匿名化基準に関する欧米公的文書7選の考察

千田 浩司[1]　吉浦 裕[2]　島岡 政基[3]

概要：匿名化基準に関して先進的な欧米公的文書 7 選について，各文書が定める再識別に対するリスク評価方法および匿名化技術の調査を行った．本稿では，代表的な再識別リスク指標である k-匿名性を中心に各文書が定めるリスク評価方法や匿名化技術を要約し，比較や特徴の考察を行う．またリスク評価方法と匿名化技術の対応関係が必ずしも明らかでないことを課題として挙げ，その対策方針を提案する．

キーワード：匿名化基準，再識別，k-匿名性

A Survey on Selected Western's 7 Official Documents for De-identification Standard

Koji Chida[1]　Hiroshi Yoshiura[2]　Masaki Shimaoka[3]

Keywords: De-identification standard, Re-identification, k-anonymity

1. はじめに

ビッグデータの活用が進む中，パーソナルデータを適切に保護するための匿名化 (De-identification)[*1]に関する基準が国内外で整備されつつある．特に匿名化したパーソナルデータから特定の個人のデータを識別する**再識別** (Re-identification) の問題が深刻化しており [3], [4], [5]，再識別に対するリスク評価や，当該リスクを軽減する匿名化技術の確立が求められている．我が国においても，昨年 9 月に成立した個人情報保護法改正に伴い，本人の同意無くパーソナルデータの流通を認める「匿名加工情報」の基準が検討されているところである [6]．そこで我々は，匿名化基準に関する現状や課題を明らかにするため，匿名化基準に関して先進的な欧米公的文書を対象に，各文書が定める再識別に対するリスク評価方法および匿名化技術の調査を行った．

調査対象の文書は，再識別に対するリスク評価方法や匿名化技術について明記されている以下 7 選である．

(1) 'Best Practice' Guidelines for Managing the Disclosure of De-Identified Health Information, Canadian Institute for Health Information (2010, カナダ)[7].

(2) Anonymisation: managing data protection risk code of practice, Information Commissioner's Office (2012, 英国)[8].

(3) Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Office for Civil Rights, U.S. Department of Health and Human Services (2012, 米国)[9].

(4) Anonymisation Standard for Publishing Health and Social Care Data Specification, National Health Service (2013, 英国)[10].

(5) Opinion 05/2014 on Anonymisation Techniques,

[1] NTT セキュアプラットフォーム研究所
NTT Secure Platform Laboratories
[2] 電気通信大学
The university of Electro-Communications
[3] セコム株式会社 IS 研究所
Intelligent Systems Laboratory, SECOM CO., LTD.

[*1] ISO/TS 25237(Health informatics — Pseudonymization)[1] では，匿名化 (Anonymization) を「身元が分かるデータの集合 (Identifying Data Set) とデータ主体 (Data Subject) の間の関連を取り除くプロセス」と定義しており，匿名化手法には Masking と De-identification の 2 種類がある [2].

ARTICLE 29 DATA PROTECTION WORKING PARTY (2014, EU)[11]

(6) Privacy and Data Protection by Design - from policy to engineering, ENISA (EU，2014)[12]

(7) De-Identification of Personal Information, NIST (2015, 米国)[13].

上記 7 文書は何れも，匿名化したパーソナルデータ (以降，単に匿名化データと呼ぶ) に対する再識別リスクの指標として k-匿名性 [4] を挙げている．k-匿名性は，匿名化データから特定の個人のデータを k 個未満に絞り込めるならばリスクが高いと考える指標である．そのため，k-匿名性を満たさないパーソナルデータは，k-匿名性を満たすように各種匿名化技術を用いて加工する必要がある．しかし汎用的な指標や匿名化技術の確立は困難であり，k-匿名性の利用方法や想定する匿名化技術は各文書で異なる．また k-匿名性が適用できない状況も存在する．そこで本稿では，k-匿名性を中心に各文書が定めるリスク評価方法や匿名化技術を要約し，比較や特徴の考察を行う．またリスク評価方法と匿名化技術の対応関係が必ずしも明らかでないことを課題として挙げ，その対策方針を提案する．

2. 調査対象文書

2.1 文書 (1)[7]

カナダの国・地方政府と民間の保健情報を扱う機関 Canadian Institute for Health Information (CIHI) が発行した文書である．再識別に関する透明で再現可能なリスク評価に重点が置かれている．カナダにおける保健情報の匿名化に関するベストプラクティスをまとめたガイドラインであり，保健関係省庁，病院，および地域の保健機関等に対して実務上の影響力がある文書と考えられる．

2.2 文書 (2)[8]

英国における Data Protective Direction (EU データ保護指令)，Privacy and Electronic Communications Regulation (EC 指令)，Freedom of Information Act の実施を監督し，罰金を科す権限を持つ，情報に関わる権利保護機関 Information Commissioner's Office (ICO) が発行した文書である．匿名化に関するケーススタディが豊富に例示されている．パーソナルデータを開示する機関一般を対象とした実務指針 (Code of Practice) だが，現状は実務上の影響力があるかどうか不明である．

2.3 文書 (3)[9]

U.S. Department of Health and Human Services (米国保健福祉省) 傘下の Office for Civil Rights (OCR) が発行したガイダンス文書である．1996 年に制定された米国 HIPAA 法のプライバシールールにおける匿名化基準 (De-identification Standard) を満たすために使われる二つの方法 (Expert Determination, Safe Harbor) に関して，Q&A の形式で解説している．HIPAA 法の対象事業者（ヘルスケアプロバイダ，ヘルスケア情報センター，医療保健関係者等）に対して法的な強制力を持つ文書である．

2.4 文書 (4)[10]

英国の国営医療サービスである National Health Service (NHS) が発行した文書である．英国における個人情報保護やデータ公開に関する法律を踏まえ，ヘルスケア情報に関する明確で実行可能な匿名化基準を提示しており，保健省やヘルスケア情報を開示する機関一般に対して実務上の影響力がある文書と考えられる．

2.5 文書 (5)[11]

EU のパーソナルデータの取り扱いに関する助言機関である Article 29 Data Protection Working Party (29 条作業部会) が発行した文書である．データ保護法や EU 指令との整合性から匿名化の要件を整理し，代表的なリスク指標や匿名化技術の限界とミスユースを明示している．欧州委員会や EU 加盟諸国に対する匿名化基準の見解を示した文書であり，現状は実務上の影響力があるかどうか不明である．

2.6 文書 (6)[12]

EU の情報セキュリティのアドバイザリ機関である European Union Agency for Network and Information Security (ENISA) が発行した文書である．市民のプライバシーが知らない間に侵害されるのを Privacy by Design により防止することを狙いとしており，Privacy-Enhancing Technologies (PETs) の広範な技術紹介に重点が置かれている．政策立案者，規制当局，データ保護機関，研究者，標準化組織等の多岐に渡る機関を対象とし，EU のパーソナルデータの扱いに関するリファレンスを提供している．現状は実務上の大きな影響力は無いものと思われる．

2.7 文書 (7)[13]

米国の連邦政府機関の一つで，科学技術に関連する標準化を行う機関である National Institute of Standard and Technology (NIST) が発行した文書である．当局や権利擁護団体を対象に，既知の再識別に対するリスク評価方法や匿名化技術の紹介，および匿名化に関する課題や専門用語の要約を提供している．HIPAA 法のプライバシールールにおける匿名化基準を取り上げているが，本文書自体は具体的な匿名化プロセスや推奨技術を提示するものではない．

3. 調査結果

3.1 リスク評価

パーソナルデータの開示におけるリスクは，識別 (Iden-

tification) と属性暴露 (Attribute Disclosure) に分類される [14]．前者はこれまで触れてきた再識別の問題である．属性暴露は，特定の個人の機微な情報が知られてしまうリスクである．本稿では再識別を焦点に議論を進めるが，比較的新しい文書である文書 (5), (6), (7) では，属性暴露に対するリスク指標，具体的には ℓ-多様性 [15]，t-近似性 [16]，差分プライバシー [17] も取り上げている．

以下では，各文書が想定している再識別に対するリスク評価について要約する．なおリスク評価に基づき適用する匿名化技術は，一般にパーソナルデータの品質/有用性を低下させる．そのためリスク評価はデータの品質/有用性を考慮して行う必要がある．

3.1.1 文書 (1)[7]

パーソナルデータ開示のリスクを許容可能なレベルにするため，以下の具体的なステップを挙げている．

(1) パーソナルデータに含まれる準識別子（Quasi-identifier：確率的に個人を特定し得る属性）を決定する．
(2) 再識別リスクの評価を行う．
(3) データの品質/有用性を考慮した匿名化技術の適用とリスク再評価を繰り返す．

再識別リスク評価の具体例：先ずデータ要求者の目的および能力，そして軽減コントロールについてレベル分け（Low, Medium, High の 3 段階）を行う．目的の判断要素として，データ提供者との関係，再識別による財政的な利得，再識別することの非財政的な理由の有無等が挙げられる．能力については，再識別に関する技術専門性，財源，関連性のある他のデータベースへのアクセス可能性等が挙げられる．軽減コントロールはプライバシやセキュリティの実践が良ければ High，悪ければ Low となる．そして図 1 に基づき，データ要求者が再識別を試みる確率を Remote (最も低い), Occasional, Probable, Frequent (最も高い) の 4 段階で評価する．

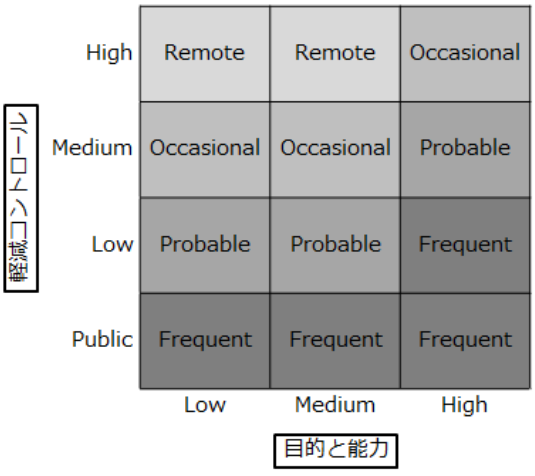

図 1 データ要求者が再識別を試みる確率の評価 ([7], Figure 5)

次に，プライバシ侵害の潜在リスクを 3 段階にレベル分けする．判断要素として，データの詳細度や機微性，意図的でないまたは許可されていない利用や後続の開示があったときの個人の損害度合い，データ要求者のロケーション（データ提供者との管轄域の差異）等が挙げられる．

最後に図 2 に基づき，k-匿名性の k の値となるリスクの閾値を 4 段階で求める．

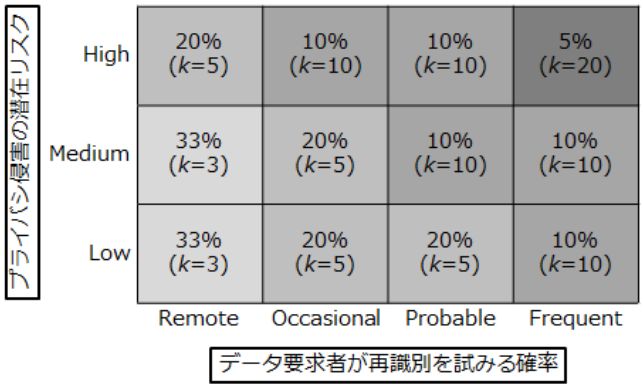

図 2 リスクの閾値の評価 ([7], Figure 6)

3.1.2 文書 (2)[8]

匿名化データは他の公開情報等のデータと突き合わせることで再識別されるリスクがあることを明確に述べている．そしてリスク評価の手段として，再識別テスト (Re-identification Testing)，すなわち匿名化データを過去に開示したデータやウェブサーチ等で得られる公開情報と突き合わせて再識別できるかどうか確認する方法が有効であるとして，次のようなケーススタディを提示している．

ある研究機関 USRC は，特別給付金の期間と，個人の年齢や BMI の関係を調査しており，調査対象者の氏名，住所，誕生日，特別給付金の期間，BMI，およびコホート研究の参照番号（以降，単に参照番号と呼ぶ）を所有している．そして別の研究機関 NRC が，USRC に対してデータ開示を要求したとする．このとき，USRC は開示する匿名化データとして，2 年以下や 5 年以上といった粗い区分の特別給付期間，BMI，5 歳刻みに丸めた年齢，および参照番号からなるデータを作成する．この匿名化データについて，先ず識別子に対する再識別リスクを評価する．参照番号は，USRC 固有の識別子であり，参照番号に紐づく氏名や住所を開示しない限りは問題ないと判断される．ただし参照番号が国民保険番号のような本人特定性の高い識別子の場合は問題となる．

その他，別の一意データとの照合可能性による再識別リスクの確認も行う．USRC は各調査対象者の月毎の BMI の変化のみを記したデータを開示しており，2011 年のデータと，2011 年 10 月から 2012 年 3 月にかけたデータが開示されている．このとき，2011 年 10 月から 12 月までのデータが一意であることから，それらを照合して 2011 年 1 月

から 2012 年 3 月までのデータを作成できてしまう．ただし特定の個人を識別できるデータは含まれていないため，再識別のリスクは高くないと判断される．

3.1.3 文書 (3)[9]

米国 HIPAA 法のプライバシールールにおける匿名化基準を満たすため，Expert Determination と Safe Harbor の何れかの方法が用いられる．

Expert Determination は，専門家が以下の原則に基づき再識別リスクの評価と軽減策を実施する．

- 再出可能性（Replicability）
  Low: 患者の血糖レベルテストの結果
  High: 患者のデモグラフィック情報
- データソースの入手可能性 (Data Source Availability)
  Low: 実験報告
  High: 患者の名前やデモグラフィック情報
- 区別可能性（Distinguishability）
  Low: 誕生年，性別，ZIP コード 3 桁の組み合わせ（ユニーク率 0.04%）
  High: 患者の誕生日，性別，ZIP コード 5 ケタの組み合わせ（ユニーク率 50%超）
- アセスリスク（Assess Risk）: 上記 3 つが高いほど再識別リスクも高くなる

また，匿名化データの一意性や別の公開情報とのリンク可能性もリスク評価の要素となり，例えば k-匿名性を用いて評価する．

Safe Harbor は，予め定められた 18 種類の識別子を抑制し，特定の個人を識別できる実知識 (Actual Knowledge) が無ければ，匿名化データとして開示可能となる．実知識とは，開示データから得られる知識であり，例えば容易に個人を特定できる特殊な職業や，親族に関する情報は匿名化技術を適用する必要がある [2]．

3.1.4 文書 (4)[10]

再識別リスクをノーマルリスクとハイリスクに分けて評価している．ハイリスクは例えば以下のような状況である．

- 攻撃者の動機が強い（例：前の妻を探したい）
- データに偏りがある（例：地域依存の病気）
- 多くの人が興味を持って探す（例：芸能人のデータ）
- 公開データとの照合が容易（例：SNS の公開データ）

そして以下の基準にしたがってリスク評価および匿名化技術の適用を行う．

ノーマルリスクの場合：

- $k = 3$ とする．
- 以下の属性は全て k-匿名化もしくは削除する．
  - 誕生日（年齢），性別，民族カテゴリ，郵便番号，イベント日（来院日等），雇用者，職業．

ハイリスクの場合：

- $k = 5$ とする．
- 1 つを除いてすべての属性を k-匿名化する．
- ただし郵便番号，生年月日，民族カテゴリは k-匿名化を必須とする．
- k-匿名化しない属性についても完全なものを出すべきでない．

その他，集計表を開示する場合においてもノーマルリスクとハイリスクに分けて明確な基準を設定している．集計表の開示によるリスクを低減させるため，公的統計（国や自治体等が実施する統計）において実利用されている統計的開示抑制（SDC: Statistical Disclosure Control）の技術を適用する．

3.1.5 文書 (5)[11]

匿名化基準において考慮すべき以下の 3 つのリスクを挙げ，各匿名化技術の当該リスクに対する適用効果を評価している．

- Singling out: 全体のデータから特定の個人のレコードを抽出できる．
- Linkability: 単一または複数のデータセットから，同一個人のレコードを抽出できる．
- Inference: ある属性の値から他の属性の値を高い確率で推定できる．

なお仮名化については，元データとの Linkability は低減できるが，これは匿名化ではなくセキュリティ手法として区別している．

上記リスクに対し，k-匿名化の適用効果を次のように評価している．

- Singling out: 同じ属性値を持つ k 人以上のグループに対する再識別はできない．
- Linkability: $1/k$ を超える確率でレコードを対応付けることはできない．
- Inference: 推定のリスクは残る．ある属性の値が同一の k 人以上のグループにおいて，推定を試みる属性値が全員同じであれば，属性値は特定される．

また，k-匿名性に関するよくある間違いとして以下を挙げている．

- 準識別子の欠落：準識別子としなかった属性から再識別されるリスクが生じる．
- k の値が小さい：一般に k の値が小さいほどデータの品質/有用性が高くなるといわれるが，同一グループにおける個々人のウェイトが高くなり，推定のリスクも高くなる．
- 同じウェイトでグルーピングしない：属性値の分布が等しくないグルーピングは，それによって各レコードの影響度が変化する問題がある．

3.1.6 文書 (6)[12]

再識別のリスク評価や匿名化技術を SDC として位置付けており，ミクロデータ，集計表，およびデータベースへの統計クエリに分けてそれぞれの匿名化技術を紹介している．k-匿名性についてはその指標の特徴の説明にとどまっ

ている．

3.1.7 文書 (7)[13]

攻撃モデルについて，以下のように攻撃の動機，再識別シナリオ，攻撃者の能力に分けてそれぞれ具体例を挙げ，リスク評価の算出要素としている．

攻撃の動機：

- 匿名化の評価
- 再識別による名声・専門性の獲得
- データ提供者に対する嫌がらせや危害
- 再識別することによる直接的な利益
- 再識別によって得られた機微情報を用いた嫌がらせや危害

再識別シナリオ：

- 検察官：匿名化データにターゲットが含まれていることを知っている．
- ジャーナリスト：誰でもいいので少なくとも一人を再識別する．
- マーケッター：なるべく多くの人を再識別する．
- 差分識別可能性：特定の個人を含むデータセットと含まないデータセットの識別可能性

攻撃者の能力：

- 一般大衆：公開されたデータにアクセスできる．
- 専門家：匿名化技術を保有している．
- 内部犯：データ提供側の構成員．
- 内部犯（精通者）：データ要求側で，一般大衆よりも詳しい情報を持つ．
- プロの犯罪者：内部利用や再販のために元データと匿名化データの両方を入手できる．
- 詮索好きな隣人：ターゲットの家族，友人，かかりつけ医者等．

そして文書 (1) 同様，パーソナルデータ開示のリスクを許容可能なレベルにするための具体的なステップを挙げている．

3.2 匿名化技術

3.2.1 文書 (1)[7]

再識別リスクを軽減する加工方法として，準識別子に適用する詳細データの簡略化 (Reduction in Detail)，識別子に適用する置換 (Substitution) や仮名化，識別子と準識別子の両方に適用できる抑制 (Suppression)，そして同じく両方に適用できるが推奨はできないノイズ付加が挙げられている．

詳細データの簡略化は，例えば郵便番号を最初の 3 桁のみとしたり，生年月日を生年に丸めたりする処理であり，一般化 (Generalization) あるいは再符号化 (Recoding) とも呼ばれる．準識別子が組み合わせ一意となるデータの数を削減することが目的であり，最も一般的な加工方法である．

表 1 元のパーソナルデータ ([8], p.86)

Age	Sex	Postcode	Income	Expense
22	F	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	M	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

表 2 大域的再符号化を適用したパーソナルデータ ([8], p.86)

Age	Sex	Postcode	Income	Expense
20–24	F	SO17–19	low	low
25–29	M	SO17–19	low	low
30–34	M	SO14–16	medium	medium
35–39	F	SO17–19	medium	medium
40–44	F	SO14–16	high	high
50–54	M	SO14–16	medium	low

抑制は，変数（属性）単位，レコード単位，セル単位に分類される．再識別リスクの高いレコードの削除や，珍しい診断結果，個人を一意に特定できる職業等の外れ値の削除等が該当する．

置換は，値のランダム化や別のレコードの値との置換 (スワッピング) がある．主な課題として，再識別リスクの評価方法が挙げられる．

ノイズ付加は，一定範囲内の乱数を元の値に足し合わせる．例えば，年齢が変わらないよう誕生日を変化させたり，おおよその位置は保ったまま郵便番号を変化させる．ただしデータ要求者は，データの信頼性の点でノイズ付加を好まず，再符号化を好むと主張している．

3.2.2 文書 (2)[8]

再識別リスクを軽減する加工方法を，データの簡略化 (Data Reduction)，データの撹乱 (Data Perturbation)，および非撹乱手法 (Non-perturbation Methods) に分類している．

データの簡略化には，変数の削除，レコードの削除，大域的再符号化（Global Recoding），局所秘匿（Local Suppression）の方法がある．変数の削除は直接的または間接的な識別子，そして非常に機微なものが対象となる．レコードの削除は，攻撃者（Intruder）が容易に個人を特定できるような特殊なレコードが対象となる．大域的再符号化は，事前に定義されたデータの一般化を行う．例えば表 1 のパーソナルデータについて，表 2 のように郵便番号，収入，出費の値を一般化する．局所秘匿は，例えば変数の組合せの値が一意となり攻撃者が個人を特定できるリスクが高いときに，値の一部を秘匿する．表 2 において，攻撃者が 20 代女性の情報を持っているとしたとき，年齢の 20–24 を ”missing” に置き換える．

データの撹乱には，ミクロアグリゲーション，スワッピング，PRAM (Post-Randomisation Method)，ノイズ付

表 3 ミクロアグリゲーションを適用したパーソナルデータ ([8], p.91)

Group	Age	Sex	Postcode	Income	Expense
G1	22	F	SO17	£23,333	£1,100
G1	25	M	SO18	£23,333	£1,300
G2	30	M	SO16	£43,833	£1,800
G2	35	F	SO17	£43,833	£2,000
G2	40	F	SO15	£43,833	£3,500
G1	50	M	SO14	£23,333	£1,200

表 4 PRAM を適用したパーソナルデータ ([8], p.95)

Age	Sex	Postcode	Income	Expense
22	M	SO17	£20,000	£1,100
25	M	SO18	£22,000	£1,300
30	F	SO16	£32,000	£1,800
35	F	SO17	£31,500	£2,000
40	F	SO15	£68,000	£3,500
50	M	SO14	£28,000	£1,200

表 5 ノイズ付加を適用したパーソナルデータ ([8], p.97)

Age	Sex	Postcode	Income	Expense
22	F	SO17	£19,828	£1,100
25	M	SO18	£23,862	£1,300
30	M	SO16	£32,960	£1,800
35	F	SO17	£28,957	£2,000
40	F	SO15	£66,951	£3,500
50	M	SO14	£27,692	£1,200

加，リサンプリングの方法がある．ミクロアグリゲーションは，閾値 k を設定し，k 個以上のレコードを含むグループを決定する．そして同一グループ内で各値を平均等の代表値に置き換える．表 3 は，表 1 をグループ 1(G1) とグループ 2(G2) に分け（すなわち $k = 3$），収入を平均に置き換えた例である．スワッピングはレコード間で一部の値を置換する．置換の割合は 1〜10%程度が代表的とされている．PRAM はある属性の値に対してマルコフ遷移確率行列に基づき確率的にランダムな値に変換する．表 4 は性別を確率的にランダムな値とした例である．ノイズ付加は一部の値に対してある分布にしたがって生成されたランダムな値を付加する．表 5 は収入にノイズを付加した例である．リサンプリングは，元のデータからランダムに標本を抽出することで得られる統計的推定の方法であり，例えば数値属性に対して平均を保つよう値を変換する．

非撹乱手法には，サンプリングとクロス集計がある．サンプリングは元のデータから一定の抽出割合でランダムにデータを抽出するランダムサンプリングと，レコード毎に独立に一定の確率で抽出するベルヌーイサンプリングがある．サンプリングデータは一般に特定の個人のデータが存在するか否か判別し難くなるため，再識別リスクは軽減すると考えられる．ただし，カテゴリ値はサンプリングが有効とされるが，数値については明らかでない．クロス集計は，個々のデータを集約するため一般に再識別リスクは軽減すると考えられる．ただし集計値がゼロや少数の場合は注意が必要となる．

3.2.3 文書 (3)[9]

開示リスクとデータ有用性のバランスが重要として，抑制 (Suppression)，一般化 (Generalization)，撹乱 (Perturbation) を用いた匿名化について例示している．その内容については 3.2.2 節と同様のため省略する．

3.2.4 文書 (4)[10]

ミクロデータに対する匿名化技術としては，k-匿名性を満たすための抑制 (Suppression) および簡略化 (Reduction) の説明にとどまっている．

3.2.5 文書 (5)[11]

再識別リスクを軽減する加工方法を，ランダム化 (Randomization) と一般化 (Generalization) に分類している．また各加工方法について Singling out, Linkability, Inference の 3 つのリスクに対する適用効果を評価しており，利用時のよくある間違いを列挙している．ミクロデータに対するランダム化には，ノイズ付加と置換が挙げられている．

3.2.6 文書 (6)[12]

様々な匿名化技術が紹介されている．ミクロデータに対する匿名化技術として，撹乱的マスキング (Perturbative Masking) と非撹乱的マスキング (Non-perturbative Masking) に分類している．

撹乱的マスキングは，ノイズ付加，ミクロアグリゲーション，スワッピング，PRAM の方法がある．非撹乱的マスキングは，サンプリング，一般化，再符号化 (Top/Bottom Coding)，局所秘匿の方法がある．

3.2.7 文書 (7)[13]

匿名化技術として他の文書同様，抑制，一般化，撹乱，スワッピング，サンプリングを挙げている．ただしこれらの技術を準識別子に適用することを明確に述べており，k-匿名化の要素技術として位置付けている．

4. 考察

4.1 リスク評価

既に実務上の影響力または法的な強制力を持つ文書と考えられる文書 (1), (3), (4) は，具体的かつ実践的な評価方法が提供されている．一方でその他の文書は，パーソナルデータの開示について再識別に限定しない様々な潜在的リスクが検討されており，将来的な対策の必要性を示唆している．しかし匿名化基準は開示リスクの軽減とともにデータの品質/有用性もなるべく損ねないことが重要であるため，それらを高いレベルで両立させる技術開発や運用手順の確立が今後益々重要になるものと考えられる．

4.2 匿名化技術の比較

3.2.1 節から 3.2.7 節までに挙げた匿名化技術を表 6 にま

とめた．k-匿名化の要素技術である抑制や再符号化は，ほぼ全ての文書が利用対象として想定していることが分かる．またノイズ付加も多くの文書が触れているが，文書 (1) は分析結果を歪めることを問題視しており，文書 (3) は平均や分散等の統計量を損ねないようなノイズ付加を推奨している．特に文書 (1), (3), (4) は医療分野を対象とした文書であり，スワッピング，ミクロアグリゲーション，PRAM 等の撹乱的な手法にほとんど触れていないことからも，分析結果の歪みに特に敏感であると考えられる．しかし一方で，PRAM やノイズ付加を適用した匿名化データから，元のデータの集計値を推定する方法も知られており [18], [19]，匿名化データの品質/有用性の維持の観点からも，撹乱的な手法の有効性は再考の余地がある．

4.3 課題

本調査で挙げたリスク評価方法と匿名化技術について，それらをどのように組み合わせて匿名化データを作成すればよいか，手順やユースケースを明確化することが課題として挙げられる．例えば文書 (3) では専門家が作成するとあり，その具体的な手順は必ずしも明らかではない．専門家以外でも匿名化データを作成できれば，適切な匿名化の実施，パーソナルデータの利活用促進につながることが期待できる．前記の課題について，k-匿名性の指標としての限界が挙げられる．すなわち，k-匿名性は抑制や再符号化の充足度を測ることはできるが，撹乱的な手法については明らかでなく，専門家の経験則等に委ねざるを得ない場合がある．しかし最近では，撹乱的な手法に対する k-匿名性と同等の指標も提案されてきており [18], [19]，これらをリスク評価や匿名化プロセスに組み込む方法を確立することで，手順の明確化が進むものと考えられる．

5. おわりに

匿名化基準に関して先進的な欧米の公的文書 7 選を対象に，リスク評価方法と匿名化技術の調査を行った．既に実務上の影響力または法的な強制力を持つと考えられる文書は，具体的かつ実践的な評価方法が提供されており，その他の文書は，将来的なリスク対策の必要性を示唆している傾向がみられた．匿名化技術には，リスクの軽減には有効だが分析結果を歪める撹乱的な手法がいくつか存在する．これらの適切な利用や技術改善が今後の大きな課題と考えられる．特に撹乱的な手法に対するリスク評価方法の確立は急務である．

今回はパーソナルデータの開示における再識別リスク評価方法と匿名化技術に重点を置いて調査を行ったが，属性の推定に対するリスク評価，集計表やデータベースへの統計クエリにおいて有効な匿名化技術，そして匿名化プロセスに関する調査や考察等，検討すべき項目は多く，今後の課題である．

謝辞 本稿は，2016 年 7 月 12 日に電気通信大学にて開催された 2016 年度第 2 回 PWS 勉強会 (タイトル：海外匿名化基準サーベイ，発表者：吉浦，島岡，千田) の発表内容に基づき執筆しました．同 PWS 勉強会で熱心にご議論頂いた参加者の皆様に感謝いたします．また，2015 年 5 月 28 日に明治大学中野キャンパスにて開催された 2015 年度第 1 回 PWS 勉強会（タイトル：匿名性規準に関する国内外事例，発表者：美馬正司氏（株式会社 日立コンサルティング)）の発表内容も本執筆にあたり参考となりました．美馬氏に感謝いたします．

参考文献

[1] ISO/TS 25237:2008 (en), Health informatics — Pseudonymization, https://www.iso.org/obp/ui/#iso:std:iso:ts:25237:ed-1:v1:en
[2] K. El Emam and L. Arbuckle (木村映善，魔狸，笹井崇司 訳), データ匿名化手法, オライリー・ジャパン, 2015.
[3] L. Sweeney, “Uniqueness of simple demographics in the U.S. population,” LIDAP-WP4, Carnegie Mellon University, Laboratory for International Data Privacy, 2000.
[4] L. Sweeney, “k-anonymity: a model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, **10(5)**, pp. 557–570, 2002.
[5] A. Narayanan and V. Shmatikov, “How to break anonymity of the Netflix Prize Dataset,” arXive.org. Retrieved 19 January 2014.
[6] 個人情報保護委員会，匿名加工情報に関する委員会規則等の方向性について （平成 28 年 6 月 3 日) http://www.ppc.go.jp/files/pdf/280603_siryou2.pdf
[7] ‘Best Practice’ Guidelines for Managing the Disclosure of De-Identified Health Information, Canadian Institute for Health Information (2010)
[8] Anonymisation: managing data protection risk code of practice, Information Commissioner’s Office (2012)
[9] Guidance Regarding Methods for De-identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule, Office for Civil Rights, U.S. Department of Health and Human Services (2012)
[10] Anonymisation Standard for Publishing Health and Social Care Data Specification, National Health Service (2013)
[11] Opinion 05/2014 on Anonymisation Techniques, ARTICLE 29 DATA PROTECTION WORKING PARTY (2014)
[12] Privacy and Data Protection by Design - from policy to engineering, ENISA (2014)
[13] De-Identification of Personal Information, NIST (2015)
[14] D. Lambert, “Measure of disclosure risk and harm,” Journal of Official Statistics, **9(2)**, pp. 313–331, 1993.
[15] A. Machanavajjhala, J. Gehrke, D. Kiefer, and M. Venkitasubramanian, “ℓ-diversity: privacy beyond k-anonymity,” ACM Trans. Knowl. Discov. Data, **1(1)**, 2007.
[16] N. Li and T. Li, “t-closeness: privacy beyond k-anonymity and ℓ-diversity,” Proc. of ICDE 2007, IEEE, 2007.
[17] C. Dwork, “Differential privacy,” ICALP 2006, Lecture Notes in Computer Science, **4052**, pp.1–12, 2006.

表 6 各文書が対象としている匿名化技術の比較

文書	抑制	再符号化	ノイズ付加	スワッピング	ミクロアグリゲーション	PRAM	サンプリング
文書 (1)[7]	✓	✓	✓	✓			
文書 (2)[8]	✓	✓	✓	✓	✓	✓	✓
文書 (3)[9]	✓	✓	✓				
文書 (4)[10]	✓	✓					
文書 (5)[11]		✓	✓	✓			
文書 (6)[12]	✓	✓	✓	✓	✓	✓	✓
文書 (7)[13]	✓	✓	✓	✓			

[18] D. Ikarashi, R. Kikuchi, K. Chida, and K. Takahashi, "k-anonymous microdata release via post randomisation method," IWSEC 2015, Lecture Notes in Computer Science, **9241**, pp. 225–241, 2015.

[19] 五十嵐大，長谷川聡，納竜也，菊池亮，千田浩司，「数値属性に適用可能な，ランダム化により k-匿名性を保証するプライバシー保護クロス集計」，コンピュータセキュリティシンポジウム 2012 (CSS2012)，情報処理学会，2012.