第 1、2 組 200 萬人抽樣檔介紹：

衛生福利部**衛生福利資料科學中心**在 100 年度開放申請使用資料，但各檔案資料筆數及欄位繁多、資料使用不易，對於較少使用相關資料庫經驗的使用者而言，有相當的困難，爰本中心開始提供 200 萬人之就醫及死因資料供使用者申請，以資料起始年分爲兩種，一爲 **89 年**開始往後 10 年之資料，一爲 **94 年**開始往後 5 年之資料，每次申請直接提供健保資料及死因資料之常用欄位。

使用 200 萬人抽樣檔之優點如下：

1. 資料量較小，程式執行時間短，較快獲得結果。
2. 無須選擇使用檔案和勾選欄位，申請手續較快。
3. 不用等待資料篩選的時間。
4. 收費便宜。
5. 可以用抽樣檔的結果做爲使用全人口資料之參考。

另外，使用本檔案須注意以下幾點：

1. 因爲資料起始年之後不會再補充新增人口，所以**新生兒**的資料只有起始年才有。
2. 僅提供**常用欄位**，需使用常用欄位以外之欄位需另外申請及計費。

200 萬人抽樣檔抽樣方法：

1. 將衛生福利部統計處整理的 89 年和 94 年之**個人屬性檔**經檢誤後(將重複資料、性別不詳、生日不詳、年齡不合邏輯、地區不明者剔除)，以該檔爲抽樣母體。
2. 將每筆個人資料之**性別**、**年齡**、**地區**分層：性別分爲男女兩層；年齡分 20 層，每 5 歲一層到 85 歲以上，唯 5 歲以下再分出 28 天以下、28 天至 1 歲、1 歲至 5 歲三層；地區以健保分局分層，共 6 層。總共分爲 **240(=2×20×6)**層。
3. 計算每種分層在母體中的比例，以這個比例計算出各層在 200 萬人中會有多少人，作爲該層之抽樣數。
4. 在各分層中抽出該層的抽樣數，抽樣方法爲**隨機抽樣**。
5. 將全部的資料垂直合併，得到 200 萬人抽樣檔之個人屬性檔，89 年檔有 2,000,118 人，94 年檔有 2,000,120 人。
6. 再以個人屬性檔以**身分證字號**比對出包含這些人的資料檔，包括**健保門診明細檔**(H_HNI_OPDTE)、**健保門診醫令檔**(H_HNI_OPDTO)、**健保住院明細檔**(H_HNI_IPDTE)、**健保住院醫令檔**(H_HNI_IPDTO)、**健保藥局明細檔**(H_HNI_DRUGE)、**健保藥局醫令檔**(H_HNI_DRUGO)、**健保承保檔**(H_NHI_ENROL)、**死因檔**(H_OST_DEATH)等資料檔。

200 萬人抽樣檔常用欄位：

200 萬人抽樣檔僅提供每個檔案之常用欄位，各檔常用欄位如下表：

檔案名稱	常用欄位			
健保門診明細檔 H_NHI_OPDTE	費用年月	FEE_YM	國際疾病分類號三	ICD9CM_3
	申報類別	APPL_TYPE	主手術代碼	ICD_OP_CODE1
	申報日期	APPL_DATE	合計點數	T_DOT
	案件分類	CASE_TYPE	部分負擔點數	PART_DOT
	流水號	SEQ_NO	申請點數	T_APPL_DOT
	就醫科別	FUNC_TYPE	身分證字號	ID
	就醫日期	FUNC_DATE	性別	ID_S
	出生年月	BIRTH_YM	醫師身分證字號	PRSN_ID
	部分負擔代號	PART_NO	醫療機構代號	HOSP_ID
	國際疾病分類號一	ICD9CM_1	醫療機構縣市鄉鎮市區代碼	CITY
	國際疾病分類號二	ICD9CM_2	醫療機構權屬別	HOS
健保住院明細檔 H_NHI_IPDTE	費用年月	FEE_YM	主手術(處置)	ICD_OP_CODE1
	申報類別	APPL_TYPE	次手術(處置)一	ICD_OP_CODE2
	申報日期	APPL_DATE	次手術(處置)二	ICD_OP_CODE3
	案件分類	CASE_TYPE	次手術(處置)三	ICD_OP_CODE4
	流水號	SEQ_NO	次手術(處置)四	ICD_OP_CODE5
	出生年月	BIRTH_YM	葯點數	DRUG_DOT
	就醫科別	FUNC_TYPE	醫療點數	MED_DOT
	入院年月日	IN_DATE	部分負擔點數	PART_DOT
	出院年月日	OUT_DATE	申請費用點數	APPL_DOT
	急性病床天數	E_BED_DAY	部分負擔註記	PART_NO
	慢性病床天數	S_BED_DAY	身分證字號	ID
	主診斷代碼	ICD9CM_1	性別	ID_S
	次診斷代碼一	ICD9CM_2	醫療機構代號	HOSP_ID
	次診斷代碼二	ICD9CM_3	醫療機構縣市鄉鎮市區代碼	CITY
	次診斷代碼三	ICD9CM_4	醫療機構權屬別	HOS
	次診斷代碼四	ICD9CM_5		
健保藥局明細檔 H_NHI_DRUGE	費用年月	FEE_YM	合計點數	T_DOT
	申報類別	APPL_TYPE	身分證字號	ID
	申報日期	APPL_DATE	性別	ID_S
	案件分類	CASE_TYPE	醫師身分證字號	PRSN_ID
	流水號	SEQ_NO	調劑醫療機構代號	HOSP_ID
	就醫(處方)日期	FUNC_DATE	調劑醫療機構縣市鄉鎮市區代碼	CITY
	出生年月	BIRTH_YM	調劑醫療機構權屬別	HOS
	給藥日份	DRUG_DAY	原處方醫療機構代號	R_HOSP_ID

	用藥明細點數小計	DRUG_DOT		
健保門診醫令檔 H_NHI_OPDTO	費用年月	FEE_YM	單價	UNIT_P
	申報類別	APPL_TYPE	總量	TOTAL_Q
	申報日期	APPL_DATE	點數	TOTAL_DOT
	案件分類	CASE_TYPE	醫療機構代號	HOSP_ID
	流水號	SEQ_NO	醫療機構縣市鄉鎮市區代碼	CITY
	醫令類別	ORDER_TYPE	醫療機構權屬別	HOS
	藥品(項目)代號	DRUG_NO		
健保住院醫令檔 H_NHI_IPDTO	費用年月	FEE_YM	醫療機構權屬別	HOS
	申報類別	APPL_TYPE	醫令類別	ORDER_TYPE
	申報日期	APPL_DATE	醫令代碼	ORDER_CODE
	案件分類	CASE_TYPE	醫令數量	ORDER_Q
	流水號	SEQ_NO	醫令單價	ORDER_P
	醫療機構代號	HOSP_ID	醫令點數	ORDER_DOT
	醫療機構縣市鄉鎮市區代碼	CITY		
健保藥局醫令檔 H_NHI_DRUGO	費用年月	FEE_YM	醫療機構縣市鄉鎮市區代碼	CITY
	申報類別	APPL_TYPE	醫療機構權屬別	HOS
	申報日期	APPL_DATE	藥品代號	DRUG_NO
	案件分類	CASE_TYPE	單價	UNIT_P
	流水號	SEQ_NO	總量	TOTAL_Q
	醫療(調劑)機構代號	HOSP_ID	點數	TOTAL_DOT
健保承保檔 H_NHI_ENROL	單位屬性	ID1_UNIT	出生年月	ID_BIRTH_YM
	地區代號	ID1_CITY	身份別	ID1_TYPE
	個人身分證字號	ID	眷屬稱謂	ID_RELATION
	個人身分證字號性別	ID_S	投保金額	ID1_AMT
	被保險人身分證字號	ID1	身分屬性	ID1_IDENT
	被保險人身分證字號性別	ID1_S		
死因檔 H_OST_DEATH	身分證字號	ID	死亡種類	D_TYPE
	性別	SEX	婚姻狀況	MARRIAGE
	戶籍	COUNTY	死因分類	D_CODE
	出生年月	BIRTH_YM	外傷分類	TRAUMA
	死亡年月	D_DATE_YM	死因分類	ICD10
	死亡地點區域代碼	D_LOCA_CODE	新生兒日齡	NB_DAGE
	死亡場所	D_PLACE		
個人屬性檔 PERSON_DATA	身分證字號	ID	出生年月	BIRTH_YM
	性別	ID_S	地區別	CITY_CODE
	婚姻狀況	MARR	教育程度	EDU

200 萬人抽樣檔驗證：

說明：挑選出**健保門診明細檔**、**健保住院明細檔**、**承保檔**、**死因檔**，計算各檔案的內容之**次數分配(FREQ)**，比較以原始檔或是抽樣檔計算出來的結果是否相同；**死因檔**則用兩種方式驗證，第一種是比較原始檔和抽樣檔在抽樣年後一年的十大死因所占之死亡人數比例，第二種則是比較抽樣年之後每一年死亡人數占全人口之比例。每個計算結果再去計算 Root MSE 驗證原始檔和抽樣檔的差距。

方法：因抽樣時有做**性別**檢誤，所以各原始檔中也一樣有做性別檢誤，但其他條件如年齡檢誤等等則因較繁複但筆數不多而不做刪除動作。

先從原始檔和抽樣檔中需整理出來的資料，然後跑出**原始檔**該資料之次數分配，將次數分配**筆數小於 5 筆**或是**百分比小於 1%者**合併，再用一樣的次數分配去對照**抽樣檔**，並計算 Root MSE 來比較抽樣檔的次數分配是否相同。

死因檔的第一種驗證方法是以抽樣年的後一年，分別計算原始檔跟抽樣檔的**十大死因**，然後把各死因死亡人數除以總死亡人數去計算各死因所占比例，比較其差異；第二種驗證方法是以原始檔和抽樣檔之抽樣年母體檔，去計算**每年死亡人數**除以該年**年初存活人數**，比較其異同。

各檔案檢查的列表如下：

健保門診明細檔 H_NHI_OPDTE	**年齡**：以**出生年月**(BIRTH_YM)和**就醫日期**(FUNC_DATE)計算，用 10 歲分層。 **性別**：以**性別**(ID_S)計算。 **地區**：以**醫療院所所在地區**(CITY)計算，用健保分局分層。
健保住院明細檔 H_NHI_IPDTE	**年齡**：以**出生年月**(BIRTH_YM)和**住院日期**(IN_DATE)計算，用 10 歲分層。 **性別**：以**性別**(ID_S)計算。 **地區**：以**醫療院所所在地區**(CITY)計算，以健保分局分層。
健保承保檔 H_NHI_ENROL	**性別**：以**個人身分證字號性別**(ID_S)計算。 **地區**：以**組別**(ID1_DIVISION)計算，以健保分局分層。
死因檔 H_OST_DEATH	**十大死因** **每年死亡人數**

結果：

資料起始年爲 **89 年**之各檔案欄位檢定結果如下：

健保門診明細檔西醫(H_NHI_OPDTE)：

年齡		**0**	**1-9**	**10-19**	**20-29**	**30-3**	**40-49**	**50-59**	**60-69**	**70-79**	**80~**	**其他**
	抽樣檔比例	1.21	19.62	9.30	10.52	12.94	13.48	10.39	10.77	9.34	2.42	0.00
	原始檔比例	1.19	19.41	9.22	10.41	12.82	13.45	10.37	10.82	9.48	2.83	0.00
	Root MSE	0.02	0.21	0.08	0.12	0.12	0.03	0.03	0.05	0.14	0.40	0.00

性別		**男性**	**女性**
	抽樣檔比例	45.07	54.93
	原始檔比例	45.16	54.84
	Root MSE	0.09	0.09

地區		**台北**	**北區**	**中區**	**南區**	**高屏**	**東區**
	抽樣檔比例	29.90	13.96	20.74	15.38	17.69	2.34
	原始檔比例	29.92	13.99	20.65	15.39	17.68	2.37
	Root MSE	0.04	0.04	0.09	0.03	0.03	0.04

健保門診明細檔牙醫(H_NHI_OPDTE)：

年齡		**1-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**其他**
	抽樣檔比例	16.21	14.98	17.46	15.84	15.23	8.86	6.57	4.13	0.72
	原始檔比例	16.23	14.96	17.37	15.78	15.17	8.91	6.63	4.16	0.80
	Root MSE	0.03	0.03	0.09	0.07	0.07	0.05	0.07	0.03	0.08

性別		**男性**	**女性**
	抽樣檔比例	45.98	54.02
	原始檔比例	45.96	54.04
	Root MSE	0.04	0.04

地區		**台北**	**北區**	**中區**	**南區**	**高屏**	**東區**
	抽樣檔比例	35.55	12.91	21.34	12.99	15.30	1.91
	原始檔比例	35.46	12.97	21.33	12.92	15.42	1.91
	Root MSE	0.10	0.06	0.03	0.07	0.12	0.01

健保門診明細檔中醫(H_NHI_OPDTE)：

年齡		**1-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80~**	**其他**
	抽樣檔比例	8.53	11.58	15.68	20.26	19.05	10.58	7.99	5.14	1.10	0.09
	原始檔比例	8.54	11.62	15.54	20.15	19.15	10.51	7.99	5.20	1.21	0.09
	Root MSE	0.02	0.04	0.15	0.11	0.11	0.08	0.02	0.06	0.11	0.00

性別		**男性**	**女性**
	抽樣檔比例	41.99	58.01
	原始檔比例	42.15	57.85
	Root MSE	0.17	0.17

地區		台北	北區	中區	南區	高屏	東區
	抽樣檔比例	25.87	10.79	30.87	14.81	15.66	1.99
	原始檔比例	25.89	10.94	30.79	14.81	15.60	1.97
	Root MSE	0.03	0.14	0.09	0.03	0.06	0.02

健保住院明細檔(H_NHI_IPDTE)：

年齡		0	1-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80~	其他
	抽樣檔比例	3.16	10.18	4.71	14.31	15.07	12.36	9.97	12.27	13.24	4.73	0.00
	原始檔比例	2.96	9.36	4.36	13.11	13.94	12.01	9.97	12.63	14.86	6.80	0.00
	Root MSE	0.21	0.82	0.36	1.20	1.13	0.35	0.02	0.36	1.62	2.07	0.00

性別		男性	女性
	抽樣檔比例	49.68	50.32
	原始檔比例	50.62	49.38
	Root MSE	0.95	0.95

地區		台北	北區	中區	南區	高屏	東區
	抽樣檔比例	28.61	16.31	20.30	13.00	18.44	3.35
	原始檔比例	28.76	16.35	20.07	13.08	18.31	3.44
	Root MSE	0.15	0.06	0.23	0.09	0.14	0.08

健保承保檔(H_NHI_ENROL)

性別		男性	女性
	抽樣檔比例	50.43	49.57
	原始檔比例	50.44	49.56
	Root MSE	0.04	0.04

地區		台北	北區	中區	南區	高屏	東區
	抽樣檔比例	36.21	14.04	18.50	13.80	15.18	2.27
	原始檔比例	36.20	14.06	18.46	13.81	15.17	2.28
	Root MSE	0.03	0.03	0.04	0.03	0.03	0.02

死因檔(H_OST_DEATH)

十大死因	惡性腫瘤	腦血管疾病	心臟疾病	事故傷害	糖尿病	慢性肝病及肝硬化	腎炎、腎徵候群及腎性病變	肺炎	自殺	高血壓性疾病	其他
	26.05	10.37	8.69	7.51	7.19	4.14	3.20	2.96	2.20	1.39	26.30
	27.01	10.15	8.79	7.52	7.55	4.25	3.33	2.81	2.06	1.27	25.26
	0.96	0.23	0.11	0.02	0.36	0.11	0.13	0.15	0.14	0.12	1.04

每年死亡人數		90	91	92	93	94	95	96	97	98
	抽樣檔比例	0.52	0.54	0.57	0.59	0.61	0.61	0.63	0.65	0.66
	原始檔比例	0.52	0.54	0.56	0.58	0.61	0.61	0.64	0.66	0.66
	Root MSE	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01

資料起始年爲 **94 年**之各檔案欄位檢定結果如下：

健保門診明細檔西醫(H_NHI_OPDTE)：

年齡		**1-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80~**	**其他**
	抽樣檔比例	15.97	8.43	10.03	11.24	13.60	13.46	11.27	10.88	4.41	0.70
	原始檔比例	16.30	8.38	9.97	11.19	13.53	13.36	11.21	10.88	4.47	0.71
	Root MSE	0.33	0.06	0.06	0.06	0.08	0.10	0.06	0.02	0.06	0.02

性別		**男性**	**女性**
	抽樣檔比例	45.65	54.35
	原始檔比例	45.61	54.39
	Root MSE	0.05	0.05

地區		**台北**	**北區**	**中區**	**南區**	**高屏**	**東區**
	抽樣檔比例	30.56	13.54	20.36	15.29	17.78	2.46
	原始檔比例	30.53	13.54	20.38	15.30	17.78	2.47
	Root MSE	0.05	0.02	0.03	0.03	0.03	0.02

健保門診明細檔牙醫(H_NHI_OPDTE)：

年齡		**1-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80~**	**其他**
	抽樣檔比例	15.66	11.98	17.27	14.12	15.65	12.50	7.05	4.54	1.23	0.00
	原始檔比例	15.63	11.97	17.25	14.16	15.66	12.48	7.07	4.53	1.23	0.00
	Root MSE	0.03	0.02	0.03	0.05	0.03	0.03	0.02	0.02	0.01	0.00

性別		**男性**	**女性**
	抽樣檔比例	46.59	53.41
	原始檔比例	46.53	53.47
	Root MSE	0.07	0.07

地區		**台北**	**北區**	**中區**	**南區**	**高屏**	**東區**
	抽樣檔比例	34.99	12.96	20.83	13.29	15.86	2.08
	原始檔比例	35.08	12.93	20.85	13.20	15.85	2.10
	Root MSE	0.10	0.04	0.03	0.09	0.03	0.02

健保門診明細檔中醫(H_NHI_OPDTE)：

年齡		**1-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80~**	**其他**
	抽樣檔比例	6.60	9.89	15.64	17.99	20.03	14.45	8.22	5.59	1.54	0.04
	原始檔比例	6.67	9.89	15.67	17.97	20.04	14.33	8.19	5.61	1.59	0.04
	Root MSE	0.07	0.02	0.04	0.03	0.03	0.12	0.04	0.02	0.05	0.00

性別		**男性**	**女性**
	抽樣檔比例	40.20	59.80
	原始檔比例	40.17	59.83
	Root MSE	0.04	0.04

地區		**台北**	**北區**	**中區**	**南區**	**高屏**	**東區**
	抽樣檔比例	27.02	11.63	27.66	15.17	16.67	1.85
	原始檔比例	27.06	11.59	27.69	15.17	16.60	1.90
	Root MSE	0.04	0.05	0.05	0.03	0.07	0.05

健保住院明細檔(H_NHI_IPDTE)：

年齡		**0**	**1-9**	**10-19**	**20-29**	**30-39**	**40-49**	**50-59**	**60-69**	**70-79**	**80~**
	抽樣檔比例	1.34	9.26	3.41	10.59	12.36	12.67	12.82	12.44	15.38	9.74
	原始檔比例	1.36	9.59	3.40	10.70	12.20	12.53	12.67	12.10	15.23	10.22
	Root MSE	0.02	0.33	0.01	0.12	0.17	0.15	0.15	0.34	0.15	0.47

性別		男性	女性
	抽樣檔比例	51.87	48.13
	原始檔比例	52.16	47.84
	Root MSE	0.30	0.30

地區		台北	北區	中區	南區	高屏	東區
	抽樣檔比例	28.79	14.23	21.14	14.02	18.38	3.45
	原始檔比例	28.74	14.26	20.96	14.28	18.33	3.43
	Root MSE	0.06	0.04	0.17	0.26	0.06	0.03

健保承保檔(H_NHI_ENROL)

性別		男性	女性
	抽樣檔比例	49.84	50.16
	原始檔比例	49.89	50.11
	Root MSE	0.06	0.06

地區		台北	北區	中區	南區	高屏	東區
	抽樣檔比例	35.85	14.23	18.41	14.20	15.01	2.28
	原始檔比例	35.80	14.37	18.43	14.17	14.95	2.28
	Root MSE	0.06	0.14	0.03	0.04	0.07	0.01

死因檔(H_OST_DEATH)

十大死因	惡性腫瘤	腦血管疾病	心臟疾病	糖尿病	事故傷害	肺炎	慢性肝病及肝硬化	腎炎、腎徵候群及腎性病變	自殺	高血壓性疾病	其他
	30.00	9.94	9.70	7.65	6.32	4.26	3.99	3.72	3.48	1.43	26.14
	28.48	9.34	9.29	7.46	6.04	4.10	3.68	3.38	3.31	1.24	23.68
	1.52	0.60	0.41	0.19	0.28	0.16	0.31	0.34	0.17	0.19	2.46

每年死亡人數		**95**	**96**	**97**	**98**
	抽樣檔比例	0.58	0.60	0.62	0.62
	原始檔比例	0.57	0.59	0.62	0.63
	Root MSE	0.01	0.01	0.01	0.01