原　著

（独）労働者健康安全機構入院病職歴データベース（ICOD-R）と入院患者レセプトデータのリンケージの試み

金子　麗奈

（独）労働者健康安全機構関東労災病院消化器内科

（2022 年 8 月 8 日受付）

要旨：【目的】近年，ビッグデータ同士のリンケージによって新規知見を得る研究が増加しているが，リンケージ率は低い傾向にある．そこで，労働者健康安全機構の有する病職歴データベース（Inpatient Clinico-Occupational Database of Rosai Hospital Group：ICOD-R）と DPC レセプトデータをリンケージするにあたり，近年模索されている probabilistic linkage 手法を用いることで高いリンケージ率を確保することを目的とした．

【方法】2014 年から 2019 年に登録された関東労災病院の DPC レセプトデータを EF ファイル，様式 1 ファイルとして，ICOD-R と病歴サマリファイルを個人を特定する識別子を有しない状態で取得した．データクリーニング後，ICOD-R と病歴サマリファイル，EF ファイルと様式 1 ファイルをそれぞれ deterministic linkage で統合した．更に両者を probabilistic linkage を用いてリンケージした．

【結果】入手したデータは，DPC レセプト 26,117,438 件，ICOD-R 7,050,800 件，病歴サマリファイル 107,511 件であった．データクリーニング後の DPC レセプト 25,604,413 件のうち 95.1% を占める 24,346,125 件が，最終的に ICOD-R 78,434 件中の 74,497 件にリンケージされた．

【結語】本研究では，ICOD-R とレセプトデータを高い確率でリンケージすることができた．今後，リンケージデータの幅広い利活用の検討が期待される．

（日職災医誌，71：30―35，2023）

―キーワード―

病職歴データベース，ビッグデータ，DPC レセプト

緒　言

ビッグデータの利用可能性が日々急速に拡大する中，複数のデータをリンケージする研究が重要視されている[1)~3)]．しかし，代表的なビッグデータであるレセプト情報・特定検診等情報データベース（National Data Base：NDB）を使用した学術論文でも，アカデミアですら成果を出すのに苦労しており，Impact Factor 等を指標として高い評価を獲得できていない状況がある[4)]．また，研究の発想からデータ取得を行い，研究成果を出すまでにいずれも 4，5 年が経過しており，レセプトデータの入手の問題やデータハンドリングの困難さを浮き彫りにしている[5)]．

（独）労働者健康安全機構が保有するビッグデータである ICOD-R は，1984 年から蓄積された，全国労災病院の 700 万件を超える入院病歴の集積であり，入院患者の入院時より遡って 4 種類の職歴が記録されているという特徴がある．これまで，ICOD-R 単独から解析された職歴と疾病に関する研究結果[6)~8)]が報告されている．ICOD-R と他のデータを突合して得られた知見として，著者らはこれまでに，ICOD-R の匿名性を保持したまま各労災病院の対象患者の病歴を取り寄せ，職業性胆管癌の臨床的特徴について報告した[9)10)]．しかし，これはビッグデータ同士のリンケージではなく，症例レベルでのリンケージであった[9)10)]．一方，著者らによる，ICOD-R と神奈川県地域がん登録とのリンケージデータから解析した職業間の癌罹患後の生命予後の差についての報告は，ICOD-R を他のビッグデータとリンケージした最初の，かつ唯一の報告である[11)]．以後，ICOD-R を用いたリンケージデータによる報告は無い．

2003 年に特定機能病院から始まった DPC（Diagnosis Procedure Combination）は現在では急性期病院の殆どが採用している[12)]．DPC データとは，DPC/PDPS（Per-Diem Payment System）の対象病院に義務付けられてい

る，「DPC 導入の影響評価に関する調査」で集計されるデータを指し，対象医療機関は，退院患者の病態や実施した医療行為の内容を，標準化された形式で厚生労働省に発表する必要がある[13]．DPC レセプトデータは，従来のレセプトデータから改良され，最も医療資源を投入した傷病名，入院契機病名，入院時併存症，入院後続発症の区分など傷病名の区分がされたことで，臨床に即した統計ができるようになった[12]．更に，平成 22 年改訂で患者所在地の郵便番号が付与され，患者の受療動向の把握が可能となった[5]．近年では DPC データによる病院指標の公開も始まり，その価値と有用性は高まり続けている．

この様なレセプトデータに ICOD-R をリンケージすることは，職業歴が医療資源や医療費の分配に与える影響や，入院経路，入院形態の違いについての解析を可能にすると考える．しかし，前述の我々の報告[11]で使用したように，異なる機関，異なる目的で集積されたビッグデータ同士のリンケージは既存の一般的なリンケージ手法（deterministic linkage）[14]ではリンケージ率が低く，リンケージデータの信頼性が課題となった[11]．

そこで本研究では，近年ビッグデータの突合の際に提唱されるようになった手法（probabilistic linkage）を用いて，レセプトデータと ICOD-R をリンケージし，高いリンケージ率を確保することを目標とした．このリンケージデータは，将来に於いて，職業による医療費や在院日数の差，職業と受診経路や退院先の傾向等，職業によって形成される個人の社会生活基盤が，受領する医療にどのような幅をもたらすかに関する解析が可能になると予想する．

本研究は（独）労働者健康安全機構　関東労災病院の倫理委員会にて承認された（2019-26 号）．

方　法

＜対象データ＞

ICOD-R は，1984 年から全国労災病院に入院した患者一人一人について，医療者が確認したカルテ情報（基本的な属性，現病名と既往症，診断，治療，転帰を含む）を集積したデータである．ICOD-R に含まれる性別，年齢，職業など患者背景構成は，日本の全国データに近似している[15)~18)]．入院時のアンケートから，各患者の職歴（現在と直近の 3 つの仕事，開始・終了年齢を含む），喫煙・飲酒習慣などがデータベース化されている[15)~18)]．職歴は，日本標準職業分類と日本標準産業分類を用いてコード化されており，それぞれ改訂に伴い，以前のコードは随時更新されている．アンケート記入に際し，患者から書面によるインフォームドコンセントを得ており，データの登録は診療情報管理士等の登録専門担当者と看護師が行っている．データベースは現在，700 万件を超える．集積されたデータは患者の匿名性を保持するため，生年月日の日を除いた状態で提供される．本研究では 1984 年 4 月 1 日から 2019 年 9 月 30 日までに（独）労働者健康安全機構の入院患者病職歴データベースに登録された全数を入手した．そのうち，2014 年 4 月 1 日から 2019 年 9 月 30 日を入院日とする関東労災病院で登録された症例を抽出した．

DPC データは，様式 1，様式 3，様式 4，D ファイル，EF 統合ファイルで構成されるが，データ分析に最も重要な役目を果たすのは様式 1 ファイルと EF 統合ファイルである．様式 1 ファイルは，年齢等の属性や，入退院経路などの入退院情報，病名情報，各種病状のスコアなど，カルテ情報の一部で構成されている．EF 統合ファイル（以下 EF ファイル）は，診療報酬の算定情報で，E（診療明細情報）と F（行為明細情報）が含まれている[13]．DPC レセプトデータは，2014 年 4 月 1 日から 2019 年 9 月 30 日を退院日とする関東労災病院の入院データを，様式 1 ファイルと EF ファイルの形式で girasol®（株式会社　ヒラソル）を用いて抽出した．

＜リンケージの方法＞

リンケージ手法は，deterministic linkage と probabilistic linkage を使用した[1)19)]．

両方法とも，変数の組み合わせでデータ内の全症例を個々に識別し得なければならいため，生年月日や入院時年齢が必須の変数となる．しかし，レセプトデータに含まれるのは入院時年齢のみであり，生年月日を含まない．従って，リンケージには ICOD-R に入院時年齢のデータを作成することが必要であった．入院時年月日と生年月から入院時年齢を作成することになるが，ICOD-R は，生年月までの情報に限られるため，±1 の誤差が生じる場合がある．そこで，ICOD-R に生年月日の「日」の情報を加えるため，年齢以外の変数を用いて ICOD-R にリンケージ可能であり，かつ生年月日を含むという条件を満たした病歴サマリファイルを利用することとした．病歴サマリファイルは，退院時に作成される退院時サマリの情報の一部である．このうち，入院年月日，退院年月日，性別，診療科，郵便番号のみを診療情報管理室にて加工し，提供を受けた．

レセプト EF ファイル，様式 1 ファイル，ICOD-R，病歴サマリファイルのリンケージ手順は以下のとおりである．

① EF ファイルと様式 1 ファイルは，“データ識別番号（レセプト用にランダムに振付されている 1 患者 1 個の番号）”＋“入院年月日”で分析用 ID を作成し，この ID の完全一致で deterministic linkage を行った．

② 病歴サマリファイルと ICOD-R は，入院年月日，退院年月日，郵便番号，性別，診療科の変数を全て一致させて deterministic linkage を行った．病歴サマリファイルの生年月日を入院年月日より減じ，365.25 日で除したもので新たに入院時年齢を作成した．

③ ①（EF ファイルと様式 1 ファイルの突合したもの）

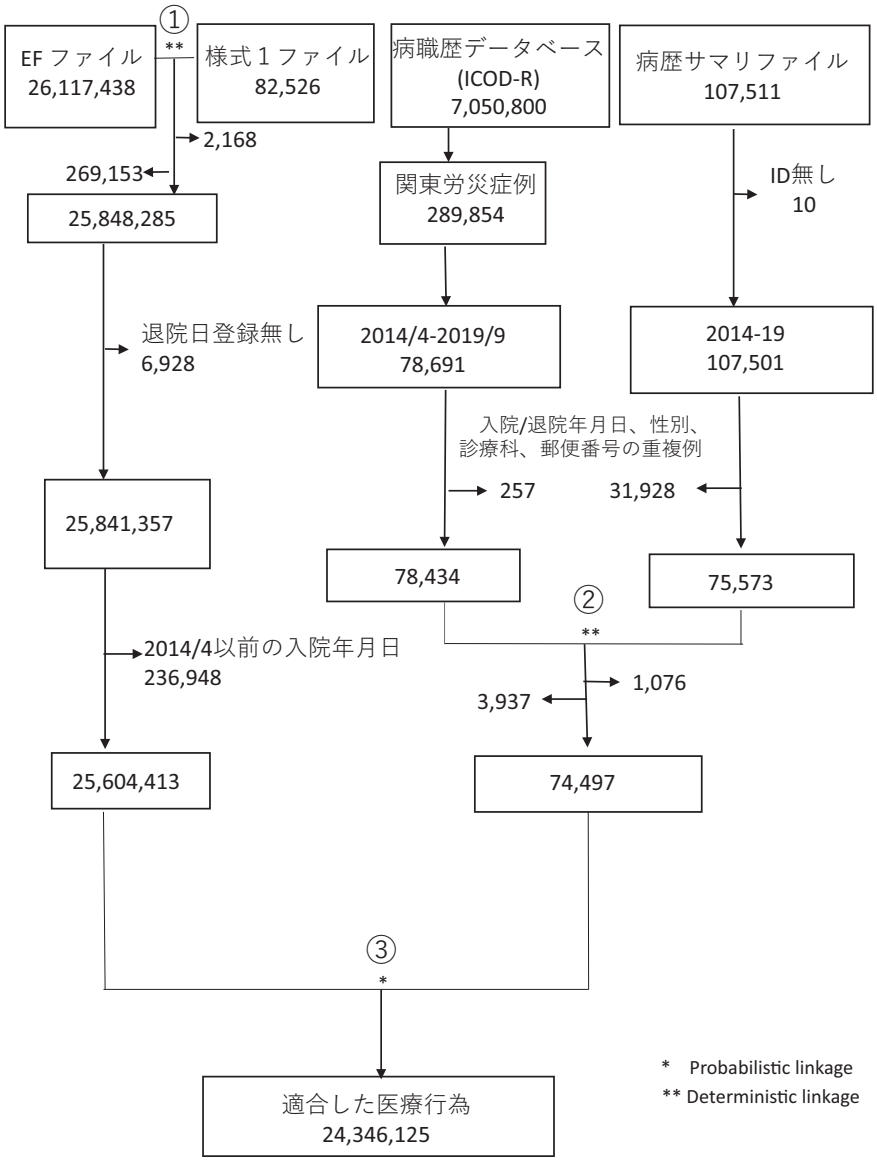

図 1 Linkage のフローチャート

と②(ICOD-R と病歴サマリファイルの突合したもの)のリンケージは，性別，郵便番号，診療科の完全一致の上，入院年月日，退院年月日，年齢を用いて probabilistic linkage を施行した．入院年月日，退院年月日は caliper 0 とし，年齢は，②で作成された年齢が除算した解となるため小数点の端数が生じることから，レセプトデータで登録された整数の±1 歳を合致範囲とし，caliper 2 を採用した．すなわち，性別，郵便番号，診療科が一致した上で，かつ「入院年月日の一致」「退院年月日の一致」「年齢の前後 1 歳の誤差での一致」を各 1 点とし，不一致の場合を−1 点とした場合，各マッチペアに 3 点(全項目合致)，1 点 (2 項目合致)，−1 点 (1 項目合致)，−3 点 (全項目非合致)のスコアリングがされ，3 点と 1 点のペアを採用した．

分析は Stata VER15 (Stata Corp. Texas, USA) を用いて施行した．

結　果

Linkage 作業の結果を図 1 に示す．ICOD-R は，対象期間において入院病歴約 705 万件を有した．そのうち，関東労災病院で登録されたものは 28 万 9,000 件であった．病歴サマリファイルは 107,511 件を取得した．レセプトデータは EF ファイル約 2,612 万件，様式 1 ファイル 82,526 件を取得した．

EF ファイルと様式 1 ファイルでリンケージできたものは，約 258 万件であり，EF ファイルの 99.0% が合致した(以下 EF−様式 1 ファイル)．EF−様式 1 ファイルは，退院年月日が 2014 年 4 月 1 日以降を対象として抽出したため，入院年月日では 2013 年 9 月 14 日から含まれており，EF−様式 1 ファイルのうち，2014 年 4 月 1 日以前の入院に該当する 236,948 件を除外した．

ICOD-R から，入院年月日 2014 年 4 月 1 日から 2019

年 9 月 30 日の期間に関東労災病院で登録されたうち，重複登録例を除外すると，78,434 件となった．同期間の病歴サマリファイルから重複登録例を除くと 75,573 件となった．両者の deterministic linkage の結果，74,497 件が一致した（以下 ICOD−サマリファイル）．

最後に，EF−様式 1 ファイルと ICOD−サマリファイルの probabilistic linkage を行い，EF−様式 1 ファイルのうち，95.1% に当たる 24,346,125 件が ICOD−サマリファイルに合致した．すなわち 74,497 件の職歴データに 24,346,125 件の診療行為が紐付いた．

考　察

本報告は，ICOD-R とレセプトデータをリンケージすることで，新たな知見を探索する土台を作成する基礎的作業の工程と結果である．リアルワールドのデータベース同士は，同じ目的で収集されていないため，100% のリンケージは不可能であり[14]，National Data Base などの公的データを使用しても，リンケージ率は 50% 程度に止まる事例もある[20]．リンケージ手段はデータの取得可能性と質により選択される[14]．リンケージの精度を上げるには，作業前の誤登録の入念なクリーニングと，管理上もしくは技術的エラーで集積されたデータの重複の除去が重要である[1)21)]．

レコードリンケージの手段には，deterministic linkage と probabilistic linkage がある[2)19)22)]．前者は，2 つのデータセットの特定の識別子について，直接的にレコードが同じ個人に属するかどうかを決定する方法である[14]．Probabilistic　linkage は変数の一致の強さを考慮して同じ個人に属する可能性を表すスコアをレコードペアに与える手法である[1)3)]．Probabilistic linkage では，直接的に個体を同定するような識別子を有さないレコード同士から，居住地域，年齢，治療施設，入院・処置・退院日など，多くの間接的な識別子を併用することで，一致する可能性のあるペアを探し出してスコア付けする[23]．しかし，この作業は極めて複雑であり，変数名称の不一致や値の欠落によって様々なエラーが生じる[1]．また，合致させるためには，一致作業をする複数の変数によって一つ一つの個体が識別同定できなければならず，そのような変数（識別子）の組み合わせを探す作業から始まる．変数の組み合わせを決定した後，①ペアになる可能性のある 2 個体の組み合わせを全て抽出する　②可能性のあるペアの変数を全て比較する　③全てのペアにスコアリングする　④高いスコアリングのものを残す，という作業を経る．

Probabilistic linkage は柔軟性が高く，症例識別子の誤りや分類ミスがあってもリンケージを可能にする．よって，deterministic linkage よりも，同じ個人に属さないレコードをリンクしてしまう可能性も高い点で注意を要する[19)22)]．

本報告で施行したリンケージ作業に於いて，問題は数多くある．

第一に，有効な入退院年月日の無いもの等，登録時もしくは抽出時の人的，技術的問題がある．これはデータ件数が失われる主な原因となる．本研究は，EF−様式 1 ファイルのうち，6,928 件をデータ内容の不備で排除せざるを得なかった．全体に対するデータ損失の割合は，リンケージデータの validity を低下させるため，データの登録や抽出における人為的なミスを極力減らすような対策が必要である．

第二にビッグデータはリンケージすることを前提として蓄積されていないため，同じ内容に対する変数名や，定義名が異なることが多い．例えば本研究でも，「診療科」は，3 種のデータベースでそれぞれ異なる変数名で登録されており，かつ，科を示す番号も異なっていたため，変換，統一する作業が必要であった．また，一つのデータベースの中でも，登録が長期間にわたる場合，採用されている電子カルテの変更などのため途中で記載形式が変更になっていることもある．例えば，本研究では，患者の病院 ID 番号は提供されていないが，関東労災病院において ID 番号は電子カルテのメーカーが変更になった 2017 年から，それ以前とは連続しないテンプレートとなっている．この様なデータハンドリングの問題に加え，ビッグデータを扱う際は，研究者が使用できる機器の性能の限界が付き纏う．筆者は，リンケージ作業に 3.6GHz 32GB　4core の一般使用としては高性能の PC を使用したがそれでもなお，今後の解析で使用する可能性が低い変数，登録件数が著しく少ない等の多くの変数を削除した上で統計作業を行わないと，CPU やメモリの問題の限界により作業が処理されない．従って，リンケージ作業を完結することを重視し，極めて多くの変数を削除した上で作業を遂行したため，今後の知見を得るための解析時には，必要な変数を適宜追加したデータベースでの作業が必要である．

以上のような限界や課題はあるものの，今回，直接的な識別子を付与せずに提供された 3 つのデータを用い，2,600 万件のレセプトデータの 95% を超える高い合致率でリンケージすることができた．しかし，最終段階で使用したのは probabilistic linkage であるため，異なった個体をリンケージしている可能性は残存している．Probabilistic linkage は，deterministic linkage に劣らず，高い正確性を持っていることが報告されている[20]．しかし，probabilistic linkage は，その手法を詳細に解説をしているものも少なく，体系的に記載された書籍も見受けられない段階であるため[1]，リンケージの精度について validity の評価が個々の研究で必要となる．リンケージエラーを定量化するためのアプローチは 1）ゴールドスタンダードサンプルとの比較　2）感度分析　3）リンクしたデータとリンクしていないデータとの比較　4)疑わしい

マッチの特定などがあげられる[1]．本リンケージでは，解析の③probabilistic linkage を行った段階で，別途に，入院年月日，退院年月日，性別，診療科，郵便番号を用いた deterministic linkage を行っており，19,744,180 件（77.1%）の合致を得た．この場合，合致したものはリンケージエラーを伴いにくいため，今後，リンケージデータを使用して行う研究の際に，deterministic linkage 後のデータを用いて同様の分析を行うことで，2)の感度分析とすることが望ましいと考える．今後，本リンケージデータを使用して職歴に関する入院医療の格差に関する知見を得る解析が期待される．

結　語

労災病院入院病職歴データベースを，レセプトデータベースと高い確率でリンケージすることができた．このような他ビッグデータとのリンケージは，ICOD-R の新たな利活用の場を提供するものと考える．

謝辞：本研究で使用したレセプトデータは，関東労災病院事務部門の経営企画課 鎌柄友子氏が快諾の上，DPC 分析ソフト等を用いて試行錯誤して抽出して下さった．全件の抽出には途中様々な障壁が存在し，通常業務の傍ら大変なご苦労であったと推測される．病歴サマリデータは，診療情報管理室　宇田嘉子氏により抽出頂いた．この様な貢献を惜しまない人材に恵まれたことは大変有難いことである．両氏に深謝申し上げる．

［COI 開示］本論文に関して開示すべき COI 状態はない

文　献

1）Sayers A, Ben-Shlomo Y, Blom AW, Steele F: Probabilistic record linkage. Int J Epidemiol 45 (3): 954—964, 2016.
2）Blake HA, Sharples LD, Harron K, et al: Probabilistic linkage without personal information successfully linked national clinical datasets. J Clin Epidemiol 136: 136—145, 2021.
3）Bohensky MA, Jolley D, Sundararajan V, et al: Data linkage: a powerful research tool with potential problems. BMC Health Serv Res 10: 346, 2010.
4）厚生労働省：第三者提供の成果物集計について．2017-3-15. https://www.mhlw.go.jp/file/05-Shingikai-12401000-Hokenkyoku-Soumuka/0000155475.pdf,（参照 2022-10-23）.
5）奥村泰之，佐方信夫，清水沙友里，松居宏樹：ナショナルデータベースの学術利用促進に向けて：レセプトの落とし穴. Monthly IHEP（医療経済研究機構　Institute for Health Economics and Policy） 268：16—25, 2017.
6）Fukai K, Kojimahara N, Hoshi K, et al: Combined effects of occupational exposure to hazardous operations and lifestyle-related factors on cancer incidence. Cancer Sci 111 (12): 4581—4593, 2020.
7）Kojimahara N, Hoshi K, Tatemichi M, Toyota A: The relationship of hospital stay and readmission with employment status. Ind Health 59 (1): 18—26, 2021.
8）Kaneko R, Zaitsu M, Sato Y, Kobayashi Y: Risk of cancer and longest-held occupations in Japanese workers: A multicenter hospital-based case-control study. Cancer Med 8 (13): 6139—6150, 2019.
9）Kaneko R, Nakazaki N, Tagawa T, et al: Preliminary analysis of labour effect on genesis of cholangiocarcinoma. Nihon Shokakibyo Gakkai Zasshi 111 (3): 510—511, 2014.
10）Kaneko R, Kubo S, Sato Y: Comparison of Clinical Characteristics between Occupational and Sporadic Young-Onset Cholangiocarcinoma. Asian Pac J Cancer Prev 16 (16): 7195—7200, 2015.
11）Kaneko R, Sato Y, Kobayashi Y: Inequality in cancer survival rates among industrial sectors in Japan: an analysis of two large merged datasets. Environmental and Occupational Health Practice 3 (1): 1—12, 2021.
12）藤森研司：電子レセの弱点を補完する DPC データ. 将来は統一も．全日病ニュース．829．2014-8-1. https://www.ajha.or.jp/news/backnumber/pdf/2014/140801.pdf,（参 照 2022-8-7）.
13）伏見清秀，今井志乃ぶ：すべて Excel でできる！経営力・診療力を高める DPC データ活用術，序章 DPC の基礎知識とデータ活用の意義．増補改訂版．東京，日経 BP 社，2014, pp 10—11.
14）Zhu Y, Matsuyama Y, Ohashi Y, Setoguchi S: When to conduct probabilistic linkage vs. deterministic linkage? A simulation study. J Biomed Inform 56: 80—86, 2015.
15）Zaitsu M, Cuevas AG, Trudel-Fitzgerald C, et al: Occupational class and risk of renal cell cancer. Health Sci Rep 1 (6): e49, 2018.
16）Zaitsu M, Kaneko R, Takeuchi T, et al: Occupational inequalities in female cancer incidence in Japan: Hospital-based matched case-control study with occupational class. SSM Popul Health 5: 129—137, 2018.
17）Zaitsu M, Kaneko R, Takeuchi T, et al: Occupational class and male cancer incidence: Nationwide, multicenter, hospital-based case-control study in Japan. Cancer Med 8 (2): 795—813, 2019.
18）Zaitsu M, Kato S, Kim Y, et al: Occupational Class and Risk of Cardiovascular Disease Incidence in Japan: Nationwide, Multicenter, Hospital-Based Case-Control Study. J Am Heart Assoc 8 (6): e011350, 2019.
19）Méray N, Reitsma JB, Ravelli AC, Bonsel GJ: Probabilistic record linkage is a valid and transparent tool to combine databases without a patient identification number. J Clin Epidemiol 60 (9): 883—891, 2007.
20）Setoguchi S, Zhu Y, Jalbert JJ, et al: Validity of deterministic record linkage using multiple indirect personal identifiers: linking a large registry to claims data. Circ Cardiovasc Qual Outcomes 7 (3): 475—480, 2014.
21）Randall SM, Ferrante AM, Boyd JH, Semmens JB: The effect of data cleaning on record linkage quality. BMC Med Inform Decis Mak 13: 64, 2013.
22）Tromp M, Ravelli AC, Bonsel GJ, et al: Results from simulated data sets: probabilistic record linkage outperforms deterministic record linkage. J Clin Epidemiol 64 (5): 565—572, 2011.
23）Lawson EH, Ko CY, Louie R, et al: Linkage of a clinical surgical registry with Medicare inpatient claims data using indirect identifiers. Surgery 153 (3): 423—430, 2013.

別刷請求先　〒211–8510　神奈川県川崎市中原区木月住吉町1—1
（独）労働者健康安全機構関東労災病院消化器内科
金子　麗奈

Reprint request:
Rena Kaneko
Department of Gastroenterology, Kanto Rosai Hospital, 1-1, Kizukisumiyoshi-cho, Nakahara-ku, Kawasaki, Kanagawa, 211-8510, Japan

An Attempt to Link the Inpatient Clinico-Occupational Database of Rosai Hospital Group (ICOD-R) with an Inpatient Receipt Database

Rena Kaneko
Department of Gastroenterology, Kanto Rosai Hospital

Aim: Recently, an increasing number of studies have been conducted to gain new knowledge through the linkage of large datasets. Therefore, we linked the Inpatient Clinic-Occupational Database of Rosai Hospital Group (ICOD-R) with inpatient DPC receipt data, with the aim of constructing a new database combining information on occupation and medical care that cannot be inferred from the ICOD-R alone.

Methods: Inpatient DPC receipt data from Kanto Rosai Hospital registered from 2014–2019 were obtained, as well as ICOD-R and inpatient medical record summary files, both without personal identifiers. After data cleaning, ICOD-R and medical record summary files, E and F file from inpatient DPC receipt data were merged using deterministic linkage respectively. The two merged datasets were then further matched using probabilistic linkage.

Results: The data obtained included 26,117,438 DPC receipts, 7,050,800 records in ICOD-R and 107,511 in medical record summary files. Of the 25,604,413 DPC receipts after data cleaning, 24,346,125 receipts (95.1%) were ultimately linked to 74,497 of the 78,434 ICOD-R cases.

Conclusion: ICOD-R and receipt data could be linked with high probability. In the future, it is necessary to evaluate the validity of the linkage data and consider a wide range of ways in which the data can be utilized.

(JJOMT, 71: 30—35, 2023)

—Key words—
ICOD-R, big data, Diagnosis Procedure Combination