[Original Article]

医学研究に二次利用するための医療情報データの特徴・性質調査
―大阪大学医学部附属病院 単施設の状況―

■

Survey of the Characteristics and Properties of Medical Information Data for Secondary Use in Medical Research
―Osaka University Hospital Single Center Status―

■

荒木　浩之[*1]　惟高　裕一[*2]　佐藤　倫治[*3]　長谷川貴大[*2]
小林　典弘[*2]　山田　知美[*4]　飛田　英祐[*3]

ABSTRACT

Background In recent years, there has been an increasing expectation to use real-world data in the development of drugs and medical devices. However, medical information from electronic medical records is not data collected following research plans. This may lead to problems in which the data is not collected with the quality expected for research, or the data is biased and is different from the target population. Therefore, when using medical information data, it is important to understand the characteristics and properties of the data in advance and to confirm and select the data for the research purpose. Therefore, we conducted a survey to clarify the characteristics and properties of medical information data from Osaka University Hospital（OUH）.

Method We extracted items frequently used in clinical studies from medical information data of all patients who had their first medical examination at OUH in the past 10 years. Among the data of each item, we examined the distribution and transition over time of patient background and the frequency of missing or abnormal data.

Results and Conclusions This study showed that the medical information data of OUH includes some items that were entered according to the status of missing data and the original rules. When using a database, it is necessary to check the error data and examine the cause and handling. Understanding the characteristics and properties of medical information data in advance is expected to facilitate identifying the cause of errors and the selection of databases will be more efficient.

（Jpn Pharmacol Ther 2022；50 suppl 1：s37-50）

[*1]大阪大学 大学院医学系研究科医療データ科学共同研究講座（現 一般社団法人 JBCRG データセンター）　[*2]塩野義製薬株式会社
[*3]大阪大学 大学院医学系研究科医療データ科学共同研究講座　[*4]大阪大学 医学部附属病院 未来医療開発部 データセンター
Hiroyuki Araki[*1], Yuichi Koretaka[*2], Tomoharu Sato[*3], Takahiro Hasegawa[*2], Norihiro Kobayashi[*2], Tomomi Yamada[*4], Eisuke Hida[*3]: [*1]Osaka University Graduate School of Medicine Department of Biostatistics & Data Science（Current affiliation：JBCRG Data Center）, [*2]SHIONOGI & CO., LTD., [*3]Osaka University Graduate School of Medicine Department of Biostatistics & Data Science, [*4]Osaka University Hospital Department of Medical innovation

KEY WORDS　RWD（real world data）, medical information data, secondary use, EMR（electronic medical record）

はじめに

これまでの医療，臨床上のエビデンスは，RCT（randomized controlled trial）などに基づく臨床試験を中心として創出されてきた。しかし，臨床試験の計画，実施，解析，報告には多大なコストと時間が必要であり，近年の新たな医薬品や医療機器の開発費は膨張の一途をたどっている[1)]。この問題に対する解決方法のひとつとしてリアルワールドデータ（RWD: real-world data）の利用に期待が高まっている。特に，RWD のなかでも電子カルテデータは，従来広く利用されてきた医事請求データより詳細な臨床情報が含まれるため，医薬品・医療機器開発への効果的な利活用が注目されている。

RWD を用いる研究では，その研究目的に即したデータベースを選択することが重要であるが，実際に解析を行うまでには，さらにデータチェックや加工処理などの作業が必須となる。しかし，医療情報データは研究目的や研究計画に沿って収集されるデータではないため，研究に必要な項目が観察されていない場合や観測された検査項目でも方法や単位，粒度が異なる，データの誤入力や施設独自のルールに従い入力されたデータやカテゴリ化などによる情報の変質・欠落など研究に期待される品質でデータが収集されていない問題がある。さらに，複数の異なる医療施設が有する医療情報データを集約したデータベースでは，施設ごとの医療情報データの状況の違いや地域医療で果たす役割および病床数など規模の違いにより，研究で関心のある対象集団とは異なり，偏ったデータの集まりとなる可能性がある。

臨床試験では，解析対象集団を明確にし，データをどのように収集したかを踏まえたうえで，解析結果を解釈することが一般的である。これと同様に，医療情報データを利活用する際にも，そのデータが有する先に述べた諸問題を含む特徴や性質を事前に把握したうえで，臨床研究などの二次利用の目的に対して適当なデータであるかの確認およびデータベースを選択することができれば，その後のデータチェックや加工段階における作業の効率化につながることが期待できる。このデータベースを選択する前段階において，医療情報データの特徴や性質を把握しておくことにより，臨床試験よりも一般化可能性が高い RWD による解析結果やヒストリカルデータを医薬品・医療機器開発へ効果的に利用できることが期待される。現在，日本で疫学研究や臨床研究に応用可能なデータベース一覧[2)]が公開されているが，特定の施設が有する医療情報データの特徴や性質について調査された報告は限られている。そこで，われわれは地域中核病院として 1000 を超える病床を有する大阪大学医学部附属病院（以下，阪大病院）の医療情報データを二次利用する観点から，その特徴や性質を明確化することを目的とした調査研究を実施した。

対象と方法

1　対　象

現在の病院情報システム（HIS: hospital information system）*と同じデータ構造として運用が開始された 2010 年 1 月 1 日～2019 年 12 月 31 日の 10 年間に阪大病院を受診（初診）した全患者を対象とし，医療情報データを用いた調査・研究で利用頻度の高い以下のファイルデータについて，同期間内のデータを DWH（data warehouse）から収集した（本研究の対象データセット：**表 1**）。

ただし，臨床検査ファイル中の臨床検査値と看護データファイル中のバイタルサインは，対象とする全患者で確実に検査が実施される外来患者の初回検査データおよび入院患者の入退院時の検査データを対象とした。

なお，本研究は大阪大学医学部附属病院観察研究倫理審査委員会で承認を得て実施した（承認番号 19411）。また，本研究の対象データセットは，阪大病院の HIS の構築・運用を行う医療情報部が研究計画に従って DWH から抽出した（匿名化済み）。学内の情報セキュリティ対策規程および阪大病院の個人情報の取り扱いに準じて管理し，解析に用いた。

2　方　法

二次利用の観点から医療情報データの特徴や性質を明らかにするため，上記の対象データセットを用いて，データの分布やバラツキの大きさ，正確性（accuracy）[4)]などに焦点を絞って，以下の検討を行った。なお，解析には SAS® 9.4 を使用し，図表の作成には Tableau® Desktop 2020.1 を使用した。

*阪大病院の HIS は，おもに基幹システム・部門システム・医事会計システム・物流管理システムで構成されており，電子カルテデータの二次利用に向けて業務用に使うデータベースと分析用のデータベース（DWH: data warehouse）に分けられ，業務用データベースで発生するデータが定期的に分析用データベースに移行されている[3)]。

1）データの分布と年次推移の確認

①患者背景

患者集団の特徴を把握する目的で，患者の男女比を算出し，出生年や入院時の年齢，入院日数の分布について，経時的な推移図を作成した。

②病　名

病名は，研究対象集団の特徴やアウトカムの発生状況を把握するうえで重要なデータである。本対象データセットにおける疾患の構成を確認するために，「保険病名ファイルの診断名」（以下，保険診断名）と「入院患者ファイルの主傷病名（病名区分）の DPC（diagnosis procedure combination）病名」（以下，主 DPC 傷病名）に付与された ICD-10 コード（3 桁分類）の件数に対する上位 20 位までの順位付けを行った。また，診療科ごとの主 DPC 傷病名の件数の年次推移について分析した。

③臨床検査およびバイタルサイン

臨床検査項目の実施件数の経時的な推移図を作成し，上位 20 位までの件数をもつ臨床検査項目とバイタルサイン（血圧，脈拍，呼吸数，体温）について要約統計量を算出した。

なお，本対象データセットの特徴を相対的に把握するため，メディカル・データ・ビジョン株式会社の構築したレセプトデータである，DPC 対象病院データ由来の施設病院診療データベース（以下，MDV データ）の 2008 年 4 月～2021 年 1 月のデータを利用して，上記の患者背景，病名，臨床検査およびバイタルサインと同様の検討を行った。

2）データの品質に関する確認

①欠測数と欠測理由

欠測の頻度とその理由を特定することは，欠測データによってもたらされる潜在的なバイアスを理解するうえで重要である[4]。そこで，本研究では，入院患者での「入院日のフィールドが未入力」などのあるべき箇所に値がない，あるいは利用不可能，判別不能な内容のものすべてを欠測と定義して，本対象データセットにおける欠測値のリストアップを行った。また，欠測理由については追加調査を行った。

②日　付

日付は，イベントの発現日，検査値の推移，薬剤の使用状況などを調査・解析するうえで重要なデータであり，二次利用する際にはその正確性が求められる。本対象データセットのうち，日付が入るべきフィールドに日付以外の形式の値が入力されていないか確認するとともに，測定日が本研究の対象期間外である 2009 年 12 月 31 日以前など，ありえない値が入力されていないか確認した。さらに，本対象データセットのうち，患者個人ごとに入院日・退院日，薬剤の処方開始日・処方終了日などの日時の前後関係の不整合（例：入院日と退院日が逆転している）についても確認した。

表 1　本研究の対象データセット　各ファイルとその構成項目

患者情報ファイル（269370 レコード）	性別，生年月日
保険病名ファイル（2175467 レコード）	診断名，病名コード，ICD-10 コード，病名開始日
入院患者ファイル（1047809 レコード）	入院日，退院日，入院時診療科，退院時診療科，DPC 病名，ICD-10 コード，病名区分（主傷病名，入院契機病名，医療資源 1，医療資源 2，入院後発症傷病名，入院時併存傷病名）
外来患者ファイル（5124106 レコード）	外来受診日，外来科
臨床検査ファイル（5820631 レコード）	検査日，検査項目，検査値
看護データファイル（1599743 レコード）	項目名（血圧，脈拍，呼吸数，体温），測定日，測定値
処方（注射）ファイル（10693427 レコード）	薬剤コード，薬品名，YJ コード，投与日，診療科，投与量，単位，投与経路，回数，投与方法
処方（注射以外）ファイル（17604889 レコード）	薬剤コード，薬品名，YJ コード，処方開始日，処方終了日，投与量（1 日量），単位

結　　果

1　データの分布や年次推移のバラツキの確認

1）患者背景

本対象データセットの患者総数は 269370 名で，男性 121555 人（45.13%），女性 147815 人（54.87%）であった。全患者の出生年の分布は，厚生労働省が報告している人口動態総覧の出生データ[5]と同じ傾向を示していた（**図 1-1**）。入院時の年齢分布は男女ともに 0 歳が突出して多く，また，男性では 70 歳前後にピークがある単峰性を，女性では 37 歳と 70 歳前後にピークがある二峰性の分布を示した（**図 1-2** 上）。入院日数が判明している 191445 名について分析した結果，0～3 日以内の患者頻度が高く，右裾が長い分布を示しており，入院日数が 1000 日以上の長期入院症例が 29 件あった（**図 1-3** 上）。

2）病　名

本対象データセットの「保険診断名」と「主 DPC 傷病名」，および MDV データの「主傷病名」に付与された ICD-10 コード（3 桁分類）の件数の上位 20 項目を**表 2** に示した。なお，MDV データは，診療行為情報の受診

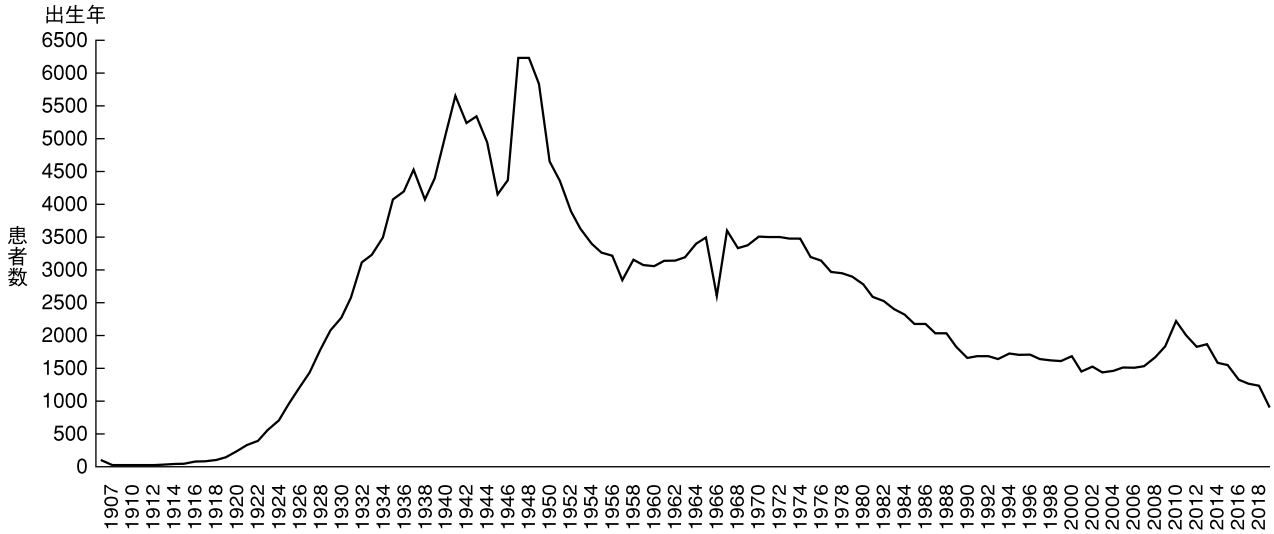

図 1-1　全患者の出生年の分布

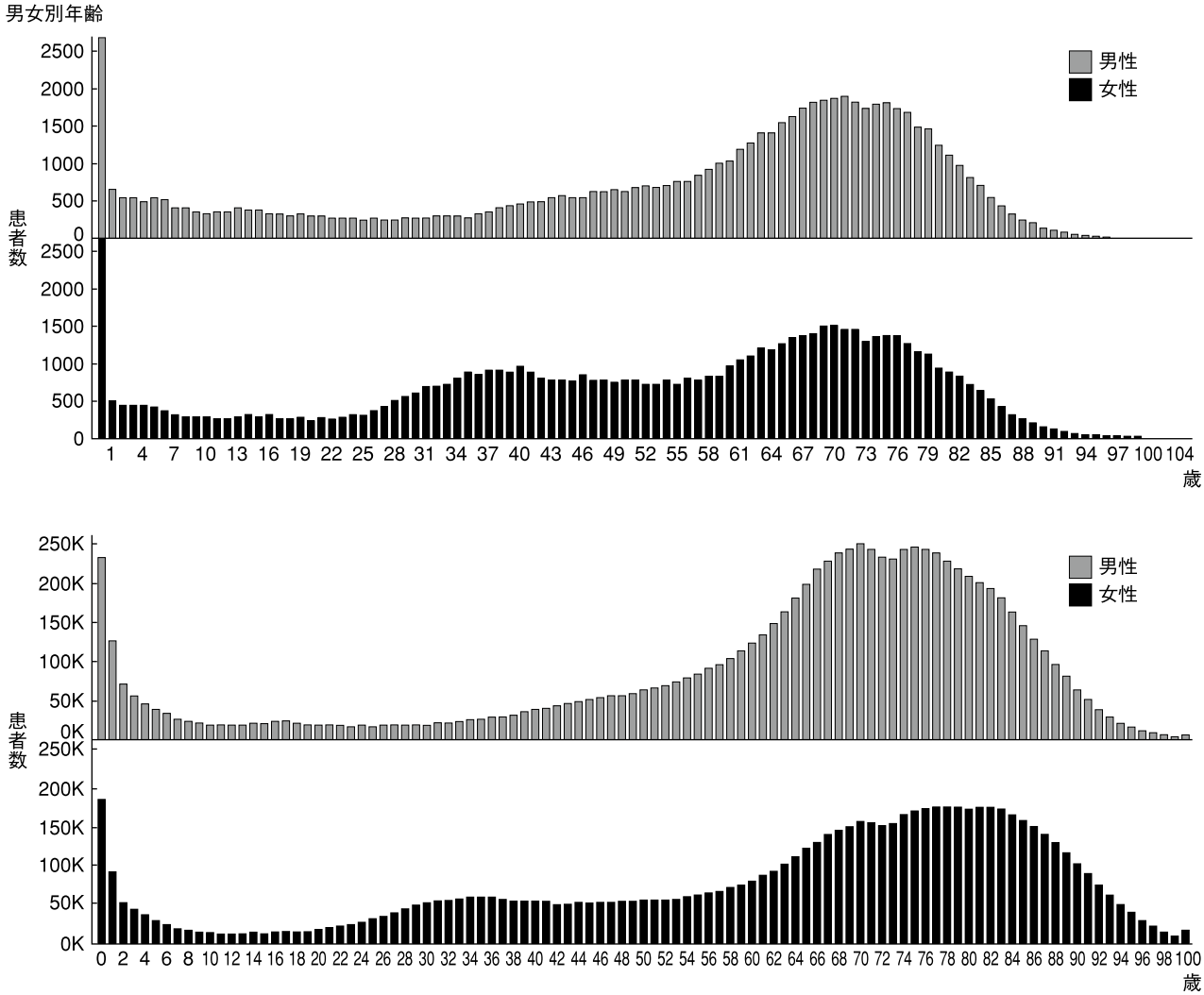

図 1-2　入院時の年齢分布
上）対象データセット，下）MDV データ

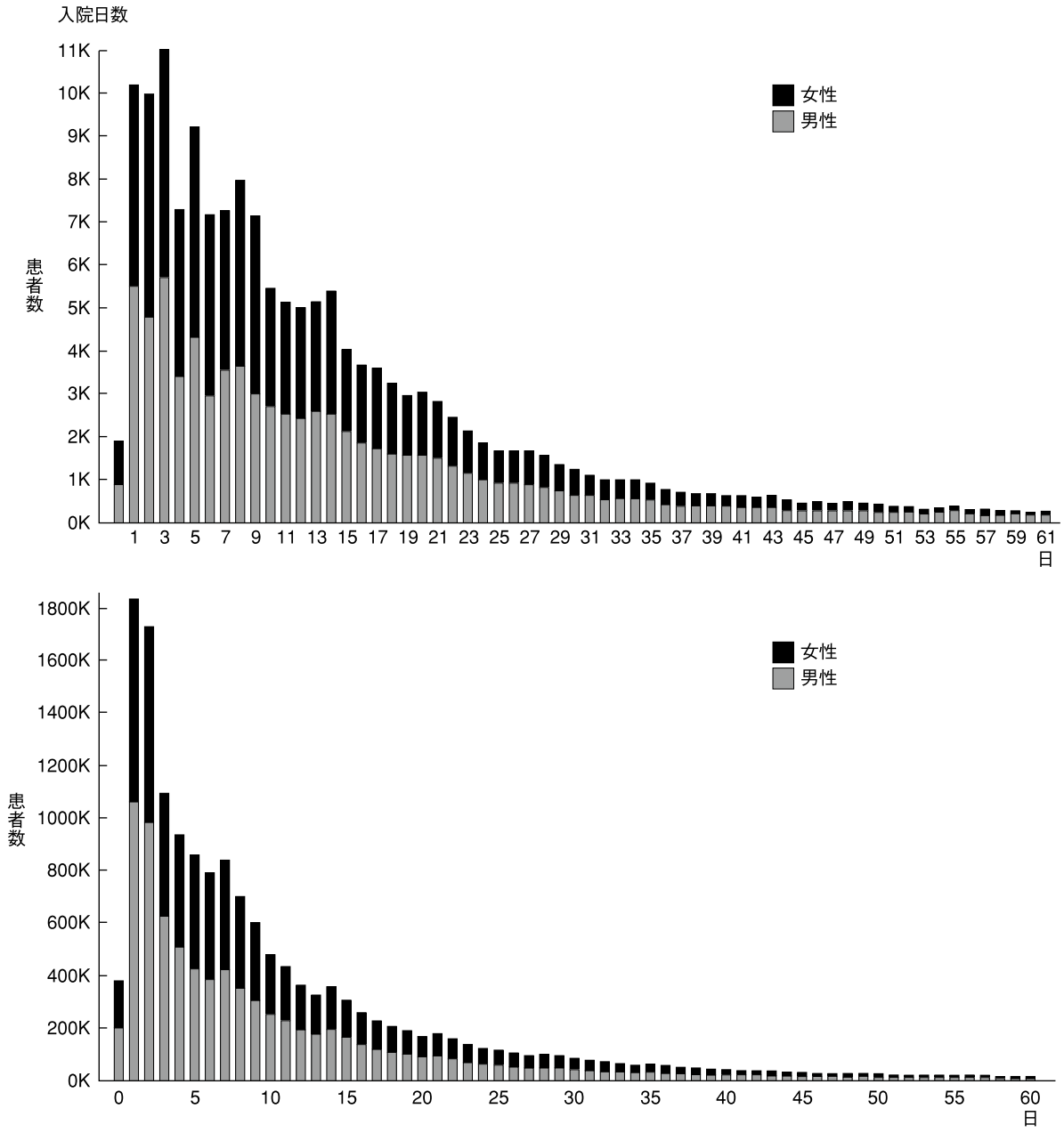

図 1-3　入院日数の分布
上）対象データセット，下）MDV データ

月と病名情報の受診月が一致するとき，当該診療科による診断があったものと定義した。本対象データセットで病名コードが付与されている 2175467 件中，保険診断名での上位 3 項目は，H52（屈折および調節の障害）71251 件（3.28%），I10〔本態性（原発性＜一次性＞）高血圧（症）〕46780 件（2.15%），K59（その他の腸の機能障害）39755 件（1.83%）であった。一方，「主 DPC 傷病名」においては全 191506 件中，主 DPC 傷病名での上位 3 項目は，H26（その他の白内障）6807 件（3.55%），C50（乳房の悪性新生物＜腫瘍＞）5998 件（3.13%），C34（気管支および肺の悪性新生物＜腫瘍＞）5108 件（2.67%）であった。保険診断名と主 DPC 傷病名では，それぞれの役割や定義が異なることから，病名件数の順位が異なるため，研究の目的や関心のある母集団に最も適した病名データであるかの検討は必要である。

入院患者における疾患別の患者数の年次推移については，患者数の顕著な変動が認められた消化器外科と消化器内科において疾患の詳細な検討を行ったところ，特に食道癌や慢性ウイルス性肝炎などの 4 つの消化器系疾患で件数の大きな変動が認められた（**図 2**）。

表 2　病名件数〔ICD-10 コード（3 桁分類）〕の上位 20 項目

	保険病名データ			大阪大学 DPC データ（主傷病名）		
	ICD10 3 桁分類	病　名	レコード数	ICD10 3 桁分類	病　名	レコード数
1	H52	屈折及び調節の障害	71251	H26	その他の白内障	6807
2	I10	本態性（原発性＜一次性＞）高血圧（症）	46780	C50	乳房の悪性新生物＜腫瘍＞	5998
3	K59	その他の腸の機能障害	39755	C34	気管支及び肺の悪性新生物＜腫瘍＞	5108
4	E78	リポタンパク＜蛋白＞代謝障害及びその他の脂血症	38276	C15	食道の悪性新生物＜腫瘍＞	4263
5	E14	詳細不明の糖尿病	36794	C22	肝及び肝内胆管の悪性新生物＜腫瘍＞	3898
6	H26	その他の白内障	35230	I20	狭心症	3777
7	K29	胃炎及び十二指腸炎	34439	C16	胃の悪性新生物＜腫瘍＞	3629
8	K21	胃食道逆流症	33977	C54	子宮体部の悪性新生物＜腫瘍＞	3599
9	K25	胃潰瘍	33400	C56	卵巣の悪性新生物＜腫瘍＞	2991
10	M54	背部痛	31506	C61	前立腺の悪性新生物＜腫瘍＞	2786
11	L30	その他の皮膚炎	30472	H35	その他の網膜障害	2752
12	H10	結膜炎	28822	I71	大動脈瘤及び解離	2637
13	G47	睡眠障害	26581	I35	非リウマチ性大動脈弁障害	2610
14	H35	その他の網膜障害	25903	E11	2 型＜インスリン非依存性＞糖尿病＜NIDDM＞	2421
15	I50	心不全	24982	H40	緑内障	2419
16	I20	狭心症	22436	I42	心筋症	2196
17	J30	血管運動性鼻炎及びアレルギー性鼻炎＜鼻アレルギー＞	22430	C25	膵の悪性新生物＜腫瘍＞	2099
18	M81	骨粗しょう症＜オステオポローシス＞，病的骨折を伴わないもの	22240	H33	網膜剝離及び裂孔	2041
19	D50	鉄欠乏性貧血	21062	I25	慢性虚血性心疾患	2017
20	H40	緑内障	16833	N18	慢性腎臓病	2006

3）臨床検査およびバイタルサイン

本対象データセットでは，1003 の臨床検査の種類があり，件数は RBC（赤血球），Ht（ヘマトクリット），WBC（白血球），Hb（ヘモグロビン）が最も多く，PLT（血小板），クレアチニン，AST（アスパラギン酸アミノトランスフェラーゼ），ALT（アラニンアミノトランスフェラーゼ），UN（尿素窒素）が次いで多い状況であった。臨床検査項目のうち特に件数の年次変化が大きかった項目の推移を**図 3** に示した。「推定 GFR（糸球体濾過量）」の件数が特徴的な変動を示していたため医療情報部に確認したところ，2013 年 10 月に新しい検査方法が導入され，従来の方法による結果と区別するため，新しい方法による結果を示す項目名として「eGFRcreat」が利用されたことが判明した。また，2011〜2013 年にかけてリンパ球などいくつかの目視法による検査件数が減少し，機械法による件数が増加していた。

次に，本対象データセットの件数の多かった臨床検査上位 20 項目およびバイタルサインの測定値の要約統計

MDV DPC データ（主傷病名）		
ICD10 3 桁分類	病　名	レコード数
I63	脳梗塞	538751
I50	心不全	538480
C34	気管支及び肺の悪性新生物＜腫瘍＞	513373
I20	狭心症	460463
C16	胃の悪性新生物＜腫瘍＞	432669
S72	大腿骨骨折	413564
J69	固形物及び液状物による肺臓炎	401537
C18	結腸の悪性新生物＜腫瘍＞	375444
K80	胆石症	363831
J18	肺炎，病原体不詳	339508
J15	細菌性肺炎，他に分類されないもの	239821
C50	乳房の悪性新生物＜腫瘍＞	234963
E11	2 型＜インスリン非依存性＞糖尿病＜NIDDM＞	229797
C22	肝及び肝内胆管の悪性新生物＜腫瘍＞	226231
N18	慢性腎臓病	221820
C20	直腸の悪性新生物＜腫瘍＞	215807
H25	老人性白内障	201441
C67	膀胱の悪性新生物＜腫瘍＞	193380
K56	麻痺性イレウス及び腸閉塞，ヘルニアを伴わないもの	190964
C25	膵の悪性新生物＜腫瘍＞	182737

量と分位点を算出した（**表 3-1**）。臨床検査値のうち，CRP（C-reactive protein）では全測定値 103852 件のうち，33.13%にあたる 34406 件で数値データではなくテキストデータが入力されていたが，その多くは検出限界未満を示す「0.04 ミマン」あるいは「0.04＞」であった。その他の項目でもテキストデータは散見されたが，入力された数値データには明らかな異常値は認められなかった。一方，バイタルサインの測定値には，測定者が電子カルテに入力し，確認のプロセスが入らないため，体温ではたとえば 26.5℃などの明らかな入力ミスの値や“0”や“99”などの通常取りえない値が 665/427032 件（0.16%）含まれていた。

臨床検査値およびバイタルサインにおける項目ごとの変動係数は，最小値 2.57（ナトリウム）から，最大値 619.20（AST）であり，幅の広い変動が認められた。臨床検査値の変動は疾患による影響があるが，網羅的に主要な臨床検査項目の要約統計量を把握することで，データの変換の必要性や測定精度に関する特徴の確認が可能となった。

2　データの品質に関する確認

1）欠測数と欠測理由

欠測数と欠測理由について，患者情報ファイルの性別と生年月日に欠測は認められなかった。保険病名ファイルにおける欠測は，全 2175467 件のうち診断名で 47 件（プライバシー保護を目的とした加工処理のため），ICD-10 コードでは 13 件（輸血後感染症の未コードのため）に認められた。病名開始日に欠測は認められなかった。

外来患者ファイルでは，外来科が 3669/5124106 件（0.07%）欠測しており，その多くが 2010 年 1 月に外来を受診している患者であった。

処方ファイルでは，「処方：注射ファイル」で 82351/10693427 件（0.77%），「処方：注射以外ファイル」で 42407/17604889 件（0.24%）に YJ コード[†]の欠測が認められた。そのほか，「処方：注射ファイル」の単位に認められた 15 件の欠測以外，投与量，投与日などすべての項目に欠測は認められなかった。

2）日　付

日付のフィールドにテキストなどの形式が入力されているデータは認められなかった。異常データとしては，生年月日のデータに“1878/11/11”（2010 年 1 月 1 日現在で 130 歳を超える）と入力されている患者が 85 例存在した。それ以外，明らかな異常データは認められなかった。入院日と退院日などの日付の不整合についても認められなかった。

考　　察

1　データの分布や年次推移のバラツキの確認

受診した患者の出生年の分布は，日本の人口動態と類

[†]YJ コードとは，厚生労働省医政局経済課が医療用医薬品で薬価基準に収載された全品目について発表する薬価基準収載医薬品コードとは別に，品目を管理するために統一名収載品目の個々の商品に対して設定されている個別医薬品コード

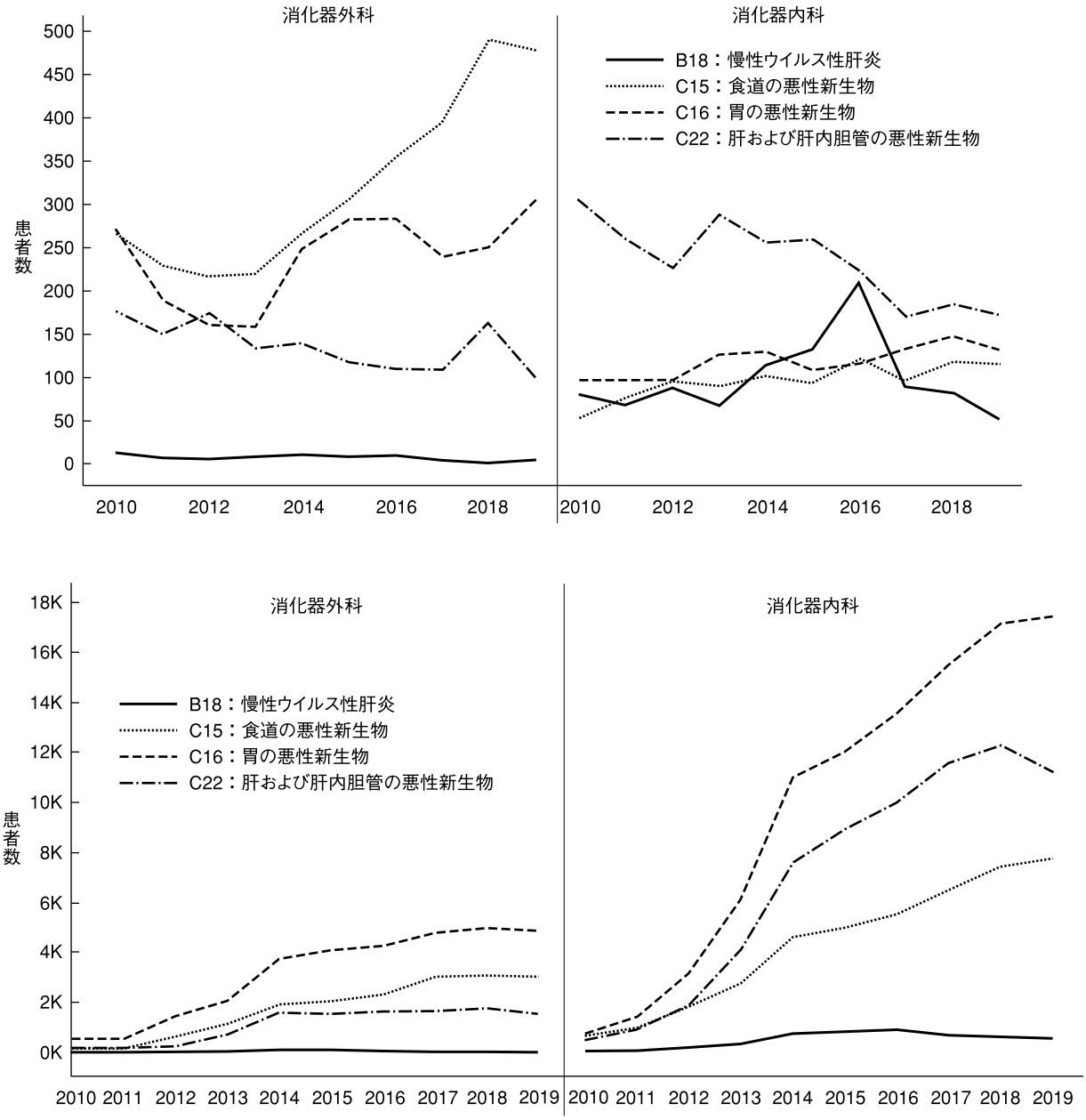

図 2　消化器系疾患による入院患者の疾患別（ICD-10 コード（3 桁分類））件数の年次推移
上）対象データセット，下）MDV データ

似しており，特定の年齢層に偏った集団ではなかった（**図 1-1**）。一方，入院時の年齢では，0 歳児の件数が突出して多く，MDV データでも同様の傾向であった（**図 1-2**）。阪大病院は，ハイリスクや合併症妊娠への適切な診断・治療が行え，かつ無痛分娩にも取り組む大阪府下に 6 施設しかない総合周産期母子医療センターを有することが，0 歳児の件数が多かった原因のひとつであると推察された。入院期間の平均日数の分布は，1〜3 日以内の患者頻度が多く，30 日以上の頻度が低く，その平均値は 18.0 日，中央値は 9.0 日であった（**図 1-3**）。本対象データセットの平均値および中央値は，「一般病床」の現状について〔急性期医療に関する作業グループ第 3 回（1/26）資料〕[6]に記載されている DPC 対象病院の平均在院日数 15.0 日の前後であり，急性疾患または重症患者の治療を 24 時間体制で行う急性期病院としての特徴が示されていた。なお，少数ではあるが長期入院を要する患者も認められ，1000 日を超えた 29 件の主 DPC 傷病名の内訳は，小児の特発性拡張型心筋症など心臓疾患が 17 件と半数以上であり，高度の医療・医療技術を提供する特定機能病院としての傾向も示唆された。

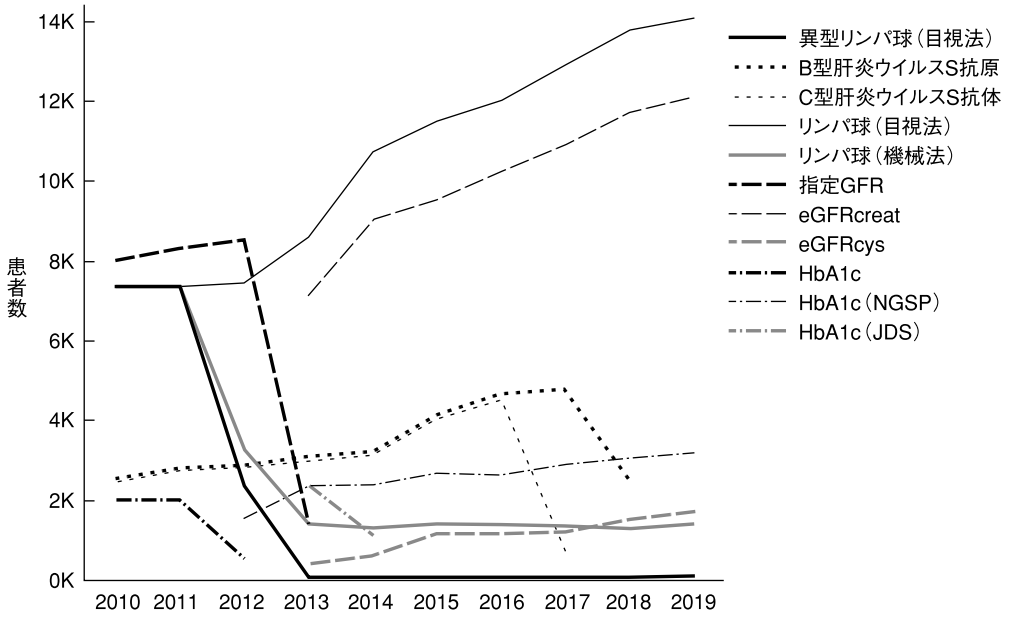

図 3　臨床検査項目の実施件数の年次推移

入院患者の診療科別かつ疾患別患者数の推移では，食道癌や慢性ウイルス性肝炎患者の件数が経時的に大きく変動していた（**図 2**）。たとえば，慢性ウイルス性肝炎の入院件数は，2015 年に経口抗ウイルス薬（DAA 薬）が承認され，販売が開始された翌年の 2016 年を境に，大きく減少していた。そのほか，当該疾患の専門性の高いスタッフの異動や革新的な設備導入の時期と件数が変動する時期が近く，これらの要因が患者数の変動に影響を与える原因のひとつと考えられた。参考として MDV データでも当該疾患の推移を確認したが，本対象データセットのような変動は認められなかった。これらのことから，臨床研究で対象患者をリクルートする場合，統合された医療情報データでは対象患者数の経時推移が一定であっても，特定の施設で患者数が減少してしまう原因の有無やその時期などを知らなければ，研究期間内に十分な被験者の確保ができなくなる可能性がある。

臨床検査項目の推移では，GFR やリンパ球の件数が経時的に大きく変動していた（**図 3**）。たとえば，リンパ球では 2012 年に機械法が導入されたため，以降は機械法の件数が増加していたのに対し，目視法では減少していた。その他の白血球分画など検査方法の変更により，その件数はリンパ球の目視法と機械法の件数と同様に推移していた。HbA1c は 2012 年 4 月に JDS（Japan Diabetes Society）値から国際標準の NGSP（National Glycohemoglobin Standardization Program）値へ変更されたが，NGSP 値は JDS 値から換算できるが同じ値ではない。そのため，本対象データセットでは，変更前後で検査項目が識別できるように JDS 値を HbA1c，NGSP 値を HbA1c（NGSP）と検査名が分けられて記録されていた。仮に，変更前後の検査名が区別されず，ともに HbA1c として取り扱われた医療情報データを用いて臨床研究を行う場合，患者背景や治療効果の推定が正しく行えず，結果や解釈にバイアスを与える可能性がある。したがって，各施設の特徴を明確にするために，検査方法の変更時期や検査名についても十分に確認しておくことが必要である。

2　データの品質に関する確認

本対象データセットのうち，患者情報ファイルと臨床検査ファイルには欠測がまったくなく完全性（completeness）[4]が保たれていた。一方，その他のファイルで認められた欠測データの理由の多くが，病名や薬品名の情報はあるもののコード化する際の対応表にコードが存在しないためであった。たとえば，「処方：注射ファイル」では薬品名および施設で規定されている薬剤コードには欠測なく情報が入力されていたが，医療用医薬品の薬価基準収載医薬品コードである YJ コードへの対応がない，82351/10693427 件（0.77%）が欠測となっていた。この約 8 万件の薬品名は，対応表にない院内製剤やワクチン類，治験薬調剤用の生理食塩水などであった。医療情報データを用いて，臨床試験における対象患者の背景情報を収集する際，標準的なコードにより主な合併症や使用されている治療薬に関する情報は得られたとしても，標準的なコード化がされていない院内製剤やワクチン類の薬剤情報は欠落してしまう可能性があるため，試験治療との相互作用や併用禁止などの確認漏れにつながり，被験者の安全性が損なわれる危険性がある。

また，入院患者ファイルで認められた退院日の欠測

表 3-1　臨床検査値とバイタルサインの要約統計量（対象データセット）

		RBC（赤血球数）	Ht（ヘマトクリット）	WBC（白血球数）	Hb（ヘモグロビン）	PLT（血小板数）
	単位	$\times 10^6 \mu$L	%	$\times 10^3 \mu$L	g/dL	$\times 10^3 \mu$L
	全件数（集計件数）*	114563（114561）	114563（114561）	114563（114561）	114563（114561）	114563（114561）
基本統計量	平均	4.04	37.32	7.44	12.29	226.37
	標準偏差	0.77	6.70	5.97	2.33	99.63
	変動係数（%）	19.04	17.95	80.30	18.93	44.01
分位点	100%最大値	10.22	87.8	625.6	29.6	3492
	75% Q3	4.55	41.6	8.58	13.8	274
	50%中央値	4.09	37.7	6.3	12.4	217
	25% Q1	3.57	33.3	4.82	10.8	166
	0%最小値	0.2	1.9	0.01	0.6	0

		単球（機械法）	好酸球（機械法）	好塩基球（機械法）	好中球（機械法）	K（カリウム）
	単位	%	%	%	%	mEq/L
	全件数（集計件数）*	105788（105776）	105788（105776）	105788（105776）	105788（105776）	105756
基本統計量	平均	6.62	2.52	0.58	65.56	4.19
	標準偏差	3.28	3.06	0.60	14.60	0.59
	変動係数（%）	49.59	121.08	104.01	22.27	14.06
分位点	100%最大値	92.7	87.8	68.4	99.2	22.6
	75% Q3	7.7	3.3	0.7	75.4	4.4
	50%中央値	6.1	1.7	0.5	66.4	4.1
	25% Q1	4.8	0.7	0.3	57.1	3.9
	0%最小値	0	0	0	0	1.4

＊全件数（集計件数）：全件数には一部テキストデータが含まれていることから，統計量は数値データ（集計件数）だけで算出した

		呼吸数	脈拍数	収縮期血圧	拡張期血圧	体温
	単位	回/分	回/分	mmHg	mmHg	℃
	全件数（集計件数）	62191	396873	360341	353305	427032
基本統計量	平均	30.86	78.42	118.80	69.56	36.59
	標準偏差	13.74	18.50	22.55	15.52	1.53
	変動係数（%）	44.53	23.60	18.98	22.31	4.18
分位点	100%最大値	99	992	999	887	99
	75% Q3	42	86	132	79	36.9
	50%中央値	26	76	117	69	36.6
	25% Q1	20	67	104	60	36.3
	0%最小値	0	0	0	0	0

クレアチニン	AST（GOT）	ALT（GPT）	UN（尿素窒素）	リンパ球（機械法）
mg/dL	U/L	U/L	mg/dL	%
108279	107966（107964）	107821（107490）	107227（107225）	105776
1.02	48.52	34.96	18.80	24.72
1.32	300.46	156.35	13.92	12.95
129.84	619.20	447.24	74.08	52.39
35.69	23200	15072	316	97.4
0.96	33	28	20	31.7
0.74	23	17	15	23.3
0.58	18	12	12	15.6
0.05	3	3	1	0

Na（ナトリウム）	CRP	Alb（アルブミン）	Cl（クロール）	TP（総タンパク）
mEq/L	mg/dL	g/dL	mEq/L	g/dL
105357	103852（69446）	100707	97965	95925
139.09	1.99	3.81	104.85	6.84
3.57	4.32	0.64	4.05	0.83
2.57	216.56	16.92	3.86	12.17
262	57.94	6.5	168	15.4
141	1.48	4.3	107	7.3
139	0.3	3.9	105	6.9
138	0.1	3.5	103	6.5
44	0.04	0.1	42	0.2

695 件のうち 99 件は，当該研究のデータ抽出期間の終了時点において入院を継続していた対象患者のものと考えられる。そのほか，退院日の残り 596 件や入院診療科および退院診療科の 51 件，外来患者ファイルの外来科 3669 件，看護データの項目名 1 件の欠測については，いずれも欠測理由を本対象データセットから明確に特定することはできなかった。

日付に関するデータとして，検査日や投与日，正確な疾患の発症日とは異なるが代替として利用が可能である病名開始日などには欠測は認められなかった。ただし，生年月日には，本対象データセットでは取りえない“1878/11/11”と入力されたデータが散見された。これは，妊娠中の患者が受診し，胎児のカルテが作成された場合に，胎児の生年月日として，一時的に“1878/11/11”が入力され，出産後の来院時にその児の正しい生年月日に修正するという，阪大病院の独自ルールによるものであった。このような各施設独自ルールを知らずに医療情報データを利用すると，誤って本来は存在しない 140 歳代の患者が含まれた集団を解析に利用し，誤った解釈につながる危険性がある。そのほか，入院日と退院日などの前後関係をもつ日付の不整合や明らかな異常値は認められず，日付のデータは信頼性（validity）[4]が高いことが示された。

臨床検査ファイルについては，阪大病院では臨床検査部の担当者が HIS に入力される臨床検査値を確認する作業を行っているため，生物学的にありえない値は認められず，論理的妥当性（plausibility）[4]が保証されたデータであった。なお，臨床検査値の CRP は，本対象データセットでは検出限界未満の値の場合に阪大病院の独自ルールでテキストデータが入力可能とされていた。その

表 3-2　臨床検査値の要約統計量（MDV データ）

		RBC（赤血球数）	Ht（ヘマトクリット）	WBC（白血球数）	Hb（ヘモグロビン）	PLT（血小板数）
	単位	×10000/μL	%	/μL	g/dL	×10000/μL
	全件数	15114201	15112534	15108386	15102394	15102576
基本統計量	平均	399.59	36.58	6805.80	12.16	21.93
	標準偏差	77.46	6.44	10800.24	2.29	9.81
	変動係数（%）	19.38	17.62	158.69	18.85	44.73
分位点	四分位範囲	103	8.7	3250	3.1	10.1
	100%最大値	45400	430	16700000	1200	2160
	75% Q3	453	41.1	7900	13.7	26.3
	50%中央値	405	37.1	6000	12.3	20.9
	25% Q1	350	32.4	4650	10.6	16.2
	0%最小値	0	0	0	0	0

		単球（機械法）	好酸球（機械法）	好塩基球（機械法）	好中球（機械法）	K（カリウム）
	単位	%	%	%	%	mEq/L
	全件数	8119961	7801441	7522105	7150036	14114823
基本統計量	平均	6.86	2.84	0.55	63.62	4.21
	標準偏差	3.86	3.13	0.57	14.12	1.08
	変動係数（%）	56.25	110.10	103.00	22.19	25.66
分位点	四分位範囲	3.2	2.7	0.5	18.4	0.6
	100%最大値	113	100	100	100	3367.8
	75% Q3	8	3.7	0.7	73.2	4.5
	50%中央値	6.2	2	0.4	63.6	4.2
	25% Q1	4.8	1	0.2	54.8	3.9
	0%最小値	0	0	0	0	0.2

ため，たとえば，医療情報データが統合される際に，数値であるべきデータに入力されたテキストデータが除かれ，欠測データとされていても施設の独自ルールにより実際は検出限界以下であるという重要な情報が欠落し，結果や解釈にバイアスを与える可能性がある。施設ごとにルールの有無や内容が異なるため，医療情報データが統合される前に，各施設の独自ルールの情報を入手しておく必要がある。

バイタルサインに該当する看護ファイルの「測定値」については，患者の診察や治療に利用する本来の目的としては十分な内容が記載されているものの二次利用する際にはデータクリーニングや異常値の取り扱いや採否などの注意が臨床検査データ以上に必要であることが示唆された。

3　本研究の限界

本研究は阪大病院単施設のデータを対象にした研究で

クレアチニン	AST（GOT）	ALT（GPT）	UN（尿素窒素）	リンパ球（機械法）
mg/dL	U/L	U/L	mg/dL	%
14964246	14338076	14286295	13937894	8134175
1.22	33.64	29.33	19.37	26.00
2.35	124.04	76.12	21.93	13.44
192.08	368.76	259.54	113.20	51.69
0.4	12	16	9	17.6
5947.3	40854	17355	61764	202
1	30	28	21	33.9
0.8	22	18	15.5	25.1
0.6	18	12	12	16.3
−0.1	−44	−24	−0.1	0

Na（ナトリウム）	CRP	Alb（アルブミン）	Cl（クロール）	TP（総タンパク）
mEq/L	mg/dL	g/dL	mEq/L	g/dL
14006951	9729518	11377248	13491473	11115403
139.78	2.38	3.69	104.19	6.71
9.76	4.56	0.77	8.20	2.53
6.98	191.75	20.85	7.87	37.70
4	2.3	1	5	0.9
33754	95	730.8	25879	7982
142	2.4	4.2	107	7.2
140	0.4	3.9	105	6.8
138	0.1	3.2	102	6.3
23	−1.7	0	12	0

あることから，ある一定の特徴や性質については明確にできたものの，より詳細に特徴や性質を明確にするには，多施設のデータやそれらを統合した医療情報データとの比較が必要であると考える。

結　　論

現在では多くの病院が電子カルテを導入し，容易に日常診療下で収集される医療情報が蓄積されつつある。その豊富に収集された医療情報データを医学研究や医薬品・医療機器の開発等に活用して，効率的に公衆衛生の向上につなげることが期待されている。しかしながら，医療情報データでは，収集されたデータの内容や妥当性，収集された元々の理由など研究に対する質の問題のほかに，データベースの集団と研究対象となる母集団との関係を明確にする必要があるため，いかに研究に適した医療情報データを選択するかが重要となる。また，多

施設の統合されたデータベースを選択し利用する際には，データチェックの段階でテキストデータが混在することでエラーが発生する場合や年齢が異常値を取るような場合，データ加工処理において元データまで原因を探索し，対処方法を検討する必要がある。

今回の結果から医療情報データの特徴や性質を事前に把握することで，エラーの原因の特定や対処方法の検討は容易になることが期待される。また，データチェックや加工処理を行う前の段階で，各施設の医療情報データの特徴や性質を把握しておくことで，研究対象の母集団に近いデータベースの集団に対する適格基準の設定やデータベースの選択も効率的に行えることが期待される。

今後，今回の結果をもとにデータベースを選択する段階で，データチェックやデータ加工処理が効率的に行うことが可能な医療情報データの特徴や性質に関するチェックリストを開発していく必要がある。将来的には，開発したチェックリストを利用し，研究目的に適切なデータベースの選定，データ抽出，データチェック，データ加工が効率的に進められ，医薬品・医療機器開発等に有益な情報が提供される土壌が構築されていくことを期待している。

抄　録

背景　近年，医薬品・医療機器開発などにリアルワールドデータ（RWD）を利活用することへの期待が高まっている。しかし，電子カルテデータから構成される医療情報データは研究目的や研究計画に沿って収集されるデータではないため，研究に期待される品質でデータが収集されていない問題や関心のある対象集団とは異なる偏ったデータとなる可能性がある。そのため，医療情報データを利活用する際には，そのデータが有する特徴や性質を事前に把握したうえで，研究目的に対して適当なデータであるかの確認および選択が重要となる。さらに，データベース選択の前段階にその特徴や性質を把握することで，その後のデータチェックや加工段階における作業の効率化につながることが期待できる。そこで，われわれは大阪大学医学部附属病院（阪大病院）の医療情報データを二次利用する観点から，その特徴や性質を明確化することを目的とした調査研究を実施した。

方法　阪大病院が有する過去 10 年間に初診受診した全患者の医療情報データのなかから，調査・研究で利用頻度の高い項目を抽出した。各項目のデータのうち，患者背景情報である出生年や入院時年齢，疾患名，臨床検査データなどの分布や経時推移，欠測データや異常データの頻度とその理由などについて検討を行った。

結果・結論　本研究の結果，阪大病院の医療情報データには，いくつか施設独自ルールにより入力された項目やテキストデータと数値データが混在する項目が含まれているほかに，収集されたデータの欠測状況についても網羅的に確認することができた。データベースを利用する際には，データチェックの段階でテキストデータが混在することでエラーが発生する場合や年齢が異常値を取るような場合，データ加工処理において元データまで原因を探索し，対処方法を検討する必要がある。本研究結果から得られた医療情報データの特徴や性質を事前に把握しておくことで，エラーの原因の特定や対処方法の検討は容易になることが期待される。また，データチェックや加工処理を行う前の段階に，各施設の医療情報データの特徴や性質を把握しておくことで，研究対象の母集団に近いデータベースの集団に対する適格基準の設定やデータベースの選択も効率的に行えることが期待される。

【利益相反】

荒木浩之，佐藤倫治，飛田英祐（塩野義製薬株式会社の共同研究講座に所属）。惟高裕一，長谷川貴大，小林典弘（塩野義製薬株式会社社員）。山田知美（大阪大学医学部附属病院，開示すべき利益相反はない）。

文　献

1）Joseph A DiMasia, Henry G Grabowskib, Ronald W Hansenca. Innovation in the pharmaceutical industries: New estimates of R & D costs. J Health Econ 2016; 47: 20–33.

2）日本薬剤疫学会．日本における臨床疫学・薬剤疫学に応用可能なデータベース調査．ver. 1.0_2020［cited 17 January，2022］. https://sites.google.com/view/jspe-database-ja2020/%E3%83%9B%E3%83%BC%E3%83%A0

3）武田理宏，真鍋史郎，松村泰志．電子カルテデータ二次利用の現状と課題．生体医工学 2017; 55（4）: 151–8.

4）Gregory Daniel, Christina Silcox, Jonathan Bryan, et al. Characterizing RWD Quality and Relevancy for Regulatory Purposes［cited 27 September, 2021］. https://healthpolicy.duke.edu/sites/default/files/2020-03/characterizing_rwd.pdf

5）厚生労働省．令和元年（2019）人口動態統計の年間推計［cited 27 September，2021］. https://www.mhlw.go.jp/toukei/saikin/hw/jinkou/suikei19/dl/2019gaiyou.pdf

6）厚生労働省．「一般病床」の現状について（急性期医療に関する作業グループ第3回(1/26)資料)［cited 27 September, 2021］. https://www.mhlw.go.jp/stf/shingi/2r9852000002nakz-att/2r9852000002naqf.pdf