公募シンポジウム
# 公募シンポジウム8
# 医療情報の Public Use File提供にむけた検討
2018年11月25日(日) 09:00 〜 10:30  B会場 (4F 409+410)

## [4-B-1-1] 医療情報の Public Use File提供にむけた検討
○木村 映善 （愛媛大学大学院医学系研究科）

次世代医療基盤法の施行により，認定事業者による医療情報収集および医療分野の研究開発に資するような匿名加工医療情報の作成が可能になった．レセプト情報・特定健診等情報の提供に関しても，公益性の高いものにたいして NDBデータの民間提供の試みが進められており， NDBオープンデータも版を重ねるにつれて収録項目や件数も拡大している．また，国が実施した統計調査データについても，公益性の高い研究にかんして一般の研究者も利用可能にする改正統計法が成立している．以上のように「公益性」のある二次利用への情報提供の機会が拡大している．

しかし，医療分野における Public Use File(PUF)の提供の試みは進んでいない．先述した試みにみられるのは，審査を通して提供される「事実上の匿名性」が施されたミクロデータあるいはオーダメイド集計である．「絶対的な匿名性」を確保すると，データの品質が著しく低下するものとして，我が国ではデータ提供の形態検討から外される傾向がある．しかし， PUFを利用した分析はただちにエビデンスに結びつかないものの，下記の点において期待がかけられる．(1)今後の医療情報の対象範囲の拡大に伴う医療情報モデルの複雑化により，標準医療情報モデルに従って記述されたデータに習熟する必要があるが， PUFにより事前にプロトタイプ構築を始められること，(2)変数選択による影響を確認できるため， Scientific Use File提供時の審査やリスク評価のプロセスを迅速化させること， (3)探索的検討やデータマイニング，機械学習等の多くの変数を利用する解析・開発に供することができること，である．本論では医療情報分野における PUFの提供にむけた課題の提示を行う．

# 保健医療情報の Public Use File 作成にむけて

木村映善[*1]
*1 国立保健医療科学院
Eizen Kimura[*1]
*1 National Institute of Public Health

By enforcing 'the next-generation medical infrastructure law', it became possible to create anonymous medical information that will contribute to the research and development in the medical field. However, attempts to provide Public Use File (PUF) in the medical field have not progressed yet in Japan. PUF tends to be excluded from consideration of the form of data provided because the data quality is considerably deteriorated by securing "absolute anonymity". The analysis using PUF does not immediately lead to sound evidence. However, the expectation is given in the following points. (1) It is necessary to familiarize the standard medical information model due to the complication of the medical information model. It is possible to start the prototype construction beforehand of the study using real data, (2) Since it is possible to check the influence of variable selection, accelerate the review process and risk assessment process when providing Scientific Use File (SUF), (3) PUF is able to provide many variables for such as exploratory study, data mining, machine learning etc. In this paper, we will present the challenges for providing PUF in the medical information field.

Keywords: Please input 3 to 5 keywords in English.

## 背景

我が国では進行する高齢化社会において，保健医療制度を維持するべく，多角的な医療情報を用い，分析することで保険運営の改善，予防医療の推進，エビデンスにもとづく医療政策の画策を実現するデータヘルスの推進を図るべく，データヘルス改革推進本部が設置された．

データヘルスの実現には，全国からの医療情報を収集し，分析者がアクセスできることが前提である．本稿では，我が国での医療情報の提供のありかたについて確認し，これまで取られていない提供形態であり，今後の研究開発に貢献するであろう Public Use File の提供に向けた提案をする．

## 統計データの種類

医療情報の提供形態について整理する．医療情報は要配慮個人情報を含むセンシティブなデータで構成されている．すなわち，データ収集時はオリジナルのデータ（個票データ）のかたちで蓄積されるが，ただちに第三者に開示することは認められない．そこで集計処理等の処理を施して高次元の情報に集約する統計表と，オリジナルのデータにちかい形態でありながら個人の識別特定を困難にする加工や，アクセスに制約を加えることで提供を許可する形態であるミクロデータがある．

現在，統計表は政府統計を提供する e-Stat で公開されている．また，厚生労働省からは NDB を有用性の高い指標に関して集計した統計表が NDB オープンデータとして公開されている．ミクロデータは事前に公開様式が定められたオーダメイド集計，要求に応じて分析，結果のみを抽出するリモート集計，個人を特定できないように匿名化が施された Scientific Use File(SUF),Public Use File(PUF)の形態[1]，そしてオンサイトで閲覧，分析環境を提供し集計結果のみを提供する形態がある．SUF は匿名化を施した結果，著しく時間と経費，労力をかけない限りファイルに含まれる個体を識別できない状態である「事実上の匿名性」を確保したものである．一方，PUF は確実に個体識別の可能性が取り除かれた状態である「絶対的な匿名性」の状態を確保したものである[2]．SUF に比較して PUF はランダムな要素を持つ処理を加える攪乱，模造的手法を取り入れて固体識別の可能性を取り除く分，データ品質の低下が生じる．

## 医療情報のミクロデータ提供状況

現在は，厚生労働省によるレセプト情報等の提供は，研究の公共性，分析対象の変数選択の必然性，情報管理体制等を吟味の上，SUF に相当するファイルが作成されて研究者に提供される．あるいは要望に応じて集計を実施して集計表を提供するオーダメイド集計が行われている．

また，探索的な研究を可能にするために，レセプト情報等オンサイトリサーチセンターが開設され，オンサイトでの閲覧，集計が可能になるように進められている．

次世代医療基盤法下の認定事業者による匿名加工医療情報も SUF に分類されるであろう．

以上のように，我が国では医療情報における統計表，SUF の提供，オンサイトの環境整備が進められているが，PUF への取り組みは未着手である．

## 海外の PUF を利用した事例

Medicare のレセプトデータの研究，教育目的に匿名加工処理を加えた Synthetic Public Use Files(SynPUFS)が公開されている[*3]．変数の数は大元より減らされているものの，SynPUFS のデータ構造は CMS Limited Data Set に類似しているため，CMS のデータファイルを活用するプログラムのプロトタイピングに利用可能である．この SynPUFS を利用して様々な研究がなされている．例えば，てんかんの脳波計側中の心停止へのリスク対処の必要性の分析[4]，クラウド上での喘息患者の再入院リスク因子を特定する機械学習の実装とビッグデータに対応できるスケーラビリティの検証[5],数種類の機械学習を適用して医療費増大に寄与する因子を探索する試み[6]等，SynPUFS を利用した論文が発表されている．これらの論文等で確認できることは，(1)大局的な傾向を俯瞰し，各々のデータの精度をさほど要求しないもの，(2)ビッグデータのスケーラビリティに対応するシステムの実装に関するフィ

ージビリティスタディ，(3)機械学習やデータマイニング等，多次元かつ大量のデータセットを利用した探索的な分析等，SUF では困難な研究があるということである．SUF の性質上，リスクを軽減するために変数の数の制限は行われるし，処理がスケールしやすいパブリッククラウド上に展開するにも制約が伴う可能性がある．つまり，PUF はビッグデータを利用した分析からエビデンスを得ようとする一部の試みや，機械学習等，変数選択において事前に仮定をおかない探索的分析の用途に向いている．オンサイトリサーチセンターで一定程度の探索的分析は可能かも知れないが，大規模な分析環境を提供できるパブリッククラウドとは計算機資源が比べものにならない．開発の速度が要求されるような分野ではオンサイトリサーチセンターが提供する計算機資源では追い付かない可能性がある．現在，我が国は AI の開発を支援する体制が推進されているが，その支援策の一環として高性能な計算機環境に投与できる PUF データの開発を推進することも必要ではないかと考える．

## PUF 作成に向けた課題

データ工学的に完全に匿名化することは，自然文章等を含まない数値，カテゴリカル変数のみで構成されたデータにおいて可能である．筆者らは Pk-匿名化という，k-匿名性をみたすように攪乱を施して元データを破壊すること，特定個人の識別特定可能性を消去する手法の適用を試みた[7]．一方で元データがもつ統計的特徴は維持しているため，元データを用いた分析の結果に近い結果が得られることを期待したものである．既に単純な統計量や集計については高い精度を実現している.しかしながら，どのようなデータ分析に対しても高い精度を提供できる万能な匿名加工処理はない．つまり，PUF も単一の匿名化手法の適用ではなく，目的別に重要視する統計量を保護するような匿名加工処理を施したうえで，有用な分析が可能であるかの実証的な試みが不足している．

また公開したものは回収が不可能であるため，PUF をリリースするプロセスに瑕疵が生じないように配慮が必要である．PUF の匿名化手法の検討，有用性やリスクの評価方法，PUF の作成とリリースの体制についての検討が必要である．

## まとめ

本稿では日本での事例がない医療 PUF の作成を検討することを公開した．その根拠として海外での PUF を用いた事例から機械学習や医療政策の検討等への有用性が確認されていることを挙げた．PUF たるために「絶対的な匿名化」を施すにあたり，様々なユースケースを想定して，それぞれの重要視する統計量に最適化された手法を選択することの知見の蓄積，PUF の作成やリリースの体制構築にむけた実証実験の必要性について指摘した．今後，匿名加工コンテスト等を通して PUF 公開への準備を進めていきたいと考えている．

## 参考文献

1.伊藤 伸．諸外国における政府統計ミクロデータの提供の現状とわが国の課題．中央大学経済研究所年報．2016(48):233-49.
2. 星野 伸．公的統計ミクロデータ提供制度の課題．日本統計学会誌 シリーズ J. 2010;40(1):23-45.
3.Centers for Medicare and Medicaid Services. Medicare Claims Synthetic Public Use Files (SynPUFs). https://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/index.html 2015
4.Malik AA, Ullah N, Adil MM, Qureshi AI. Risk of in-hospital cardiac arrest among medicare beneficiaries undergoing video electroencephalographic monitoring. Journal of vascular and interventional neurology. 2015 Oct;8(4):39.
5.Chen R, Su H, Khalilia M, Lin S, Peng Y, Davis T, Hirsh DA, Searles E, Tejedor-Sojo J, Thompson M, Sun J. Cloud-based predictive modeling system and its application to asthma readmission prediction. InAMIA Annual Symposium Proceedings 2015 (Vol. 2015, p. 406). American Medical Informatics Association.
6.Lahiri CB, Agarwal N. Predicting healthcare expenditure increase for an individual from medicare data. InProceedings of the ACM SIGKDD Workshop on Health Informatics 2014.
7.Kimura Eizen, Chida Kouji, Ikarashi Daisuke, Hamada Kouki, Ishihara Ken. Statistical disclosure limitation of health data based on pk-anonymity. Stud Health Technol Inform. 2012;180:1117-9.