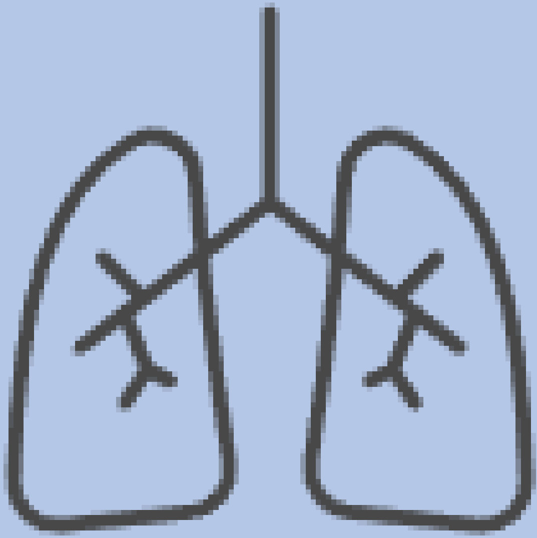
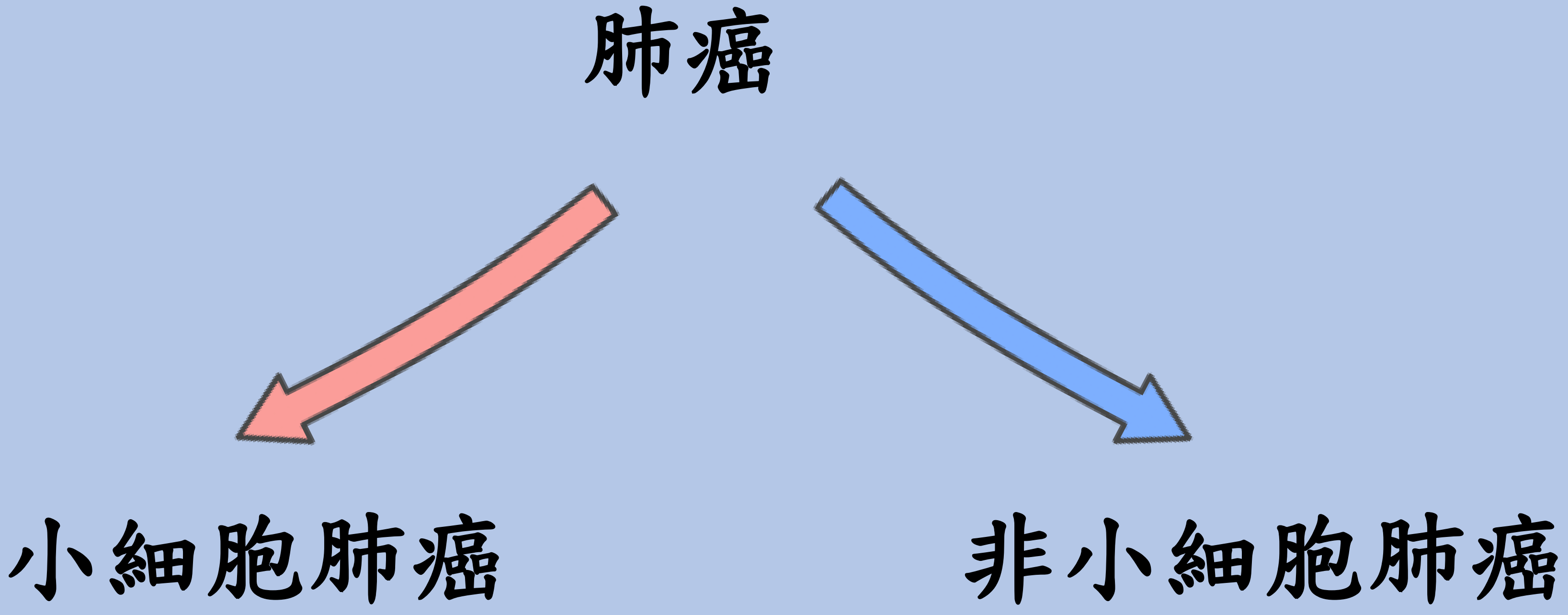
# 非小細胞肺癌五年存活率預測

科系 ： 大數據科技及管理研究所 碩一

姓名 ： 張哲安





肺癌

小細胞肺癌

非小細胞肺癌

其中非小細胞肺癌最為常見
約佔所有肺癌病例的85%

# 根據GLOBOCAN 2022年統計

全球男性最常被診斷的癌症
女性則居第二位

該年度全球新診斷肺癌病例達248萬人
死亡人數為123萬人，為所有癌症死因之首

# 衛生福利部國民健康署2021年癌症登記統計報告顯示

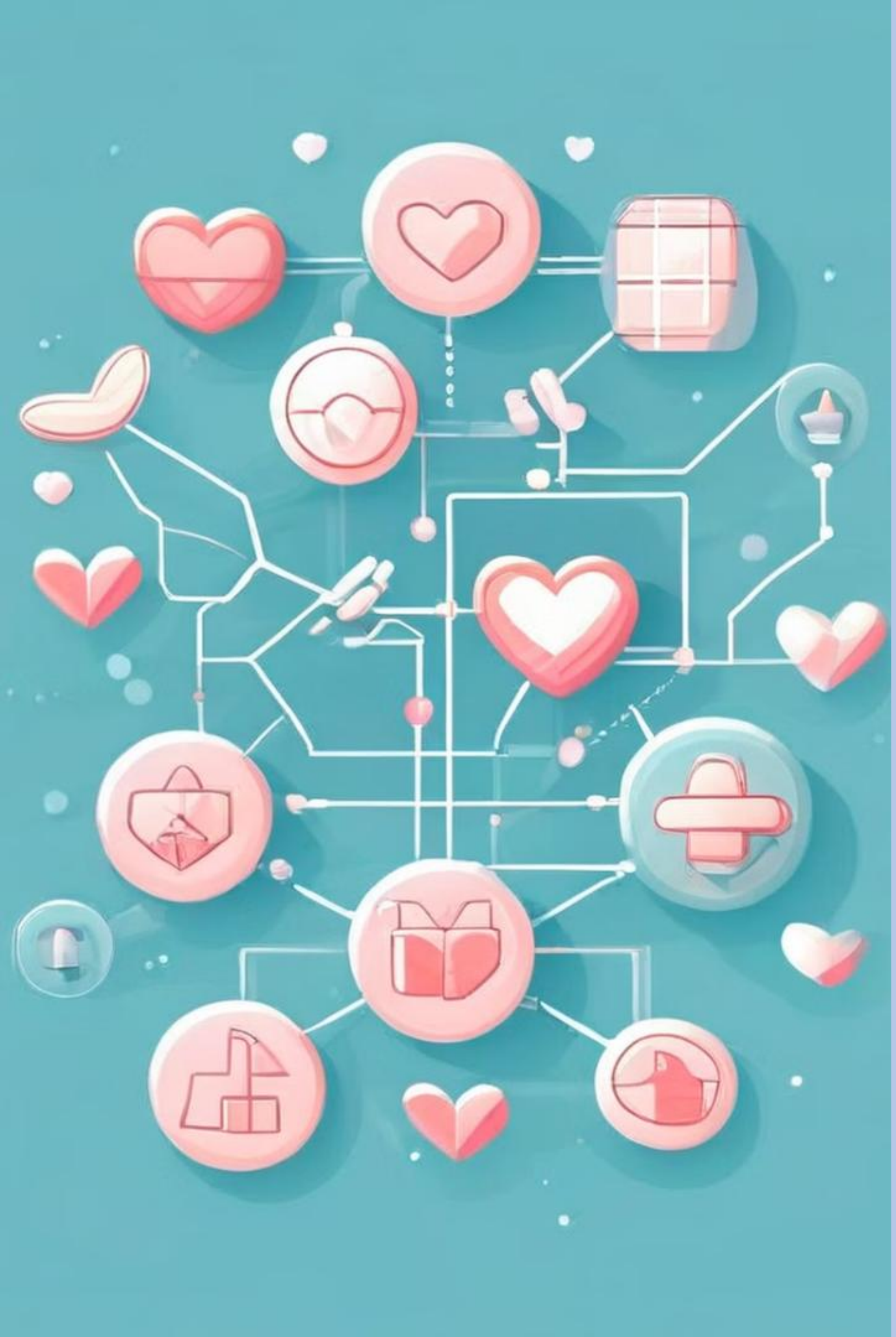
## 肺癌在新診斷癌症及死亡率均位居第一

近數十年

1 早期低劑量電腦斷層掃描篩檢

2 標靶治療等治療方式的進展

2年及5年存活率有所提升，但死亡率仍為首冠

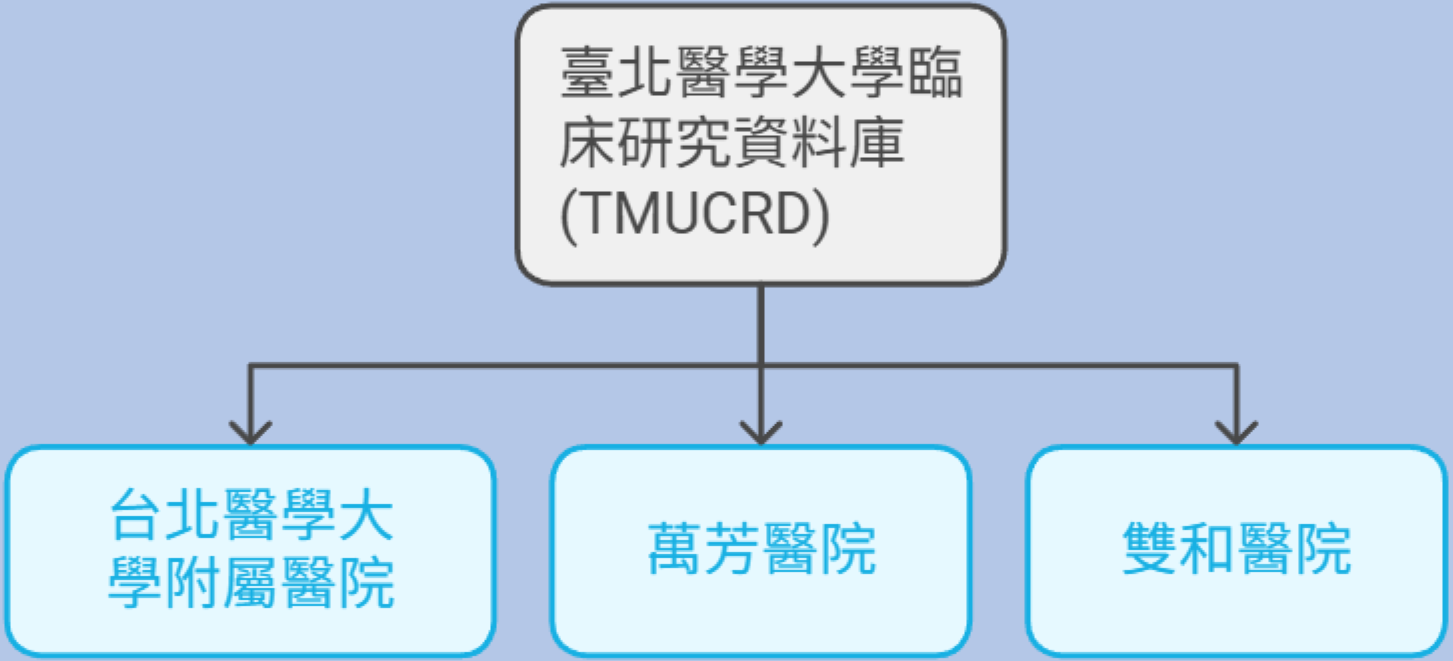


# ■ 建立存活預測模型

協助醫師提供有效的治療計畫建議

及早與病人及家屬進行安寧照
護的討論，以減輕病人的痛苦

# 資料來源



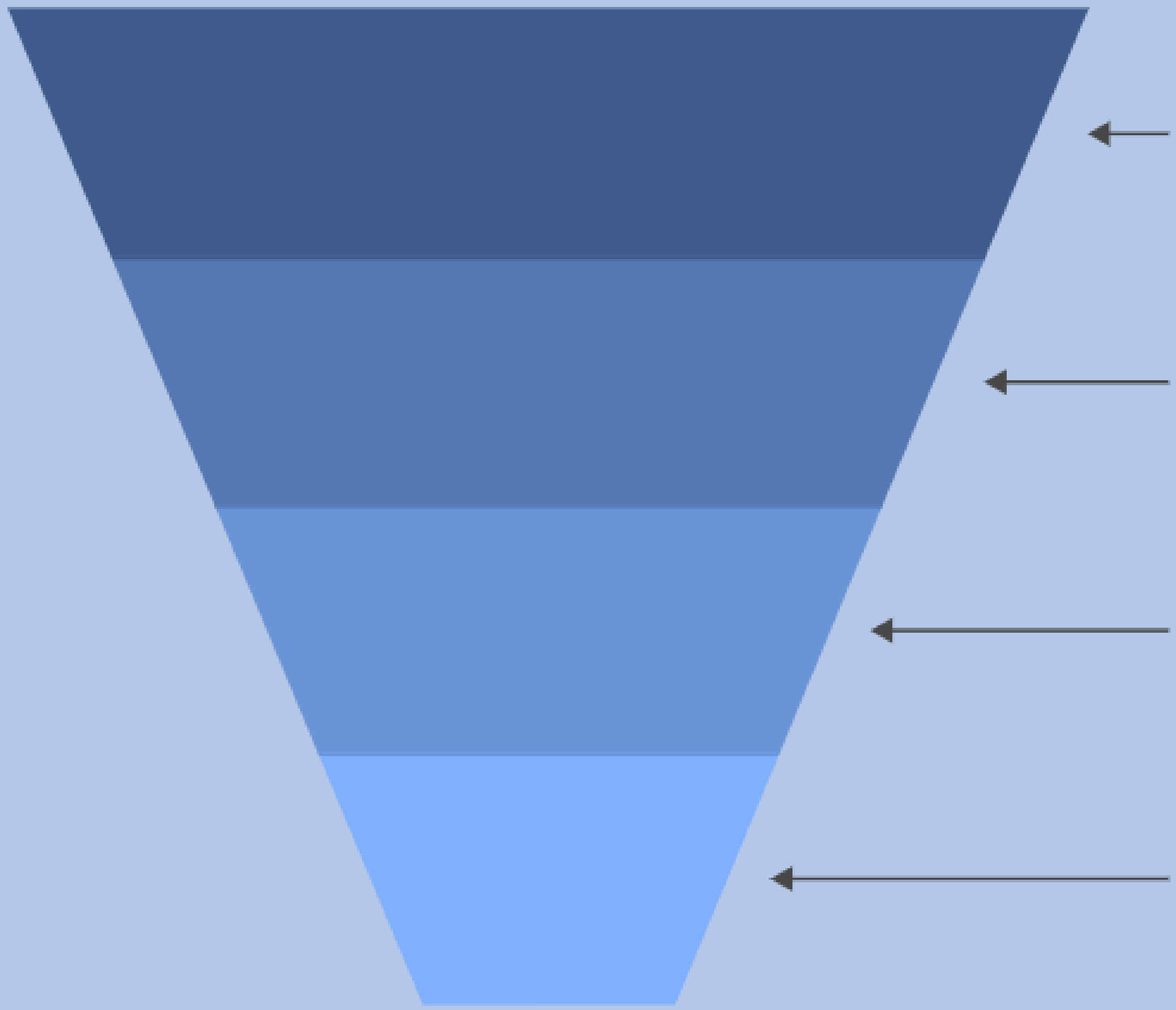
臺北醫學大學臨床研究資料庫(TMUCRD)

台北醫學大學附屬醫院

萬芳醫院

雙和醫院

# 納入



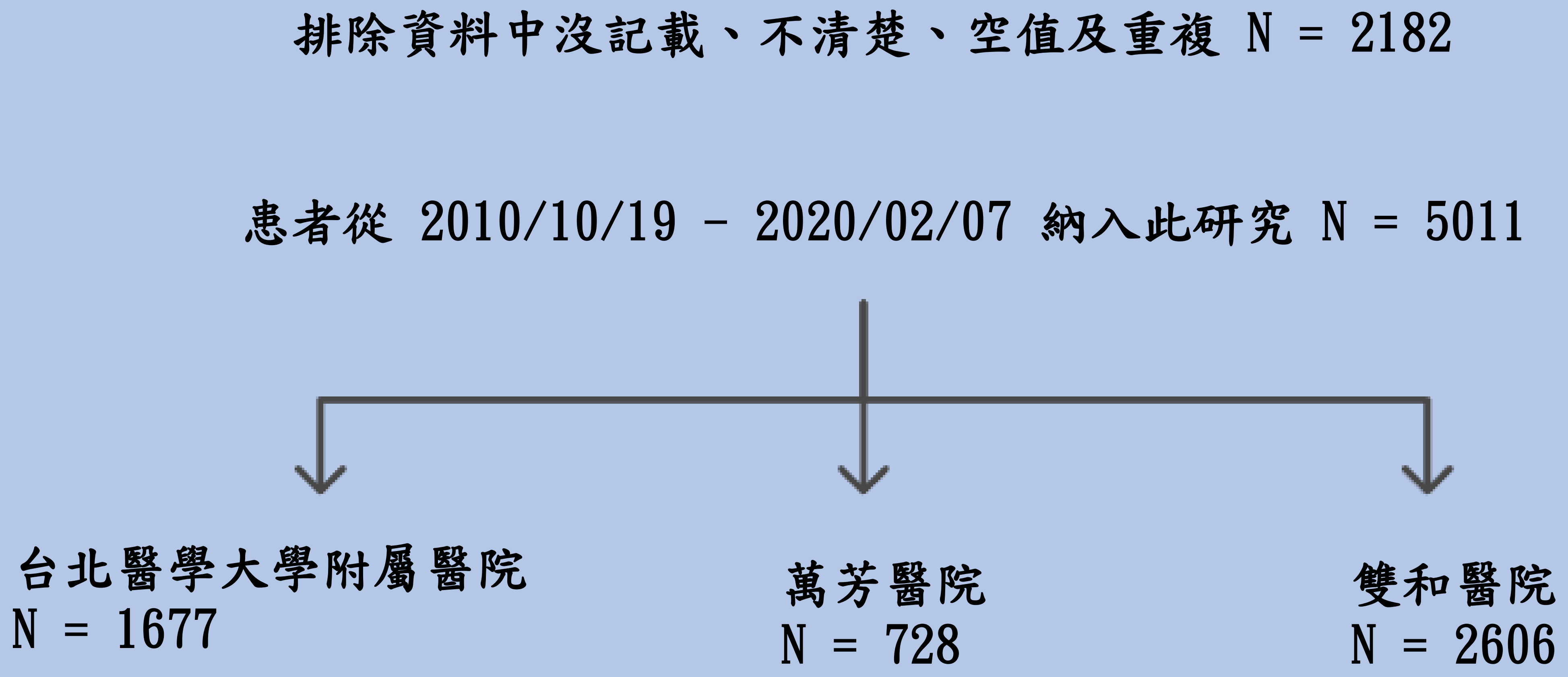
住院入院主診斷ICD-9、ICD-10為肺癌

有被登記在癌登資料中

ICD-O-3
原發腫瘤部位為肺癌
組織型態為肺腺癌、鱗狀細胞癌、大細胞癌

臨床分期I、II、III、IV

N = 7193



排除資料中沒記載、不清楚、空值及重複 N = 2182

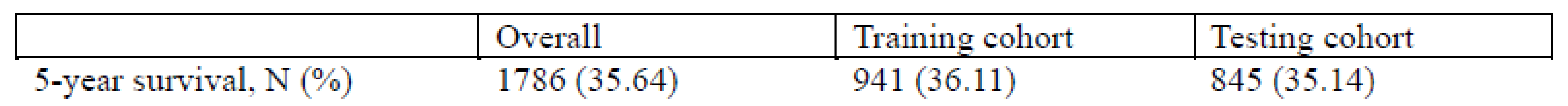
患者從 2010/10/19 - 2020/02/07 納入此研究 N = 5011

台北醫學大學附屬醫院
N = 1677

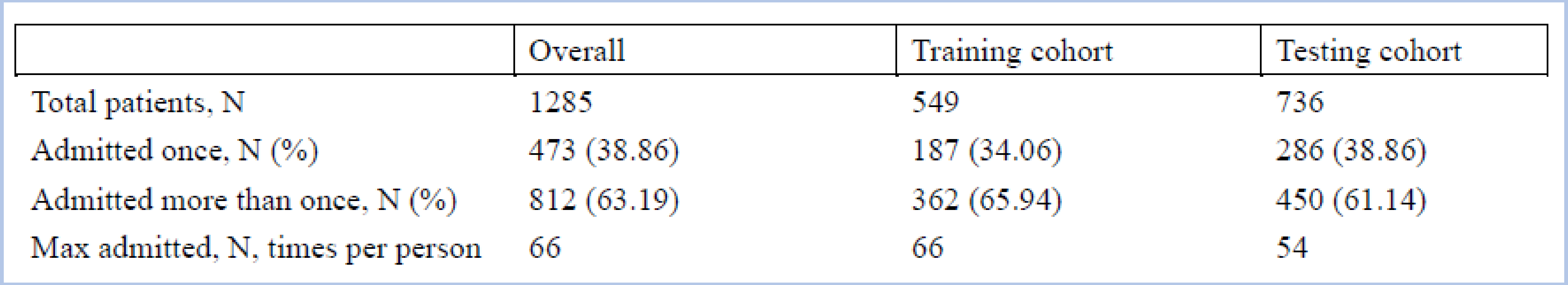
萬芳醫院
N = 728

雙和醫院
N = 2606

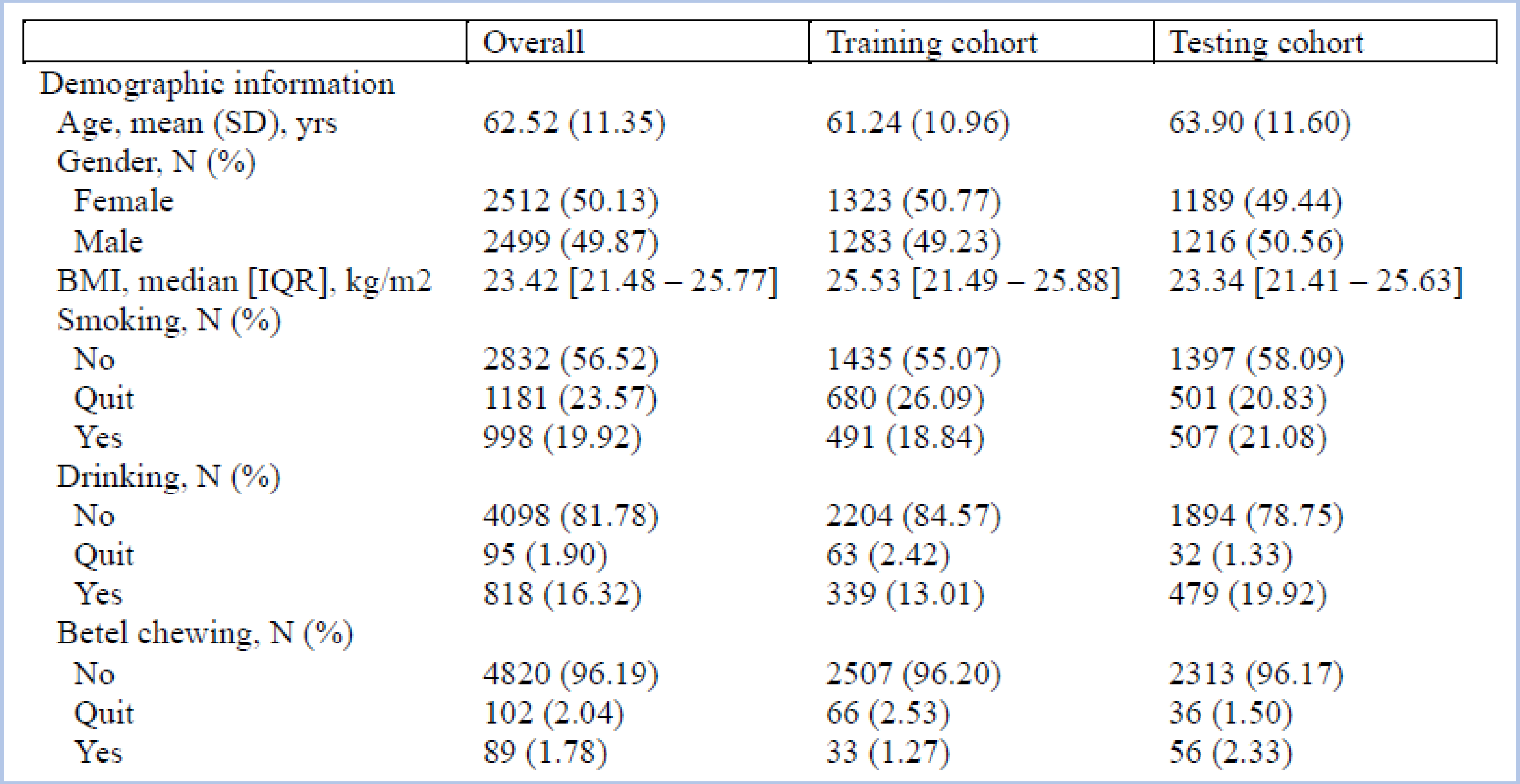
# 最初診斷日期往後計算5年時間，是否存活



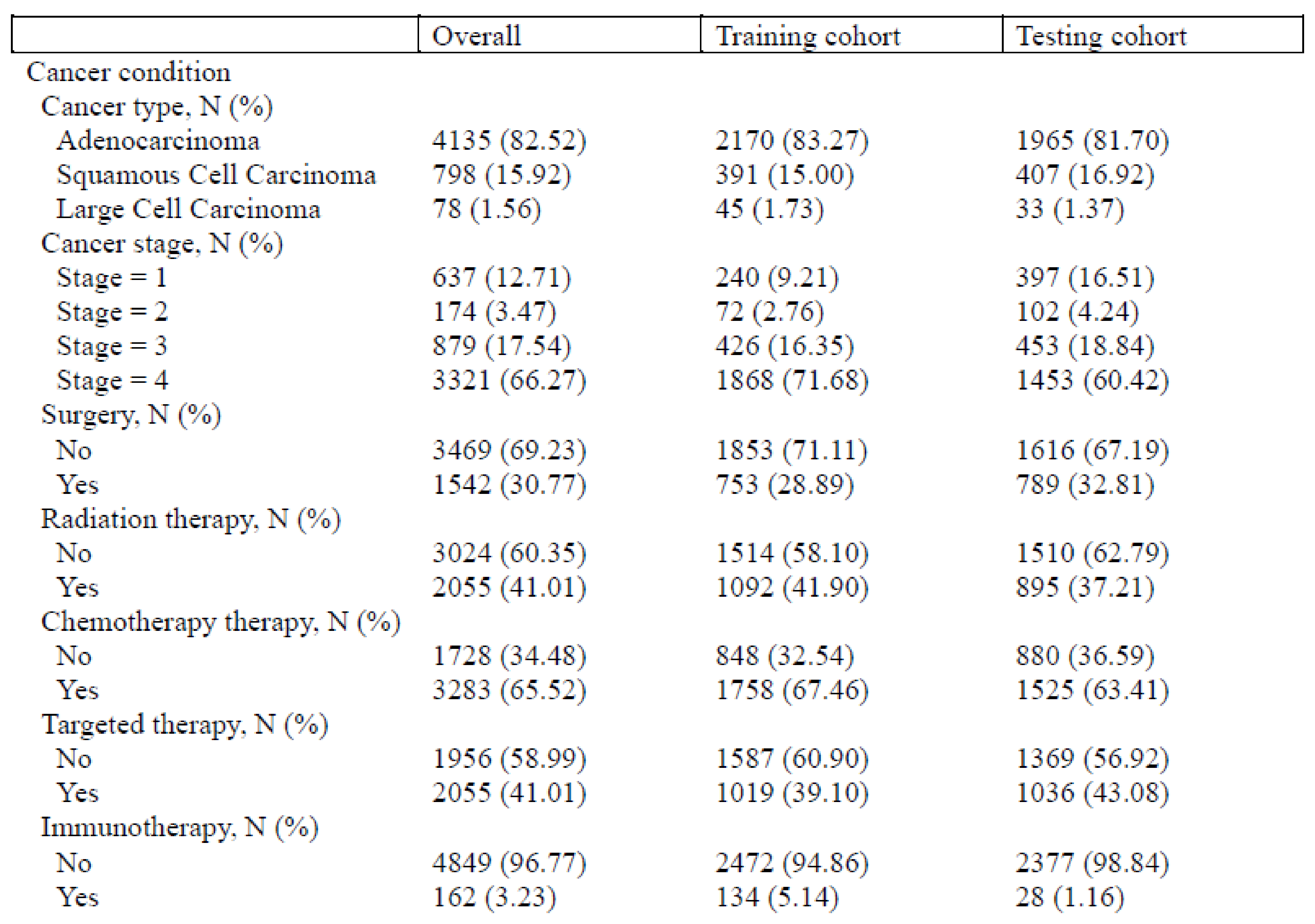
| | Overall | Training cohort | Testing cohort |
|---|---|---|---|
| 5-year survival, N (%) | 1786 (35.64) | 941 (36.11) | 845 (35.14) |



| | Overall | Training cohort | Testing cohort |
|---|---|---|---|
| Total patients, N | 1285 | 549 | 736 |
| Admitted once, N (%) | 473 (38.86) | 187 (34.06) | 286 (38.86) |
| Admitted more than once, N (%) | 812 (63.19) | 362 (65.94) | 450 (61.14) |
| Max admitted, N, times per person | 66 | 66 | 54 |

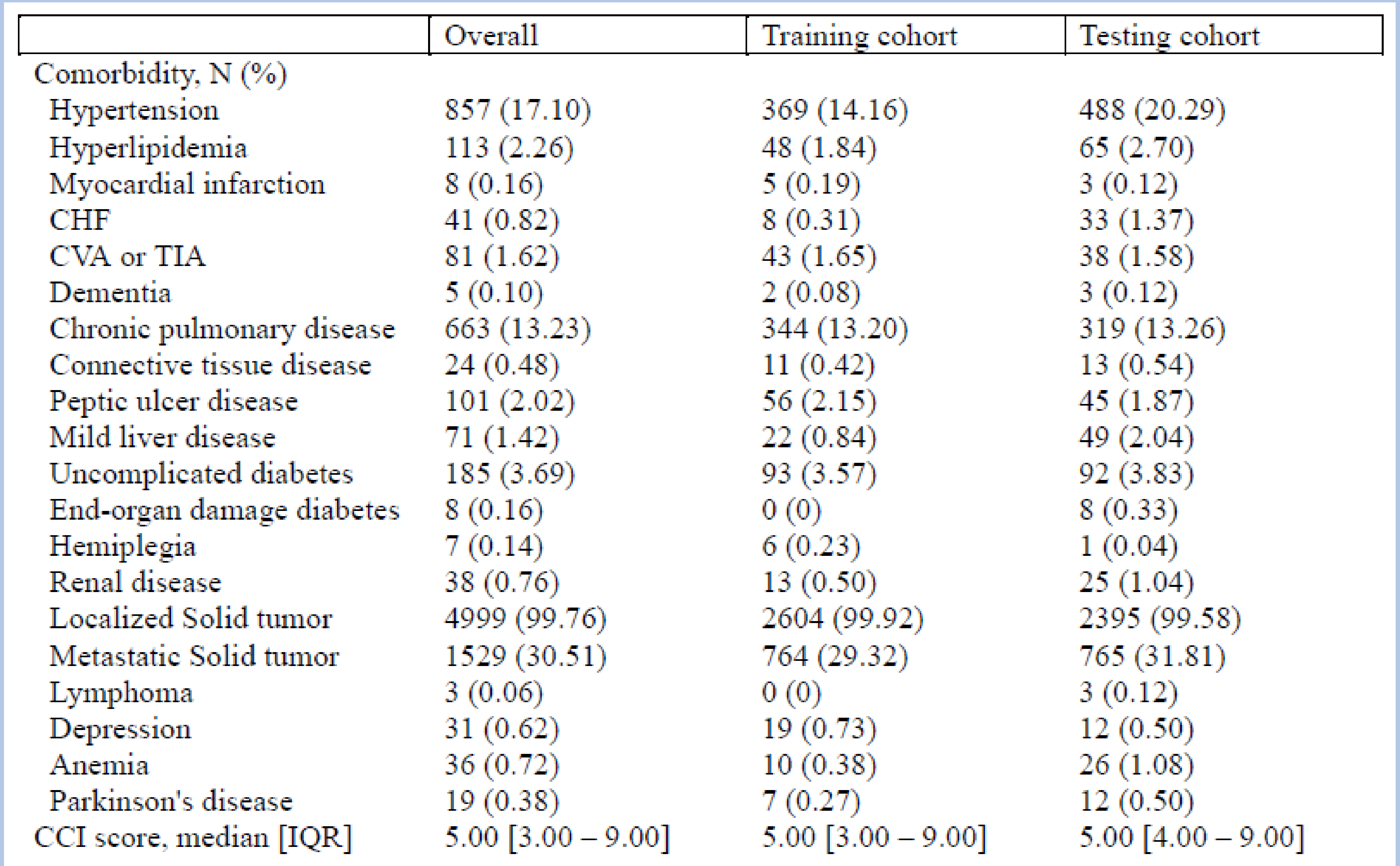


| | Overall | Training cohort | Testing cohort |
|---|---|---|---|
| Demographic information | | | |
| Age, mean (SD), yrs | 62.52 (11.35) | 61.24 (10.96) | 63.90 (11.60) |
| Gender, N (%) | | | |
| Female | 2512 (50.13) | 1323 (50.77) | 1189 (49.44) |
| Male | 2499 (49.87) | 1283 (49.23) | 1216 (50.56) |
| BMI, median [IQR], kg/m2 | 23.42 [21.48 – 25.77] | 25.53 [21.49 – 25.88] | 23.34 [21.41 – 25.63] |
| Smoking, N (%) | | | |
| No | 2832 (56.52) | 1435 (55.07) | 1397 (58.09) |
| Quit | 1181 (23.57) | 680 (26.09) | 501 (20.83) |
| Yes | 998 (19.92) | 491 (18.84) | 507 (21.08) |
| Drinking, N (%) | | | |
| No | 4098 (81.78) | 2204 (84.57) | 1894 (78.75) |
| Quit | 95 (1.90) | 63 (2.42) | 32 (1.33) |
| Yes | 818 (16.32) | 339 (13.01) | 479 (19.92) |
| Betel chewing, N (%) | | | |
| No | 4820 (96.19) | 2507 (96.20) | 2313 (96.17) |
| Quit | 102 (2.04) | 66 (2.53) | 36 (1.50) |
| Yes | 89 (1.78) | 33 (1.27) | 56 (2.33) |



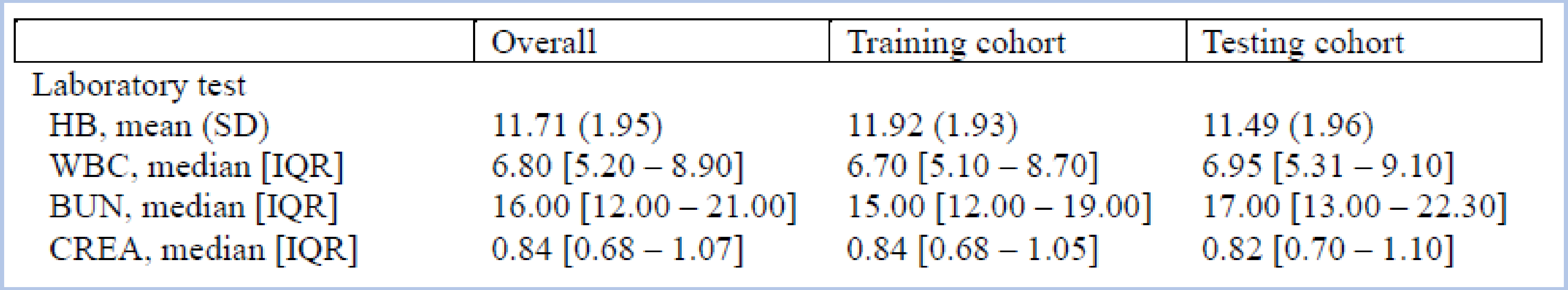
| | Overall | Training cohort | Testing cohort |
|---|---|---|---|
| Cancer condition | | | |
| Cancer type, N (%) | | | |
| Adenocarcinoma | 4135 (82.52) | 2170 (83.27) | 1965 (81.70) |
| Squamous Cell Carcinoma | 798 (15.92) | 391 (15.00) | 407 (16.92) |
| Large Cell Carcinoma | 78 (1.56) | 45 (1.73) | 33 (1.37) |
| Cancer stage, N (%) | | | |
| Stage = 1 | 637 (12.71) | 240 (9.21) | 397 (16.51) |
| Stage = 2 | 174 (3.47) | 72 (2.76) | 102 (4.24) |
| Stage = 3 | 879 (17.54) | 426 (16.35) | 453 (18.84) |
| Stage = 4 | 3321 (66.27) | 1868 (71.68) | 1453 (60.42) |
| Surgery, N (%) | | | |
| No | 3469 (69.23) | 1853 (71.11) | 1616 (67.19) |
| Yes | 1542 (30.77) | 753 (28.89) | 789 (32.81) |
| Radiation therapy, N (%) | | | |
| No | 3024 (60.35) | 1514 (58.10) | 1510 (62.79) |
| Yes | 2055 (41.01) | 1092 (41.90) | 895 (37.21) |
| Chemotherapy therapy, N (%) | | | |
| No | 1728 (34.48) | 848 (32.54) | 880 (36.59) |
| Yes | 3283 (65.52) | 1758 (67.46) | 1525 (63.41) |
| Targeted therapy, N (%) | | | |
| No | 1956 (58.99) | 1587 (60.90) | 1369 (56.92) |
| Yes | 2055 (41.01) | 1019 (39.10) | 1036 (43.08) |
| Immunotherapy, N (%) | | | |
| No | 4849 (96.77) | 2472 (94.86) | 2377 (98.84) |
| Yes | 162 (3.23) | 134 (5.14) | 28 (1.16) |

# 門診或住院所有出入院診斷ICD-9、ICD-10中相同診斷碼加總起來超過兩次



| | Overall | Training cohort | Testing cohort |
|---|---|---|---|
| Comorbidity, N (%) | | | |
| Hypertension | 857 (17.10) | 369 (14.16) | 488 (20.29) |
| Hyperlipidemia | 113 (2.26) | 48 (1.84) | 65 (2.70) |
| Myocardial infarction | 8 (0.16) | 5 (0.19) | 3 (0.12) |
| CHF | 41 (0.82) | 8 (0.31) | 33 (1.37) |
| CVA or TIA | 81 (1.62) | 43 (1.65) | 38 (1.58) |
| Dementia | 5 (0.10) | 2 (0.08) | 3 (0.12) |
| Chronic pulmonary disease | 663 (13.23) | 344 (13.20) | 319 (13.26) |
| Connective tissue disease | 24 (0.48) | 11 (0.42) | 13 (0.54) |
| Peptic ulcer disease | 101 (2.02) | 56 (2.15) | 45 (1.87) |
| Mild liver disease | 71 (1.42) | 22 (0.84) | 49 (2.04) |
| Uncomplicated diabetes | 185 (3.69) | 93 (3.57) | 92 (3.83) |
| End-organ damage diabetes | 8 (0.16) | 0 (0) | 8 (0.33) |
| Hemiplegia | 7 (0.14) | 6 (0.23) | 1 (0.04) |
| Renal disease | 38 (0.76) | 13 (0.50) | 25 (1.04) |
| Localized Solid tumor | 4999 (99.76) | 2604 (99.92) | 2395 (99.58) |
| Metastatic Solid tumor | 1529 (30.51) | 764 (29.32) | 765 (31.81) |
| Lymphoma | 3 (0.06) | 0 (0) | 3 (0.12) |
| Depression | 31 (0.62) | 19 (0.73) | 12 (0.50) |
| Anemia | 36 (0.72) | 10 (0.38) | 26 (1.08) |
| Parkinson's disease | 19 (0.38) | 7 (0.27) | 12 (0.50) |
| CCI score, median [IQR] | 5.00 [3.00 – 9.00] | 5.00 [3.00 – 9.00] | 5.00 [4.00 – 9.00] |

# 住院中檢驗結果報告時間最早的一筆



| | Overall | Training cohort | Testing cohort |
|---|---|---|---|
| Laboratory test | | | |
| HB, mean (SD) | 11.71 (1.95) | 11.92 (1.93) | 11.49 (1.96) |
| WBC, median [IQR] | 6.80 [5.20 – 8.90] | 6.70 [5.10 – 8.70] | 6.95 [5.31 – 9.10] |
| BUN, median [IQR] | 16.00 [12.00 – 21.00] | 15.00 [12.00 – 19.00] | 17.00 [13.00 – 22.30] |
| CREA, median [IQR] | 0.84 [0.68 – 1.07] | 0.84 [0.68 – 1.05] | 0.82 [0.70 – 1.10] |

## ■ 模型訓練及測試

訓練資料使用雙和

外部驗證測試資料使用北醫及萬芳

使用Stratified 5-fold、調整權重

## ■ 特徵篩選

特徵重要性篩選前加總至80%特徵

## ■ 模型使用

Logistic Regression

Random Forest

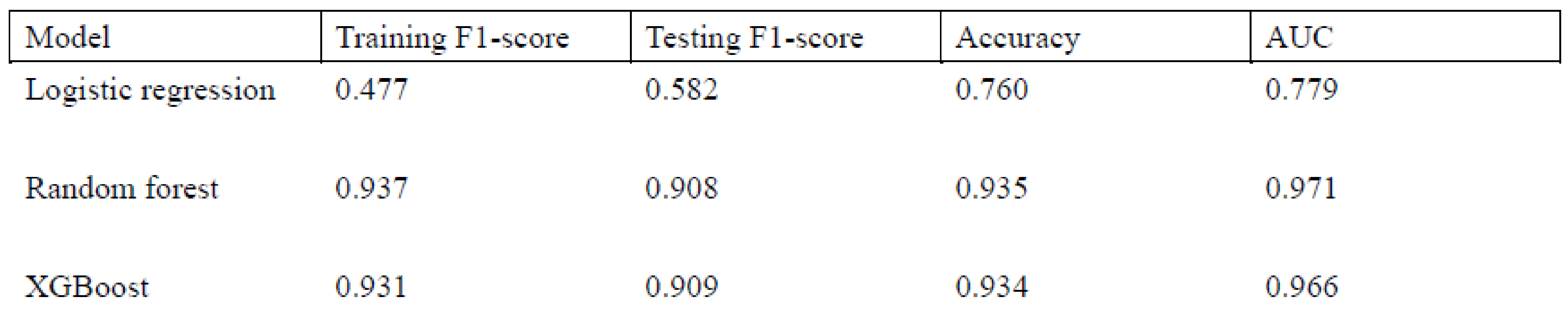
XGBoost

## ■ 模型評估指標

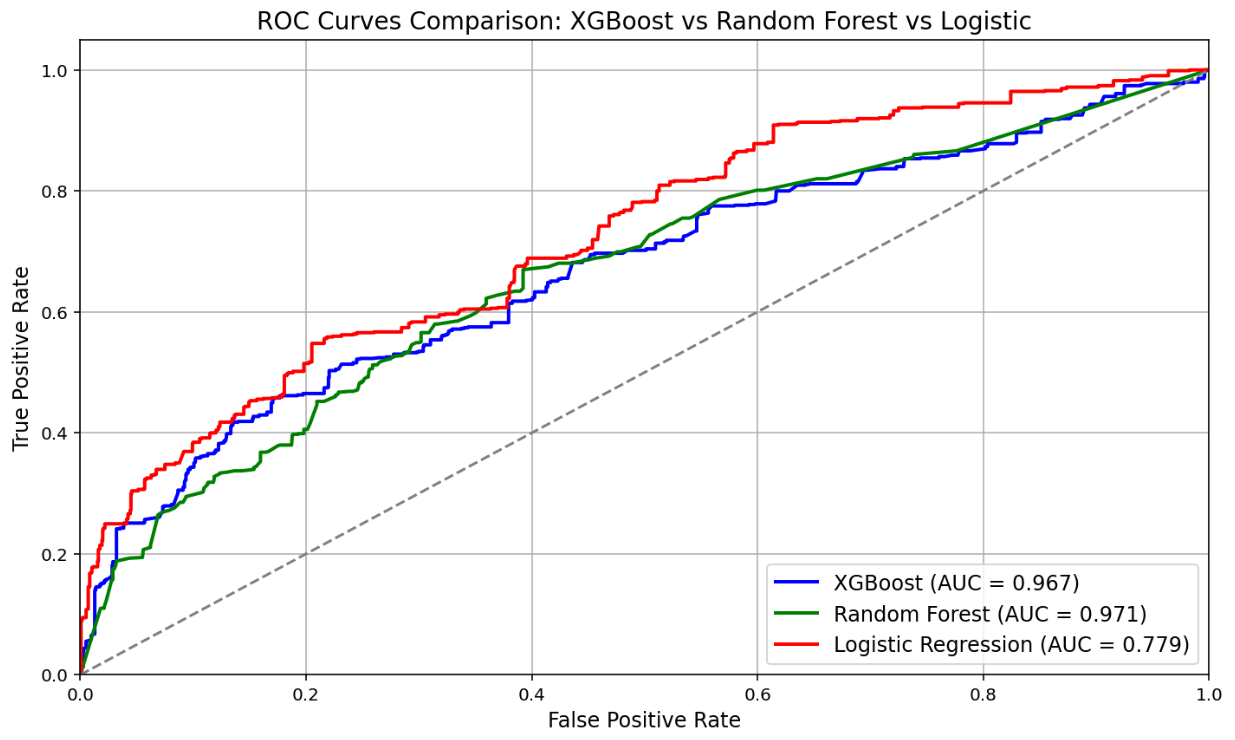
F1-score

Accuracy

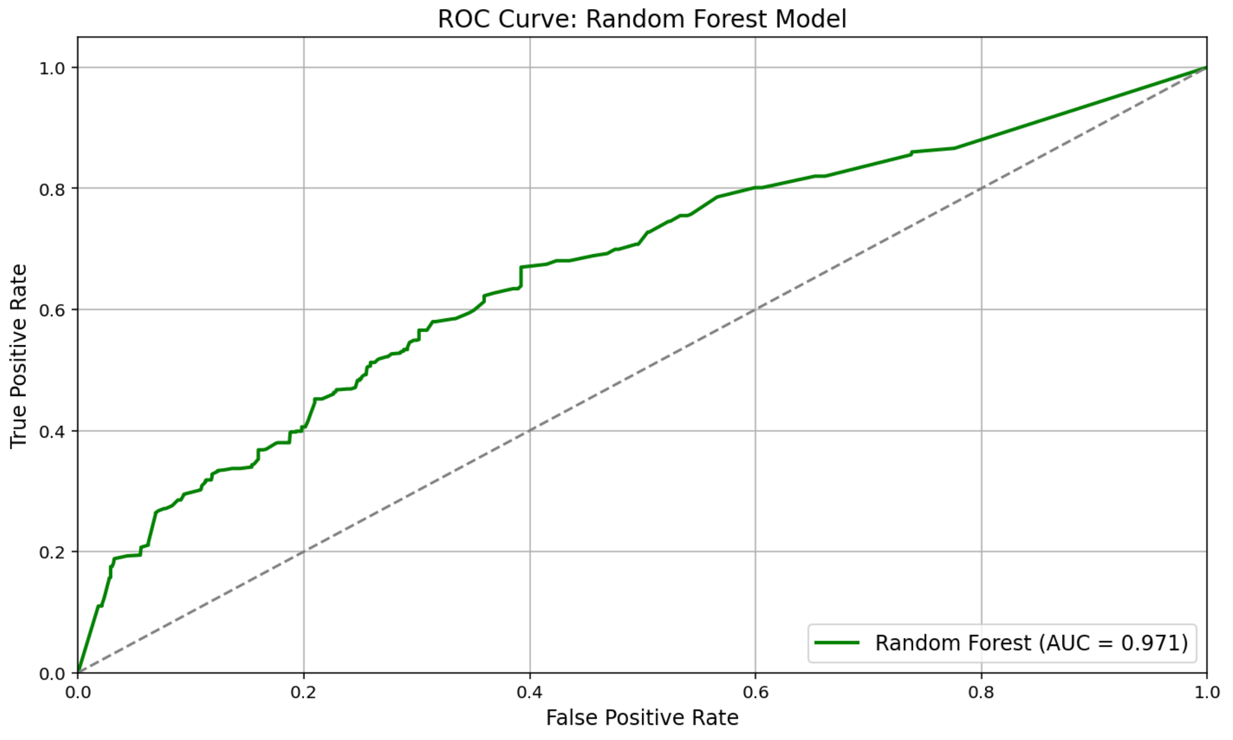
AUC



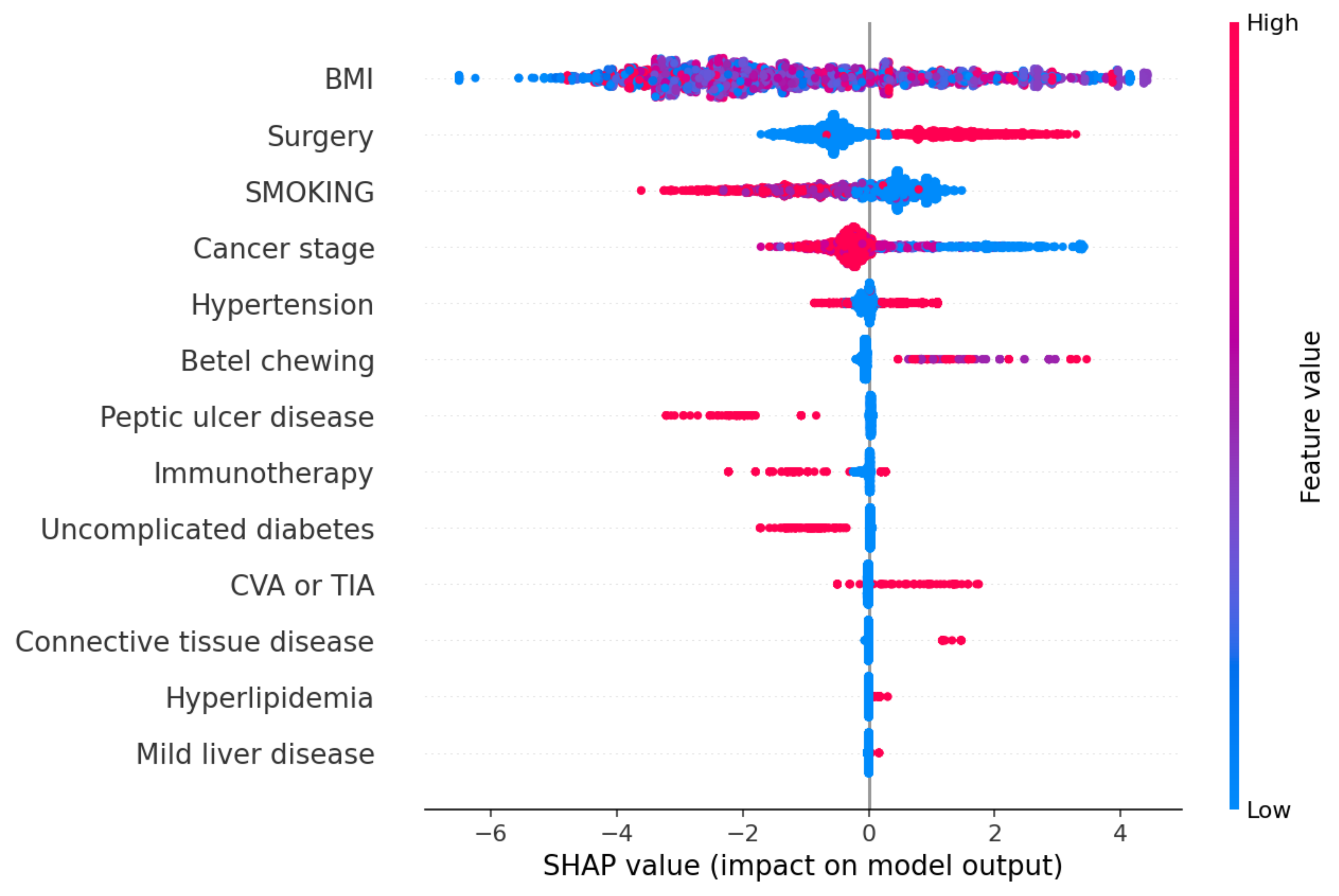
| Model | Training F1-score | Testing F1-score | Accuracy | AUC |
|---|---|---|---|---|
| Logistic regression | 0.477 | 0.582 | 0.760 | 0.779 |
| Random forest | 0.937 | 0.908 | 0.935 | 0.971 |
| XGBoost | 0.931 | 0.909 | 0.934 | 0.966 |



ROC Curves Comparison: XGBoost vs Random Forest vs Logistic



ROC Curve: Random Forest Model



預測5年存活的機率越大

有進行手術的

不抽菸

Stage越低

# 參考文獻資料

Kratzer, T. B., Bandi, P., Freedman, N. D., Smith, R. A., Travis, W. D., Jemal, A., & Siegel, R. L. (2024). Lung cancer statistics, 2023. *Cancer*, *130*(8), 1330–1348.

Filho, A. M., Laversanne, M., Ferlay, J., Colombet, M., Piñeros, M., Znaor, A., Parkin, D. M., Soerjomataram, I., & Bray, F. (2024). The GLOBOCAN 2022 cancer estimates: Data sources, methods, and a snapshot of the cancer burden worldwide. *International journal of cancer*, 10.1002/ijc.35278. Advance online publication.

Nguyen, P. A., Hsu, M. H., Chang, T. H., Yang, H. C., Huang, C. W., Liao, C. T., Lu, C. Y., & Hsu, J. C. (2024). Taipei Medical University Clinical Research Database: a collaborative hospital EHR database aligned with international common data standards. *BMJ health & care informatics*, *31*(1), e100890.

Nguyen, Q. T. N., Nguyen, P. A., Wang, C. J., Phuc, P. T., Lin, R. K., Hung, C. S., Kuo, N. H., Cheng, Y. W., Lin, S. J., Hsieh, Z. Y., Cheng, C. T., Hsu, M. H., & Hsu, J. C. (2023). Machine learning approaches for predicting 5-year breast cancer survival: A multicenter study. *Cancer science*, *114*(10), 4063–4072.

國民健康屬（2024）。*中華民國111年癌症登記報告*。台北市 ： 衛生福利部

感謝