以台灣健保申報資料進行糖尿病相關研究個案定義作法之差異

顧芳萍[1] 李昇暾[2] 李中一[1] 呂宗學[1,*]

目標：本描述性研究探討以台灣健保申報資料進行糖尿病相關研究個案定義作法之差異。**方法：**搜尋文獻資料庫找出2001至2020年發表的相關論文，整理出每一篇研究使用的個案定義作法。個案定義作法的組成包括：1)診斷編碼，2)使用資料（譬如門診，住院或藥物處方紀錄），3)診斷碼最少就診次數，4)是否有時間間隔要求。本研究參考已發表準確度研究的陽性預測值高低將不同作法進行分類。**結果：**本研究找出611篇相關研究，整理出8類30種不同個案定義作法。陽性預測值第1類最低，第8類最高。數目最多的作法依序是第8類「診斷碼且有藥物處方」194篇（32%），第1類「1次門診診斷碼」119篇（20%）與第3類「2次以上門診診斷碼」111篇（18%）。第7,8類「有使用藥物處方為個案定義條件」的研究論文篇數顯著增加，由2001-2011年的18篇增加到2018-2020年的89篇，而且86篇是用較嚴格的「有診斷碼且有藥物處方」。反之，第1,2類「1次門診或住院診斷碼」的比例由2001-2011年的31%，下降到2018-2020年的14%。**結論：**近二十年來，越來越多研究使用較嚴格（高陽性預測值）的個案定義作法。未來應該有研究進一步產出台灣不同個案定義作法的準確度指標數據，提供後續研究者引用，以符合國際學術期刊對於使用例行收集行政資料研究的報告規範要求。（台灣衛誌 2021；**40(6)**：725-733）

關鍵詞：糖尿病、台灣健保申報資料、個案定義作法

前　言

公共衛生研究（尤其是醫療照護研究）與臨床醫學研究（尤其是藥物流行病學與藥物安全研究）越來越重視使用全人口行政資料與真實世界資料，可以產出傳統單一醫院樣本與臨床試驗研究無法產出的重要資訊[1-7]。健保申報資料是台灣非常重要的全人口行政資料與真實世界資料，使用健保資料發表在國際學術期刊的論文數目越來越多[8,9]。截至2021年8月20日PubMed收錄使用台灣健保申報資料發表的論文篇數已達7,325篇，由2005年的23篇，增加到2020年的838篇[10]。根據一篇分析2000至2015年4,473篇使用台灣健保申報資料發表論文，糖尿病，腦中風與失智症是頻率最高的三個疾病主題，數目分別是263,189與139篇[9]。再者，因為申報資料沒有檢驗結果數值，因此必須以國際疾病分類（International Classification of Disease, ICD）編碼來進行糖尿病個案定義，會有一些準確度的問題（後詳述）。因此，本研究將針對使用台灣健保申報資料探討糖尿病相關研究以ICD編碼進行糖尿病個案定義作法的差異進行探討。

[1] 國立成功大學醫學院公共衛生研究所
[2] 國立成功大學管理學院資訊管理研究所
[*] 通訊作者：呂宗學
地址：台南市東區大學路1號
E-mail：robertlu@mail.ncku.edu.tw
投稿日期：2021年9月8日
接受日期：2021年12月3日
DOI:10.6288/TJPH.202112_40(6).110120

隨著研究論文數目快速大量增加，學術期刊也開始對於使用行政資料與真實世界資料研究的方法與結果的報告有較嚴格的要求，於是有「使用觀察性例行收集健康資料研究的報告」規範（RECORD, REporting of studies Conducted using Observational Routinely-collected health Data）的提出[11,12]。該規範在材料與方法部分的6.1關於個案定義的描述，建議要詳細寫出所使用的編碼與作法（codes and algorithms）（註一），6.2建議要寫出前述編碼與作法的準確度數據。

RECORD之所以會有這樣的要求，是因為許多行政資料的收集目的不是為了研究，許多資料的準確度可能不是很好。譬如健保申報資料主要目的是為了申請費用，因此與申請費用相關的醫令碼會比較準確，許多診斷碼沒有涉及費用申請，準確度就會較差。再者，因為健保單位有核刪機制，所以也常常出現為了檢驗而下診斷碼的作法。譬如王先生有三多（多飲、多食、多尿）症狀，醫師懷疑是糖尿病，於是處方醣化血色素與空腹血糖檢驗。因為怕被核刪，於是鍵入糖尿病診斷碼。後來檢驗結果顯示不是糖尿病，醫師以後的診斷就不會鍵入糖尿病診斷碼。但是，也有可能第二次就醫時，醫師沒有刪除糖尿病診斷。到第三次就診時，醫師才刪除糖尿病診斷。

研究者可以使用「門診出現1次糖尿病診斷碼」，「門診出現2次糖尿病診斷碼」與「門診出現3次糖尿病診斷碼」三種糖尿病個案定義作法（case definition algorithms）。過去驗證糖尿病編碼準確度的文獻中，使用的金字標準主要有藥物處方資料、檢驗資料、問卷調查、病歷摘要或特定糖尿病註冊資料等。金字標準定義糖尿病的條件根據使用的資料不同。有些研究只使用一個資料來源，例如藥物處方資料中有降血糖藥物處方紀錄定義為糖尿病個案。有些研究則使用多個資料來源，例如有降血糖藥物處方紀錄或2次血糖檢驗結果異常或病歷有糖尿病文字診斷者定義為糖尿病個案。接著比對不同的診斷碼的個案定義作法，計算陽性預測值以及敏感度等準確度指標。陽性預測值代表個案定義作法判斷有糖尿病中金字標準定義為糖尿病個案的比例。敏感度代表金字標準定義為糖尿病個案中個案定義作法判斷有糖尿病的比例。因此，王先生在前兩個較寬鬆的個案定義作法會被定義為有糖尿病而納入研究，這時就會發生偽陽性錯誤分類。如果研究者採用第三種較嚴格的個案定義作法，王先生就不會被納入研究。使用較嚴謹個案定義作法也是有代價的，譬如使用「有診斷碼且有處方降血糖藥物」做為個案定義，陽性預測值（分母是個案定義作法認定有糖尿病個案數，分子是金字標準判定真的有糖尿病病個案數）非常高。但是代價就是，許多輕症糖尿病患不會被納入研究，敏感度（分母是金字標準判定真的有糖尿病個案數，分子是被個案定義作法判斷為有糖尿病個案數）會降低。

RECORD要求使用行政資料的研究論文在方法描述時要明確寫出個案定義使用的診斷編碼與作法，最好還能寫出這個作法的準確度數據（譬如陽性預測值，陰性預測值，敏感度與特異度等），讓讀者可以判斷該個案定義作法偽陽性與偽陰性的機率。RECORD關於討論限制的寫作提醒19.1，建議能寫出錯誤分類偏差對結果方向性上的影響，這時就要引用前述準確度數據進行推估。本描述性研究的目的是要探討以台灣健保申報資料進行糖尿病相關研究個案定義作法之差異。本研究的結果對於使用行政資料進行研究者在研究設計個案定義作法時，能有實務上的參考。

材料與方法

資料來源

本研究在2021年3月20日使用PubMed資料庫搜尋2001-2020年使用台灣健保資料發表的糖尿病相關研究論文。搜尋關鍵字與條件為：（“Taiwan” or ”Taiwanese”）AND （“insurance” or ”health insurance” or ”national health programs” or ”national health insurance” or ”claim data” or ”administrative claim” or ”administrative data” or ”claim

database” or ”medical claim” or ”nationwide”）AND （“diabetes” or “diabetic” or ”diabetes mellitus”），共獲得1,450篇。閱讀每一篇論文摘要，排除不是使用健保申報資料的研究，再排除糖尿病是共病（控制變項）的研究，最後剩下611篇是針對糖尿病個案為主要自變項或依變項的相關研究。

糖尿病個案定義作法

個案定義作法的組成包括：1）診斷編碼，2）使用資料（譬如門診，住院或藥物處方紀錄），3）診斷碼最少就診次數，4）是否有時間間隔要求。目前收集到的研究都是使用國際疾病分類編碼第九版250以及早期台灣健保局使用簡碼A181，所以本次分類不會考慮使用不同診斷編碼這個組成。本研究的重點是要探討不同研究對於個案定義的不同作法：譬如出現1次門診診斷碼就算是糖尿病個案，還是要出現2次門診診斷碼才算是糖尿病個案，還是更嚴格要出現3次門診診斷碼才算是糖尿病個案。前述三種作法的陽性預測值會越來越高，但是敏感度相對地越來越低。

關於組成二「使用資料」，健保申報資料包括門診申報資料，住院申報資料，藥物處方紀錄，重大傷病檔（第一型糖尿病）或是特殊給付方案資料（譬如糖尿病論質計酬方案）。組成三是「診斷碼至少就診次數」，要求至少就診次數越多，個案定義越嚴格，陽性預測值越高。組成四是要求「時間間隔」，有研究者定義兩次出現時間要超過一個月才定義是個案，一個月內連續出現多次不算。也有研究者定義上述多次診斷出現必須侷限一年內，兩年後才出現第二次診斷就不算。

因為上述組成二至組成四可以排列組合出相當多種作法，為了避免太複雜交叉分析，我們將不同的個案定義作法進行合併分組。本研究回顧十多篇糖尿病個案定義作法準確度研究，選擇一篇加拿大安大略省全人口研究結果做為分類參考[13]。該研究是所有準確度研究中個案定義作法種類最多（22種作法）的研究，表一依照陽性預測值大小排序不同作法。由表一可以發現，使用的資料來源相同，陽性預測值愈高，敏感度愈低。譬如使用門診或住院申報資料，準確度結果皆為陽性預測值愈高，敏感度愈低。比較不同資料之間的準確度，可能受到個案是否接受藥物或有無住院而影響敏感度。譬如「1次住院診斷」作法敏感度低，因為並不是所有病患皆有住院，雖然陽性預測值高，但敏感度僅有37%。

本研究將所有個案定義作法分類成八類：第1類：1次門診診斷碼，第2類：1次住院診斷碼，第3類：2次門診診斷碼，第4類：1次住院或2次門診診斷碼，第5類：3次門診診斷碼，第6類：1次住院或3次門診診斷碼，第7類：診斷碼或藥物處方，第8類：診斷碼且有藥物處方。雖然本研究未驗證不同個案定義作法的準確度，但根據過去準確度研究的研究結果，第8類陽性預測值93-98%較高，而第1類與第7類陽性預測值為58%-59%相較於其他組較低。相反的，第1類與第7類的敏感度皆在94%以上，第8類敏感度只有50%較其他類低[13]。

論文特徵相關分析

本研究將論文特徵與前述個案定義作法分類進行卡方檢定交叉分析。論文特徵包括：發表年代，研究主題與期刊影響係數。發表年代區分為2001-2011，2012-2014，2015-2017與2018-2020年。研究主題區分為三類，一是「疾病風險」，譬如探討糖尿病病患是否有較高風險罹患其他疾病，或是探討某些疾病是否有較高風險罹患糖尿病等。二是「藥物效果與風險」，譬如探討降血糖藥物的效果或是對於其他疾病發生的風險，或是探討其他藥物對於新發生糖尿病的風險。三是「其他」，譬如糖尿病盛行率，併發症發生率，糖尿病照護品質，糖尿病藥物使用描述性研究，糖尿病醫療服務利用等。期刊影響係數採用Journal Citation Reports（JCR）於2020年發布的分數區分為小於5分與大於等於5分。

表一　加拿大安大略省全人口不同糖尿病個案定義作法準確度[13]

糖尿病個案定義作法	陽性預測值%	敏感度%
1次門診診斷且1次藥物處方紀錄	98.5	50.0
1次診斷碼且有1次藥物處方	98.5	50.2
1次藥物處方紀錄	97.3	50.7
1次住院診斷或1次藥物處方	93.3	61.3
1次診斷碼且有特定負擔代碼	92.6	77.2
1次診斷碼且有1次藥物處方或特定負擔代碼	92.5	84.2
1次住院診斷	91.9	36.7
1年內3次門診診斷	91.4	79.9
2年內3次門診診斷	90.1	83.1
1年內1次住院診斷或3次門診診斷	89.1	82.4
2年內1次住院診斷或3次門診診斷	88.0	84.9
2年內1次住院診斷或3次門診診斷或1次藥物處方	87.5	87.4
1年內2次門診診斷	84.6	87.2
（1年內2次門診診斷）或（1次門診診斷且1次藥物處方）	84.6	88.6
2年內2次門診診斷	83.4	88.4
1年內1次住院診斷或2次門診診斷	83.0	88.4
1年內1次住院診斷或2次門診診斷或1次藥物處方	82.6	90.0
2年內1次住院診斷或2次門診診斷	81.9	89.3
2年內1次住院診斷或2次門診診斷或1次藥物處方	81.5	90.7
1次門診診斷	58.5	93.6
1次門診診斷或1次藥物處方紀錄	58.4	94.4
1次住院診斷或1次門診診斷	58.0	94.0

結　果

本研究找出611篇相關研究，整理出8類30種個案定義作法（表二）。數目最多的作法依序是第8類「診斷碼且有藥物處方」194篇（32%），第1類「1次門診診斷碼」119篇（20%）與第3類「2次門診診斷碼」111篇（18%），糖尿病主題式資料庫屬於本研究中的第6類。

表三是八類個案定義研究論文篇數的年代分布（卡方檢定$p<0.001$），圖一將八類個案定義作法合併為四類，呈現不同年代的分布。可以看到第7,8類「有使用藥物處方為個案定義條件」的研究論文篇數顯著增加，由2001-2011年的18篇增加到2018-2020年的89篇，而且86篇是用較嚴格的「有診斷碼且有藥物處方」。反之，第1,2類「1次門診或住院診斷碼」的比例由2012-2014年的19%，下降到2018-2020年的14%。

表四是不同個案定義類型與研究主題與期刊影響係數的分布。此部分目的在於探討不同研究主題、期刊與使用的個案定義作法是否有相關，譬如藥物效果與風險相關的研究傾向使用包含藥物條件的個案定義作法，發表在影響係數高的期刊的研究使用較嚴格的個案定義作法。研究結果顯示不同研究主題使用的個案定義作法有統計上顯著差異（卡方檢定$p<0.001$），藥物效果與風險為主題的研究超過四成（46%）的個案定義有包括有藥物處方為條件。風險相關研究也有四分之一是採用有藥物處方為條件的個案定義作法。不同個案定義作法在不同影響係數期刊分布沒有統計顯著差異（卡方檢定$p=0.685$）。

討　論

本研究發現近二十年來以台灣健保申報

表二　台灣使用健保申報資料進行糖尿病相關研究611篇論文，有30種不同個案定義作法，依照資料來源及作法嚴謹度分成八類

個案定義作法	篇數
第1類：1次門診診斷碼	119
1次診斷碼（不限門診或住院）	112
1次門診診斷碼	7
第2類：1次住院診斷碼	4
1次住院診斷碼	3
1次住院診斷碼且有1次門診診斷碼	1
第3類：2次門診診斷碼個案定義作法	111
2次診斷碼[a]	5
2次門診診斷碼	5
1年內有2次診斷碼	9
1年內有2次門診診斷碼	72
1年內有2次門診診斷碼且相隔30天	20
第4類：1次住院或2次門診診斷碼	52
1次住院或2次門診診斷碼	29
1年內1次住院或2次門診診斷碼	23
第5類：3次門診診斷碼個案定義作法	32
3次診斷碼[a]	8
3次門診診斷碼	3
1年內有3次診斷碼[a]	2
1年內有3次門診診斷碼	12
1年內有4次以上門診診斷碼	7
第6類：1次住院或3次門診診斷碼個案定義作法	79
1次住院或3次門診診斷碼	31
1年內1次住院或3次門診診斷碼	48
第7類：診斷碼或藥物處方紀錄	20
1次門診診斷碼或1次藥物處方紀錄	13
2次診斷碼或有1次藥物處方紀錄[a]	2
1次住院或2次門診診斷碼或1次藥物處方紀錄	5
第8類：診斷碼且有藥物處方紀錄	194
1次藥物處方紀錄	18
1次診斷碼且有1次藥物處方紀錄[a]	76
2次診斷碼且有藥物處方紀錄[a]	34
3次診斷碼且有藥物處方紀錄[a]	12
1次住院或2次以上門診診斷且有藥物處方	14
1年內1次住院或2次以上門診診斷且有藥物處方	14
重大傷病診斷碼	21
糖尿病論質計酬醫令代碼	4
診斷碼且有檢驗醫令碼[a]	1

註：[a]不限門診或住院診斷碼。

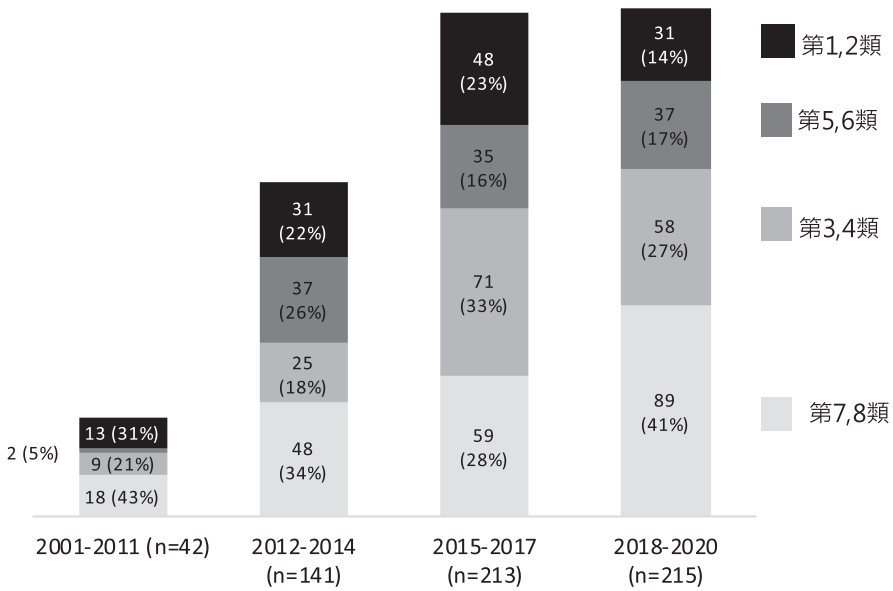
資料進行糖尿病相關研究有六百多篇，有四成的個案定義作法有使用藥物處方為條件，是陽性預測值最高的作法，而且越晚進行的研究，越多使用這類個案定義作法。

促成越來越多論文使用較嚴謹個案定義作法的可能解釋有二：一是探討糖尿病藥物

表三　不同糖尿病個案定義作法年代分佈

糖尿病個案定義作法	總計		2001-2011		2012-2014		2015-2017		2018-2020	
	No	%	No	%	No	%	No	%	No	%
總篇數	611	100.0	42	100.0	141	100.0	213	100.0	215	100.0
第1類：1次門診診斷碼	119	19.5	13	31.0	27	19.1	48	22.5	31	14.4
第2類：1次住院診斷碼	4	0.7	0	0.0	4	2.8	0	0.0	0	0.0
第3類：2次門診診斷碼	111	18.2	8	19.0	18	12.8	51	23.9	34	15.8
第4類：1次住院或2次門診診斷碼	52	8.5	1	2.4	7	5.0	20	9.4	24	11.2
第5類：3次以上門診診斷碼	32	5.2	2	4.8	16	11.3	10	4.7	4	1.9
第6類：1次住院或3次以上門診診斷碼	79	12.9	0	0.0	21	14.9	25	11.7	33	15.3
第7類：診斷碼或藥物處方	20	3.3	5	11.9	8	5.7	4	1.9	3	1.4
第8類：診斷碼且有藥物處方	194	31.8	13	31.0	40	28.4	55	25.8	86	40.0

卡方檢定p<0.001



圖一　不同糖尿病個案定義作法類型研究論文篇數年代分布

第1,2類：1次門診或1次住院診斷碼
第3,4類：2次門診或1次住院診斷碼
第5,6類：3次門診或1次住院診斷碼
第7,8類：包括藥物處方為個案定義條件

的相關研究逐年增加，這類研究大多會納入使用糖尿病藥物處方為個案定義條件。二是2011年國家衛生研究院發行糖尿病人抽樣歸人檔（有67篇論文使用此資料檔），2017年釋出糖尿病主題資料庫，這些資料檔都是採用較嚴謹的個案定義作法。糖尿病主題式資料庫個案定義作法為一年內有1次住院或3次門診診斷編碼，屬於本研究的第6類。本研究僅納入一篇使用糖尿病主題式資料庫的研究，但從研究結果可以發現，使用第6類研究定義者有79篇，且使用的趨勢逐年增加。根據加拿大準確度研究結果，第6類的陽性

表四　不同研究主題與期刊影響係數（IF）糖尿病個案定義作法分佈

糖尿病個案定義作法	總計		風險相關		藥物		其他		IF＜5		IF≧5	
	No	%	No	%	No	%	No	%	No	%	No	%
總計	611	100.0	247	100.0	254	100.0	110	100.0	411	100.0	200	100.0
第1類：1次門診診斷碼	119	19.5	65	26.3	43	16.9	11	10.0	82	20.0	37	18.5
第2類：1次住院診斷碼	4	0.7	3	1.2	1	0.4	0	0.0	2	0.5	2	1.0
第3類：2次門診診斷碼	111	18.2	50	20.2	43	16.9	18	16.4	68	16.5	43	21.5
第4類：1次住院或2次門診診斷碼	52	8.5	21	8.5	17	6.7	14	12.7	38	9.2	14	7.0
第5類：3次以上門診診斷碼	32	5.2	12	4.9	12	4.7	8	7.3	21	5.1	11	5.5
第6類：1次住院或3次以上門診診斷碼	79	12.9	29	11.7	23	9.1	27	24.5	58	14.1	21	10.5
第7類：診斷碼或藥物處方	20	3.3	6	2.4	14	5.5	0	0.0	13	3.2	7	3.5
第8類：診斷碼且有藥物處方	194	31.8	61	24.7	101	39.8	32	29.1	129	31.4	65	32.5

研究主題卡方檢定p<0.001，期刊影響係數卡方檢定p=0.685

預測值以及敏感度皆在80%以上[13]。

由於使用台灣健保申報資料進行糖尿病相關研究的研究者大多是臨床醫師與臨床藥物流行病學學者，所以傾向使用較嚴格個案定義，比較可以看到糖尿病對於其他疾病之影響與糖尿病藥物的影響。使用較嚴格個案定義作法的代價，會排除不少輕症或初期糖尿病個案。如果研究目的是公共衛生全人口觀點探討糖尿病盛行率，糖尿病進程與併發症發生率與疾病負擔估計等議題，這時候就應該納入輕症與初期糖尿病個案，所以不應該使用敏感度太低的個案定義作法。考量陽性預測值與敏感度的取捨，第6類「1次住院或3次門診診斷碼」個案定義作法應該是比較折衷的作法。

由實務觀點，本研究特別區分個案定義的四個組成，組成二「使用資料」可以提供研究者實務參考。目前使用健保申報資料進行研究都必須向衛生福利部資料科學中心購買資料，門診與住院申報資料的診斷碼與藥物醫令碼。使用較嚴格個案定義作法，所需要勾選購買的變項數目就會較多，當然費用也更高。

本研究的強項是針對台灣使用健保申報資料研究的方法學實務問題進行實證描述性分析，提醒研究者能留意國際學術期刊對於使用行政資料撰寫論文RECORD規範要求，必須對於個案定義的診斷編碼與個案定義作法有詳細描述。

本研究有不少限制，解釋本研究發現時要有所保留。一是本研究已經使用相當寬廣的關鍵字搜尋，但是還是有可能有遺漏符合條件的論文沒有納入。二是第一作者人工閱讀上千篇論文摘要進行篩選，也可能誤刪一些論文。三是判斷每篇論文的個案定義作法，有時候因為論文沒有詳細描述，所以有可能判斷錯誤。通訊作者有選擇部分論文來進行信度評估，雙盲比對結果大多是一致。本研究有記錄每一篇論文的分類，如果讀者想要重複驗證，可以來函索取相關紀錄。四是本研究對於不同作法的分類是依據加拿大研究的陽性預測值來分類。台灣的健保給付規定，醫療體系與醫師行為與加拿大可能都有差異，台灣不同個案定義作法的陽性預測值可能不同於加拿大。台灣目前只有一篇糖尿病診斷碼的準確度研究，可惜該篇研究是以問卷病患當金字標準，而且所探討的個案定義作法種類不多，年代也較早期[14]。台灣應該要有使用更好金字標準針對不同個案定義作法進行準確度評估的研究，產出台灣本土不同個案定義作法的準確度數據，讓相關研究者可以引用，才能符合RECORD規範的要求。

結語

本研究回顧二十年來使用台灣健保資料庫進行糖尿病相關研究論文，發現越近期研

究使用個案定義作法越嚴謹（偽陽性糖尿病患數目減少）。近三年研究，約有四成的研究加入使用抗糖尿病藥物當個案定義條件之一。使用越嚴謹（陽性預測值高）的個案定義作法代價就是排除許多輕症或初期糖尿病患（敏感度降低）。公共衛生研究者的研究目的可能不同於臨床醫師與藥物流行病學學者，可要採取較折衷的「1次住院或3次門診診斷碼」個案定義作法。為了符合國際學術期刊規範要求：要寫出個案定義作法的準確度數值佐證。台灣應該多準確度研究，產出本土不同醫療層級不同糖尿病個案定義作法的準確度指標，方便相關研究者引用，以符合RECORD規範要求。

註一：Algorithm通常翻譯為「演算法」，但是在本研究關於個案定義並沒有複雜的程式語法或是步驟，所以簡單翻譯為「作法」。

參考文獻

1. Sarrazin MS, Rosenthal GE. Finding pure and simple truths with administrative data. JAMA 2012;**307**:1433-5. doi:10.1001/jama.2012.404.
2. Ambroggio LV, Shah SS. Administrative data: expanding the infrastructure for pediatric research. J Pediatr 2013;**162**:681-4. doi:10.1016/j.jpeds.2012.10.040.
3. Garland A, Gershengorn HB, Marrie RA, Reider N, Wilcox ME. A practical, global perspective on using administrative data to conduct intensive care unit research. Ann Am Thorac Soc 2015;**12**:1373-86. doi:10.1513/AnnalsATS.201503-136FR.
4. Leopold Z, Dave P, Menon A, et al. Trends in the use of administrative databases in urologic oncology: 2000-2019. Urol Oncol 2021;**39**:487-92. doi:10.1016/j.urolonc.2021.01.014.
5. Sherman RE, Anderson SA, Dal Pan GJ, et al. Real-world evidence - what is it and what can it tell us? N Engl J Med 2016;**375**:2293-7. doi:10.1056/NEJMsb1609216.
6. Corrigan-Curay J, Sacks L, Woodcock J. Real-world evidence and real-world data for evaluating drug safety and effectiveness. JAMA 2018;**320**:867-8. doi:10.1001/jama.2018.10136.
7. Basch E, Schrag D. The evolving uses of "real-world" data. JAMA 2019;**321**:1359-60. doi:10.1001/jama.2019.4064.
8. Lin LY, Warren-Gash C, Smeeth L, Chen PC. Data resource profile: the National Health Insurance Research Database (NHIRD). Epidemiol Health 2018;**40**:e2018062. doi:10.4178/epih.e2018062.
9. Sung SF, Hsieh CY, Hu YH. Two decades of research using Taiwan's National Health Insurance claims data: bibliometric and text mining analysis on PubMed. J Med Internet Res 2020;**22**:e18457. doi:10.2196/18457.
10. 國立成功大學健康資料加值應用研究中心：看見健康數據網站「論文搜尋」。https://visualizinghealthdata.idv.tw/?route=article/thesis。引用2021/8/20。
    NCKU Research Center for Health Data. Visualizing health data-search. Available at: https://visualizinghealthdata.idv.tw/?route=article/thesis. Accessed August 20, 2021. [In Chinese]
11. Benchimol EI, Smeeth L, Guttmann A, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Med 2015;**12**:e1001885. doi:10.1371/journal.pmed.1001885.
12. Langan SM, Schmidt SA, Wing K, et al. The reporting of studies conducted using observational routinely collected health data statement for pharmacoepidemiology (RECORD-PE). BMJ 2018;**363**:k3532. doi:10.1136/bmj.k3532.
13. Lipscombe LL, Hwee J, Webster L, Shah BR, Booth GL, Tu K. Identifying diabetes cases from administrative data: a population-based validation study. BMC Health Serv Res 2018;**18**:316. doi:10.1186/s12913-018-3148-0.
14. Lin CC, Lai MS, Syu CY, Chang SC, Tseng FY. Accuracy of diabetes diagnosis in health insurance claims data in Taiwan. J Formos Med Assoc 2005;**104**:157-63. doi:10.29828/JFMA.200503.0002.

Variations in case definition algorithms in diabetes-related studies using Taiwan National Health Insurance claims data

Fang-Ping Ku[1], Sheng-Tun Li[2], Chung-Yi Li[1], Tsung-Hsueh Lu[1,*]

Objectives: This descriptive study examined the variations in case definition algorithms in diabetes-related studies using Taiwan National Health Insurance claims data. **Methods:** We searched the PubMed database to retrieve relevant papers published between 2001 and 2020. The components of a case definition algorithm included 1) diagnostic codes, 2) data used, 3) minimum number of visits with diagnostic codes, and 4) time intervals required. We grouped the algorithms according to their positive predictive value (PPV) derived from a published validity study. **Results:** We identified 611 studies with 30 distinct case definition algorithms and classified them into 8 groups. The PPV was lowest in Group 1 and highest in Group 8. The three most frequently used algorithm appeared in Group 8, ("diagnostic code AND antidiabetic drugs prescribed," 194 papers, 32%), followed by Group 1 ("at least one outpatient diagnostic code," 119 papers, 20%) and Group 3, ("at least two outpatient diagnostic codes," 111 papers, 18%). The number of papers in Groups 7 and 8 that used antidiabetic drugs as a condition for case definition increased prominently, from 18 between 2001 and 2011 to 89 between 2018 and 2020. Furthermore, 86 papers used the more rigorous definition "diagnostic codes AND medications." However, the proportion of papers in Groups 1 and 2 decreased from 31% between 2001 and 2011 to 14% between 2018 and 2020. **Conclusions:** The number of diabetes-related studies using more rigorous (higher PPV) case definition algorithms increased between 2001 and 2020. Additional studies, which are requested through the reporting of studies conducted using observational routinely collected health data (RECORD), on the validity of these algorithms are required. (*Taiwan J Public Health. 2021;**40(6)**:725-733*)

Key Words: diabetes mellitus, Taiwan National Health Insurance claims data, case definition algorithm

[1] Department of Public Health, College of Medicine, National Cheng Kung University, No. 1, University Rd., East Dist., Tainan, Taiwan, R.O.C.

[2] Department of Industrial and Information Management, College of Management, National Cheng Kung University, Tainan, Taiwan, R.O.C.

[*] Correspondence author E-mail: robertlu@mail.ncku.edu.tw

Received: Sep 8, 2021 Accepted: Dec 3, 2021

DOI:10.6288/TJPH.202112_40(6).110120