公募ワークショップ
# 公募ワークショップ2
# 日本での EHR-Phenotypingとアウトカム指標のバリデーション
2018年11月23日(金) 14:20 〜 15:50 C会場 (4F 411+412)

[2-C-2-1] データ駆動型医学研究における Phenotypingの重要性
○中島 直樹 （九州大学病院メディカルインフォメーションセンター）

レセプト電算化に続き、電子カルテ化、特定健診の普及などが進み、保健医療領域にも大量のデータが蓄積しつつある。また、厚生労働省はがんゲノム医療中核拠点病院事業を立ち上げ、 Precision Medicineをがん領域で推進し始めた。つまり実医療データ（ Real World Medical Data, RWMD）としてゲノム情報も蓄積し始めた。さらに今後は、 IoT（モノのインターネット）を含めて医療施設外での生体情報の電子化も進むであろう。この蓄積した RWMDを適切に2次利用して医学研究に用いる手法をデータ駆動型医学研究という。

データ駆動型医学研究を進めるにあたり、データ品質の上での大きな２つの課題が、１）標準コードへの適切なマッピングが進んでいないこと、２） RWMDに正確な Phenotypeがタイムリーに記録されていないこと、である。後者は、現在の電子カルテ、およびレセプトシステムが保険病名ベースで開発が進められてきた弊害であり、根本的な是正には制度改正を含めた大きな改革が必要である。しかしながら、既に蓄積した RWMDを活用するために、 Phenotyping手法を用いて可能な限り精緻な Phenotype を描出することが有効である。その目的で日本医療情報学会では、2016年に課題研究会として ePhenotyping研究会を発足した。

自験例を述べると、ある医療施設では1型糖尿病の保険病名で患者を抽出し専門医がレビューしたところ、真の1型糖尿病は55%であった。もし遺伝子研究などの目的で1型糖尿病をカルテから保険病名で抽出したら、45%は偽症例が混入するわけである。このことから、現在既に蓄積しているデータを少しでも精緻な状態で、あるいは精度は低くとも正確に把握した上で、データ駆動型医学研究を進めるべきことが理解できる。本発表では、具体的な RWMDからの Phenotyping開発手法を含めて発表する。陽性的中率を上げるために機械学習を用いて網羅的に Phenotypingのアルゴリズムを構築する一方で、できる限り感度を正確に把握するために、構造化データのみならず非構造化データも用いて真の症例を抽出し解析している。

日本での EHR-Phenotyping とアウトカム指標のバリデーション

平松達雄[*1]、小出大介[*2]、宇山佳明[*3]、中島直樹[*4]
*1 東京大学医学部附属病院、*2 東京大学大学院医学系研究科、
*3 独立行政法人医薬品医療機器総合機構、*4 九州大学病院

EHR-Phenotyping and Validation of Outcome Definition in Japan

Tatsuo Hiramatsu[*1], Daisuke Koide [*2], Yoshiaki Uyama[*3], Naoki Nakashima[*4]
*1 The University of Tokyo Hospital, *2 Graduate School of Medicine, The University of Tokyo,
*3 Pharmaceuticals & Medical Devices Agency, *4 Kyushu University Hospital

When analyzing health insurance claims data or electronic health records (EHR), it is often necessary to judge the disease condition without clear description in the data. For example, even if exact information for requesting medical fee is stated in the claims data, it does not necessarily mean that the state of the patient is medically represented appropriately. EHR data also often do not contain required information as it is unlike patient registries. In order to solve this problem, the state of the required information is determined from other items in the data. This is called EHR-Phenotyping in medical informatics field, and almost the same thing is done in pharmacoepidemiology field where DB research progresses actively. The validity evaluation of the determination is called Validation of Outcome Definition in pharmacoepidemiology field. In this article, we will explain the activities of the four research groups on the theme of these initiatives and examine the essence and future issues of EHR-Phenotyping and Validation of Outcome Definition. The biggest challenge of these efforts is difficulty of obtaining the correct information set for evaluating the determination and it is desirable to solve the problem at an early stage in the data flow process.

Keywords: Real World Data, Phenotyping, Outcome Validation, Common Data Model.

1. はじめに

レセプト電算化に続き、電子カルテ化、特定健診の普及などが進み、保健医療領域にも大量のデータが蓄積しつつある。さらに今後は、IoT（モノのインターネット）を含めて医療施設外での生体情報の電子化も進むであろう。

これら蓄積された実診療データを二次利用するための準備は3段階のデータの流れ（データフロー）に区分できる。最初の段階（ステージ1）は診療情報の発生時におけるものであり一次利用と共通の段階である。ここでは二次利用時に必要になる情報を欠落させずに記録しておくことが重要だが、実診療や病院業務においてはその手間とコストが課題となる。次のステージ2では記録されたデータを二次利用できるように整理してDB等に格納する。データ品質の担保やデータ形式、データ表現の標準化はこの段階で行われる。ステージ3は蓄積整理されたデータを分析等に活用するためのデータの適切な解釈の段階である。各データ列の由来を把握し、そこで表現されている意味範囲を解釈する。EHR-Phenotyping やアウトカム指標バリデーションはステージ3における取り組みとなる。（図1）

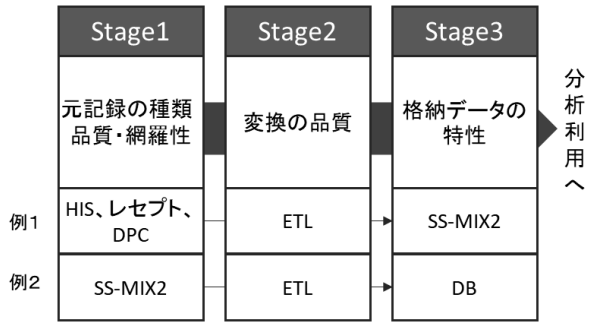

図 1 診療データを二次利用するための3段階

レセプトや診療記録のデータを利用する時に、データに明確な記載のない病態や患者情報を判断する必要がしばしば生じる。たとえばレセプトには診療報酬の請求のための情報は正確に記載されていても、それが必ずしも医学的に患者の状態を適切に表しているとは限らない。診療記録にも臨床レジストリとは異なり、取得したい情報の項目がそのままは入っていないことも多い。臨床レジストリにおいても、取得していなかったこの情報がほしかったと後になって判明することも、欲しいことは最初からわかっていたが取得項目数の制限であきらめたものなどもある。こういった、欲しいデータが直接的には利用可能なデータの中にはない問題に対応するため、データ中の参考になる他の複数の項目を使って取得したい情報を推定判別することが行われ、これを医療情報学では EHR-Phenotyping と呼ぶ。必ずしも複数の項目からとは限らないが、ほぼ同様のことは国内では MID-NET 等の DB 利用が進む薬剤疫学においても、特にアウトカム（医薬品投与によりもたらされた状態）の情報に対して行われ、その妥当性評価はアウトカム指標バリデーションと呼ばれる。

EHR-Phenotyping（以下 Phenotyping と省略する）はゲノム医学研究の一環として 2000 年代後半に米国で誕生した。[1)] 遺伝子情報は試料が一旦取得できれば、あとは組織的機械的に続々と得られる時代となってきたのに対し、臨床情報は患者一人一人について医療者が個々に状態を判断してレポジトリに登録する方法のままでは、情報蓄積速度に圧倒的な違いが生じてしまいボトルネックとなる。臨床情報も素早く大量に得て遺伝子情報とリンクさせてその関連を統計的に解析する研究を進めるために、既に存在するカルテ情報から人手による判断より多少荒くても自動的に臨床状態を判別することができないだろうか、という発想がもとになり取り組みが始まった。Phenotyping 自体はゲノム医学的手法に基づくものではなく、応用先のひとつとしてのゲノム医学分野からの誕生由来であるために「phenotype」というゲノム分野を想起させる名称がついている。

薬剤疫学分野でのアウトカム指標バリデーションは、臨床試験（治験）を行う分野であることが背景にある。臨床試験ではその結果の保証のために、試験で扱われるあらゆるものに対して信頼性の確保が求められる。同様に臨床試験ではなくDB を使う場合でも、分析結果は医薬品の規制等に用いられるものでありデータの信頼性を担保することは当然とされる。そのような分野において診療情報から有害事象等のアウトカムを抽出する方法にも信頼性・妥当性の評価は必須となる。臨床試験の分野では用いるコンピュータシステムが正しく開発・導入・運用されることを確認するコンピュータ化システムバリデーション(CSV)という用語があるが、アウトカム指標バリデーションは CSV とは異なるもので、システムではなくデータから条件式等のある定義により導かれた患者区分の妥当性を評価するものである。有害事象ではないが大規模な診療請求データからの有病者定義を対象にカルテや臨床検査値と比較する研究は世界的には 2000 年代初頭から行われてきている。[2)]

薬剤疫学分野で扱う有害事象も患者の状態の一つであり、Phenotyping で主に扱う病名等と本質的に違いはない。診療情報に存在する病名が必ずしも信頼できないことだけでなく、診療録の元来の性質として、分析のため後日着目することになった病名や患者の状態に対してその有無が明示的に記載されているわけではない。欲しい情報があるのに当てになる明示的な記載はないという課題を、データセット中の他の情報を駆使して判別することにより解決したいという作業上の目論見は両者同じである。とはいえその背景の違いから、実際の取り組み内容には大な違いがみられる。本稿では、国内で取り組みが進む研究班の各活動を解説し、そこから読み取れる Phenotyping とアウトカム指標バリデーションの本質と広がり、今後の課題について検討する。

2. 国内で活動する Phenotyping とアウトカム指標バリデーションに関する研究班

2.1 日本医療情報学会・ePhenotyping 研究会（まとめ研究）

蓄積した実医療データ（Real World Medical Data, RWMD）を適切に 2 次利用して医学研究に用いる「データ駆動型医学研究」が進展している。データ駆動型医学研究を進めるにあたり、データ品質の上での大きな2つの課題が、1）標準コードへの適切なマッピングが進んでいないこと、2）RWMD に正確な Phenotype がタイムリーに記録されていないこと、である。後者は、現在の電子カルテ、およびレセプトシステムが保険病名ベースで開発が進められてきた弊害であり、根本的な是正には制度改正を含めた大きな改革が必要である。しかしながら、既に蓄積した RWMD を活用するために、Phenotyping 手法を用いて可能な限り精緻な Phenotype を描出することが有効である。その目的で日本医療情報学会では、2016 年に課題研究会として ePhenotyping 研究会を発足した。

たとえば、ある医療施設では 1 型糖尿病の保険病名で患者を抽出し専門医がレビューしたところ、真の 1 型糖尿病は55％であった。もし遺伝子研究などの目的で 1 型糖尿病をカルテから保険病名で抽出したら、45％は偽症例が混入するわけである。このことから、現在既に蓄積しているデータを少しでも精緻な状態で、あるいは精度は低くとも正確に把握した上で、データ駆動型医学研究を進めるべきことが理解できる。

陽性的中率を上げるために機械学習を用いて網羅的に Phenotyping のアルゴリズムを構築する一方で、できる限り感度を正確に把握するために、構造化データのみならず非構造化データも用いて真の症例を抽出する方法を進めている。

また、1 型糖尿病、およびインスリン依存に陥った 1 型糖尿病に対して電子カルテ DB およびレセプト DB を用いた Phenotyping 開発を行い、その結果を用いて NDB から全国における有病者数の推定をおこなった。同様に、末期腎不全患者調査を利用して血液透析者数、腹膜透析者数、腎移植症例数などの Phenotyping 研究を行った。

AMED 事業「MID-NET を用いた医薬品等のベネフィット・リスク評価のためのデータ標準化の普及に関する研究（2016～2018 年度・研究開発代表・中島直樹）」により、本邦において厚生労働科研、MIHARI 事業、MID-NET 事業などで考案された Phenotyping 事例の収集を継続し、42 の Phenotyping 事例を収集し年次研究報告をおこなっている。今後 Web サイトで日本において開発されたアウトカム定義リストとして公開する予定である。

2.2 薬剤疫学会・レセプト情報から得られる指標のバリデーションに関するタスクフォース（まとめ研究）

医療情報 DB から得られる情報、特に傷病名の正確性をより適切に評価するためのバリデーション研究について、日本の環境下でいかに実施するかについての指針を与えることを目的に、日本薬剤疫学会にて 2016 年 7 月にタスクフォース(TF)が設立され、その報告書が 2018 年 5 月に公表された。

当 TF では過去のバリデーション研究の総括（100 余りのバリデーション研究を検討）、日本で実施可能なバリデーション研究の問題点・有り方の明確化、バリデーション研究の実施に必要な事項（手順・チェック項目）の洗い出しをおこなった。

バリデーションとは、アウトカム（や研究対象集団や交絡因子）の定義をゴールドスタンダードと照らし合わせ妥当性を確認することである。それぞれの疾患毎に妥当性を検証してみないことにはどの疾患が研究のアウトカムとして使えるかわからない。正しい相対リスクを求めるためには、アウトカム定義の高い陽性的中度が必須である。

教科書では Strom の Pharmacoepidemiology 第 5 版にバリデーション研究に関する記載が見られた。ガイドラインでは、バリデーション研究の方法が部分的に言及されているものは見出されたが、バリデーション研究それ自体を主たる対象にするものは見出されなかった。

バリデーション研究実施の手順・チェック項目 に関する結果としては、以下のようであった。(1)診療報酬請求データ、電子カルテ情報のほか多様な DB が対象にされていた。(2)国・地域・被保険者集団を包括する DB からランダムサンプリングした患者を対象に行った”population-based”と考えられるバリデーション研究も見られたが、少数の医療機関を受診した患者集団や研究者が独自に設定した集団において実施されたバリデーション研究も散見された。(3)ICD コード単独、または薬・検査・診療行為との組み合わせなどがアウトカム定義に用いられていた。(4)リンケージは多くの研究でアウトカム定義の評価に必須の要素であり、多くの研究で利用されていた。(5)ゴールドスタンダードにはカルテレビューのほか、疾患登録などが使われていた。(6)サンプル方法、サンプルサイズは何をゴールドスタンダードとするかに強く依存していた。(7)感度、特異度、陽性的中度、陰性的中度の全てを求めているもののほか、陽性的中度のみを評価したものも多かった。(8)指標の閾値を示したものはほとんどなく、少数、特定の研究の目的を達成する上で必要な「感度○%以上の指標を見出

すことを目標とした」などの記載が見られた。

日本においては、バリデーション研究において重要な役割を果たすリンケージ（照合）は一般に困難であるため、病院（診療所）単位で、保管されている過去の診療報酬明細書（レセプト）を院内のカルテ情報や院内疾患登録などとリンケージ（照合）する病院（診療所）ベースのバリデーション研究が中心とならざるをえない。バリデーション研究にあたっては、DPC データまたは DPC レセプトの傷病名については医科レセプトの傷病名よりも特異性が高いことが期待され、今後のバリデーション研究でその妥当性が評価されるべきであろう。また、病院（診療所）単位のバリデーション研究以外の方法として、大規模コホート研究における国保レセプトを用いたバリデーション研究も可能な方法と考えられる。

バリデーション研究はデータベース研究の質を大きく高める（特に、より適切な比較を可能にすることを通じて内的妥当性を高める）ことを可能にする。質の高いデータベース研究は、医薬品などの安全性・有効性に関して、信頼性が高くかつ他では得難い知見の獲得を可能とし、その結果は国民の健康と福祉に貢献しうる。現状、国内でのバリデーション研究の実施は困難を伴う場合が多いが、今後、医療情報 DB の適切な利用を進めようとする者は、データベース研究の有用性を高めるためにも、医療情報 DB の二次利用に対する国民的理解を得ながら、バリデーション研究を実施しやすい環境の向上に寄与することが求められる。

2.3 MID-NET データの特性解析及びデータ抽出条件・解析手法等に関する研究（実施研究）

MID-NET は全国 10 拠点 23 病院に御協力いただきながら構築を進めてきた医療情報データベースで、処方、病名等の他、臨床検査結果を利活用できるという特徴を有しており、リアルワールドにおける医薬品の副作用発現状況を客観的に評価できる手法として期待されている。本データベースは、品質管理等を実施した結果データの高い品質が確認され、平成 30 年 4 月より本格運用を開始した。

データベースに基づく調査においては、副作用等の目的とするイベントを適切に同定するため、イベント定義が重要であり、処方、病名、臨床検査等を組み合わせて定義されることが一般的である。現在、日本医療研究開発機構（AMED）の研究班「MID-NET データの特性解析及びデータ抽出条件・解析手法等に関する研究」において、MID-NET を活用した適切な調査実施を促進するため、副作用として検討されることが多いイベントに対して適切なアウトカム定義を設定するための研究を進めている。

本研究では、より妥当性の高いアウトカム定義を効率的に作成するための機械学習法を加えたアウトカム定義の作成とその作成プロセスを確立する研究、複数の医療機関で利用可能なアウトカム定義と医療機関毎に妥当性に大きな差異が認められる場合の要因の検討、病院情報システムデータとしての MID-NET、保険請求のデータベースとしてレセプトの類似性や差異からの医薬品の安全対策における複数のデータベースを用いた調査の有用性や留意点等の検討などをテーマとしている。本研究班での成果は、MID-NET の適切な利活用を促進するだけでなく、アウトカムバリデーション研究の一つのモデルになることも目指している。

2.4 レセプトデータから phenotyping を行う各種方法の評価（実施研究）

多くの研究者や実務者が利用可能なレセプトデータを対象にして、データからは明確な識別が困難な状態の患者群を判別抽出しようとするときの良好な抽出アルゴリズムを選択するための指針となり得る基礎的な知見を見出すことを目的に研究を行なっている。抽出アルゴリズムの評価には正解セットが必要となるが、人手で正解セットを用意するのは作業が膨大になるため、Phenotyping やアウトカム指標バリデーションの研究がなかなか進まない原因になっている。本研究では、検査値でほぼ判断できる病態もしくは検査値異常そのものを判別対象に限定することで、電子カルテ由来で入手できる臨床検査値を活用して、Phenotyping の検証実施をコンピュータ自動化できるようにすることを大きな特徴としている。これにより大量の Phenotyping を試行評価し、その集積により Phenotyping そのものの基礎的な特性を明らかにしようというものである。応用目的で実践される Phenotyping やアウトカム指標バリデーションからのまとめ知見というよりも、実データを使った実験室研究に近い位置にある。

抽出アルゴリズム研究は、個々の医療施設特有の事情によるバイアスを避けるため複数の医療施設のデータで行なう必要がある。レセプト全体つまり参加施設の全受診者を対象とする研究であるため、各施設から全データを1箇所に集めるのではなく、複数施設で同一の抽出アルゴリズムを実行して結果を比較検討する方法を採用した。このためには同一プログラムにより各施設のデータを対象に同一の抽出動作を行なうための共通基盤が必要となるためその整備をおこなった。

共通基盤開発にあたっては仮想化機能を利用することで各医療施設へ容易に配布でき、医療施設の既存の PC 設備でも実施しやすいものとなるよう心がけた。データ形式については Phenotyping 特化ではなく、汎用の分析二次利用向けに診療データを格納するための国際的に研究利用されている形式(OMOP Common Data Model)を、日本国内での利用に適合させた中間段階を定めて使用している。そのため将来の国際研究にもつなげることができる。SS-MIX2 にフルデータが格納されていない医療施設が現状かなり多いため、臨床検査値の取得には院内 DWH から取り出した csv ファイルを利用できるようにも対応した。検査結果単位で外字が使われている場合の通常文字への変換も可能である。院内での診療情報管理部門と利用部門とが異なる場合に対応するため、レセプトの形式はそのままで匿名化のみを行う機能も実装した。これらの対応の結果、レセプトだけでなく SS-MIX2 や DWH もデータソースとして利用できる分析のための汎用の共通基盤となっており、大病院だけでなく検査結果値が扱えるタイプのレセコンがあるクリニックまで、広い範囲の医療施設で利用できるポテンシャルがある。本基盤は Phenotyping の発展に協力していただける医療施設を対象に今後公開予定である。

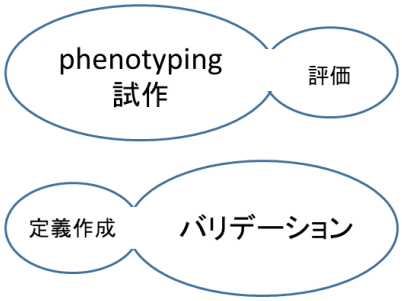

図2 取り組みによる重点の違い

3. 考察

４つの具体的研究班の活動から Phenotyping やアウトカム指標バリデーションについて特徴を整理すると次のようになる。Phenotyping は分析対象のデータセットに直接的には記載がない、あるいは十分な記載がない情報を、記載がある情報から計算することにより判別する。アウトカム指標バリデーションはアウトカム定義に基づいて使用する判別値が真の値に対して妥当なものであるかを評価する。Phenotyping も正しく判別できるかどうかの評価は行う。したがって、データセットの記載あるいはそこから導いた判別を、データセット外から他の方法で得られる正解（より正しい値）と比較して精度の検証を行うという一連の流れはどちらも同じである。

アウトカム指標バリデーションは使用するアウトカム定義の評価を行うことに焦点をあてている。正しさの評価を目的とするものであり、判別がアルゴリズムによるものか、単独の項目なのか等は問わない。100%正確な判別は存在しないため、高度なアルゴリズムを追い求めるよりも、利用目的を考慮した適正な判別であることに力点がおかれている。一方、Phenotyping は大量の EHR データから判別する機能自体に焦点を当てており、そのため判別アルゴリズムの高度化や一連の作業の自動化が目的になりやすい。薬剤疫学と医療情報学のそれぞれの分野の特徴がよく現れている。（図2）

Phenotyping も DB からのアウトカム指標バリデーションも、医療ビッグデータの時代になり登場した概念ではあるが、そこで行われている判別行為に類似のことは、以前から既に何かしら行われている。たとえば臨床診療において患者の状態の評価のために、複数の項目への当てはまり具合を合計するスコア計算方式は広く利用されている。また、インフルエンザの広がりを推定するためにタミフル等の処方の情報を使うことも類似の行為であろう。ケース定義のバリデーションとしては、公衆衛生学分野で疫学調査からの該当者定義の妥当性研究が 1990 年代初頭には行われている。[3)4)] そもそも患者抽出を行うのに複数の条件を適用する方法は、DB が登場する以前から一般に行われている手法であろう。そういう意味では両者とも新たな行為を行っているというよりも、医療ビッグデータに適合して再包装された概念だと言える。

どちらの取り組みも主に病名等の患者の医学的状態が判別対象になっているが、それ以外の事項の判別にも利用できる可能性がある。現在 DB 利用できる診療情報はオーダや実施の羅列記録であり、そのような診療行為が行われた背景や診療コンテキストに関する情報がないため、記載がある情報の意味判断がつかないことがある。今後カルテ文書を自在に利用できるようになったとしても、コンテキストを理解するための十分な情報は記載されていないことも多い。こういったことにも Phenotyping の手法が応用できる可能性はないだろうか。

最大の課題は、判別結果の評価のための正解セット（より正しいとわかっている判別）の取得が難しいことである。たとえば、カルテを見ての判断を正解とするのであれば、該当分野に詳しい医療者が個々のカルテ調査を行う必要があり、大量の正解づけは手間的に困難となる。別の例では、他の DB に正解があることがわかっていても、2つの DB のリンケージが技術的あるいは制度的にできない場合が多いことがあげられる。この状況の中では ePhenotyping 研究会で紹介された 1 型糖尿病の例は巧みな方法であり、たとえば疾患レジストリ等で正解がわかる一部のデータが存在する場合に、そのデータでアルゴリズム開発と評価をおこなったあと、NDB のような全体データに適用する。ただし正解がわかるデータの代表性には注意が必要である。

こういった正解セットの取得を容易にしていく取り組みは重要である。ただし本当に容易になれば正解セットの項目を目的の情報として使用すれば良いため Phenotyping は必要なくなる。逆にいうと Phenotyping が必要な状況はすなわち正解取得が難しい状況であるとも言え、Phenotyping にとっては正解セットの取得は課題でありつづける宿命である。正解セットの取得容易化は Phenotyping の課題解決というよりも研究負荷の低減あるいは信頼性がより高いデータの利用という、より大きなメリットになる。本稿の最初で触れた3段階のステージ1（データ発生時）での対応によりデータセットの中に最初から正解情報が含まれるようにすることで Phenotyping が必要な状況を減らしていくことができ、本質的には望ましい方向性である。とはいえ将来必要になるすべての情報をステージ1で記録しておくこともやはり不可能で、Phenotyping が位置するステージ3とのバランスを取りながら進めていくことになるであろう。

4. まとめ

４つの研究班の活動を解説し Phenotyping やアウトカム指標バリデーションについてそれぞれの特徴や共通点を整理した。両者とも判別結果の評価のためのより正しい値の取得が課題であり、可能なものについてはデータ発生時に取得され伝達されるようにデータフローを整備することが望まれる。

謝辞

本研究の一部は下記の各助成を受けたものです。

厚生労働科研「1型糖尿病の実態調査、客観的診断基準（H29-循環器等-一般-006）」（研究代表者・田嶋尚子）

AMED 事業「MID-NET を用いた医薬品等のベネフィット・リスク評価のためのデータ標準化の普及に関する研究」（2016～2018 年度・研究開発代表・中島直樹）

AMED 事業「MID-NET データの特性解析及びデータ抽出条件・解析手法等に関する研究」（2017～2019 年度・研究開発代表・宇山佳明）

JSPS 科研費 JP15K08710「大規模医療データベースを用いた2型糖尿病における急性膵炎の合併リスクの評価」（2015～2018 年度・研究代表者・小出大介）

JSPS 科研費 JP17K09226「レセプトデータから phenotyping を行う各種方法の評価」（2017～2019 年度・研究代表者・平松達雄）

参考文献

1) Wei WQ, Denny JC. Extracting research-quality phenotypes from electronic health records to support precision medicine. Genome Med. 2015 Apr 30;7(1):41.

2) Leong A, Dasgupta K, Bernatsky S, Lacaille D, Avina-Zubieta A, Rahme E. Systematic review and meta-analysis of validation studies on a diabetes case definition from health administrative records. PLoS One. 2013 Oct 9;8(10):e75256.

3) Katz JN, Larson MG, Fossel AH, Liang MH, Validation of a surveillance case definition of carpal tunnel syndrome. Am J Public Health 1991; 81:189-193.

4) Matte TD, Baker EL, Honchar PA: The selection and definition of target workrelated conditions for surveillance under SENSOR. Am J Public Health 1989; 79(suppl):21-25.