101年健康資料加值應用分享研討會

資料品質提升與資料加值

1

吳肖琪
國立陽明大學衛生福利研究所教授
台灣公共衛生學會常務理事
台灣長期照護專業協會理事長

101年10月5日(五) 10:40-11:20
台北醫學大學 醫學綜合大樓後棟16F國際會議廳

誌謝

- 統計室
- 中國醫藥大學 鄭光甫教授之團隊
- 研究群洪燕妮、陳慧珊…等協助

政府建置健康資料加值應用的重要性

3

- **國民健康資訊建設計畫(National Health Informatics Project, NHIP)**
  - **建置衛生醫療資訊實證性決策、研究或研發**之基礎建設
  - 在**保護隱私**的前提下，妥善運用健康資料、創造加值應用
  - 將政府健康資料檔加值成**具應用價值之集體資訊**
  - **達到健康資訊共享**的目標
  - 促進公共衛生決策品質、相關學術研究及醫療保健服務業等相關產業研發創新之參據

建置NHIP前提

4

1. **資訊共用平台**
   - 健康資料加值應用協作平台

2. **安全議題是推展健康資料加值應用前須優先考慮的議題**
   - 資料的安全性（軟硬體及空間的安全性、加密保護個資、不予攜出小於五單位之結果）
   - 使用的申請審核
   - 資料的保密與運用規範（依政府資訊公開法、個人資料保護法及其相關法規辦理）

3. **資料品質**

NHIP資料品質之工作

5

- 任務
  - 資料品質提升
  - 降低進入門檻
- 作法
  - 建立檢核模式
  - 開發資料庫使用手冊
  - 以不刪資料(改設新變項)為原則

資料品質

6

1. 資料檔及欄位要如何命名？
2. 要優先處裡哪些資料檔的加值？
3. 資料的品質要考慮哪些？
4. 編製資料庫使用手冊

檔案的命名

7

資料庫命名影響到其管理及使用的便利性及正確性，資料檔來自各政府部門，且隨時代變遷可能會增刪，因此命名應考慮**未來性**

- **原則**
  - **一、兼顧「辨視性」與「擴充性」**
  - **二、以英文字母縮寫，分三層次編碼， 共計9碼，每一層以「下底線」(underline, _)隔開**，例：H_NHI_OPDME
- **編碼**
  - **主分類碼1碼，代表資料內容之主要涵蓋之領域**
    - **H**代表「**health** related database」(健康資料相關資料庫)
    - **S**代表「**social** and demographic database」(社會、人口相關資料庫)
    - **E**代表「**economy** related database」(經濟層面相關資料庫)
  - **次分類碼3碼，代表資料提供單位**
  - **細分類碼5碼，以檔案之英文名稱中取5碼縮寫，或國內慣用之通稱**(如：門診為OPD，住院為IPD等)**便於記憶為原則**(代碼末加年月)
    - 取代**DD**(住院醫療費用清單明細檔)、**DO**(住院醫療費用醫令清單明細檔)、**CD**(門診處方及治療明細檔)、**OO**(門診處方醫令明細檔)等檔名

欄位的命名

8

由於健康相關資料庫欄位龐雜，**訂定明確之欄位命名原則，能降低錯誤發生，增進溝通及使用資料庫之效率**

- **原則**

  一、**欄位名稱盡可能表達欄位內容**，避免使用var1, var2, item1, item2…等名稱

  例：承保檔

  二、**不同檔案之同一內容欄位** (如: ID、SEX、BIRTHDAY、CITY...)**欄位名稱應能辨識所屬檔案，以避免造成檔案串聯後資料覆蓋的問題**

  例：出生登記檔 – 胎別代碼 (DEL_NO)

  三、原來的欄名中之”_”(underline)予以刪除

檔案的優先性

9

序號	完成時間	第一優先檔案	檔案代碼
1-3	98年	全民健保處方及治療明細檔─西醫、中醫及牙醫門診	H_NHI_OPDME H_NHI_OPDDE H_NHI_OPDCE
4		全民健保處方及治療明細檔─住院	H_NHI_IPDTE
5		全民健保處方及治療明細檔─藥局	H_NHI_DRUGE
6-8		全民健保處方及治療醫令明細檔─西醫、牙醫及中醫門診	H_NHI_OPDMO H_NHI_OPDDO H_NHI_OPDCO
9		全民健保處方及治療醫令明細檔─住院	H_NHI_IPDTO
10		全民健保處方及治療醫令明細檔─藥局	H_NHI_DRUGO
11		全民健保承保檔	H_NHI_ENROL
12		全民健保重大傷病檔	H_NHI_CATAS
13		醫事機構基本資料檔	H_NHI_CONMF
14		戶籍資料檔	S_MOI_REGHH
15		原住民身分檔	S_CIP_ ABORI

序號	完成時間	第二優先檔案	檔案代碼
16	99年5月	癌症登記檔-CRS(68-95年,20欄位)	H_BHP_CRSSF
17	99年5月	癌症登記檔-TCDB(91-95年,65欄位)	H_BHP_CRSLF
20	99年8月	死因統計檔	H_OST_DEATH
21	99年8月	醫事機構現況檔	H_OST_RESMF
22	99年8月	醫事機構服務量檔	H_OST_UTIMF
18	99年11月	癌症登記檔-SF(96年以後,33欄位)	H_BHP_CRFSF
19	99年11月	癌症登記檔-LF(96年以後,95欄位)	H_BHP_CRFLF
23	99年11月	醫療院所評鑑等級	H_DOH_ACCMF
24	99年11月	出生通報檔	H_BHP_BIRTH
25	100年3月	出生登記檔	S_MOI_BIRTH
26	100年8月	醫事機構病床檔	H_NHI_BED
27	101年6月	全民健保處方及調劑明細檔– 特約醫事檢驗機構及特約醫事放射機構	H_NHI_LAB
28	101年6月	全民健保處方及調劑明細檔–特約物理(職能)治療所	H_NHI_THER

- **主管單位同意提供資料**
- **使用頻度**
  - 健保相關檔
  - 醫事機構檔
  - 戶籍檔
  - 原住民身分檔
  - 癌登
  - 出生檔
- **98年**：15個**第一優先檔案**
- **99-101年**：13個**第二優先檔案**

資料品質驗證

10

- 不同資料庫有**不同設立目的**及**品質良莠**的問題
- 進行資料加值應用前，必須先確認各健康相關資料庫的資料品質，**若無正確之資料，則加值運用會受到質疑與使用限制**
  - **Garbage in, garbage out**
- 品質驗證程序
  - 檢視各檔案欄位本身邏輯
  - 檢視填報品質(**一致性**、**正確性**及**完整性**)
  - 檢視同一檔案**跨年度**的一致性及變化情形
  - **跨欄位**、**跨檔案之**資料品質
  - 研擬改善建議(**提案單**)

如何驗證資料正確性 - 欄位基本資料

11

- **異常性**
  - 歷年(各月)資料筆數
  - 欄位數
  - 名稱(中文欄名與英文欄名)、型態、長度、順序
  - 鍵值(唯一化)

    例：**死亡檔 - 逐一比對死亡證明書更誤ID重號**
    - 死因檔唯一鍵值(Primary Key)，60-73年之「連續號碼(s_num)」及74-98年之「身分證字號(ID)」，經檢覈其唯一化，發現重號情形

    例：**出生登記檔 - 研議辨別外配身分之對策、以確認主要鍵值之正確性**
    - 新生兒身份證字號 (ID)、生父身份證字號(FA_ID)、生母身份證字號(MO_ID)為加值應用之重要鍵值，然**加密後ID有重複多筆**之情形，生父身分證字號檢誤 (FA_ID_ROC)之代碼「0」表身分證字號不符合編碼原則占3.70%-4.46%
    - 加密後生父ID與生母ID重複，為**部分鄉鎮以同樣代碼代替外配ID**
- **合理性、正確性**
  - Max、Min、Range、Mean、Mode
  - **譯碼類別**(例：內政部出生登記檔-同胎次序代碼)
  - **遺漏值**(資料庫使用手冊註記說明**遺漏比率%**)

如何驗證資料正確性 - 同檔跨欄位交叉驗證

12

資料庫品質問題僅由個別欄位分析無法真實呈現，須透過不同欄位交叉檢覈能進一步找出資料庫之問題

- 先針對個別檔案中可進行交叉檢覈之欄位分析

  例：癌登的性別，由癌登性別與癌登ID第二碼轉出之性別不符

  例：癌篩補助年齡合理性，癌篩就醫日期計算年齡與實際年齡不符(健保就醫生日落在補助範圍，但實際年齡尚未符合)。是否為自費？

  例：2009年出生登記檔 - 出生場所性質代碼(PLACE_CHA) vs. 接生者身份代碼(DELIVER)

  - **檢視合理性** (紅字所示)

如何驗證資料正確性 - 跨檔同欄位交叉驗證

13

透過資料庫連結來進行檢覈，分析不同來源同一變項的差異

- **鍵值**
  - 以**出生通報檔**之**產婦身份證字號(ID_M)**為例，比對**戶政檔**，**比對不到的比率**由**90年**的**11.5%逐年下降**至**98年**的**8%**
- **跨檔檢覈資料完整性**
  - 呼吸照護年度新個案應與**當年度戶籍檔**加值應用為佳
  - 呼吸照護年度新個案應與**前一年度承保檔**加值應用為佳
- **跨檔檢覈資料一致性與正確性**
  - **生日** - 擷取優先順序為**戶籍** > **重大傷病檔** > **承保檔** > **住院申報** > **門診申報**，依此遞補，期使資料之完整性與正確性最大化
  - **性別** - **一致性近100%**，建議仍優先以**戶籍檔**為主
  - **婚姻狀況** - 建議**以戶籍檔為主**

如何驗證資料正確性 - 跨檔跨欄位交叉驗證

由研究議題出發，檢視資料庫品質

- 驗證資料之一致性與合理性

以呼吸照護個案為例：

- 醫療機構服務量檔(H_OST_UTIMF)中與呼吸照護有關的欄位 vs. **醫療院所評鑑等級檔**(H_DOH_ACCMF)」之特約類別，分析91-99年之呼吸照護病床數(開放床、使用床)及呼吸治療師之變化情形
  - 各層級醫院之家數及**ICU(開放床)**病床數變化
  - 各層級醫院之家數及**RCW(開放床)**病床數變化
  - 各層級醫院之家數及**執業醫事人員–呼吸治療師**人數變化

醫療機構服務量檔之呼吸照護相關欄位

中文欄名	英文欄名	型態	長度
開放病床數-加護	o_bed07	Num	8
使用病床數-加護	u_bed07	Num	8
健保病床-加護	h_bed07	Num	8
開放病床數-呼吸照護病床	o_bed12	Num	8
開放病床數-呼吸照護中心	o_bed13	Num	8
使用病床數-呼吸照護病床	u_bed12	Num	8
使用病床數-呼吸照護中心	u_bed13	Num	8
健保病床-呼吸照護病床	h_bed12	Num	8
健保病床-呼吸照護中心	h_bed13	Num	8
住院人日-加護	dd_day07	Num	8
住院人日-呼吸照護病床	dd_day12	Num	8
住院人日-呼吸照護中心	dd_day13	Num	8
住院人次-加護	dd_times07	Num	8
住院人次-呼吸照護病床	dd_times12	Num	8
住院人次-呼吸照護中心	dd_times13	Num	8
執業醫事人員數-呼吸治療師	pra22	Num	8

資料庫使用手冊 – 提醒使用者避免誤用

15

- **瞭解資料庫蒐集用途目的、方式、來源，以免誤用**
- 健保申報資料庫原本僅為申請醫療給付之用，部分臨床相關訊息品質可能較差，故須進行一些細緻化的動作來提高準確度及其價值
  - **檔案屬性** 可提醒
    - 例：**重大傷病檔**中**有效迄日(VALID_E_DATE)**欄位於90-98年有2.7%~10.1%為空白
      若為空白則該筆資料屬無效，因為沒核准給付，故不會有迄日
      若重大傷病類之有效期限為永久(如hv_type=2,先天性凝血因子異常)會註記為" 29991231"
  - **寬鬆 vs. 嚴格個案定義標準** 無法提醒
    - 比較出現**一次診斷**與出現**至少三次診斷**兩種個案定義的分析結果差異
  - **透過交叉比對** 無法提醒
    - 例：定義精神病人，以臨床診斷(三次診斷)與**醫令(藥品處方)**來精緻化個案定義
    - 例：定義長期使用呼吸器病人，以醫令(呼吸器與P code)且領有**重卡**來確認長期依賴個案
    - 例：定義血友病病人，以領有重卡(hv_type=2,先天性凝血因子異常)與**用藥情形**來納入後天性血友病，因為這類病患不用領重卡，避免低估個案
- **變項使用說明**盡可能註記於**資料庫使用手冊**

加值應用

16

- Evidence-Based Health Policy 研究

政策與決策	• 使相關政策擬定與計畫評價有客觀數據依據 • 進行健康資訊與衛生決策相關學術研究
醫務管理	• 品質、資源配置、成效分析，與相關學術研究
流行病學	• 找出各種疾病盛行率、發生率、危險因子，與相關學術研究
臨床醫學/護理	• 疾病臨床表現、關聯性和最佳檢查方法、治療及照護模式(包括藥物、器材及手術)，作為臨床決策之參考或相關學術研究

Thanks for your attention