財團法人演譯基金會

美兆健康資源中心

美兆健康數據資料庫
資料檢誤與整理與流程概要

技術報告編號：MJHRF-TR-04

2016/03/18

引用本文參考文獻格式：
莊淵傑(2016)。**美兆健康數據資料庫資料檢誤與整理與流程概要**。財團法人演譯基金會技術報告(編號：MJHRF-TR-04)。

美兆健康數據資料庫資料檢誤與整理與流程概要

一、前言

美兆健康數據資料庫自1994年開始，已累積了超過20年的健康問卷與健檢數據資料，然而資料庫中所儲存的資料與資訊，在歷經長時間的資料蒐集、題目變革以及不可避免的資料記錄疏漏，難免會降低資料庫的品質，進而影響研究數據的分析。

為避免因資料品質導致研究上的誤差，美兆健康資源中心（以下簡稱本中心）進行一系列的資料檢誤與整理工作，除可達到降低誤差、提供高品質的數據資料，透過系統性的問卷題目整理，更能增進資料使用者的便利性。後續將說明本中心所進行的資料檢誤與整理之程序。

二、建立歸人基本資料檔

1. 美兆健檢的特色，是採取會員制度，同一位會員，可能每隔一至兩年，或一段時間，就會定期到美兆診所進行健康檢查，因此美兆資料庫所蒐集到的資料，會包含會員歷次健檢的資本資料、健檢數據與健康問卷內容，為了日後分析使用的方便，我們會將基本資料進行歸人處理，也就是每位會員只保留一筆資料；主要是透過歷次資料的交叉比對，確認個人基本資料（如：身分證字號、性別、出生年等），健利完整正確的會員資料檔，也可避免因會員某次健檢時的誤填或漏填等因素，而造成基本資料的缺失。

2. 身分證字號、姓名與出生年月日比對

   在整理基本資料中，會發現有相同的身分證字號（外籍人士則為出生年月日加上英文姓氏前三個字母），但姓名或出生年月日卻不一樣，主要原因是改名、資料輸入錯誤，或是受檢者非會員本人的狀況；在處理上，會依據受檢者所留的電話、地址、Email等資訊來逐一判斷是否為同一人，若為同一人，則更正其誤植的資訊（姓名或出生日期）；若判斷非同一人，則查詢所有會員資料檔中，是否有相同姓名與出生日期的受檢者，如果有，就更正錯誤的身分證字號，如果沒有，因無法判斷哪一位才是身分證字號的真正擁有者，因此個別編給其不同的基金會ID，但原本的身分證字號則不更動。

3. 身分證字號與性別比對

   由於中華民國的身分證字號第二碼，可識別個人的性別，因此可用來比

對受試者資料檔中的性別欄位；程序上會先判斷受檢者的身分證字號是否為中華民國的身分證字號，如果是的話，則會進一步比對性別欄位，當發現矛盾狀況時，會先檢視健檢數據中，特定性別的檢驗項目，例如：乳房檢查、攝護線檢查等，並據此更正性別欄位，如果相關的檢驗項目都無數據的話，則會參考問卷中，只有男性或女性該答的題目，來判斷受檢者的性別。

4. 基本資料的儲存與釋出

上述會員基本資料的儲存、處理與運用，皆依照相關資訊安全規定辦理（可參閱美兆健康資源中心資訊安全技術手冊），透過實體隔離、權限管理、稽核追蹤等方式，確保這些個人資料的安全。

而個人資料的運用，主要是作為內部的身分識別與確認，並不會全部都對外釋出，其中只有與分析相關的重要資訊：性別、年齡，會對外提供分析研究使用。

三、問卷資料整理

1. 不合理值確認

美兆健康問卷題目，全部都是封閉式問卷，每一題都有合理的回答範圍，例如：你的血型是(1)A；(2)B；(3)O；(4)AB；(5)不知道，答案只能出現1到5之中的數字，但可能因為資料鍵入的錯誤，或是資料庫轉換的問題等，會出現極少數的不合理值，處理上會將這些選項外的數值，更改為遺漏值。

2. 跳答相關題目確認

在美兆健康問卷中，部分題目有跳答設計，但因為採用受試者自填紙本問卷的方式，無法強制跳題，所以會有少數誤答不應作答題的狀況發生，茲舉例說明如下。

（1）題目：

〔抽菸習慣〕
87.您抽菸嗎？(選擇“不抽”者，請跳答第91題)
○ 不抽 ○ 不抽，但經常吸二手菸 ○ 以前抽，現已戒菸 ○ 偶爾抽 ○ 每天抽
88.您抽菸已抽幾年了？(戒菸者依過去情形回答)
○ 未滿一年 ○ 一年至三年 ○ 三年至五年 ○ 五年至十年 ○ 十年至二十年 ○ 二十年以上
89.以前抽但現已戒菸者，您戒菸已戒多久了？
○ 未滿一年 ○ 一年至三年 ○ 三年至十年 ○ 十年至二十年 ○ 二十年以上
90.您平均每天抽多少菸？ (戒菸者依過去情形回答) ○ 半包以內 ○ 半包至一包 ○ 一包以上

（2）檢查內容：

➢87題回答”不抽”者，需跳答91題，不需做答88至90題。

➢87題回答”已戒菸者”，才需回答89題。

受檢者忘了跳答，可能是沒有注意跳答提示，而接續做答，對於與自身狀況不符的問題，就以最接近的答案來填答，例如：不抽菸的人，在回答抽幾年時，就選擇”未滿一年”來做答，以勉強符合自身的經驗，但為了資料結構的一致性，我們還是將應跳答卻回答的數值更改為遺漏值。

3. 複選題內容檢查

在問卷中有複選題的設計，並通常會伴隨著互斥的選項，例如：

44.您長期持續服用的藥物有哪些？ (可複選)(平均每日服用一次以上的藥物稱之)
○ 無服用任何藥 ○ 鎮定劑或安眠藥 ○ 尿酸藥物 ○ 類固醇藥物 ○ 心臟病藥物 ○ 荷爾蒙
○ 高血壓藥物 ○ 止痛藥 ○ 糖尿病藥物 ○ 腸胃藥 ○ 甲狀腺藥物 ○ 中藥 ○ 高血脂藥物
○ 精神科藥物 ○ 氣喘藥物 ○ 自購成藥 ○ 其它

當選擇”無服用任何藥”時，就不應該選擇後續的任何一種藥物，當發現此答題矛盾時，就會以勾選的藥物為主，而將”無服用任何藥”的答案改為遺漏值。

4. 題目間邏輯檢查

有些題目雖然沒有跳答的提示，但題目之間確實有邏輯關係，這部份在資料處理時，也會特別檢查，以確認回答的一致性。例如：

(限女性作答)
37.是否曾經生育？○否 ○是
38.懷孕次數？ ○0次 ○1次 ○2次 ○3次 ○4次 ○5(含)以上
39.生產次數？ ○0次 ○1次 ○2次 ○3次 ○4次 ○5(含)以上
40.您幾歲時生產第一胎？ ○19歲以下 ○20-24歲 ○25-29歲 ○30-34歲 ○35歲以上 ○不適用

在以上四個題目間，存有絕對的邏輯關係，如果未曾生育，生產次數就是”0”，生產第一胎的年齡也應該回答”不適用”；相反的，如果曾經生育過，懷孕次數與生產次數就不應該回答”0次”，生產第一胎的年齡也不應該回答”不適用”；另外，懷孕次數也不應該少於生產次數；以上若有不一致的狀況發生，就會綜合這四題的回答，來修正矛盾的答案。

5. 歷年相似資料整理

美兆健康問卷資料的收集，已超過20年以上，這期間會有數次問卷內容的改版，可能是增減題目，或是調整選項，而這些更動，可能會造程資料使用者的不便，因此，先經過資料的整理與重新編碼，並製作過錄編碼簿，才能提供給使用者更友善與便利的檔案。

以婚姻狀況題目為例（見表一），此題目在1998年與2014年初都進行了選項的調整。在原始的資料檔中，婚姻狀況係儲存於同一個變項，因此同樣是回答第(3)個選項，但在1997年與2013年的意義是不一樣的，所以

我們將此變項依據選項更改年度，一分為三，讓資料使用者保有資料使用的彈性，同時並製作如下的過錄編碼簿，供使用者清楚瞭解每個變項之間的差異。

表一：歷次婚姻狀況問卷題目編碼簿

變項名稱	題目	選項	題目年度
marriage_96	婚姻狀況	【1996.02-1997】(1)未婚　(2)已婚 (3)再婚　(4)鰥或寡　(5)離婚　(6)分居	96-97
marriage_98	婚姻狀況	【1998-2013】(1)未婚　(2)有偶　(3)離婚 (4)喪偶	98-13
marriage_14	婚姻狀況	【2014-】(1)未婚　(2)已婚、再婚、同居 (3)離婚　(4)喪偶	14-

四、健檢資料整理

1. 不合理值確認

在美兆一百多項的健檢數據中，有少數的資料是經由人工輸入，例如：早期的身高、體重，以及腰圍等，這些變項中，難免有人為鍵入錯誤的情況發生，對於這類不合理值的出現（如身高超過250公分、成年人體重小於30公斤等），會依據該受試者歷次的相關健檢數據來進行修正，如果沒有相關數據可以參考，則將不合理值改為遺漏值。

2. 極端值處理

在整理健檢數據時，我們會列出所有變項的極端值（如血壓收縮壓超過200mmHg），並請檢驗科與護理部相關同仁一起檢視這些極端值，對於有疑慮的數值，會請診所同仁調出該筆健檢的詳細相關記錄，以確認資料的正確性，通常需更正數據的案例非常的少。

3. 性別相關項目檢查

健檢項目中，有些是女性專屬的檢查（如子宮頸抹片檢查），我們會與性別變項進行核對，這部份除了是驗證性別變項的正確性之外，有可以檢驗資料欄位是否有整批位移的情形。