

平成28年6月29日	資料1-2
第31回レセプト情報等の 提供に関する有識者会議	

# レセプト情報等オンサイトリサーチ センターの試行的利用に関する 中間報告

20160629

松居宏樹<sup>1</sup>, 佐藤大介<sup>2</sup>

1:東京大学大学院公共健康医学専攻臨床疫学・経済学分野

2:東京大学医学部附属病院 企画情報運営部

# 本日も話します内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

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1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)



# 本日も話します内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

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 menu categories: 【医科】 (Medical), 【DPC】 (DPC), 【歯科】 (Dental), and 【調剤】 (Pharmacy). Below these are sections for 【特定健診・特定保健指導】 (Specific Health Examination/Special Health Guidance) and 【レセプトビューワ】 (Receipt Viewer). The main area shows a detailed report titled '【調剤】特定健診報告書 男 各区分別集計' (Pharmacy Specific Health Examination Report Male by District Summary). The report includes a table with columns for '区分' (District), '件数' (Number of Cases), '合計金額' (Total Amount), and '診療日数' (Number of Treatment Days). A bar chart on the right shows '月別集計グラフ (件数)' (Monthly Summary Graph (Number of Cases)).

区分	件数	合計金額	診療日数
合計	21	104,940,421	1,696
08	1	15	99
10	1	12	99
18	1	18	99
08	1	16	99
78	1	12,345,678	13
08	1	999	99
08	1	10	10
10-108	2	2	198
10-198	1	1	99
20-208	2	2,235,566	198
25-298	1	10	10
30-308	1	1	99
35-398	1	12,345,678	13

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

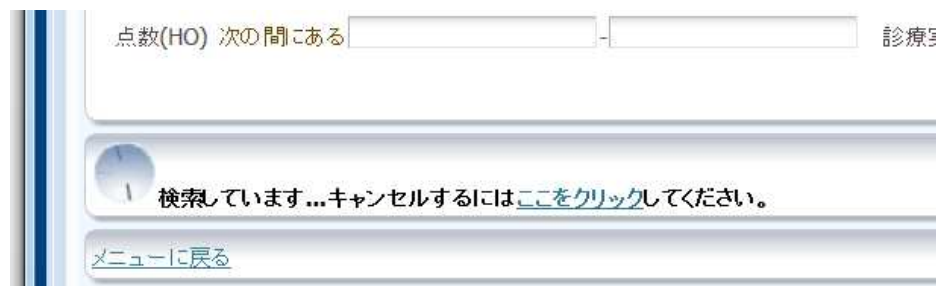
The screenshot shows a search interface for a BI tool. It features a grid of search criteria, each with a dropdown menu and a search icon. The criteria include:

- 請求年月(IR) 次の間にある
- 制道行県(IR)
- 診療科(IR)
- 病区分(IR)
- \* 診療年月(RE) 2015年10月
- 男女区分(RE)
- 年齢階層(実)(RE)
- 入院区分(RE)
- フタフタ(実)(RE)
- 法別番号(HO)
- ICD10-1コード(SB) 次で始まる
- ICD10-2コード(SB) 次で始まる
- 傷病(SB)
- 病名(SB)
- 主傷病(システム設定)(SB)
- 主傷病(医療機関設定)(SY)
- 診療識別(SI) 次で始まる
- コード表用番号(SI) 次で始まる
- 診療行為(SI)
- 診療識別(IY) 次で始まる
- 薬物分類コード(IY) 次で始まる
- 医薬品(IY)
- 検査品(IY)
- 剤形(IY)
- 診療識別(TO)
- 特定器材(TO)
- 点数(HO) 次の間にある
- 診療実日数(HO) 次の間にある

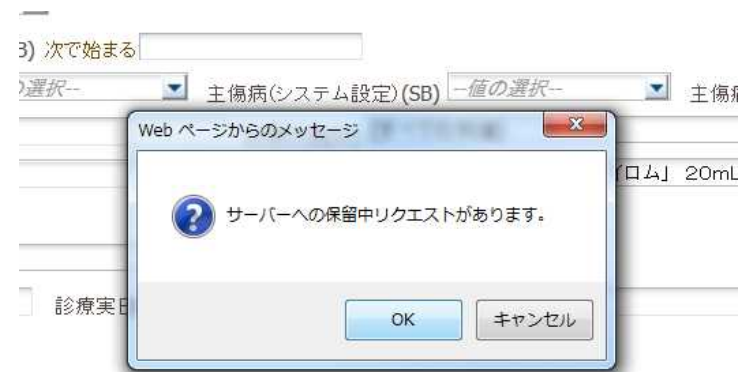
At the bottom of the screen, there is a summary line: [DPC] 年齢階層別・男女区分別集計, 実行時刻: 2016/02/01 16:36:15.

# 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。



# Oracle BI の注意点

利用者の認識と齟齬が生じるので、解決を希望する。

- テーブル間を結合した上で集計する場合
- 本来1件しかないレセプトをn件と返す処理がなされている。
- 技術的解決手法はすでに富士通より厚労省に報告済み。

○連結するテーブルで、それぞれ同一レセプトで複数レコードが存在する場合



富士通提供資料

# 本日も話します内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

# ORE を用いた抽出・集計

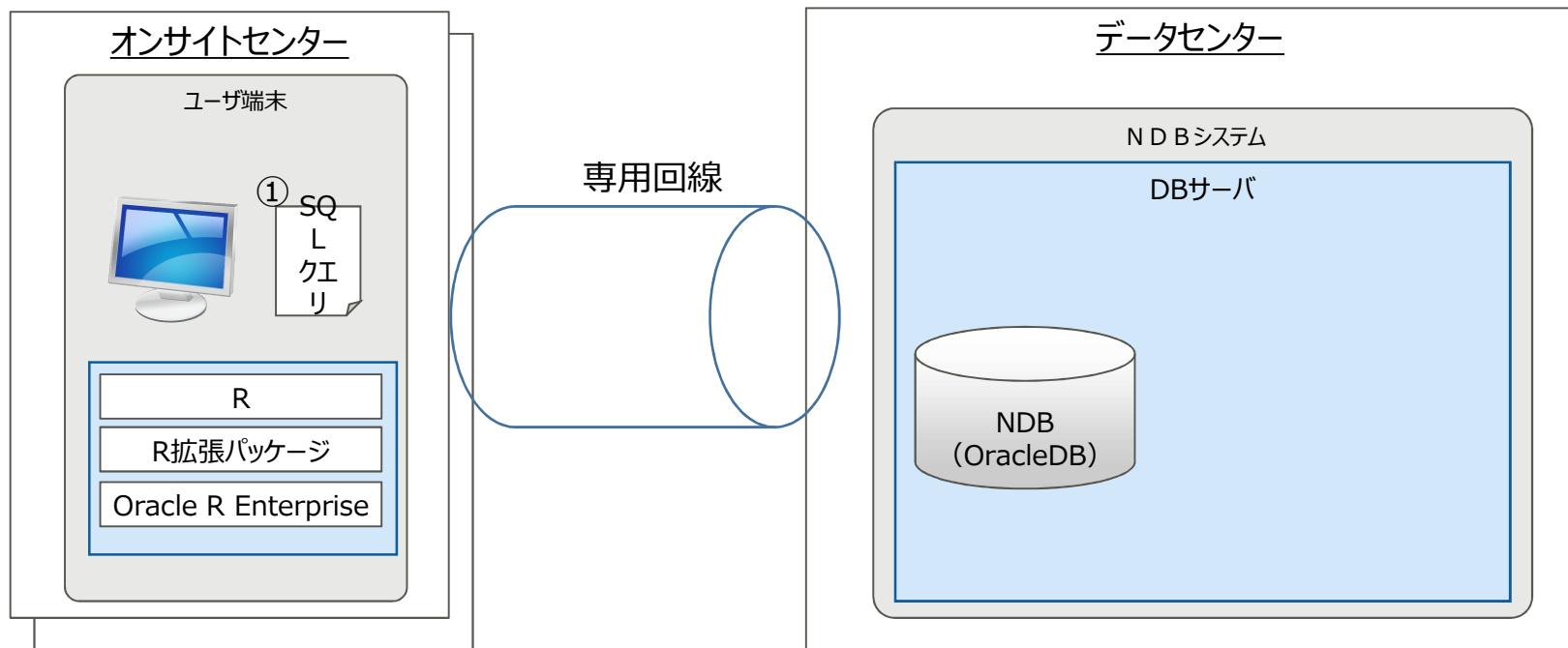
## Oracle R Enterprise

- 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



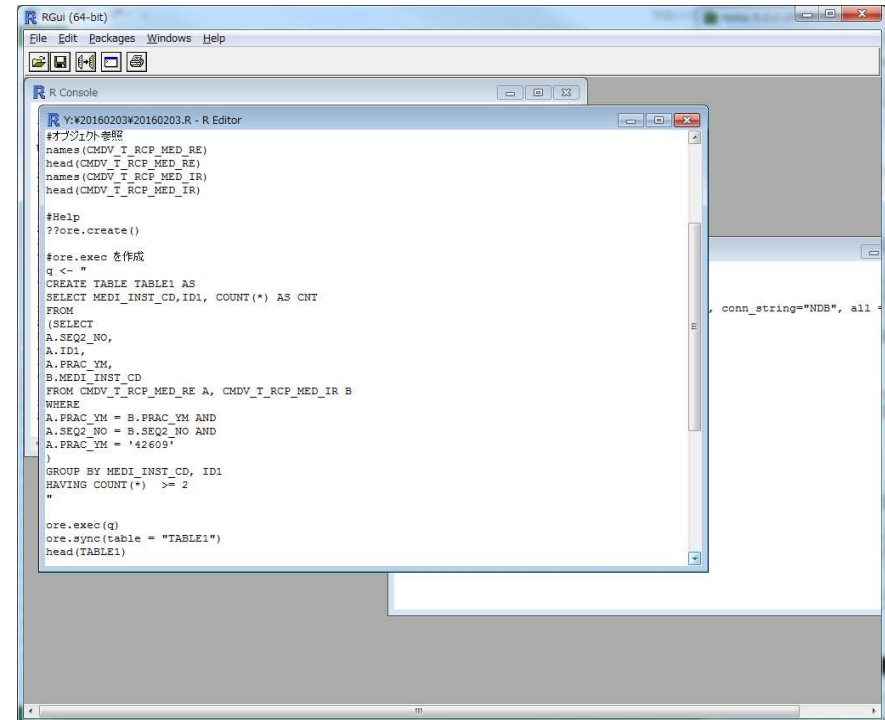
- ・ オンサイトセンター内にて、SQLのクエリを作成した。

# 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

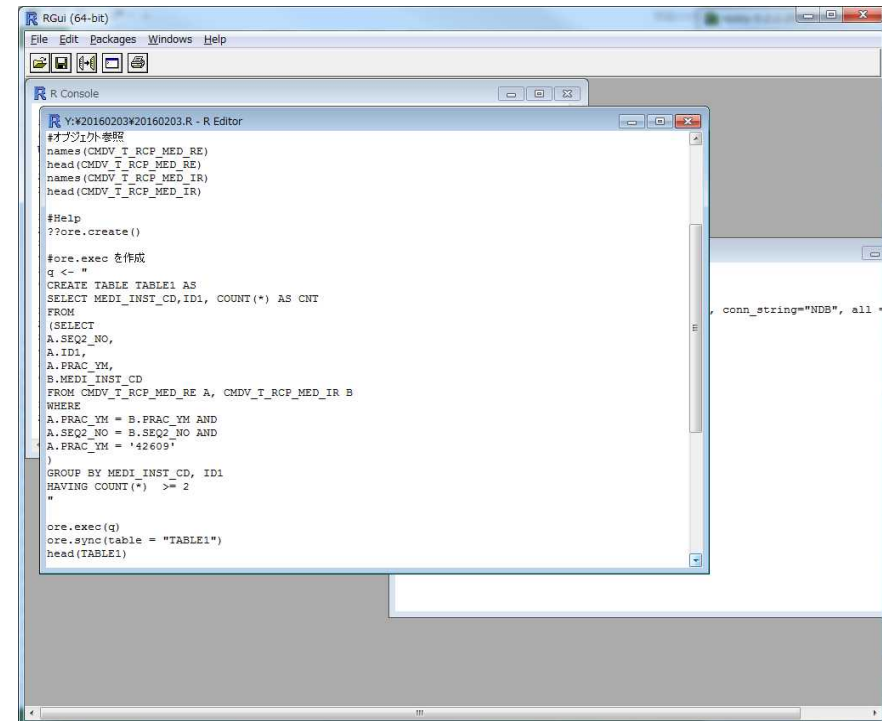
# ORE (SQL クエリを書く)

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。



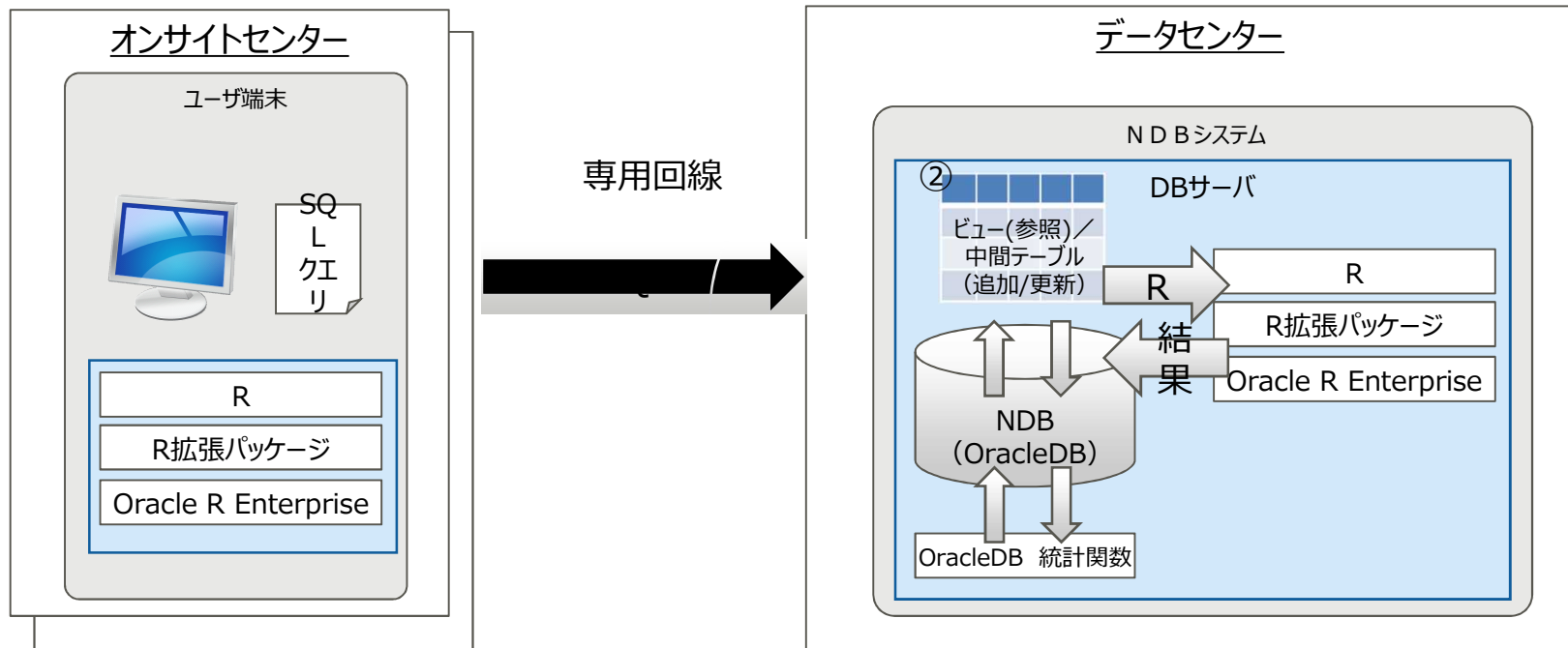
```
#オブジェクト参照
names(CMDV_T_RCF_MED_RE)
head(CMDV_T_RCF_MED_RE)
names(CMDV_T_RCF_MED_IR)
head(CMDV_T_RCF_MED_IR)

#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_RCF_MED_RE A, CMDV_T_RCF_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)

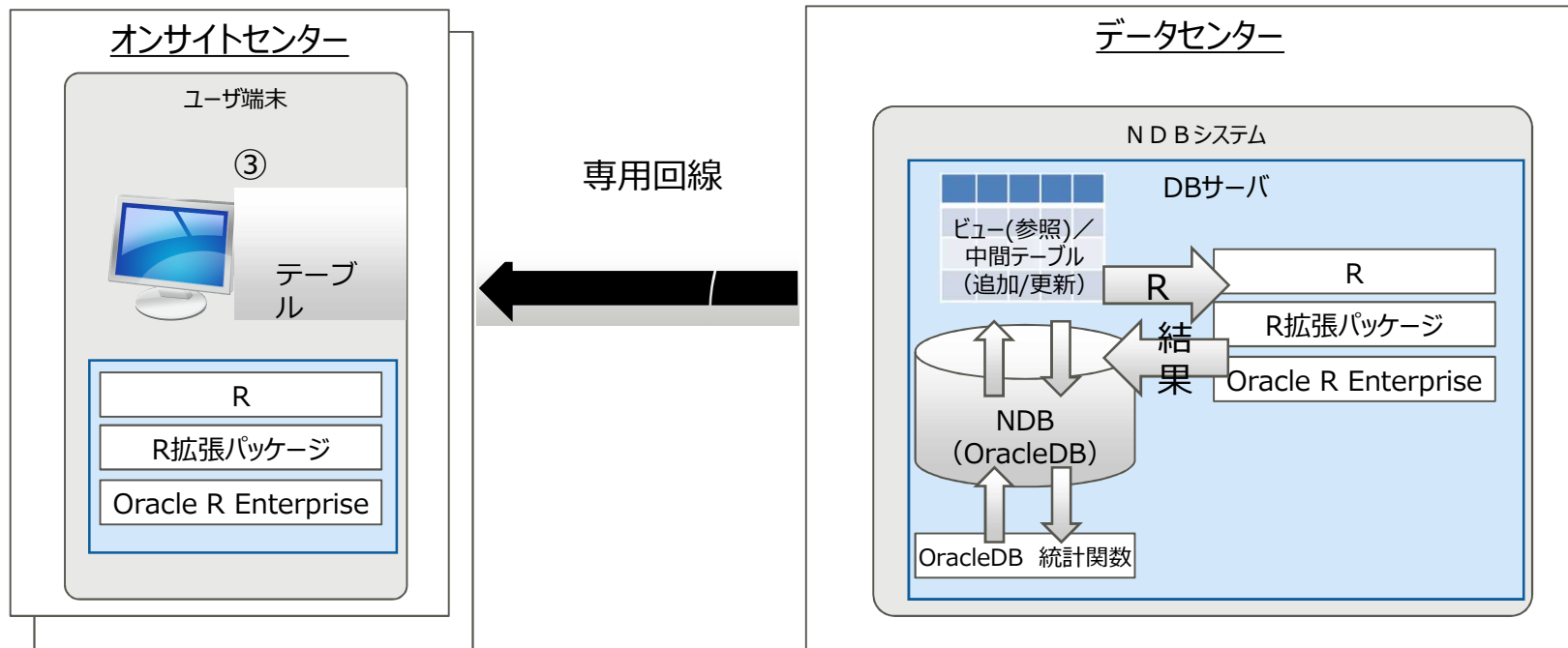
処理	クエリ詳細	処理時間(秒)		
		C34\$ (400名)	I63\$ (50000名)	I50\$ (90000名)
処理 1	H26.10 のDPC SBレコードを対象にICD10 コードがC34\$, I63\$, I50\$ のSEQ2_NO (レセプトID) を抽出する。	0.31	6.77	6.81
処理 2	H26.10 のDPC REレコードを対象に処理 1 で抽出した SEQ2_NO に紐づくID1 (患者ID) を抽出する。	0.41	4.05	2.12
処理 3	H26.10~H27.03 の期間のDPC レセREレコードから、処理 2 で抽出した ID1 に紐づく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。

# ORE の使いどころ

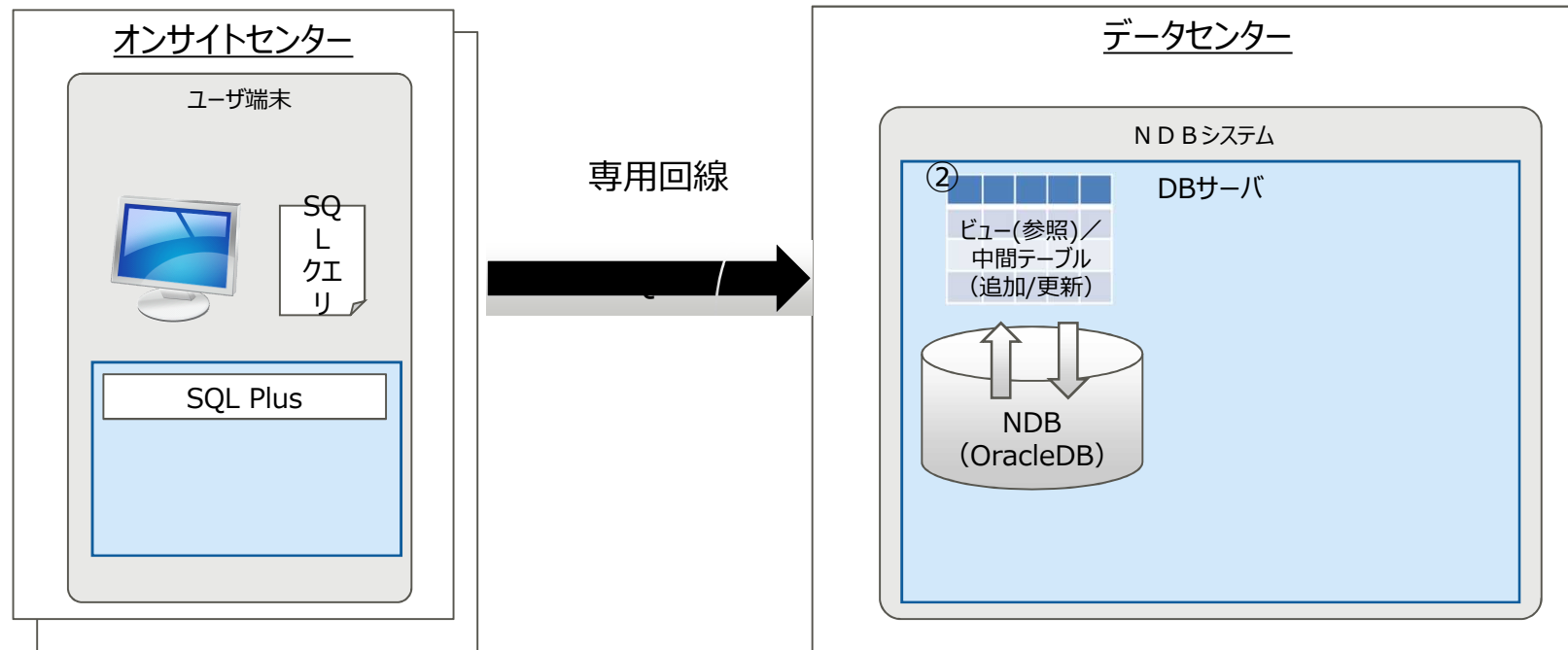
## データハンドリングよりもサーバーサイドでの解析実行

- ORE を用いてSQL を書くことは出来るが、エラーなどがラップされてしまい実用的ではない。
- ORE を用いてサーバー側のデータをローカルに持ち込み解析する事は出来るが、通信速度が制約因子になる。
- ORE を用いることでサーバーサイドで解析プログラムを走らせることが出来るので、解析時に真価を発揮するものと考えられる。
- さらなる検証が必要。

# 本日も話します内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

# SQL クエリを直接実行する。

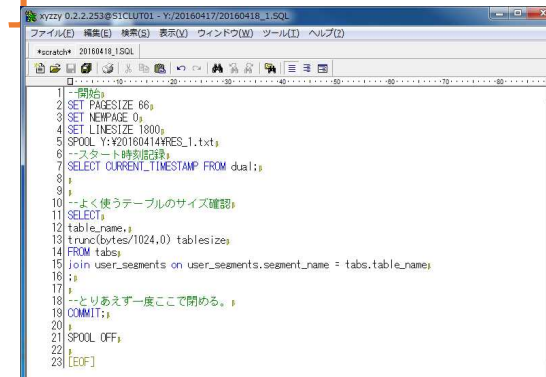


・SQL Plusを介してクエリを実行した。

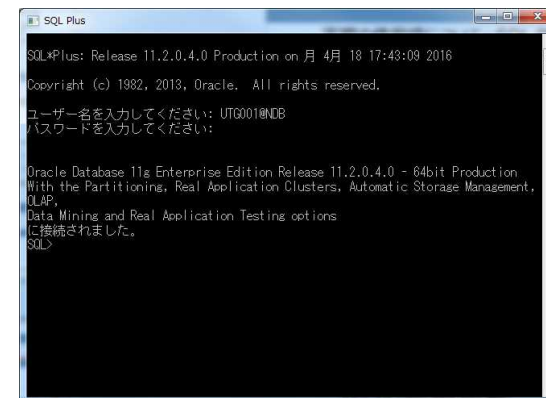
# SQL クエリを書く

## 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 からも参照可能
- エラーの内容が表示される。



```
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: 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: UTG001@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>
```

# 本日も話します内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

# 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するのに必要な時間とデータサイズ

抽出人数	RE	SY	IY	SI	TO	合計
686人 (1:00)	4096 (3:44)	2048 (3:44)	1024 (2:09)	5120 (10:1)	64 (00:08)	12MB (20.7min)
5213人 (1:02)	28672 (3:47)	14336 (3:45)	8192 (2:12)	34816 (10:48)	320 (00:09)	84MB (21.7min)
22803人 (1:03)	131072 (3:54)	64512 (3:59)	34816 (2:23)	155648 (11:15)	2048 (00:09)	379MB (22.7min)
178106人 (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)

# 本日も話している内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

# 個人追跡率

## 症例追跡率について

- NDB の個人の追跡はID1, ID2 によって行われる。
- ID1とID2 が同時に変わる事は少ない（寿退社など）と想定されている。
- ID1, ID2を用いた個人の追跡可能期間などは明らかになっていない。
- ID1, ID2 の何れかが一致していれば同一人物として扱い、個人の追跡可能期間とおおよその追跡中断率を検討した。

# 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リストを抽出するクエリ

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。）

# 個人追跡率 (処理方法)

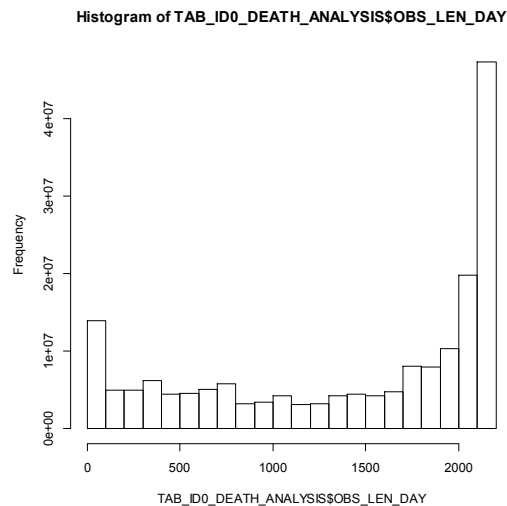
## 処理方法

- 診療年月H22.1–H27.12 の医科・DPC・調剤・歯科 のREレコードからID1, ID2 のペアを重複無しで抽出
- ペアIDを通番としてふる。
- ID1, ID2 基準で見ると親となるペアIDを各行に追加
- 再帰クエリを用いて全ての子ID の祖先ID を見つけ、それをID0 (独自ID) とした。

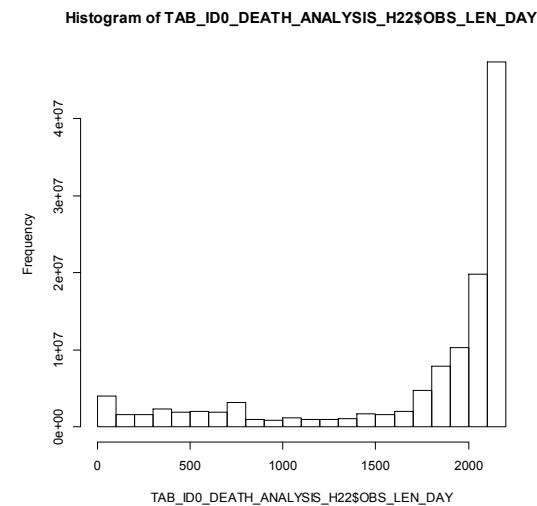
匿名化ID1	匿名化ID2	ペアID	親ID	祖先ID (ID0)
A	a	1	-	1
B	a	2	1	1
B	b	3	2	1
C	b	4	3	1
D	c	5	-	5
E	d	6	-	6

# 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 $4.73 \pm 1.7$  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。
- 現在、死亡アウトカムの妥当性と悉皆性の調査を行っている。



# 本日も話します内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

## その他課題

- データ追加時期
  - 毎月わずかにデータが追加されるため、抽出結果の再現性を確保するには、抽出時期をあらかじめ指定するなど工夫が必要となる。
- データの更新
  - 一部データは更新が行われるため、結果の再現ができない場合がある。これは、他の大規模データベース研究でも問題となっており、研究者側のコンセンサスが必要である。
- アクセス過多の問題

# 想定されるユーザー層

## ユーザーに求められるスキルセット

Group 名	統計/機械学習の知識	プログラミングスキル (主としてR)	ソフト利用(R, SAS, etc)	DB(SQL)	レセプトに関する理 解	解析の中身
Group A	○	○	○	○	○	大規模個票データの 解析
Group B			○	○	○	大規模個票データか らのサンプリングデー タ・集計データの解析
Group C			○		○	抽出済みデータの解 析
Group D					○	集計データのみを利用

# 本日も話しする内容

- 試行的利用に至る経緯および位置づけ
- 試行的利用における検討について
  - 1) BIツールを用いた集計
  - 2) Oracle R Enterpriseを用いた集計
  - 3) SQL Plus を用いた集計
- 試行的利用の検討結果と考察
  - 1) テーブルサイズ、各テーブルに対する検索・抽出パフォーマンス
  - 2) 個人追跡率、死亡追跡率のパフォーマンス
  - 3) その他の課題
- 本格利用に向けて
  - 1) 模擬申出研究テーマの実施準備
  - 2) 第三者提供に向けた管理規程および利用規定等の見直し

# 模擬申出研究テーマの実施

## 解析結果の公表を前提とした個別研究

- 第24回レセプト情報等の提供に関する有識者会議にて審議
  - 小林廉毅（東大）  
「後発医薬品の普及状況および関連要因に関する研究」
  - 大江和彦（東大）  
「レセプト情報等オンサイトリサーチセンターのレセプト情報等を用いた脳血管疾患の実態に関する研究」
  - 康永秀生（東大）  
「周術期口腔機能管理による術後肺炎発症予防の効果」

# 管理規程および利用規程等の見直し

- 本格利用に向け、以下の整備・見直しが必要となる。
  - 第三者提供に係る管理規程
  - 利用規程ならびにガイドラインの策定
  - 厚労省・利用者・オンサイトセンターの責任体制
  - 罰則規程
- 本試用期間で明らかになった課題について、引き続き有識者会議にて検討をお願いしたい。