医療情報利活用における匿名化技術ガイド

2015/05/30

Ver1.0

JIRA　医用画像システム部会　新 IMIT-WG　 SWG3

１．はじめに

医療の進歩のためには医療情報の利活用は不可欠である。現在の医療水準が達成されているのは、過去の医療情報の蓄積と分析、すなわち利活用が行われてきた成果である。もし医療情報の利活用が制限されるようなことがあれば、それは今後の医療の進歩を妨げるものであり、社会的利益の喪失を意味する。

他方、患者のプライバシーを守ることは、医師に課せられた守秘義務に代表されるように、医療情報管理の必須要件である。誰しも自身の不利益になる情報の公開は望まないし、場合によってはその患者の家族にまで影響が及ぶ可能性がある。したがって、医療情報管理には高度な安全性が要求される。

これら利活用と安全管理を両立させる対策として情報の「匿名化」がある。医療情報のみならずパーソナルデータを含むビッグデータの利活用の推進の上で、「匿名化」の重要性が高まっている。

本書では、医療情報の「匿名化」に関しての社会的背景や技術的内容について解説する。

２．目的と適用範囲

本書は、JIRA 会員企業に対し、医療情報の利活用における匿名化についての社会的な要求と技術的な内容について解説するものである。JIRA 会員企業向けということで、医療情報の内、DICOM 画像データ、読影レポートを対象としている。JIRA 会員企業が製造・販売している機器から出力される画像データやレポートデータを、医療機関が第三者へ提供を行う際に、医療機関から製造企業に対して情報の匿名化の技術的対応要求があるという前提で、それに対応するための情報提供を目的としている。

JIRA 会員企業になじみの深い利活用ユースケース例には、

- 臨床画像・レポートの症例データベースへの提供
- 臨床画像・レポートの学会での発表
- 臨床画像・レポートの教育への利用

が考えられる。

利活用にあたって、匿名化したデータを提供する時のリスク分析は、情報提供者（医療機関）が用途・提供先組織・利用形態で判断することであるが、医療情報システム事業者として関与する場合には、本書4章･5章の資料等により理解を進めておくことを推奨する。

さらに、JIRA 会員企業が「医療機器・システム機能のテストデータとして診療情報の提供を受ける」場合は、JIRA 会員企業側が当事者になることから、手法や制度的体制につ

いての理解が必須である。すなわち、内閣府高度情報通信ネットワーク社会推進戦略本部（IT 総合戦略本部）で平成 25 年 12 月 20 日に決定された「パーソナルデータの利活用に関する制度見直し方針」の内容に従うこととなる。このケースは民間事業者による事業用途であるが、市場への機能提供によって最終的には社会的利益＝公益が発生することを提供側に理解してもらうことが望まれる。

３．医療情報の匿名化について

３．１　個人情報の匿名化とは

医療情報の匿名化に関して、現行の主なガイドラインは以下の 2 点が存在する。

・「医療・介護関係事業者における個人情報の適切な取扱いのためのガイドライン」[1]
・「医療情報システムの安全管理に関するガイドライン」[2]

「医療・介護関係事業者における個人情報の適切な取扱いのためのガイドライン」にて、下記定義が記載されている。

・個人情報：「個人情報」とは、生存する個人に関する情報であって、当該情報に含まれる氏名、生年月日、その他の記述等により特定の個人を識別することができるもの（他の情報と容易に照合することができ、それにより特定の個人を識別することができることとなるものを含む。）をいう。
・個人情報の匿名化：当該個人情報から、当該情報に含まれる氏名、生年月日、住所等、個人を識別する情報を取り除くことで、特定の個人を識別できないようにすることをいう。

「医療情報システムの安全管理に関するガイドライン」にて、匿名情報の取り扱いに関して規定が記載されている。

３．２　国内の動向

現在、多分野でのパーソナルデータ利活用にあたってのルール明確化・環境整備の検討が、内閣官房高度情報通信ネットワーク社会推進戦略本部（IT 総合戦略本部）の「パーソナルデータに関する検討会」[3]を中心に進められている。

医療情報（診療記録、検査記録（画像、画像診断レポート）、など）の臨床研究、疫学的利用、医学教育、医療行政用途等での利活用の有用性については、異論の無いところであるが、有用であることを重視するあまり、不十分な匿名化処理のままで開示さらには公開することは個人のプライバシー侵害にもなり得る。一方、現行での匿名化指針についても、曖昧で不十分との指摘があり、上記検討会での対象になっている。

「パーソナルデータに関する検討会」における検討内容を紹介する。

・健康情報パーソナルデータ利活用の問題点（山本委員）：第 2 回　資料 1-3

・医療は過去の診療情報の蓄積の上に成り立っているため利活用は必須である
・完全な匿名化（例、HIPAA のプライバシールール）では疫学的な利活用に支障がある
・第三者提供時には匿名化の程度を判断する必要がある
・医療機関の設立者によって適用法令が異なり情報連携に支障がある
・利活用できないことの対策がなされていない
・個人情報の定義があいまいで匿名化が定義できない

などの現状での問題点が挙げられている。

・医療等分野におけるパーソナルデータの利活用の類型及び考察（松本構成員）：第 2 回技術検討ワーキンググループ　資料 4

現状の利用例と利活用にあたっての医療情報の特性が述べられている。

上記の認識を踏まえて、平成 25 年 12 月 20 日に高度情報通信ネットワーク社会推進戦略本部（IT 総合戦略本部）において「パーソナルデータ利活用の制度見直し方針」が出されている。

「パーソナルデータ利活用の制度見直し方針」において、匿名化された情報を本人同意なしに第三者提供するための法的整備（個人情報保護法の改訂）に向けた提言が示されている。

すなわち、第三者提供における本人同意原則の例外条件の追加を提示している。例えば、個人が特定される可能性を低減した個人データの扱いが、個人再特定される危険に対して、提供先組織が“他のデータと突合して再特定しないこと、再提供先にも求めることを公表している場合（いわゆる FTC 三原則）“等である。

３．３　匿名化技術

匿名化技術について、「パーソナルデータに関する検討会」技術検討ワーキンググループ資料の記載を紹介する。

・匿名化技術の現状について（高橋構成員）：第 1 回技術検討ワーキンググループ　資料 2-3
・個人識別できない匿名データは作成できるか（高橋構成員）：第 2 回技術検討ワーキンググループ　資料 1
・技術検討ワーキンググループ報告書：第 5 回資料 2-1

この中では、

・用語「匿名化」を、個人「識別」と「特定化」に区分している。
・どのような技術的手法をとっても一般的意味での「完全な匿名化」はあり得ないこと
・匿名化の程度に関して、匿名化には幾つかのレベルがあること

が記載されている。

匿名化レベル内容を、下記「1」「2」「3」と、図 4.1 で紹介する。
個人情報とされる範囲の境界は、「2」のレベル中にあるとされるが、明確な規定は難しいとされている。

「1. 連結可能匿名データ」
氏名等を削除するが、元の情報と照合することで個人と連結可能なもの（仮名化と呼ばれる）

「2. いわゆる匿名データ」
氏名等を削除し、元の情報と対応できないようにしたもの（無名化と呼ばれる）

「3. 高度な匿名データ」
高度な匿名処理により、特定の個人の識別が困難になるようにしたもの

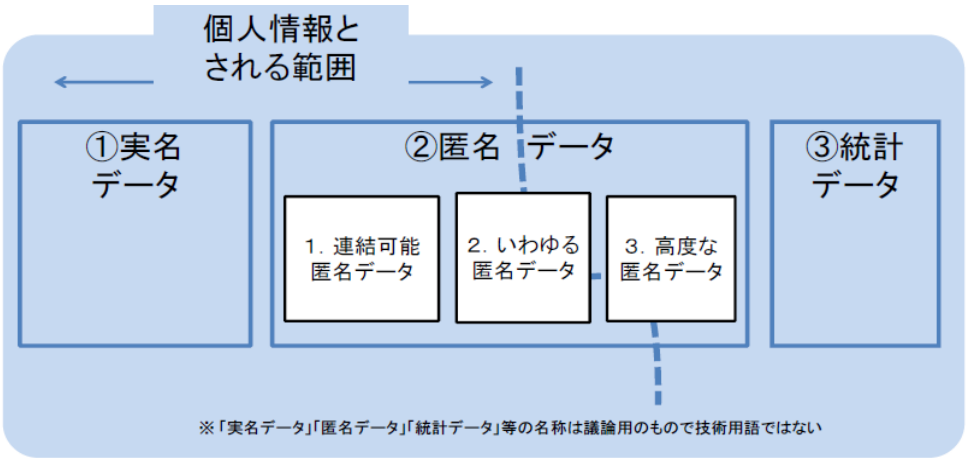

図 3.1　匿名化のレベル
（「匿名化技術の現状について」第 1 回技術検討ワーキンググループ　資料 2-3　p6 より引用）

「3」としての高度な匿名化処理の代表例として“k-匿名化”がある。ここでの k とは、特定個人のデータを k 個未満に絞り込めないかどうかを示す匿名性の指標であり、k を大きくすると個人特定リスクは減るが、大きくしすぎると情報量が落ち、実用上使えないデータになってしまうとの問題が存在する。k-匿名化に関しては、4.2 で説明する。

３．４　海外の事例

３．４．１　米国

HIPAA が規定した匿名化ルール（二つの方法が選択可：１）規定された 18 属性を削除、２）統計分析の専門家により個人が特定されるリスクを評価し、十分低いことを判断した分析の経過および結果を文書化）に従い、利活用が数多く実施されている。[4]

３．４．２　英国

EHR を利活用するための仕組みとして SUS（Secondary Uses Service）が構築されている。また、医療情報の利活用に向けたガイドライン整備も進んでおり、医療分野における情報保護と公開の両立を図るため、NHS が 2013 年 2 月に“Anonymization Standard for Publishing Health and Social Care Data”というガイドラインを策定し、匿名化のルールが明確化されている。[4]

【3 章の参考資料】

[1]厚生労働省「医療・介護関係事業者における個人情報の適切な取扱いのためのガイドライン」 http://www.mhlw.go.jp/topics/bukyoku/seisaku/kojin/

[2]厚生労働省「医療情報システムの安全管理に関するガイドライン」4.2 版
http://www.mhlw.go.jp/stf/shingi/0000026088.html

[3]内閣官房高度情報通信ネットワーク社会推進戦略本部「パーソナルデータに関する検討会」 http://www.kantei.go.jp/jp/singi/it2/pd/index.html

[4]NTT データ経営研究所「医療情報に関する海外調査報告書」
http://www.keieiken.co.jp/medit/pdf/240423/0-report_2.pdf

４　JIRA に関係する医療情報の匿名化について

４．１　DICOM における匿名化

DICOM 規格[1][2]における匿名化については、DICOM 規格書 15 巻の「セキュリティとシステム管理のプロファイル（Security and System Management Profiles）」の、付属書 E「属性の秘匿プロファイル（Attribute Confidentiality Profiles）」に規定されている。この付属書では、暗号化による秘匿や、匿名化された情報を元に戻すことも規定されているが、ここでは主に匿名化に関して説明する。

DICOM データは、図 4.1 のように、患者や検査に関する属性情報の値（Value）が、タグ（Tag）、値の形式（Value Representation、以下 VR）とデータの長さ（Value Length）が付帯したデータ要素(Data Element）として記録されている。タグの値（グループ値、エレメント値）は情報の項目ごとに決められており、例えば、患者の ID は(0010,0020)、検査日付は(0008,0020)である。項目ごとのタグ値は、DICOM 規格書 6 巻の「データ辞書（Data Dictionary）」に規定されている。

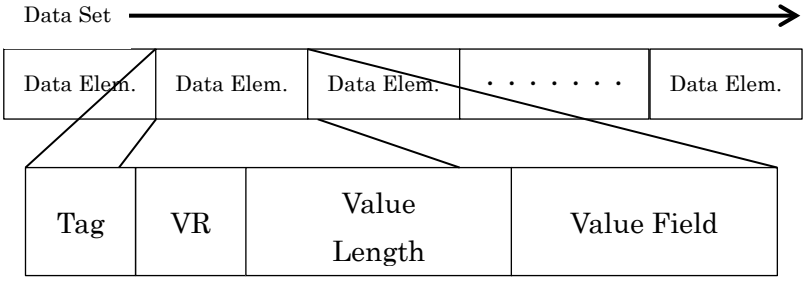

図 4.1 DICOM データ構造

DICOM データ内の患者に関する識別情報を匿名化するということは、単純にその識別情報のタグのデータ要素を取り除けば良いわけではない。DICOM 規格は、タグごとにタイプや値の形式を決めており、単に削除すると DICOM 規格違反のデータになってしまい、その後の処理に使えないデータになってしまうので注意が必要である。

タイプ（Type）に関しては、タイプ 1（Type 1）は長さがゼロでなく何らかの値が入っていること、タイプ 2（Type 2）は値が不明ならば長さゼロで値なしにすること、等と決められている。したがって、匿名化する際は、Type1 のデータ要素には長さがゼロでないダミー値をセットする必要がある。

また、値の形式（VR）に関しては、例えば検査日付(0008,0020)は、VR が DA（Date）と定義されており、8 文字固定の文字列で、使える文字は数字のみである。したがって、検査日を匿名化したいということで、"20140101"と入っている値を、"XXXXXXXX"といった数字以外の値に置き換えると DICOM 規格違反となってしまう。

その他に DICOM データの匿名化には、以下の点に注意が必要である。

1）匿名化の際、どのデータ要素を処理するかは、匿名化したデータの利用目的に関連するため、ケースごとに匿名化処理者の責任で判断しなければならない。DICOM 規格書 15 巻付属書 E では基本アプリケーションレベル秘匿プロファイルとして、教育用ファイル作成等の目的のために処理するデータ要素のリスト例が提示されている。

2）処理対象のタグ値のデータ要素は、シーケンスにアイテムとして含まれている場合がある。そのデータ要素に対しても同じ処理が必要である。

3）処理時にセットされるダミー値の内容は規定しないが、それが患者を識別不能な値であることは、ダミー値を生成する匿名化ソフトウェアか、それを入力する操作者の責任である。

4）画像データに識別情報が埋め込まれている場合があるため、処理者は埋め込まれていないことを確認すること。なお、埋め込まれたものを削除することは規格の適用外である。

5）収集コメント(0018,4000)のように、現在はリタイアされているデータ要素にも対応が必要である。

6）メディアに保存されている場合に付加されるメタ情報についても同様な匿名化の処置が必要である。

7）装置ベンダが独自に定義したプライベート属性については、その中に識別情報が含まれるかどうかは不明のため、原則削除すべきである。

8）オーバレイデータ(60xx,3000)に識別情報がビットマップデータとして含まれている

可能性があるため、原則削除すべきである。

DICOM 規格書 15 巻付属書 E では、匿名化を行う際にどのように処理するかを、以下の 6 つのアクションコードで示し、匿名化対象のデータ要素ごとに表に示している。

D - ゼロでない長さのダミーの値に置換する。値の形式は VR と一致させる。

Z - 長さをゼロにして値をセットしない、あるいは、ゼロでない長さのダミーの値と置換する。値の形式は VR と一致させる。

X - データ要素を削除する。

K - 保持する。（シーケンスでない要素は変更なし。シーケンス要素は消去する）

C - 消去する。識別情報を含まない値に置き換える。値は VR と一致させる。

U - インスタンスとして一貫性のある UID に置き換える。ここでの一貫性とは、例えば同じ検査のデータの場合、Study Instance UID が同じに保つことである。利活用目的によっては一貫性が必須のケースがあるため、注意が必要である。

DICOM 規格書 15 巻付属書 E の表 E.1-1 から、主な属性データ要素の処理方法を以下に抜粋した。他のデータ要素については DICOM 規格書を参照のこと。

属性	タグ値	リタイア	処理方法
患者の ID	(0010,0020)		Z
患者の名前	(0010,0010)		Z
患者の生年月日	(0010,0030)		Z
患者の年齢	(0010,1010)		X
患者の性別	(0010,0040)		Z
検査日付	(0008,0020)		Z
受付番号	(0008,0050)		Z
検査 ID	(0020,0010)		Z
検査内容	(0008,1030)		X
患者コメント	(0010,4000)		X
検査コメント	(0032,4000)		X
収集コメント	(0018,4000)	Y	X
オーバーレイデータ	(60xx,3000)		X
施設名	(0008,0080)		X/Z/D(Type3/Type2/Type1)
装置名前	(0008,1010)		X/Z/D(Type3/Type2/Type1)
操作者名	(0008,1070)		X/Z/D(Type3/Type2/Type1)
検査インスタンス UID	(0020,000D)		U
SOP インスタンス UID	(0008,0018)		U

プライベート属性	グループ番号が奇数		X

表 4.1 主な属性データ要素の処理例

なお、匿名化を実施した場合、患者識別削除（0012,0062）を値"YES"で追加し、匿名化方法コードシーケンス（0012,0064）に、DICOM 規格書 16 巻の CID7050 で定義されている匿名化方法のコードをセットするか、匿名化方法（0012,0063）に匿名化の手法について記述すること。ただし、必須ではない。DICOM 規格は頻繁に改定が行われており、最新の規格書を参照すること。

【4.1 の参考資料】

[1]NEMA　The DICOM Standard
http://medical.nema.org/standard.html
[2] JIRA「DICOM の世界」
http://www.jira-net.or.jp/dicom/index.html

４．２　k-匿名化

米国　HIPAA では個人の同定が不可能なデータとは「個人を同定せず、かつ個人を同定するのに用いることができる情報であると思うような合理的根拠が全くない」データとして定義されるが[1]、この要件を満たすためには公開されたデータが単独あるいは他の公開情報との組み合わせに対しても、個人の同定のリスクが小さいことが要求される。

例えば、次のような例が考えられる[3]。下の図では医療データが氏名と住所を削除された状態で公開されている。しかし、氏名および住所が公開されている選挙人リストを利用することで誕生日、性別、ZIP コード（郵便番号）から個人を特定することができ、したがって個人の医療データを特定することが可能となる。

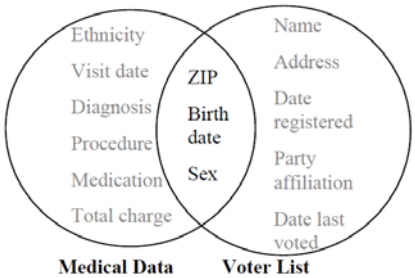

Figure 1 Linking to re-identify data

図 4.2-1

L. Sweeney, “Guaranteeing anonymity when sharing medical data, the Datafly system”

こういった組み合わせによる同定可能を防ぐための技術が k-匿名化(*k*-ANONYMITY)であり、公開情報から非公開情報の推論を制御するためのプライバシー保護手法の 1 つである。

k-匿名化は上述のようなデータテーブルから特定のレコードが同定されることを防ぐために、同じ保護属性（上記の例では削除した氏名および住所にあたる情報が相当）の組み合わせをもつレコードが、少なくとも k 個存在すること（k-匿名性）を保証するために、レコードの削除・修正方法を提供する技術である。例えば、図 4.2-2 は k-匿名性を保持したテーブルの例である[2]。図 4.2-2 ではそれぞれ t3,t5,t7 の生年, t4,t6 の性別データの一部あるいは全体を削除し、（一般化された）テーブル RT を作成している。このテーブル RT はいずれのデータもすべての項目が一致するレコードを自分も含めて最低 2 個以上はもっており、したがってこのテーブルは 2-匿名性を満たしている。

k-匿名化のアルゴリズムはすでに既存手法の改良版を含めいくつか存在しているが[4]、k-匿名化アルゴリズムは処理時間や過度のデータの変更の回避・低減などの観点から改善が継続されている状況である。

	Race	BirthDate	Gender	ZIP
t1	Black	1964	f	02138
t2	Black	1964	f	02138
t3	Black	1967	m	02141
t4	White	1971	f	02139
t5	White	1967	m	02141
t6	White	1971	m	02139
t7	White	1965	m	02141

(a) 初期テーブル PT

	Race	BirthDate	Gender	ZIP
t1′	Black	1964	f	02138
t2′	Black	1964	f	02138
t3′	Person	196*	m	02141
t4′	White	1971	*	02139
t5′	Person	196*	m	02141
t6′	White	1971	*	02139
t7′	Person	196*	m	02141

(b) 一般化テーブル RT

図 4.2-2

村本俊祐，上土井陽子，若林真一，k 匿名性を利用したデータ一般化によるプライバシー保護

【4.2 の参考資料】

[1] J.ヴィイダヤ, C.W.クリフトン, Y.M.ズー著, 嶋田茂, 清水將吾訳, プライバシー保護データマイニング, 丸善出版株式会社, 2011

[2]村本俊祐，上土井陽子，若林真一，k 匿名性を利用したデータ一般化によるプライバシー保護，DEWS2007, A7−10,2007

[3] L. Sweeney, “Guaranteeing anonymity when sharing medical data, the Datafly system”, Journal of the. American Medical Informatics Association, pp.1-5,. 1997

[4] k-匿名化手法の効率向上に関する一提案. 渡邉奈津美 ,土井洋 ,趙晋輝; 全国大会講演論文集 2013(1), 519-521, 2013.

５．まとめ

JIRA 会員が関係する医療情報の利活用に関しての現状と技術的対応について解説を行った。本書の内容については、今後とも充実させていく方針である。

匿名化についての社会的状況は、まだまだ制度的に過渡的である。だからこそ匿名化についての正しい知識が重要である。

なお、利活用、匿名化に関する内容については、最新のものを参照願いたい。

６．参照法規、規格、ガイドライン

1) 日本経済再生本部(2013.6.14)：2016 以降に「医療情報の番号制度の導入」

2) 厚生労働省(2012 年度)：「社会保障分野サブワーキンググループ及び医療機関等における個人情報保護のあり方に関する検討会」報告書

3) 英国(NHS)でのリスク分析ルール“Anonymization Standard for Publishing Health and Social Care Data”

4) FTC 三要件

http://www.ftc.gov/reports/protecting-consumer-privacy-era-rapid-change-recommendations-businesses-policymakers

・略語一覧（記載順）

JIRA	Japan Medical Imaging and Radiological Systems Industries Association （一般社団法人　日本画像医療システム工業会）
DICOM	Digital Imaging and Communication in Medicine
HIPAA	Health Insurance Portability and Accountability Act
FTC	Federal Trade Commission
EHR	Electronic Health Record
NHS	National Health Service
NEMA	National Electrical Manufacturers Association