大規模電子レセプト情報の解析のための データベース基盤の性能ベンチマークの検討

合田　和生† 　山田　浩之† 　喜連川　優† 　満武　巨裕††

† 東京大学 生産技術研究所　〒 153–8505 東京都目黒区駒場 4-6-1
†† 医療経済研究機構　〒 105-0003 東京都港区西新橋 1-5-11 第 11 東洋海事ビル 2F
E-mail: †{kgoda,hiroyuki,kitsure}@tkl.iis.u-tokyo.ac.jp, ††mitsutake@ihep.jp

あらまし　本論文では，大規模電子レセプト情報を解析するためのデータベース基盤の性能を制約条件下で測定し，評価するためのベンチマークを検討する．著者らは，平成 24 年より，大規模電子レセプト情報を解析するためのデータベース基盤「高速レセプト解析システム」を開発し，今日に至るまでの 5 年間に亘って，国内の医療分野・公共政策分野の研究者や実務者に対する実験的な解析サービスを提供してきた．平成 28 年 12 月現在，当該システムは平成 21 年度から平成 26 年度に亘る 6 年間に我が国の公的医療保険によって生じたほぼ全ての電子レセプト情報をホストしている．データセットの規模は 2,000 億レコードに昇り，世界最大級の健康医療ビッグデータに他ならない．本論文では，当該システムの開発と運用を通じて，著者らが得た知見を元に，大規模データベース基盤の開発と運用に資するための新たな性能ベンチマークを検討する．とりわけ，電子レセプト情報は，入れ子タプル構造を有し，多様なデータベースのスキーマ構造やシステムアーキテクチャによる実装が想定され，関係データモデルを前提とする従前の性能ベンチマークでは，評価することのできない実装が存在する．データベース基盤に求められる要件を高水準に規定することにより，多様な実装を評価可能とすることを試みる．本論文では，このための初期検討の一環として著者らが実施した電子レセプト情報に対する問合せを規定する枠組みの検討と著者らの「高速レセプト解析システム」に於ける測定事例を示す．

キーワード　診療報酬明細書（レセプト），公的医療保険，システム開発，性能ベンチマーク

1. はじめに

医療費の高騰を抑制しつつ，如何に高品位な医療を国民に提供し続けるかが，我が国が抱える最も深刻な課題の一つであることは周知の通りである [1,2,4,7,13,14]．国家レベルで医療の実態を徹底的に分析し尽くすことが，1 つの有力なアプローチであると考え，著者らは，平成 24 年より，大規模電子レセプト情報を解析するためのデータベース基盤「高速レセプト解析システム」[9,10] を開発し，これまでの 5 年間に亘って，国内の医療分野・公共政策分野の研究者や実務者に対する実験的な解析サービスを提供してきた．

当初は，厚生労働省が保有する「レセプト情報・特定健診等情報データベース」[12,15–17] より，1 年分の我が国ほぼ全ての電子レセプト情報の提供を受けて，データベースを構築し，これに基づく解析サービスを開発した．我が国の医療制度は，公的医療保険によるユニバーサルサービス（所謂，国民皆保険制度）を基礎としていることから，我が国で行われるほぼ全ての医療行為 (注1) は，公的医療保険によってカバーされている．「高速レセプト解析システム」は，我が国の医療の全容を把握可能である点が特徴的であり，その後の平成 28 年には，新たな電子レセプト情報の提供が認められ，本論文執筆時点においては，平成 21 年 4 月から平成 27 年 3 月までの 6 年分のほぼ全ての電子レセプト情報をホストし，これを解析可能としている．データセットの規模は 2,000 億レコードに到達しており，著者の知る限り，世界最大級の健康医療ビッグデータに他ならない．

本論文では，当該システムの開発と運用を通じて，著者らが得た知見を元に，大規模データベースの開発と運用に資するための新たな性能ベンチマークを検討する．とりわけ，電子レセプト情報は，入れ子タプル構造を有し，多様なデータベースのスキーマ構造やシステムアーキテクチャによる実装が想定され，関係データモデルを前提とする従前の性能ベンチマークでは，評価することのできない実装が存在する．データベース基盤に求められる要件を高水準に規定することにより，多様な実装を評価可能とすることを試みる．本論文では，このための初期検討の一環として著者らが実施した電子レセプト情報に対する問合せを規定する枠組みの検討と著者らの「高速レセプト解析システム」に於ける測定事例を示す．

本論文の構成は以下の通りである．2. では，我が国の公的医療保険に於ける支払い実務を支える電子レセプト情報を説明する．3. では，大規模電子レセプト情報を解析するためのデータベース基盤に求められる性能管理性を纏める．4. では，著者らが検討している性能ベンチマークを示すと共に，著者らの「高速レセプト解析システム」に於ける測定事例を示す．5. に於いて本論文を纏めるとともに，今後の課題と展望を述べる．

(注1)：評価療養，不妊治療における生殖補助医療，正常な妊娠・分娩，健康の維持・増進を目的とした健康診断や予防接種等は除く．

2. 我が国の公的医療保険に於ける支払い実務を支える電子レセプト情報

我が国では，全ての国民(注2)は，いずれかの公的医療保険制度に加入することが義務付けられており，国民が医療機関等(病院，診療所，調剤薬局等) において医療サービスを受けるにあたっては，通常，公的医療保険が適用され，国民が医療費の一部を払えばよい．この際，医療機関等は，請求書である診療報酬明細書（レセプト）を作成し，保険者に送付し，レセプトを受け取った保険者は，請求を審査し，医療費のうち保険者負担分を医療機関等に支払う．レセプトには，医療サービスを提供した医療機関等の情報，患者の個人情報，傷病名，提供した医療サービス，診療報酬ならびに参考情報等が記載される．通常は，医療機関等が提供した医療サービスに対して，対価である診療報酬が定められており，これを合算した金額が請求される．医療機関等の種別等に応じて，レセプトには記載様式が定められており，主に以下の 4 つに分類することができる．

- **医科用レセプト**：病院や診療所に於いて患者が外来診療もしくは入院診療を受けた際に発生する（ただし，後述の疾病群別包括払い制度が適用される場合と歯科診療は除く）
- **DPC 用レセプト**：一部の急性期病院等に於いて患者が疾病群別包括払い制度（diagnosis procedure combination; DPC）が適用された入院診療を受けた場合に発生する
- **歯科用レセプト**：病院や診療所に於いて患者が歯科診療を受けた際に発生する
- **調剤用レセプト**：調剤薬局に於いて患者が調剤を受けた際に発生する

レセプトは，古くは紙媒体によって処理されていたが，現在では，電子的な手段によるレセプトの記述と管理が標準化され，利用されている [8,11]．本論文では，このような電子的な手段によって記載されたレセプトを，電子レセプト情報と称することとする．電子レセプト情報の一例を図 1 に図示する．また，電子レセプト情報を構成するレコードの種別の一例を表 1 に纏める．電子レセプト情報を編成するファイルに於いては，RE レコードを先頭とする一連のレコードが，1 件の請求を構成する．即ち，1 件の請求は，1 件の RE レコードから開始し，次の RE レコードの直前で終了する．また，傷病名を表す SY レコードや医療行為を表す SI レコード等は，1 件の請求に複数が含まれることがある．このように，電子レセプト情報は，入れ子タプル (nested tuple) 形式の構造を有する．図 2 に，電子レセプトファイルの一例（医科レセプト）を示す.

レセプトは，本来は医療サービスの対価である診療報酬の請求に用いられる情報であるものの，そこには，医療機関等が提供した医療サービスの種別と数や，医療サービスを必要と判断した医療機関等による診断情報等が含まれている．診療録のように，臨床現場に於ける厳密な検査値等の値は含まれないものの，統一的な様式によって，国内で提供された医療サービスを網羅的に把握することが可能であり，国家の医療政策の立案等に活用するためには，我が国で生成される当該電子レセプト情報を一元的に収集し，解析することが，極めて有力なアプローチと考えられる．

一方で，我が国の電子レセプト情報は，年間あたり約 16 億件，約 350 億レコードの規模を誇る，世界最大級の健康医療ビッグデータに他ならない．これまでその大規模性故に，電子

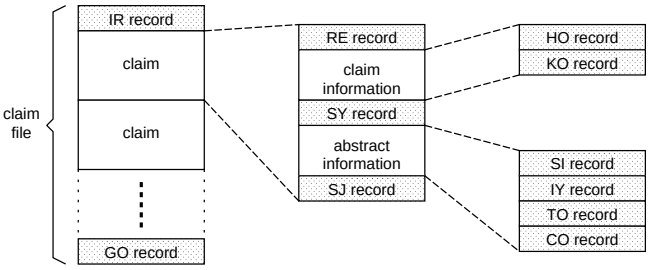

図 1　電子レセプト情報を格納するファイルの編成の一例（医科用レセプトの場合）(注4)

表 1　電子レセプト情報を構成する主なレコードの種別（医科用レセプトの場合）

種別	識別子	レコードの意味
医療機関情報	IR	保険医療機関に関する情報（医療機関コード，医療機関名称，請求年月等）を表す．医療機関単位のデータの先頭を意味する．
レセプト共通	RE	レセプトの基本情報（診療年月，レセプト種別，患者の氏名等の個人情報等）を記録する．患者単位のデータの先頭を意味する．
保険者	HO	医療費の請求先保険者に関する情報（保険者番号，被保険者番号，診療実日数，合計点数等）を記録する
公費	KO	医療費の公費負担に関する情報（負担者番号，受給者番号，診療実日数，合計点数等）を記録する
傷病名	SY	患者の傷病名に関する情報（傷病名コード，診察開始日，転帰区分等）を記録する
診療行為	SI	患者に行った診療行為に関する情報（診療行為コード，数量データ，点数，回数等）を記録する
医薬品	IY	患者に使用した医薬品に関する情報（医薬品コード，使用量，点数，回数等）を記録する
特定機材	TO	患者に使用した特定機材に関する情報（特定機材コード，使用領，点数，回数等）を記録する
コメント	CO	請求に関する医療機関のコメント（コメントコード，文字データ等）を記録する
症状詳記	SJ	患者の症状の詳細情報（症状詳記区分，症状詳記データ等）を記録する
診療報酬請求	GO	保険医療機関の診療報酬請求の合算情報（総件数，総合計点数等）を記録する．医療機関単位のデータの末尾を意味する．

(注2)：生活保護の受給者等を除く国内に居住する全国民と一定の在留資格のある外国人を被保険者とする.

(注4)：本論文の議論を越えることから，国保連固有情報レコード（KH），包括評価対象外理由レコード（GR），臓器提供者関連レコード（TI, TR, TS）は簡単のために省いている.

```
1  IR,1,13,1,9999939,,サンプル医科病院,42811,00,03-9999-9999
2  RE,3,1118,42805,サンプル医科０
   ２,1,3180717,,,,,,500,,,,,,,, ,01,,,,,,,,,,,,,
3  HO,06132013,１２３４５６７,１,1,141,,,,,,,,
4  SY,8845988,4280527,1,,,01,
5  CO,01,1,810000001,他科受診中にて初診料算定せず
6  SI,12,1,112011310,,73,1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,1
   ,,,,
7  SI,80,1,120002910,,68,1,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,1
   ,,,,
8  CO,,1,810000001,内科
9  IR,1,13,1,9999939,,サンプル医科病院,42811,00,03-9999-9999
10 RE,9,1124,42805,サンプル医科０
   ６,1,4240223,,,,,,500,,,,,,,, ,27,,,,,,,,,,,,,
11 HO,06132013,１２３４５６７,１,2,222,,,,,,,,
12 KO,88132121,2222222,,2,222,,,,,,
13 SY,8840014,4271216,1,,,01,
14 SY,3829027,4271216,1,,,,
15 SY,4742002,4271216,1,,,,
16 SY,4739014,4271224,1,,,,
17 SY,3814010,4280106,1,2058,,,
18 SY,8832434,4280323,1,,,,
19 SY,0389004,4280427,1,,,,
20 SY,4640013,4280427,1,,,,
21 SI,12,2,112011310,,,2,,,,,,,,,,,,,,,,1,,,,,,,,,,,,,,,1,,
   ,,,,
22 SI,,2,112006270,,111,2,,,,,,,,,,,,,,,,1,,,,,,,,,,,,,1,
   ,,,,,
```

図 2　電子レセプト情報の一例 (医科用レセプト，2 件の請求分)

レセプト情報の本格的な活用には至っていなかった．電子レセプト情報の利活用を促進するには，大規模性を超越し得る高性能なデータベース基盤を如何に構築するかが鍵と言える．本論文は，当該課題の解決の一端として，性能ベンチマークの在り方を議論するものである．

3. 大規模電子レセプト情報解析データベース基盤のための性能ベンチマーク

3.1 大規模電子レセプト情報に適した性能ベンチマークの必要性

性能ベンチマークは，データベース基盤が備える性能を議論する際の共通の指標となるものである．当然のことながら，性能の評価を行うには，性能ベンチマークは必須のものであるが，これ以外にも，性能ベンチマークは設計上の基準を明確に示すものとして，とりわけ，高速性を追求する基盤の開発工程には欠かせないものと言える．データベース基盤のための性能ベンチマークとしては，TPC が広く利用されており，中でも，TPC-H [5] は意思決定支援系，即ち，解析系データベース基盤のための性能ベンチマークとして，事実上の標準的なベンチマークである．現に TPC-H は，その前身である TPC-D の提案から 20 年を経過しているものの，今なお，産業界から性能値の登録が行われている他，学術論文においても TPC-H は広く指標として用いられている．

しかしながら，TPC-H は，以下のような点において，大規模電子レセプト情報解析のためにデータベース基盤の性能ベンチマークとして，十分に適したものではないことが判ってきた．

- TPC-H は，関係データモデルに基づき構成されたスキーマを前提とする．これに対して，電子レセプト情報は，上述の通り，入れ子タプル構造を有する．関係データモデルへの変換は一意ではなく，多様な変換による実装が可能であり，また，関係データモデルに変換する実装以外の実装も選択肢としてあり得る．現状の TPC-H では，そういった実装を評価することができない．
- TPC-H は，人工的に生成されたデータセットを前提としており，当該データセット内に於ける鍵属性の値は一様分布が仮定されている．これに対して，電子レセプト情報は，現実のデータであることから，著しい偏りが存在する．現状の TPC-H では，このような偏りのあるデータセットに対する性能を評価することができない．
- TPC-H が規定する問合せは，1%以上の選択率による対象データの全体的な傾向の把握を目的とする問合せが主である．しかしながら，著者らの経験では，解析のワークフローは，一般に，アドホックな問合せが繰り返され，当初は選択的でない問合せによって，概観を確認するのに対して，その後は徐々に選択性が高まる傾向があり，最終的にはより低い選択率（例えば，0.01%程度）を持つ問合せが頻発する．また，この際に，多段の自己結合が頻出する傾向がある．このような問合せは，TPC-H では看過されている．

大規模電子レセプト情報に適した新たな性能ベンチマークを以って，データベース基盤の性能を評価可能とすることが肝要と考え，著者らは後述のベンチマークを検討しているところである．

3.2 大規模電子レセプト情報解析データベース基盤のための性能ベンチマークの構成

著者らが検討している大規模電子レセプト情報のための解析データベース基盤の性能ベンチマークは，

（1） 電子レセプト情報の様式
（2） 電子レセプト情報のデータセット
（3） 電子レセプト情報に対する問合せ

から構成される．

電子レセプト情報の様式については，我が国で標準化された様式 [8] をベースとすることとする．ただし，実際に電子レセプト情報を解析する際には，電子レセプト情報の所有者（例えば，保険者）がデータベース基盤を直接構築することは現実的ではなく，秘密保持契約の下に，匿名化された電子レセプト情報を第三者に提供してデータベース基盤を構築させることとなるだろう．匿名化に際しては，氏名や被保険者番号等の極めて機密性の高い情報は削除され，また，所在地や生年月日等の機密性の高い情報は一定のルールに従って一般化され，代わりにハッシュ等の手段によって生成された識別子が与えられるのが一般的である．性能ベンチマークに於ける電子レセプト情報の様式としても，上記の加工が施されたものとする．

また，電子レセプト情報のデータセットとしては，当面は著者らに対して厚生労働省から提供されたデータセットを検討の材料として用いることとする．しかし，一般に電子レセプト情報を入手することは困難を伴うことから，TPC-H を参考に，人工的なデータセット生成器を開発している．当該生成器の仕様，構成法ならびに有効性については，別稿にて議論したい．

最後に，電子レセプト情報に対する問合せは，過去に著者らに医療系の研究者から寄せられた解析要求を参考に，現時点で

表 2　性能ベンチマークに於ける問合せの一覧候補

番号	概要
Q1	月別の医療費の請求動向を報告する
Q2	外来で特定の疾患に診断された患者群の治療に要した医療費を報告する
Q3	ある年に特定の疾患に診断された患者の数とその後に当該患者の治療に要した医療費を報告する
Q4	ある年に特定の疾患に診断された患者の数とその後 2 年に重症化した患者の数を報告する
Q5	特定の疾患に診断された患者への特定の医薬品の処方の動向を報告する
Q6	特定の疾患に診断された患者群について，他の特定の疾患の診断の状況を報告する

20 個程度の問合せ候補を規定しており，そのうちの主要なものの概要を表 2 に纏める．なお，TPC-H では，規定された関係データベーススキーマを前提として，SQL により問合せが定義されている．これに対して，電子レセプト情報の場合，先述の通り，多様な実装が想定されることから，同様に SQL を以って定義することは適切ではない．このため，性能ベンチマークでは，電子レセプト情報の様式を前提として，これに対する問合せを自然言語で規定することとした．ただし，自然言語には曖昧性が伴う可能性があり，これを除外するために，同時に関係論理式を提示することにより，問合せを定義することとした．具体例を次節に示す．

なお，上記で議論した問合せは，いずれもアドホックな解析問合せであり，現実のユーザの解析プロセスを観察すると，一定の仮定の下に行う仮説検証は，このような問合せを数回程度実行することにより完了する．他方，解析手法や課題を模索する課程に於いては，可視化ツール等を用いて，類似の問合せを繰り返すことにより，対象をドリルダウンし，最終的に解を得る傾向が見られる．このような過程の性能ベンチマークへの組み込みは，今後の研究課題としたい．

3.3　電子レセプト情報に対する問合せの一例

表 2 に示した Q2 を例に示す．

3.3.1　自然言語による問合せの規定例

表 3 に自然言語（英語）による問合せ Q2 の規定例を示す．なお，MED, DPC, DEN ならびに PHA は，それぞれ医科，DPC，歯科，調剤の各レセプト種別を意味する．この際の，電子レセプト情報の解釈ルールを表 4 に示す．

3.3.2　関係論理式による問合せの規定例

関係論理式による問合せ Q2 の規定例を以下に示す．

$$\begin{aligned} Q2:\{p,n1,n2\} = \{&a1.p, a1.n1, a2.n2 \mid \\ &a1 \in A1 \wedge a2 \in A2 \wedge \\ &(\exists a1 \wedge \exists a2)(a1.p = a2.p)\} \end{aligned}$$

ここに，$A1$ ならびに $A2$ は関係表である．以下に，$A1$ の定義を示す．なお，F_t はレセプト種別 t のレセプトを格納するレセプトファイルを，$c(c \in F)$ はレセプトファイル F に格納される 1 つのレセプトを，$r(r \in c)$ はレセプト c に含まれる 1 つのレコードを，r_a はレコード r に含まれる属性 a を意味する．また，記述の簡単のため，$r^{\mathrm{TYPE}}(r^{\mathrm{TYPE}} \in c)$ は，集合 $\{r | r \in c \wedge r(p)_{\text{record type}} = \mathrm{TYPE}\}$ の元を意味することとする．また，$\mathcal{A}(R)$ は，関係表 R に対する集約演算の適用を表す．ここでは，簡略的に記載しているが，より厳密な公式化については，文献 [3] 等を参考に，別に議論したい．

表 3　自然言語による問合せ Q2 の規定例

List, for each prefecture,
- a prefecture code of the prefecture; and
- the aggregate number of points that any medical organizations located in the prefecture claimed, in MED outpatient claims, to care patients satisfying **Condition 1** on an intervention date during the term satisfying **Condition 2**; and

the aggregate number of points that any dispensing pharmacies located in the prefecture claimed, in PHA claims, to care patients satisfying **Condition 1** on a dispensiing date during the term satisfying **Condition 2**;

where
- **Condition 1** requires that the patients shall be declared, by any medical organizations in any MED outpatient claims, to be diagnosed with a primary disease specified by a disease code matching any of `DISEASE CODE LIST` on a diagnosing date during the term satisfying **Condition 2**; and
- **Condition 2** requires that the term shall match [`DATE`, `DATE` + 1 year).

表 4　自然言語による問合せ Q2 に於ける電子レセプト情報の解釈ルール

- A claim type (MED, DPC, DEN or PHA) shall be identified by auxiliary information. Further classification of claim types (outpatient or inpatient) shall be identified by a 'claim type' attribute in a RE record (MED).
- A patient shall be identified by a combination of 'name', 'gender' and 'birthday' attributes in a RE record and 'symbol part of insured ID' and 'number part of insured ID' attributes in a HO record (MED, DPC, DEN and PHA). If a claim contains an alternative attribute useful for identifying a patient, the use of the attribute is allowed.
- A intervention/dispensing date shall be identified by an 'intervention year and month' attribute in a RE record (MED, DPC or DEN) or a 'dispensing year and month' attribute in a RE record (PHA).
- A prefecture shall be identified by a 'prefecture' attribute in a IR record.
- The number of points shall be identified by a 'total number of points' attribute in a HO record (MED, DPC, DEN and PHA). Note that a claim of MED, DEN and PHA contains a single HO record. In contrast, a claim of DPC may contain plural HO records. For each claim of DPC, the first HO record or the HO record with the highest total number of points shall be evaluated.
- A primary disease shall be identified by a 'disease code' in a SY record in which a 'primary disease' attribute is flagged on (MED and DPC) or a 'disease code' in a SB record in which a 'primary disease' attribute is flagged on (DPC). For a DPC claim, a disease identified either in a SY record or in a SB record is valid.
- A diagnosing date shall be identified by a 'intervention start date' attribute in a SY record (MED or DPC) or by a 'intervention start date' attribute in a SB record (DPC). For a DPC claim, a diagnosing date identified either in a SY record or in a SB record is valid.

$$A1 : \{p, n1\} = \mathcal{A}(R1)\{p \to \sum n1\}$$

$$\begin{aligned} R1 : \{p, n1\} = \{ & r1^{\mathrm{IR}}_{\text{prefecture code}}, r1^{\mathrm{HO}}_{\text{number of points}} | \\ & c1 \in F_{\mathrm{MED}} \wedge c3 \in F_{\mathrm{MED}} \wedge \\ & (\exists r1^{\mathrm{RE}} \wedge \exists r1^{\mathrm{IR}} \wedge \exists r1^{\mathrm{HO}} \wedge \exists r3^{\mathrm{RE}} \wedge \exists r3^{\mathrm{SY}})(\\ & r1^{\mathrm{RE}} \in c1 \wedge r1^{\mathrm{IR}} \in c1 \wedge r1^{\mathrm{HO}} \in c1 \wedge \\ & r3^{\mathrm{RE}} \in c3 \wedge r3^{\mathrm{SY}} \in c3 \wedge \\ & r1^{\mathrm{RE}}_{\text{claim type}} = \text{outpatient} \wedge \\ & r1^{\mathrm{RE}}_{\text{intervention year and month}} \in [\texttt{DATE}, \texttt{DATE} + 1 \text{ year}) \wedge \\ & r1^{\mathrm{RE}}_{\text{name}} = r3^{\mathrm{RE}}_{\text{name}} \wedge r1^{\mathrm{RE}}_{\text{gender}} = r3^{\mathrm{RE}}_{\text{gender}} \wedge \\ & r1^{\mathrm{RE}}_{\text{birthday}} = r3^{\mathrm{RE}}_{\text{birthday}} \wedge r1^{\mathrm{HO}}_{\text{insured ID}} = r3^{\mathrm{HO}}_{\text{insured ID}} \\ & r3^{\mathrm{SY}}_{\text{diagnosing date}} \in [\texttt{DATE}, \texttt{DATE} + 1 \text{ year}) \wedge \\ & r3^{\mathrm{SY}}_{\text{primary disease}} = \text{primary} \wedge \\ & r3^{\mathrm{SY}}_{\text{disease code}} \in \texttt{DISEASE\_CODE\_LIST})\} \end{aligned}$$

同様に，以下に $A2$ の定義を示す．

$$A2 : \{p, n2\} = \mathcal{A}(R2)\{p \to \sum n2\}$$

$$\begin{aligned} R2 : \{p, n2\} = \{ & r2^{\mathrm{IR}}_{\text{prefecture code}}, r2^{\mathrm{HO}}_{\text{number of points}} | \\ & c2 \in F_{\mathrm{PHA}} \wedge c4 \in F_{\mathrm{PHA}} \wedge \\ & (\exists r2^{\mathrm{RE}} \wedge \exists r2^{\mathrm{IR}} \wedge \exists r2^{\mathrm{HO}} \wedge \exists r4^{\mathrm{RE}} \wedge \exists r4^{\mathrm{SY}})(\\ & r2^{\mathrm{RE}} \in c2 \wedge r2^{\mathrm{IR}} \in c2 \wedge r2^{\mathrm{HO}} \in c2 \wedge \\ & r4^{\mathrm{RE}} \in c4 \wedge r4^{\mathrm{SY}} \in c4 \wedge \\ & r2^{\mathrm{RE}}_{\text{claim type}} = \text{outpatient} \wedge \\ & r2^{\mathrm{RE}}_{\text{intervention year and month}} \in [\texttt{DATE}, \texttt{DATE} + 1 \text{ year}) \wedge \\ & r2^{\mathrm{RE}}_{\text{name}} = r4^{\mathrm{RE}}_{\text{name}} \wedge r2^{\mathrm{RE}}_{\text{gender}} = r4^{\mathrm{RE}}_{\text{gender}} \wedge \\ & r2^{\mathrm{RE}}_{\text{birthday}} = r4^{\mathrm{RE}}_{\text{birthday}} \wedge r2^{\mathrm{HO}}_{\text{insured ID}} = r4^{\mathrm{HO}}_{\text{insured ID}} \\ & r4^{\mathrm{SY}}_{\text{diagnosing date}} \in [\texttt{DATE}, \texttt{DATE} + 1 \text{ year}) \wedge \\ & r4^{\mathrm{SY}}_{\text{primary disease}} = \text{primary} \wedge \\ & r4^{\mathrm{SY}}_{\text{disease code}} \in \texttt{DISEASE\_CODE\_LIST})\} \end{aligned}$$

3.4 高速レセプト解析システムを用いた問合せ処理性能の計測例

著者らは，東京大学で開発したアウトオブオーダ型高速データベースエンジン [6] を基盤とする実験システム「高速レセプト解析システム」[9, 10] を，平成 24 年に東京大学生産技術研究所内に構築し，その後，5 年間に亘って，国内の医療分野・公共政策分野の研究者や実務者に対する実験的な解析サービスを提供してきた．平成 29 年 3 月より，利用ユーザの拡張と安定稼働を目的として，新システムの設計と構築を進めてきた結果，同年 12 月より新システムが始動している．当該新システムのコアであるデータベース基盤のハードウェア諸元を表 6 に示す (注5)．また，ホストしている電子レセプト情報の概数を表 5 に纏める．データベース基盤を構成するデータベースシステムとしては，アウトオブオーダ型データベースエンジンの商用実装である HADB を用いた．

(注5)：新システムは，エンドユーザの利便性向上のための仮想デスクトップ端末や，セキュリティの向上を目的とした監査システム等，多様なサブシステムから構成されている．

表 5　提供を受けた平成 21 年度から平成 26 年度分の電子レセプト情報の諸元

レセプト種別	請求件数	レコード数
医科	5,330,562,780 件	111,357,539,283 件
DPC	64,018,540 件	11,625,759,091 件
歯科	504,593,167 件	7,067,753,410 件
調剤	3,417,052,688 件	48,593,978,069 件
合計	9,316,227,175 件	178,645,029,853 件

表 6　実験システムのハードウェアの諸元

サーバ	
プロセッサ	4x Intel Xeon E7-8890v4 (2.2GHz, 24 cores)
主記憶	2,048GB
NIC	2x 1Gbps ports
HBA	12x 16Gbps FC ports
OS	Redhat Enterprise Linux
ストレージ	
コントローラ	2x (256GB cache memory)
HBA	12x 16Gbps FC ports
ストレージデバイス	9x 12.8TB flash modules 17x 6.4TB flash modules

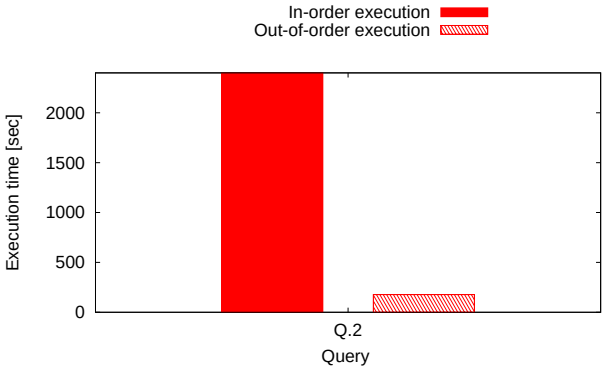

図 3　性能試験の結果

著者らは，性能ベンチマークの検討の一環として，上記の新システムを用いて，前述の問合せを実行し，その実行にかかる時間を計測した．この際，アウトオブオーダ型実行を行わない場合と，行う場合を比較した．なお，測定にあたっては，両者とも本論文執筆時点までに可能な範囲で，十分な性能チューニングを行った．結果を図に示す．アウトオブオーダ型実行を行わない場合は，実行を 2,400 秒で打ち切った．

4. おわりに

本論文では，大規模電子レセプト情報を解析するためのデータベース基盤「高速レセプト解析システム」を開発し，国内の医療分野・公共政策分野の研究者や実務者に対する実験的な解析サービスを提供して著者らの知見をもとに，新たに行っている大規模データベースの開発と運用に資するための新たな性能

ベンチマークの検討を示した．当該特性を有する電子レセプト情報に対する問合せを規定する枠組みを提案するとともに，著者らの「高速レセプト解析システム」に於ける測定事例を示した．本論文での議論ならびに医療分野・公共政策分野の研究者や実務者からの意見をもとに，性能ベンチマークの開発を進めているところである．いずれは，一定の議論を経て，仕様や関連ツールを公開したい．

謝　　辞

本研究の一部は，厚生労働科学研究費政策科学推進研究「汎用性の高いレセプト基本データセット作成に関する研究」，厚生労働科学特別研究事業戦略研究「レセプト情報・特定健診等情報データベースを利用した医療需要の把握・整理・予測分析および超高速レセプトビッグデータ解析基盤の整備」，内閣府最先端研究開発支援プログラム（FIRST）「超巨大データベース時代に向けた最高速データベースエンジンの開発と当該エンジンを核とする戦略的社会サービスの実証・評価」，内閣府革新的研究開発推進プログラム（ImPACT）「社会リスクを低減する超ビッグデータプラットフォーム」，日本医療研究開発機構（AMED）臨床研究等 ICT 基盤構築研究事業「エビデンスの飛躍的創出を可能とする超高速・超学際次世代 NDB データ研究基盤構築に関する研究」の助成に依る．レセプト情報・特定健診等情報データベースからの電子レセプト情報の第三者提供に掛かる手続きに関しては，厚生労働省保険局保険システム高度化推進室から丁寧なご指導を頂いた．

文　　献

[1] Victor R. Fuchs. The Gross Domestic Product and Health Care Spending. *New England Journal of Medicine*, Vol. 369, No. 2, pp. 107–109, 2013.

[2] Peter Groves, Basel Kayyali, David Knot, and Steve van Kuiken. The 'big data' revolution in healthcare: Accelerating value and innovation. McKinsey & Company, 2013.

[3] Gultekin Özsoyoglu, Z. Meral Özsoyoglu, and Victor Matos. Extending relational algebra and relational calculus with set-valued attributes and aggregate functions. *ACM Transactions on Database Systems (TODS)*, Vol. 12, No. 4, pp. 566–592, 1987.

[4] Sheila Smith, Joseph P. Newhouse, and Mark S. Freeland. Income, Insurance, And Technology: Why Does Health Spending Outpace Economic Growth? *Health Affairs*, Vol. 28, No. 5, pp. 1276–1284, 2009.

[5] Transaction Processing Performance Council. TPC-H, an ad-doc, decision support benchmark. `http://www.tpc.org/tpch/`.

[6] 喜連川優, 合田和生. アウトオブオーダ型データベースエンジン OoODE の構想と初期実験. 日本データベース学会論文誌, Vol. 8, No. 1, pp. 131–136, 2009.

[7] 厚生労働省. 厚生労働統計：国民医療費. `http://www.mhlw.go.jp/toukei/list/37-21.html`. 2017 年 1 月 15 日に参照.

[8] 厚生労働省保険局. 診療報酬情報提供サービス. `http://www.iryohoken.go.jp/shinryohoshu/`. 2017 年 1 月 15 日に参照.

[9] 合田和生, 山田浩之, 喜連川優, 満武巨裕. 我が国の公的医療保険の悉皆分析を可能とする高速レセプト解析システムの開発と今後の展望. 電子情報通信学会第 9 回データ工学と情報マネジメントに関するフォーラム／第 15 回日本データベース学会年次大会 (DEIM2017), pp. E3–2, 2017.

[10] 山田浩之, 合田和生, 喜連川優. 128 ノード規模のストレージインテンシブクラスタ環境におけるアウトオブオーダ型並列データ処理系の性能評価. 電子情報通信学会論文誌, Vol. J98-D, No. 5, pp. 728–741, 2015.

[11] 社会保険診療報酬支払基金. レセプト請求形態別の請求状況（平成 28 年度）：平成 28 年 10 月診療分. `http://www.ssk.or.jp/tokeijoho/tokeijoho_rezept/tokeijoho_04_h28.files/seikyu_2810.pdf`. 2017 年 1 月 15 日に参照.

[12] 藤森研司. レセプトデータベース（NDB）の現状とその活用に対する課題. 医療と社会, Vol. 26, No. 1, pp. 15–24, 2016.

[13] 内閣府. 経済・財政一体改革推進委員会，社会保障ワーキング・グループ資料. `http://www5.cao.go.jp/keizai-shimon/kaigi/special/reform/wg1/281013/shiryou1-2.pdf`. 2017 年 2 月 17 日に参照.

[14] 内閣府. 国民経済計算（GDP 統計）. `http://www.esri.cao.go.jp/jp/sna/menu.html`. 2017 年 1 月 15 日に参照.

[15] 満武巨裕. 日本のレセプト情報・特定検診等データベース（NDB）の有効活用. 情報処理, Vol. 56, No. 2, pp. 140–144, 2015.

[16] 満武巨裕. レセプトビッグデータ解析の現状と将来. 実験医学, Vol. 34, No. 5, pp. 799–804, 2016.

[17] 満武巨裕, 大江和彦, 今中雄一. NDB オープンデータを研究利用に活用する：医療技術（CT, MRI, PET）の利用に関する国際比較の試み. 社会保険旬報, Vol. 2661, pp. 12–16, 2016.