CTDS（Clinical Trial Data Sharing）に係る

データ非特定化の手法検討

日本製薬工業協会

データサイエンス部会

2020 年度 継続タスクフォース 3

2021 年 2 月

目次

略語一覧表

略語	英語	日本語
CDISC	Clinical Data Interchange Standards Consortium	―
CSDR	ClinicalStudyDataRequest.com	―
DSA	Data Sharing Agreement	データ共有合意書
EMA	European Medicines Agency	欧州医薬品庁
HIPAA	Health Insurance Portability and Accountability Act of 1996	医療保険の携行性と責任に関する法律
IOM	U.S. Institute Of Medicine	―
PDS	Project Data Sphere	―
PhUSE	Pharmaceutical Users Software Exchange	―
SNS	Social Networking Service	ソーシャル・ネットワーキング・サービス
SDTM	Study Data Tabulation Model	―

1　はじめに

臨床試験の個別被験者データの共有（Clinical Trial Data Sharing：CTDS）にあたり，研究者に臨床試験データを提供するためには，通常，個人情報保護に配慮したデータの匿名化／非特定化処理が必要となる。CTDS の昨今の進展については既報の製薬協医薬品評価委員会データサイエンス部会（以下，DS 部会）による CTDS についてのレポート[1] 及び製薬協ニューズレター[2] を参照いただきたい。また，CTDS に関連する規制の動向とその対応方法については別報告書を参照いただきたい[3]。

データの匿名化／非特定化処理にはさまざまなものがあり，処理されたデータが個人情報にあたるか否かその解釈を含めて幅があるが，本報告書においてはデータにひもづいた本人が特定されるリスクを低減するため本人の特定を困難にする処理の意味で用いることとし，単に「非特定化」と記載する。

前述のレポートでも触れている通り，再特定化のリスク（匿名化／非特定化したデータから個人が特定されてしまうリスク）を下げようとすると研究者にとってのデータの有用性が損なわれ，逆に有用性を高めようとするとリスクも高まる。非特定化の具体的な方法や，どの程度のリスクを受け入れ可能とみなすかは，非特定化を行う企業の判断に委ねられているのが現状である。その一方で，非特定化の処理がどのように行われ，どの程度有用性に影響するのか，具体的にイメージすることは難しい。これは企業が非特定化のプロセスを定めるうえでの障害となり，また非特定化されたデータを利用する者がデータにどのような制限があるのかを理解するうえでも障害となりうる。そこで，本報告書は，具体的に非特定化の処理とはどのように行われ，どの程度有用性に影響するのかを具体的にイメージできるようになることを目的として作成した。

非特定化の方法論には各種あるが，本報告書ではルールベースの方法とリスクベースの方法について触れる。2 章では非特定化の手法の概要として，非特定化のプロセスと方法論について述べる。3 章及び 4 章では，非特定化のプロセス（2 章で紹介）の手順の 1 つである，念頭に置くべき攻撃パターン，受入れ可能なリスクの閾値についてそれぞれ触れる。5 章では非特定化を行った結果を示し，リスクベースの方法については，主に k-匿名化の方法を適用することにより，再特定化のリスクがどの程度変動するのか，検討した結果も併せて示す。

2　非特定化の手法の概要

2.1 非特定化のプロセス

個人情報の取り扱いは，各国・地域で満たすべき要件がある。US に関しては，HIPAA からプライバシールールとして，セーフハーバー方式の考え方が出されている。セーフハーバー方式は，18 種類の識別子を削除すれば，非特定化されたデータ（非特定化された健康情報）として取り扱えるが，準識別子の組み合わせにより個人が特定されるリスクがあり，一般的に 18 種類の識別子を削除するだけでは十分ではない[4]。また，HIPAA には，Expert Determination による非特定化の手法も示されている。2015 年に PhUSE から，HIPAA も含むガイダンス類を考慮した非特定化標準が出されている[5]。臨床試

験データの非特定化，特に SDTM データに対しての非特定化には大変参考になる。JPMA でも PhUSE の非特定化を要約した報告書[1]を作成しているため，詳細は参照いただきたい。

臨床試験データの非特定化を実施するには，複数のプロセスが必要となる。そのプロセスは，米国医学アカデミー（National Academy of Medicine：NAM）（旧 Institute of Medicine：IOM）のレポート[6]に示されており，前述のレポートで紹介されている。本報告書では，NAM のレポートの翌年にリリースされた Health カナダのガイドライン[7]を参考に簡潔に表 1 に示す。

表 1： 非特定化のプロセスの概要

番号	プロセス	概要
1	データの共有モデルの決定： determine the release model	データを公共の場，半公共の場（semi-publicly：データの共有先は特定されているが，追加のプライバシー保護やセキュリティ対策を要求する利用規約を要求していない場合），非公共の場のいずれで共有するかを確認する。共有する場により非特定化のレベルを考慮する。
2	変数を識別する： classify variables	データセットに含まれる変数を確認する。直接識別子は，個人の特定につながる変数であるため削除や再コード化を行う。準識別子は単独では個人の特定にはつながらないが，組み合わせにより個人の特定につながることに留意する必要がある。攻撃者はデータセット内のデータだけでなく，公開されているデータと紐づけるかもしれないことにも留意する。
3	受け入れ可能な再特定化のリスクの閾値を決定する： determine an acceptable re-identification risk threshold	個人のプライバシーを保護するために，データの非特定化が必要となる。再特定化のリスクをふまえ非特定化の度合い，すなわち，再特定化のリスクの閾値を決める。 プライバシーの侵害 / 再特定化リスクの閾値 / セルサイズ* Low / 0.1 / 10 Medium / 0.075 / 15 High / 0.05 / 20 *準識別子の値が同じであるレコード数
4	データのリスクを計算する： measure the data risk	再特定化のリスクの閾値を決めたら，実際のデータセットの再特定化のリスクを計算する。まず，データごとの再特定化のリスクを計算し，データの共有モデルにより適切なリスク測定方法（最大リスク，厳格な平均リスク）を適用する。 （各データの再特定化のリスク） データの再特定化のリスク＝1÷等価クラスのサイズ 等価クラスは当該レコードと準識別子の値が同じデータを指し，当該レコードの再特定化のリスクは上記の式で計算される。等価クラスのレコードが 5 であれば，再特定のリスクは 0.2 となる。最大リスクは，各データの再特定化のリスクの最大値により算出される。平均リスクは，すべてのレコードの再特定化のリスクの平均により算出される。

番号	プロセス	概要
		（リスク測定方法） 公共の場/半公共の場：最大リスク 非公共の場：厳格な平均リスク（Strict Average Risk）［厳格な平均リスクでは，平均リスクに加えて最大リスクも同時に考慮する。各データの再特定化のリスクを考慮し，0.33（等価クラスのサイズ 3）がよく提案されるが，実際には，0.5（ユニークなレコードが存在しない）も使われうる］
5	コンテキストリスクを計算する：measure the context risk	コンテキストリスクとは，再特定化の攻撃にさらされるリスク（確率）であり，データをどのように共有するか等の状況に左右される。 公共の場：コンテキストリスク 1 非公共の場：次の 3 つを考慮してリスクを決める必要がある。①内部の者による意図的な攻撃［Data Sharing Agreement によるプライバシーやセキュリティのコントロール及びデータ受領者の再特定化の動機や能力を加味して 0.05～0.6 が考慮される］，②個人を知る者による意図しない認識（inadvertent recognition of an individual in the data set by an acquaintance）［データの中に友人・同僚・家族・知人が含まれていれば，意図せず個人を再特定してしまうことがある。その確率は，任意のデータ受領者がデータの中の誰かを知る確率として定義される］，③データの漏洩（data breach）［データ受領者の業界での利用可能なデータ漏洩率が参照される］ 半公共の場： 非公共の場で考慮されたリスクを同様に検討できる。ただし①のリスクは高く見積もる必要がある。
6	全体のリスクを計算する：calculate the overall risk	データのリスク（Step 4）とコンテキストのリスク（Step 5）の積により全体のリスクを計算する
7	データの非特定化を行う：de-identify the data	全体リスクが再特定化リスクの閾値内におさまるように非特定化の調整を行う（データの一般化（generalization）や抑制/削除（suppression）により等価クラスのサイズを調整する）。
8	データの有用性を評価する：assess data utility	非特定化加工を適用した量とデータの有用性はトレードオフの関係にある。全体リスクが閾値内におさまるために適用したデータの一般化や抑制/削除は，異なるやり方や組み合わせでも達成できる。もし 5%超のデータに対して抑制/削除を適用していないのであれば，一般的には，抑制/削除を一般化の前に考慮した方がデータの利用可能性が高くなる。5%超のデータに対して適用した抑制/削除や一般化はその組み合わせを見直すことにより，全体リスクを維持したままデータの利用可能性を高められる可能性がある。
9	プロセスを文書化する：document the process	データ共有のたびに適用した非特定化は違うこともあり，どのように非特定化したか説明できるように文書化しておく（社内外へ説明責任を果たすためなど）。

2.2 リスク評価 「リスクベースの方法」と「ルールベースの方法」

前節のプロセスに基づいて試験ごとに非特定化データを作成する方法を本報告書では「リスクベースの方法」と呼ぶことにする。一方，プロセス 3 から 8 を試験ごとのデータに応じて実施するのではなく，個々の企業で定めた 1 つのルール，つまり，変数毎に決められた特定の手法を適用し，非特定化したデータを作成する方法も考えられる。本報告書ではこれを「ルールベースの方法」と呼ぶことにする。リスクベースの方法を用いる場合は，後述する「企業の判断で設定しなければならない値」が多く，企業が行った全ての臨床試験でリスクベースの方法を適用するのは非常に難易度が高く，また作業が煩雑になるため企業の負担が大きいことが想定される。一方，セキュアなサイトを通じてデータ共有合意書（Data Sharing Agreement：DSA）を結んだ研究者にのみデータを提供する場合，ルールベースの方法で十分という判断もあるのかもしれない。仮にルールベースの方法だったとしても，①いくつかの人種のカテゴリをまとめて Other と定義する，②1 つの試験施設から 10 例未満しか症例が参加していない場合には施設の情報をブランクにする，③試験規模が一定の例数を下回る場合はそもそも共有対象としないなど，事前に取り決めておく必要事項はあるが，具体的なリスク評価の方法やリスクの閾値の設定が各企業に委ねられている現状において，この簡便なルールベースの方法を用いることは，十分に選択肢と思われる。

再特定化のリスクを評価して非特定化されるという面で，望ましいとされるのはリスクベースの方法でデータを作成することであろう。下記にリスクベースの方法（表 1）のプロセス 1，2，4 について補足する。プロセス 3 は，4 章，プロセス 5 は，3 章，プロセス 6～8 は，5 章を参照されたい。

プロセス 1 にあるように，データをどこで共有するのかは重要である。各種 CTDS ポータルサイトの内，CSDR（ClinicalStudyDataRequest.com）[8]や Vivli[9]などは，研究リクエストを提出した研究者にのみデータを提供する close な環境を提供している。一方，Project Data Sphere[10]，オーストラリアの DB（Population Health Research Network）[11]などは，データのダウンロードを認めており，open な環境を提供している。どのような環境でデータを提供するかにより，再特定化のリスクも変わり，要求される非特定化のレベルも変わることに留意する必要がある。

プロセス 2 にあるように，データを非特定化するためには，各変数に対して，直接識別子，準識別子，センシティブ・データの区別をする必要がある。 PhUSE のガイダンス（De-Identification Standard for CDISC SDTM3.2）[12]での用語の解説を表 2 で示す。直接識別子は，削除もしくは，再コード化，準識別子は，削除・一般化などを行い，他の情報と関連づけて再特定化されないようにする必要がある。

表 2： PhUSE ガイダンスの用語の解説

用語	定義
直接識別子（Direct Identifier）	単独もしくは複数で個人を一意的に特定することができる情報 例：症例番号，社会保障番号，電話番号，正確な住所など いかなる直接識別子は，削除もしくは再コード化をしなければならない。
準識別子（Quasi Identifier）	準識別子は，単独では個人の識別はできないが，他の情報と関連づけて使うと高い確率で個人を識別することができる背景情報 例：ベースラインの年齢，人種，性別，イベント，特定の調査結果など
準識別子レベル 1（Quasi Identifier Level 1）	時間の経過により変化する可能性がない情報で，他のソースから明らかもしくは利用可能な情報。典型的には，被験者背景情報 例：性別，ベースラインの年齢，国，BMI など
準識別子レベル 2（Quasi Identifier Level 2）	時間の経過とともに変化する可能性がある，経時情報 例：測定値，イベントなど
センシティブ・データ（Sensitive Data）	データの漏えいや再特定された場合に，雇用，評判，保険加入権利，自尊心や収入の減少に関して個人に損害をもたらすあらゆるデータ（例：アルコール中毒/薬物乱用やあらゆる危険な行動の既往，性病の既往など）

PhUSE のガイダンスでは，各変数に対して，基本ルール（primary rule）と代替ルール（alternative rule）を提案しており，提供者が研究目的や再特定化のリスクを勘案し選択していく必要がある。

プロセス 4 では，再特定化のリスクを計算する。一般的には，k-匿名化の考え方（最小のセルサイズと同じ）に基づき行う。ただ，次節に示すように k-匿名化が行われていたとしても，万全ではないことに留意しなければならない。

2.3 非特定化の方法論

データの非特定化方法論[13] は，2017 年 12 月に製薬協 DS 部会で開催したシンポジウムの内容（統計数理研究所 南和宏先生の発表内容）が参考になる。非特定化の方法論を理解するため，性別・年齢・身長という準識別子を含んだデータ（表 3）を用いる。「k-匿名化」は，同一の値を取る準識別子（性別・年齢・身長など）の組み合わせを持つレコードが必ず k 個以上となるように作成し，レコードの識別を k 個以下に絞り込ませないという非特定化で，そのレコードに対する再特定化のリスクを 1/k にすることができる。この理由は，直接識別子のみを削除すれば，直接個人を特定できるデータはデータ上からはなくなるが，準識別子の組み合わせで個人を特定する属性の組み合わせが存在し，それにより個人を特定できるからである。たとえば，男性，45 歳，175 cm，…と準識別子が追加されていくことで，該当する個人が一意に特定できてしまう（表 3）。そこで，年齢や身長をカテゴリ化することにより同一の値を取る準識別子のレコード数を増やし，2-匿名化とした。これにより，各レコードの再特定化のリスクを 1/2 以下にすることができる（表 4）。ただし，k-匿名化を当てはめれば，再特定化のリスクが 1/k になることを期待するが，絶対ではない。例えば，準識別子のグループ（性別・年齢・身長）の中で病気（センシティブ・データ）が全員違えば，ある個人の病気（センシティブ・データ）を特定するには 1/k の確率となるが，もし全員の病気が同じであれば，同一の値を取る準識別子のレコードが k 個あったとしても，個人の病気（センシティブ・データ）を特定する確率が 1（自動的に判明）となってしまい非特定化をしたことにはならない（表 5）。その場合に「l-多様性」の考え方が重要となる。準識別子グループ（性別・年齢・身長）が病気（センシティブ・データ）に関して少なくとも l 個の異なる値を含めば l-多様性を満たす。例えば，データに含まれる全ての準識別子グループが 2 つ以上のデータの多様性を満たすとき，そのデータは 2-多様性を満たす（表

6）。ただ，l-多様性を実現することが困難な場合がある。たとえば，全レコードに対して病気（センシティブ・データ）の割合が 1%（つまり HIV でない 99%，HIV 1%）のデータであれば，そもそも準識別子のグループすべてで l-多様性の実現（すべての準識別子のグループ内で，全体で 1%の病気を含むこと）は困難であろう。また，全体集団での病気の有無が半々（なし 50%，あり 50%）であった場合，その病気の有無がある特定の準識別子グループ（たとえば若年者）で全体と異なり（なし 10%，あり 90%），攻撃者が当該人物の準識別子グループを知っていた場合は，病気の有無を高い確率で推測できてしまう。

給与を事例として考えた場合（表 7），準識別子グループ 1 では全体の集団に対して給与が下位に位置している。特定の人物について，給与が低いという事実を知っている場合，その情報を元に病気（センシティブ・データ）［胃に病気があること］が類推される可能性がある。全体分布と各準識別子グループの分布が同様の分布であればそのような観点からの再特定化のリスクを減らすことができる。「t-近似性」は全体の分布と各準識別子のグループの分布の距離が t 内であることを意味する。よって t が大きいと分布の乖離が大きいことを意味する。距離 t を計算する方法は，地球移動距離 (EMD: Earth Mover's Distance)，Kulback-Leibler 距離[14]などがある。

表 3：準識別子（性別・年齢・身長）を持つデータ（当該する個人がユニークに特定される事例）

k-匿名化は，同一の値を取る準識別子（性別・年齢・身長）のレコードを k 個以上にすることであるが，この事例では k=1 となっており準識別子の情報から当該する個人がユニークに特定される。

ID	性別	年齢	身長	病気（センシティブ・データ）
1	Male	45	175	肝炎
2	Male	40	171	HIV
3	Male	48	179	ループス腎炎
4	Male	49	178	クローン病
5	Female	…	…	…

表 4：2-匿名化の事例（準識別子をカテゴリ化）

準識別子をカテゴリ化することにより k-匿名化の k を調整できる。表 3 のデータの年齢を 10 歳刻み，身長を 5 歳刻みのカテゴリデータにすることにより，同一の値を取る準識別子のレコードが 2 個となった。

ID	性別	年齢	身長	病気（センシティブ・データ）
1	Male	40〜49	171〜175	肝炎
2	Male	40〜49	171〜175	HIV
3	Male	40〜49	176〜180	ループス腎炎
4	Male	40〜49	176〜180	クローン病
5	Female	…	…	…

表 5：同一の 値を取る準識別子（性別・年齢・身長）の組み合わせを持つレコードを 4 個の事例

k-匿名化により，同一の値を取る準識別子のレコードに対する再特定化のリスクを 1/k 以下にすることができる。ただし，下記の事例は，同一の値を取る準識別子のレコードで病気（センシティブ・データ）が同一となっており，k-匿名化が無効となっている。

ID	性別	年齢	身長	病気（センシティブ・データ）	k=4 であるが病気（センシティブ・データ）が同じであり,非特定化の意味がない状態
1	Male	40～49	170～180	HIV	
2	Male	40～49	170～180	HIV	
3	Male	40～49	170～180	HIV	
4	Male	40～49	170～180	HIV	
5	Female	…	…	…	

表 6：l=2 の l-多様性を満たす事例

同一の値を取る準識別子グループ（性別・年齢・身長）が病気（センシティブ・データ）に関して少なくとも 2 個の異なる値を含むデータとなっている場合，2-多様性を満たす。

ID	性別	年齢	身長	病気（センシティブ・データ）	すべての準識別子のグループで病気（センシティブ・データ）の種類が 2 つ以上であれば，l-多様性は，l=2 となる。
1	Male	40～49	170～180	HIV	
2	Male	40～49	170～180	HIV	
3	Male	40～49	170～180	インフルエンザ	
4	Male	40～49	170～180	インフルエンザ	
5	Female	…	…	…	

表 7：準識別子のグループによって給与の類推がしやすくなる事例

準識別子グループ 1 では全体の集団に対して給与が下位に位置している。特定の人物について，給与が低いという事実を知っている場合，その情報を元に病気（センシティブ・データ）［胃に病気があること］が類推される可能性がある。

ID	給与	全体の集団	準識別子グループ 1	準識別子グループ 2	準識別子グループ 3	病気（センシティブ・データ）
1	3K	○	○			胃潰瘍
2	4K	○	○			胃炎
3	5K	○	○			胃がん
4	6K	○		○		胃炎
7	7K	○			○	気管支炎
6	8K	○		○		気管支炎
9	9K	○			○	肺炎
10	10K	○			○	胃がん
5	11K	○		○		インフルエンザ

3　念頭に置くべき攻撃パターン

3.1 コンテキストリスクの評価

リスクベースの非特定化をおこなうにあたり，まずはデータがどのように共有されるかといった状況に応じて発生しうる再特定化攻撃のリスク（コンテキストリスク）がどの程度あるのかを測っておく。コンテキストリスクの定量化のためには，データセットがおかれている環境から，「3.2 攻撃の種類とリスクの考え方」に例示するように，どのような攻撃（再特定化）が考えられるかを仮定しておくことが必要となる。

これには法規制に関しての専門家の意見のみならず，当該試験や疾患領域について理解しているクリニカルサイエンティスト，試験統計家，メディカルライター，メディカルドクターらの関与も必要となる。試験が対象としている疾患が慢性疾患か，致死性の疾患か，希少疾患か，遺伝する病気かなどの条件によって，どの情報がセンシティブ・データに該当するかが異なる可能性もある。具体的に関与が必要と想定される部門・関係者および必要な文書やルールについては既報のニューズレター[2]にまとめているので参照いただきたい。

3.2 攻撃の種類とリスクの考え方

起こりそうな攻撃として，以下の 4 つの攻撃を仮定する。これにより全体のリスク（再特定される確率：Pr(re-id)）を算出することができる[15]。攻撃 1〜3 はデータが非公共の場で共有，攻撃 4 はデータが公共の場で共有されていることを想定している。数式の Pr (X)とは X の確率，Pr (Y|X)とは，X であるという条件の下で Y の確率をそれぞれ示す。

1. データ受領者が故意にデータの再特定を試みる場合

   Pr(attempt)×Pr(re-id|attempt)

   攻撃が試みられる確率（攻撃者の動機と能力が影響）と再特定化が成功する確率の積

2. データ受領者が故意でなく（あるいは無意識に）データを再特定してしまう場合

   Pr(acquaintance)×Pr(re-id|acquaintance)

   知人（隣人，親戚，有名人など）が含まれる確率と再特定化が成功する確率の積

3. データ受領者の施設でデータ侵害があり，データが外部に漏れてしまう場合

   Pr(breach)×Pr(re-id|breach)

   データ侵害が発生する確率と再特定化が成功する確率の積

4．攻撃者がデータに対して，デモンストレーション攻撃（データから個人を再特定可能であることを攻撃者が誇示したいために行われる攻撃）を仕掛ける場合

Pr(re-id)

再特定化確率のみを考える。

上記により個々のレコードのリスクを評価し，データ全体に対するリスク（Pr(re-id|attempt)，Pr(re-id|acquaintance)，Pr(re-id|breach)または Pr(re-id)）を算出する。算出にあたり，データが非公共の場にある想定の攻撃 1～3 の場合は平均リスクを適用し，データが公共の場にある想定の攻撃 4 の場合は最大リスクを適用する。

- 最大リスク：データ全体でリスクが最大のレコードのもの
- 平均リスク：各レコードのリスクの平均

攻撃 1～3 を想定した場合に Pr(attempt)，Pr(acquaintance)，Pr(breach)の各確率を設定するうえで明確な根拠となる数字はなく，非特定化を行う状況を踏まえて常識や経験をもとに各企業はその試験にふさわしい値を決める責任がある。常識や経験の中には，試験から得られる情報以外を考慮することも含まれる。米国の選挙人名簿は購入できることはよく知られている話であるし，日本でも公職選挙法に基づいて選挙人名簿抄本の閲覧は可能である。これは選挙の透明性を確保することが目的の一つで同様の規制は各国にあるが，日本の場合は，各市区町村で所定の手続きを行えば氏名・住所・生年月日・性別の閲覧が可能である。

偶然に知り合いが含まれる可能性を考慮する攻撃 2 を例にとると，「人間が安定的な社会関係を維持できるとされる人数の上限が 150 人程度である」という仮説（この数はダンバー数と呼ばれる）を設定することで，他の 2 つよりも比較的容易に直接計算することができる。当該疾患の有病率 p とダンバー数が分かると，$Pr(acquaintance) = 1 - (1 - p)^{150}$ となる。性別に特有の乳癌や前立腺癌のような疾患であれば，そのことを考慮して 150 の代わりに 75（知り合い 150 人のうち男性，もしくは女性が半分と仮定）を用いることも妥当と言えるかもしれない。

3.3 SNS との連携によるリスクの増大

個人の情報にアクセスする方法として，公的な手続きによる方法のほかに，最近は Facebook，Twitter，Instagram 等の SNS の普及により多くの人が手軽に個人の情報を投稿し，不特定多数がその情報にアクセスできる状況にある。

SNS で入力される主な情報としては以下のようなものがある。

- プロフィール：氏名，誕生日，出身地，経歴，メールアドレスなど
- 投稿：住所等の位置情報，家族構成，（写真，動画）顔認識データなど

また，公開を制限していても不正な操作によるアカウントの乗っ取りや情報漏洩が起こる事例も多く発生している。ジェムアルト社の調査[16]によると，2018 年上半期はケンブリッジ・アナリティカ社の facebook データの不正取得事件等により SNS 上の 25 億のレコードが侵害を受けている。攻撃者が悪用することで再特定化のリスク増加につながるため，IoT の進化を踏まえたリテラシーを臨床試験参加者自身も持ち，以下のような適切な対策が必要である。

- 不要なアカウントは削除する
- プライバシー設定（公開範囲など）を見直す
- 二要素認証など，セキュリティの高いログイン方法に変更する
- 不審な SNS サービス，アプリの利用を控える
- プライバシー情報の書き込み，写真の投稿に留意する
- 他者が投稿したリンクを安易に開かないよう注意する

4　受入可能なリスクの閾値

リスクベースのアプローチでは，実際に非特定化された臨床試験データから算出される再特定化の全体リスクが，閾値よりも低いことが求められる。設定する閾値は，それぞれの国や地域によって法律や省令，ガイドライン等に従うことが望ましい。例えば EMA のガイダンス[17]と Health Canada のガイダンス[18]では 0.09 という値を推奨している。ただし，この値は画一的なものではなく，症例数が少ない試験や希少疾患の試験等，状況により考慮する必要があることに留意する必要がある。

また，地域によって異なる閾値が推奨されている場合，①最も厳格な基準に合わせて対応する，②それぞれの地域に沿った基準で対応する，の 2 パターンが考えられるが，それぞれのデメリットについても理解しておく必要がある。①の場合は，基準が緩い地域では必要以上にデータの有用性が低くなってしまう点，②の場合は，非特定化されたデータが複数存在することで利用者の混乱を招いてしまう点がデメリットとして挙げられる。

一方，PhUSE のガイダンスでは，公共の場に共有する場合は 0.09（最小セルサイズが 11），非公共の場に共有する場合は 0.2（最小セルサイズが 5）という値が一つの目安として提示[19]されている。これは，公共の場で共有する場合において，前述した EMA のガイダンスと Health Canada のガイダンスで推奨されている値と同じである。

留意事項として，PhUSE のガイダンスでは CDISC 標準に対応した SDTM 形式のデータをソースとしているのに対し，EMA や Health Canada のガイダンスではそれらに限定しているものでない。また，PhUSE で示されている閾値は，データのリスク閾値に着目したものである。

このように，受入可能なリスクの閾値について様々提唱されているが，ルールベースの方法かリスクベースの方法かという点も含めて，最終的には企業側の判断に委ねられている。

5　適用結果及び考察

5.1 適用データ

過去に行われた臨床試験データを Project Data Sphere から入手した。Project Data Sphere とは，2001 年に創設された CEO Roundtable on Cancer[20] により実現した取り組みで，過去に実施された抗がん剤の臨床試験データを共有（share），統合（integrate），解析（analyze）するためのプラットフォームである。

本検証で使用するデータとして，以下の観点で 2 つの試験データを選択した。

- データ形式が SDTM であること
- 症例数が極端に少なくないこと
- DM ドメインに検証に必要な準識別子が十分に含まれていること

データの有用性の評価をするにあたり，すべてのドメインを検討・確認することは冗長なので，被験者背景にあたる DM ドメインおよび死亡情報を含む DS ドメインを評価対象とした。

以降，2 つの試験をそれぞれ試験 A，試験 B と記載する。データはすでに非特定化の処理が行われた後のものであること，意図せず症例を特定できてしまうことをリスクととらえる前提に立つ。その上で，データを提供する企業側がどういった考察に基づいてどういった処理を行うか／行ったか，どういったリスクが残っているか，それによって解析を行う上での有用性がどれだけ損なわれるか／維持されるかを具体的に考察する。

試験 A データへの考察

試験 A は大腸癌を対象とした国際共同試験であり，共有されたプラセボ群の例数は 1604 例（無治療の 21 例を含む）の大規模な臨床試験である。実薬群のデータは共有されていない。

DM ドメインからは暦日情報（xxxDTC），症例番号（SUBJID），フリーテキスト情報（人種でその他を選んだ場合の記載欄），医師名（INVNAM），民族（ETHINC），施設番号（SITEID）が列ごと削除されている。ただし，症例番号は，他のデータセットと結合する際の重要な要素であるため，別のランダムな番号が与えられていることから，番号を置換したとみなすことができる。

まず暦日を削除したのは，症例の探索範囲が極端に狭くなることを避けるためである。なお PhUSE のガイダンスでは FPI の初回投与日等にあわせて全症例の暦日をシフトさせるオフセット処理を推奨している。同様の理由で施設の特定につながる医師名や施設番号も削除されている（医師個人に配慮したという要素もあると推察される）。症例番号は，国・施設や症例の登録順を考慮して発番されることが多いため，やはり暦日や施設を推定しうるものとして削除したと考えられる。一般に解析を行う上では割付や初回投与日からの日数が用いられ，暦日が必要になることは稀であることからデータの有用性はあまり落ちないと考えることはできる。一方で，これらのデータ加工によって，探索的な検討として特定の施設や国が外れ値に該当することがないかの検討，月毎の解析やシーズン毎の解析など，特定の期間毎の解析ができないことになる。

国の情報は表 8 で示すカテゴリに基づいて地域に置き換えられて提供されている。表 8 の情報はデータと共に提供される補足文書（de-idenfiticaiton notes）から抜粋した。この試験の場合，特に被験者が

1 例しかいないエストニア，香港，台湾では，試験に参加したことが分かれば，それが個人の特定につながるリスクが非常に高い。一方で，国や地域は解析において有用な情報になりうるため，有用性とリスクのバランスを踏まえて地域という形で丸めた情報を提供していると推察できる。イスラエルや南アフリカ，オーストラリアは地域で丸めたとしても例数が少なく，再特定化リスクが高いと判断したのでその他（Rest of World）というカテゴリを設けたのであろう。

表 8：試験 A の参加地域，国及び例数の内訳

地域（Region）	国（Country）*	例数 **
西欧	オーストリア（19），ベルギー（43），スイス（8），ドイツ（227），デンマーク（15），スペイン（66），フィンランド（4），フランス（102），英国（72），ギリシャ（11），アイルランド（9），イタリア（184），オランダ（32），ノルウェー（3），ポルトガル（17），スウェーデン（35）	847
東欧	ブルガリア（32），ベラルーシ（65），チェコ（45），エストニア（1），クロアチア（52），ハンガリー（171），リトアニア（19），ラトビア（108），ルーマニア（141），ロシア（323），セルビア（25），スロバキア（5）．スロベニア（15），ポーランド（173），ウクライナ（125）	1300
北アメリカ	カナダ（53），米国（182）	235
南アメリカ	アルゼンチン（22），ブラジル（159），チリ（20），コロンビア（75），メキシコ（40），ペルー（112）	428
アジア	中国（98），香港（1），インドネシア（15），インド（333），韓国（94），マレーシア（17），台湾（1）	559
その他（Rest of World）	イスラエル（12），南アフリカ（21），オーストラリア（22）	55

** ()内は国別の参加数*
*** 実薬群を含めた例数*

PhUSE のガイダンスを踏まえると，非特定化された試験 A データの内，準識別子として含まれているのは以下の 4 つとなる。

- AGE （数値変数）
- RACE （カテゴリ変数）
- REGION （カテゴリ変数）
- SEX（カテゴリ変数）

年齢は特にカテゴライズされておらず 1 歳刻みで情報を残しているが，高齢者（85 歳以上）は 1 つのカテゴリに分類しており，再特定化リスクに配慮していると思われる。HIPAA では 89 歳超 は 1 つのカテゴリにまとめることを要求しているので，いずれの試験もその条件は満たしているが，85 歳以上と定義した理由までは推測の域を出ない。

人種は一部の症例でブランクに置換されている。特定の人種の症例が少なく，再特定化リスクが高かったことからブランクに置換されていると推測される（年齢や人種に対する非特定化方法は補足文書に記載されている）。

DS ドメインからはフリーテキスト情報が削除されている。また，暦日情報は初回投与日からの日数（Study Day）に変換されている。ただし，死亡日は初回投与日からの日数ではなくさらに週数（Study Week）に変換されている。Overall Survival を算出・再現することが研究計画の主たる目的であれば影響は少なからずある。特に Overall Survival の比較的短い膵臓癌の場合は，結果に与える影響は無視できないとも考えられるが，この処理をしていないと，当該試験のいずれかの暦日が試験外の情報から明らかになった場合には，死亡日の暦日が算出できてしまうというリスクがある。死亡は稀なイベント，もしくは世の中に公表されるデータに含まれる可能性が高いイベントと判断し，Study Week を用いることでリスクを下げたのであろうと推測できる。

試験 B データへの考察

試験 B は，膵臓癌を対象とした国際共同試験であり，共有されたプラセボ群には 273 例が含まれていた臨床試験である。実薬群のデータは共有されていない。

DM ドメインからは暦日情報（xxxDTC），症例番号（SUBJID），フリーテキスト情報（人種でその他を選んだ場合の記載欄），医師名（INVNAM），施設番号（SITEID）が列ごと削除されている。

削除の理由は A と同様であることが推察でき，試験 B では試験 A のように各国の例数の内訳までは開示されていない（表 9）。また，国の情報は地域に置き換えられているが，当該試験の公開情報（Clinicaltrial.gov）から国名を特定することは可能であるものの，どの国がどの地域に該当するのかは明示されておらず推測に留まる。例えば，メキシコを例にとると「北米」「中南米」というカテゴリを設定した場合には，「北米」に分類するのが一般的と思われるが，試験 A ではメキシコを南アメリカに含めていた。この試験では「北米」に含めたのか，それとも「その他」に含めたのかまではわからず，その他の例数が相対的に少ないことを考慮して，あえてメキシコを「北米」に含めていない可能性もある。「西欧」と「東欧」についても厳格な境界線は定義されていないことから，どちらの地域に分類することも可能な国は存在する。少数集団が出来てしまうことを避けるため，定義にはある程度の自由があると考えるべきであろう。ごく少数例しか参加していない国があった場合には，プラセボ群に 1 例も症例がいない（すなわちこのデータセットに当該国の患者が含まれていない）可能性もある。各国がどの地域カテゴリに含まれているか分からないことが再特定化のリスクを下げているともいえる。研究者が地域の情報を用いて解析を行う際には，補足文書を確認し，どの国がどの地域に含まれるか慎重に確認する必要があるであろう。

表 9：試験 B の参加地域，国及び例数の内訳

地域（Region）	国（Country）	例数
西欧	オーストリア，ベルギー．フランス，ドイツ，ギリシャ，イタリア，スペイン，スイス	85
東欧	ブルガリア，キプロス，チェコ，ハンガリー，ポーランド，スロバキア，ルーマニア，ロシア	87
北米	カナダ，米国	86
	メキシコ	
その他	アルゼンチン，コロンビア，インド，プエルトリコ	15

PhUSE のガイダンスを踏まえると，非特定化された試験 B データの内，準識別子として含まれているのは以下の 4 つとなる。

- AGE （数値変数）
- RACE （カテゴリ変数）
- REGION （カテゴリ変数）
- SEX（カテゴリ変数）

試験 A と同様に年齢は特にカテゴライズされておらず 1 歳刻みで情報を残しているが，再特定化リスクに配慮して高齢者（84 歳以上）は 1 つのカテゴリに分類している。

人種は一見処理がされていないように見えるが，BLACK OR AFRICAN AMERICAN と ASIAN を OTHER に統合し，小集団が残ることを回避している（年齢や人種に対する非特定化方法は補足文書に記載されている）。

DS ドメインは試験 A と同様の処理がされている。

5.2 k-匿名化での評価結果

試験 A，試験 B データの DM ドメインに対して k-匿名化手法を適用し，平均リスク，最大リスクを算出する。準識別子は AGE，RACE，REGION，SEX の 4 つとした。

平均リスク，最大リスク算出の為には，まず各レコードに対し，同じ準識別子を持つレコード数を求め，逆数を取ることで各レコードの再特定化リスクを算出する。次にレコード全体での再特定化リスクの平均と最大を計算する．この値がそれぞれ，データの平均リスク，最大リスクとなる。例えば下表の場合，準識別子を SEX, AGE とすると，ID:1 は同じ準識別子を持つレコードが 3 レコード存在するため，再特定化リスクは 1/3 となる．各レコードの再特定化リスクを計算し，レコード全体の再特定

化リスクの平均を計算すると，(1/3+1/3+1/3+1/2+1/2)/5=2/5 となる．この値が下表のデータの平均リスクである．また，最大リスクは 1/2 である．

ID	SEX	AGE	再特定化リスク
1	**Male**	**45**	**1/3**
2	**Male**	**45**	**1/3**
3	**Male**	**45**	**1/3**
4	**Female**	**32**	**1/2**
5	**Female**	**32**	**1/2**

試験 A，試験 B で同様の方法を用いて算出した平均リスク，最大リスクは以下の通りであった。

Study	Number of patients	Mean risk	Max risk
試験A	1604	0.671	1
試験B	273	0.608	1

また，これら 2 試験に対して，比較的詳細な情報が残っている年齢に対してカテゴリ化を適用し，平均リスク，最大リスクを再度計算した結果は以下の通りであった。

Study	Method	Number of patients	Mean risk	Max risk
試験A	年齢5歳毎にカテゴリ化	1604	0.340	1
	年齢10歳毎にカテゴリ化		0.233	1
	年齢20歳毎にカテゴリ化		0.156	1
試験B	年齢5歳毎にカテゴリ化	273	0.253	1
	年齢10歳毎にカテゴリ化		0.165	1
	年齢20歳毎にカテゴリ化		0.114	1

いずれの適用でも最大リスクは 1 となり，最大リスクが 1 となる要因としては，比較的年齢の低い症例，年齢の高い症例が同値類数 1 となることであった。最大リスクを 1 より小さくするには強い非特定化処理が必要で，臨床試験の例数規模でそれを実現するのは困難である。また，データが欠測の場合，”Other”等に置き換えることで既存のカテゴリにまとめることが必要だが，非特定化により置き換えられたのか，元々カテゴリに属していたのか，データの受領者は見分けることが出来ないため解析時には注意が必要である。3 章で記載した攻撃が試みられる確率（Pr(attempt)，Pr(acquaintance)，Pr(breach)）との積を取る場合でも，EMA や Health Canada が推奨している再特定化リスク閾値 0.09 を下回るか否かの判断には，おのずと専門家の見解が必要となる。研究目的によっては，共有するデータを研究者が必要とする変数のみに絞る等の対策も有効である。

なお，k-匿名化を適用してリスクを評価する場合，一般に以下の仮定が置かれている[21]。

1. 加工・開示される非特定化データは 1 つの元データに対し 1 種類のみである
   - 非特定化データは共有されるがその加工アルゴリズムは共有されない
2. 非撹乱的マスキング（Non-perturbative masking）のみ適用されている
   - 非撹乱的マスキングは，データの一般化，消去，ある値以上（以下）を 1 つのカテゴリにまとめるトップ（ボトム）コーディングなどの方法がある

3. 攻撃者が元データにおける準識別子の値をすべて知っている
4. 攻撃者は攻撃対象が当該データに含まれていることを把握している

これを臨床試験データに適用する場合，3 番目と 4 番目の仮定は強すぎるという点に留意すべきである。3 番目の仮定関しては，攻撃者は準識別子の値のすべてを把握しているケースは稀であろう。また，4 番目の仮定に関しても，攻撃者の攻撃対象が当該データに存在するか把握しているケースは稀である。

5.3 l-多様性での評価結果

DS ドメインから死亡の有無を取得し，2 値変数（Y or N）である DTHFL を作成し，DM ドメインの変数に加えた。DTHFL を仮にセンシティブ・データとする。準識別子は AGE，RACE，REGION，SEX の 4 つとした。

ID	AGE	RACE	REGION	SEX	DTHFL
1	**45**	**45**	**Asia**	**ASIAN**	**Y**
2	**45**	**45**	**Asia**	**ASIAN**	**Y**
3	**32**	**32**	**Asia**	**ASIAN**	**N**
4	**32**	**32**	**Asia**	**ASIAN**	**N**
5	**32**	**32**	**Asia**	**ASIAN**	**Y**

準識別子が同じレコード（準識別子グループ）毎にセンシティブ・データの種類を算出する。例えば，ある準識別子グループの全てのレコードの DTHFL の値が Y であれば，この準識別子グループのセンシティブ・データの種類は 1 として種類=1 のグループ数に含める。DTHFL の値が Y のレコードも N のレコードも存在すれば種類=2 のグループ数に含める。

準識別子グループ	AGE	RACE	REGION	SEX	種類
1	**45**	**45**	**Asia**	**ASIAN**	**1**
2	**32**	**32**	**Asia**	**ASIAN**	**2**

まとめた結果は以下の通りであった。

Study	Method	準識別子グループ数	種類=1のグループ数	種類=2のグループ数
試験A	年齢5歳毎にカテゴリ化	184	103	81
	年齢10歳毎にカテゴリ化	117	59	58
	年齢20歳毎にカテゴリ化	72	29	43
試験B	年齢5歳毎にカテゴリ化	70	30	40
	年齢10歳毎にカテゴリ化	46	19	27
	年齢20歳毎にカテゴリ化	32	14	18

いずれの適用でも，センシティブ・データの種類が 1 種類である準識別子グループが存在していた。準識別子グループにおけるセンシティブ・データの多様性を担保する，つまり，2-多様性を満たすに

は，準識別子のカテゴリ化を見直し，全ての準識別子グループが 2 種類のセンシティブ・データを有する状態にする，またはセンシティブ・データが 1 種類となっている準識別子グループの DTHFL の値をブランクに置換し，センシティブ・データを分からなくするなどの処理が必要となる。 準識別子のカテゴリ化を見直した場合，データの有用性を低くする可能性があるため，研究に不要であるならばデータからセンシティブ・データを除くことも検討する必要があるであろう。

6　おわりに

実際の臨床試験データを用いて，非特定化の適用・再特定化リスクの算出を行った。本報告書執筆時点で，推奨される非特定化の方法や許容出来る再特定化リスクの閾値は明確に定められておらず，ルールベース・リスクベースのいずれの方法を用いるか，リスクベースの方法を用いるとした場合の受入可能なリスクの閾値の設定は個々の企業に委ねられている。そのような状況下においてなお，CTDS の取り組みは求められているのである。

推奨される方法や許容できる再特定化リスクの閾値がない状況において，非特定化を実施する者にとって重要なことは，再特定化が行われるリスクと研究者にとっての有用性のバランスを取り，適用した非特定化手法，再特定化リスクの設定根拠等（根拠に乏しいながら仮定した値があるのなら，そのことも明確にして），許容可能なリスクとして判断した閾値をきちんと作業記録として残しておくことである。非特定化データが再特定されるリスクは決して 0 にはならない。少なからずリスクが存在する中で非特定化データを提供するとした理由を，企業が説明出来ることが最も重要であろう。

資料作成者

日本製薬工業協会　医薬品評価委員会 データサイエンス部会　2020 年度 継続タスクフォース 3

大塚 渉　　中外製薬株式会社（～2020/03）

成宮 大貴　　あすか製薬株式会社

持永 浩二　　株式会社大塚製薬工場

宮澤 昇吾　　塩野義製薬株式会社

佐土原 和宏　　鳥居薬品株式会社

澤田 克彦　　大鵬薬品工業株式会社（～2020/03）

タスクフォースリーダー兼推進委員

青木 真　　アステラス製薬株式会社

加藤 智子　　サノフィ株式会社（～2020/03）

大塚 渉　　中外製薬株式会社（2020/04～）

担当副部会長

酒井 弘憲　　エーザイ株式会社（～2020/03）

加藤 智子　　サノフィ株式会社（2020/04～）

7　参考文献

[1] 臨床試験の個別被験者データの共有 CTDS（Clinical Trial Data Sharing）（2017 年 6 月）: available at http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/ctds.html

[2] 臨床試験の個別被験者データの共有にあたって：最近の動向も交えて（2019 年 7 月号 No.192）: http://www.jpma.or.jp/about/issue/gratis/newsletter/html/2019/92/92cm-01.html

[3] CTDS（Clinical Trial Data Sharing）に関連する規制と対応の留意点（2020 年 3 月）http://www.jpma.or.jp/medicine/shinyaku/tiken/allotment/pdf/ctds_points_to_remember.pdf

[4] De-identification and Anonymization of Individual Patient Data in Clinical Studies – A Model Approach: available at http://www.transceleratebiopharmainc.com/wp-content/uploads/2015/04/TransCelerate-De-identification-and-Anonymization-of-Individual-Patient-Data-in-Clinical-Studies-V2.0.pdf

[5] PHUSE De-Identification Working Group paper “Providing De-Identification Standards to CDISC Data Models”: available at https://www.pharmasug.org/proceedings/2015/DS/PharmaSUG-2015-DS10.pdf

[6] Institute of Medicine (IOM), “Sharing Clinical Trial Data; MAXIMIZING BENEFITS, MINIMIZING RISK”: available at http://www.nap.edu/catalog/18998/sharing-clinical-trial-data-maximizing-benefits-minimizing-risk

[7] De-identification Guidelines for Structured Data: available at https://www.ipc.on.ca/wp-content/uploads/2016/08/Deidentification-Guidelines-for-Structured-Data.pdf

[8] https://www.clinicalstudydatarequest.com/

[9] https://vivli.org/

[10] https://www.projectdatasphere.org/

[11] https://www.phrn.org.au/

[12] Pharmaceutical Users Software Exchange (PhUSE), De-Identification Standard for CDISC SDTM 3.2: available at https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/De-identification+Standard+for+SDTM+3.2+Version+1.0.xls

[13] 南 和宏. データの非特定化方法論 (2017) ：available at http://www.jpma.or.jp/medicine/shinyaku/tiken/symposium/pdf/20171219/20171219_05.pdf

[14] Li, N., Li, T. and Venkatasubramanian, S.: t-Closeness: Privacy Beyond k-Aonymity and l-Diversity. Proc. ICDE 2007, pp. 106-115, (2007).

[15] haled El Eman, Luk Arbuckle 著, 笹井崇司 訳, データ匿名化手法 ― ヘルスデータ事例に学ぶ個人情報保護, オライリー・ジャパン, 2015.

[16] https://www.gemalto.com/press/Pages/Data-Breaches-Compromised-3-3-Billion-Records-in-First-Half-of-2018.aspx

[17] https://www.ema.europa.eu/en/human-regulatory/marketing-authorisation/clinical-data-publication/support-industry/external-guidance-implementation-european-medicines-agency-policy-publication-clinical-data

[18] https://www.canada.ca/en/health-canada/services/drug-health-product-review-approval/profile-public-release-clinical-information-guidance.html

[19] Pharmaceutical Users Software Exchange (PhUSE), De-identification standards for CDISC SDTM 3.2 Appendix2: available at https://phuse.s3.eu-central-1.amazonaws.com/Deliverables/Data+Transparency/Data+De-Identification+Standard+for+SDTM+3.2+%E2%80%93+Appendix+2%3A+Low+Frequencies+Version+1.0.pdf

[20] CEO Roundtable on Cancer: available at http://www.ceoroundtableoncancer.org/

[21] 南 和宏. プライバシー保護とデータ有用性確保の両立を目指したデータの非特定化技術, 第 16 回 DIA 日本年会 (2019)