健保資料與抽樣調查

余清祥　　簡于閔　　梁穎誼

摘要

人口資料是國家施政的基礎，大多數國家會以普查（Census）為全國性資料的主要蒐集方法，但近年戶口普查遭遇不少挑戰，許多國家考量不進行全國普查，改以其他方法取得資料品質可媲美普查的全國人口估計值。美國、法國的滾動式普查是方法之一，每年輪流抽取不同 3%的樣本，除了可取得不錯的資料品質，又能獲得五年連續調查結果，打破十年一次靜態人口的限制。

本文也以探討普查替代方法為目標，評估以健保就醫資料取得我國國民經常活動資料的可行性，比較幾種常見的常住地判斷方式，包括感冒就醫、以及本文提出的幾種方法，評估哪種方法較能反映國人就醫地區的特性，並討論未來以健保就醫作為取得常住人口時，必須考慮的配套措施及限制。另外，我們也將常住地判斷方法應於探討人口遷移，將立委選區訂為分析單位，以避免人數差異太大的問題，分析發現臺灣大致可分為幾個生活圈，其中有幾個地區的遷入及遷出相對活躍。

關鍵詞：常住人口、全民健康保險、大數據、普查、遷移

Abstract

Census usually is the only method for collecting the information of national population and it is conducted every 10 years for most countries. However, there are potential problems in the traditional census, such as low response rates, data quality, and rising survey costs. A lot of countries have been seeking alternative methods for collecting national data. Rolling census is one of the new methods and it was first adapted by the U.S. and France in 2010. The rolling census in the U.S., American Community Survey, collected 3% of national population annually for 5 years, and this can provide a time series national estimates for 5 years.

In this study, we aim to explore the new data collection methods which can serve as an alternative to the traditional census. In particular, our goal is to evaluate the possibility of using the data from National Health Insurance (NHI) Research Database for acquiring the information of de jure population in Taiwan. In addition to the records of upper respiratory tract infection, we also propose three other methods to identify the place of normal residence. We use the outpatient rates and other criteria from the NHI data for method evaluation. We also apply the proposed methods to explore the domestic migration in Taiwan. The results show that Taiwan can be separated into three sub-regions and the migration is active within each region, not as active between regions.

Keywords: De Jure Population, National Health Insurance, Big Data, Census, Migration

第一節 研究動機與目的

我國官方的人口統計方式大致分為常住人口與戶籍人口，兩者各有其特色及適用之處，常住人口透過戶口普查蒐集，戶籍人口則依賴戶籍登記。其中常住人口較能反映各地區的實際生活人數，可提供政策訂定、資源規劃的參考，但相關的統計工作較為棘手；戶籍人口則是法理上的各地人數，作為納稅、就學、兵役、選舉等的依據，戶籍資料的記錄及更新較為容易，但未必能反映經常活動於某地的人口。我國戶籍人口大多由民眾主動申報，出生、死亡、遷移等都是記錄項目，近年政府採取許多便民措施，像是當事人（包括戶長）可委託他人代為申報，但常住人口資料的取得維護較為困難，每隔十年一次的戶口普查只能提供靜態人口，無法紀錄兩次普查間的人口的動態變化，只能藉由問卷中的回溯性問項蒐集，很難確定資料的可信度（尤其是由戶長代為填寫普查問卷）。

由於大眾愈來愈注重個人隱私權，各國的戶口普查在二十世紀末以來遭遇不少挑戰，包括普查問卷回覆率低、資料品質不佳、調查成本上升等，許多國家嘗試不直接進行全國普查，改以其他方法取得資料品質可媲美普查的全國人口估計值。美國在2010年普查採用的美國社區調查（American Community Survey；顏貝珊、余清祥　2010），就是新型態普查的著名範例，我國在2010年採取抽樣代替普查原意也是基於類似原因，以大約16%的樣本推估全國人口的主要特徵。另外，即便是戶口普查也未必能取得準確及可靠的全國資料。以澳洲為例，普查調查對象限於普查當日停留該地點者，澳洲政府為求能取得接近真實的人口活動（常住人口）資料，除了普查結果外，會再以人口推估取得較準確之常住資料，並於發布資料後仍保留修訂之空間，確保資料時效性及資料品質（楊雅惠 2015）。

除了透過戶口普查，國內學者也嘗試不同作法取得常住人口的資訊，全民健康保險資料庫是其中的方法之一。全民健康保險（以下簡稱健保）制度實施至今

超過20年，幾乎所有國人都加入了全民健保，現今各鄉鎮市區至少有一個醫療院所，確保了國人可在地就醫（或是醫療可近性；Accessibility），由於健保已融為日常生活的一部分，國人就醫地點與經常活動地點有很高的吻合度。例如：學術界一般認為小病就醫會在經常活動地區附近，如蔡文正等人（2003）研究民眾對基層診所評價與就醫影響因素時，發現民眾生小病時有82.48%選擇診所（基層醫療院所）；吳依凡（2004）認為基層醫療院所設置廣泛，就醫可近性高，小病就醫地往往在常住地較近的醫療院所。因此有學者建議將小病就醫地視為常住地，如林民浩等人（2011）使用上呼吸道感染就醫地與投保地推估常住地，林敬昇（2016）則以上呼吸道感染就醫地推估常住地。

至今對小病尚無統一的定義，過去曾採用上呼吸道感染（俗稱感冒）、皮膚病、牙科等，藉由這些疾病的經常就醫地作為我國國民常住地的參考。以感冒就醫為例，國人每年平均將近約70%國人因感冒而就醫（資料來源：健保署），以普及率的角度來看還算合理。然而，因為各年齡層的就醫習性不同，感冒就醫的比例隨年齡下降，許多高齡人口甚至不到六成，與各年齡的整體就醫率不同，單以感冒就醫的樣本作為就醫地或常住地的判斷依據，可能會有以偏蓋全的疑慮，感冒就醫樣本的人口結構等特性未必和全體國人一致（亦即樣本代表性）。

我們也認為可藉由健保資料庫取得常住人口的資訊，根據國人的就醫習發掘經常就醫地、甚至常住地的判斷準則，探究國人就醫習慣及其變化，希冀可藉由感冒就醫紀錄評估作為經常就醫地的可行性。另外，感冒就醫率在成年族群較低，為避免樣本代表性的疑慮，本研究也探索加入其他疾病填補感冒就醫率較低的年齡層，減少就醫年齡層不均而衍生的資料歧異性，提升憑藉就醫紀錄即可推估常住人口的可能。當然，不同疾病的就醫地間未必有一致特性，如何比對不同疾病就醫地得出合理估計，將在以下各節詳細說明。

本文將於第二節整理相關文獻與介紹健保資料庫及其資料特性，包括資料偵錯及清理；第三節整理國人健保門診就醫的主要特性，透過探索性資料分析尋找適合作為經常就醫地的可能疾病，並評估各種準則的特性及使用限制；第四節應用本文方法於探討人口遷移（或移動），將立委選區（每個選區20～50萬人）訂為分析單位；第五節則提供本文方法的使用限制，討論未來如何繼續結合不同資料來源，包括普查、抽樣、其他資料庫等，提升我國人口資料的完整性。

第二節　文獻回顧與健保資料庫介紹

普查是許多國家獲取全國人口資料的唯一方法，為各國施政的重要依據。然而普查在每個時代遭遇不同挑戰，例如早期普查仰賴人力蒐集及紀錄資料，資料品質會受到操作方式不一致的影響，而資料在傳輸、合併、儲存過程也會有毀損之虞，這些問題因為電腦及網路科技進步而大幅改善。近年傳統普查遭遇新一波挑戰，包括個人資料及隱私權使得完訪率（拒訪率）日益降低（增加），資料插補（Imputation）也未必能處理遺漏值（Missing Values）或拒訪造成的資料空缺。另外，兩次普查間隔過長（十年），無法反映現代人快速變化的生活步調，而人力及費用也使得傳統普查愈發困難，臺灣 1990 年、2000 年兩次普查動輒運用九萬人以上訪員，無論人力招募、訪員訓練都是非常巨大的工程。

有鑑於這些傳統普查的限制，近年有不少國家採用新的普查方法，美國及法國的滾動式普查（Rolling Census）就是其中的範例，以每年全國 3%的抽樣調查，連續五年取得全國 15%的樣本，如此可降低成本及取得較為即時的資料（顏貝珊、余清祥，2010）。我國也於 2010 年以抽樣調查代替傳統普查，透過全國 16%的樣本一窺全國母體的特性，但這並未解決拒訪率及即時資訊的問題，而且主計總

處也未公開抽樣設計，無法確定 16%樣本能夠反映全臺灣人口的特性，學術界對於 2010 年常住人口有諸多討論。[1]

首先對官方普查資料的疑慮，歷年有不少相關研究，例如：洪永泰（1995）發現戶籍人口與常住人口在以「人」為單位時，戶籍地與常住地吻合之比例不超過九成，若以「戶」為單位，其比例甚至會更低。陳肇男與劉克智（2002）認為臺灣的戶籍人口與常住人口差異日漸明顯，並指出在都市化程度不高時，戶籍確實能夠反映臺灣人口分佈，但伴隨著時代進步的人口流動會改變結構，使戶籍與實際情形產生差異。顏貝珊、余清祥（2010）認為近年如遷移之類的社會結構改變更加頻繁，使得民眾於普查上的配合意願越來越低，而 2010 年的抽樣調查代替全面性普查也造成調查品質的下降。

隨著時空環境變遷，人口流動因為交通便利及生活安排更為頻繁，戶籍所在不見得是經常活動地，傳統戶口普查的資料品質、資料豐富度又頗多限制，近年學者提出不少替代方案，以取得較為可靠的常住人口估計值。其中全民健康保險制度（以下簡稱健保）是臺灣引以為傲的社會保險，實施至今二十餘年，國人已非常熟悉健保相關規定，就醫時多半會個人需求及居住附近的特色選擇醫療院所，就醫地與常住地間的關聯性頗高。為了學術研究的需要，衛福部釋出不同主題的健保抽樣資料，提供各界提出需求申請計畫、審核後付費使用，我國除了健保服務及民眾民意程度傲視全球，根據健保資料衍生出之學術產出也聞名國際，研究成果可作為國家訂定政策時的參考，是不可或缺的研究資源。

正因為健保資料庫內容的多樣性、涵蓋率高，許多學者利用健保資料推估常住地。[2]廖建彰等人(2006)探討 2000 年臺灣腦中風發生率與盛行率的城鄉差異，

[1] 2010 年戶口普查與 1966 年以來的普查有不少差異，除了抽樣代替普查外，2010 年因為縣市長選舉延至 12 月 26 日開始，與先前 12 月 16 日的普查標準日不同。

採用投保地作為投保人的常住地，然而企業組織納保時通常以總部為投保地，可能會扭曲常住人口（或戶籍人口）的數值，像是臺北市投保人口接近 500 萬人，但戶籍人口約 260 萬人。林民浩等人（2011）認為全民健保高納保率（99%）涵蓋幾乎所有人，而且多數臺灣人小病就醫傾向選擇距離較近的醫療院所，他們以上呼吸道感染（或感冒）的就醫地作為患者的常住地，發現無論是以上呼吸道感染就醫地的推估結果，與普查的相關性皆較以投保地作為推估準則高。不過，因為各地區特性不同，他們認為若能再結合被保險人身分、投保類別、年齡，在不同都市化程度的鄉鎮與不同年齡層會有更好的表現。

然而，結合愈多資料庫、衍生的問題也愈多。例如：遺漏值、離群值之類的發生可能性更高，需要依賴插補的機會也愈多，但對於資料品質的衝擊也愈大。另外，上呼吸道感染就醫的比例和性別、年齡、地區有關，年齡愈高、就醫率反而愈低，每年高高齡人口（85歲以上）的上呼吸道感染就醫率不到五成。這也是引發本文的研究動機之一，希望能找到精進健保紀錄作為常住人口的可能作法，我們將在第三節詳細說明本文方法。本節以下將大略介紹健保資料庫來龍去脈，以及本文將使用的抽樣資料庫及其基本特性。

全民健保與1995年開辦，1998年中央健康保險局（現為中央健康保險署）委託國家衛生研究院（簡稱國衛院）建立健保資料庫，並於2000年開放各界申請使用。健保資料庫依照母體大小分為兩種類型：抽樣檔與普查檔，若母體較大則提供抽樣檔（如全國人口），若母體較小則提供全部母體資料（如重大傷病人口），開放的資料庫名稱及其內容，包括檔案資料欄位名稱和資料描述，可查閱國衛院「譯碼簿」。[3]本研究主要依據2005年承保抽樣歸人檔（LHID2005），以2005年

[2] 除了健保資料外，也有藉由結合其他公務統計推估常住人口，如陳豔秋、楊雅惠（2017）藉由連結戶籍、學籍、綜合所得稅、健保及醫療等公務資料，推估臺灣未來的常住人口。

[3] 譯碼簿，資料來源：國家衛生院全民健康保險研究資料庫，http://nhird.nhri.org.tw/date_02.html (2019/03/01)

承保資料檔中「2005年在保者」隨機抽取100萬人，擷取其2005年之後每年就醫資料。

由於健保資料庫的資料數量多且內容複雜，輸入錯誤等問題十分常見，因此本研究使用的資料庫已經過除錯、正規化等處理，以確保資料的正確性。本研究需透過患者的就醫紀錄探討就醫特性，以門診處方及治療明細檔（簡稱CD檔）串連承保資料檔（簡稱ID檔），取得兩性、年齡別的各疾病就醫紀錄，同時以CD檔中就醫紀錄的醫事機構代號串連醫事機構基本資料檔（簡稱HOSB檔），取得就醫的醫療院所看診地點。以下大略說明本文使用的健保資料庫檔案內容：

1. 承保資料檔：ID（2005年~2012年）

   紀錄ID、出生年月、性別、加保日期、加保類別、退保日期、退保類別等。

2. 門診處方及治療明細檔：CD（2005年~2012年）

   紀錄ID、ICD-9（國際疾病分類碼第九版；International Classification of Diseases, Ninth Revision）、性別、醫事機構代號、出生日月、就醫日期、金額等。

3. 醫事機構基本資料檔：HOSB（2005年~2012年）

   記錄醫事機構代號、縣市區碼、特約類別、型態別等欄位。

表1則為本文分析的資料量筆數與資料容量，ID為抽樣資料庫中的涵蓋人數，人數因為死亡等因素逐年下降；CD為每位國民各年度的門診紀錄，平均每年大約十餘次；HOSB為醫療院所的紀錄，可以看出臺灣醫療院所的家數逐年上升，可以預期醫療供給也在增加中。上述三個資料庫的資料容量總和超過100GB，已經達到大數據分析的層級，一般資料庫軟體已無法負擔，例如：微軟公司的Access資料庫只能處理至多5GB的資料。有鑑於此，本研究使用SQL資料庫軟體，搭配統計分析軟體R，整理分析健保就醫的相關訊息。

表1、各年度ID、CD、HOSB資料筆數與記憶容量（MB）

檔案	ID		CD		HOSB	
年份	筆數	MB	筆數	MB	筆數	MB
2005	1011467	271.016	15036998	9036.664	42582	28.031
2006	1004477	270.383	14275312	8578.914	44802	29.430
2007	997485	267.555	14324281	8608.344	46397	30.508
2008	996216	267.844	14213998	8542.070	48024	31.586
2009	991502	267.164	14668611	8815.273	49811	32.711
2010	989188	258.930	14598852	8773.352	51587	33.922
2011	988298	258.695	14978654	9001.602	53564	35.242
2012	987507	258.492	14973208	12997.578	55389	36.406

使用健保資料庫推論全國人口時，首先我們必須確保納保人口具有代表性。全民健康保險制度實施至今已逾20年，納保率初期僅有59%。經過政府及全民各界努力，幾乎所有國人都已納保。依照衛生福利部的統計，國人納保率在2010年已高達99.4%，[4]而健保署的報告也顯示2017年仍舊維持在99.6%（全民健康保險年報，2017-2018 Annual Report），[5]由此可見健保納保人口與全臺灣人口間的差異不大。由於國人對全民健保滿意度很高，我國醫療品質聞名世界，加上我國各鄉鎮市區皆有醫療院所，且醫療院所與健保署的特約率亦高達93%，因此可從健保資料庫推估出良好的常住地資訊。

以樣本推估全人口資訊時，需先確定樣本與母體的結構接近，亦即確認樣本代表性，本文以健保資料庫的抽樣檔判斷常住人口，同樣也需檢查樣本代表性。

[4] https://www.gender.ey.gov.tw/gecdb/Stat_Statistics_Field.aspx (2019/03/01)
[5] https://www1.nhi.gov.tw/Nhi_E-LibraryPubWeb/CustomPage/P_Detail.aspx?CP_ID=207 (2019/03/01)

本文樣本代表性的檢查分為三個層次：納保人口與百萬抽樣檔、所有就醫人口與百萬抽樣檔、全體抽樣檔就醫與各疾病就醫（例如：上呼吸道感染人口）。其中第一項檢查在於驗證健保抽樣檔與全臺灣母體類似，確定可由健保抽樣檔這些樣本推估全臺灣的特性，檢查項目為人口結構。第二項檢查補充第一項的不足，因為人口結構類似也無法保證可由樣本推論母體特性，因為健保抽樣檔紀錄國人就醫行為，並由這些紀錄推敲國人的常住地，與官方的醫療公務統計比對後，才能確定健保抽樣檔的資料是否足以反映全臺灣人口的就醫行為。第三項檢查主旨在於以部分抽樣檔的資料代表全體抽樣檔樣本，像是驗證健保抽樣檔的子樣本（部分集合）是否可代表全體樣本，這部分檢查將在下一節討論。

第一項檢查可由官方網站取得分析結果。百萬抽樣檔由全臺灣母體依據簡單隨機抽樣取得，其中納保人口與百萬抽樣檔的比較，由健保署檢查性別、年齡、平均投保金額等因子，分析發現樣本與母體差異很小，在顯著水準0.05之下，不拒絕兩者相同之虛無假設，確定樣本代表性無虞。[6]換言之，百萬抽樣檔的人口結構與全臺灣人口結構接近。

由於國人的就醫習慣不盡相同，有些人就醫頻繁、有些人不選擇不看病，除了人口結構外，本文也檢查有就醫紀錄者的人口結構，確定百萬抽樣檔的就醫者與全人口就醫者兩者特性接近，其中全人口就醫率資料來自於衛福部，在此僅以2005年為例說明。在性別比例上，百萬抽樣及全人口的就醫紀錄中，男性分別出現了49.5%和48.3%的比例，抽樣的男性比例高了約1%左右。年齡結構的比較可參考圖1，百萬抽樣檔中幼齡人口（0～14歲）就醫率比全人口同年齡層比例少了約0.8%，抽樣檔的25到49歲人口則少了約1.3%，其他年齡層的人口結構則較為接近。由於抽樣檔人數相當多，亦即百萬樣本與母體稍有差異，就無法通過卡方

[6] 抽樣與樣本代表性的資訊可參考網頁 https://nhird.nhri.org.tw/date_cohort.html (2019/03/01)

樣本代表性檢定（p-value小於0.001），即便如此，上述這些差異在實務考量上仍不算太大。

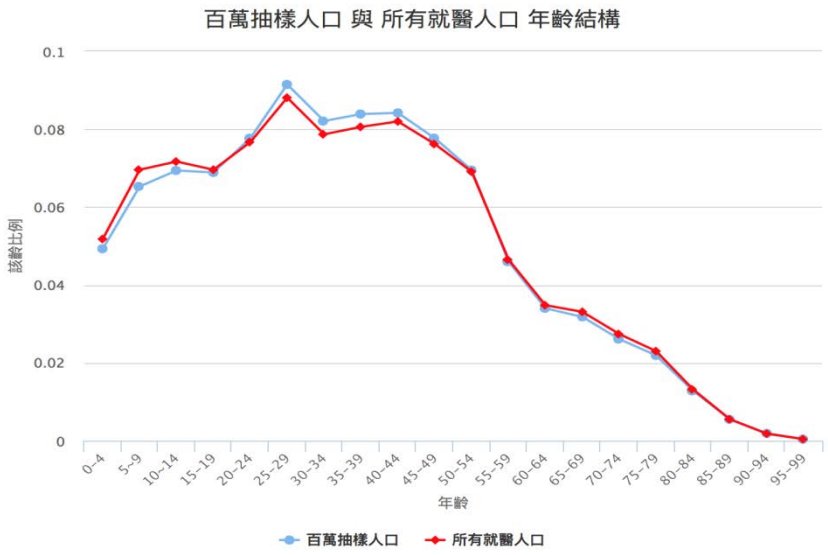

圖1、2005年百萬抽樣檔與所有人口的就醫者之年齡結構

第三節　健保就醫紀錄與常住人口

健康保險資料庫中有兩種與地點有關的紀錄：投保地與就醫地，兩者皆可追溯至縣市、鄉鎮市區等層級。投保地與投保類別有關，以人數最多的第一類投保人口為例，投保地為所屬機關、學校、公司、團體或雇主的所在地，總公司設籍於臺北市的比例最高，因此投保人口的高估總數最多，而且無法獲知眷口（依附於主被保險人）的資訊。就醫地則為門診醫療院所的所在地，與民眾的就醫習慣有關，而且醫療資源較為豐富的縣市也會有高估的疑慮，雖然臺灣各鄉鎮市區至少都有一個醫療院所，但對於癌症等重大傷病、或是手術等多半會選擇醫學中心（或教學醫院）、區域醫院等設備較為完善的醫療院所。為了避免這些問題，過去研究如林民浩等人（2011）選擇上呼吸道感染等小病，民眾可就近到鄰近醫療院所，不需專程至都會地區等醫療資源較為豐沛的地區；李虹映等人（2014）也

使用類似的方式，運用基層醫療使用率，以投保地區做為居住地的參考標準。上述這些研究都運用了全民健保的醫療可近性，認為民眾傾向選擇鄰近的醫療院所就診，因此健保就醫紀錄可用以推估國人經常活動的地區，彌補兩次戶口普查間十年的常住人口資料空窗期。

過去研究之所以採用上呼吸道感染就醫紀錄推估常住地，是因為這類疾病普遍較不嚴重，一般醫療院所都有足夠能力處置，不像癌症之類的重大傷病需要更完善的設備與醫護人力。由於我國醫療服務品質舉世聞名，加上各鄉鎮市區均設有至少一個以上的醫療院所，以往研究大多認為民眾罹患小病時，多半選擇在經常活動地看診，這也是感冒就醫地被視為常住地的原因。我們可延伸上述這種小病的概念，定義其他類似經常就醫地的判斷方式，像是基層醫療院所的門診量比例、門診費用、跨區就醫（或轉診）比例等。然而必須注意的地方是，以上作法需要有穩定且數量充足的就醫紀錄，才能保證樣本的解析度，從就醫人口可推敲出整個母體，換言之，上述小病定義的就醫率不應太低，避免沒有出現就醫紀錄的人數過多，亦即出現過多遺漏值，無法確定以就醫樣本推論全體樣本是否合理。本文將以三個因素作為評斷某種小病的定義，實務上是否可用於推估常住地：就醫率、基層醫療院所的門診量比例、門診費用。

首先，民眾是否因為某種疾病而就醫，依據CD檔中的門診主診斷碼（ICD-9），根據本研究認為可行之疾病做為分析標的，像是上呼吸道感染之就醫紀錄，[7]或是消化系統疾病等之就醫紀錄；[8]而就醫率除了整體民眾的就醫比例不應太低外，也會考量各年齡層的比例相差不大（類似樣本代表性的檢查）。基層醫療院所就醫的判定需要連結CD檔與HOSB檔，CD檔的就醫紀錄中有門診的醫療院所之紀錄，也就是醫事機構代號（HOSP_ID）欄位與HOSB檔串連，取得每位患者門診

[7] 上呼吸道感染的疾病代碼欄位 ICD9 內容為 460-466、480-487。
[8] 消化系統疾病代碼欄位 ICD9 內容為 520-579。

就醫地之所在地（鄉鎮市區、縣市）。如有多次就醫紀錄或就醫地次數相同者，本文建議採用機器學習（Machine Learning）常見的投票（Voting）多數決，若剛好有兩個或多個就醫地出現次數相同，則以就診日期最晚者為病患的就醫常住地。

除了定義某些疾病為標準，小病另一種可能解釋是醫療費用較低者，這也符合社會的共識。其中醫療費用可藉由CD檔健保補助的合計點數（T_AMT），[9]除了每次門診的平均花費金額之外，單次就醫金額的百分位數或是金額的累積分布函數（CDF, Cumulative Distribution Function）也可用於判斷小病。例如：如果就醫金額不超過500點可算是小病，則先估算500點等於就醫金額的多少百分位數，再以此評估500點是否可行。另外，我們也考慮醫事機構代號（HOSP_ID）欄位與HOSB檔串連，加入該醫事機構之特約類別（HOSP_CONT_TYPE）等資訊，評估就醫金額的上限與範圍，以確保就醫金額的資料品質。

本文先檢視抽樣檔之上呼吸道感染與所有人口的就醫率，同樣以2005年的結果為範例。抽樣檔上呼吸道感染與所有就醫人口中，分別有48.3%與46.6%的男性，抽樣檔的上呼吸道感染者多了約1.7%。年齡結構的差異則更大（參考圖2），上呼吸道感染就醫在19歲以下的比例較高，其他年齡層則是全人口較多，年齡結構的差異明顯高於圖2，這些差異當然也無法通過卡方樣本性檢定。進一步檢視抽樣檔的上呼吸道感染之就醫率，各年度整體就醫率介於65%~70%，每年大約有1/3人口不會因為感冒就醫，亦即若以感冒就醫地視為常住地，需要確定感冒就醫的2/3人口能否反映全人口的特性。

[9] 健保資料中的「點數」並不等於新台幣金額，與時間、醫療院所諸多因素有關，健保署會根據醫療單位紀錄、檢查項目必要性等作為核定標準，通常每個點數的核定金額介於 0.8~0.9 元。

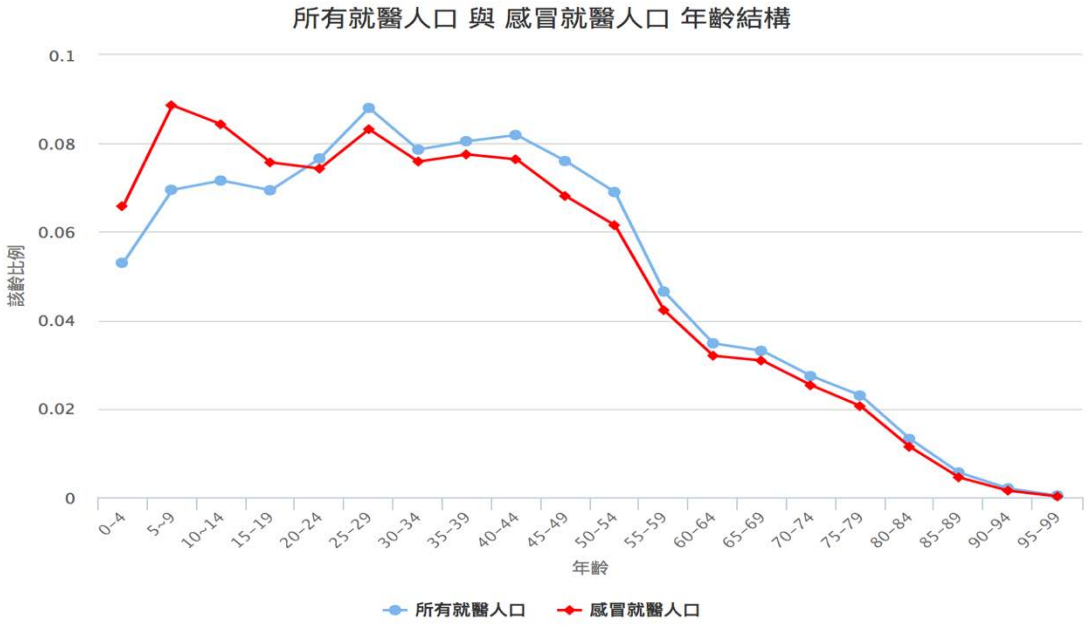

圖2、2005年抽樣檔上呼吸道感染、全人口就醫之年齡結構

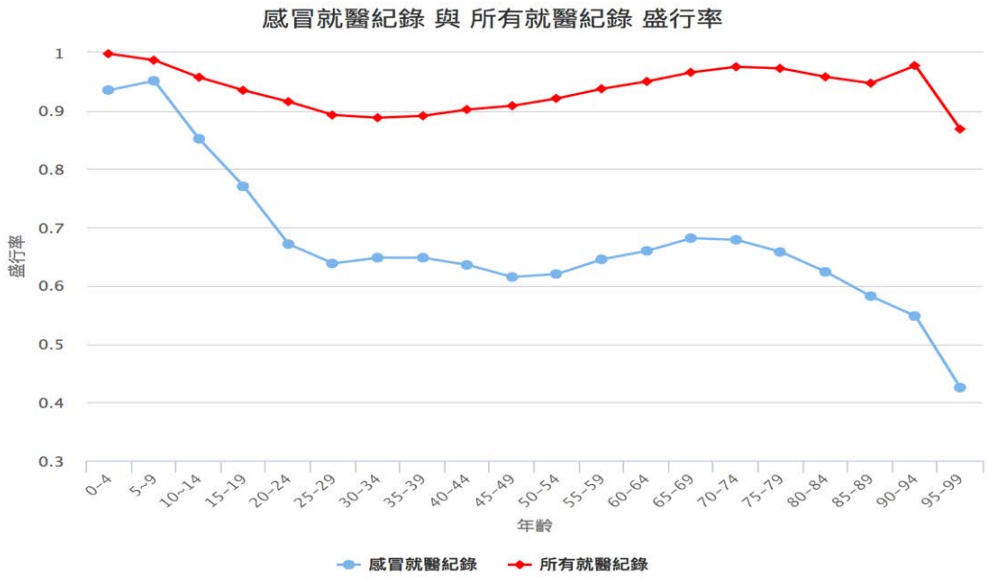

圖3、2005年抽樣檔上呼吸道感染、全人口之就醫率

接著再以年齡結構檢視感冒就醫的2/3人口（圖3）。明顯可見上呼吸道感染人口整體就醫率較低，僅有幼齡人口（0～9歲）的感冒就醫率與全人口就醫率相

當，兩者差異隨年齡越來越大；比較有趣的現象是健保抽樣檔85歲以上的感冒就醫率，其數值隨年齡而下降，與高齡人口較常看病的印象不同，值得未來繼續深入探討。整體而言，雖然上呼吸道感染的人口結構與全人口就醫類似，但其就醫率卻與年齡有非常高的關聯，尤其是高齡人口的感冒就醫相對較低，在缺乏更進一步資訊之下，以感冒就醫地作為常住地確實存在疑慮。

抽樣檔中20歲以上因為上呼吸道感染而就醫的比例太低，直接的解決方法為加入感冒以外疾病，類似增補樣本的想法，藉由新增疾病提升各年齡就醫率，尤其是感冒就醫率偏低的高齡人口，如何選擇新增疾病，還是以小病為主要考量。因此在填補樣本時，除了注意填補樣本後提升了多少就醫率之外，也要注意是否將較嚴重的疾病併入考量，避免誤判真正的常住地。

表2、2005年四個醫療院所層級的門診比例

	基層院所	地區醫院	區域醫院	醫學中心
上呼吸道感染	0.902	0.047	0.032	0.018
所有疾病	0.717	0.098	0.104	0.080

註：因為四捨五入，四個層級醫療院所的比例總和為0.999。

接著繼續以基層醫療院所的門診量比例、門診費用兩個因素，檢視上呼吸道感染是否符合小病的預期。[10]表2為2005年上呼吸道感染、所有疾病在四個層級醫療院所的門診比例，感冒有九成左右的門診都在基層醫療院所就醫，僅約5%門診選擇醫學中心（及教學醫院）、區域醫院；反觀所有疾病的門診在基層醫療院所接近72%，比感冒少了約18%。換言之，感冒大多選擇在基層醫療院所就診，

[10] 轉診制度自 2005 年 7 月 15 日後從資料庫欄位才有紀錄，加上門診的轉診比例低於 1%，而且只有註明轉入（沒有轉出）的醫療院所，因此本文不將轉診列入「小病」的討論。

由於基層醫療院所總數在2005年超過一萬八千家，其密度幾乎是所有便利商店（約一萬餘家）的兩倍，[11]推測因為感冒而跨區就醫的動機並不高。

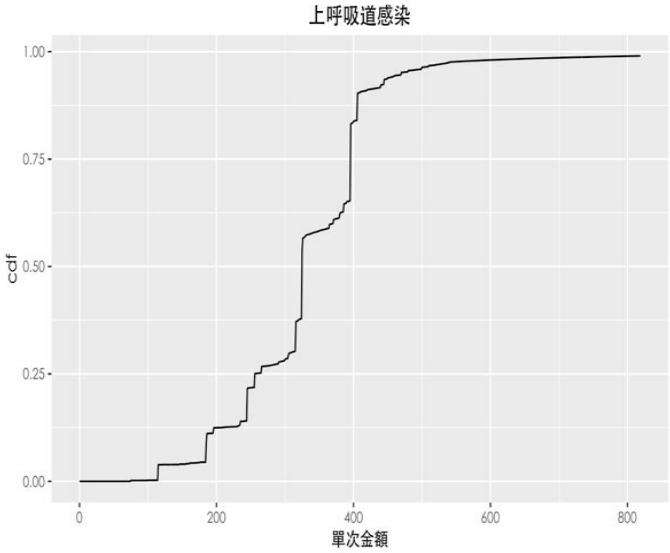

圖4、2005年上呼吸道感染單次門診金額的累積機率分佈

另外，本文也計算2005年感冒門診費用（點數）的累積機率分佈（如圖4），感冒的就醫金額分佈較低，大多數（約98%）集中在600點以下，而且無論是基層醫療院所或醫學中心，感冒的平均門診點數都介於500到600點。對比於所有疾病的單次門診點數，基層醫療院所約500點，地區醫院、區域醫院、醫學中心都大於1000點。也就是說，以基層醫療院所的門診比例、門診金額兩個因素考量，感冒確實合乎小病的定義，可惜20歲以上的感冒就醫率偏低，必須尋求其他條件用以判斷常住地。（註：若累計三年的門診紀錄，則感冒的整體就醫率可達八成，但國內縣市層級的遷移在2005年為140餘萬人次，遷移率約6%，鄉鎮市區層級的

[11] 便利商店總家數參閱維基百科，網址：https://zh.wikipedia.org/wiki/%E5%8F%B0%E7%81%A3%E4%BE%BF%E5%88%A9%E5%95%86%E5%BA%97%E5%88%97%E8%A1%A8。(2019/03/01)

遷移更為頻繁，[12]如果以累積三年的資料判斷常住地，某些地區容易受到頻繁國內遷移之扭曲。）

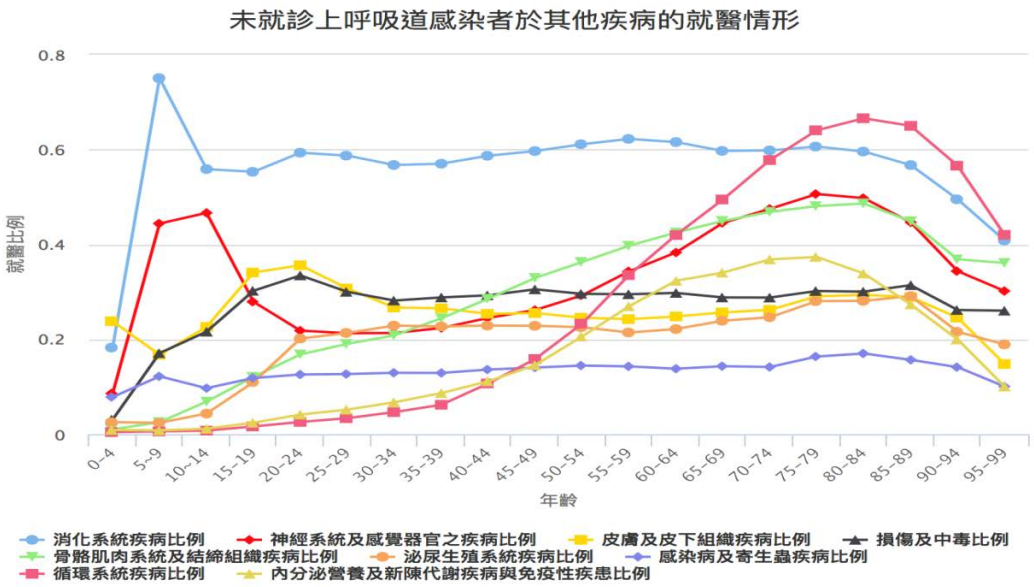

圖5、2005年未就診上呼吸道感染者於其他疾病的各年齡就醫率

上呼吸道感染算是不錯的小病判斷標準，我們嘗試另加入其他疾病，希望在不影響基層醫療院所、就醫金額兩個因素下，盡量提升整體就醫率。首先比對感冒以外疾病的各年齡就醫率，其中衛福部將疾病分為十八類，在此只顯示就醫率較高的九類疾病，以消化系統（全年齡都較高）、循環系統（高齡人口較高）比較適合補充上呼吸道感染的不足（圖5）。然而無論是消化或是循環系統疾病，單次門診金額都遠高於上呼吸道感染，而且這些疾病在基層醫療院所的門診量比例也偏低，不適於直接用於擴增上呼吸道感染的就醫率。以消化系統的單次門診金額的CDF為例（圖6），消化系統與整體疾病的金額分配很接近，但明顯高於上呼吸道感染的金額，大約98%的上呼吸道感染門診點數不高於600點（圖4），

[12] 參考中華民國統計資訊網 https://www.stat.gov.tw/ct.asp?xItem=15410&CtNode=3624&mp=4 (2019/03/01)

卻有接近40%的消化系統門診點數高於600點。經過多次的試誤法（Trial and Error），我們最後選擇加入消化系統疾病，並且限定消化疾病的門診點數不多於840點者，作為判定常住地的標準。

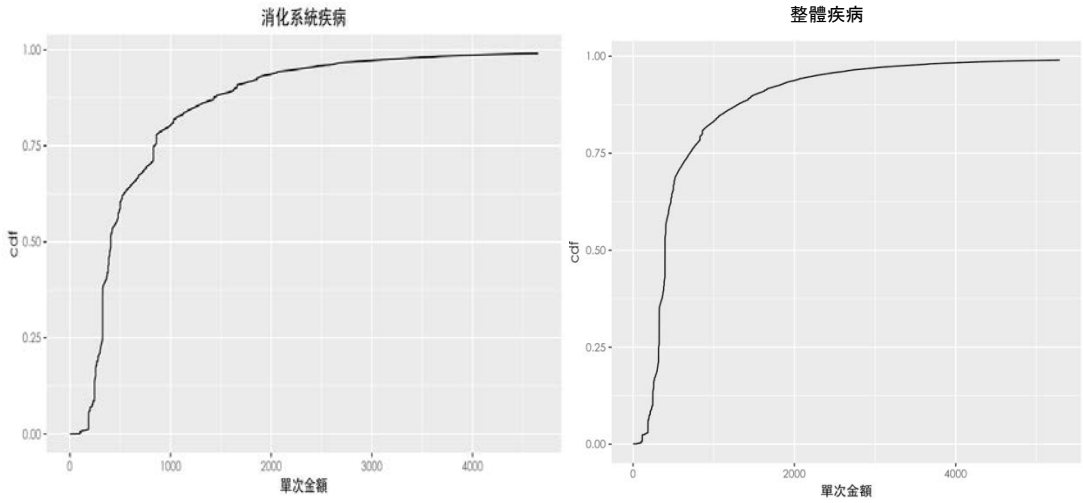

圖6、2005年消化系統（左）、整體疾病（右）單次門診金額的累積機率分佈

此外，我們也評估以門診金額、基層醫療院所就醫兩種準則作為小病的判定標準。小病和金額較小關聯性頗高，建議將門診金額少於某個數值的疾病視為小病，根據圖6判斷，門診金額的CDF累積達70%時開始減緩，因此將CDF為70%時所對應到的單次就醫總金額555點作為金額篩選的門檻，代表超過555點後的就醫總金額分佈變得分散。至於基層醫療院所就醫則較為直接，連結CD檔及HOSB檔，確定門診在基層醫療院所的所有就醫紀錄。表3紀錄上呼吸道感染、以及本研究提出的三種常住地判斷準則之比較，檢視整體就醫率、基層醫療院所的門診量比例，作為評判準則的優劣。本文提出的方法的就醫率高於上呼吸道感染，但於基層醫療院所的門診量比例上有些不如（基層醫療院所就醫不列入比較），我們建議讀者採用金額低於555點當作常住地的判斷準則，因為這個方法得出之各

年齡就醫率幾乎都是最高，整體就醫率最接近所有門診的結果，尤其在高齡組的就醫率更是優於其他方法（圖7）。

表3、四種小病判斷準則的比較（2005年）

判斷準則	就醫率	各層級醫療院所比例			
		基層院所	地區醫院	區域醫院	醫學中心
上呼吸道感染	0.701	0.902	0.047	0.032	0.018
上呼吸道感染＋消化(篩選金額)	0.798	0.838	0.051	0.061	0.050
金額低於555點	0.964	0.844	0.071	0.052	0.034
基層醫療院所	0.958	1.000	---	---	---

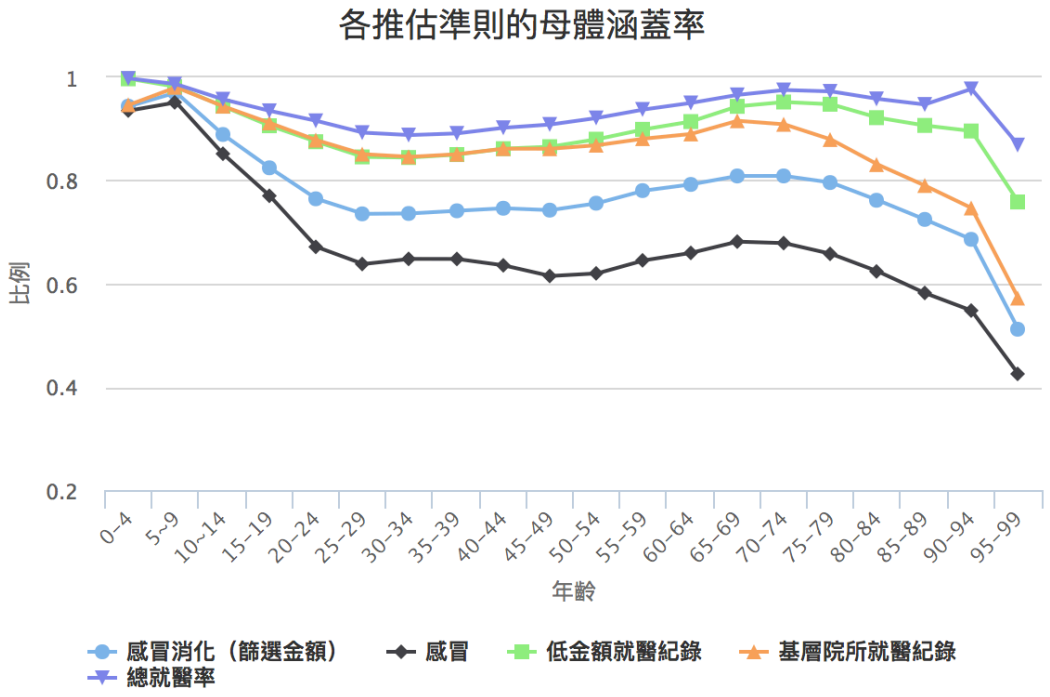

圖7、2005年所有門診、四種判斷準則的年齡別就醫率比較

第四節 進階比較與實務應用

除了考量就醫率（或是樣本涵蓋率），我們也以類似樣本代表性檢查上節提出的幾種方法，如果上述方法得出之人口結構，與健保抽樣檔中所有門診的人口結構一致，則估算出的常住人口更有說服力。但本文不採用卡方樣本代表性檢定，因為得出之結果只能用於假設檢定（拒絕或不拒絕結構的一致性），尤其當樣本數很大時，檢定結果在稍有些微差距時通常得出拒絕的結論（如第三節）。在此仿造死亡率等模型的評比，採用估計誤差作為比較標準，像是計算各年齡人口比例的 MAPE（平均絕對值誤差率；Mean Absolute Percentage Error）：

$$MAPE = \frac{100\%}{n} \times \sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}$$

其中 y_i 及 $\hat{y}_i$ 分別為全體門診、四種估計方法的第i個年齡層之人口比例（$1 \le i \le n$）。MAPE值愈小代表人口結構與母體愈接近，反之則代表結構愈不相同。

表4、2010年所有門診與各準則估計的各年齡層人口比例

年齡	所有門診	感冒	感冒＋消化	低金額	基層院所
5~19歲	0.1998	0.2463	0.2232	0.2031	0.2058
20~44歲	0.3898	0.3843	0.3843	0.3860	0.3922
45~64歲	0.2848	0.2750	0.2752	0.2842	0.2821
65歲以上	0.1256	0.1084	0.1173	0.1268	0.1199

上一節已經比較四種方法與所有門診的2005年就醫率（圖7），本節繼續比較2010年的門診人口結構。因為健保抽樣檔為封閉人口，不會加入新的樣本，2005

年的 x 歲觀察值在2010年變成 $x+5$ 歲，亦即不會有5歲以下的觀察值。另外，為了避免某些年齡層人數過少，我們將所有觀察值分成四個年齡組：5～19歲、20～44歲、45～64歲、65歲以上，除了65歲以上的高齡人口，各組人口比例約有20%以上（表4）。詳細五齡組的人口比例可參考附錄1，85歲以上者的五齡組比例都小於1%，用於計算MAPE時容易產生較大的震盪，建議以MAPE為比較標準時，分母每一項都不要太小，可仿造卡方檢定要求不得小於5%。

各方法在20～44歲及45～64歲這兩組的估計誤差最小，在5～19歲幼齡及青少年有較大的估計偏差，且四種方法皆高估幼齡組的人口比例，其中又以感冒高估比例最多，符合感冒就醫率在這個年齡較高的預期。平均而言，低金額的年齡比例估計在各年齡都與所有門診最為接近，表5的估計誤差也呈現相同的結果，低金額的估計誤差最小、上呼吸道感染的誤差最大。表5的誤差大小與表4整體就醫率的結果頗為一致，就醫率愈高者、誤差也愈小。綜合本研究的幾種比較標準，雖然上呼吸道感染提供不錯的就醫地估計，本文提出的三種小病判斷方法又優於上呼吸道感染，讀者可根據問題及實務概況、配合資料分析的便利性，選擇適合研究需求的判斷方法。另外，由於上述比較僅使用一次資料紀錄作為推估依據，原則上無法計算推估誤差及信賴區間，可使用拔靴法（Bootstrap）之類的重複捕取（Resampling）等電腦模擬方法，得出常住人口推估值的信賴區間。

表5、2010年四種方法估計值的估計誤差（單位：MAPE）

判斷方法	感冒	感冒＋消化	低金額	基層院所
估計誤差	14.56%	7.94%	1.17%	4.80%

除了提出小病判斷方法，作為估算常住人口的參考外，本文也考量常住人口估計值的實際應用，特別是人口遷移的研究。我國戶籍登記的人口資料的品質相當好，許多項目的紀錄超過60年，不止我國產官學界經常引用人口資料，外國學術研究、新聞報導也會提到。然而，戶籍資料中缺乏人口遷移的詳細紀錄，尤其是在鄉鎮市區層級，縣市層級僅在院轄市紀錄較為完整，但仍缺乏遷入地、遷出地、性別、年齡等欄位，因此學術界較少著墨於國內遷移研究，連帶影響我國縣市層級的人口推估及政策規劃。有鑑於此，本文將套用上述探討之常住地判斷準則，根據就醫行為推估我國國內遷移，探討縣市層級以下的遷移概況。

然而，因為本文根據百萬人健保抽樣檔，全臺灣的鄉鎮市區總數350餘個，平均每個鄉鎮市區的樣本數接近三千人，再細分男女兩性、不同年齡組，樣本數恐怕有不足的疑慮。另外，鄉鎮市區的人口差異甚大，即使得出鄉鎮市區層級的遷移估計值，有些地區資料可信度高、有些則很難評估。為了避免上述這些潛在問題，本文提議以立法委員選舉區為基本研究單位，全臺灣共有73個選舉區，每個選舉區人數介於20萬～50萬人，[13]區與區之間的人數差異較小（甚至小於縣市層級），如有需要也可透過內外插法、或是空間統計的Kriging，推估出鄉鎮市區層級的遷移結果。使用立委選區的另一優點是各區面積、形狀比較一致（除了少數偏遠地區與大都會地區），而且也比較符合空間統計的規則格子（Regular Lattice）佈局，有利於觀察短、中、長遷移距離（Schwartz 1973)。

我們以2005年至2010年為例，說明如何透過門診就醫資訊推估各選舉區的人口遷移。如果採用上呼吸道感染判斷常住地，先找出每個樣本在2005-2007年及2008-2010年兩個區間的常住地，如果前後兩個時期的地點不同，則判定該樣本在這兩時期間改變了經常活動的地點，同時將前後時期的地點定為遷出地、遷入

[13] 臺灣選舉區的資訊請參考中選會選舉資料庫 http://db.cec.gov.tw/。(2019/03/01)

地。接著透過各選舉區的人口資料、遷移紀錄，估算出該區的總遷移率，以及男女兩性、各年齡別的細部遷移率。圖8顯示臺灣本島73個選舉區（不考慮澎湖縣、金門縣、連江縣）的總遷移率，水平軸為遷入地、垂直軸為遷出地，顏色愈深代表遷移率愈高。從圖8可判斷出臺灣遷移概況大略可分為三個區塊：北部、中部、南部，區內的遷移比跨區遷移活躍，而且距離愈近遷移愈明顯。另外，三個區塊內也有子區塊，像是北部又可分為北北基桃、竹苗兩小塊；遷出較明顯的選舉區有花蓮縣、臺東縣，遷入較為頻繁者有新竹市、臺中市各一個選舉區。

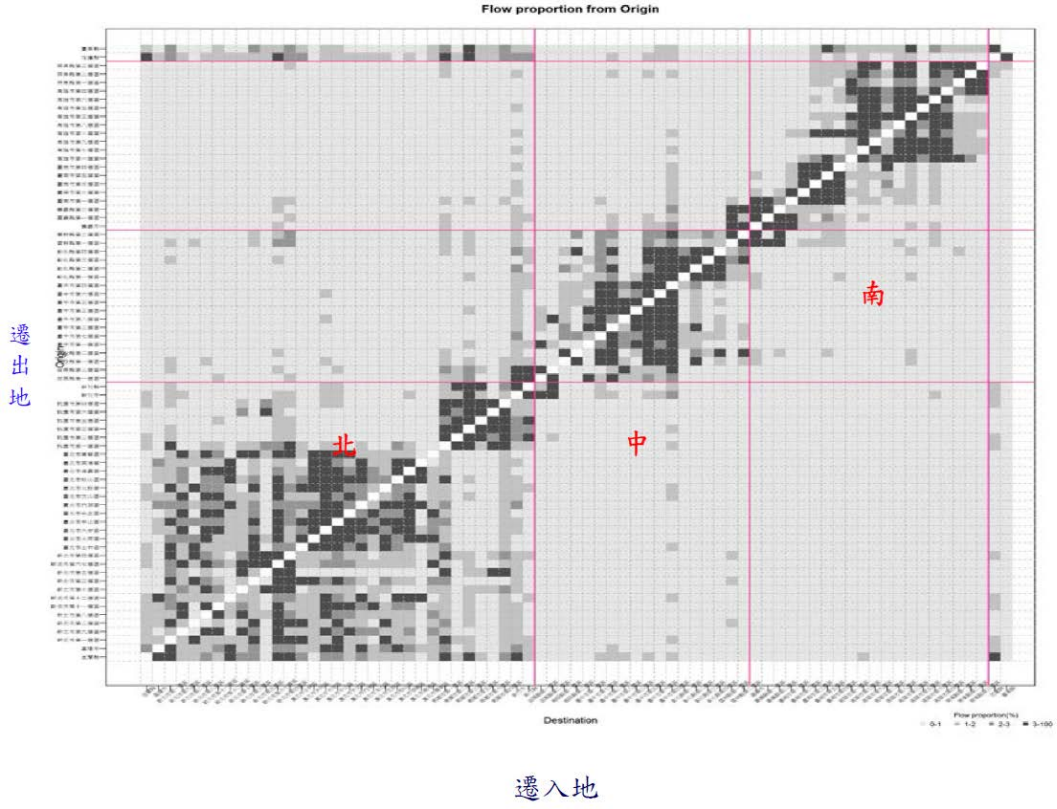

圖8、2005年各選舉區的遷移比例

上述各選舉區間的遷移推估結果可套入社群網絡（Social Network）、遷移模型（Migration Models）、或是視覺化呈現工具，進一步找出遷移的趨勢。圖9

將各區間的遷移結果以圖像呈現，兩區之間較為頻繁者以直線連結，其遷移人數則以藍色圈圈註記，圈圈愈大者遷移人數愈多。遷移大致以鄰近選舉區較為活躍，沒有跨越兩三個選舉區、或是北部、中部、南部，這個結果和圖8很接近，但圖9顯示跨區遷移沒有鄰區遷移踴躍，建議我們可嘗試引力模型（Gravity Model）估算各區之間的遷移人數（Greenwood 1985 & 1997）。另外，北部圈圈最大者為新北市、桃園市間的遷移，南部則以臺南縣、臺南市間較大。

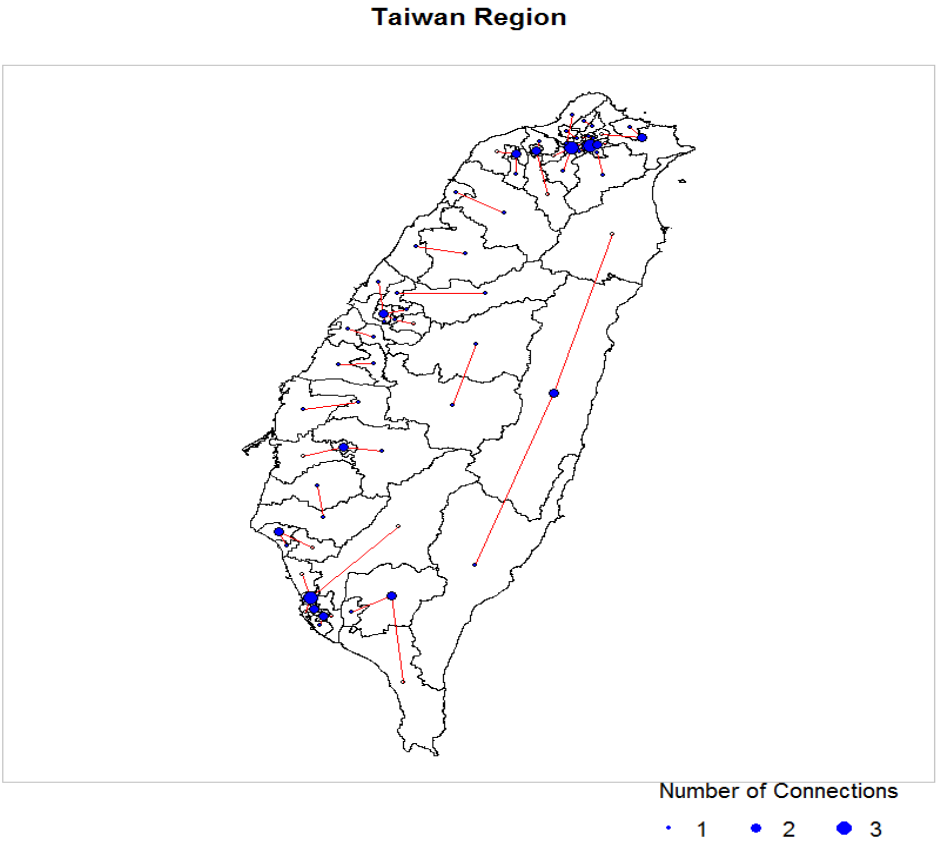

圖9、2005-2010年遷移較為頻繁地區

另外，本文重點在於探討藉由健保資料推估常住地，同時示範實務應用時的可能產出，像是於本節的人口移動，我們發現除了原先北中南三區的生活圈外，還能觀察出各區內的移動特性。由於國內學者大多使用上呼吸道感染做為常住地

的判斷依據，為了方便比較，本文也使用這個判斷標準，而沒有使用前一節的較低門診金額。

第五節 結論與建議

本文探討以健保就醫紀錄推估常住地的可行性。社會變遷使得傳統普查愈來愈難執行，包括資料品質及問卷回覆率偏低都亟需解決，而且每十年更新一次資料不符合需求，近年各界積極研發可即時反映常住人口的方法。由於我國實施全民健保超過20年，大多數醫療院所參加健保體系，全國各地都可找到醫療院所，我國國民就醫比其他先進國家便利，健保已經成為了日常生活的一部分，因此有不少學者建議以健保就醫紀錄推估我國常住人口的資訊，其中又以上呼吸道感染（俗稱感冒）最為常見。本文探討以門診就醫記錄判斷常住人口的可行性，評估現有方法的優缺點，並提出更合時宜的判斷方法。研究發現上呼吸道感染具有小病的特質，但仍需謹慎使用上呼吸道感染推估常住人口，其中感冒就醫率偏低（尤其是高齡老年人口），類似普查問回覆率太低的問題，擔心感冒就醫無法取得具有代表性的樣本。

本文提出三種判斷小病的方法，包括加入其他疾病填補中高齡樣本，彌補上呼吸道感染就醫者年齡偏低的問題。分析發現這三種方法都優於依據上呼吸道感染，整體及年齡別就醫率都較佳，其中又以門診就醫點數不大於550點最佳，可取得健保抽樣檔中96%以上樣本的常住地，即便是高齡人口也有八成以上。結合感冒及消化系統疾病（但金額不多於840點）、基層醫療院所就醫也是不錯的常住地判斷方法，可視情況與門診就醫點數不大於550點交互使用。

本文也比較以上呼吸道感染、上呼吸道感染與消化系統疾病兩者合併的差異。若以2010年普查為標準，各縣市人口比例為研究目標，經過22個縣市綜合平均

後，發現上呼吸道感染、上呼吸道感染與消化系統疾病兩者合併與普查資料的差異分別為6.3%與5.5%。雖然單以上呼吸道感染判斷仍與普查資料較為相近，但上呼吸道感染與消化系統疾病兩者合併能夠多提供了約20%的樣本人數，且其年齡族群較為平均，未來用於探討人口特性及移動等研究時較具優勢。另一方面，雖然本文考量國內學者大多使用上呼吸道感染，為了方便比較，也以這個判斷標準為依據。不過，我們也計算了門診低金額/基層院所就醫的人口結構，結果發現這兩者的年齡比例，與全體國民的年齡結構差異不大。

以健保就醫紀錄推估現住/常住人口時，因為就醫率的不同，需謹慎處理其中的差異，加入性別、年齡等因素，可套用類似抽樣調查中的事後加權方法處理，像是套用反覆重複加權（Raking）或事後分層加權（Post-stratification）依照性別、年齡等因素調整加權。由於健保抽樣檔透過分層隨機抽樣取得，以就醫紀錄推估全國或各縣市人口時也應採事後分層，如果以簡單隨機抽樣的方式推估，反而失去分層抽樣的特色。另外，如有可能，推估人口時也應將地區資訊列入考量，但套用時需格外謹慎，因為我國常住人口僅能在普查年度獲得，會有時間上的落差；而健保納保人口、或是戶籍登記的地區資訊，未必能完全反映實際的現住/常住人口。

在常住地的判斷結果方面，相較於普查提供的各縣市人口比例，本文的常住地判斷法與上呼吸道感染判斷法大多具有相同的趨勢（高估、低估），原因是即使本文已透過小病多於常住地附近就醫的特性進行控制，就醫資源的不平均仍會造成一定程度的偏差，可以透過這個性質發現各縣市的醫療資源差距，並可大致上看出各縣市間的跨區就醫現象。

過去研究人口移動、就醫需求或跨區就醫等議題時，多以上呼吸道感染所推估的常住地進行討論。但如本文所示，上呼吸道感染的患者年齡層偏低，且往往

不是具有人口移動或跨區就醫行為者，造成研究結果容易顯示出不同於現實的人口移動現象，如過低的人口移動率。本文透過加入消化系統疾病提升中老年人於樣本中的就醫率，能在未來進行人口移動相關研究時，增加較有移動潛力的人數。未來將透過此常住地判斷準則，繼續研究各年齡患者就醫地數量的多寡，討論不同年齡在選擇就醫地時的特性，並研究常住於不同都市化程度地區之患者跨區就醫之特性，藉由患者跨區就醫之距離與疾病類型討論各地患者在罹患不同疾病時的就醫習慣。

在研究限制方面，以健保就醫紀錄推估現住/常住人口時，因為各個年齡層的就醫率的不同，需謹慎處理其中的差異。由於20到44歲的青壯年人口本身就醫率較低，因此在以就醫紀錄推估常住人口時較難克服該年齡層的樣本問題，建議未來進行相關研究時可經由權數的調整修正該年齡層的比例，而加入消化系統疾病後較高的高齡比例經過青壯年人口的加權後，可望修正到與普查結果接近的比例。另外，本文在估計各年齡層結構時，顯示女性的估計結果較男性準確，原因是20到64歲的女性在上呼吸道感染與消化系統疾病的就醫率較高，進而分散了幼齡人口與高齡人口的比例，與普查結果較接近。例如，建議未來再進行相關研究時可對20到64歲的男性給予加權，平衡男性在該年齡層較少就醫衍生的問題。

參考文獻

一、 中文部份

1. 李虹映、黃信忠、許怡欣、林文德，2014，〈台灣急重症跨區就醫之變化情形：2001-2010 年〉。《臺灣衛誌》30(1)：64-74。
2. 洪永泰，1995，《戶籍登記常住人口與非常住人口之差異研究》。國科會專題研究報告。
3. 陳豔秋、楊雅惠，2017，《常住人口推計方法之研究》。國勢普查處，行政院主計總處研究報告。
4. 吳依凡，2004，《醫療資源可近性對個人醫療利用的影響─台灣地區的實證研究》。中壢：國立中央大學產業經濟研究所碩士論文。
5. 林民浩、楊安琪、溫在弘，2011，〈利用地區差異與人口學特徵評估全民健保資料庫人口居住地變項之推估原則〉。《臺灣衛誌》30(4)：347-361。
6. 林敬昇，2016，《以全民健保資料庫探討臺灣人口特性與變遷》。台北：國立政治大學商學院統計學系碩士論文。
7. 陳肇男、劉克智，2002，〈台灣 2000 年戶口普查結果的評價：常住人口與戶籍登記人口的比較分析〉。《人口學刊》25：1-56。
8. 楊雅惠，2015，《澳洲常住人口推計技術及人口普查資料蒐集方法》。行政院主計總處研究報告。
9. 廖建彰、李采娟、林瑞雄、宋鴻樟，2006，〈2000 年台灣腦中風發生率與盛行率的城鄉差異〉。《臺灣衛誌》25(3)：223-230。
10. 蔡文正、龔佩珍，2003，〈民眾對基層診所評價與就醫選擇影響因素〉。《臺灣衛誌》22(3)：181-193。
11. 顏貝珊、余清祥，2010，〈2010 年各國人口普查制度之研究〉。《人口學刊》

40：203-229。

二、英文文獻

1. Greenwood, M.J., 1985, “Human Migration: Theory, Models, and Empirical Studies.” *Journal of Regional Science* 25(4): 521-544.
2. Greenwood, M.J., 1997, “Internal Migration in Developed Countries.” Pp.647-720 in *Handbook of Population and Family Economics Volume 1*, edited by Rosenzweig Mark R & Stark Oded. North Holland: Elsevier Science Ltd.
3. Schwartz, A. 1973, “Interpreting the Effect of Distance on Migration.” *Journal of Political Economy* 81(5): 1153-1169.

附錄 1、2010 年所有門診與四種推估方法的五齡組人口比例

年齡	整體就醫	感冒	感冒消化	低金額	基層院所
5~9	0.0554	0.0765	0.0676	0.0572	0.0575
10~14	0.0715	0.0872	0.0798	0.0729	0.0738
15-19	0.0728	0.0826	0.0759	0.073	0.0745
20-24	0.0663	0.0687	0.0667	0.0657	0.0673
25-29	0.0755	0.0768	0.0752	0.0747	0.0764
30-34	0.0876	0.0888	0.0868	0.0868	0.0880
35-39	0.0785	0.0777	0.0770	0.0777	0.0787
40-44	0.0819	0.0763	0.0786	0.0810	0.0818
45-49	0.0846	0.0767	0.0811	0.0839	0.0841
50-54	0.0797	0.0712	0.0766	0.0793	0.0790
55-59	0.0722	0.0651	0.0701	0.0723	0.0714
60-64	0.0483	0.0441	0.0475	0.0487	0.0475
65-69	0.0358	0.0328	0.0349	0.0363	0.0352
70-74	0.0325	0.0291	0.0312	0.0330	0.0320
75-79	0.0252	0.0216	0.0236	0.0256	0.0241
80-84	0.0191	0.0154	0.0171	0.0192	0.0175
85-89	0.0093	0.0070	0.0077	0.0092	0.0081
90-94	0.0030	0.0021	0.0023	0.0029	0.0025
95-99	0.0007	0.0004	0.0005	0.0007	0.0005