20120102 9A

厚生労働科学研究費補助金
政策科学総合研究事業（政策科学推進研究事業）

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

（課題番号H24－政策－一般－002 ）

平成２４年度　総括研究報告書

研究代表者　　満武　巨裕

平成 25 年（2013）年 3 月

一般財団法人　医療経済研究・社会保険福祉協会
IHEP 医療経済研究機構

目　次

厚生労働科学研究補助金（政策科学総合研究事業（政策科学推進研究事業））
総括研究報告書

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

研究代表者　満武巨裕
一般財団法人　医療経済研究・社会保険福祉協会　医療経済研究機構　副部長

研究要旨

分担報告書1の目的は、レセプト情報等データベース（以下、NDB）の利用促進のために、汎用性の高いデータセットの設計と作成を行うことである。初年度は、NDBの利用申請に加えて、保険者からもNDBと同様のデータ提供を受け、データの検証、基本データセットを試作する。

データを授受の後、レセプト情報等解析システムを構築した。構築後は、データの検証を行い、基本データセットとして研究利用しやすい(1)一レセプト一レコード、(2)一定期間のレセプトを個人ID毎に統合する等の処理をし且つ診療行為も含むデータセットを設計・試作した。また、(3)特定健診・保健指導データとレセプトとリンケージしたデータセットも試作した。データの検証、基本データセットの設計等は、本研究班と厚生労働省の関連部局から構成される委員会で検討した。疾患情報、医療費の傾向などの情報を作成し、研究計画を企画・立案できる方式についても検討した。

NDBデータの検証結果から、IDに関しては、被保険者証番号に半角・全角や氏名に関しては漢字表記の揺れ等の理由で、実際には同一の被保険者にも関わらず、複数のIDが発生している可能性があることが伺えた。そのために、試行的に、ID1とID2の特性を生かし、複数紐づくものに関しては同一人物とみなして、新しいIDを試作した。また、特定健診データとレセプトデータとの突合率は約2割であった。

本年度は、データ提供日から報告書作成までの期間が限られたことから、レセプト情報等解析システムとデータの検証作業に多くの労力を費やした。基本データセットについては、詳細なレコードフォーマットを引き続き更新していくものとし、厚生労働省・保険局と情報共有し、ホームページなどでも情報提供を行うことも検討する。また、利用者へ個人情報が特定されない範囲で、対象とする疾患の人数、医療費の傾向などの情報を提供し、研究計画を企画・立案できる方式について検討したところ、ウェブサイトにおいてパラメータを指定することで対象とする疾患の人数などがブラウザ上に表示されるダッシュボード機能は、個人情報は含まず、研究計画書を作成する際の有益な情報提供手段となると考えられた。

現在、NDBから提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプト

を研究者の要望に応じて、ある程度分析し易い形式に加工して提供している。汎用性のある基本データセットを事前作成しておくことで、提供件数の増加、厚生労働省および研究者の負担軽減にもつながると考えられる。

分担報告書 2 における要旨は以下である。レセプト等ナショナル・データベースは社会にとって極めて大きな潜在的価値を有しており、これを最大限に活用するしくみを推進していく必要がある。当研究は、レセプト等ナショナル・データベースのデータ処理や活用のあり方について検討し、より効率的で効果的な運用・活用を提案し、より大きな社会的価値を生み出すようにすることを目的とする。そのために、有識者ならびに現在システムを担当するベンダーと検討会合を持ち、データベース等に強いベンダー各社と検討会合を持ち、各種の専門家等との意見交換、文献、海外の先行事例の動向などから課題と解決策・向上策を検討した。その結果、以下が重要と考えられた。

1. 中期的なアウトプットを具体的にデザインすることで効率的なシステム設計を可能にする。
2. ベンダーに依存しないシステムを構築する。
3. 定型、非定型を分けて、同時処理を可能にし、データ処理を効率的に行う。
4. 研究に活用可能な定期集計データを集めたデータテーブルを充実、公開する。
5. NDB 運用を担当する組織を常設する。
6. 各種マスターを維持管理する組織を持つ。
7. データ形式は研究に適した DPC E/F ファイル形式で構築する。
8. 利用者の支援・教育のための組織を常設する。
9. システム担当組織の中で自らをレビューし改善していく仕組みを持つ。
10. システム担当ベンダーの外部に、当該ナショナル・データベースのシステム運用とデータ活用を継続的にレビューし課題を整理し、継続的に改善を提言する仕組みを持つ。

定期的・不定期を含め数々あるアウトプットとその出力タイミングと必要なハードとソフトを明確に設計し、特定のベンダーに依存しないことが求められる。定期・非定期のタスクのもとに、各種マスターやデータテーブル、システム全体を組織的に常時見直して継続的に完全していく仕組みの確立、などが極めて重要である。

研究分担者
（分担報告書 1）
大江和彦、東京大学医学部附属病院・企画情報運営部、教授
喜連川優、東京大学生産技術研究所、教授
伏見清秀、東京医科歯科大学大学院・医療政策情報学、教授
岩崎　学、成蹊大学・理工学部、教授
印南一路、慶應義塾大学・総合政策学部、教授
清水沙友里、一般財団法人　医療経済研究・社会保険福祉協会　医療経済研究機構、主任研究員
合田和生、東京大学・生産技術研究所、特任准教授（研究協力者）
山田浩之、東京大学大学院・情報理工学系

研究科、博士後期課程（研究協力者）

（分担報告書 2）
今中雄一、京都大学 大学院医学研究科 医療経済学分野 教授
猪飼 宏、京都大学 大学院医学研究科 医療経済学分野 講師
大坪徹也、京都大学 大学院医学研究科 医療経済学分野 助教
研究協力者
後藤 悦、京都大学 大学院医学研究科 医療経済学分野
小林大介、京都大学 大学院医学研究科 医療経済学分野
森島敏隆、京都大学 大学院医学研究科 医療経済学分野
國澤 進、京都大学 大学院医学研究科 医療経済学分野
佐々木典子、京都大学 大学院医学研究科 医療経済学分野
田中将之、京都大学 大学院医学研究科 医療経済学分野
宇川直人、京都大学 大学院医学研究科 医療経済学分野

A. 研究目的

分担研究報告書 1 の目的は、レセプト情報等データベース（以下、NDB）の利用促進のために、研究用途における汎用性の高いレセプト基本データセットの設計と作成することである。初年度は、NDB の利用申請を行って基本データセットを作成するためのデータソースを入手し、試行版を作成することとした。

分担報告書 2 では、レセプト等ナショナル・データベースのデータ処理や活用のあり方について検討し、より効率的で効果的な運用・活用を提案し、より大きな社会的価値を生み出すようにすることを目的とする。

B. 研究方法

分担報告書 1 では、初年度は、NDB の利用申請と同時に保険者から NDB と同様のデータの提供を受けて、基本データセットの試行版を作成する。NDB のレセプト（電子レセプト）は、詳細な診療行為データ（診療行為コード(SI)や医薬品コード(IY)等）が含まれる。そこで、診療行為記録を含み、且つ研究者が利用しやすい (1)一レセプト一レコード、(2)一定期間のレセプトを個人 ID 毎に統合する等の処理をしたデータセットを試作する。本邦の NDB には、諸外国には存在しない特定健診・保健指導データを含むことも特徴の一つである。そこで、(3)特定健診・保健指導とレセプトとリンケージしたデータセットも作成する。

NDB の利用申請によるデータ抽出は、被保険者マスター（台帳あるいは個人 ID 一覧）が利用できないために、電子レセプトレコード種別のレセプト情報(HO)を利用して抽出する条件を設定する。

対象期間は、2009 年度～2011 年度である。（抽出データとの比較対照用に、2011 年度の全保険者のデータも入手した。）対象レセプトは、医科（入院、入院外）、調剤、DPC、歯科および、特定健診・特定保健指導のデータである。

レセプト情報等から汎用性の高いレセプト基本データセットを作成するためには、当該レセプト情報等が潜在的に備える

情報の価値を研究班が実験的に見出しながら（把握しながら）進める必要があり、本格的なレセプト情報等の解析システムが欠かせない。研究班では、東京大学の研究室において「レセプト情報等解析システム」と称する解析用 IT プラットフォームを構築し、当該プラットフォームを用いて、基本データセットの作成を進めた。

基本データセット作成を円滑に進めるためには、レセプト情報等に対する解析処理を機動的に行うことが欠かせず、最先端の情報技術を投入することとし、レセプト情報等の管理・解析のためのデータベースエンジンとしては、研究分担者である喜連川らが内閣府最先端研究開発支援プログラムにおいて開発を進めている超高速データベースエンジンを用いることとした。また、研究者による解析クエリの発行ならびに結果の確認を容易とするために、レセプト情報等の解析のためのビジネスインテリジェンスツールを新たに開発し、グラフィカルインターフェース上で解析クエリの構成と発行、実行状況の確認、グラフや表による結果の表示とダウンロード等、一連の作業を統合的に実施できることとした。当該ツールの開発に関しては、研究班において保険医療分野の研究者と情報技術分野の研究者の密接な連携の下、機能追加を機動的に行う等により、研究の推進に大いに資した。

分担報告書 2 においては、レセプト等ナショナル・データベースのデータ処理や活用のあり方について、以下の方法で、検討し提案を構築する。

1）有識者ならびに現在システムを担当するベンダーと会合を持ち、グループでヒアリングと意見交換を行う。それを以て、解決策・向上策を検討する。

2）データベース等に強いベンダー各社からヒアリングを行い、課題と解決策・向上策を検討する。

3）各種の専門家等との意見交換、文献、海外の先行事例の動向などから、課題と解決策・向上策を検討する。

C. 研究結果

分担研究報告書 1 では、対象期間のレコード件数を外来・入院・DPC・調剤・歯科別に示したところ、2010 年度からデータ件数が安定した。レセプト種別ごとに、NDB の ID に関する検証を行った結果、外来レセプトの ID1 に対する ID2 の数は、一対一対応が 75.4%、ID1 に対して ID2 が二つは 22.7%であった。また、外来レセプトの ID2 に対する ID1 の数は、一対一対応が 80.4%、ID1 に対して ID2 が二つは 15.6%であった。調剤レセプトは、外来診療に付随して発生するものであるが、調剤レセプトに対応した外来レセプトが存在しない件数が ID1 では 5.6%、ID2 では 12.9%存在した。医療費に関しては、NDB から得られたレセプトデータを医科（入院、外来）・調剤ごとの集計値と国民医療費（厚生労働省・統計情報部）との比較を行った。その結果、同様の傾向を示した。疾患別医療費も同様の傾向を示した。特定健診データと外来・入院・DPC のレセプトデータとの突合を ID1 および ID2 を用いて行った。その結果、ID1 では突合率 18.5%、ID2 では、19.9%であった。

基本データセットの構成は、現時点では

五段階とした。第一段階（データセットA）は、データ項目がID、性別、年齢階級からなる提供データの管理用データセットである（被保険者台帳（マスタ）を想定しており、第三者提供は行わない予定である）。第二段階（データセットB）は、患者毎に紐付けを行い、3年分のデータ、および単年度分毎のサマリ情報からなるデータセットであり、基本的属性情報に加え集計された医療費、主傷病名等の限られたデータ項目から構成される。第三段階（データセットC）は、1回の入院、あるいは一連の外来診療をひとまとまりの単位として整備したデータセットであり、データ項目はデータセットBと同水準の情報を含む。第四段階（データセットD）は、診療をひとまとまりの1回の入院、あるいは一連の外来単位として整備したデータセットデータセットCの情報に加え、診療行為別の点数情報から構成されている。第五段階（データセット最明細）は、データセットDに加え、傷病名や診療行為、医薬品等のデータが含まれ、現在のNDBが保有する源データに近い。同様に2010年度のデータに関しても、データ検証を行った。

分担研究報告書2では、重要点を要約すると以下の如く整理された。ただし、断言の形で言い切っているが、実運用に適用する際には、現実的な状況に合わせて最善の手段をとる必要がある。

1.「中期的なアウトプットを具体的にデザインすることで、効率的なシステム設計を可能にする」

・NDBの特徴を生かせる定期集計（月次、年次、社会医療診療行為別調査等を含む）、政策上の要求事項、想定される非定期集計、研究領域を明確にして、それに適した新しいシステムの設計を、現行システムにとらわれることなく行う。

2.「ベンダーに依存しないシステムを構築する」

・データ処理の手順やプログラムをNDBの運営組織に引き継ぎ、また蓄積できるような仕組みづくりを行う。

・データ処理過程の全体の仕様において、特定のベンダーによるブラックボックス化を避けるようなオープンシステムを構築する。

3.「定型、非定型を分けて、同時処理を可能にし、データ処理を効率的に行う」

・定型（月次処理）、非定型（NDB）に適したシステムをそれぞれ独立に構築し、並列処理を行うことで効率化を図る。

・例えば、定期処理はNoSQL、非定型処理はRDBを用いる。

4.「研究に活用可能な定期集計データを集めたデータテーブルを充実、公開する」

・全件をカバーするNDBの特徴として定期集計データだけでも研究に資するものがあり、早急な公開が望まれる。

・どのような集計データが有用であるかは、学会等の意見も聞き、検討していく必要がある。

（実現には予算を伴うNDB周辺重要事項）

5.「NDB運用を担当する組織を常設する」

・データの将来活用計画、保守運用、利用申請への対応を厚労省の少人数の担当者で行うことは限界があるため、今後いっそうの活用を進める上では、常設の組織がNDBの運用を担当することが望まれる。

6.「各種マスターを維持管理する組織を持つ」

・必要とされるマスターテーブルは多数あり、例えば診療報酬改定等で随時更新されるものや、市町村合併等により変更されるものもある。

・マスターのバージョン管理(更新および追加)に責任を持ってあたる専門組織が必要である。

・研究対象に即して最適なバージョンのマスター類が適用されることで、データ解析の精度が上がることが期待できる。

7.「データ形式は研究に適したDPC E/Fファイル形式で構築する」

・既にDPC参加病院では作成されているファイル形式であり、E/Fファイルを厚生労働省に提出している。

・DPCに参加していない医療機関においては、レセコンにすでにE/Fファイルを作成する機能はあり、その機能を追加導入することで、ファイル作成が可能になる。

・各審査支払機関において一時的なシステム変更コストを要する一方、E/Fファイル形式は審査においても一層の効率化に資する可能性が大きい。

8.「利用者の支援・教育のための組織を常設する」

・申請者や申請を考慮している人へのサポート、データの一部を取り扱うことが可能なオンサイトセンターの運営、人材育成等を担う組織を常設する。

9. システム担当組織の中で自らをレビューし改善していく仕組みを持つ。

10. システム担当ベンダーの外部に、当該ナショナル・データベースのシステム運用とデータ活用を継続的にレビューし課題を整理し、継続的に改善を提言する仕組みを持つ。

D. 考察

(分担研究報告書1)

2009年度のレセプト件数が他の年度と比較して少なかったのは、医科(診療所)のレセプト電子率が医科(病院)や調剤よりも低かったことが原因と考えられる。よって、NDBのデータでは、若干捕捉できないレセプトがあることを考慮しなければならない。IDの重複の結果から、被保険者番号や記号等の数値の入力に関して半角・全角入力や空白の取り扱い、氏名欄の記載の揺れ等の影響があることが示唆された。調剤レセプトが外来レセプトと突合できずに単独で存在するレセプトが5.6%存在しているが、その理由としても、IDの問題が指摘できる。特定健診データとレセプトデータとの突合率は、約2割であった。本研究に協力している国民健康保険においては、特定健診を受診した約87.6%が医療機関を受診していたために、特定健診とレセプトデータの突合方法についても引き続き検討していくべき課題である。

基本データセット設計に関しては、ID1とID2の両方をベースに作成したが、試行的にID1とID2の特性を生かし、双方が複数紐づくものに関しては同一人物とみなして、新しいID3を構築した。ID3によって、ある程度は記載情報の揺れを補正できたと考えられた。

本年度は、データ提供日から報告書作成までの期間が限られたことから、データの検証作業とシステム構築に多くの時間と

労力を費やしたが、次年度の重要な検討課題としてはIDの利用が第一に挙げられる。では、諸外国では研究提供用データセットのIDをどのように管理しているのだろうか。例えば、米国のCMS(Centers for Medicare & Medicaid Services)では、メディケアとメディケイドでIDが異なるために、複数のデータソースから個人を識別する独自のID(BENE_ID: Beneficiary Unique Identifier)をつくることで対応している。このBENE_IDは、CMSのCCW(Chronic Condition Data Warehouse：慢性疾患データウェアハウス)上で運用されており、個々の受給者が、どの年でも、診療請求レセプトの種類が違っても、制度（メディケア・メディケイド）が違っても一意のIDとなっている。また、CMSではメディケア・メディケイドの診療請求データ（日本のレセプトに相当）に加えて、被保険者の「資格(eligibility)ファイル」、「加入記録(enrollment)ファイル」が収集されている。現在、NDB上ではIDの整備・管理はされていないが、米国のようにID管理を行うためにも、今後は保険者が管理している被保険者台帳などの情報も入手して、IDの管理機能を持つべきである。また、基本データセットの設計については、詳細なレコードフォーマットを作成中であり、引き続き更新していくものであり、厚生労働省・保険局と情報共有すると同時に、ホームページなどでも情報提供をすることも検討する。

利用者へ個人情報が特定されない範囲で、対象とする疾患の人数、医療費の傾向などの情報を提供し、研究計画を企画・立案できる方式について検討したところ、ウェブサイトにおいてパラメータを指定することで対象とする疾患の人数などがブラウザ上に表示されるダッシュボード機能は、個人情報は含まず、研究計画書を作成する際の有益な情報提供手段となると考えられた。

現在、レセプト情報等データベース（以下、NDB)から提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易い形式に加工して提供している。汎用性のある基本データセットを事前作成しておくことで、提供件数の増加、厚生労働省および研究者の負担軽減にもつながると考えられる。

E. 結論

（分担研究報告書1）

NDBの集計データは、国民医療費と同様の傾向を示したことから、一定の精度が保たれていると考えられた。IDに関しては、被保険者証番号には半角・全角や氏名に関しては漢字表記の揺れ等の理由で、実際には同一の被保険者にも関わらず、複数のIDが発生している可能性がある。調剤レセプト単独で存在するレセプトが存在しており、特定健診データとレセプトデータとの突合率は約2割であった。

基本データセットのレコードフォーマットは、引き続き更新していくものとし、厚生労働省・保険局と情報共有すると同時に、ホームページなどでも情報提供をすることも検討する。

利用者が研究計画を企画・立案できる方式について検討したところ、対象疾患の人数、医療費の傾向などの情報をブラウザ上

に提供するダッシュボード機能の活用が有効と考えられた。

現在、NDBから提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易い形式に加工して提供している。汎用性のある基本データセットを事前作成しておくことで、提供件数の増加、厚生労働省および研究者の負担軽減にもつながると考えられる。

（分担研究報告書2）

レセプト等ナショナル・データベースは社会にとって極めて大きな潜在的価値を有する。これを最大限に活用するしくみを導入していく必要がある。そのためには、定期的・不定期を含め数々あるアウトプットとその出力タイミングと必要なハードとソフトを明確に設計し、特定のベンダーに依存しないことが求められる。定期・非定期のタスクのもとに、各種マスターやデータテーブル、システム全体を組織的に常時見直して継続的に完全していく仕組みの確立、などが極めて重要である。

F. 健康危険情報

なし

G. 研究発表

該当なし

H. 知的所有権の取得状況

該当なし

Ⅱ．分担研究報告書

厚生労働科学研究補助金（政策科学総合研究事業）
分担研究報告書 1

基本データセットの作成

満武　巨裕
一般財団法人　医療経済研究・社会保険福祉協会　医療経済研究機構、副部長
大江　和彦
東京大学医学部附属病院・企画情報運営部、教授
喜連川　優
東京大学生産技術研究所、教授
伏見 清秀
東京医科歯科大学大学院・医療政策情報学、教授
岩崎　学
成蹊大学・理工学部、教授
印南　一路
慶應義塾大学・総合政策学部、教授
清水沙友里
一般財団法人　医療経済研究・社会保険福祉協会　医療経済研究機構、主任研究員
合田和生（研究協力者）
東京大学・生産技術研究所、特任准教授
山田浩之（研究協力者）
東京大学大学院・情報理工学系研究科、博士後期課程

研究要旨

本研究では、レセプト情報等データベース（以下、NDB）の利用促進のために、汎用性の高いデータセットの設計と作成を行う。初年度は、NDB の利用申請に加えて、保険者からも NDB と同様のデータ提供を受け、データの検証、基本データセットを試作することを目的とする。

データ授受の後、レセプト情報等解析システムを構築した。構築後に、データの検証を行い、基本データセットとして研究利用しやすい (1)一レセプト一レコード、(2)一定期間のレセプトを個人 ID 毎に統合する等の処理をし且つ診療行為も含むデータセットを試作した。また、(3)特定健診・保健指導データとレセプトとリンケージしたデータセットも試作した。データの検証、基本データセットの設計等は、本研究班と厚生労働省の関連部局から構成される委員会で検討した。疾患情報、医療費の傾向などの情報を作成

し、研究計画を企画・立案できる方式についても検討した。

NDB の集計データは、国民医療費と同様の傾向を示したことから、集計データとしては一定の精度が保たれていると考えられた。ID に関しては、レセプト種別毎の ID1 と ID2 のユニーク件数が、入院、DPC、調剤では ID2 の方が少なく、外来では逆の結果であった。被保険者証番号には半角・全角等の混在、氏名に関しては漢字表記の揺れ等のため、同一被保険者にも関わらず複数 ID が発生していることが考えられた。そのために、試行的に、ID1 と ID2 の特性を生かし、双方が複数紐づくものに関しては同一人物とみなして、新しい ID3 も試作した。また、調剤レセプトには、外来レセプトと突合できないレセプトが存在していた。特定健診データとレセプトデータとの突合率は約 2 割であった。

本年度は、データ提供日から報告書作成までの期間が限られたことから、データの検証作業に多くの労力を費やした。基本データセットについては、詳細なレコードフォーマットを引き続き更新していくものとし、厚生労働省・保険局と情報共有すると同時に、ホームページなどでも情報提供をすることも検討する。また、利用者へ個人情報が特定されない範囲で、対象とする疾患の人数、医療費の傾向などの情報を提供し、研究計画を企画・立案できる方式について検討した。例えば、ウェブサイトにおいて複数のパラメータを指定することで、対象疾患の人数などを表示するダッシュボード機能は、個人情報は含まず、研究計画書を作成する際の有益な情報提供手段となると考えられた。次年度は、 (4) 希少疾患や長期観察のデータセット、(1) ～ (3) データセットの更新について検討する。

現在、NDB から提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易い形式に加工して提供している。そのために、汎用性のあるデータセットを事前作成しておくことで、提供件数の増加につながると共に厚生労働省および研究者の負担軽減にもつながると考えられる。

A. 研究目的

本研究の目的は、レセプト情報等データベース（以下、NDB）の利用促進のために、研究用途における汎用性の高いレセプト基本データセットの設計と作成を行うことである。初年度は、NDB の利用申請を行って基本データセットを作成するためのデータソースを入手し、試行版を作成することとした。また、NDB からデータの提供を受けるまでに、時間を要することから、保険者からもレセプトと特定健診のデータ提供を受けて、基本データセットを試作する。

B. 研究方法

B-1：NDB データの利用申請およびデータ抽出条件

保険者は、被保険者の加入・退出を記録した被保険者台帳（マスタ）を整備することで、医療保険の利用にあたっての資格照合を行っている。そのために、被保険者台帳（マスタ）からランダムサンプリングを行うことで対象とする被保険者を指定し、その被保険者の関連するレセプト情報等

を抽出することが望ましい。しかし、NDBは被保険者台帳（マスタ）を保有していない（保険者、医療機関などのマスタ等の作成もしていない）。実態は、巨大なCSVファイルの格納庫ともいえる。

今回は、被保険者番号等の個人IDを指定した抽出ができないために、電子レセプトレコードの保険者情報(HO)を利用する抽出条件を設定した。

協会けんぽ、市町村国保、国保組合、後期高齢者医療制度は、保険者レコード(HO)の都道府県情報を利用して、特定地域への偏りを極力避けるため各都道府県ブロックから1つをランダム抽出した（ただし、北海道と沖縄は、それぞれ東北ブロックと九州ブロックに組み入れた。県数が最小の中部と四国ブロックからは、1県が抽出されるようにした）。

組合健保、共済組合については、都道府県による区別ができないので、保険者(保険者レコード(HO))を約25%で抽出（ランダムサンプリング）した。

ランダムサンプリングは、研究班以外の第三者が実施した。その結果（対象県、対象保険者のリスト）をNDBデータの利用申請書に添付した。対象期間は、2009年度～2011年度とした。対象レセプトは、医科（入院、入院外）、調剤、DPC、歯科および特定健診・特定保健指導のデータである。また、抽出データとの比較対照用として、2010年度の全保険者のデータを入手した（職域保険および地域保険のレセプトデータであり、公費単独のレセプトデータは除外されている）。

B-2：レセプト情報等解析システム

レセプト情報等から汎用性の高いレセプト基本データセットを作成するためには、当該レセプト情報等が潜在的に備える情報の価値を研究班が実験的に見出しながら（把握しながら）進める必要があり、本格的なレセプト情報等の解析システムが欠かせない。研究班では、東京大学の研究室において「レセプト情報等解析システム」と称する解析用ITプラットフォームを構築し、当該プラットフォームを用いて、基本データセットの作成を進めた。

レセプト情報等解析システムの概要を参考資料5の図1に示す。当該システムは、研究室内のデータセンタに設置したサーバ装置群と、研究室内の専用パーティションに設置したクライアント端末から構成される。サーバ装置群は、厚生労働省から研究班に開示されたレセプト情報等の堅牢な管理と、レセプト情報等に対する解析クエリの実行を担い、クライアント端末は、通常の研究者へのインターフェースであり、レセプト情報等の管理ならびに解析の命令をサーバ装置群に発行し、また、その結果を取得して出力する。

参考資料5の表1にはサーバ装置群のハードウェア構成を示す。通常時においては、サーバ装置群は1台の常設のサーバ装置を以って構成しており、当該ハードウェア資源を以ってレセプト情報等の管理ならびに解析は円滑に行うことができている。しかしながら、研究の進展状況によっては、多数の解析クエリを短期間に機動的に処理していく需要が見込まれたことから、研究室の備える最大で128台の臨時のサーバ装置を追加投入可能とすることとし、解析処理能力を柔軟に増強できるシステム

を実現した。

参考資料5の図2にレセプト情報等解析システムのソフトウェア構成を示す。基本データセット作成を円滑に進めるためには、レセプト情報等に対する解析処理を機動的に行うことが欠かせず、最先端の情報技術を投入することとし、レセプト情報等の管理・解析のためのデータベースエンジンとしては、研究分担者である喜連川らが内閣府最先端研究開発支援プログラムにおいて開発を進めている超高速データベースエンジン（参考資料5の文献1,2）を用いることとした。また、研究者による解析クエリの発行ならびに結果の確認を容易とするために、レセプト情報等の解析のためのビジネスインテリジェンスツールを新たに開発し、グラフィカルインターフェース上で解析クエリの構成と発行、実行状況の確認、グラフや表による結果の表示とダウンロード等、一連の作業を統合的に実施できることとした。当該ツールの開発に関しては、研究班において保険医療分野の研究者と情報技術分野の研究者の密接な連携の下、機能追加を機動的に行う等により、研究の推進に大いに資した。

なお、レセプト情報等解析システムの構築ならびに運用におけるレセプト情報等の取り扱いに際しては、厚生労働省のガイドラインに従い、運用管理規定ならびに内部監査規定を設け、情報セキュリティの確保と個人情報の保護に努めた。例えば、サーバ装置群については、研究室へのICカード扉錠による立ち入り制限に加えて、研究室内に「制限区画A」と称する保安区画をに設け、専用セキュアラック扉錠によるアクセス制限を行ったほか、論理的なセキュリティ手段として、IPアドレス認証と公開鍵方式のユーザ・端末認証を組み合わせたアクセス制限やアクセス監査ログ取得を施した。また、クライアント端末については、同じく研究室へのICカード扉錠による立ち入り制限に加えて、研究室内に「制限区画B」と称する保安区画をに設け、パーティションにより物理的に窃視を防ぐ措置を実施した他、共通鍵方式のユーザ認証、セキュリティソフトウェアの導入、出力デバイスの制限、アクセス監査ログ取得を施した。制限区画Aと制限区画Bを接続するネットワークについては、保守作業時を除き、他とは独立させることとしている。このように、多重的に物理的ならびに論理的な保安措置を行っており、年度末には情報セキュリティ専門家による監査を受けた。（監査の終了時には、非公式ながら「特に問題はない。徹底したセキュリティ管理がなされている」との評を頂いた。）次年度も、情報セキュリティの確保と個人情報の保護に努めながら、基本データセットの作成の資するべく、レセプト情報等解析システムを発展させていく予定である。

B-3：データ検証

NDBには、2009年4月から日本の全保険医療機関の電子レセプトデータが蓄積されているが、蓄積されたデータの検証作業はこれまで行われていない。そこで、データ検証を行った。

B-4：基本データセットの試作

基本データセットは、(1)一レセプト一レコード、(2)一定期間のレセプトを個人ID毎に統合する等の処理をしたデータセ

ットを試作した。(3)特定健診・特定保健指導とレセプトとリンケージしたデータセットも作成した。

C. 研究結果

C-1：レコード件数

外来、入院、DPC、調剤、歯科別にNDBから得られた電子レセプトの件数を月別に示したところ、2009年度は件数が上昇傾向にあったが、2010年度からデータ件数が（比較的）安定した(図表.1)。

C-2：ID1とID2のユニーク件数

レセプト種別ごとに、NDBのID1とID2の一意(ユニーク)の件数を調べた。(ID1は、保険者番号・記号番号・生年月日、性別から生成したハッシュ値。ID2は、氏名・生年月日・性別から生成したハッシュ値である)。外来レセプトのID2のユニーク件数のほうがID1より多かった。一方で、入院・DPC・調剤に関しては、ID1のユニーク件数のほうがID2より多く、逆の結果であった（図表.2)。

C-3：ID1とID2の重複件数

ID1に対するID2の重複件数、ID2に対するID1の重複件数を外来、入院、調剤、DPC、歯科毎に調べた。外来レセプトのID1とID2が一対一対応している割合は75.4%、ID1に対してID2が二つ対応している割合は22.7%であった(図表.3)。また、外来レセプトのID2とID1が一対一対応している割合は80.4%、ID2に対してID1が二つ対応している割合は15.6%であった(図表.4)。入院レセプトのID1に対するID2が一対一対応している割合は96.7%、ID1に対してID2が二つ対応している割合は3.2%であった(図表.5)。また、入院レセプトのID2に対するID1が一対一対応している割合は96.5%、ID2に対してID1が二つ対応している割合は3.3%であった(図表.6)。DPCレセプトのID1に対してID2が一対一対応している割合は97.7%、ID1に対してID2が二つ対応している割合は2.2%であった(図表.7)。また、DPCレセプトのID2に対してID1が一対一対応している割合は97.2%、ID2に対してID1が二つ対応している割合は2.8%であった(図表.8)。調剤レセプトのID1に対してID2が一対一対応している割合は85.6%、ID1に対してID2が二つ対応している割合は13.4%であった(図表.9)。また、調剤レセプトのID2に対するID1が一対一対応している割合が82.9%、ID2に対してID1が二つ対応している割合は14.2%であった(図表.10)。

C-4：外来・調剤レセプトのマッチング

調剤レセプトは、外来診療に付随して発生する。そのため、ID1及びID2を利用して、外来レセプトと調剤レセプトのマッチングを行った。その結果、ID1では外来レセプトに対応しない調剤レセプト件数が5.6%存在した。これは、調剤医療費の2.2%に該当した。ID2では、外来レセプトに対応しない調剤レセプト件数件数が12.9%発生し、調剤点数の8.3%を占めた(図表.11)。

C-5：国民医療費との比較

医療費に関して、NDBから得られたデータの医療費と国民医療費（厚生労働省・統

計情報部）との比較を行った(図表. 12)。その結果、2010 年度の入院・外来・調剤別に医療費構成割合が同様の傾向を示した(外来医療費の割合は、微減していた)。疾患別医療費との比較に関しても、同様の傾向を示した(図表. 13)。また、参考資料 1 には、疾患分類表（中分類）での集計表も掲載した(図表. 14-20)。

C-6：特定健診データとレセプトデータの突合

特定健診データと外来・入院・DPC レセプトデータとの突合を ID1 および ID2 を用いて行った。その結果、ID1 では特定健診受診者の 18. 5%が医療機関を利用していた。ID2 では、19. 9%であった。

C-7：基本データの試作

本年度、施行した基本データセットは、五段階の構成とした（参考資料 3:基本データセットの構成とレコードフォーマットを参照)。第一段階（データセット A）は、データ項目が ID、性別、年齢階級からなる管理用データセットである(被保険者台帳（マスタ）を想定しており、第三者提供は行わない予定である)。第二段階(データセット B）は、患者毎に紐付けを行ったデータセットである。各年および 3 年分の統合データからなる。データ項目は、年齢と性別の属性に加え、医療費、傷病名など限られたデータ項目からなるデータセットである。第三段階（データセット C）は、1 回の入院、あるいは一連の外来診療をひとまとまりの単位として整備したデータセットである。データ項目はデータセット B と同水準の情報を含む。第四段階（データセット D）は、診療をひとまとまりの 1 回の入院、あるいは一連の外来単位として整備したデータセットデータセットＣの情報に加え、診療行為別の点数情報から構成されている。第五段階（データセット最明細）は、データセット D に加え、傷病名や診療行為、医薬品等のデータが含まれ、現在の NDB が保有する源データに近い。

また、基本データセットには、保険者の都道府県番号と医療機関の都道府県番号が存在している。そこで、研究班では、レセプトごとに保険者と医療機関の距離を算出して新たなデータ項目として設定した(被保険者が同一県内の医療機関を受診した場合の距離はゼロである)。

C-8:2010 年度データの検証

同様に 2010 年度のデータに関するデータ検証結果を参考資料 2 に示した。

D. 考察

2009 年度のレセプト件数が他の年度と比較して少なかったのは、レセプト電子率が低かったことが原因だと思われる(医科（診療所)のレセプト電子率が医科(病院)や調剤よりも低かった)。よって、2009 年度の NDB のデータは、捕捉できないレセプトが存在することを考慮しなければならない。特に、外来レセプトと DPC レセプトに影響があると考えられる。

国民医療費との比較から、電子レセプトに対応していない医療機関の影響が、外来レセプトに多少見られるものの、集計データに関しては、国民医療費と同様の傾向を

示した。疾病別医療費も同様の傾向を示したことから、ID を利用した個人毎のデータには課題があるとしても、医療費の集計データとしては一定の精度が保たれていると考えられる。

レセプト種別毎の ID1 と ID2 のユニーク件数は、入院、DPC、調剤は ID2 の方が少ない一方で、外来は逆の結果であった。保険者が変更になった被保険者は、複数の ID1 が発生する。また、名前が変更になった被保険者は、複数の ID2 が発生する。検証結果は、外来レセプトからは 19.6%の被保険者の保険者が変更になり、25.6%の被保険者の氏名が変更になったことを示してが、この割合は現実よりも高いと考えられる。おそらく、被保険者証番号には、半角・全角などが混在しており、実際には同一の被保険者にも関わらず、複数の ID1 が発生している可能性がある。加えて、氏名に関しては漢字表記の揺れ（外事、漢字変換ミス、ひらがな・カタカナ併記等）のために、同一被保険者にも関わらず複数発生していることが考えられる。

今後、ID2 に対して ID1 が複数発生しているケースは同一人として判断してもよいのか、あるいは、ID1 に対して複数の ID2 がある場合は保険者の変更によるものと判断して同一の被保険者と定義するのか、他のデータ項目との整合性も観察しながら、検討を深めていかなければならない。今回、NDB には 2 種類の ID1 と ID2 が存在するという利点を生かし、双方が複数紐づくものに関しては同一人物とみなして、新しい ID3 を試作した。図表 35 に 2010 年のデータを利用して、ID1 と ID2 の発生数（外来、入院、DPC を統合）を示した。ID1 が 113,970,732 件、ID2 が 122,425,816 件だった。全国民が医療機関を利用しているわけではないので、おそらく ID1 および ID2 は実際の人数以上、データベース上に存在していることが考えられる。そこで、ID1 と ID2 の双方が複数紐づくものに関しては同一人とみなし ID3 を試作した結果、109,779,461 件となった。この数値の妥当性についても、今後検討していかなければならない。

調剤レセプトが外来レセプトと突合できずに単独で存在するレセプトが 5.6%存在しているが、その原因としても、ID の問題が指摘できる。もちろんこの 5.6%の中には、外来診療後の翌月に調剤薬局にいったケースも含まれている可能性があるが、一般に調剤レセプトが単独で発生することはない。

特定健診データとレセプトデータとの突合率は、約 2 割であった。特定健診は 40 歳以上 75 歳以下の被保険者を対象としている。この年齢層の医療機関受診率は、2 割は超えるものと考えられる。本研究に協力を得た国民健康保険において、同様の突合を行ったところ、約 87.6%が医療機関を受診していた（三重県の 15 国民健康保険（市町）の特定健診受診者 43,536 人の調査）。

本年度は、データ提供日から報告書作成までの期間が限られたことから、データの検証作業に多くの労力を費やした。

その結果、次年度の重要な検討課題としては ID の利用が第一に挙げられる。では、諸外国では研究提供用のデータセット ID をどうしているのだろうか。例えば、米国の CMS(Centers for Medicare & Medicaid

Services)では、メディケアとメディケイドでIDが異なるために、複数のデータソースから個人を識別する独自のID(BENE_ID: Beneficiary Unique Identifier)をつくり対処している。BENE_IDは、CMSのCCW(Chronic Condition Data Warehouse：慢性疾患データウェアハウス)上で運用されており、個々の受給者が、どの年でも、診療請求タイプが違っても、制度（メディケア・メディケイド）が違っても一意（ユニーク）のIDとなるような仕組みがある。CMSのデータベースの中に新たな受給者発生すると新しいBENE_ID交付されるが、その際に既存データとの確認も行われIDの整合性を保つ仕組みが確立している。加えて、CMSではメディケア・メディケイドの診療請求データ(日本のレセプトに相当）に加えて、被保険者の「資格(eligibility)ファイル」、「加入記録(enrollment)ファイル」が収集されている。今後、日本もID管理を行うためにも、保険者が管理している被保険者台帳などの情報も入手すべきである。例えば、被保険者台帳を基に作成したハッシュIDのリスト内に、レセプトのデータを基にしたハッシュIDが存在するかについての検証機能を持たせるべきである。

基本データセットの設計については、詳細なレコードフォーマットを作成中であり、引き続き更新していくものである。その都度、厚生労働省・保険局と情報共有すると同時に、ホームページなどでも情報提供をすることも検討する。

最後に、利用者へ個人情報が特定されない範囲で、対象とする疾患の人数、医療費の傾向などの情報を提供し、研究計画を企画・立案できる方式について検討した。米国CMSのCCWにおいては、ウェブサイトにおいてパラメータを指定することでブラウザ上に、対象とする疾患の人数などが表示される機能を提供している（参考資料4参照)。これは、ダッシュボードとよばれ、個人情報は含まれないが、研究計画書を作成するうえで、初期の情報提供手段となる例である。日本においても、（個人情報を含まない）性別、年齢、疾病、地域等のパラメータを選択し、医療費、人数、入院日数などをダッシュボードの形で提示する機能を備えることも有益であると考えられる。

次年度は、基本データセットでは対応できない希少疾患や長期観察が必要な研究のためのデータセット（以下、特殊データセット）の検討も行う。台湾における特殊データセットには、癌、糖尿病、精神疾患、高額な医療費を要する疾患といった15種類のデータセットがある。米国においては、メディケア加入者に関連が深い21種類の慢性疾患が指定されて、追跡調査が可能となっている。

現在、NDBから提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易い形式に加工して提供している。そのために、汎用性のある基本データセットを事前作成しておくことは、提供件数の増加につながると共に厚生労働省および研究者の負担軽減にもつながると考えられる。

E. 結論

NDBの集計データは、国民医療費と同様

の傾向を示したことから、医療費の集計データとしては一定の精度が保たれていると考えられた。ID に関しては、被保険者証番号には半角・全角や氏名に関しては漢字表記の揺れ等の理由で、実際には同一の被保険者にも関わらず、複数の ID が発生している可能性がある。そのために、試行的に、ID1 と ID2 の特性を生かし、複数紐づくものに関しては同一人物とみなして、新しい ID3 も試作した。また、調剤レセプト単独で存在するレセプトが存在しており、特定健診データとレセプトデータとの突合率は約 2 割であった。

本年度は、データ提供日から報告書作成までの期間が限られたことから、データの検証作業に多くの労力を費やしたが、基本データセットについては、詳細なレコードフォーマットを引き続き更新していくものとし、厚生労働省・保険局と情報共有すると同時に、ホームページなどでも情報提供をすることも検討する。

利用者へ個人情報が特定されない範囲で、対象とする疾患の人数、医療費の傾向などの情報を提供し、研究計画を企画・立案できる方式について検討したところ、ウェブサイトにおいてパラメータを指定することでブラウザ上に、対象とする疾患の人数などが表示される機能を提供するダッシュボード機能は、研究計画書を作成する際の有益な情報提供手段となると考えられた。

現在、NDB から提供されるデータセットは、厚生労働省側で複雑な構造の電子レセプトを研究者の要望に応じて、ある程度分析し易い形式に加工して提供している。汎用性のある基本データセットを事前作成しておくことで、提供件数の増加、厚生労働省および研究者の負担軽減にもつながると考えられる。

F. 研究発表

該当なし

G. 知的所有権の取得状況

該当なし