資料１－１

調査票情報の二次利用の現状と解決策

重岡 仁 (しげおか　ひとし)
東京大学 公共政策大学院 教授

規制改革推進会議
医療・介護・感染症対策ワーキング・グループ
2023年3月6日

本日の内容

1. 調査票情報の二次利用の現状
2. 現状がいかにEBPMの推進の足かせか
3. 具体的な解決策の提案

調査票情報の二次利用の現状

- EBPMの推進のためには、調査票情報の円滑・迅速な二次利用が不可欠
- しかし、データ申請からデータ取得まで、一般的に早くて数か月、場合によっては1年以上かかる。
- 私自身の経験:
  - 昨年3月に二次利用申請したが、役所の担当者と「なぜこの変数が必要か？」というやり取りを5、6回続け、申請する変数を大幅に減らした上、申請からデータ入手まで約1年かかった。
- なぜこんなことが起こるのか？

一般的な調査票情報の利用の流れ

研究計画書の作成

科研費申請

約6か月

科研費
KAKENHI

専門家による審査

採択率は分野によるが
高くても30-40%程度

データ利用申請

早くて3か月？

1年以上かかることも頻出。

一般的な調査票情報の利用の流れ

研究計画書の作成

科研費申請

約6か月

科研費
KAKENHI

専門家による審査

採択率は分野によるが
高くても30-40%程度

データ利用申請

早くて3か月？

1年以上かかることも頻出。

- 政府内の他省庁等によるEBPMのための利用（単年度の委託事業等を含む）でも同様の手続き

申請手続きに必要な書類

01-1. 申出書

01-2. 申出書別紙

02. 調査票情報に係る管理簿

03. 別添 1-1 利用する調査事項

04. 別添 2-1 研究概要

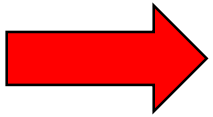
04. 別添 2-2 集計様式

05. 依頼書

06. 誓約書

07. 室内図

08. 科研費に関する書類



1. (まだ見ぬ)データをどのように分析し、
2. 最終的にどのような表やグラフを作成するか、
3. そのために必要な変数と、それが必要な理由の詳細な記載が必要。

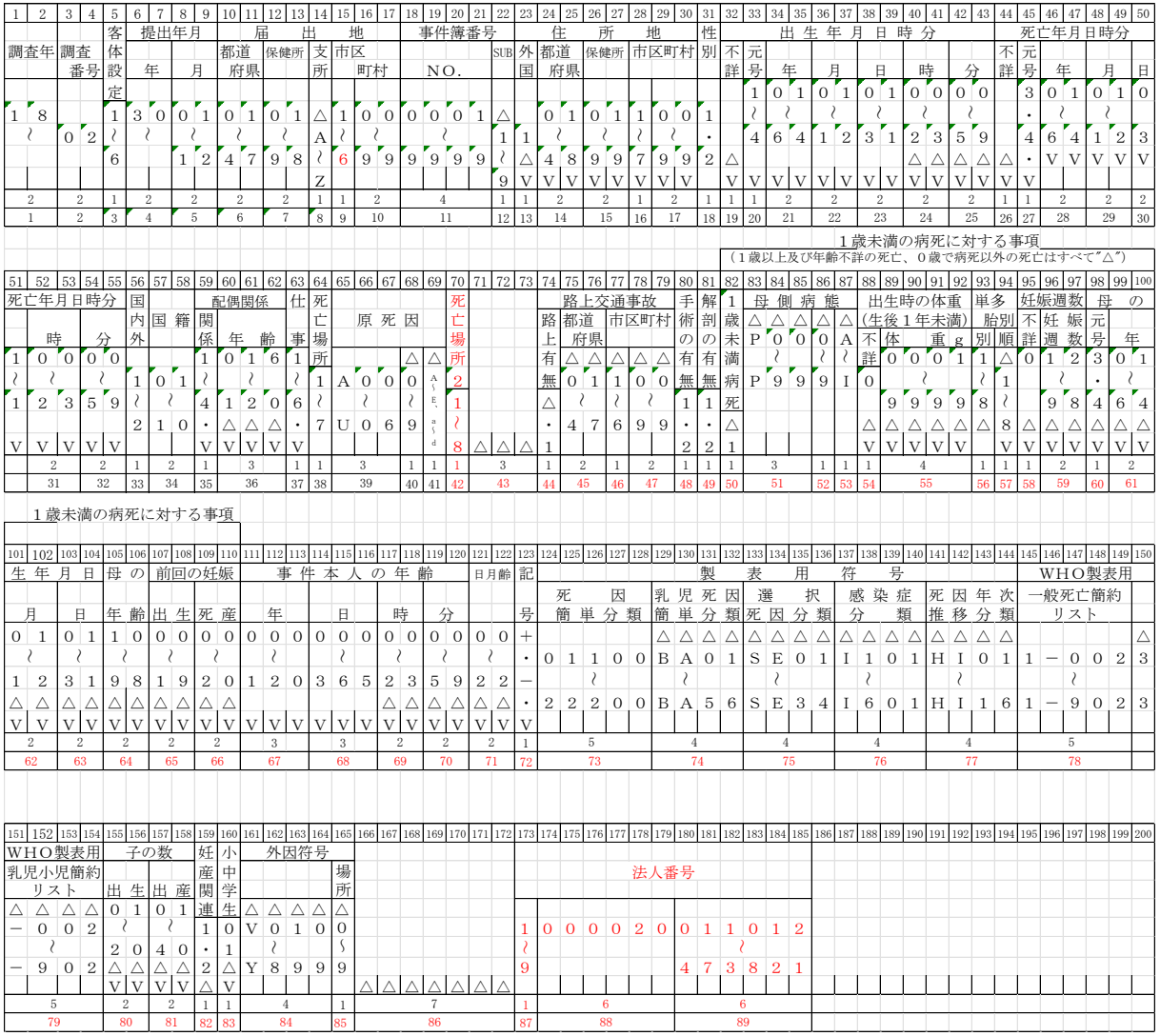
⇒ 申請から漏れた表やグラフを論文に掲載できない

⇒ リスク回避で、申請者は(事後的に)必要となるよりも、はるかに多くの表やグラフを申請書に記載せざる得ない。

申請手続き

例： 人口動態調査（死亡票）

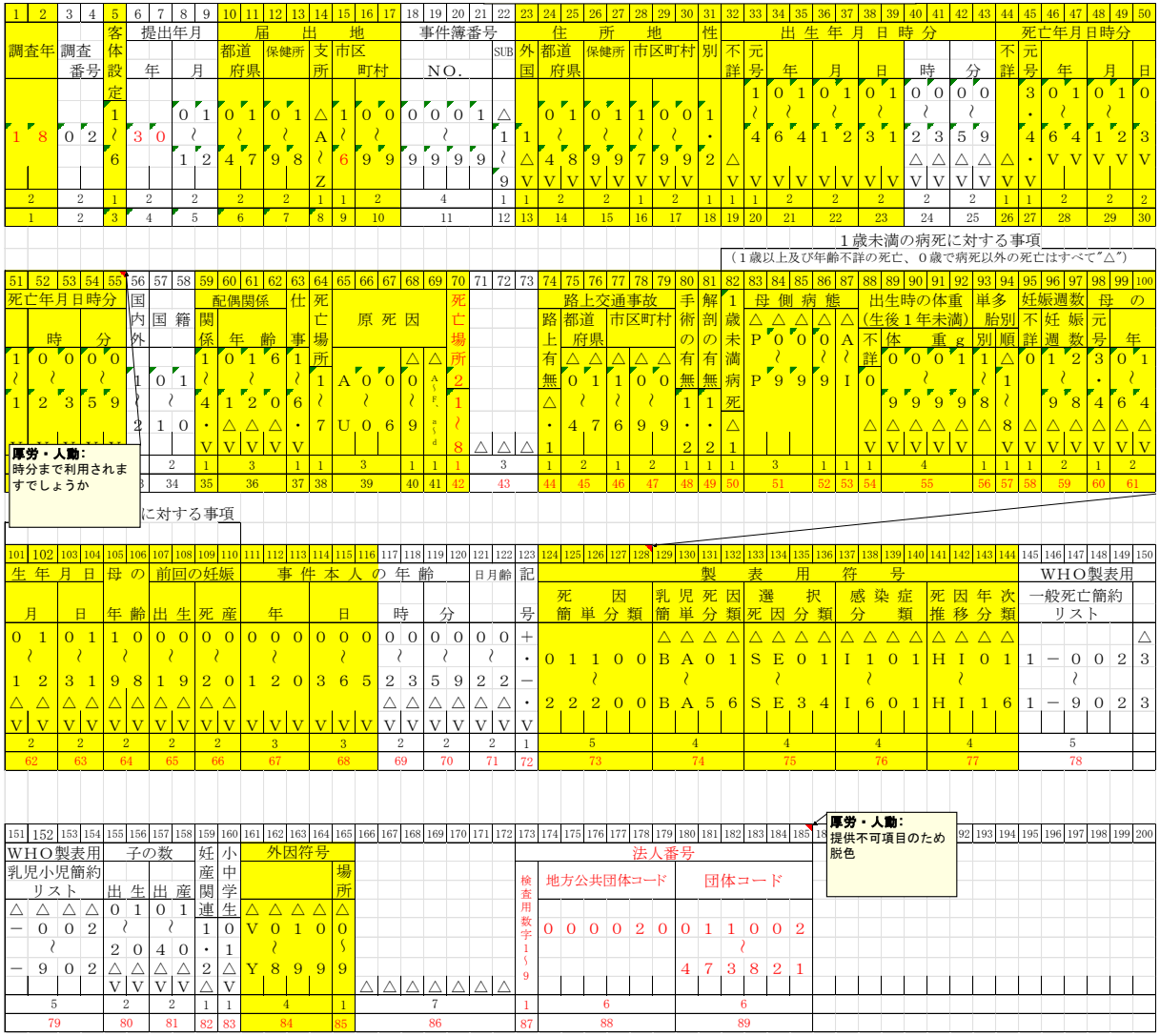
- 必要な変数に色を塗る(!?)方式
- (ひと昔まではすべての変数を抜き出し、列挙させていた)



提出

- この作業を**全ての**調査票、年度について行う必要あり
- 年度によって調査票が異なる(変数の位置も違う)ので、申請時に変数を見落とす可能性大

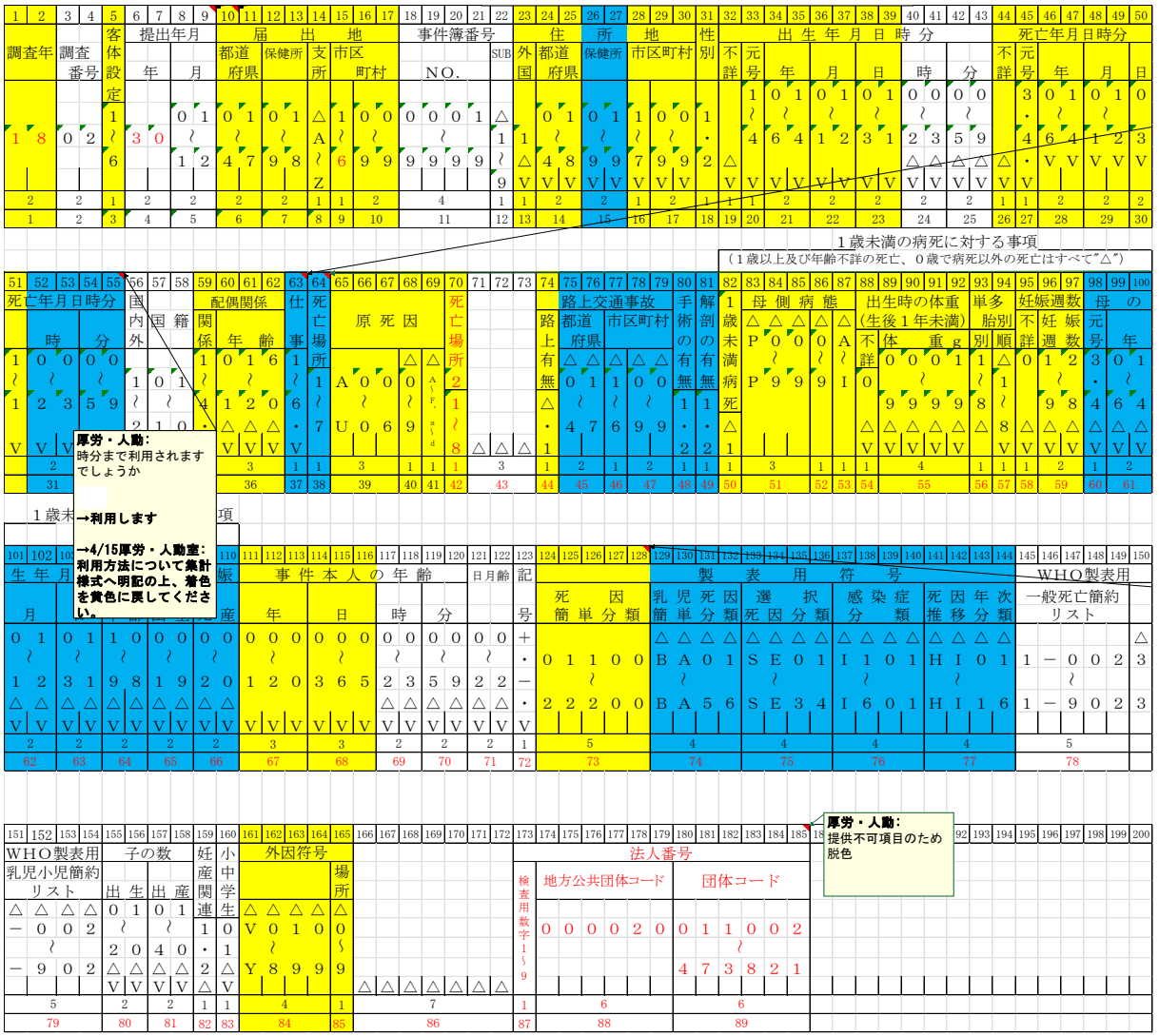
例： 人口動態調査（死亡票）



返却

- 青色の変数は、「利用する**必要性が計画書から読み取れない**」ので修正するようにとの依頼を受ける
- 「提出→返却」を何往復も繰り返す
- 順調な場合は週に1度の頻度で返信。ただし、返信が滞ることもしばしば（**担当者に大きく依存**）。

➡ 2次利用申請→データ入手まで、1年以上かかることが頻出



甚大なる弊害

1. **まず第一に、無駄に時間がかかる!!!**
   - 1年以上待たされて、どうしてタイムリーな研究などできようか？
   - コロナのように一刻を急ぐ場合はさらに致命的
   - 研究者側、役所側ともに、時間と人材の浪費としか言いようがない

2. **いつデータを入手できるが不確実** (平均ではなく分散の話)
   - 特にパーマネント職のない若手研究者が利用を躊躇

   ⇒ 手続きの煩雑さ、不確実性が理由で、多くの研究者が実際に申請を断念

3. **変数の抜き出しにミスが起こる。**
   - データ提供側も申請ごとにいちいちデータを切り出す必要あり
   - 申請した変数が出てこないミスが起こる、時間もかかる

4. **研究者間で知識の蓄積が進まない。**
   - 利用者ごとに使用データが異なる ⇒データ前処理の蓄積が全く進まず
   - 同一データであれば、データの読み込みやクリーニング、頻繁に使う変数の作成などについて、研究者間でプログラムを共有できる
   - 例）出生票（米国) ⇒ 異なる統計ソフトに対応したプログラムを共有
     https://www.nber.org/research/data/vital-statistics-natality-birth-data

具体的な解決方法

1. リモートアクセス
⇒期限を定めて、至急進めるべき

2. 現存の磁気データによる提供
⇒ 1は時間がかかることが予想される
⇒ 手続きの簡素化が不可欠

- ただし、公開可能なデータは積極的に公開！
  - 例えば、このサイトで、世界中の国勢調査(センサス)がワンクリックでダウンロードできるが、日本は未だ参加せず
  - リモートアクセスのみになるとサーバーのキャパや審査の人員が足りず、システムがパンクする恐れあり

1. リモートアクセス

- 仕組み: 各自の研究室や自宅より、総務省統計局管理のサーバにアクセス
  ⇒ その中ですべての統計分析を行う
  ⇒ 分析の結果だけを取り出す

- 利点: **全ての変数、調査票情報**を提供可能
  ⇒ 安全性が確保されているリモートアクセスで変数を制限する理由がない

- 懸念点: 審査者による出力結果のレビューが必要
  ⇒ 米国では当日少なくとも翌日には完了
  - **2営業日以内**を遵守すべき（ボトルネックになる可能性大）
  - 審査者の人材確保が必要。本当に可能か?
  - レビューを大学に委託することも一案か?

2. 磁気データによる提供①

- **ガイドラインの改訂(最善、個人情報保護法に範囲内で)**
  - **科研費の審査済みなので、以下の文言を即刻削除**
  - 総務省 調査票情報の提供に関するガイドライン
    “ウ 利用する調査票情報の名称及び範囲
    統計調査の名称及び年次並びに調査票情報の名称、地域及び属性的範囲が利用目的から判断して、~~必要最小限となっており、不要と考えられるものが含まれていないことが必要である。~~”
    ⇒ すべての変数を一体として提供
    ⇒ 担当者による解釈の余地をなくせる(不確実性の解消)
    ⇒ 再審査に従事する人材を他に配置できる(資源の有効活用)

- **ガイドラインの再解釈 (次善策)**
  - 最小限の単位を変数ではなく統計表とする(例えば、H17年度人口動態調査 (死亡票)を単位として申請)
    ⇒そこに含まれる変数については審査で立ち入らない
    ⇒統計表に含まれる全ての変数を一体として提供できる

2. 磁気データによる提供②

- **現行のホワイトリスト⇒ブラックリスト方式への移行**
  - ホワイトリスト：申請した変数のみ利用可能
  - ブラックリスト：リストに記載されている変数以外は全て利用可能（発想の転換）⇒海外で多く採用
  - ブラックリスト⇒個人の特定可能性のある変数のみ
    (例: 出生票の誕生「時」、「分」、一方「日」は含めない)
  - これらの変数のみ申請理由書が必要、それ以外は全て申請時に**無条件で提供** (色塗り廃止)
  - ただし、ブラックリストに入れる**必要最低限**に絞るべき。
  - これらの変数も申請から、2週間をめどに提供。

    ⇒ほとんどの変数がリストに入ることを懸念
    ⇒実際に利用する研究者を選定作業に参加させるべき

2. 磁気データによる提供③

- **個人情報保護の罰則規定を制定**
  - データの利用契約に反したデータ利用や開示を行った場合⇒
  - 米国：$10,000未満の罰金または5年未満の刑事罰
  - 日本： 同様の規定はなし。データの提供禁止や違反者の公表にどとまる。

  ⇒ **技術**（個人が特定出来ないデータをいかにして提供するか）で縛るよりも **ルール**で縛る、という発想の転換が必要！

  - そもそも、①漏洩する、②本人が特定される、③悪用される、は別々の話。研究者へのデータ提供で、①があったとして、②や③が起こった例はあるのか？

その他：手続きに関して改善案

1. 審査の日数への**上限の導入**（⇒不確実性を減らす)
2. 一定の条件（科研費の利用）の下でのセーフハーバールール的な自己宣誓チェック方式などの簡易化
3. 申請書類は**各省庁で統一**
4. 統計ごとに利用可能な変数と定義を整理
   ⇒各年度にどの変数が存在し、どの変数がブラックリストに含まれるか、が一目でわかる**リストの作成**
5. 研究計画、アウトプットなどはすべて科研費の申請書類で代用　(⇒ 2度手間を省略）
6. 科研費取得が調査票情報申請の実質の条件となっている
   ⇒ 科研費が取れない人（海外在住の研究者や大学院生）も利用できるよう検討すべき
7. データ送付: （現状) DVD による物理的送付
   ⇒データ送付システムなどを利用 ⇒**安全性が向上**

オンサイト利用について

- 総務省はオンサイト利用を推奨

  ⇒「オンサイト利用ならば変数ごとの審査は必要ない」

- **しかし、オンサイト利用は問題点が多すぎる**
  1. 施設数、利用時間が限定 (監視の人件費)
  2. 同時利用できる人数が限定
  3. 自然災害や感染症(コロナ禍) に対し非常に脆弱
  4. 格納されているデータが限定されている
  5. 中間生成物持出しの審査に時間がかかりすぎる

  ⇒ 質の高い研究は実質的に不可能!

その他のデータ利用方法も同様

- **匿名化データ**
  - 個人の在住県に関する情報が得られない
  - 年齢が5歳刻みでしか開示されない
- **オーダーメイド集計**
  - あくまで集計値

⇒ 信頼性の高いEBPMの分析を行うためには、どの利用方法も**使い物にならない**

⇒ 選択肢を現状以上に増やすことに、限りある資源を費やすべきではない

結論

1. **リモートアクセスを期限を決めて(!)進める**
   - 審査専門職員を育成、雇用（再審査の人材を回す）
   - 分析内容ではなくあくまでルールに従っているかを審査
2. **磁気データ提供の手続きの簡素化**
   - ガイドラインの改訂（または、再解釈）
   - 変数のブラックリスト化
   - 個人情報保護の罰則規定の制定
3. **リモートアクセスが軌道に乗り次第、オンサイトは閉鎖** (⇒リソースの分散を避ける)
4. **公開できるデータは積極的に公開**

- **「EBPMを着実に推進する」**において、現在の統計法の運用は足かせ以外の何物でもない。
- 日本人研究者が日本の政府統計(調査票情報)の使用を避ける悲劇につながっている。
- 最後はEBMPをやる気、**覚悟**が本当に政府にあるか。

追加スライド

我が国の公的データの質について

公的統計の質や量について

- 個人情報保護の枠組みの中で、
  1. 既存の公的統計へのアクセスを改善
  2. 新規のデータを整備

  を進める必要あり
- この2つは分けて考える必要がある。
- 本日は、主に1のアクセスの話、以下は2の話。

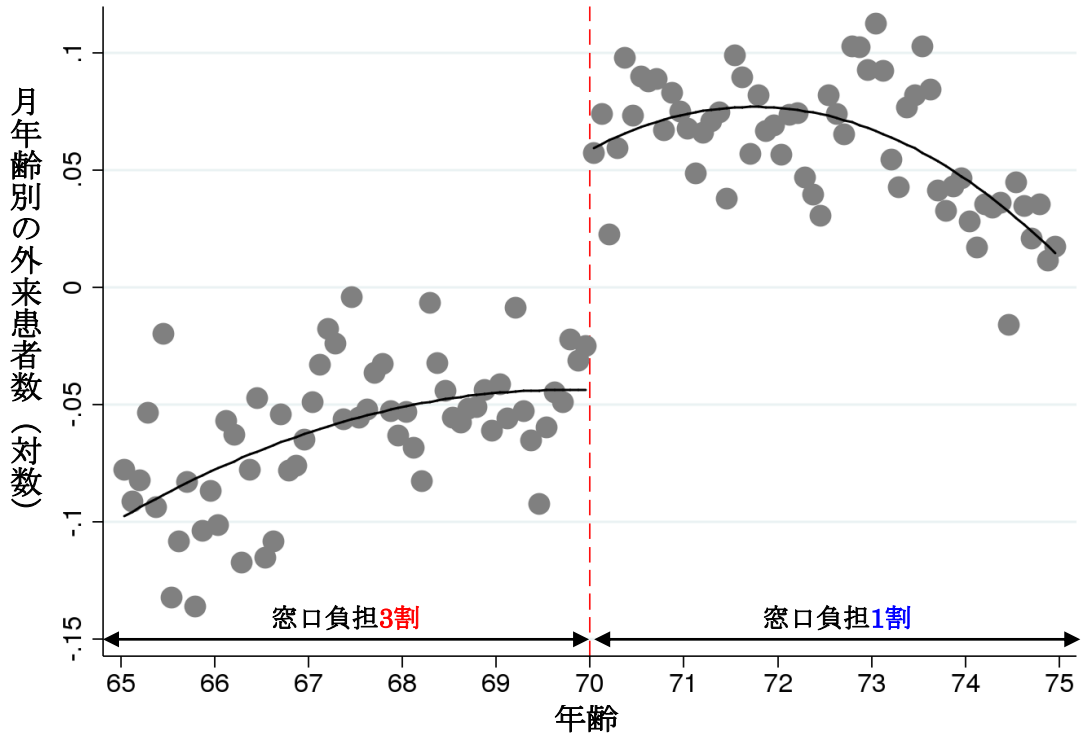
我が国の公的統計の現状

- ①質や量, ②アクセス 共に先進国で最低

- その結果、日本におけるEBPMの成功例は知る限りほぼ**皆無** (“証拠”に基づく政策???)
  - 例) 2022年10月より、75歳以上の後期高齢者の医療費窓口負担を1割から2割に引き上げ

    ⇒ 合計所得320万円以上に適用
    ⇒ では、320万は一体どこから???

例）窓口負担の分析



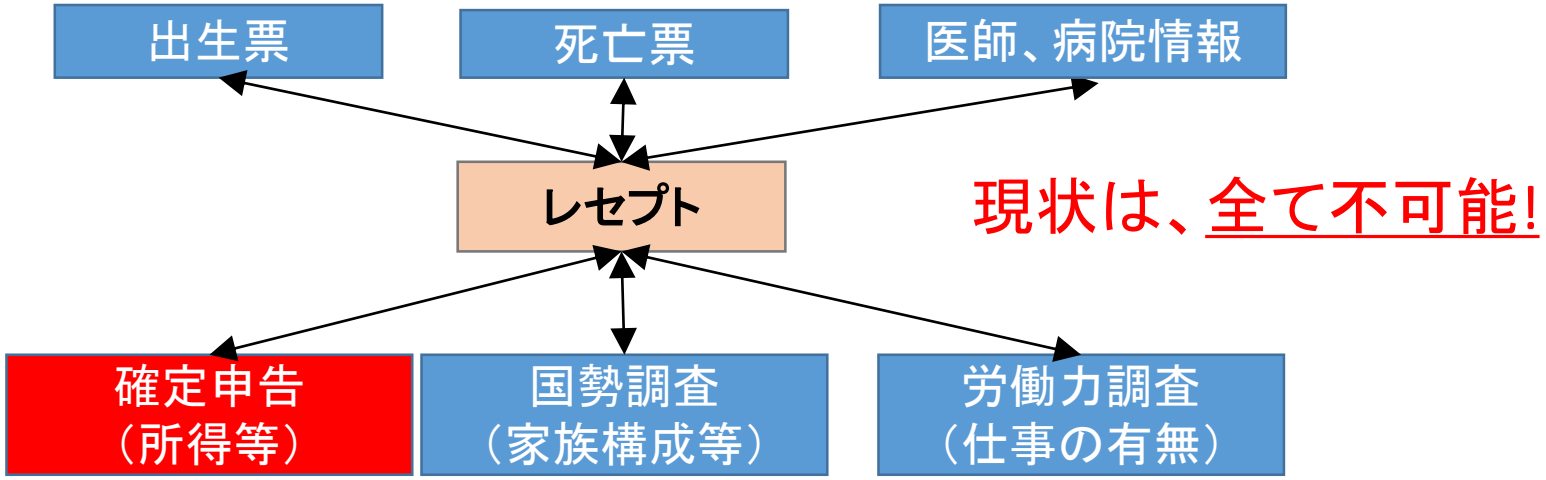
重岡 (2014)

- 政府統計の「患者調査」（レセプトデータ）を利用
- 70歳で医療サービス需要が劇的に増えるが、短期的な健康には影響がなかった
  ⇒ 窓口負担を上げても問題ない？

その結論は早計

- しかし、患者調査には所得の情報がない
  ⇒ 所得によって効果が違うのか、を分析できない
- もし、確定申告のデータと患者調査を結合できれば、所得ごとに窓口負担の効果を分析できる
  ⇒ どの所得で区切るべきなのか、の科学的根拠を提示できる

注) 以下のデータは全て、公的統計として既に存在



我が国の公的統計の現状

- ①質や量, ②アクセス 共に先進国で最低
  - このサイトで、世界中の国勢調査(センサス)がワンクリックでダウンロードできるが、日本は未だ参加せず。

Italy	National Institute of Statistics
Jamaica	Statistical Institute
Jordan	Department of Statistics
Kenya	National Bureau of Statistics
Republic of Korea	Statistics Korea
Kyrgyz Republic	National Statistical Committee

Japan?

  - 日本の出生データには変数が数10、米国では数百
    - 日本は最低限の基本情報、米国は出産方法(自然分娩vs.帝王切開)、母親のタバコや飲酒歴、保険支払額等、非常に細かい
  - 出生地(県レベルでさえ)を含むデータが皆無
    - 幼少期の環境が将来(収入、寿命等)に与える研究はほぼ皆無

新規データの整備

1. **複数の既存データを結合させる**
2. **新規のデータ収集**
3. 行政データの利用 (レジストリーデータ)
   - 既に存在するが、整備が進んでおらず使用されてない
   - 「税務データを中心とする自治体業務データの学術利用基盤整備と経済分析への活用」(学術変革領域研究B)
   - 「行政データ活用に向けて」→海外の事例、日本の現状のまとめ

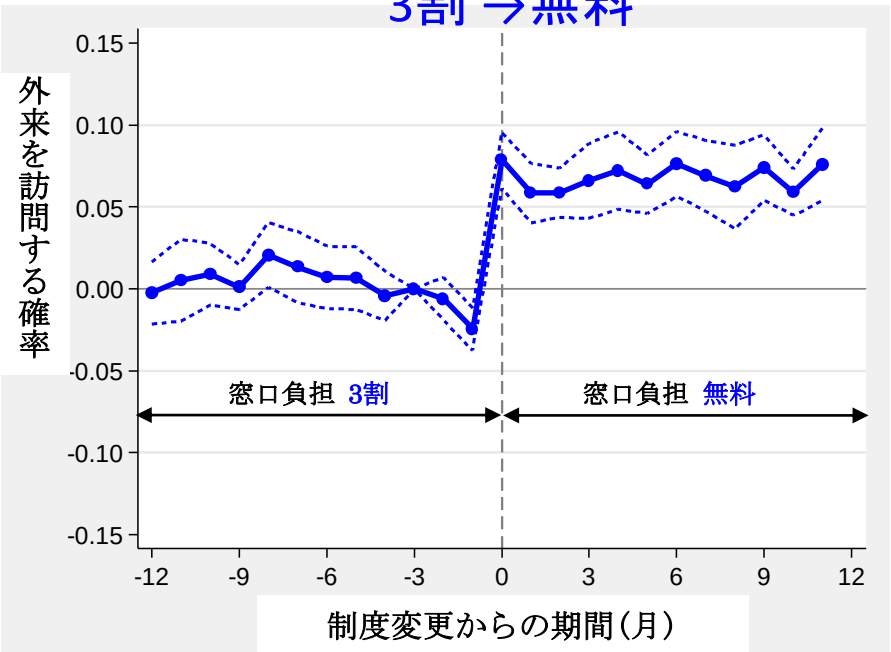
- **1, 2について医療の分野を例に**
  - 他の分野(労働、教育、犯罪等)でも全く同様
- そもそも集計値を公開しているデータに対して、個別データの提供を依頼するも断られることは日常茶飯事
  - 例) 労災のデータ → 厚労省に個別データの開示を求めたが、「対象外なので提供は不可」との返事のみ。

子供の窓口負担の分析

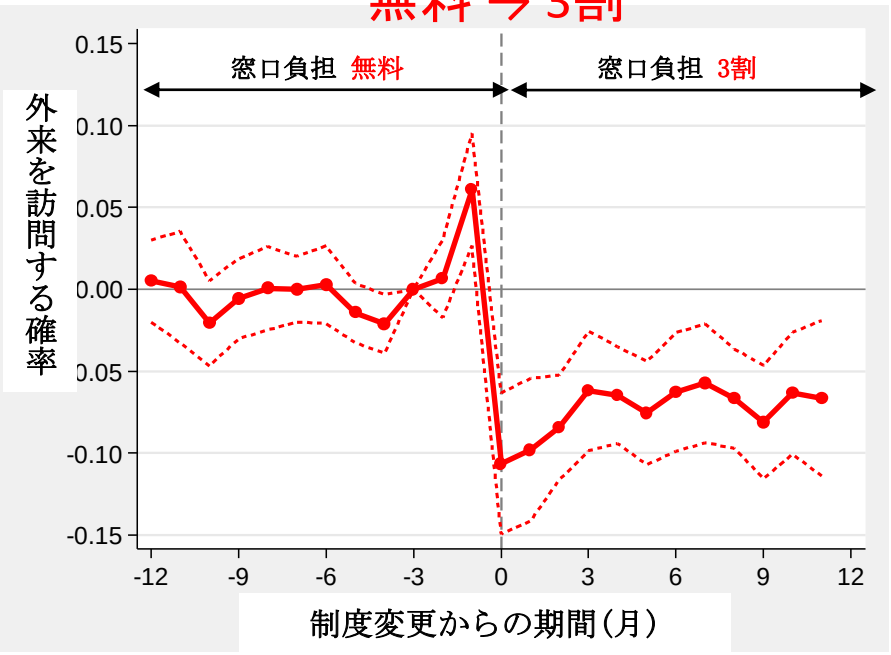
飯塚・重岡 (2022)

- 窓口負担は本来3割（6歳まで2割）
- ただし、多くの市町村で無料
  - 地域間競争、再選目的（エビデンスあり）

3割 →無料



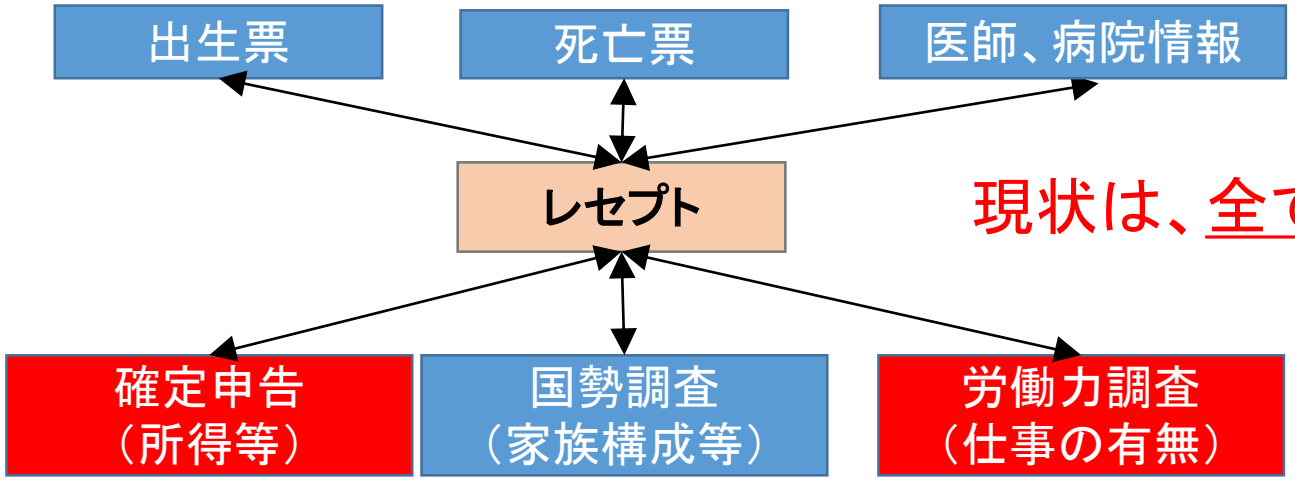
無料 → 3割



- 無料により劇的に医療需要が増える。健康に影響なし。
- わずか200円/回の窓口負担でも需要を大幅に抑えることが可能。
- 少なくとも我々のデータでは、過剰医療。年間1200億円が無駄。

問題点1: データが結合できない

- 現状は、レセプトデータが単独でしか使えない



現状は、全て不可能!

- 所得によりその効果が違うのか→「確定申告」と結合
- 共働きの家庭では効果が違うのか→「労働力調査」と結合

⇒ マイナンバーの活用（私の生存中には実現不可能!?）

⇒ マイナンバーの活用が倫理的に難しい場合、氏名、性別、住所、出生地、等の情報を使い、2つのデータを確率的に結合することは可能（世界中で行われているが、日本では前例なし）

  - 出生地の情報がないことは、致命傷

問題点2: データの整備が不十分

- 全国規模のデータの整備が不十分
  - 子供医療の分析の際に協会けんぽ、国保連合会にデータ提供を打診も断られる
  - 理由は、「提供のメリットがない」⇒政府の主導が必要
  - 最終的には、民間会社からレセプトデータを購入(ただし、企業の組合健保のレセプトなので低所得者を含まず)
- NDB (ナショナルデータベース)
  - 2013年より第三者提供開始。大きな進歩。
  - ただ、例えば、前述の子供医療の分析は不可能
    - 分析には、市町村IDと年齢(月齢)の情報が必要だが、
    - ⇒ 市町村のIDがない(2次医療圏レベルのみ)
    - ⇒ 年齢が5歳刻み
  - 手続きが非常に面倒かつ時間がかかる

特に医療データに欠如しているもの

①「所得」に関する情報

- レセプトデータ（患者調査、NDB）に所得情報がない。
- また、他の所得データ（例えば、確定申告）と結合もできない。

②「医療の質」に関する情報

- 医療政策の効果分析には、「質」に関する分析が必須
- 例えば、「死亡」は客観的なアウトカムであるが、現状はレセプトデータと死亡データ（人口動態調査の死亡票）を結合できない。
- また、出生データと死亡データに結合できない。
  ⇒低体重児の治療が乳幼児死亡の減少につながったか等の、先進国では出来て当然の分析ができない。
- 米国: 社会保障番号(SSN)を用いてレセプトデータと死亡データの結合が可能（北欧、チリ等の中所得国でも）

特に医療データに欠如しているもの

③「医療供給側」（つまり医師や病院）の情報

- 近年の研究: 地域ごとの医療費の違いは、「需要側」（患者）だけではなく、「供給側」（医師や病院）の影響が大きい。
  - 例えば、我々の子供医療の分析では、18%の子供が不必要な抗生物質を処方されていた。

- しかし、現状、医師情報（卒業大学、研修場所、勤務病院等）、病院の情報（財政状況等）がレセプトに結合できない。

- さらに、詳細な医師情報に関してはそもそも日本では集計されていない。

- 結果として、供給側への政策、例えば、「提供された医療の質に合わせた診療報酬の支払い」等、が不可能

⇒ 新規のデータ収集が必要
（医師会等の反対を抑える政治力が必要）