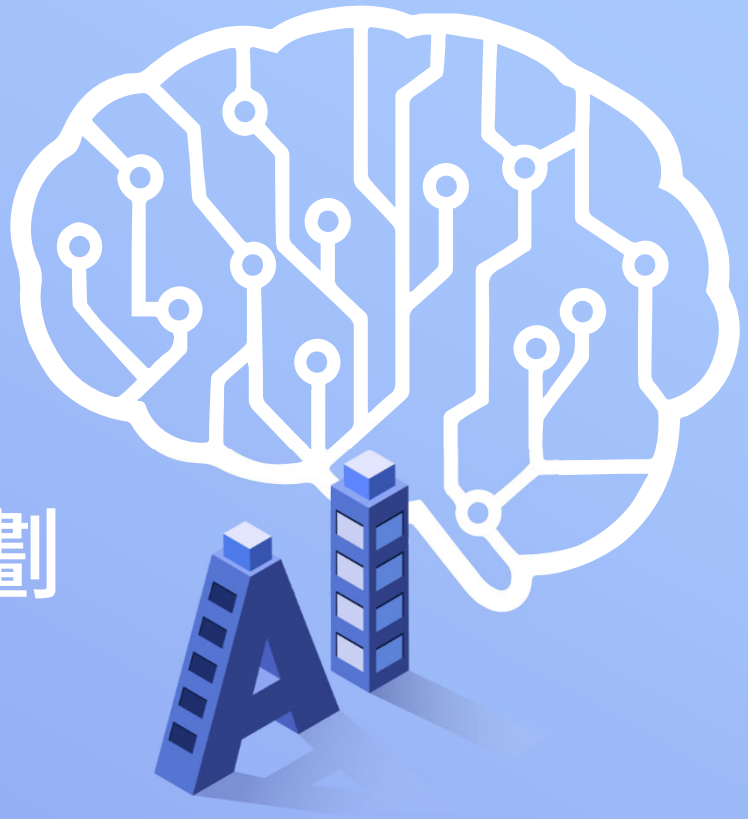
探討死因編碼運用AI技術之可行性

衛生福利部統計處
吳姿慧專員

大綱



01 現況及問題分析

1.1 死因統計目的

死亡證明書

(十一) 死亡原因：(儘量不要填寫症狀或死亡當時之身體狀況；如心臟衰竭、身體衰弱)

1. 直接引起死亡之疾病或傷害:

甲、敗血症

先行原因：(若有引起上述死因之疾病或傷害)

乙、(甲之原因) 肺炎

丙、(乙之原因) 腦梗塞 根本死因

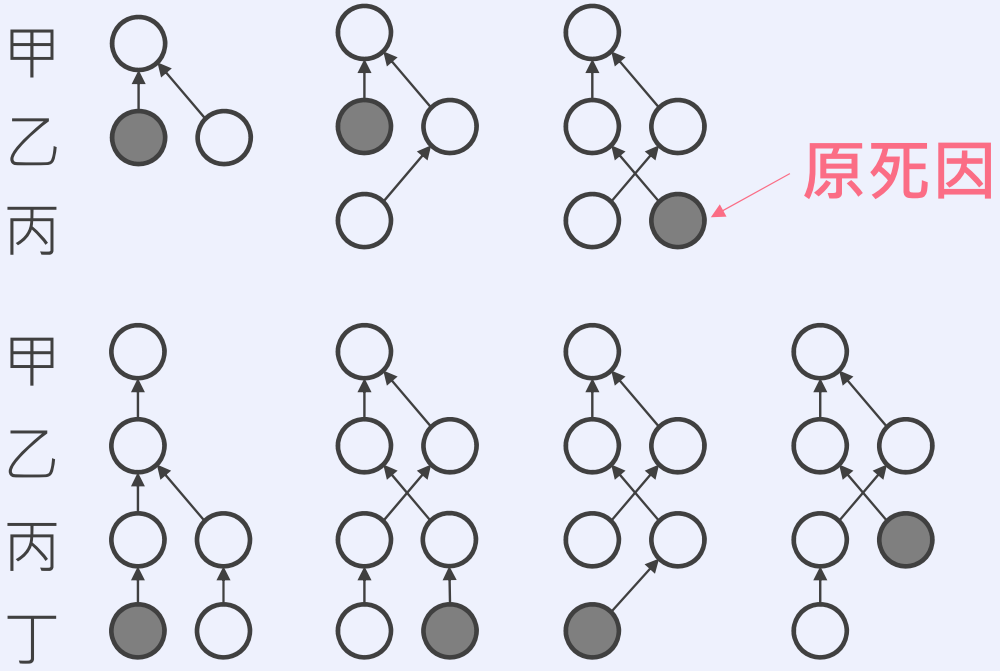
丁、(丙之原因)

2. 其他對於死亡有影響之疾病或身體狀況 (但與引起死亡之疾病或傷害無直接關係者)

- 世界衛生組織(WHO)建議死因填法為按死前疾病後、先，原則**一行只填一個診斷**
- 於預防醫學理念下，以國際死因判別軟體IRIS、ACME，找出「導致死亡的**原死因**」歸類統計
- 以「**原死因**」統計為國際比較基礎

1.2 原死因判断

第1部分



第2部分

第1部分

甲　**心臓懸停**

乙

丙

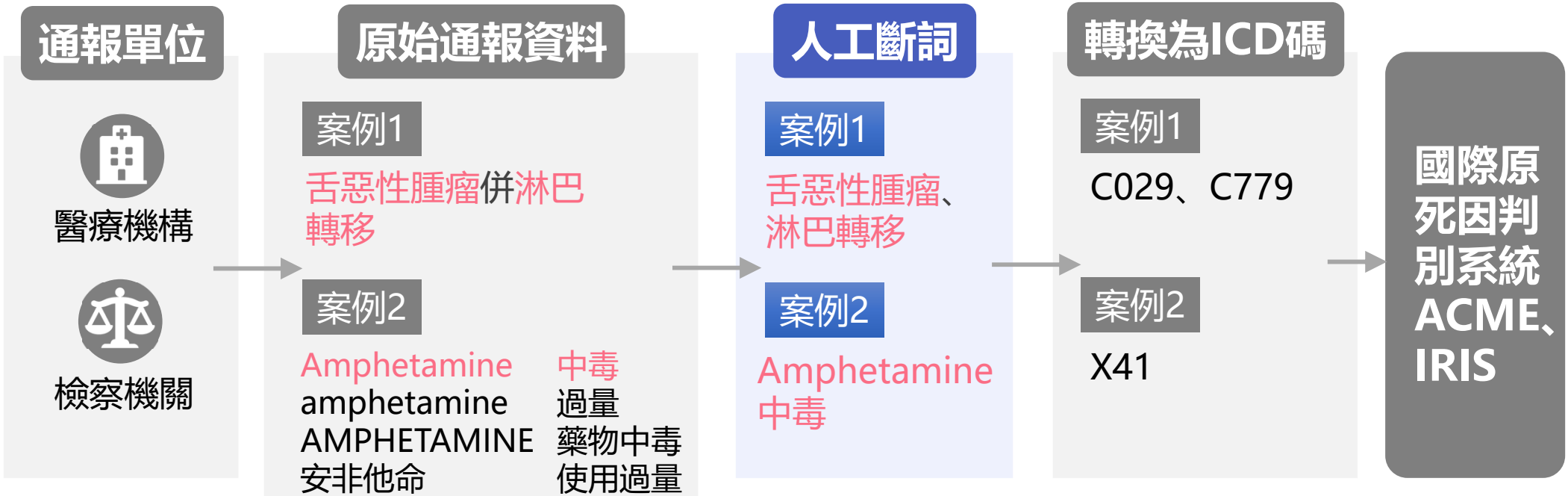
丁

原死因

第2部分　**腸繋膜栓塞**

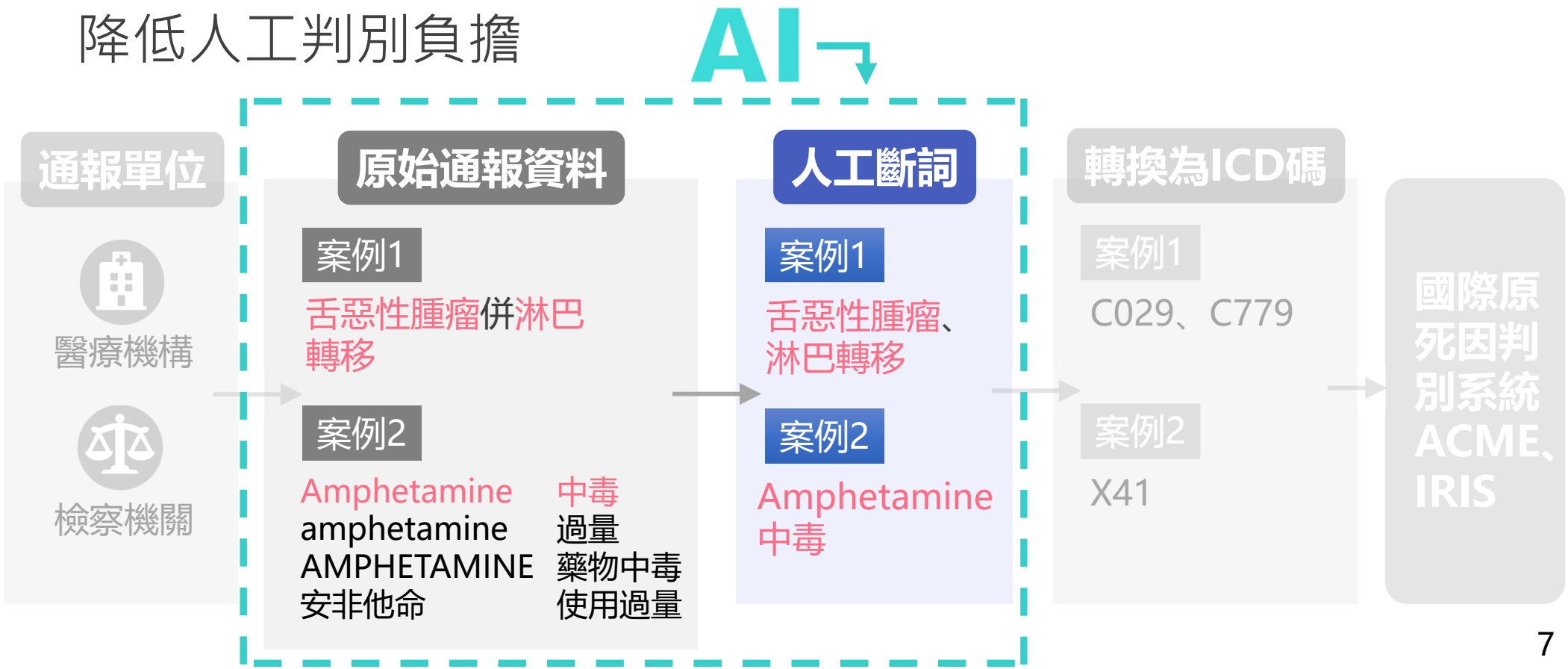
1.3 資料處理流程

- 原始通報資料多依個人詞彙行為填報，需將臨床診斷進行**疾病標準語、斷詞及詞語價值判別**，再轉換為國際死因分類碼(ICD)統計



1.4 本研究處理之核心問題

- 運用AI技術協助死因斷詞作業，增進死因註碼準確度，降低人工判別負擔



1.5 新冠肺炎診斷

227種

- WHO命名為COVID-19，本部公告為「**嚴重特殊傳染性肺炎(COVID-19)**」
- 醫師或法醫開具新冠肺炎診斷詞語計**227種**

02 研究架構

2.1 研究架構

1 現況分析

回溯死因診斷重複且低效率樣態，評估可用AI處理者

2 建置模型

建立「**死因自定義詞典**」，導入Jieba中文分詞、BERT中文預訓練、LSTM等模型

3 模型評估

利用111年死因資料進行**模型驗證**，探討AI與人工斷詞差異

4 優化模型

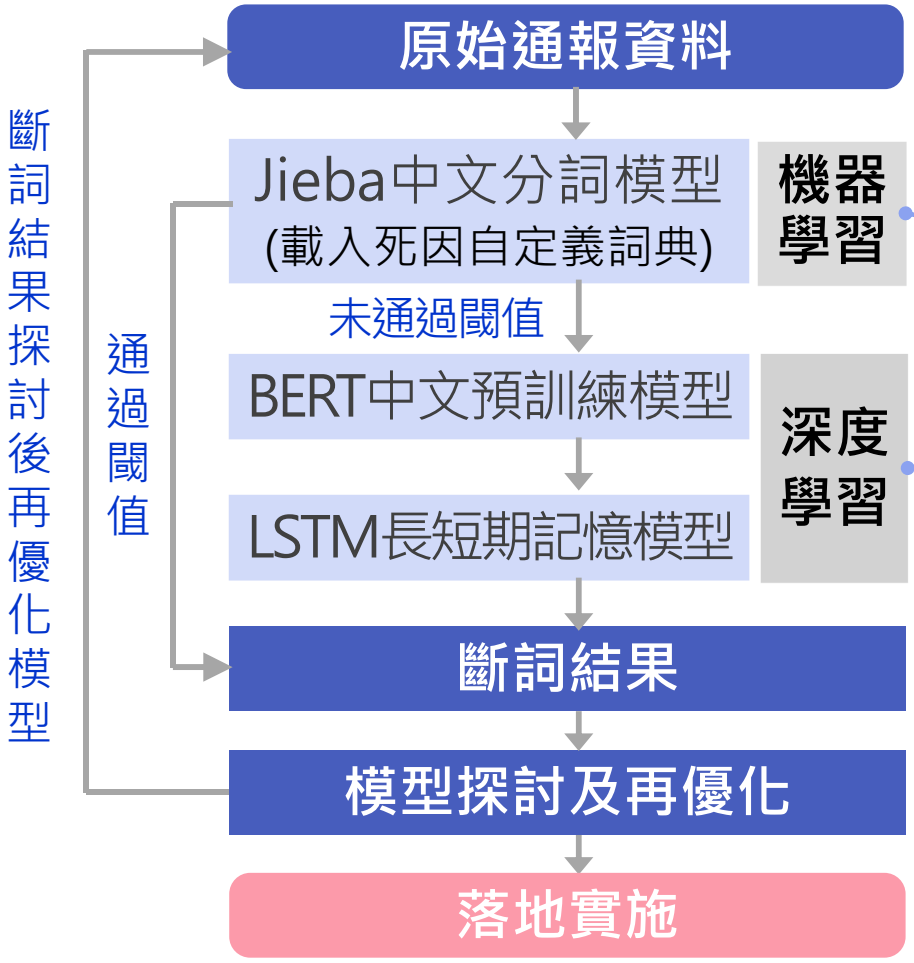
歸納6種不一致樣態，建置**關鍵字及停用字詞典**，進行模型優化升級

5 落地規劃

規劃AI技術落地實施策略，包括模型持續優化方向及實務操作步驟

2.2 研究流程

AI死因斷詞模型建置流程



人工智慧(AI)

讓電腦像人類一樣思考和行動

機器學習(ML)

利用過往資料找出規則

深度學習(DL)

模擬大腦神經網絡，從過往資料找出規則

03 建構模型

3.1 死因自定義詞典

- 蒐集110年以前醫師及法醫師診斷詞語與人工判詞對照資料，建立符合我國臨床醫學之**自定義詞典**，迄今已建置**46,822個**常用診斷詞語

個數

ICD代碼長度

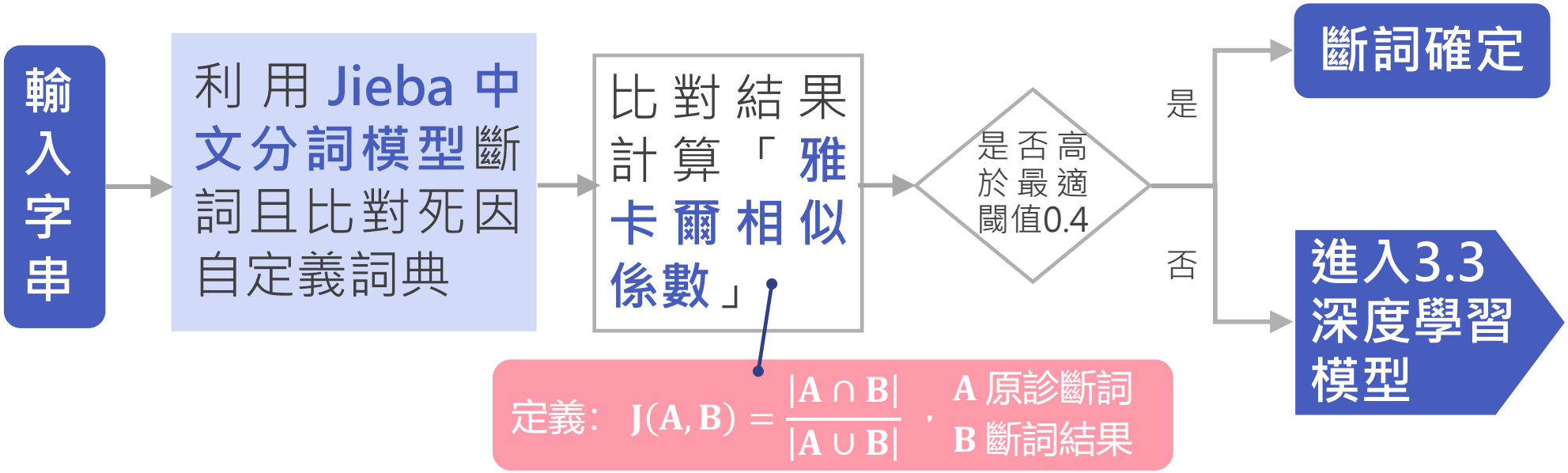
死因編碼 ≤ 4碼　　約8,800個

序號	死因疾病診斷	ICD碼	序號	死因疾病診斷	ICD碼
1	COVID-19	U071	10	新冠肺炎感染	U071
2	COVID-19疾管署發布確診	U071	11	新冠病毒	U071
3	COVID-19國外死亡	U071	12	新冠病毒肺炎	U071
4	SARS-CoV-2病毒感染	U071	13	新冠病毒感染	U071
5	冠狀病毒肺炎	U071	14	嚴重特殊傳染性肺炎	U071
6	第五類法定傳染病嚴重特殊傳染性肺炎	U071	15	嚴重特殊傳染性肺炎(COVID-19)	U071
7	第五類嚴重特殊性傳染病肺炎	U071		⋮	
8	新冠肺炎	U071			
9	新冠肺炎病毒感染	U071	46,822	霍亂	A009

3.2 Jieba中文分詞模型(1/3)

- **Jieba中文分詞模型：**

  利用**自定義詞典**與**Jieba中文分詞模型**進行死因診斷詞語處理，並計算「雅卡爾相似係數」篩選出最佳結果

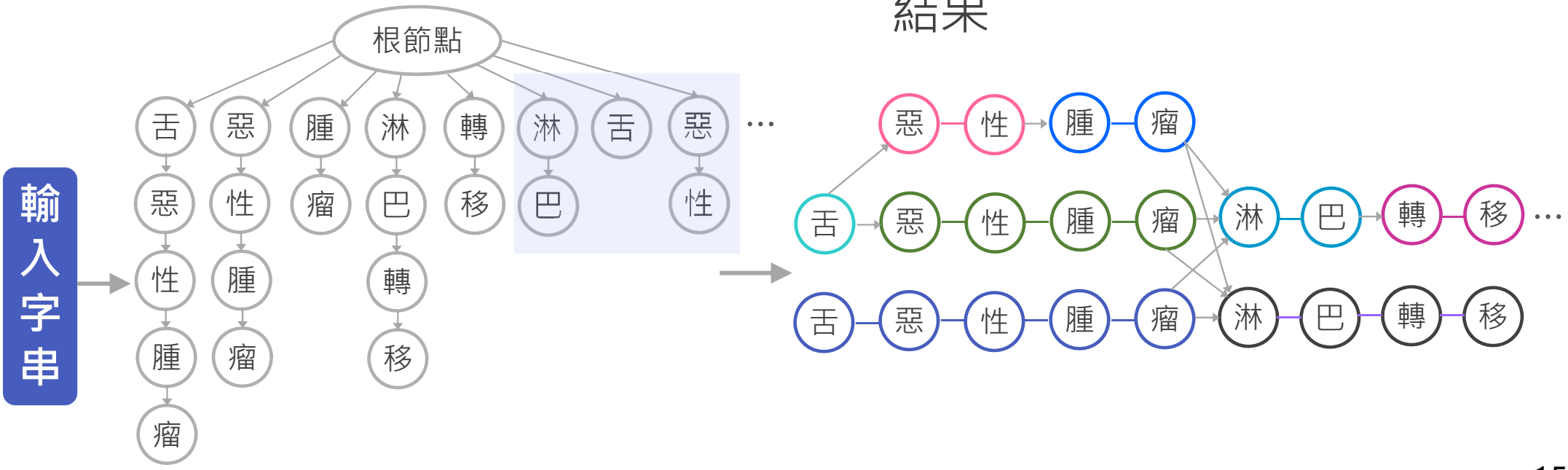


3.2 Jieba中文分詞模型(2/3)

- 以字串「**舌惡性腫瘤併淋巴轉移**」為例

載入Jieba斷詞庫及自定義詞典建立**字典樹(Trie)**

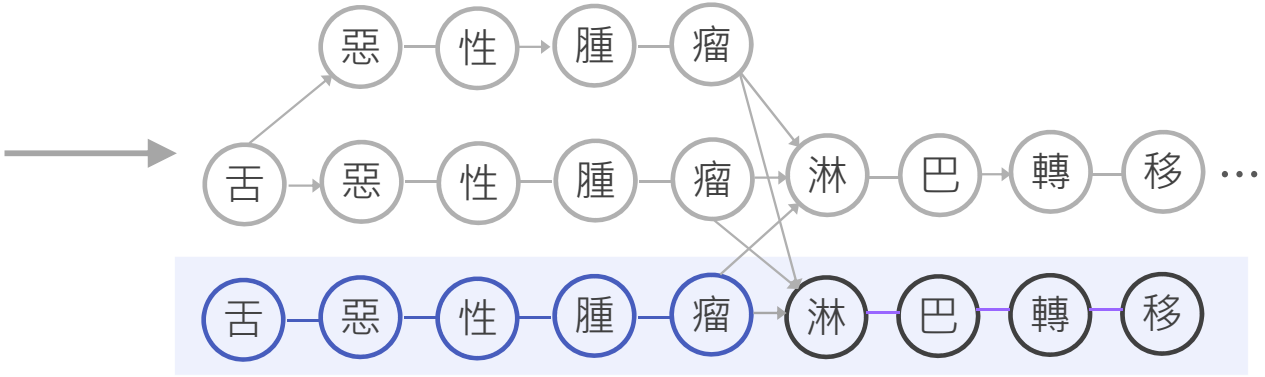
利用Trie建立**有向無環圖(DAG)**，走訪所有可能斷詞結果



3.2 Jieba中文分詞模型(3/3)

- 以字串「**舌惡性腫瘤併淋巴轉移**」為例

比對自定義詞典，找出匹配之**最長詞語最少病症**



斷詞結果

舌惡性腫瘤

淋巴轉移

- **雅卡爾相似係數 J(原診斷詞, 斷詞結果)**

$$= J(\text{舌惡性腫瘤併淋巴轉移}, \text{舌惡性腫瘤 淋巴轉移})$$

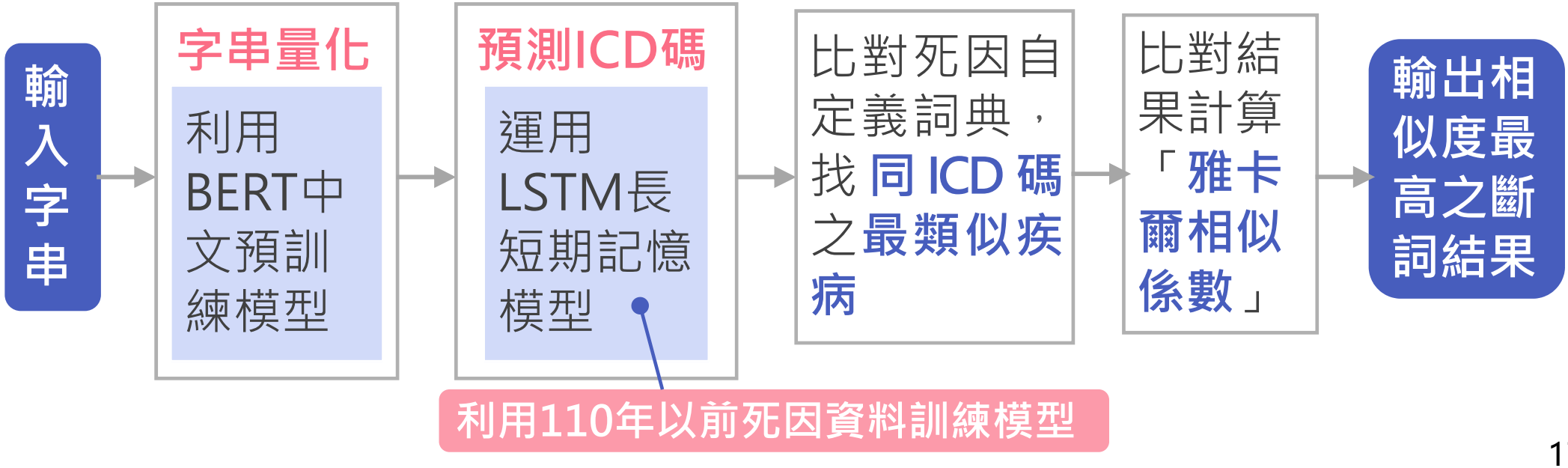
$$= \frac{|\text{舌惡性腫瘤併淋巴轉移} \cap \text{舌惡性腫瘤 淋巴轉移}|}{|\text{舌惡性腫瘤併淋巴轉移} \cup \text{舌惡性腫瘤 淋巴轉移}|} = \frac{9}{10} = 0.9$$

高於閾值0.4，輸出斷詞結果

3.3 深度學習模型(1/3)

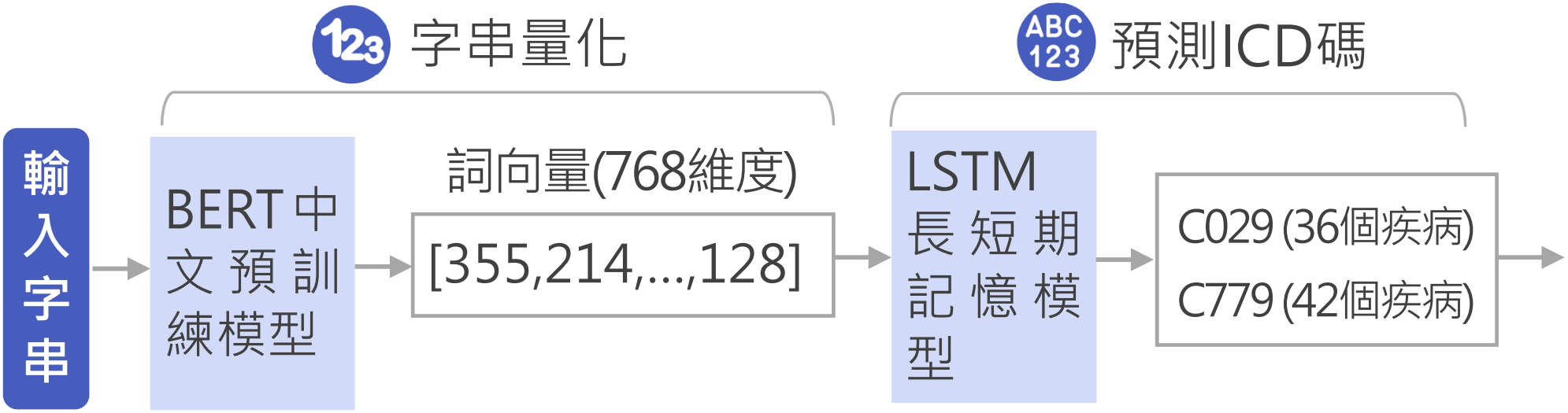
- **深度學習模型：**

  採用**BERT中文預訓練模型**將字串量化後，運用**LSTM長短期記憶模型**進行ICD編碼預測，利用相似度查找最佳可能斷詞



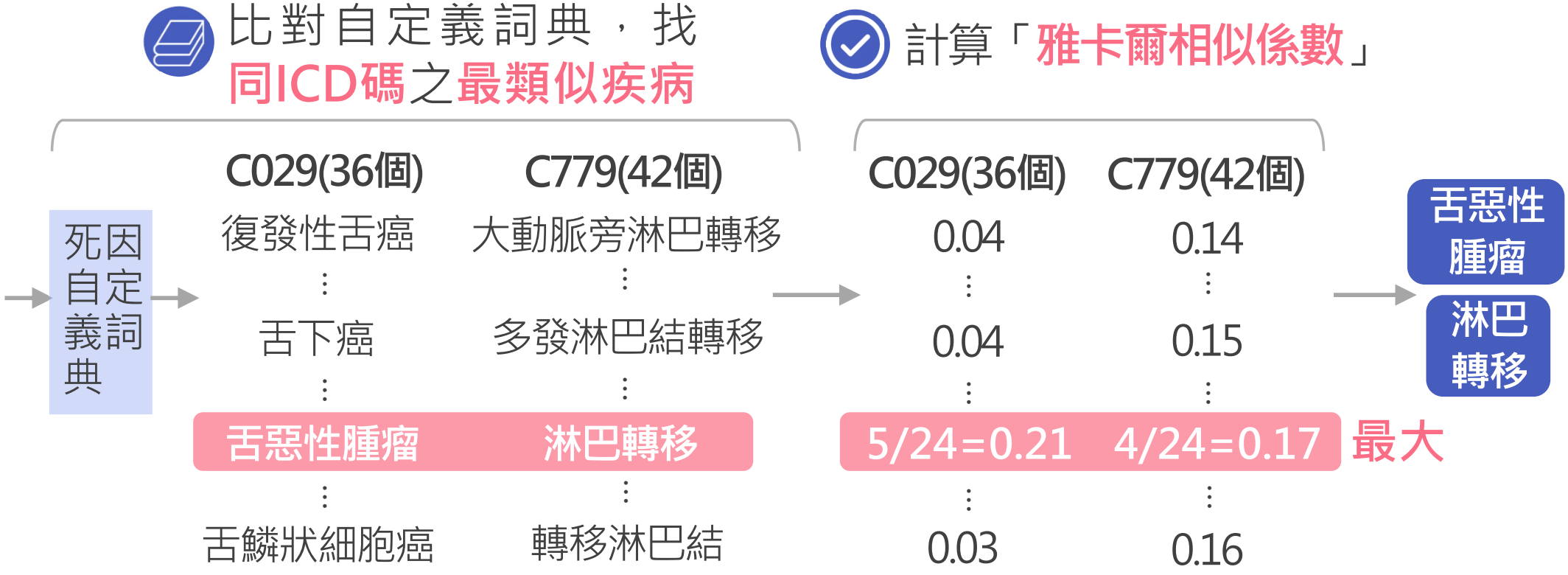
3.3 深度學習模型(2/3)

- 以字串「**舌惡性腫瘤第三期併淋巴轉移參考醫院診斷證明書開立**」為例，經Jieba中文分詞模型，計算相似係數為 $\frac{9}{24}$ = 0.38，低於閾值，故進入深度學習模型
- **深度學習模型：**



3.3 深度學習模型(3/3)

■ 以「**舌惡性腫瘤第三期併淋巴轉移參考醫院診斷證明書開立**」為例



04 模型評估及落地規劃

4.1 模型驗證

- 利用**111年死因資料**進行驗證，全面檢視19.3萬筆死因診斷詞語，探討AI與人工斷詞差異

83.2%

結果相同

16萬298筆

(完全相同15萬8,832筆，近似相同1,466筆)

16.8%

結果不同

3萬2,413筆

(不含近似相同1,466筆)

4.2 AI與人工斷詞差異

■ 整理AI與人工斷詞**差異樣態**計12種(3萬3,879筆)

00 結果近似相同 (1,466筆)	01 缺身體部位及修飾詞 (8,965筆)	02 未拆診斷 (5,744筆)
03 否定修飾詞 (194筆)	04 不當拆解 (155筆)	05 斷詞結果空白 (8,290筆)
06 新興疾病 (4,239筆)	07 須特殊處理 (2,430筆)	08 過度拆解 (1,891筆)
09 因果順序不同 (369筆)	10 腎臟病缺期別 (78筆)	11 早產兒資訊完整 (58筆)

4.3 再優化模型

00~05

建置
關鍵字詞典
停用字詞典

■歸納6種不一致樣態資料，建置**關鍵字詞典**及**停用字詞典**，提供成大進行模型優化升級

關鍵字詞典 (1,850個詞語)	停用字詞典 (565個詞語)
■ **身體部位及修飾詞**：女性右側乳房、瀰漫性巨大B細胞...等 ■ **未拆診斷**：及...出血、併...衰竭、...等 ■ **否定修飾詞**：COVID-19陰性、非創傷性顱內出血...等 ■ **不當拆解**：急性病毒性B型肝炎...等	■ **結果近似相同**：~~死亡~~、~~病史~~、~~呼吸器使用~~...等 ■ **斷詞結果為空白**：口腔~~部位~~腫瘤、肺~~腔~~積液、乙狀結腸~~惡性~~癌、~~老化性~~失智症...等

4.4 維持人工處理類型

06~11

特殊樣態維持人工處理

■ 如**交通事故**、**早產兒**等歧異性大之診斷，或**新興疾病**、**過度拆解**、**因果順序不同**、**腎臟病缺期別**等特殊樣態，維持人工處理

交通事故	早產兒	其他樣態
■ 需包括死者身份、車輛種類 ■ 例：車禍B2行人＊汽車	■ 需轉換妊娠週數、出生體重 ■ 例：早產兒<999克、早產兒少於28週	■ 新興疾病 ■ 過度拆解 ■ 因果順序不同 ■ 腎臟病缺期別

4.5 模型準確率

Jieba模型 53.2%

↓

Jieba+深度學習模型

83.2%

＋

模型再優化升級

→

扣除須人工斷詞類別

準確率(111年)

約96%

05 結語

5. 結語

01 Jieba斷詞成功**53.2%**，經深度學習模型，整體準確率**83.2%**，加入關鍵字及停用字詞典，扣除須人工斷詞類別，準確率可達**96%**

02 已完成**自定義詞典**；**關鍵字**及**停用字詞典**尚在測試歷年資料補齊中

03 研發初衷係為解決6名斷詞編碼之約僱人員出缺不補案，因行政院同意改以臨時人力遞補，故**人工與AI雙軌進行**

04 ICD死因碼以下拉式或鍵入關鍵字填報，恐影響死因填報確度，WHO不建議資料蒐集之初有任何系統或自動化引導醫師開立診斷

謝謝聆聽