台第13769號聲請釋憲案說明會意見

王大為
中研院資訊所

自我介紹

- 中央研究院資訊科學研究所研究員
- 中央研究院資訊服務處處長(2010-2017)
- 台灣醫學資訊學會理事長(2011-2015)
- 研究專長: 隱私強化技術, 醫學資訊
- 主持
  - 衛生署八十八年下半年及八十九年度委託研究計畫「健保資料庫個人隱私保護機制研究」(DOH89-NH-032)
  - 衛福部「健保署一一零年度民眾資料自主權框架與醫療資源共享模式委辦案」(案號C1100039499)

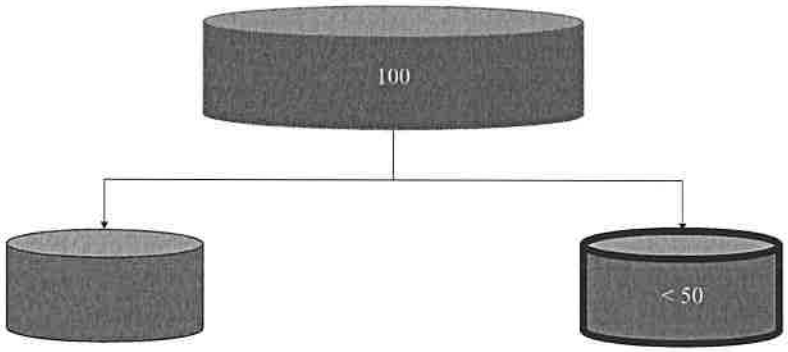
K-匿名化

- K-匿名化
  - 是一個資料集的性質
  - 每組個人資料都可以發現有K-1個人與其資料一模一樣
  - 用K-匿名化來做去識別化，意思是將一個資料集經過處理後可以產生一個K-匿名化的資料集
  - 刪除資料或將資料模糊化
- 隱私保護程度與資料效用(utility)之間的權衡(tradeoff)
  - 如果刪去的資料越多或是模糊的程度越大，那麼所產出的滿足K-匿名化的資料集的用處就越小
- 健保資料庫包含了許多細緻的個人醫療相關的資料
  - 詳細的資料可以透過資料分析產生重要洞見(insight)
  - 但這樣的資料幾乎不可能產生出符合K-匿名化的資料集且有不錯的效用
- 資料中心的模式:
  - 在分析的過程中所使用的資料集並不符合K-匿名化的要求
  - 而是在資料攜出時候對於攜出資料做了限制
  - 這樣的做法的確增加了隱私的保護，但是是否與K-匿名化有相同的效果，似乎不是簡單的論證

甕模型（Polya's urn model）

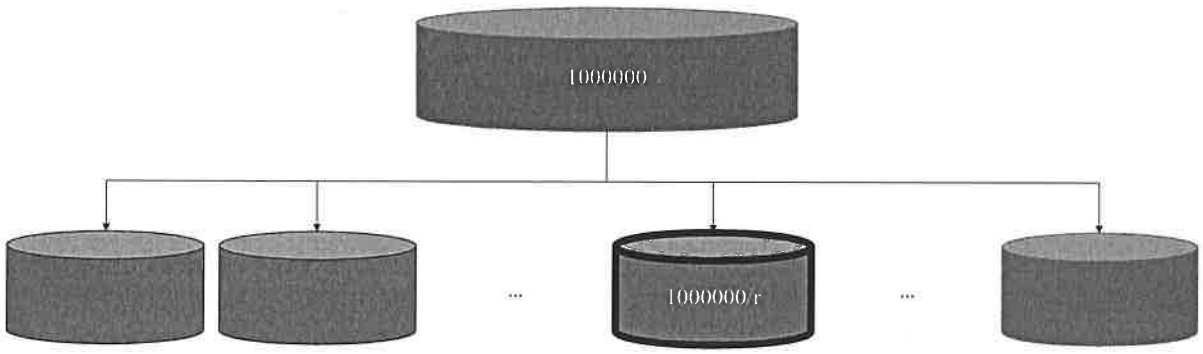
- 想像每個人的資料是一顆顆具有不同特性的球（大小、顏色等）。
- 想像把所有的球都放在一個**甕**裡。
  - 因為同一個**甕**中有太多球，分不出個別特性，所以隱私有保障。
- 現在我們把甕中的球進行分**類**，不同特性的球放在不同的甕中。
- 若是分**類**後的甕中只有一顆球，則代表我們可以根據分**類**的方式，分辨出這一顆球與眾不同之處。
  - 換言之，這個人的資料已被識別。
- 把**甕**中所有的球進行最仔細的分**類**，一定可以分辨出每一顆球。
- 若不進行最仔細的分類，分辨出一顆球有多難？

甕模型（Polya's urn model）



在有效分類的前題下，把盒子中n個球分2份，一定有一份的球數小於n/2

甕模型（Polya's urn model）

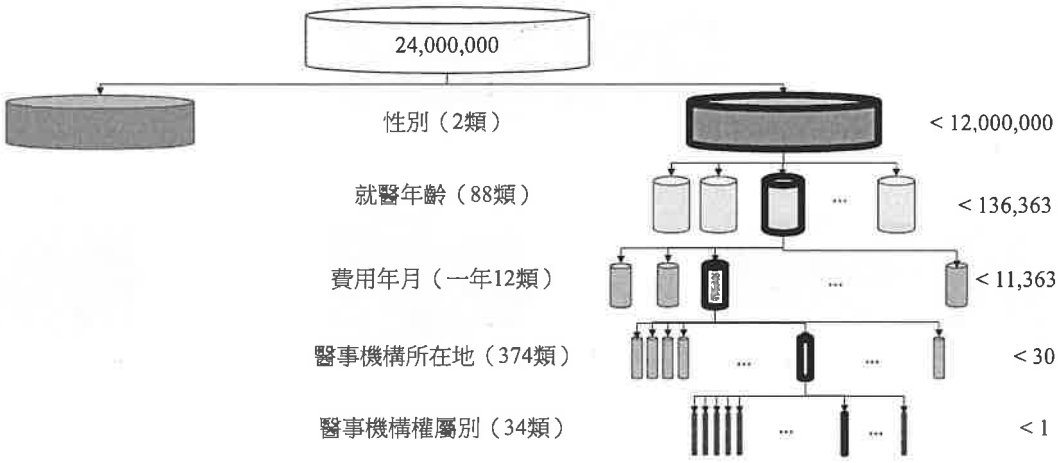


在有效分類的前提下，把盒子中n個球分r份，一定有一份的球數小於n/r

全民健保處方及治療明細檔（門急診）

- 常見欄位
  - 性別—1男性；2女性；9不詳
  - 就醫年齡—0至14天、15至28天、29天至未滿一歲、1歲、2歲、...、84歲、85歲以上（共88類）
  - 費用年月：過去一年（共12類）
  - 醫事機構所在地：0101臺北市松山區、0102臺北市大安區、...、9104連江縣東引鄉（共374類）
  - 醫事機構權屬別：1署立及直轄市立醫院、2縣市立醫院、4公立醫學院校附設醫院、...、41醫療社團法人診所（共34+45類）
    - 其中45類如：42醫療財團法人其他醫療機構、43醫療社團法人其他醫療機構、...、Z7私立營養諮詢機構

全國人口甕模型分析



去識別化的效果

- President's council of advisors on science and technology, PCAST Big Data and Privacy: A Technological Perspective, May 2014
  - 去識別化技術要成為隱私保護的基石會有困難
  - 很不幸地有些法律誤以為匿名化即可讓資料不再是個人資料而不在該法之範疇內
- Rocher, L., Hendrickx, J.M. & de Montjoye, YA. Estimating the success of re-identifications in incomplete datasets using generative models. Nature Communication 10, 3069 (2019)
  - 即使是抽樣的資料，只要資料夠多人數夠多仍然會有不可忽略的風險能夠再識別個人

個資風險分析範例

- 2011 加州的醫療照護組織Heritage Provider Network(NPH) 舉辦了一個健康資料分析競賽，目的是建構模型來預測病人明年住院的天數。所提供的資料為去前年去年與今年的申報資料(claim data)。資料集包括了十三萬三千名病人
- 分析結果發現約0.84%的人遇到八卦的鄰居可能會被識別出來
- 另一位學這做了稍微不同的假設,得出的結論為12.5%的人有被再識別的風險
- 小結：
  - 去識別化過程與其風險分析並沒有一致的標準
  - 以保密為由而使去識別化方法與風險分析無法由第三方檢視，是相當不合適的作法
  - 利用風險認定都有爭議的方法來論述這些資料無從識別個人因此不再視為個資，並不恰當

尊重自主權的可行技術

- 尊重個人自主權的機制可能行政成本太高?
- 利用提供選擇退出(option-out)的機制
  - 尊重個人自主
  - 社會在以資料為核心的競爭中保有競爭力
  - 選擇退出機制就是一種意願表達的機制，目前疫苗注射登記系統也有相同的功能
  - 去識別化技術可以做為增加的保障，而不需要勉強做為已經無法識別個人因此非個資的論述基礎的技術基礎

結論

- 去識別化與資料效用有權衡關係
- 大量且細緻的資料要做到可接受的去識別化程度並同時保有資料效用，幾乎不可能
- 全民健保資料
  - 強制保險
  - 如何進行去識別程度與效用的權衡，定非能由一單位在不透明的情形下單獨為之
  - 去識別的風險評估可因些許假設的差異而得到不同的結果，完成可受公評的評估所需的資源與人力，恐遠大於提供選擇退出機制來表達對個人自主權的尊重
- 以上為我的意見，也是建構民眾資料自主權框架與醫療資源共享模式思考的起點