99 年 200 萬人世代追蹤抽樣檔建置及驗證報告

一、 緣起

為提供研究經費額度有限或研究時間有限之計畫能應用衛生資料進行學術研究，於 100 年 8 月建置 200 萬人世代追蹤抽樣檔供外界申請使用，並規劃每 5 年為建置週期，目前已有 89 年及 94 年 2 組抽樣檔，內容包括健保資料及死因資料之常用欄位。其中 89 年抽樣檔往後追蹤 14 年資料(89-102 年)；94 年抽樣檔往後追蹤 9 年資料(94-102 年)及往前追溯至 89 年(89-93 年)。

本部衛生福利資料科學中心第 3 組 200 萬人世代追蹤抽樣檔係以 99 年為起始年抽取 200 萬人，提供往後追蹤 5 年(99-103 年)及往前追溯至 89 年(89-98 年)之健保資料、癌症登記資料及死因資料之常用欄位。

二、 99 年 200 萬人世代追蹤抽樣檔建置說明

(一) 抽樣母體

以 99 年 12 月 31 日之內政部戶籍檔人口為母體，比對 99 年健保在保人口，依下列檢誤條件剔除不合宜資料後即為抽樣母體。

1. 非 99 年內政部戶籍檔現住人口。
2. 非 99 年健保在保者，99 年健保在保者之定義為 99 年 1 月至 99 年 12 月間，任一月在保者。
3. 身分證字號重複者(同身份證字號但生日不同)。
4. 性別不詳者。
5. 出生日期有誤者(包含欄位缺漏、格式錯誤、99 年 12 月 31 日後出生等)。
6. 年齡非 0-120 歲者。
7. 98 年 12 月 31 日(含)前死亡者。
8. 戶籍地不明者。

(二) 抽樣設計

將抽樣母體以性別、年齡及戶籍地為分層依據，依各層比例於各層內隨機抽取樣本，本抽樣檔所需抽樣總人數為 200 萬人。

1. 分層準則與層數

(1) 性別：分為男、女，共 2 層。

(2) 年齡：分為日齡小於 28 天、日齡大於等於 28 天但未滿 1 歲、1 歲至未滿 5 歲、5

歲至未滿 10 歲、…、80 歲至未滿 85 歲、85 歲以上，共 20 層。

(3) 地區：依戶籍地所屬之健保分局分為台北分局、北區分局、中區分局、南區分局、高屏分局及東區分局，共 6 層。

2. 抽取樣本數

所需抽樣總人數為 200 萬人，依上述各層人口數佔抽樣母體人口數之比例分配各層所需樣本。公式如下：

各分層所需抽樣人數 $f_{ijk} = \frac{n_{ijk}}{N} \times 2{,}000{,}000$

f_{ijk}：第 ijk 層需抽樣人數

$\frac{n_{ijk}}{N}$：各分層抽出率

N：抽樣母體總人口數

n_{ijk}：各層人口數

i：1、2，性別分層; j：1-20，年齡分層; k：1-6，地區分層

3. 抽樣方法

採分層隨機抽樣法，以性別、年齡及戶籍地為分層依據，計算出各層所需抽取樣本數後，進行抽樣，所需抽樣總人數為 200 萬人。

(三) 樣本代表性驗證

進行 200 萬人世代追蹤抽樣檔代表性驗證，驗證 200 萬人抽樣檔與抽樣母體間是否具顯著性差異，其中類別變項驗證採用卡方檢定，連續變項驗證採用 K-S(Kolmogorov-Smirnov)檢定。

三、第 3 組 200 萬人抽樣檔相較前 2 組抽樣檔之差異情形

(一) 比對 99 年健保承保檔在保人口，以避免比對健保資料後人數不足 200 萬人。

(二) 除進行十大死因及當年死亡人數分布驗證外，新增全民健保投保金額、投保地區、身分屬性、每年出生人數分布及健保住院就診率前五大疾病之年齡分布為驗證變項，驗證 99 年 200 萬人世代追蹤抽樣檔與抽樣母體間是否具顯著性差異。

(三) 新增癌症登記檔為常用欄位。

四、 99 年 200 萬人世代追蹤抽樣檔之驗證結果

(一)類別變項

表 1.全民健保承保檔(H_NHI_ENROL)類別變項驗證結果

變項	抽樣檔(%)	母體檔(%)	p-value
投保金額			**0.9454**
<=22,800	74.16	74.15	
22,801-28,800	5.68	5.69	
28,801-36,300	6.37	6.36	
36,301-45,800	6.62	6.61	
45,801-57,800	3.00	3.00	
57,801-72,800	2.34	2.35	
72,801-87,600	0.89	0.90	
87,601-110,100	0.57	0.58	
110,101-150,000	0.36	0.36	
>150,000	0.00	0.00	
投保地區			**0.9992**
台北	36.10	36.10	
北區	15.10	15.11	
中區	18.24	18.24	
南區	13.76	13.77	
高屏	14.61	14.60	
東區	2.19	2.19	
身分屬性			**0.9857**
第一類第一目	2.57	2.57	
第一類第二目	27.70	27.74	
第一類第三目	1.44	1.44	
第一類第四目	0.00	0.00	
第一類第五目	0.00	0.00	
第二類第一目	11.30	11.32	
第二類第二目	0.00	0.00	
第三類第一目	6.82	6.82	
第三類第二目	1.37	1.37	
第五類(社福機構)	0.03	0.03	
第五類(公所)	1.18	1.18	
第六類第一目	1.74	1.73	
第六類第二目	9.31	9.28	
其他	36.54	36.51	

每年出生人數分布

檢驗抽樣母群體與 200 萬樣本的每年出生人數分布，年度分布自 1893 年至 2010 年，以卡方檢定分析，p-value 為 0.9899，表示兩分布之差異未達統計上顯著。

表 2.死因統計檔(H_OST_DEATH)類別變項驗證結果

變項	抽樣檔(%)	母體檔(%)	p-value
前十大死因			
99 年			**0.9544**
惡性腫瘤	28.67	28.54	
心臟性疾病(高血壓性疾病除外)	10.78	10.85	
腦血管疾病	7.06	7.04	
肺炎	6.33	6.18	
糖尿病	5.50	5.70	
事故傷害	4.73	4.62	
慢性下呼吸道疾病	3.72	3.61	
慢性肝病及肝硬化	3.41	3.42	
高血壓性疾病	2.97	2.87	
腎炎、腎症候群及腎病變	2.95	2.84	
前十大死因			
100 年			**0.9462**
惡性腫瘤	28.93	28.35	
心臟性疾病(高血壓性疾病除外)	10.50	10.89	
腦血管疾病	7.00	7.16	
糖尿病	5.99	6.04	
肺炎	6.14	6.02	
事故傷害	4.37	4.37	
慢性下呼吸道疾病	4.03	3.98	
慢性肝病及肝硬化	3.31	3.39	
高血壓性疾病	3.03	3.05	
腎炎、腎症候群及腎病變	2.91	2.89	
101 年			**0.7581**
惡性腫瘤	28.27	28.61	
心臟性疾病(高血壓性疾病除外)	11.19	11.20	
腦血管疾病	6.98	7.26	
肺炎	6.49	6.12	
糖尿病	6.39	6.11	
事故傷害	4.31	4.35	
慢性下呼吸道疾病	4.25	4.17	
高血壓性疾病	3.25	3.28	
慢性肝病及肝硬化	3.27	3.24	
腎炎、腎症候群及腎病變	2.84	2.85	

變項	抽樣檔(%)	母體檔(%)	p-value
102 年			**0.7032**
惡性腫瘤	29.09	29.21	
心臟性疾病(高血壓性疾病除外)	11.81	11.52	
腦血管疾病	7.63	7.40	
糖尿病	6.17	6.21	
肺炎	6.15	5.92	
事故傷害	4.02	4.20	
慢性下呼吸道疾病	3.89	3.92	
高血壓性疾病	3.29	3.30	
慢性肝病及肝硬化	2.90	3.14	
腎炎、腎症候群及腎病變	3.08	2.95	
死亡年齡分布			
99 年			**0.3156**
0-9 歲	0.51	0.50	
10-19 歲	0.46	0.51	
20-29 歲	1.29	1.28	
30-39 歲	3.01	3.01	
40-49 歲	7.02	7.04	
50-59 歲	11.78	12.20	
60-69 歲	13.53	13.35	
70-79 歲	22.50	23.31	
80 歲以上	39.89	38.80	
100 年			**0.6060**
0-9 歲	0.24	0.28	
10-19 歲	0.45	0.48	
20-29 歲	1.14	1.16	
30-39 歲	2.79	2.92	
40-49 歲	6.29	6.70	
50-59 歲	11.97	11.69	
60-69 歲	13.51	13.49	
70-79 歲	22.81	22.94	
80 歲以上	40.80	40.33	
101 年			**0.5557**
0-9 歲	0.15	0.19	
10-19 歲	0.40	0.47	
20-29 歲	1.13	1.06	
30-39 歲	2.84	2.82	
40-49 歲	6.53	6.47	
50-59 歲	11.79	11.33	
60-69 歲	13.79	13.84	
70-79 歲	21.82	22.37	
80 歲以上	41.53	41.46	

102 年			**0.3480**
0-9 歲	0.11	0.12	
10-19 歲	0.54	0.45	
20-29 歲	0.86	0.92	
30-39 歲	2.54	2.55	
40-49 歲	6.25	6.03	
50-59 歲	11.06	11.57	
60-69 歲	13.86	14.21	
70-79 歲	21.91	21.96	
80 歲以上	42.87	42.19	

(二)連續變項

表 3.全民健保承保檔(H_NHI_ENROL)連續變項驗證結果

變項	KS*	KSa*	D*	Pr>KSa*
投保金額(不分組)	0.000138	0.683916	0.000504	**0.7378**

*KS : the Kolmogorov-Smirnov statistic
KSa : the asymptotic Kolmogorov-Smirnov statistic, where KSa =$\sqrt{n}$KS
D : the two-sample Kolmogorov statistic
Pr>KSa : the asymptotic p-value for KSa

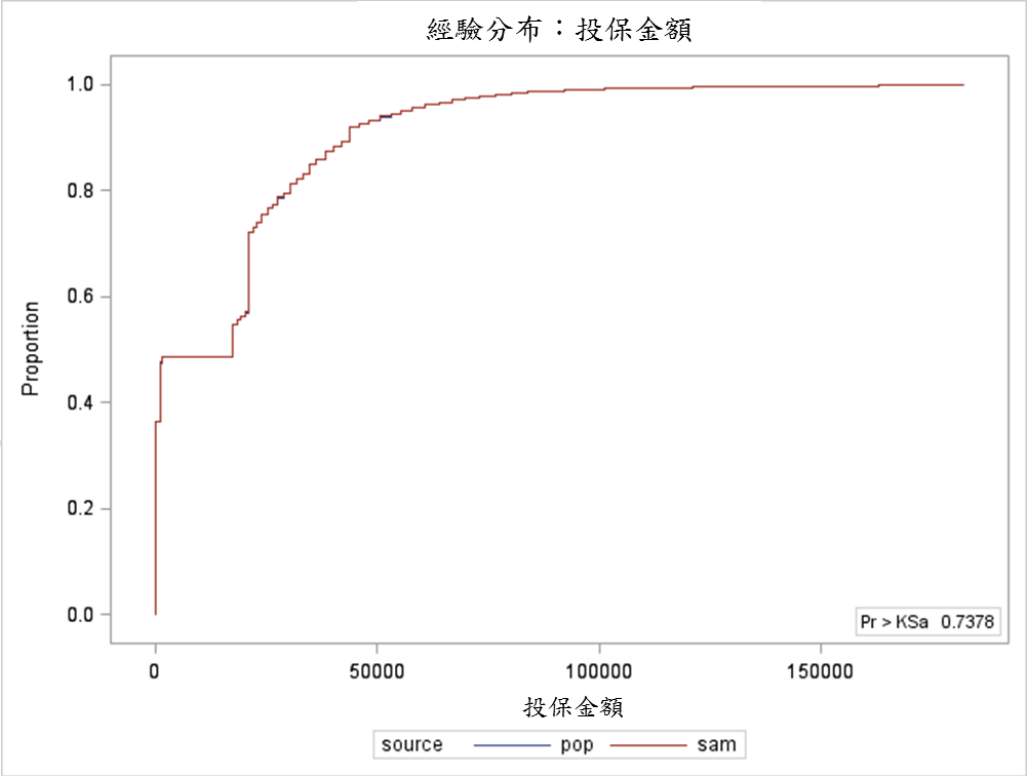

圖 1 投保金額的經驗分布圖

表 4. 全民健保處方及治療明細檔—西醫住院(H_NHI_IPDTE)連續變項驗證結果**

變項	KS*	KSa*	D*	Pr>KSa*
疾病就診年齡				
內分泌失調	0.009009	0.867911	0.032798	**0.4385**

高血壓疾病	0.005195	0.685489	0.01908	**0.7352**
呼吸道感染	0.003379	0.872206	0.012275	**0.4322**
消化系統之其他部位疾病	0.001992	1.145867	0.007271	**0.1447**
泌尿系統之疾病	0.001288	0.547414	0.004683	**0.9254**

*KS : the Kolmogorov-Smirnov statistic
KSa : the asymptotic Kolmogorov-Smirnov statistic, where KSa =$\sqrt{n}$KS
D : the two-sample Kolmogorov statistic
Pr>KSa : the asymptotic p-value for KSa
**本表採用衛生福利部統計處 99 年醫療統計年報中就診率統計，取就診率前五大疾病進行年齡驗證

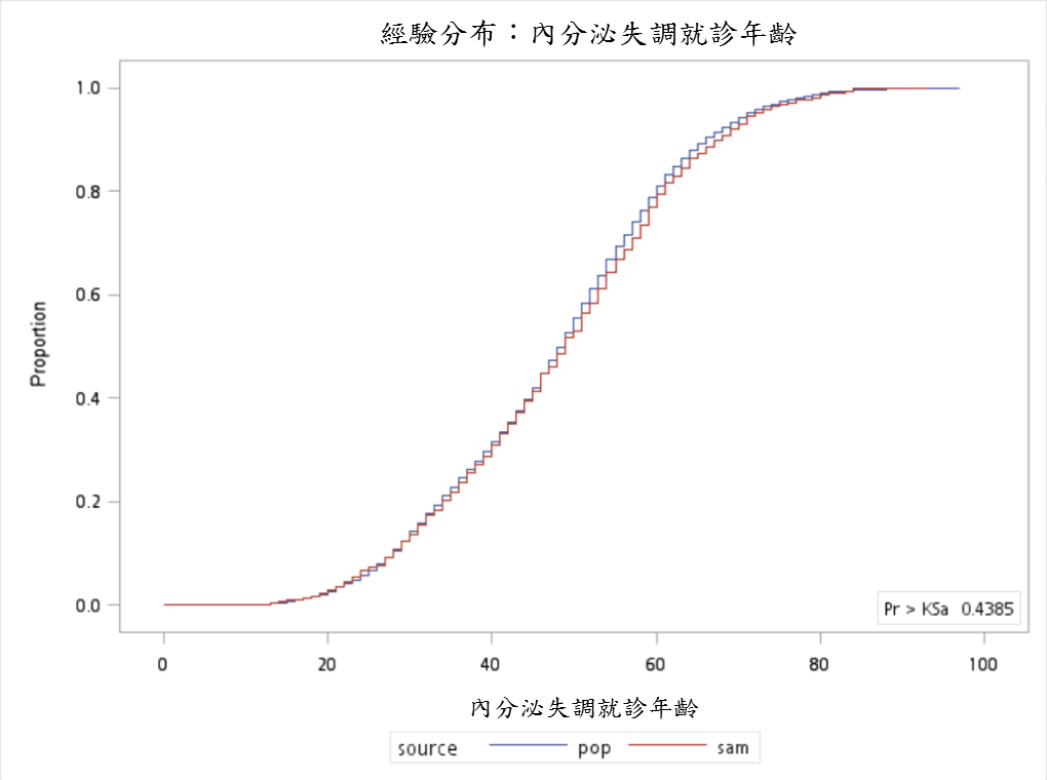

圖 2 內分泌失調就診年齡的經驗分布圖

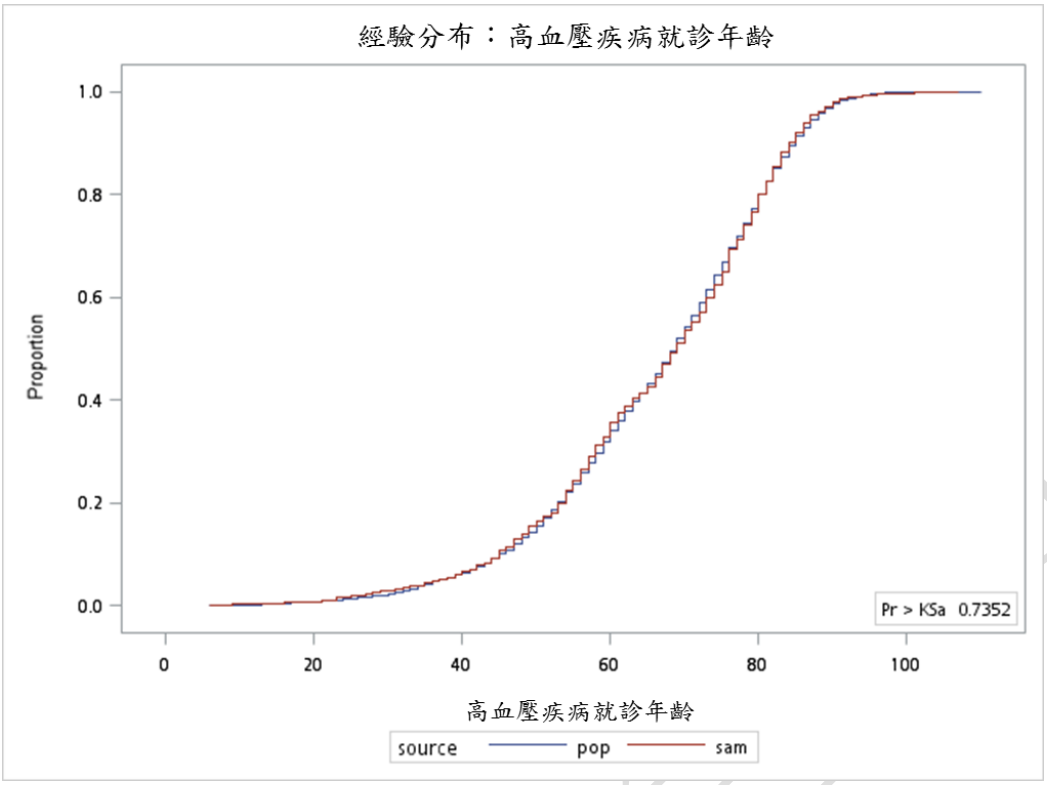

圖 3 高血壓疾病就診年齡的經驗分布圖

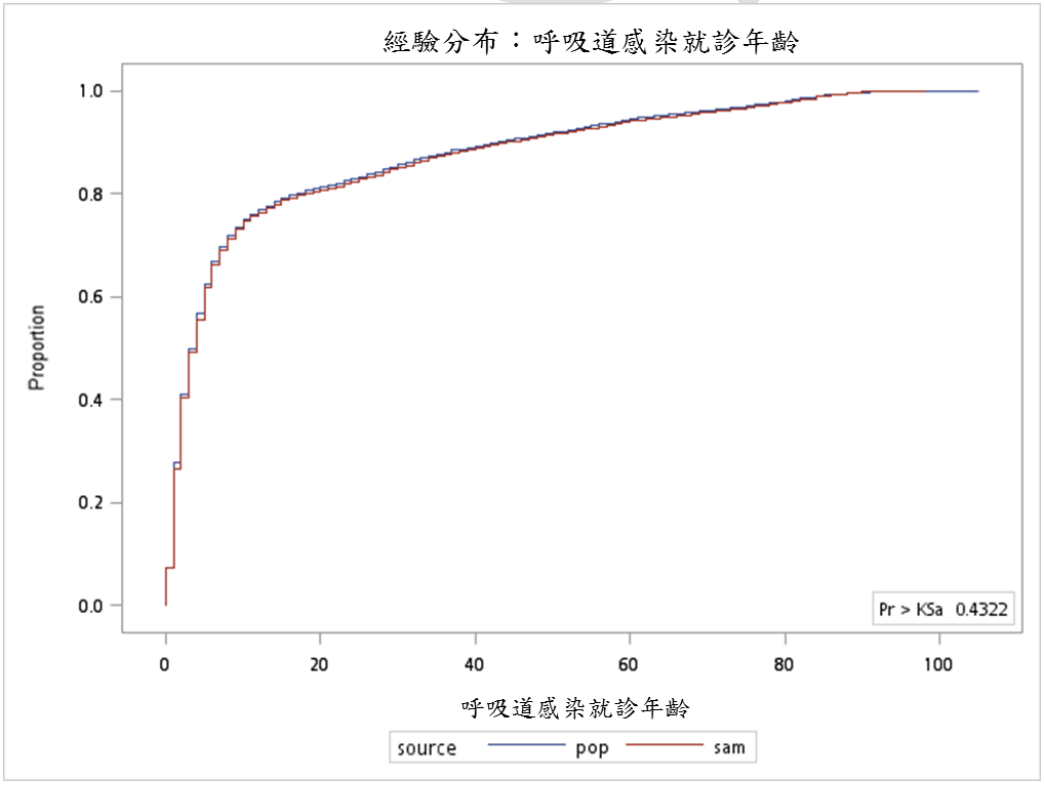

圖 4 呼吸道感染就診年齡的經驗分布圖

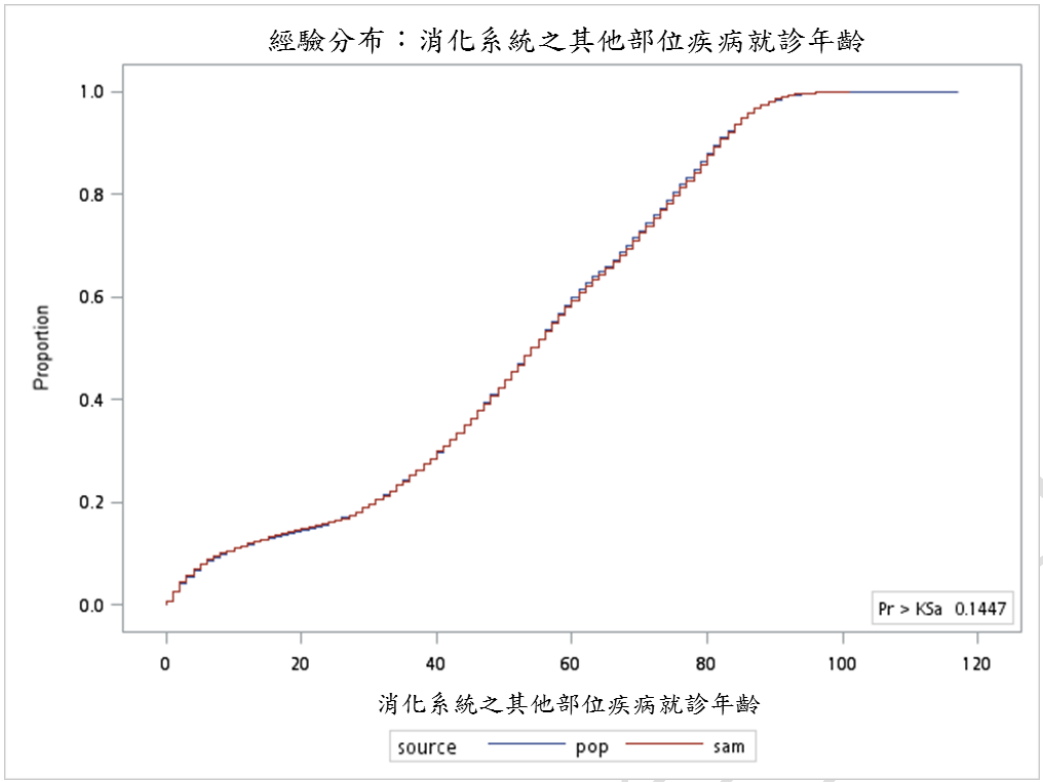

圖 5 消化系統之其他部位疾病就診年齡的經驗分布圖

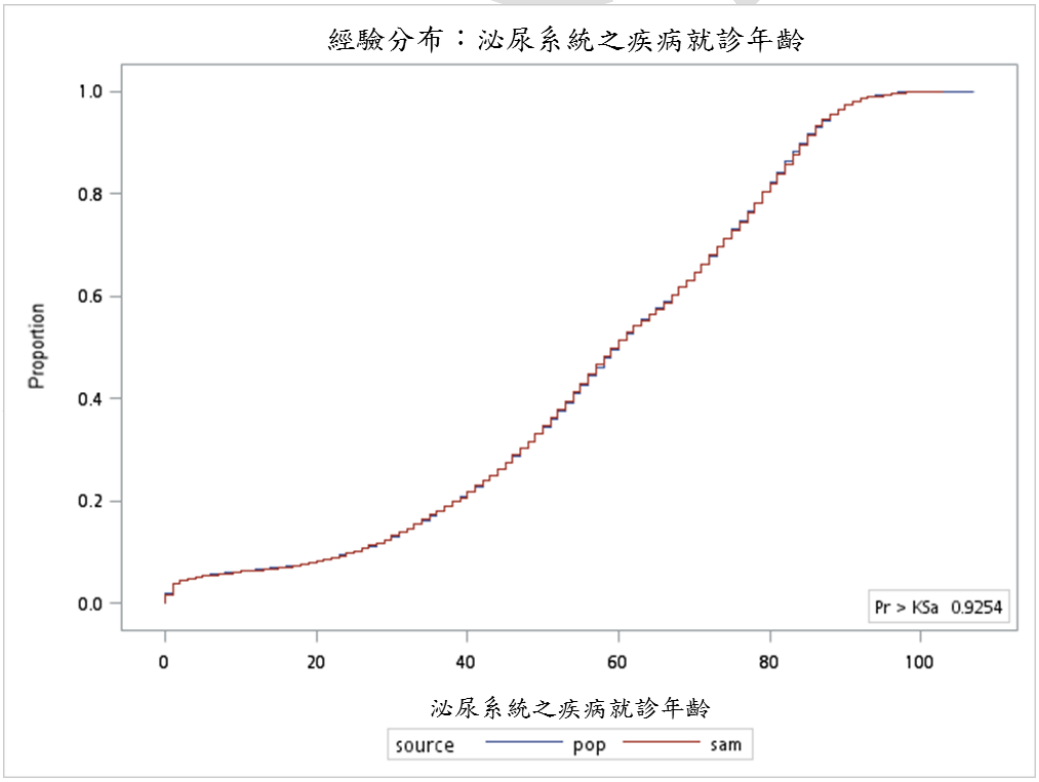

圖 6 泌尿系統之疾病診年齡的經驗分布圖

表 5.死因統計檔(H_OST_DEATH)連續變項驗證結果

變項	KS*	KSa*	D*	Pr>KSa*
死亡年齡				
99 年死亡年齡	0.003099	1.208082	0.011303	**0.1080**
100 年死亡年齡	0.002101	0.831088	0.007722	**0.4945**
101 年死亡年齡	0.001892	0.765503	0.006865	**0.6011**
102 年死亡年齡	0.002755	1.114662	0.01006	**0.1666**

*KS : the Kolmogorov-Smirnov statistic
KSa : the asymptotic Kolmogorov-Smirnov statistic, where KSa =$\sqrt{n}$KS
D : the two-sample Kolmogorov statistic
Pr>KSa : the asymptotic p-value for KSa

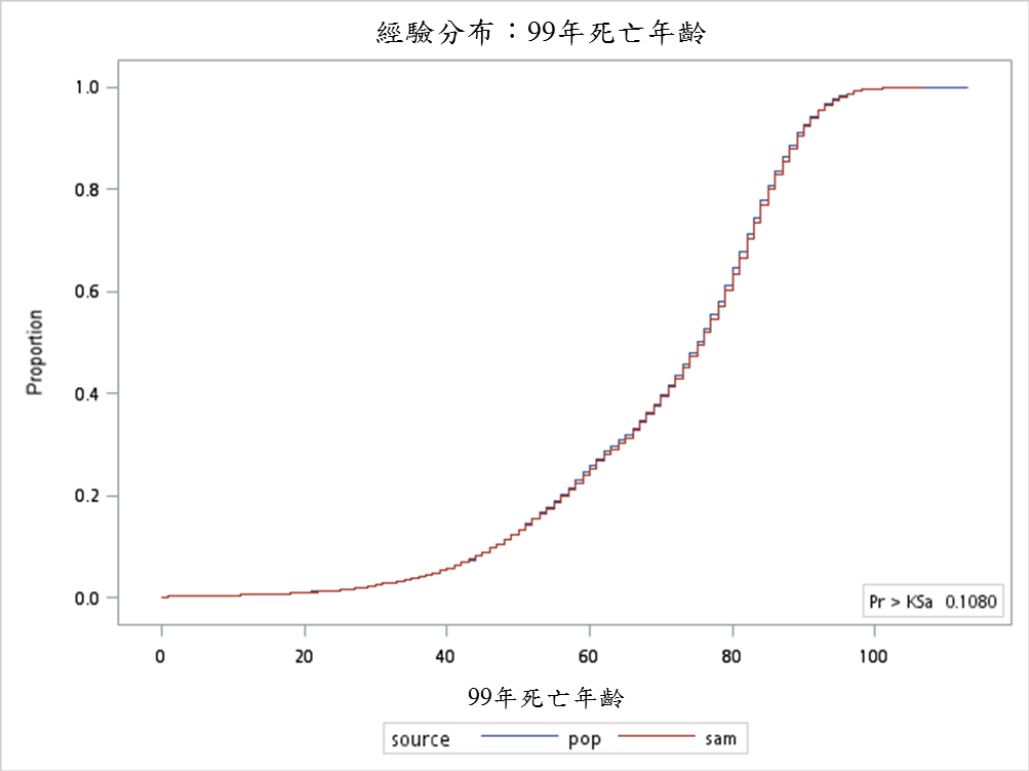

圖 7 99 年死亡年齡的經驗分布圖

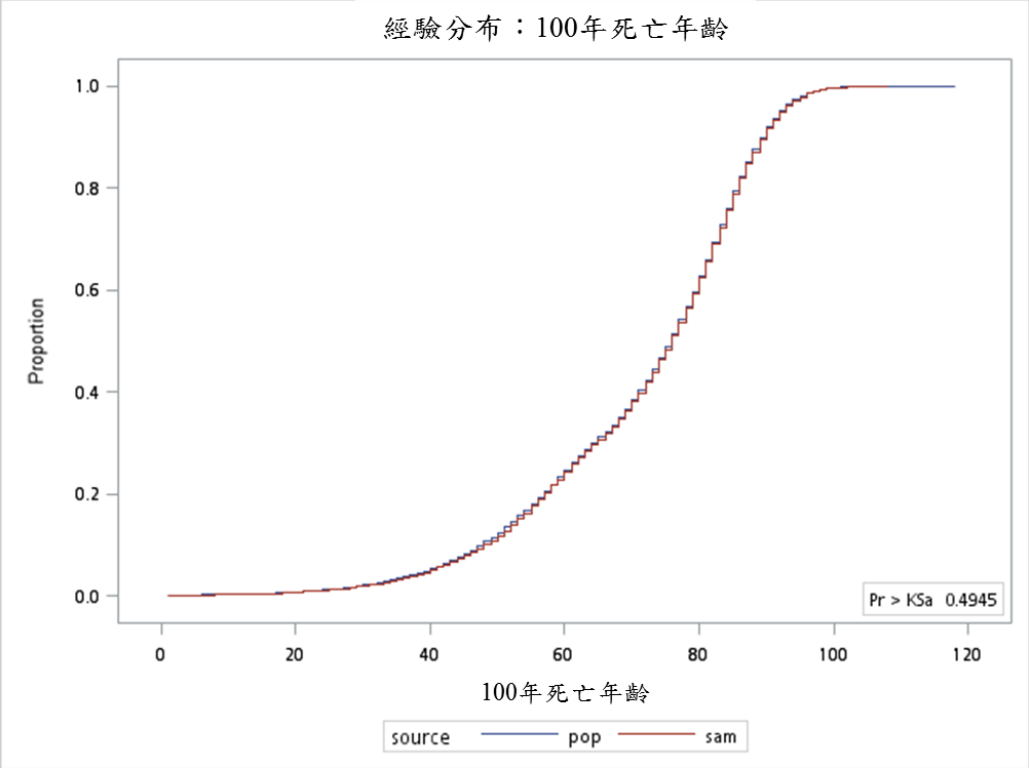

圖 8 100 年死亡年齡的經驗分布圖

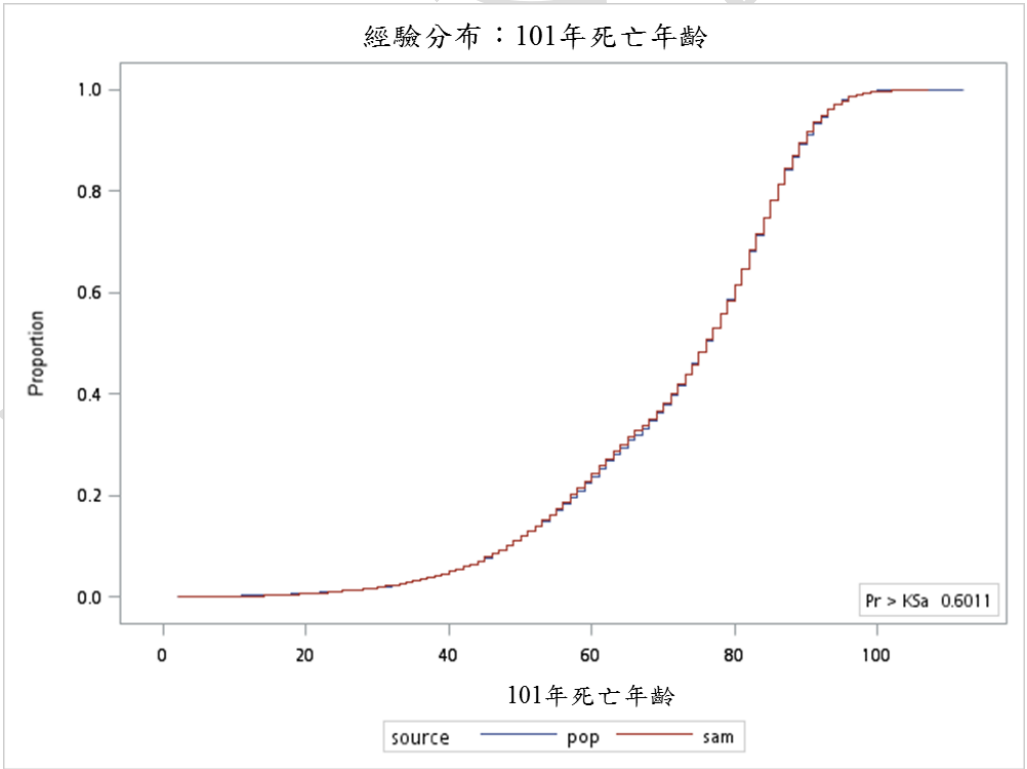

圖 9 101 年死亡年齡的經驗分布圖

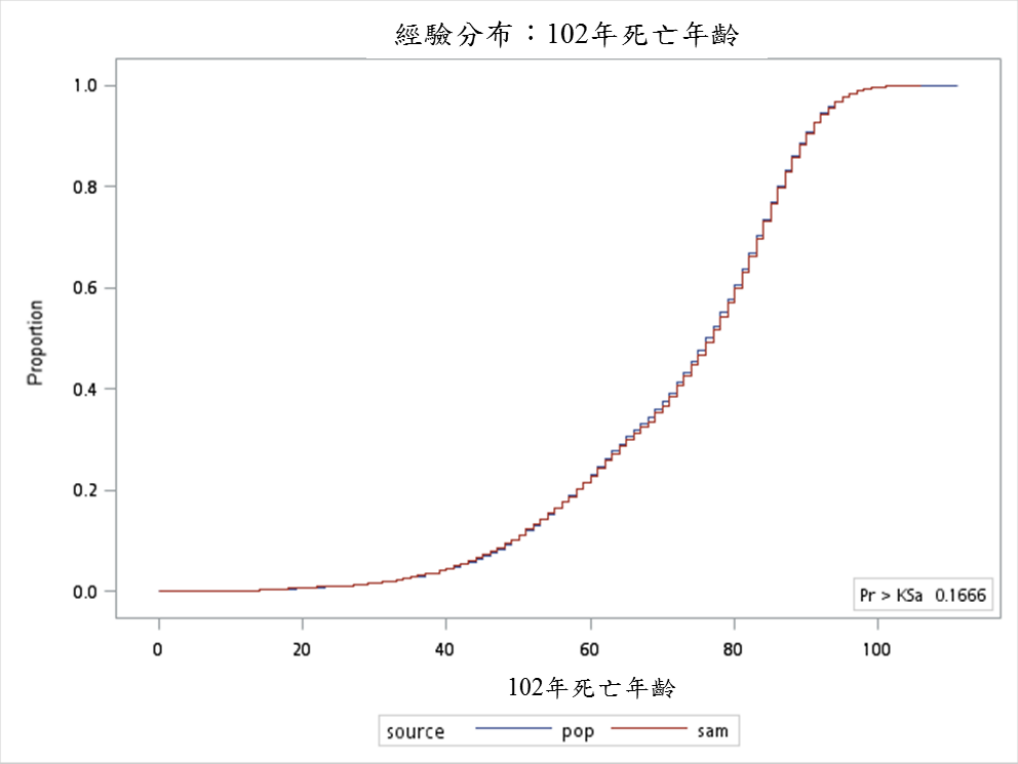

圖 10 102 年死亡年齡的經驗分布圖