にハンドリングする必要がある。
- オンサイトセンターではそれらのプロセスを研究者が行う必要があるため、研究者間の協働が求められる。