

医療ビッグデータアナリティクスプロセス —抗がん剤副作用の解明における実践—

石井 一夫¹ 野原 正一郎¹ 柴田 龍宏¹ 小原 仁¹ 宮本 貴宣¹ 角間 辰之¹ 上野 高史¹ 福本 義弘¹

¹久留米大学

医療ビッグデータには、①電子カルテや電子レセプトなどの医療電子ドキュメント、②次世代シーケンサによるゲノム多型解析やゲノム発現定量解析などの医療ゲノムデータ、③画像診断データや医療センサ、患者のモバイル端末などから収集される医療IoTデータなどがある。これらのデータの利活用は一人の専門家の知識・技能によって達成できるものではなく、情報科学や、統計科学、医療従事者などの緊密なコミュニケーションによる共同作業により初めて実現可能となる。本稿では、医療ビッグデータアナリティクスの実践として、約20万人の抗がん剤が投与されたがん患者の医療電子ドキュメント（DPCデータ/レセプトデータ）を用いて、心臓血管系疾患の副作用の実態を明らかにした分析事例を紹介し、ビッグデータアナリティクスプロセスの戦略について検証する。ビッグデータには5V（Volume（量）、Velocity（速度）、Variety（多様性）、Veracity（正確性）、Value（価値））という処理上の留意点が知られている。本研究では医療電子データの特徴からVeracity（正確性）をどうやって担保するかという点が問題となった。本稿ではこの点をどのように克服し、解決していったかということに焦点をあてて解説する。

1. はじめに

現在、医療現場ではICT化および医療情報の活用促進が進行しており、電子レセプト（診療報酬明細書）やDPC（診療群分類包括評価）データ、特定健診（特定健康診査）データ、電子カルテデータなどの診療行為にともなって出てくる医療電子ドキュメント（EMR：electric medical record）が日々大量に蓄積している。また、次世代シーケンサ（大規模DNA自動解析装置）の普及により、日常診療においてもゲノム解析が活用されることが増え、1人の患者から得られる検査などの情報量も格段に増えている。このため、個別化医療を実現化するプレジジョンメディシンの推進も進んでいる。さらには、AI技術を駆使した画像診断や、医療センサ、患者由来のモバイル端末から収集される医療IoTの活用も進んでいる。

（株）富士経済の試算した2025年における市場予測によると医療ビッグデータビジネスの国内市場は2,859億円に達すると見ている[1]。これらのデータは、患者の疾病分析による臨床現場の診療への還元や、新規医薬品の創薬のみならず、保険や病院経営などの医療経済分析や、「地域包括ケアシステム[2]」など、医療政策の立案などにも活用される。

国内においても医療ビッグデータを集めた大規模データベースがいくつか構築されている。公的なものには、厚生労働省の構築した「レセプト情報・特定健診等情報データベース（以下、NDBデータベース）[3]」や「DPCデータベース[4]」があり、商用のものでは、（株）JMDCの構築した「JMDC Claims Database[5]」や、MDV（メディカル・データ・ビジョン）社の構築した「MDV診療データ[6]」などがある。

今回筆者らは、このうちMDV社の構築したMDV診療データを用いたデータ分析を2017年4月から約3年間かけて行った。これは、MDV社が全国に約1,700あるDPC病院のうち291病院から、診療データの二次利用許諾を得て匿名加工処理した上で、データを集積してきた大規模データベースである。すなわち、MDV診療データから抽出された約20万人の抗がん剤の投与を受けたがん患者のデータを分析した。その結果、抗がん剤による心臓血管系疾患副作用の全貌を把握することが可能になった。その臨床的知見については本稿の趣旨から外れるため他の文献で紹介するので割愛し、本稿ではそのデータ分析プロセスを確立する上で問題となった点とその対応策について述べる。

ビッグデータには5V（Volume（量）、Velocity（速度）、Variety（多様性）、Veracity（正確性）、Value（価値））という処理上の留意点が知られている[7]。本稿では、医療ビッグデータの分析プロセスに関し、ビッグデータの5VのうちのVeracity（正確性）の問題に焦点を当てこれをいかに克服し、ビッグデータアナリティクスを遂行したかという点に注力して紹介する。ビッグデータアナリティクスでは、大容量メモリーサーバ（いわゆるHPC）や、並列処理システムによるビッグデータ処理技術、AI・深層学習をはじめとするデータ分析技術等が必要となる。しかし、本事例ではそのような技術的課題に加え、医療行為を運営する上でのEMRの特殊性から、そのドメイン知識が不可欠である。このため、大規模データを処理する情報処理技術者と、データ分析を専門とする生物統計家および、レセプトデータやDPCデータおよび、がんや心疾患などの臨床知識を持つ医師など医療従事者の密接なコミュニケーションによる連携により分析の目的に合ったコードマスタの作成や、用語の定義等を行って初めて可能となる。

本稿では、最初に本研究の分析対象となったEMRの概要と、それらを構成要素とする医療ビッグデータである医療大規模データベースの概要を説明する。次に、データ分析プロセスの内容に関し、医師との連携による用語定義等について医療現場の実情を反映させるために実践した事項を述べる。

医療ビッグデータの分析には、患者の個人情報扱うために患者の人権保護のための倫理的配慮やそれを実施するための法的手続きとその遵守が重要である。しかし、その手続きの確立や法整備は進行途上であり、完全にコンセンサスが取れているとは言い難い。したがって、本稿では関連法案や施設内での倫理委員会の審査を含む手続きなどについては、詳細は触れず、最小限度の記述にとどめる（本稿の第6章を参照）。

2. 医療電子ドキュメント（EMR）

最初に、本事例で利用した医療電子ドキュメント（EMR）の概要を説明する。すなわち、本事例で分析の対象となったMDV診療データのベースとなっているEMRである①レセプトデータ、②特定健診データ、③DPCデータの3種類の医療電子文書について説明する。

2.1 レセプトデータ

「レセプト」とは、医療機関が保険者（市町村や健康保険組合等）に請求する医療報酬の明細書のことである[8]。医科・歯科の場合には診療報酬明細書、薬局における調剤の場合には調剤報酬明細書、訪問看護の場合には訪問看護療養費明細書ともいう。「診療報酬」とは、診療に要した費用のことで、診

療報酬点数表に基づいて点数で算出される。「医療費」は診療報酬点数から1点=10円として金額で算出される。

従来、レセプトは紙ベースで作成されていたが、保険医療機関・保険薬局、審査支払機関、保険者の医療保険関係者すべての事務の効率化の観点から「レセプト電算処理システム」が構築された。現在では、「療養の給付および公費負担医療に関する費用の請求に関する省令」により原則として電子レセプトによる請求を行うこととされている。

電子レセプトは、厚生労働省が定めた規格・方式（記録条件仕様）に基づきレセプト電算処理マスターコードを使って、CSV形式のテキストで電子的に記録される。レセプト電算処理システム[9]には、マスターコードとして、9つの基本マスターファイルがある[10]。①医科診療行為マスター、②歯科診療行為マスター、③調剤行為マスター、④医薬品マスター、⑤特定器材マスター、⑥コメントマスター、⑦傷病名マスター、⑧修飾語マスター、⑨歯式マスターの9つである。したがって、電子レセプトからは、保険者のかかった医療機関や、居住地、年齢、性別などの情報のほか、診療情報などが読みとれる。

2.2 特定健診・特定保健指導データ

特定健康診査（特定健診）・特定保健指導[11]は、2008年度から法律に基づき、健康保険組合等の医療保険者で40歳～74歳の被保険者および被扶養者を対象にメタボリックシンドローム（内臓脂肪症候群）に着目した生活習慣病予防のために実施されている。これらの実施結果を集計データは、都道府県において、医療費適正化計画の策定にかかわる参考データとして活用される。

特定健診の検査項目には、身体計測（身長、体重、BMI、腹囲（内臓脂肪面積））、理学的検査（身体診察）、血圧測定、血液化学検査（中性脂肪、HDLコレステロール、LDLコレステロール）、肝機能検査（AST（GOT）、ALT（GPT）、 γ -GT（ γ -GTP））、血糖検査（空腹時血糖又はHbA1c検査）、尿検査（尿糖、尿蛋白）がある。さらに、生活習慣病リスクの評価、保健指導の階層化などを目的に質問票が活用される。医師が必要と判断した場合には、血液一般（ヘマトクリット値・色素量・赤血球数）、胸部X線、心電図が実施される。これらの検査結果の情報は、患者の病状を反映した生体情報そのものであり、個々の患者の病態や地域集団での健康状態の実態を把握できる。

2.3 DPCデータ

DPCデータとは、「診療群分類別包括払い（DPC）制度[12]」に基づくデータのことである。DPCとは、急性期入院医療を対象とする診断群分類に基づく1日あたり包括払い制度である。2003年4月より特定機能病院を対象に導入された。診断群分類ごとに設定される在院日数に応じた3段階の定額点数に、医療機関ごとに設定される医療機関別係数を乗じた点数を算定する仕組みである。

DPCが導入されて以来、診断群分類の開発をはじめ分類の妥当性の検証やデータの精緻化のための作業が現在も進められている。この基盤となるものが、「DPC導入の影響評価にかかわる調査[13]」で、DPC対象病院はDPC請求を行うと同時にカルテ・レセプト情報のデータを厚生労働省に提出する義務がある。提出する内容は次の3つに分けられる。

- 1) 患者単位で把握する主に診療録（カルテ）からの情報：様式1など（後に表1で示すFF1ファイルはこのデータから抽出したデータセットである）。
- 2) 患者単位で把握する主に診療報酬明細書（レセプト）からの情報：E、Fファイルなど。
- 3) 医療機関単位で把握する情報：様式3など。

DPCデータでは、2.1節のレセプトデータに加えて、より詳細な患者情報の解析が可能になる。

3. 医療ビッグデータのデータベース

先に述べた医療用電子文書は、国や公共団体、病院や保健機関、および企業などにより、データベースに登録され、患者の診療への還元、病院経営、創薬、医療政策立案などに活用される。これらデータベースのうち、本事例で使用したMDV診療データについて以下に説明する。

3.1 MDV診療データ

今回使用したMDV診療データは、データの2次利用許諾を個別に受け、匿名化した全国291施設の、実患者数約1,785万人のDPCデータおよびレセプトデータ（以後、DPCデータ/レセプトデータ）を元に作成した大規模データベースである（2017年2月末現在）。このうちがん拠点病院は129病院（約40%）で、200～499床の病院が46%、500床以上の病院が39%含まれる。

MDV診療データの更新は今も日々継続しており、同社のプレスリリースによると、本稿執筆時の2020年2月時点では、DPC病院のデータ規模は、3,000万人超となっている[14]。さらに、DPC病院のデータに加え、500万人超の健保組合が保有する診療データ（健保データ）の提供も4月から提供するようになるという。これにより、回復期と慢性期のデータも把握できるようになる。たとえば、糖尿病治療の実態調査について、DPC病院と開業医の両方の診療実態が把握できるようになるほか、開業医から病院に流れやすい疾患や病態の把握も可能になるという。

3.2 データの抽出条件とその内訳

本研究の目的のために、MDV診療データに含まれる患者のうち抗がん剤が投与されたがん患者のみを抽出したデータの提供を依頼した。患者は、2008年4月から2017年1月の8年10カ月の間に登録され、登録より前180日以内にがんに関連する受診歴がない合計197,645人分の患者データが含まれていた。

表1に、本データセットのうちデータ分析に用いたファイルの一覧を示す。これは、①ActData.txt、②DiseaseData.txt、③LaboData.txt、FF1.txtおよび、④M_Drug.txtのタブ区切りテキスト形式の4つのファイルから構成されている（以後、.txt表記は省略）。合計で、容量で約34.6GB、エントリ数で約4億件のデータであった。

表1 データ分析に用いたデータセット

| データセット名 | 容量(B) | レコード数 | 説明 |
|-----------------|-------|-------------|---------------------------------|
| ActData.txt | 29.2G | 340,144,953 | 診療情報（検査、手術、投薬など） |
| DiseaseData.txt | 3.7G | 53,594,769 | 診断情報（傷病名など） |
| LaboData.txt | 1.7G | 21,780,361 | 健診データ |
| M_Drug.txt | 1.6M | 16,381 | 医薬品コード対応表 |
| FF1.txt | 59M | 587,271 | DPCデータ:様式1からの抽出データ（身長、体重、喫煙歴など） |
| 合計 | 34.6G | 416,123,735 | |

①ActDataは、本データセットに含まれる患者の投薬、検査、手術などの診療情報をDPCデータ/レセプトデータから抽出したものである。②DiseaseDataは、本データセットに含まれる患者の疾病名情報をDPCデータ/レセプトデータから抽出したものである。③LaboDataは健診データの結果を抽出したものである。④M_Drugは、本データセットで用いられていた医薬品コードのATC分類（解剖治療化学分類）コードとの対応表である。今回の研究においては、臨床医の研究者メンバは、ATC分類をもとに、医薬品の選択を行っている。これは、WHOによって管理されている医薬品の解剖学的部位に基づいた薬効分類コードである。⑤FF1は、DPCデータの様式1等からの抽出データである。カルテからの情報で、患者の身長、体重、喫煙歴などを含む。患者の本データセットの特徴は、①ActDataというファイルが極端に大きく、約3億5千万エントリであり、これをどうやって処理するかが技術的課題であった。他に、投与された145種の抗がん剤のリストが提供された。

4. ビッグデータアナリティクスプロセスの概要と解析チームの構成

4.1 ビッグデータアナリティクスプロセスの概要

図1に、一般的なビッグデータアナリティクスプロセスを示す。ビッグデータアナリティクスでは、大規模な非定型データをHadoopやNoSQLなどのビッグデータ処理用プラットフォームで処理し、SQLで処理できるようなリレーショナルデータベースに格納できるテーブルまで落とし込む。そこから必要なデータを抽出してデータ分析ソフト（R、Python等）でデータ分析を行ってレポート化する。

今回のデータセットは、MDV診療データから条件抽出したサブセットであるため、ビッグデータ処理用プラットフォームを使用することなくリレーショナルデータベースにそのまま格納できるテーブルであった。「ローデータの収集・蓄積」の工程はMDV社で実施しているが、詳細な収集プロセスは、公開されていないので、元となったレセプト電算処理システムや特定健診データ、DPCデータなどの情報を手がかりに、「データ構造の解析」を行い、データ解析に必要なコードマスタの作成を行った。

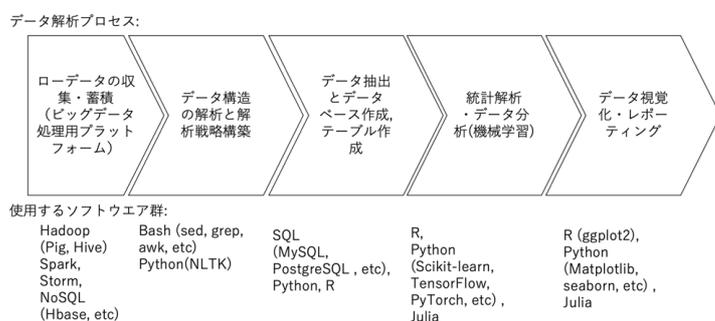


図1 ビッグデータアナリティクスプロセス

4.2 医師、情報技術者、統計家の三者からなるチーム構成

図2に本研究のビッグデータアナリティクスのチーム編成を示した。同じ施設内に勤務する今回の分析対象である心臓血管系疾患副作用に詳しい心臓血管内科医および、情報処理技術者と生物統計家の3者から構成されており、頻繁な連携による共同作業の下で分析を実施した。

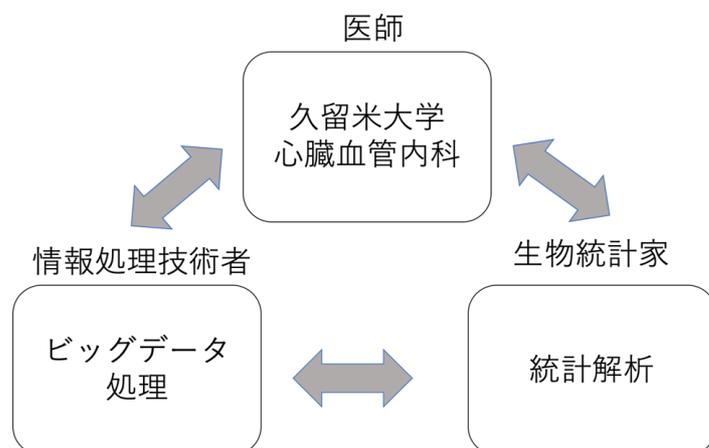


図2 ビッグデータアナリティクスのチーム構成

4.3 データ分析システムおよびセキュリティ

今回のデータは、MDV診療データという巨大なデータベースから抽出した定型データである。3億5千万エントリの大きなファイル（ActData）があるため、これを処理できる大容量メモリのサーバが必要であった。

データの全サイズは、約35GBでありビッグデータとはいえ比較的小さいので、あまり大きなストレージは必要ではなかった。数億エントリのデータをストレスなく処理するために、785GBのメモリを搭載したHP ProLian ML350 Gen9（38コアCPUs、HDD8TB）を調達した。ストレージとして12TBのNASを用意した。

データは、MDV社で使用許諾を得たデータを用いている。使用許諾の際には、利用目的の明確化、利用の制限（目的外の使用や第三者への譲渡禁止）など厳格なポリシーの下で実施している。

セキュリティに関しては、①カードキーで入退室管理および施錠管理された施設の中でデータ処理・分析を行い、②ネットワークアクセスの制限された環境でのみデータを使用し、データや、データを保存した媒体の持ち出しを禁止するなどの対策を施した。

データ処理のソフトウェアは、基本的にオープンソースソフトウェアを使用した。コスト削減と完全に自己管理、制御できるソフトウェアがデータ分析の運営上望ましいと考えたからである。Linux（Ubuntu Linux 16.04）をプラットフォームとし、SQL（MySQL、PostgreSQL）、R、Python、Perl、Ruby、Bash（awk、sed、grep）などや、関連のライブラリ、パッケージを利用した。ビッグデータ処理の場合、ソフトウェアによって適切なデータのサイズが異なる。効率的なデータ処理、データ分析のためには複数のソフトウェアを使用し、比較・評価しながら処理、分析の効率化を図っていく必要があった。

総計解析ソフトは、SASやStataなど、参加した生物統計家の使い慣れた比較的使いやすいソフトを用いている。

4.4 ビッグデータの5Vからみたデータの評価、課題と解決策

ここで、ビッグデータの5Vの観点から、本データ分析について評価をしてみる。本データセットの元のMDV診療データはデータ抽出時点で実患者数約1,785万人（2017年2月末現在）を含み、その後も急速に更新、拡張が進んでいる。データは、国内のDPC病院のあらゆる多種多様な疾患の診療行為を処置された患者のデータが含まれている。①Volume、②Velocity、③Varietyの処理上の留意点を持つ巨大データである。

一方で、本実践で用いたデータセットは、そこから条件抽出したサブセットであり、2017年1月時点で固定されたリレーショナルデータベースの体裁をした構造化データである。ビッグデータであることには変わりはないが、データ分析という点で今回の分析で技術課題となったのは④Veracity（正確性）である。

現実の診療行為によって得られたデータはしばしばリアルワールドデータと呼ばれる[15]。このようなデータは、元来データの分析を目的として収集されたデータではない。したがって、データの収集プロセスや運営形態は、データ分析の目的と乖離がある、その乖離を埋める確かなデータ分析を実施するには、実際に診療にあたっている医師や、保健制度、医療文書に詳しい者でないと判断ができない事柄も多数ある。

最後に、⑤Value（価値）という点であるが、医療ビッグデータの分析は、医療学術的研究、医療経済・経営研究、医療政策提言などに使用されるが、現在はその出力結果は記述統計的な集計データであることが多い。一方で、本研究では医師の助言の下、プログラミングを用いて患者の経時的な詳細情報を抽出し、時系列解析や、多変量解析などの数理モデリング研究も実施できた。簡単な概要は5.5節に紹介するが詳細は、本研究に関与した医師らによる学会発表[16]などを参照されたい。

5. 分析プロセスでの個別の問題点とその解決策

本章では、⑤Veracity（正確性）について、いくつかの問題と処したその解決策を紹介する。

5.1 分析に必要な情報と使用したデータセット

5.1.1 本研究の目的

本研究の目的を再確認すると、「抗がん剤を投与されたがん患者の心臓血管系疾患副作用の実態を明らかにすること」が目的である。

5.1.2 本研究の目的達成に必要な情報

提供されたデータセットから目的達成に必要な情報を抽出する必要がある。本研究で使用した「がん患者」の情報は、以下のとおりである。

- ① 身体的情報：性別、年齢、身長、体重などの身体情報、喫煙歴
- ② 疾患の情報：肺塞栓症・静脈血栓症、心筋梗塞、その他の虚血性心疾患、心室性不整脈、心房細動、心膜炎、高血圧症、糖尿病、脂質異常症、高尿酸血症、慢性腎臓病、脳血管疾患など
- ③ 薬剤投与情報：抗がん剤、高圧薬、心不全薬などのリスト
- ④ 診療行為情報：心電図検査、エコー検査など
- ⑤ 経時的情報：入院日、外来日、疾患に罹患した日（診断日）、退院日、薬剤投与日、検査など診療行為を実施した日、死亡日

5.1.3 目的達成に必要な情報を含むデータセット

これらの情報のうち、①性別、年齢に関する情報は、DiseaseDataとActDataの両方から得られた。また、身長、体重、喫煙歴は、FF1から得られた。さらに、身長と体重の情報をもとにBMI（Body Mass Index）計算した（ $BMI = \text{体重} \text{kg} / (\text{身長} \text{m})^2$ ）。本研究ではBMI 25以上を肥満とした。

②疾患に関する情報は、DiseaseDataから、ICD10コード、疾病名、疾病名コード、疑い病名フラグの形で得られた。疑い病名フラグとは、レセプトで検査や投薬を行うために病名が確定していなくても、病名が疑われる状態で病名を記入しておき、疑い病名フラグをつけておくことで検査や投薬を実施可能にするものである。病名が確定すると疑い病名フラグは外れる。病気の患者を集計する場合は疑い病名フラグを持つ患者は除外した。

③投薬および④診療行為に関する情報はActDataから得られた。ただし、投薬情報は、コードマスタであるM_Drugを、ReceiptCodeをキーとして、参照する必要があった。

⑤経時的情報については、入院日、外来日はDiseaseDataから、薬剤投与日、検査など診療行為を実施した日はActDataから直接得られた。疾患に罹患した日（診断日）については、直接は入力されておらず、医師の検討の下でDiseaseDataの該当する疾患の入院日、外来日で代用した。死亡日も直接は入力されておらず、医師の検討の下でFF1に入力されている退院日で代用した。

5.2 データセットの確認とクレンジング

5.2.1 専門家のクロスチェックによる独自コードマスタの作成

MDV社から提供された本データセットは「抗がん剤が投与されたがん患者のみを抽出した」ということであった。しかし、MDV社から提供されたデータ抽出に用いた、独自コードが付された145種類の抗がん剤のリストはあったが、医薬品マスタであるM_Drugファイルとの対応は開示されなかった。このため、抗がん剤のリストから医薬品名の語幹部分を抽出し、M_Drugファイルに正規表現を用いて検索をかけることで、抗がん剤リストとM_Drugファイルを対応づけるマスタファイルを作成し、これを間違いないか臨床医のメンバで確認をした。疾患の定義についても、ICD-10コードをもとに、データセットに格納されている傷病名のリストを作成して、臨床医のチームで確認をし、個々にマスタファイルを作成した。

DiseaseData.ファイルに含まれていたICD-10コードは、5,649コードであり、このうちがんであると定義したもの（ICD-10国際分類C00-C99コード）は354コードであった。がん以外の疾患のコードの定義については5.3節に示す。

5.2.2 独自コードマスタを用いたデータのクレンジング

作成したマスタファイルを用いて、データセットから条件に合うデータを抽出した（ICD-10国際分類C00-C99コードを持つがんの確定診断がされた患者）ところ、①約4万人（正確には38,098人）の我々の定義したがんのコード名が確認できない患者のデータが見つかったので、分析対象から外した。これには、半分の約2万人のがんの疑い病名のフラグのみを持つ患者が含まれていた。残り半分の約2万人は我々が独自のマスタファイルを作成して、データ抽出を行ったために、条件の違う患者が抽出されていた可能性がある。ICD-10国際分類では、C00-D48をがんとしている。本データセットでは、C00-D48のコードを含まない患者も多数検出したが、これらの患者がなぜ混入してきたかは不明である。我々と異なるコードががんとして定義されていた可能性がある。

さらに、確実にがんにかかった後に抗がん剤が投与されている患者とわかっている患者のみを分析するために、2008年4月から2017年1月までの間にごんと確定診断された患者のみを扱うこととした。その結果、15,698人が除外された。

また、②我々の定義した抗がん剤のコードマスタで抗がん剤の投与が確認できない患者も見つかった。この段階で、2,530人が除外された。また、稀ではあったが、③がんの傷病名の日付より前に抗がん剤が投与された例も見つかった。これは、診療行為の入力ミスによるエラーであるのか、別の効果を期待してがん治療の目的以外に抗がん剤が投与されたのか確認できなかったが、想定外の事象であるの

で除外した。その結果、854人が除外された。さらに、本研究では高齢者における抗がん剤副作用の実態調査を研究の目的としていたため、18歳未満の患者138人を解析対象から除外した。その結果、解析対象患者は、最初の20万人から14万人に減った（図3）。

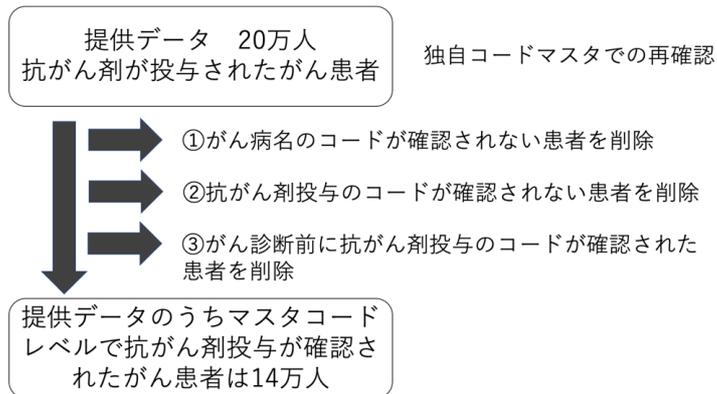


図3 実施したクレンジング工程

5.2.3 独自コードマスタを用いたクレンジングに関する考察

結果として棄てたデータが多かった原因は、次のようなことが考えられる。①抗がん剤およびがんのマスタコードを医師の確認のもと独自に再現・作成した。これは、MDV社のものより、より厳格で厳選されたものであったと思われる。②疑い病名のフラグのみがついているものは分析対象から除外し、がんと確定診断されている患者だけを扱った。③がんの確定診断の前に抗がん剤を投与されている例外がみつかった。結果的に、条件に合わないデータを多数棄てることにはなったが、より正確なデータセットが担保されると考えた。

教訓として、ビッグデータでは想定外の例外は必ず起こるため、あらゆる例外を想定して、確認する必要がある。また、想定外のノイズが見つかることもある。

5.3 疾患の診断の確定方法の検討

5.3.1 業界独自事情に依存する集計の不正確さの問題

DPCデータ/レセプトデータの各データ項目は、医療業界のしきたり的な慣習、特殊事情に基づいて書き込まれていることがあり、そのまま集計すると危険である。集計値が、実際の数値や、厚生労働省の患者調査[17]などのデータから実感する数値とかなり違った数値が得られることがあり、何らかの調整が必要である。

たとえば、ここでは傷病名コードについての2つの問題を取り上げる。1つ目は「コードの冗長性、同一性」の問題、2つ目は、疑い病名や、いわゆる「保険病名（レセプト病名）」の問題である（図4）。

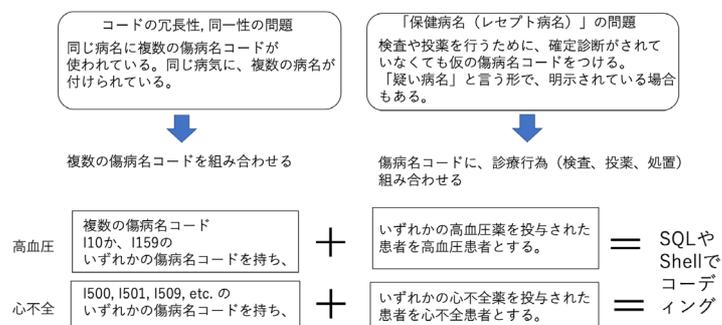


図4 傷病の診断確定に関する問題点

5.3.2 コードの冗長性、同一性の問題

これは、同じ疾病に関して、別の名前が付けられることを指す。わかりやすい例で言うと、「がん」という同じ病気に対して「がん」、「悪性腫瘍」、「悪性新生物」といった別の傷病名が付けられるケースである。たとえば、「前立腺がん」と「前立腺腫瘍」は同じ疾病であるが、別の傷病名コードが割り当てられることがある。データ分析処理を行う担当者は、傷病名に関して、背景知識を持っておらず、どれとどれをまとめるかは、臨床医でなければ判断が難しい。そこで、各傷病名コードと疾病名および、そのグループ集計結果のリストをSQLなどで作成し、心臓血管内科医の研究グループにフィードバックし、傷病の再定義を依頼した。

MDV診療データは、傷病名コードとしてICD-10と呼ばれる国際的な疾病分類を用いていた。臨床医による再定義の結果、たとえば、糖尿病患者は、E10, E11, E12, E13, E14 から始まる傷病名コードを持つ患者、高血圧患者は、I10, I159から始まる傷病名コードを持つ患者などと、分析したい各疾患のコードを再定義していった（図4左側）。

今回のデータ分析の目的のために以下の疾患のICD-10コードを医師の検討の元で定義した。

- ① 心不全: I500（うっ血性心不全）など5コード
- ② 肺塞栓症・静脈血栓症: I126など12コード
- ③ 心筋梗塞: I210など17コード
- ④ その他の虚血性心疾患: I200（狭心症）など12コード
- ⑤ 心室性不整脈: I470など4コード
- ⑥ 心房細動: I48（心房細動および粗動）
- ⑦ 心膜炎: I319, I309
- ⑧ 高血圧症: I10, I159
- ⑨ 糖尿病: E10など40コード
- ⑩ 脂質異常症: E780など8コード
- ⑪ 高尿酸血症: E790
- ⑫ 慢性腎臓病: N188, N189
- ⑬ 脳血管疾患: E610など17コード

同様の問題は、医薬品コードや診療行為コードなどでも生じる。また、1人の患者が複数の医療機関にかかった場合に、複数のIDが割り当てられることがある。これを、統合・連結して解決することを「名寄せ」という。

5.3.3 「保健病名（レセプト病名）」の問題への対応

「保健病名（レセプト病名）」とは、本来の病名ではなく保険請求上記載した病名である。疾病を疑って検査などを行い、疾病の疑いが晴れた場合でも、検査に対応する病名の記載がないと審査で検査料の請求が認められないため、慣例的に行われている。

すなわち、DPCデータ/レセプトデータに書き込まれる傷病名コードは、たとえば、感染症の薬を出すために、感染症の確定診断がされる前に感染症の傷病名コードが書き込まれたり、検査を行うために、確定診断がされる前に「仮」の傷病名コードが書き込まれたりすることがある。実際の、疾患の集計を行う場合には、患者数が多く見積もられ、病気の分析を行う混乱材料となる。

このため、傷病名コードに加え、診療行為コードを組み合わせ、疾患を定義した。たとえば、以下のように疾患を再定義した（図4右側）。

- ① 心不全の傷病名コードを持つ患者で、心不全の治療薬（ATC分類コードでC3A1など10剤のうちいずれか）を投与されたものを、心不全患者と定義した。さらに、類似疾患度として18の心筋梗塞ITC-10コードおよび胸膜炎など47のITC-10コードを心不全の診断初日に持つ患者を除外した
- ② 高血圧の傷病名コードを持つ患者で、高血圧の治療薬（ATC分類コードでC03A1など8剤のうちいずれか）を投与されたものを、高血圧患者と定義した
- ③ 虚血性心疾患の傷病名コードを持つ患者で、発症前3日以内に心電図検査が施行されているものを、虚血性心疾患患者と定義した
- ④ 心室性不整脈の傷病名コードを持つ患者で、発症前3日以内に心電図検査が施行されているものを、心室性不整脈患者と定義した
- ⑤ 心房細動の傷病名コードを持つ患者で、発症前3日以内に心電図検査が施行されているものを、心房細動患者と定義した
- ⑥ 心膜炎の傷病名コードを持つ患者で、発症前7日以内に心エコー検査が施行されているものを、心膜炎と定義した
- ⑦ 肺塞栓症・静脈血栓症の傷病名コードを持つ患者で、発症前7日以内に心エコー検査が施行されているものを、心膜炎と定義した

疾患の再定義をする際に、治療薬1剤投与か、2剤投与かのいずれが妥当かについても問題となったが、各疾患の投与例ごとに投薬患者数を集計し、確認した。また、検査を施行する日が発症前3日以内か、7日以内か、1カ月以内かのいずれが適切かについても、各疾患ごとに実施数を集計し、確認した。

このような、傷病名コードと診療行為コードを組み合わせたパターンをいろいろ集計して、心臓血管内科医の研究グループにフィードバックし、実際の既知データ（患者調査[17]など）の疾患率と整合するものを、疾患の定義として決定した。

5.4 疾患集団の厳格なマッチング条件の検討

5.4.1 リアルワールドデータの集団の不均一性の問題

データのクレンジングを行い、臨床実態に整合する疾患分布を示す疾患定義を行った後に、疾病患者集団のデータ分析を実施し、疾患の背景因子の検討を行う。今回のようなリアルワールドデータから抽出した集団は、実際には多種多様な背景の患者データを含む不均一な集団であり、そのままでは副作用の分析はうまく行かなかった。

たとえば、一般的には喫煙[18]や肥満[19]、糖尿病[20]は心疾患の危険因子（リスクファクター）であることが知られている。これらの背景を持つ患者は心疾患になりやすいことが予想される。

しかし、今回抽出した集団を用いて喫煙歴の有る人となない人で心疾患や糖尿病の罹患率を比べても、喫煙歴のない人に心疾患罹患率が高かったり、肥満の人（BMI 25以上）と肥満でない人（BMI 25未満）では、肥満でない人の方が肥満の人に比べて糖尿病や高血圧の比率が高いという予想と逆の結果が出たりした。

現実には、糖尿病末期やがん末期の患者は、痩せの傾向がある。リアルワールドデータはさまざまな背景の人が混在するため、単に病気の有り無しや、肥満の有り無しなどの背景因子で患者を分類して、その傾向を調べようとしてもうまくいかなかった（図5上の図）。

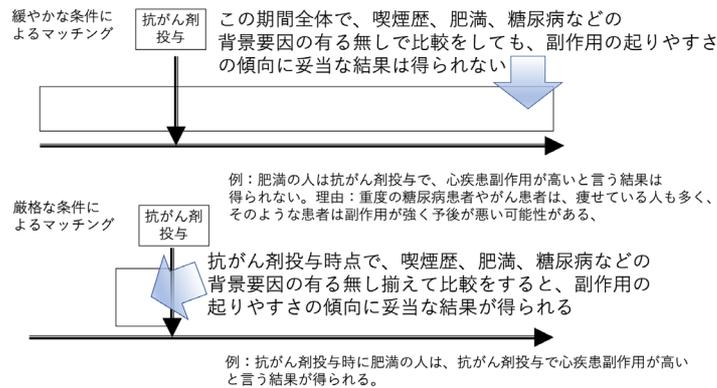


図5 喫煙歴、肥満、既往歴などの背景因子のマッチングの重要性

そこで背景因子を厳格な条件でマッチングすることとした（図5下の図）。たとえば、肥満の人と肥満でない人の分類は、抗がん剤投与開始前1カ月以内にBMIが25以上の患者を肥満とし、抗がん剤投与開始前1カ月以内にBMI25未満の患者を非肥満とした。疾患の背景、たとえば糖尿病の背景のある人は、抗がん剤投与時点での、糖尿病の既往のある人とした。高血圧も同様で、高血圧の背景のある人は、抗がん剤投与時点での、高血圧の既往のある人とした。

この場合、肥満や、高血圧、糖尿病の背景がある時期の基準を、抗がん剤投与前1カ月以内とするか、7日以内とするか、当日とするか、いずれが妥当かについても実際にその頻度を計数して確認した。

背景や条件を厳密に設定し、患者背景を厳格条件でマッチングすることで、喫煙歴のある人、肥満の人、糖尿病や、高血圧にかかっている人は心疾患になりやすいという一般に認知されるような傾向が取れるようになった。

逆に、一般的に認知されている事象がきちんと再現できる厳密な条件下で抽出できたデータセットを用いることで初めて、未知の薬剤や、未知の併存疾患の解析が可能になる。

また、本研究では、時系列解析を実施する際に、年齢、性別、背景疾患などの交絡因子となる因子を調整した上で、解析を実施した。本研究では利用しなかったがリアルワールドデータでは、傾向スコア（Propensity Score, PS）という指標^[21]を用いて交絡因子を調整する場合もある。

5.5 副作用の発生時期の問題

もう1つの論点として、副作用の発生時期で、抗がん剤投与開始の当日に心疾患になった例を副作用と考えるかという問題があった。これは、個別の疾患により変わってくる。

高血圧のような慢性疾患の場合、抗がん剤投与開始の当日に急に副作用が起こるということは考えにくく、もともと高血圧であった可能性の方が高い。一方で、急性心筋梗塞や急性心不全の場合は、抗がん剤投与開始の当日に急に発生するということもありうる。

今回、まず心不全を対象とした解析から開始しているのため抗がん剤投与開始の当日の発症は、副作用とカウントした。

5.6 データ分析方法の適用に関する検討

5.6.1 リアルワールドデータの分析方法に関して

リアルワールドデータの分析方法については、これらの研究が行われるようになって間もない事もあり、記述統計学的なデータの集計などの解析が多い[22]。

あらかじめ学術研究や臨床研究（いわゆる2次利用）を目的としてデザインされ集められたデータではないため、厳格な条件で背景のマッチした集団で実施する必要のあるCox回帰比例ハザード分析を含む生存時間解析、一般化線形モデルや機械学習など、データモデリングを用いた分析例はあまり見られない。本研究では、詳細な経時的情報を抽出し、 Kaplan-Meier 曲線の作図などを行った。このとき、臨床医の検討により、薬剤投与開始日を研究開始時点とし、退院日を死亡日として解析した。

5.6.2 本データ分析の臨床研究・臨床試験における意義

本データ分析により、約14万人の抗がん剤が投与されたがん患者において、心不全、肺塞栓症・静脈血栓症、心筋梗塞、その他の虚血性心疾患、心房細動、心室性不整脈、心膜炎、高血圧症などの、心疾患の罹患率等が明らかになった。さらに、個々の心疾患について、その詳細を調べたところ、女性より男性の方が心疾患になりやすいこと、高齢者、肥満、喫煙歴、糖尿病や高血圧、慢性腎臓病、脳血管疾患などの既往がある人も心疾患かかりやすい傾向にあること等が示された。詳細については、臨床系の専門誌に掲載される予定である。

新薬開発における臨床試験のような、非常に手続きが煩雑で、高額な研究を実施しなくても、医療ドキュメントを追跡するだけで疾患のリスク因子や副作用の分析ができることを示した意義は大きいと考える。

5.6.3 リアルワールドデータを用いた臨床研究・臨床試験における最近の動向

では、リアルワールドデータを用いた臨床研究・臨床試験はどの程度実施されているのだろうか。2019年5月に米国食品衛生局FDAは、“Submitting Documents. Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics. Guidance for Industry. [23]”という薬事申請などへのリアルワールドデータを用いた試験結果の申請に関する指針のドラフトを公表し、そのパブコメを実施した（同年7月第1週まで）。米国ではすでに2019年9月30日時点で、リアルワールドデータを転帰指標として対照に用いた臨床試験が17件実施されているという調査報告がある[24]。しかし、同調査では日本の事例はまだ確認されていない。また、日本では医薬品医療機器総合機構（PMDA）がレジストリデータ（レセプトデータ）を承認申請などに用いるための基本的な考え方や、データの信頼性に関する留意点等が記載されたガイドライン案を2019年度に作成し、2020年度に公表することを予定しているという[24]。同調査によると、「承認申請時の根拠情報として使用されるためには、データの質[25]、すなわちデータの「完全性」、「一貫性」、「正確性」の点を確保することが求められる。」ということから、本稿で検討したデータのVeracity（正確性）をいかに担保するかは、このような研究で最も留意すべき点であると思われる。

5.7 本データ分析の課題・限界

本研究では、医療分野、情報分野、統計分野の各専門家が、より正確なデータを出すためにチームワークで実施した検討内容と得られたノウハウを紹介した。これらは長期に渡る頻繁な議論の成果である。これらのプロセスの迅速化を図るために、機械学習などの導入が望まれる。

リアルワールドデータの課題・限界として1人の患者が複数の医療機関にかかった場合の追跡方法、疾患が治癒した場合や再発例の検出の問題などがある。今回は、MDV診療データのみでの分析であったが、国勢調査や、患者調査など他の統計データとの連結や整合性も課題である。いずれにしても、診療行為の結果生じるドキュメントであるため、活用の際には、その特徴や限界を見据えた分析が必要になる。

6. 医療ビッグデータを扱う場合のコンプライアンス

医療ビッグデータ、特に患者の診療情報を扱う場合は、患者の人権およびプライバシーの保護、データの倫理的取り扱いに特別の注意を払う必要がある。

今回、使用したデータ元のMDV診療データは、個人情報の保護に関する法律（2017年5月30日施行、以下「改正個人情報保護法」）が求める匿名加工情報の作成基準を満たした匿名加工処理が、事前になされている。これらのデータは、同法の匿名加工情報に該当する。MDV社は、同法の匿名加工情報取扱事業者として、同法の匿名加工情報取扱事業者の義務および国内の法令・ガイドラインを遵守していることを宣言している[26]。

また本研究は、学内の倫理委員会の審査を受け、承認を得た後に実施した。

7. まとめ

本稿では、医療ビッグデータであるDPCデータ/レセプトデータを元に作成した大規模データベースを用いた分析を通じて、その問題点とその解決策を述べた。ビッグデータの5Vのうち、今回特に問題になったのは、Veracity（正確性）をどう担保するかという点である。その主な論点は3つであった。1つ目は、データの確認とクレンジングの問題である。ビッグデータでは例外的な想定外の事象が、かなりの頻度で生じる。2つ目は、コードの冗長性・同一性と、分析項目のドメイン特異的な慣習によるバイアスである。3つ目は、データの背景因子を調整するためのマッチングの問題である。ビッグデータの患者集団は多種多様で不均一であるため、分析のための集団の背景因子を揃えることに注意を払う必要がある。これらの問題の解決には、情報処理、統計科学、医療などの各ドメインの専門知識が必要で、各専門家の密接なコミュニケーションによる検討により初めて可能になる。

謝辞 本研究は、平成29年度石橋学術研究振興基金助成金の助成により実施された。関係者のご協力に感謝を申し上げる。

参考文献

- 1) 富士経済：民間企業による医療ビッグデータビジネスが本格化, https://fuji-keizai.co.jp/market/detail.html?cid=18046&view_type=2
- 2) 厚生労働省：地域包括ケアシステム, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/hukushi_kaigo/kaigo_koureisha/chiiki-houkatsu/
- 3) 厚生労働省：レセプト情報・特定健診等情報の提供に関するホームページ, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryoku/iryohoken/reseputo/index.html
- 4) 厚生労働省：DPCデータの提供に関するホームページ, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryoku/iryohoken/dpc/index.html
- 5) 株式会社JMDC：JMDC Claims Database, <https://www.jmdc.co.jp/jmdc-claims-database/>
- 6) メディカル・データ・ビジョン株式会社、「MDV診療データ」について, <http://www.mhlw.go.jp/bunya/iryohoken/database/sinryo/dpc.html>
- 7) Demchenko, Y., de Laat, C. and Membrey, P. : Defining Architecture Components of

the Big Data Ecosystem, International Conference on Collaboration Technologies and Systems (CTS), Minneapolis, MN, 2014, pp.104-112 (2014).

- 8) 社会保険診療報酬支払基金：診療報酬の審査・支払業務—業務の流れ, <https://www.ssk.or.jp/seikyushiharai/gyomufLOW/index.html>
- 9) 社会保険診療報酬支払基金：レセプト電算処理システム, <https://www.ssk.or.jp/seikyushiharai/rezept/index.html>
- 10) 社会保険診療報酬支払基金：基本マスタ, <https://www.ssk.or.jp/smph/seikyushiharai/tensuhyo/kihonmasta/index.html>
- 11) 厚生労働省：特定健診・特定保健指導について, <https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/0000161103.html>
- 12) 厚生労働省：DPC制度（DPC/PDPS）の概要と基本的な考え方, <https://www.mhlw.go.jp/stf/shingi/2r985200000105vx-att/2r98520000010612.pdf>
- 13) 厚生労働省：DPC導入の影響評価に係る調査, https://www.mhlw.go.jp/stf/seisakunitsuite/bunya/kenkou_iryuu/iryuuhoken/database/dpc.html
- 14) メディカル・データ・ビジョン株式会社：https://www.mdv.co.jp/press/2020/detail_1262.html (2020年2月末日集計)
- 15) FDA：Real-World Evidence, <https://www.fda.gov/science-research/science-and-research-special-topics/real-world-evidence>
- 16) Nohara, S., Shibata, T., Ishii, K., Obara, H., Miyamoto, T., Kakuma, T. and Fukumoto, Y. : Cancer Therapeutics-related Heart Failure from a Cohort Study Using Big Data of Electronic Health Record in Japan, European Society of Cardiology (ESC) Congress 2018, Munich, Germany (2018 Aug. 25-29)
- 17) 厚生労働省：患者調査, <https://www.mhlw.go.jp/toukei/list/10-20.html>
- 18) Ambrose, J.A. and Barua, R. S. : The Pathophysiology of Cigarette Smoking and Cardiovascular Disease : an Update, J Am Coll Cardiol.43 (10) : 1731-7 (2004).
- 19) Rosito, G. A., Massaro, J. M., Hoffmann, U., Ruberg, F. L., Mahabadi, A. A., Vasan, R. S., O'Donnell, C. J. and Fox, C.S. : Pericardial Fat, Visceral Abdominal Fat, Cardiovascular Disease Risk Factors, and Vascular Calcification in a Community-based Sample : the Framingham Heart Study, Circulation. 117(5) : 605-13 (2008) .
- 20) Rajagopalan, S. Brook R.Canagliflozin and Cardiovascular and Renal Events in Type 2 Diabetes.N Engl J Med. 377 (21) :2098-9 (2017) .
- 21) Rosenbaum, P.R. and Rubin, D.B. : The Central Role of the Propensity Score in Observational Studies for Causal Effects, Biometrika, Vol.70, pp.41-55 (1983) .
- 22) Colin-Bracamontes, I., Pérez-Calatayud Á. A., Carrillo-Esper, R., Rodríguez-Ayala, E., Padilla-Molina, M., Posadas-Nava, A., Olvera-Vázquez, S. and Hernández-Salgado, L. : Observational Safety Study of Clottafact® Fibrinogen Concentrate : Real-World Data in Mexico, Clinical Drug Investigation, 2020;10.1007/s40261-020-00906-6.doi:10.1007/s40261-020-00906-6 (published online ahead of print, 2020 Mar. 25).
- 23) FDA : Submitting Documents Using Real-World Data and Real-World Evidence to FDA for Drugs and Biologics Guidance for Industry (2019) .
- 24) 中塚靖彦：Points of View臨床開発における治験対照群の利活用, 政策研ニュース, Vol.58, PP.66-70, http://www.jpma.or.jp/opir/news/058/pdf/no58_p66.pdf (2019)
- 25) 日本製薬工業協会, RWD : 「データの質」に関する考察, http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/pdf/rwd_quality.pdf
- 26) メディカル・データ・ビジョン株式会社,EBM事業における国内法令・ガイドライン遵守宣言, https://www.mdv.co.jp/press/2018/detail_1109.html

石井一夫 (正会員) ishii_kazuo@med.kurume-u.ac.jp

久留米大学バイオ統計センター准教授, 1995年徳島大学大学院医学研究科博士課程修了。博士(医学)。専門はビッグデータ分析, 計算機統計学。R, Python, Juliaなどのオープンソースソフトウェアを使用して医療ビッグデータ, 医療ゲノムデータ, 医療IoTデータを用いた大規模データ分析に従事。

野原正一郎（非会員）nohara_shoichiro@med.kurume-u.ac.jp

久留米大学医学部内科学講座 心臓・血管内科部門 助教 2007年久留米大学医学部医学科卒。専門:心不全, 腫瘍循環器学, 専門資格:循環器専門医, 総合内科専門医。

柴田龍宏（非会員）shibata_tatsuhiko@med.kurume-u.ac.jp

久留米大学医学部内科学講座 心臓・血管内科部門 助教, 2009年熊本大学医学部医学科卒。専門:重症心不全, 腫瘍循環器学, 心不全緩和ケア。専門資格:循環器専門医, 総合内科専門医, 緩和医療認定医。

小原 仁（非会員）obara_hitoshi@kurume-u.ac.jp

久留米大学バイオ統計センター助教。診療情報管理士指導者, 九州大学大学院 医学系学府 医療経営・管理学専攻修了, 医療経営・管理学修士(専門職), 久留米大学大学院 医学研究科 博士課程修了, 博士(バイオ統計学), 医療情報の活用・分析を中心に, 病院経営管理や臨床研究に関するデータ分析に従事している。

宮本貴宣（非会員）t_miy@med.kurume-u.ac.jp

久留米大学バイオ統計センター准教授, 博士(工学) 山口大学大学院理工学研究科システム工学専攻博士課程修了。専門は機械学習, 情報工学, 医療ビッグデータ分析。現在, 久留米大学病院にて医療ビッグデータに基づく病院経営分析に従事。

角間辰之（非会員）tkakuma@med.kurume-u.ac.jp

久留米大学バイオ統計センター所長/教授。1990年エール大学School of Public Health, Biostatistics博士課程修了(Ph.D), 1990-2000年コーネル大学医学部精神科ウエストチェスター部(准教授), Weill Cornell Institute of Geriatric Psychiatry(統計部部長), 2001-2004年日本赤十字九州国際看護大学 保健統計・情報科学教授, 専門は臨床試験全般の統計解析。

上野高史（非会員）takueno@med.kurume-u.ac.jp

久留米大学病院循環器病センター教授。久留米大学バイオ統計センター副所長。1982年長崎大学医学部医学科卒。1993年久留米大学医学博士。1999年米国アトランタ心臓血管研究所へ留学。2010年から現職。現在, 久留米大学病院副院長および久留米大学臨床研究支援センターセンター長を併任。

福本義弘（非会員）fukumoto_yoshihiro@med.kurume-u.ac.jp

久留米大学医学部内科学講座心臓・血管内科部門 主任教授。1991年九州大学医学部医学科卒。1998年九州大学医学博士。1998-2001年ハーバード大学ブリガム・ウィメンズ病院(ポスドク)。1991-2006年九州大学循環器内科および関連病院。2006-2013年東北大学循環器内科。2013年から現職。現在は久留米大学循環器病研究所所長および久留米大学病院副院長を兼任。

投稿受付: 2020年3月22日

採録決定: 2020年4月5日

編集担当: 細野 繁 (東京工科大学)