國立臺灣大學公共衛生學院流行病學與預防醫學研究所

碩士論文

Graduate Institute of Epidemiology and Preventive Medicine
College of Public Health
National Taiwan University

Master Thesis

探討台灣敗血症病患28天存活相關基因

Identification of genetic variants associated with the 28-day survival in Taiwanese sepsis patients

吳珮瑄

Pei-Hsuan Wu

指導教授：盧子彬 博士

Advisor: Tzu-Pin Lu, Ph.D.

中華民國 112 年 6 月

June, 2023

doi:10.6342/NTU202301905

國立臺灣大學碩士學位論文
口試委員會審定書

探討台灣敗血症病患 28 天存活相關基因

Identification of genetic variants associated with the 28-day survival in Taiwanese sepsis patients

本論文係 吳珮瑄 君（學號 R10849031）在國立臺灣大學流行病學與預防醫學研究所完成之碩士學位論文，於民國 112 年 6 月 27 日承下列考試委員審查通過及口試及格，特此證明。

口試委員：

（簽名）

蕭自宏（指導教授）　李建學

凃妙蓮

doi:10.6342/NTU202301905

摘要

研究背景：

敗血症定義為病原體進入人體後，若免疫系統對其產生過度反應，且無法自主調節體內免疫功能時，引起器官衰竭和死亡的症候群。早期研究指出台灣敗血症發生率和死亡率有隨著年齡增加而上升的趨勢，目前臨床以相繼器官衰竭評估分數或血液檢測數值輔助判斷患者的治療策略，然而敗血症患者之間異質性高，即使進行相同治療，仍可能出現預後不佳的情形。先前有研究透過全基因組關聯研究(Genome-wide association study)提出位於調控體內免疫功能基因的單核苷酸多型性(single-nucleotide polymorphisms) 與敗血症患者死亡風險有關。由於這些研究大部分聚焦於非亞洲族群，被提出與敗血症死亡相關的基因尚未在亞洲族群進行驗證，因此本研究採用台灣敗血症病患資料進行分析，包含三個主要目的，首先找出與 28 天敗血症存活顯著相關位點。接著為了降低臨床實務的基因分析時間，以納入 100 個以內的位點估計多基因風險分數(polygenetic risk score)，並檢測其對模型預測能力的影響。最後透過多基因風險分數將敗血症病患區分為高死亡風險與低死亡風險組。預期利用該方法可加速辨別 28 天高死亡風險病患，提早治療介入時間。

材料與方法：

由 2017 年 1 月至 2021 年 12 月從台中榮總醫院收集 3,847 位首次診斷敗血症的患者，排除基因資料和臨床資料缺漏的個案，經過品質控制共納入 373 名敗血症患者和 6,865,736 個位點。研究的主要終點(primary endpoint)定義為診斷敗血症後 28 天內發生死亡。透過每位個案的存活時間進行分組，其中 39 位(10.5%)在 28 天內發生死亡，334 位(89.5%)沒有觀察到死亡事件。以依變數的形式進行兩種全基因組關聯分析，首先利用 Cox 迴歸分析與 28 天敗血症存活時間相關之基因變異，接著將連續型的存活時間以 28 天作為切點，轉換至二元變項，以邏輯斯迴歸分析 28 天敗血症死亡風險相關之基因變異。同時收集過去六篇非亞洲族群研究，觀察被提出的顯著位點是否能在台灣族群中進行驗證。接著從 373 名敗血症患者隨機抽出 80% (n=299)作為原始資料集(base dataset)，進行全基因組關聯分析，根據 C + T 方法(剪枝與 p 值篩選)，依據每個人基因座攜帶的風險等位基因數量加權由 Cox 迴歸和邏輯斯迴歸估計的效應大小後，估計多基因風險分數，稱為 PRS-sepsis；並由過去六

doi:10.6342/NTU202301905

篇非亞洲族群研究提出與罹患敗血症或敗血症死亡相關的顯著位點，估計另一多基因風險分數，稱為 PRS-sepsis-pre。透過五倍交叉驗證檢測模型的預測能力。首先加入 PRS 變數建立基準模型(baseline model)，並控制年齡、性別和主成分。接著選入在單變數 Cox 迴歸顯著的臨床特徵建立多變數模型(multivariate model)。最後為了降低多變數模型的複雜度，利用最小絕對壓縮挑選操作迴歸(簡稱 Lasso 迴歸)和逐步迴歸，篩選合適的臨床變數建立最終模型(final model)。所有納入 PRS 的模型與只有臨床變數的傳統模型進行比較，評估加入基因變數對模型預測能力的影響。若以 Cox 迴歸建立預測模型，將由一致性指數評估模型預測能力；若由邏輯斯迴歸建立預測模型，則以 ROC 曲線下面積(area under the curve) 評估模型預測能力。並依據 PRS-sepsis 分數將病患進行 28 天死亡風險分層，利用 Kaplan-Meier 分析高風險和低風險組別是否具有顯著差異。

研究結果：

透過 Cox 迴歸校正年齡、性別和主成分進行全基因組關聯研究分析，共有 4 個顯著位點($p < 5\times10^{-8}$)，過去在非亞洲族群研究觀察到 134 個與罹患敗血症或敗血症死亡相關位點，其中 64 個可在本研究樣本中進行驗證，但皆未達統計顯著。利用 C + T 篩選留下 86 個 SNPs 估計 PRS-sepsis。另外透過過去文獻指出的 64 個顯著位點估計 PRS-sepsis-pre，並觀察其在台灣敗血症病患的應用。由結果顯示，納入 PRS-sepsis 的多變數模型 C-index 顯著高於傳統模型。接著由 Lasso 迴歸以及逐步迴歸針對多變數模型篩選潛在重要的臨床變數並建構最終模型。其模型的 C-index 與傳統模型相比，於訓練集顯著提升 5.85%和 9.3% ($p < 0.05$)，而在測試集分別顯著提升 15.18%和 15.23% ($p < 0.05$)；另外檢測加入 PRS-sepsis-pre，以及分別透過 Lasso 迴歸和逐步迴歸篩選的臨床變數建構最終模型後，與傳統模型相比，模型 C-index 於訓練集上升 1.57%和 1.92%，而在測試集上升 3.04%和 0.33%。然而這些預測能力沒有達到顯著差異($p > 0.05$)。將 PRS-sepsis 分數由低至高排序，以 67%作為切點，分數最高的 33% (前三分之一)病患分類至高分組，其餘為低分組，由統計結果顯示兩組的存活風險存在顯著差異($p < 0.0001$)。若以邏輯斯迴歸校正年齡、性別和主成分透過全基因組關聯研究分析，沒有任何位點達顯著，但仍有 19 個位點達潛在相關($p < 1\times10^{-5}$)。過去非亞洲族群研究提出的 134 個顯著位點在台灣敗血症族群中皆未顯示顯著

doi:10.6342/NTU202301905

相關。然而有 59 個 SNPs 與台灣敗血症病患的位點重疊，其中 58 個同時出現在原始資料集，並依據這些 SNPs 估計 PRS-sepsis-pre。透過 C + T 篩選方式留下 94 個 SNPs 估計 PRS-sepsis。當模型藉由 Lasso 迴歸和逐步迴歸留下合適的臨床變數後，加入 PRS-sepsis 建構最終模型，模型預測能力 AUC 比傳統模型分別增加 11.5%和 11.7%，但皆未達統計顯著差異($p > 0.05$)。若由 Lasso 迴歸留下合適的臨床變數後，加入 PRS-sepsis-pre 建構最終模型，模型預測能力 AUC 與傳統模型相同；然而臨床變數若由逐步迴歸進行篩選，最終模型的 AUC 比傳統模型增加 2.1%，但未達統計顯著差異($p > 0.05$)。

結論：

敗血症屬於綜合性症狀，患者可能同時伴隨不同症狀，若未即時接受合適治療，出現多重器官衰竭後，可能導致死亡。過去研究指出臨床特徵、宿主與病原體的交互作用、治療方式與特定基因，使敗血症患者之間出現預後差異。因此本研究以全基因組關聯研究，觀察到 4 個顯著與存活時間相關的位點，並進一步利用過去非亞洲族群相關研究提出的顯著位點進行分析，然而根據結果顯示這些位點無法在台灣敗血症患者中觀察到與存活時間顯著相關；其中透過本研究族群發現的 4 個達顯著的位點，亦無重疊於過去非亞洲族群研究提出的位點。透過千人基因組計畫發現這 4 個顯著位點，其中 3 個在美洲、非洲、歐洲和南亞洲的次要等位基因頻率皆為 0，推測這些位點可能只出現在亞洲地區敗血症患者。當模型考慮基因變數後，預測能力高於傳統單純由臨床變數建構的模型。若透過多基因風險評分方式，將敗血症患者區分為高、低分數組，結果顯示兩組的死亡風險具有顯著差異。因此患者若被診斷為敗血症，可立即透過標的基因檢測區分死亡風險，輔助臨床端提供合適治療策略，提升患者預後。

關鍵字：敗血症、28 天死亡、全基因組關聯研究、存活分析、多基因風險評分

doi:10.6342/NTU202301905

Abstract

Background:

Sepsis occurs when the immune system has an extreme response to infection by pathogens, and consequently leads to organ dysfunction and even death if the immune system cannot regulate itself normally. A previous study indicated that the incidence and mortality rate of sepsis increased with the increasing age among sepsis patients in Taiwan. The Sequential Organ Failure Assessment (SOFA) score and/or blood test results are used to make treatment decisions in the clinic. However, due to the high heterogeneity among sepsis patients, even when using the same treatments, patients suffer deterioration. A series of studies have identified the relationships between single-nucleotide polymorphisms (SNPs) and the mortality of sepsis using genome-wide association studies (GWAS). These SNPs were located in genes that regulate immune-related function. However, most of these studies focused on non-Asian populations. The reported SNPs have not been validated in Asian sepsis patients. Therefore, this study used the data from Taiwanese sepsis patients. This study had three main objectives. Firstly, we aimed to identify the significant SNPs associated with 28-day survival in sepsis patients. Secondly, to expedite the time of conducting gene analysis in clinical settings, we limited the number of SNPs used for calculating the PRS to less than 100. Prediction models were then constructed to evaluate the impact of including the PRS. Thirdly, stratifying patients based on their PRS, they were categorized into high-risk or low-risk groups for 28-day mortality. This approach held potential for discerning high-risk of 28-day mortality among sepsis patients early and facilitating timely intervention.

Methods and materials:

A cohort of 3,847 patients with a first diagnosis of sepsis at Taichung Veterans General

doi:10.6342/NTU202301905

Hospital from January 2017 to December 2021 were included as participants. Patients without genotype data or clinical data were excluded. After quality control, there were 373 sepsis patients and 6,865,736 SNPs. The primary endpoint of the study was death due to sepsis within a 28-day follow-up period. According to the survival time to classify patients, of 39 (10.5%) were with events and the rest 334 (89.5%) were without observed events in this study. The GWAS was conducted in two ways, depending on the class of the dependent variable. First, the GWAS was performed using Cox regression to identify the SNPs associated with survival time. Second, survival time with continuous format was converted to binary (1, 0) based on death within a 28-day follow-up period. Logistic regression was performed in the GWAS to identify the significant SNP associated with sepsis mortality. Six studies that reported the significant SNPs in non-Asians were selected for validation in the Taiwanese sepsis patients. Sampling 80% of the whole samples to be base dataset (n=299) was used to perform GWAS to obtain the effect sizes of the SNPs. After the C + T (clumping and thresholding) method was used to select the SNPs, individuals' PRS were estimated as PRS-sepsis using the number of risk alleles for each variant weighted by its effect size. Conversely, SNPs previously identified as significantly associated with sepsis incidence or mortality were combined to create a second polygenic risk score, PRS-sepsis-pre. Five-fold cross validation was used to verify the predictive power of the models. The baseline model was constructed by adding the polygenic risk score (PRS) and adjusting for age, sex, and principal components (PCs). Subsequently, the multivariate model was built by incorporating the clinical variables selected through univariate regression into the baseline model. Finally, to address the complexity of the multivariate model, the least absolute shrinkage and selection operator (Lasso) regression and stepwise regression were employed to select the appropriate clinical variables and construct the final model. The performances of all models with the incorporation of PRS

doi:10.6342/NTU202301905

was compared against the traditional models that included only clinical variables, in order to assess the improvement in model performance by incorporating genetic predictors. The concordance index (C-index) and the area under the curve (AUC) represented the predictive power of the model when constructed using Cox regression and logistic regression, respectively. Stratified by PRS-sepsis, Kaplan-Meier analyses was conducted on high-risk and low-risk individuals to evaluate whether the risks of 28-day survival differed between two groups.

Results:

A total of 4 SNPs were genome-wide significantly associated with survival time using Cox regression, adjusting for age, sex, and PCs ($p < 5\times10^{-8}$). Previous studies identified 134 significant ($p < 1\times10^{-5}$) SNPs, 64 of which were observed in Taiwanese sepsis patients. However, none were significantly associated with 28-day survival. The C + T thresholds allowed a total of 86 SNPs to be aggregated for PRS-sepsis construction. In addition, 64 significant SNPs from previous studies were pooled to create the PRS-sepsis-pre and tested for applicability to sepsis patients from Taiwan. The result was observed that the inclusion of PRS-sepsis in the multivariate model resulted in a significant improvement in the C-index compared to the traditional model. After applying Lasso regression and stepwise regression to select potentially important clinical variables in the multivariate model, the final models were constructed. Compared to the traditional model, the C-index exhibited a significant increase of 5.85% and 9.3% in the training set, respectively ($p <0.05$). In the testing set, the C-index also showed a significant increase of 15.18% and 15.23% ($p <0.05$). Furthermore, incorporating PRS-sepsis-pre and the clinical variables chosen by Lasso regression and stepwise regression into the final model resulted in a 1.57% and 1.92% increase of the C-index in the training set, as well as a 3.04% and 0.33% increase of the C-

doi:10.6342/NTU202301905

index in the testing set. However, all of the predictive powers of the final models were not significantly different from those of the traditional models ($p > 0.05$). The sepsis patients improved one standard deviation of PRS-sepsis, the probability of the mortality would increase nearly 4% when considering the clinical features. Based on the ascending PRS-sepsis values, patients in the top 33%, the highest one-third, were classified as the high PRS group, while the remaining patients were classified as the low PRS group. The hazard functions were significantly different between the two groups ($p < 0.0001$). Even though none of the SNPs were genome-wide significant using logistic regression, 11 SNPs were suggestively associated with sepsis mortality adjusting for age, sex and PCs ($p < 1\times10^{-5}$). A total of 134 SNPs from prior studies were deemed significant, but none of them showed significance in sepsis patients in Taiwan. Out of these, only 59 SNPs were observed in Taiwanese sepsis patients. Among them, 58 SNPs were replicated in base dataset and utilized to generate the PRS-sepsis-pre. Applying a C + T threshold, 94 SNPs were selected and aggregated as PRS-sepsis. Upon adding PRS-sepsis to the model along with the clinical variables retained through Lasso regression and stepwise regression, the final model exhibited an improvement of 11.5% and 11.7% in AUC; however, these improvements were not statistically significant compared to the traditional models ($p > 0.05$). When incorporating PRS-sepsis-pre into the final model along with the clinical variables selected via Lasso regression, the AUC of the final model remained the same as that of the traditional model. On the other hand, when the clinical variables were retained using Lasso regression, the AUC of the final model improved by 2.1% when compared to the traditional model. However, this improvement was not statistically significant ($p > 0.05$).

Conclusion:

Sepsis is a syndrome with diverse symptoms. Patients with sepsis develop organ

doi:10.6342/NTU202301905

dysfunction and then death if they could not receive the appropriate treatments. The prior study reported that the heterogeneity in sepsis is related to clinical features, host-pathogen interactions, treatments, and/or genetic predisposition, leading to the differences in prognosis among patients. In this retrospective study, we identified 4 SNPs significantly associated with 28-day survival. None of the significant SNPs from the previous studies were replicated among sepsis patients in Taiwan. In addition, the four SNPs identified in our data did not overlap with the SNPs reported in previous studies. Additionally, among the Americans, Africans, Europeans, and South Asians in the 1000 Genomes phase 3 dataset reference panel, the minor allele frequencies of three out of the four SNPs were zero. Therefore, these significant SNPs may only appear in Asian population. Comparing to the traditional model, adding polygenic risk scores improved the predictive power of the models. The high-PRS and low-PRS subgroups well discriminated the risk of sepsis mortality. In summary, the use of these significant genes in the diagnosis of sepsis can help clinicians to determine the subgroup at risk and guide therapy, thereby improving outcomes.

Key words: Sepsis, 28-day mortality, Genome-wide association study, survival analysis, polygenic risk score

doi:10.6342/NTU202301905

Contents

doi:10.6342/NTU202301905

doi:10.6342/NTU202301905

doi:10.6342/NTU202301905

List of Figures

doi:10.6342/NTU202301905

doi:10.6342/NTU202301905

List of Tables

doi:10.6342/NTU202301905

doi:10.6342/NTU202301905

Chapter 1: Introduction

1.1 The epidemiology of sepsis

Sepsis occurs when the immune system has an extreme response to an infection by pathogens. The underlying causes of sepsis include infections, non-communicable diseases, and injuries. The most common underlying cause is infection. When pathogens, including bacteria, viruses, and fungi invade the host body and cause a simple infection, they proliferate within the bloodstream and release substantial quantities of toxins. These toxins are carried by the blood to other organs, leading to dysfunction in one or more organs. In more acute and/or severe cases, patients with sepsis will develop hypotension and hyperlactatemia and progress to septic shock. Without proper treatment, sepsis can rapidly lead to organ failure and even death.

The global incidence of sepsis decreased by 37% from 1990 to 2017, but the age-standardized incidence rate of sepsis still reached 677.5 cases per 100,000 person-years in 2017. Higher incidence rates are observed in both the neonatal group and the elderly group. The global age-standardized mortality rate is 148.1 cases per 100,000 person-years [1]. Sepsis accounts for 25%-30% of hospital deaths worldwide. Nevertheless, in patients with more intricate medical conditions, the mortality rate can escalate to as high as 40%-50%. [2]. In addition, the mortality rate is higher than the average in the less developed countries [3].According to a previous study, there were 643 cases of sepsis per 100,000 person-years and 287 deaths due to sepsis per 100,000 person-years between 2010 and 2014 in Taiwan [4]. They also observed that both the incidence and/or mortality rate of males were higher than that of females while the elderly had a higher risk of mortality than adults [5].

The 28-day mortality rate for sepsis and septic shock ranges from 3.9% to 14% [6-8].

doi:10.6342/NTU202301905

However, for sepsis patients with malignancy, the 28-day mortality rate rises to between 50.9% and 70.2%. With pleural cancer, sepsis patients have the highest mortality rate at 81.5% [5] and those with high lactate levels and the severity score, Acute Physiology and Chronic Health Evaluation II score, are more prone to 28-day mortality [8]. Moreover, BMI is associated with mortality; where patients in the underweight group (BMI < 18.5) has a higher risk than their counterparts (BMI ≥ 18.5) [9].

1.2 The impact of heredity on infectious diseases

A prior study investigated the impact of heredity in infectious diseases [10] constituting of children who were separated from their biological parents and lived with adoptive parents. If either of the biological parents died of an infectious disease before the age of 50 or 70, their children exhibited a greater risk of mortality from infectious diseases compared to children whose biological parents were still alive (OR=5.81, 95% CI: 2.47–13.7, and OR=5.00, 95% CI: 1.73–14.4), however, the mortality relationships between children and adoptive parents were not significantly associated if the adoptive parents died of an infectious disease before age 50 or 70.

In the recent decades, samples from individuals can be genotyped using high throughput microarrays, which is a useful platform to obtain whole genome variants such as single-nucleotide-polymorphisms (SNPs) [11]. SNP constitutes of a variation in a single nucleotide with in the DNA sequence of a genome and is observed in more than 1% of people in a population. The genome-wide association study (GWAS) approach can identify correlated SNPs within the same block that are associated with a particular trait and/or disease. It can help pinpoint risk loci and potential genes involved in regulating the trait and/or disease [12].

Several GWASs have been applied to identify the infection related genetic variants. A

doi:10.6342/NTU202301905

previous study reported two variants (rs2856718 and rs7453920) in the HLA-DQ locus involved in the development of chronic hepatitis B (CHB). Variants in the antigen-binding regions of HLA-DQ are linked to the risk of long-term HBV infection [13]. The previous cohort studies of HIV-1 disease identified genetic variants within the MHC region significantly associated with viral load [14, 15].

1.3 The relationship between the genetic variants and sepsis mortality

According to the aforementioned studies, genetic variants have been demonstrated to be associated with infectious diseases. Previous studies have aimed to find genetic variants associated with the risk of mortality from sepsis or septic shock. Through the GWAS analysis, they successfully identified several genetic variants that exhibited significant associations with mortality, primarily in non-Asian sepsis patients. However, due to heterozygosity among sepsis patients, the significant genetic variants were poorly replicated in other sepsis studies, even though the enrolled individuals were from the same ethnicity. In a previous European study, 21 SNPs associated with 28-day mortality caused by pneumonia were reported. The most significant SNP was rs4957796, located on the *FER* gene on chromosome 5. The risk for death was 44 % lower at rs4957796 (HR=0.56, 95% CI: 0.45–0.69) [16]. Another Caucasian study identified 16 SNPs associated with 7 to 28 days mortality in severe sepsis or septic shock. Upon ranking the SNPs based on their correlation with sepsis mortality and assigning a comprehensive weighted score, the highest score was for rs143356980 situated within the *CISH* and *MAPKAPK3* gene. Both the *FER* gene and the *CISH* gene are involved in cytokine regulation [17].

1.4 Polygenic risk score (PRS)

The GWAS method has been widely used to identified potential related SNPs. However,

doi:10.6342/NTU202301905

determining the relationship between genes and disease can be challenging [18] because the significant genetic variants identified by GWAS have small or moderate effect sizes. Therefore, not only using the significant results from GWASs but aggregating a set of genetic variants correlated with a trait or disease were to examine the individual's risk. The polygenic risk score (PRS) is a value presenting the individual's risk of developing a particular trait or disease, which is calculated by evaluating the number of risk allele the individual takes and then multiplying the genetic effects derived from summary statistics of GWASs.

PRS is utilized to discriminate between different risk groups [19] and improves the predictive power. Several studies on coronary artery disease have shown that incorporating PRSs into prediction models improves their performance and enhances risk stratification [20]. Till date, there are few PRS studies on sepsis susceptibility risk or sepsis mortality. In a previous study, PRSs derived from SNPs associated with decreased granulocyte count or decreased hematocrit levels were reported to be predictive of 28-day mortality [21]. However, PRS aggregated with other SNPs clustered at multiple p-value thresholds were not significantly ($p < 0.001$) associated with 28-day sepsis survival [22].

The progression of sepsis is rapid, leading to the median time to death is approximately 7 days (IQR: 3–15) [23]. Therefore, most sepsis deaths are observed within 30 days of sepsis diagnosis as opposed to between 30 to 90 days [23]. Additionally, the time to medical intervention in sepsis patients is related to their survival probabilities. A prior study highlights that delaying the control of infection sources among sepsis patients is associated with increased 90-day mortality [24]. Specifically, patients who had their infection sources controlled within 6 hours showed a 29% reduction in 90-day mortality compared to those whose control occurred after 6 hours. Therefore, to expedite the time

doi:10.6342/NTU202301905

of conducting gene analysis in clinical settings, the number of SNPs utilized to calculate the PRS was restricted to be less than 100. These genetic variants will be compiled and incorporated into a specialized chip designed as a rapid test kit. Once sepsis is diagnosed in patients, they can undergo the rapid test for the targeted genes, providing them with their test results and individual PRS within a few hours. Based on the PRS, patients can be categorized into high-risk or low-risk groups for 28-day mortality, helping physicians make treatment decisions.

1.5 The aims of the study

The main purpose of this study was to identify the novel genetic variants associated with 28-day sepsis survival among patients from Taiwan. The second aim was to confirm and validate the significant genetic variants that were reported to be associated with sepsis incidence or mortality from non-Asian sepsis patients in Taiwanese sepsis patients. A set of fewer than 100 SNPs from all significant variants was included to generate an individual's polygenic risk score (PRS) by aggregating the effect sizes of these SNPs. This PRS was then utilized to build PRS-based prediction models. The efficacies of the models were established by the rigorous procedures and evaluate the predictive power of the PRSs in the prediction models. Subsequently, all patients were stratified into two groups, high- and low-risk groups, for 28-day sepsis mortality based on their calculated PRS.

doi:10.6342/NTU202301905

Chapter 2: Material and Methods

2.1 Data source and study population

A cohort of 3,847 patients with a first diagnosis of sepsis who were 20 years of age or older at Taichung Veterans General Hospital from January 2017 to December 2021 were included as participants. Patients with sepsis were defined depended on the guidelines of Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) [25]. Of these, 3,436 without genotype data and 31 duplicate samples were excluded (Figure 1). For the remaining 380 individuals, whole-genome genotyping and clinical data were incorporated into the study.

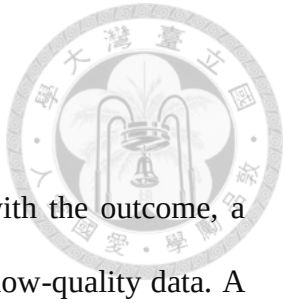
The genotyping was conducted using the Taiwan Precision Medicine Initiative (TPMI) chip (Affymetrix Axiom Genomewide TPMI array). Genotype imputation was performed to obtained untyped SNPs that were not directly genotyped [26]. East Asian samples from the 1000 Genomes phase 3 dataset reference panel was used pre-phasing and imputation utilizing SHAPRIT2 and IMPUTE2 respectively [27]. A total of 12,931,988 SNPs was obtained post imputation.

Demographic data such as age, sex, height and weight, and clinical data via Electronic Health Records (EHR) data such as comorbidities, blood test data, treatments, main infection sites, etc. were obtained for each patient included in the analysis.

2.2 Definition of study outcome

The primary endpoint was death due to sepsis within a 28-day follow-up period. Events were censored if patients were (i) alive at discharge, (ii) died after the 28-day follow-up period, or (iii) had an accidental death. A total of 39 patients (10.5%) were with events and the rest 334 patients (89.5%) were without observed events in this study.

doi:10.6342/NTU202301905



2.3 Quality control

To mitigate the risk of false positives and uncertain associations with the outcome, a quality control procedure was implemented to identify and remove low-quality data. A total of 12,931,988 SNPs was initially obtained after imputation for 380 sepsis patients (Figure 2). To include highly accurate SNPs, an imputed information metric (INFO score) of 0.7 was used as a threshold to exclude low quality imputed SNPs [26]. After that, SNPs with high missingness were excluded where SNPs with genotyping call rate < 98%, with low minor allele frequency (MAF) (<0.01), and with potential genotyping error or evolutionary selection (p-value for Hardy-Weinberg equilibrium (HWE) < 1×10^{-6}) were excluded. Individuals with missing genotypes > 2%, with high or low heterozygosity rates (heterozygosity rate >heterozygosity rate mean±3 S.D.), and with evidence of relatedness with other individuals (pi-hat > 0.1875) were excluded [28]. Post quality control, 373 people and 6,865,746 SNPs were retained for further analysis. PLINK1.9 software [29] was utilized for all quality control steps.

2.4 Principal Component Analysis

Principal component analysis (PCA) is a useful tool and has been used in genome-wide association studies (GWAS) to detect effect due to population stratification [30]. Population stratification refers to the non-similarity of samples, diversity of ancestry, and/or potential outliers. It will confound the association of SNPs with a given phenotype, leading to false-positive results [31]. PCA reduces the dimensionality in order to generate the eigenvalues and the eigenvectors of the covariate matrix of the allele frequencies. Linkage disequilibrium pruning (LD-pruning) is commonly used to retain the independent SNPs prior to conduct PCA [32]. The threshold for LD-pruning was the

doi:10.6342/NTU202301905

square of the correlation coefficient (r^2) <0.5 [33], and therefore the SNPs that were highly correlated with the index SNP in the same cluster were removed. A scree plot (Figure 4) was constructed using the first 10 principal components (PCs). These PCs were incorporated into the GWAS analysis to account for any potential population stratification.

2.5 Genome-wide association study for the analysis of 28-day sepsis survival

2.5.1 Survival genome-wide association analysis with Cox proportional hazards regression

Cox proportional hazards regression was to detect the SNPs associated with 28-day survival for patients with sepsis using the ***gwasurvivr*** package (version 1.12.0) in R with age, sex, and principal components (PCs) as covariates [34]. PLINK binary format files (.bed, .bim, and .fam files) were used in order to execute the ***plinkCoxSurv*** function. The genome-wide significant p-value was set as 5.0×10^{-8} post Bonferroni correction to control for multiple testing. To minimize the rate of false positives, Bonferroni correction was performed using the formula 0.05/n, where n represents the total number of tested SNPs [35]. In addition, a p-value of 1.0×10^{-5} was set as the threshold for suggestive association. The hazard ratio (HR) was represented as the effect size of the SNPs.

The equation for genome wide association analysis with the Cox regression was:

$$\log \boldsymbol{h}(\boldsymbol{t};\boldsymbol{X}) = \log\{\boldsymbol{h_0}(\boldsymbol{t})\} + \{\boldsymbol{\beta_1 X_1} + \cdots + \boldsymbol{\beta_p X_p}\}$$

where *t* was survival time (unit: day), X=($X_1, X_2, \ldots, X_p$) was a set of predictor variables comprised of SNPs and the covariates included in the Cox regression, h_0(t) was baseline hazard function, and $\beta_1, \ldots, \beta_k$ were the regression coefficients.

2.5.2 Genome-wide association study using logistic regression

Binary (1, 0) outcome based on whether death occurred or not within a 28-day follow-up

doi:10.6342/NTU202301905

period was used. Logistic regression was performed to select the SNPs associated with sepsis mortality using PLINK1.9, adjusting for age, sex, and principal components (PCs) as covariates. Similar genome-wide and suggestive significant p-value thresholds as that applied in the prior section, were used to identify SNPs in the logistic regression analysis. The odds ratio (OR) provided as the effect size of the SNPs.

The equation for logistic regression utilized was as follows:

$$\boldsymbol{log\left(\frac{Y}{1-Y}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p}$$

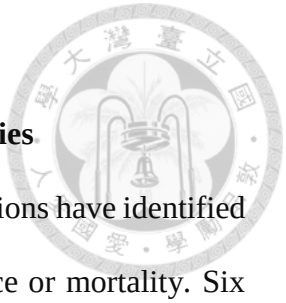
where Y was the probability of the mortality of sepsis (event=1) within 28-day follow-up, X=($X_1, X_2, \ldots, X_p$) was a set of predictor variables comprised of SNPs and the covariates incorporated into the logistic regression, β_0 was the intercept of the regression and β_1, …, β_k were the regression coefficients.

To obtain independent significant SNPs, clumping was applied to remove correlated SNPs within the same LD block and within a 250 kb window of the lead SNP (r^2 >0.5) [36].

2.6 Genomic inflation factor and quantile-quantile plot

To investigate whether an overdispersion of the association test statistics presented by GWAS was caused by the population structure and/or kinship, the genomic inflation factor (λ) and quantile-quantile (Q-Q) plot were investigated. Genomic inflation factor was calculated by dividing the median of the observed chi-squared test statistic by its expected values under the null hypothesis. P-values were divided by genomic inflation factor to correct for inflation due to population structure [37]. Q-Q plot was utilized to assess the divergence of observed p-values from the null hypothesis.

doi:10.6342/NTU202301905



2.7 Validation of variants reported as significant in previous studies

Previous studies or large-scale biobank analyses in non-Asian populations have identified several SNPs that show significant associations with sepsis incidence or mortality. Six studies were selected for validation among the Taiwanese populations. A total of 21 SNPs found by Rautanen et al. [16] have a significant association with sepsis mortality attributed to pneumonia, while 15 SNPs showed a significant association with sepsis mortality related to either pneumonia or abdominal infection. In Scherag et al. [38], 14 SNPs were reported to be associated with 28-day sepsis mortality. Rosier et al. [17] published 12 and 16 SNPs that were associated with 7-day and 7- to 28-day sepsis mortality, respectively. Neale et al. [39], utilizing UK Biobank dataset reported 40 SNPs associated with sepsis susceptibility. A total of 5 SNPs was associated with septic shock-related symptoms or mortality in the study by Hernandez-Beeftink et al. [22].

2.8 Estimation of polygenic risk score (PRS)

The polygenic risk score (PRS) was obtained by a two-step analysis. First, we sampled 80% of individuals (n=299) as base dataset and conducted genome-wide association analysis to obtain the effect sizes of SNPs. The C + T method (clumping and p-value thresholding) is widely used to calculate PRS [40, 41]. The variants were clumped by the squared correlation thresholds, r^2, representing the levels of LD of the nearby variants corrected by their index variant. Hence, the variants were removed if their p-value was larger than significant thresholds. For clumping, we utilized various levels for r^2, including 0.2, 0.4, 0.6, and 0.8. Additionally, we employed multiple thresholds for p-value, namely 1×10^{-5}, 1×10^{-4}, 0.001, 0.01, and 0.05, during our analysis [42]. Clumping and p-value thresholding were performed using PLINK 1.9 software [43]. Second,

doi:10.6342/NTU202301905

individual's PRSs were estimated from the total of risk alleles carried for each variant, weighted by the effect size.

The equation for the PRS of an individual j was:

$$\boldsymbol{PRS_j = \frac{\sum_{i=1}^{N} w_i \times SNP_{ij}}{2 \times M_j}}$$

where $\boldsymbol{N}$ is the number of SNPs included in the genome-wide association studies, $\boldsymbol{w_i}$ is the weight (effect size) of each SNP, $\boldsymbol{SNP_{ij}}$ is the number of risk alleles of the $\boldsymbol{i}$th variant that each $\boldsymbol{j}$th individual had, and $\boldsymbol{M_j}$ is the number of observable variants for each individual $\boldsymbol{j}$. Due to diploidy for human autosomes, $\boldsymbol{M_j}$ is multiplied by 2 to sum the total number of alleles on the loci.

After calculating PRSs, we standardized them by dividing each PRS by its corresponding standard deviation.

Due to the genome-wide association analysis conducted by Cox regression and logistic regression, respectively, there are two GWAS statistical results. Therefore, PRSs, PRS-sepsis, were estimated using results from each GWAS, respectively. In addition, for comparison purposes SNPs reported from previous studies were utilized to calculate another PRSs, PRS-sepsis-pre. Both PRS-sepsis and PRS-sepsis-pre were finalize through C + T methods via concordance probability and 5-fold cross validation. The discrimination ability of PRS-based models was assessed by calculating the concordance probability (Harrell's concordance index, C-index). It was determined by dividing the proportion of concordant pairs by the total number of possible evaluation pairs. [44]. The C-index with five-fold cross validation allowed to choose the optimal C + T thresholds for inclusion of SNPs towards PRS calculations. In contrast, pseudo-R^2 (Nagelkerke's R^2) was used to evaluate the performance of PRS constructed from GWAS using logistic

doi:10.6342/NTU202301905

regression. It was commonly used to indicate the amount of variance that was explained by the predictors in logistic regression. Finally, Kaplan-Meier analysis was used to illustrate the performance of the PRS in the plot and compared the survival function of two groups separated by a cut-off point of PRS score. Kaplan-Meier plot was performed via ***survival*** and ***survminer*** packages in R.

2.9 Five-fold cross-validation for the prediction models

2.9.1 Description for the classification of the prediction models (baseline model, multivariate model, final model and traditional model)

Baseline prediction model included PRS (PRS-sepsis or PRS-sepsis-pre) as the genetic factor, adjusting for age, sex and PCs. Clinical variables selected via a univariate regression were added to the baseline model to construct the multivariate model. Furthermore, to assess the improvement in predictive power offered by the PRS, we used the model containing only the clinical variables, referred to as the traditional model, as the reference. Since there were several significant clinical variables in the univariate regression, least absolute shrinkage and selection operator (Lasso) regression and stepwise regression selection were used to evaluate relatively dominant clinical variables for the final mode.

Lasso regression was commonly employed for feature selection as it incorporates regularization to mitigate overfitting and complexity in models [45]. By adding a penalty term to the sum of squared residuals, the regularization parameter was multiplied by the sum of the absolute values of the coefficients, effectively shrinking the prediction error. Consequently, if a parameter was deemed less influential to the outcome, the penalty term approached zero. This process ultimately led to an improved efficacy in the final model.

doi:10.6342/NTU202301905

Through the utilization of these variable selection methods, we compared the efficacy of the final models incorporating relevant clinical variables and assessed the association between PRS and the probability of 28-day survival. Stepwise regression is a linear regression technique that involves both forward and backward selection [46]. In this approach, each variable is sequentially added to the model and tested for statistical significance. If the variable was significant, it would be retained in the model. Meanwhile, the other variables already present in the model were also evaluated for significance. Any variable deemed insignificant was then removed from the model. This iterative process continued until all included variables remained significant in the model, while any excluded variables failed to reach significance. To summarize, the final model was constructed with the clinical variables selected via the Lasso and stepwise regression methods, along with the genetic factor.

2.9.2 Cox regression analysis and logistic regression analysis

Two outcomes were considered. The first considered of survival and mortality from sepsis within a 28-day follow-up period, and the second was binary (1, 0) outcome based on whether or not death occurred within a 28-day follow-up period. Given the two approaches for assessing the study outcomes, models were developed using survival (Cox regression) and binary outcome (logistic regression), respectively.

After Lasso regression and stepwise regression, final model using Cox regression exhibited the highest C-index. Among the models evaluated, the final model built using logistic regression was selected based on the criterion of achieving the highest area under curve (AUC), the area under receiver operating characteristic curve (ROC). To perform five-fold cross-validation, 80% of the samples were employed as the training set, with the remaining 20% allocated as the testing set for each fold. The C-index and the AUC of the

doi:10.6342/NTU202301905

models were estimated by the average of five-fold cross-validation results. The hazard ratio (HR) and odds ratio (OR) of PRSs were evaluated using whole samples in the testing sets.

doi:10.6342/NTU202301905

Chapter 3: Results

3.1 Descriptive statistics

A cohort of 373 patients diagnosed with sepsis was included in our study. Of these, 39 (10.5%) were with event (death) group and 334 (89.5%) survivors with no event within a 28-day follow-up period. Table 1 presents an overview of the demographic information and clinical features of the patients. The mean age of patients who experienced the event was 69.05 years, whereas the mean age of the survivors was 64.05 years. The mortality rate of male sepsis was higher than that of female sepsis. The group with event included subjects that were underweight, demonstrated higher Charlson comorbidity index and SOFA score, and higher incidence of septic shock and metastatic solid tumor compared to that of the 28-day survivor group. In addition, the patients with event had lower hemoglobin, a higher proportion of hemodialysis treatment, and a higher incidence of respiratory tract infection.

Principal component analysis (PCA) and scree plot were used to estimate population stratification effects if any. To control for confounding due to population stratification in genome-wide survival studies, principal components (PCs) were used as covariates in the regression models. The results of PCA were shown in Figure 3. The scree plot showed that maximum variations were accounted by the first 5 PC (Figure 4). Therefore, the first three, four, and five PCs were set as covariates in the genome-wide survival analysis, respectively, to demonstrate which would make the genomic inflation factor closest to 1.

3.2 Cox regression analysis

3.2.1 Genome-wide survival analysis with continuous survival time

After quality control, a total of 6,865,736 SNPs were used to perform GWAS, adjusted

doi:10.6342/NTU202301905

for age, sex, and the first three PCs, that showed the genomic inflation factor (λ) to be 0.921 (closest to 1). A total of 4 genome-wide significant ($p < 5\times10^{-8}$) SNPs were identified as genetic predictors of the sepsis survival (Figure 5 and Figure 6), of which rs296175 was an intergenic variant to long non-coding RNAs (*LINC01470*) and glutamate ionotropic receptor AMPA type subunit 1 (*GRIA1*), rs138138121 was in the Family With Sequence Similarity 204 Member A (*FAM204A*), rs117040844 was an intergenic variant to StAR Related Lipid Transfer Domain Containing 13 (*STARD13*) and Replication Factor C Subunit 3 (*RFC3*), and rs374290727 was in the Formyl peptide receptor 1 (*FPR1*). The identified SNPs, situated on chromosome 5, 10, 13, and 19, exhibited significant associations with an increased risk of 28-day sepsis mortality (Table 2). Furthermore, a total of 293 SNPs exhibited suggestive significance ($p < 1\times10^{-5}$).

3.2.2 Validation of significant SNPs reported from previous papers

Among the 134 SNPs reported as significantly associated with sepsis incidence or mortality in previous studies (Table 19), 64 SNPs were observed in the present study. However, none of them were validated as genome-wide significant ($p < 5\times10^{-8}$) by Cox regression in our population. Only one of them, rs1573332, had a p-value lower than 0.05.

3.2.3 The performance of polygenic risk score in Cox prediction models

The determination of the number of SNPs included in estimating the polygenic risk score (PRS) was carried out using the C + T method. The C-index was utilized to assess the predictive power of the models with PRS-sepsis (Table 3). Comparison of the discrimination ability of each of the models via C-index were demonstrated in Figure 7. PRS-sepsis with the highest C-index, 0.9523, was estimated from SNPs clumped by $r^2 <$ 0.2 and p-value < 0.01, adjusting for age, sex, the first three PCs. However, this PRS-

doi:10.6342/NTU202301905

sepsis was not feasible to be utilized in the prognostic model due to its construction involving approximately 9,000 SNPs, which did not align with the aim of this study. Therefore, the PRS with the best performance but constructed with a set of fewer than 100 SNPs were chosen as the final PRS-sepsis. The corresponding C-index was 0.9375 and the C + T thresholds were with $r^2 < 0.4$ and p-value $< 1\times10^{-5}$, allowing a total of 86 SNPs to be aggregated for PRS-sepsis construction. The HR of PRS-sepsis were 4.9488 (95% CI: 3.6248–6.7565, p-value $< 2\times10^{-16}$). Additionally, Furthermore, a total of 64 SNPs from previous studies were identified in Taiwanese sepsis patients. All of these SNPs were successfully replicated in the base dataset and subsequently aggregated to construct PRS-sepsis-pre, enabling the evaluation of their relevance and applicability to sepsis patients in Taiwan.

3.2.4 Comparison of the performance between Cox prediction models

The demographic and clinical characteristics significantly associated with survival time using univariate Cox regression are listed in Table 4. Cox prediction models were evaluated using a five-fold cross-validation approach. Baseline model and multivariate model performed high discrimination abilities with concordance value (C-index). The C-index (s.e.) for the baseline model using the training set was 0.9389 (0.0210), while for the testing set, it was 0.9402 (0.1169) (Table 5). Considering the clinical variables selected via univariate Cox regression, the C-index (s.e.) of multivariate model using training set was 0.9513 (0.0208), whereas for the testing set, it was 0.8304 (0.1394). Compared to traditional model, the inclusion of PRS-sepsis significantly improved the C-index by 8.02% in the training set (p=0.016) and demonstrated a non-significant improvement of 23.14% in the testing set (p=0.44) (Figure 8).

After performing Lasso regression selection, three clinical variables, namely Charlson

doi:10.6342/NTU202301905

comorbidity index, hemodialysis, and indirect bilirubin, were chosen to be incorporated in the final model. The model exhibited a C-index (s.e.) of 0.9254 (0.0320) in the training set and 0.8811 (0.1275) in the testing set (Table 6). An increase of one standard deviation in PRS-sepsis among sepsis patients was associated with a nearly 4% increase in the probability of mortality (HR=4.04, 95% CI: 2.78–5.86, $p=2.59 \times 10^{-13}$) (Table 7). Compared to traditional model, the inclusion of PRS-sepsis significantly improved the C-index of the final model by 5.85% in the training set (p=0.032) and by 15.18% in the testing set (p=0.032) (Figure 9). On the other hand, after performing stepwise regression selection, two clinical variables, namely metastatic solid tumor and hemoglobin, were chosen to be included in the final model. The model exhibited a C-index (s.e.) of 0.9467 (0.0234) and 0.9298 (0.1428) on the training set and testing set, respectively. As sepsis patients improved one standard deviation of PRS-sepsis, the probability of the mortality would increase nearly 4% (HR=3.96, 95% CI: 2.71–5.78, $p=9.23\times10^{-13}$). Compared to traditional model, when adding PRS-sepsis, the C-index of final model was significantly improved by 9.3% using the training set (p=0.032) and improved by 15.23% using the testing set (p=0.032).

The C-index (s.e.) for the baseline PRS-sepsis-pre model was 0.7631 (0.0494) in the training set and 0.7772 (0.1155) in the testing set. When considering the significant clinical variables via univariate Cox regression, the C-index (s.e.) of multivariate PRS-sepsis-pre model was 0.8813 (0.0372) in the training set and 0.7134 (0.1377) in the testing set. Compared to traditional model, when adding PRS-sepsis-pre, the C-index of multivariate model demonstrated a non-significant improvement of 1.02% in the training set (p=0.92), as well as a non-significant improvement of 0.46% in the testing set (p=0.69). After performing Lasso regression selection, the model exhibited a C-index (s.e.) of 0.8826 (0.0367) in the training set and 0.7597 (0.1393) in the testing set. However, the

doi:10.6342/NTU202301905

probability of mortality was not significantly different between sepsis patients with an increase of one standard deviation of PRS-sepsis-pre (HR=9.11, 95% CI: 0.32–259.54, p=0.1960). Compared to traditional model, the incorporation of PRS-sepsis-pre resulted in a 1.57% improvement in the C-index of the final model in the training set (p=0.31) and a 1.92% improvement in the testing set (p=0.62), although these improvements were not statistically significant. On other hand, after performing stepwise regression selection, the model exhibited a C-index (s.e.) of 0.8729 (0.0393) and 0.7808 (0.1393) on the training set and testing set, respectively. Nevertheless, there was no statistically significant difference in the probability of mortality when comparing sepsis patients with an increase of one standard deviation in PRS-sepsis-pre (HR=6.93, 95% CI: 0.34–142.25, p=0.2091). When compared to the traditional model, the inclusion of PRS-sepsis-pre in the final model led to a 1.92% increase in the C-index in the training set (p=0.23) and a 0.33% increase in the testing set (p=1.00). However, these improvements were not statistically significant.

3.2.5 Kaplan-Meier analysis

Sepsis patients were grouped according to the value of PRS-sepsis. According to the ascending PRS-sepsis values, patients in the top 33%, the highest one-third, were classified as the high-PRS group, the remaining 67% into low-PRS group. The survival curves of two groups did not overlap in the Kaplan-Meier (K-M) plot, and the hazard functions were significantly different between two groups ($p < 0.0001$) (Figure 10). The probability of mortality was higher among sepsis patients in the high-PRS group as opposed to those in the low-PRS group. The PRS-sepsis classes explained a great portion of the variation in 28-day survival in the K-M plot.

doi:10.6342/NTU202301905

3.3 Logistic regression analysis

3.3.1 Genome-wide survival analysis with binary death event

Genome-wide survival study with adjustment for age, sex, and the first four PCs showed the genomic inflation factor (λ) was 0.97 which was the closest to 1. However, none of these SNPs exhibited a statistically significant association at the genome-wide level ($p < 5 \times 10^{-8}$) with sepsis mortality. A total of 11 SNPs was identified as suggestively significant ($p < 1 \times 10^{-5}$) after removal of the dependent SNPs (Figure 12, Table 10).

3.3.2 Validation of significant SNPs reported from previous studies

Out of 134 SNPs significantly associated with sepsis incidence or mortality as reported in previous studies (Table 19), 59 were observed in the present study. However, all of them failed to replicate ($p < 5 \times 10^{-8}$) using logistic regression. Only one of them, rs1573332, had a p-value lower than 0.05.

3.3.3 Polygenic risk score performance in logistic regression

The C + T (clumping and p-value threshold) method was used to select the appropriate SNPs to be aggregated into a polygenic risk score. When incorporating the PRS into the logistic prediction models, the predictive ability of the models was assessed using the R^2 (Table 11). The comparison of the individual R^2 of the models based on different PRSs are shown in Figure 14. PRS-based predictive model had highest R^2, 0.795, was estimated from SNPs clumped with $r^2 < 0.8$ and p-value < 0.05 threshold, adjusting for age, sex, the first five PCs, but including almost 56,000 SNPs. Similar to the SNPs selection carried out in Cox regression, we focused only on the C + T thresholds that allowed us to select less than 100 SNPs to align with the aim of this study. Among the evaluated SNPs, a subset of 94 SNPs satisfied the threshold of $r^2 < 0.2$ and p-value $< 1 \times 10^{-4}$. Notably, these

doi:10.6342/NTU202301905

SNPs demonstrated the highest R^2 value of 0.743, which was utilized to generate the PRS-sepsis score. The OR of PRS-sepsis was 40.246 (95% CI: 9.893–163.720, p-value < 2.45×10^{-7}). Additionally, a total of 59 SNPs from previous studies were identified in Taiwanese sepsis patients. Among them, 58 SNPs were replicated in base dataset and subsequently utilized to generate the PRS-sepsis-pre. However, due to a lower minor allele frequency (MAF=0.016), one out of the 59 SNPs did not exhibit variation among the samples in the base dataset.

3.3.4 Comparison of the performance between logistic prediction models

The demographic and clinical characteristics significantly associated with sepsis mortality via univariate logistic regression is demonstrated in Table 12. To assess the performance of the logistic prediction models, a five-fold cross-validation approach was employed. The AUC of the prediction models was used to assess the predictive power. The AUC for baseline PRS-sepsis model was 0.971 (95% CI: 0.919–1.000) (Table 13). Considering the clinical variables selected via univariate logistic regression, the AUC of multivariate PRS-sepsis model was 0.818 (95% CI: 0.622–0.978). Compared to traditional model, when adding PRS-sepsis, the AUC of multivariate model was improved by 20.5% but not significant (p=1.00) (Figure 15).

After performing Lasso regression selection, five clinical variables, namely Charlson comorbidity index, metastatic solid tumor, hemoglobin, hemodialysis, and transfusion, were chosen to be incorporated in the final model. The model exhibited an AUC of 0.891 (95% CI: 0.745–0.998) (Table 14). An increase of one standard deviation in PRS-sepsis among sepsis patients was associated with a nearly 54% increase in the probability of mortality (OR=53.98, 95% CI: 8.97–324.61, $p=1.32\times10^{-5}$) (Table 15). Compared to traditional model, the inclusion of PRS-sepsis improved the C-index of the final model

doi:10.6342/NTU202301905

by 11.5%. However, this improvement was not statistically significant (p=0.43) (Figure 16). On the other hand, after performing stepwise regression selection, two clinical variables, namely hemoglobin and hemodialysis, were chosen to be included in the final model. The model exhibited an AUC of 0.918 (95% CI: 0.785–1.000). As sepsis patients improved one standard deviation of PRS-sepsis, the probability of the mortality would increase nearly 50% (OR=49.78, 95% CI: 8.88–279.14, $p=8.92\times10^{-6}$). Compared to traditional model, when adding PRS-sepsis, the AUC of multivariate model was improved by 11.7%. However, this improvement was not statistically significant (p=0.37). The AUC for the baseline PRS-sepsis-pre model was 0.75 (95% CI: 0.59–0.911). When considering the significant clinical variables via univariate Cox regression, the AUC of multivariate PRS-sepsis-pre model was 0.761 (95% CI: 0.547–0.946). Compared to traditional model, when adding PRS-sepsis-pre, the AUC of multivariate model was improved by 14.8%. However, this improvement was not statistically significant (p=1.00). After performing Lasso regression selection, the model exhibited an AUC of 0.776 (95% CI: 0.573–0.943). An increase of one standard deviation in PRS-sepsis-pre among sepsis patients was associated with a nearly 3.5% increase in the probability of mortality (OR=3.63, 95% CI: 1.57–8.38, p=0.0026). Compared to traditional model, however, the incorporation of PRS-sepsis-pre did not result in any improvement in the AUC of the final model and the improvement was not significant (p=1.00). On the other hand, after performing stepwise regression selection, the model exhibited an AUC of 0.822 (95% CI: 0.611–0.995). As sepsis patients improved one standard deviation of PRS-sepsis-pre, the probability of the mortality would increase nearly 3% (OR=2.77, 95% CI: 1.49–5.18, p=0.0014). When compared to traditional model, when adding PRS-sepsis-pre, the AUC of final model was improved by 2.1%. However, this improvement was not statistically significant (p=1.00)

doi:10.6342/NTU202301905

3.4 Investigation of correlation between genetic factor and clinical variables

The correlation matrix was employed to illustrate the relationship between the variables. We used a heat map to illustrate the correlation matrix (Figure 15). The PRS-sepsis constructed from survival-GWAS using Cox regression was highly associated with the PRS-sepsis constructed from GWAS using logistic regression and the length of hospital stay for the sepsis patients.

In addition, we evaluated the cumulative effects of the SNPs derived from survival-GWAS using Cox regression and GWAS using logistic regression to examine their associations with clinical characteristics. These associations were presented in Table 9 and Table 16.

doi:10.6342/NTU202301905

Chapter 4: Discussion

4.1 Main finding

Patients with sepsis may rapidly develop organ failure and progress to septic shock. Therefore, early recognition for patients with sepsis will have them promptly receive the treatments against aggravation. However, the risk of death can vary within the same time frame. The heterogeneity in sepsis is related to clinical features, host-pathogen interactions, treatments, and/or genetic predisposition [47], leading to the main limitation for predicting the prognosis. To account for this variability and to improve the predictive power, we included individuals' survival time in our analysis and identified 4 genome-wide significant SNPs ($p < 5\times10^{-8}$) associated with survival time. We explored the correlation between these genes and the 28-day mortality from sepsis, identifying 11 SNPs that demonstrated a significant association with sepsis mortality ($p < 1\times10^{-5}$).

Several studies reported a variety of SNPs significantly associated with either the sepsis incidence or sepsis mortality using GWAS approach, however, most of them focused on the non-Asian countries. This is the one of the first study to validate if significant associations among the Caucasians were observed among Asian patients. Moreover, none of the 4 genome-wide significant SNPs demonstrated in our data overlapped with the previous studies. Using 1000 Genomes phase 3 dataset reference panel, the minor allele frequency of three genome-wide significant SNPs even exhibited zero among non-Asians, including Americans, Africans, Europeans, and South Asians. On the other hand, most of the minor allele frequency of those significant SNPs in previous studies were different from Taiwan sepsis patients. Hence, the future studies will be conducted to verify those on the different ethnicities.

doi:10.6342/NTU202301905

The genetic factors demonstrated the improvement of the predictive power. When adding the genetic factor, PRS-sepsis, along with the clinical characteristics selected using Lasso regression in the Cox regression models, the predictive power of the final model in the training set and the testing set (C-index=0.9254 and 0.8811) increased respectively 5.85% and 15.18% in comparison with the traditional models (C-index=0.8669 and 0.8537) which constituted of only clinical predictors. If the clinical variables selected using stepwise regression, the predictive power of the final model in the training set and the testing set (C-index=0.9467 and 0.9298) increased respectively 9.3% and 15.23% in comparison with the traditional models. All of these improvements of the predictive power reached statistical significance. However, the final models incorporated PRS-sepsis-pre did not have a significantly improvement in the training set and the testing set compared with the traditional models. In logistic regression models, the addition of either PRS-sepsis or PRS-sepsis-pre, along with the clinical features selected using Lasso regression, did not result in significantly higher predictive power for the final model (AUC=0.891 and 0.776) compared to the traditional clinical features models (AUC=0.776). Similarly, when using the clinical features selected via stepwise regression, the final model's predictive power (AUC=0.918 and 0.822) did not exhibit a significant improvement compared to the traditional clinical features models (AUC=0.801). However, both PRS-sepsis and PRS-sepsis-pre demonstrated positive associations with 28-day sepsis survival.

Sepsis patients with a higher PRS were found to have an increased risk of death within 28 days of diagnosis, as indicated by the prediction models constructed using Cox regression or logistic regression. In comparison to previous studies that primarily focused on building prediction models for sepsis survival using clinical characteristics alone, our study incorporates PRS as a genetic factor. When evaluating in-hospital and ICU

doi:10.6342/NTU202301905

mortality in Vietnamese sepsis patients using the Quick Sequential Organ Failure Assessment (qSOFA) score, these models exhibited moderate discriminatory ability. (AUC=0.610 and 0.619) [48]. In a prior study, a predictive model for sepsis-associated acute kidney injury was formulated using clinical characteristics. The model has a C-index of 0.711 and 0.712 in the training and validation datasets, respectively [49].

We observed that the hazard functions of high-PRS and low-PRS subgroups were well discriminated, demonstrating the lower survival probability in the high-PRS group. Significant SNPs reported in previous studies could not replicated in our data, possibly attributable to the limited size of our study cohort [50]. After aggregating the effect sizes of these SNPs to estimate PRS-sepsis-pre, we found a significant association with increased mortality of sepsis within a 28-day follow-up period.

4.2 Extended inference

After PRS-sepsis was arranged in ascending powers, we divided sepsis patients into high-PRS and low-PRS groups using different cut-off points. These include 50%, 55%, 60%, 65%, 67% (two-third), 70%, 75%, 80%, 85%, 90%, and 95% of the individuals with the lowest scores being classified as the low-PRS group. The best cut-off was evaluated by the C-index of the prediction model. The results showed that the C-index of each cut-off point were similar (Figure 11 and Table 8). The 67% was used to be a cut-off point, since it demonstrated the relatively higher predictive power of the model. The survival probability of high-PRS group was significantly lower than low-PRS group. Therefore, identifying the patients at high risk of sepsis mortality within 28 days may lead to better interventions and treatments to reduce the severity of outcomes.

Even though the PRS was initially estimated using PLINK, we employed an additional software, PRSice-2, to verify the precision of the PRS through the C + T method. PRSice-

doi:10.6342/NTU202301905

2 followed a similar approach to PLINK, estimating the cumulative risk alleles carried for each variant, weighted by their effect size. However, PRSice-2 offered more flexibility in terms of p-value thresholding [51]. All PRSs calculated using PRSice-2 conducted standardization, where each PRS was divided by its corresponding standard deviation. Initially, PRSice-2 was utilized the same clumping and threshold criterions as PLINK for estimation, employing r^2 values of 0.2, 0.4, 0.6, and 0.8, along with p-value thresholds of 1×10^{-5}, 1×10^{-4}, 0.001, 0.01, and 0.05. For the purposes of the study, results were retained if the number of SNPs was less than 100. The best C-index and R^2 are similar to those estimated using PLINK (Figure 18). In addition, to evaluate the predictive power of Cox regression models incorporating additional PRSs estimated using different p-value thresholds, while still considering sets of fewer than 100 SNPs, we conducted simulations with increasing p-value thresholds of 1×10^{-5} by 1×10^{-6} at each step. Among these simulations, the inclusion of 76 SNPs at a p-value threshold of 1×10^{-5} and an r^2 value of 0.2 resulted in the highest C-index of 0.9290 (Table 17).

On the other hand, when using the p-value threshold of 1×10^{-5} in logistic regression, only a small number of SNPs were included. This led to a large number of simulations required when considering sets of fewer than 100 SNPs, if we were to use a similar p-value threshold as in Cox regression. Therefore, we conducted simulations with a p-value threshold increasing by 1×10^{-5} at each step. Among these simulations, the inclusion of 88 SNPs at a p-value threshold of 1.1×10^{-4} and an r^2 value of 0.2 resulted in the highest AUC of 0.760 (Table 18). Compared to the PRS estimated by PLINK, the C-index was reduced by 0.85% using PRSice-2, but the R^2 increased by 1.7%. Consequently, the two methods were similar for the estimation of the PRSs.

A total of four SNPs was significantly associated with sepsis survival time. These SNPs are rs138138121 in *FAM204A*, rs374290727 in *FPR1*, rs296175 in an intergenic variant

doi:10.6342/NTU202301905

between *LINC01470* and *GRIA1*, and rs117040844 an intergenic variant between *STARD13* and *RFC3*. A prior study has provided evidence of the involvement of the *FPR1* in bacterial detection, inflammation regulation, and cancer immunosurveillance [52]. When confronted with inflammation, FPR1 receptor swiftly responds by releasing pro-inflammatory cytokines, binding with the ligands released by bacteria [53]. Furthermore, the injured tissue has the potential to induce the recruitment of neutrophils by FPR1 receptor, ultimately resulting in systemic inflammatory response syndrome [54]. *RFC3* was found to participate the pathway of the enhancement of virus spread [55] and *STARD13* is associated with cell motility and growth [56]. However, the evidence demonstrated that *GRIA1* variants are implicated in brain development [57]. The specific function related to *FAM204A* remains uncertain at present.

4.3 Strengths and Limitations

An important strength of our study is the incorporation of the genetic factor, PRS-sepsis, into the prediction models, resulting in enhanced predictive abilities of the models. The panel of genetic variations has the potential to identify patients with a high risk of mortality during a 28-day follow-up duration. As a result, patients who were stratified based on a 70% cut-off for PRS-sepsis demonstrated a substantial differentiation in their survival probabilities. Moreover, we integrated SNPs that have previously shown significant associations with sepsis susceptibility or sepsis mortality in studies involving individuals of non-Asian ancestry. The PRS-sepsis-pre, derived from these SNPs, was discovered to associate with sepsis mortality within a 28-day follow-up period.

It is important to note the limitations inherent in our study. First, our study is conducted with the relatively small sample size. Some potential associations may not be observed by the genome-wide p-value threshold, either in Cox regression or in logistic regression.

doi:10.6342/NTU202301905

Although we sampled 80% of individuals to conduct GWAS and obtain the effect size of them, we could not use the only remaining 20% as the target dataset to calculate the polygenic risk score and build prediction models due to limited statistical power. Second, there is a lack of control over the potential relationship between sources of infection and mortality because we could not collect information on the sources of infection of patients.

doi:10.6342/NTU202301905

Chapter 5: List of figures

doi:10.6342/NTU202301905

Figure 14. The performance of PRS-sepsis using C + T method in logistic regression model.

Figure 15. Comparison of the AUC of the baseline models and the multivariate models.

ns: non-significant ($p \geq 0.05$)

Figure 16. Comparison of the AUC of the final models with Lasso and stepwise selection methods.

ns: non-significant ($p \geq 0.05$)

Figure 17. Correlation matrix.

HGB: hemoglobin, ED.stay.hour: the number of hours of ED stay, CCI: Charlson comorbidity index, dbp: diastolic blood pressure, MAP: mean arterial pressure , sbp: systolic blood pressure, PLT: platelet counts, BMI: body mass index, GLU: Glutamine, PRS_logistic: PRS-sepsis was aggregated the effect sizes of the SNPs using logistic regression, PRS_Cox: PRS-sepsis was aggregated the effect sizes of the SNPs using Cox regression, BIL_T: indirect bilirubin, BIL_D: direct bilirubin, SOFA: sequential organ failure assessment score, CREAT: creatinine, resp.rate: respiration rates, WBC: white blood cells, spo2: pulse oximetry, SF.ratio: Spo2/Fio2 (pulse oximetry divided by fraction of inspired oxygen), IP_Day: the number of days of hospital stay, SURVT: survival time (unit: days), CRP: C-Reactive protein

Figure 18. The performances of PRS-sepsis in models were evaluated using PRSice-2 software (C + T method).

(A) PRS-sepsis was calculated using Cox regression. (B) PRS-sepsis was calculated using logistic regression.

doi:10.6342/NTU202301905

Chapter 6: List of tables

doi:10.6342/NTU202301905

doi:10.6342/NTU202301905

Chapter 7: Reference

1. Rudd KE *et al.* Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study *Lancet* 2020;**395**:200-211.
2. Cohen J *et al.* Sepsis: a roadmap for future research *Lancet Infect Dis* 2015;**15**:581-614.
3. Kwizera A *et al.* National intensive care unit bed capacity and ICU patient characteristics in a low income country *BMC Res Notes* 2012;**5**:475.
4. Chen YJ *et al.* Epidemiology of sepsis in Taiwan *Medicine (Baltimore)* 2019;**98**:e15725.
5. Wang YG *et al.* High 28-day mortality in critically ill patients with sepsis and concomitant active cancer *J Int Med Res* 2018;**46**:5030-5039.
6. Abdullah S *et al.* Prognostic Accuracy of SOFA, qSOFA, and SIRS for Mortality Among Emergency Department Patients with Infections *Infect Drug Resist* 2021;**14**:2763-2775.
7. Prachanukool T *et al.* The 28-Day Mortality Outcome of the Complete Hour-1 Sepsis Bundle in the Emergency Department *Shock* 2021;**56**:969-974.
8. Song JE *et al.* Mortality Risk Factors for Patients with Septic Shock after Implementation of the Surviving Sepsis Campaign Bundles *Infect Chemother* 2016;**48**:199-208.
9. Oami T *et al.* Association between low body mass index and increased 28-day mortality of severe sepsis in Japanese cohorts *Sci Rep* 2021;**11**:1615.
10. Sørensen TI *et al.* Genetic and environmental influences on premature death in adult adoptees *N Engl J Med* 1988;**318**:727-732.

doi:10.6342/NTU202301905

11. Bumgarner R. Overview of DNA microarrays: types, applications, and their future *Curr Protoc Mol Biol* 2013;**Chapter 22**:Unit 22.21.
12. Dehghan A. Genome-Wide Association Studies *Methods Mol Biol* 2018;**1793**:37-49.
13. Mbarek H *et al.* A genome-wide association study of chronic hepatitis B identified novel risk locus in a Japanese population *Hum Mol Genet* 2011;**20**:3884-3892.
14. Pereyra F *et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation *Science* 2010;**330**:1551-1557.
15. van Manen D *et al.* Genome-wide association studies on HIV susceptibility, pathogenesis and pharmacogenomics *Retrovirology* 2012;**9**:70.
16. Rautanen A *et al.* Genome-wide association study of survival from sepsis due to pneumonia: an observational cohort study *Lancet Respir Med* 2015;**3**:53-60.
17. Rosier F *et al.* Genetic Predisposition to the Mortality in Septic Shock Patients: From GWAS to the Identification of a Regulatory Variant Modulating the Activity of a CISH Enhancer *Int J Mol Sci* 2021;**22**.
18. Dudbridge F. Polygenic Epidemiology *Genet Epidemiol* 2016;**40**:268-272.
19. Gouveia C *et al.* Genome-wide association of polygenic risk extremes for Alzheimer's disease in the UK Biobank *Sci Rep* 2022;**12**:8404.
20. Elliott J *et al.* Predictive Accuracy of a Polygenic Risk Score-Enhanced Prediction Model vs a Clinical Risk Score for Coronary Artery Disease *Jama* 2020;**323**:636-645.
21. D'Urso S *et al.* Septic Shock: A Genomewide Association Study and Polygenic Risk Score Analysis *Twin Res Hum Genet* 2020;**23**:204-213.
22. Hernandez-Beeftink T *et al.* A genome-wide association study of survival in patients with sepsis *Crit Care* 2022;**26**:341.

doi:10.6342/NTU202301905

23. Wang HE *et al.* Long-term mortality after community-acquired sepsis: a longitudinal population-based cohort study *BMJ Open* 2014;**4**:e004283.
24. Reitz KM *et al.* Association Between Time to Source Control in Sepsis and 90-Day Mortality *JAMA Surg* 2022;**157**:817-826.
25. Singer M *et al.* The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3) *Jama* 2016;**315**:801-810.
26. Zheng HF *et al.* Performance of genotype imputation for low frequency and rare variants from the 1000 genomes *PLoS One* 2015;**10**:e0116487.
27. Delaneau O *et al.* Improved whole-chromosome phasing for disease and population genetic studies *Nat Methods* 2013;**10**:5-6.
28. Marees AT *et al.* A tutorial on conducting genome-wide association studies: Quality control and statistical analysis *Int J Methods Psychiatr Res* 2018;**27**:e1608.
29. Truong VQ *et al.* Quality Control Procedures for Genome-Wide Association Studies *Curr Protoc* 2022;**2**:e603.
30. Elhaik E. Principal Component Analyses (PCA)-based findings in population genetic studies are highly biased and must be reevaluated *Sci Rep* 2022;**12**:14683.
31. Abraham G *et al.* Fast principal component analysis of large-scale genome-wide data *PLoS One* 2014;**9**:e93766.
32. Laurie CC *et al.* Quality control and quality assurance in genotypic data for genome-wide association studies *Genet Epidemiol* 2010;**34**:591-602.
33. Caro-Consuegra R *et al.* Identifying signatures of positive selection in human populations from North Africa *Sci Rep* 2023;**13**:8166.
34. Rizvi AA *et al.* gwasurvivr: an R package for genome-wide survival analysis *Bioinformatics* 2019;**35**:1968-1970.
35. Pe'er I *et al.* Estimation of the multiple testing burden for genomewide association

doi:10.6342/NTU202301905

studies of nearly all common variants *Genet Epidemiol* 2008;**32**:381-385.

36. Adam Y *et al.* Performing post-genome-wide association study analysis: overview, challenges and recommendations *F1000Res* 2021;**10**:1002.

37. van Iterson M *et al.* Controlling bias and inflation in epigenome- and transcriptome-wide association studies using the empirical null distribution *Genome Biol* 2017;**18**:19.

38. Scherag A *et al.* Genetic Factors of the Disease Course after Sepsis: A Genome-Wide Study for 28Day Mortality *EBioMedicine* 2016;**12**:239-246.

39. M. NB. UK Biobank GWAS Round 2 Results [Internet] 2018, Website: www.nealelab.is/uk-biobank/.

40. Zhang H *et al.* Novel Methods for Multi-ancestry Polygenic Prediction and their Evaluations in 5.1 Million Individuals of Diverse Ancestry *bioRxiv* 2023:2022.2003.2024.485519. 10.1101/2022.03.24.485519.

41. Privé F *et al.* Efficient Implementation of Penalized Regression for Genetic Risk Prediction *Genetics* 2019;**212**:65-74.

42. Privé F *et al.* Making the Most of Clumping and Thresholding for Polygenic Scores *Am J Hum Genet* 2019;**105**:1213-1221.

43. Choi SW *et al.* Tutorial: a guide to performing polygenic risk score analyses *Nat Protoc* 2020;**15**:2759-2772.

44. Mazurowski MA *et al.* Imaging descriptors improve the predictive power of survival models for glioblastoma patients *Neuro Oncol* 2013;**15**:1389-1394.

45. Satheeshkumar PS *et al.* Feature selection and predicting chemotherapy-induced ulcerative mucositis using machine learning methods *Int J Med Inform* 2021;**154**:104563.

46. Sudheer Kumar *et al.* Comparison of Lasso and stepwise regression technique for

doi:10.6342/NTU202301905

wheat yield prediction *Journal of Agrometeorology* 2019;**21**:188-192.

47. Fohner AE *et al.* Assessing clinical heterogeneity in sepsis through treatment patterns and machine learning *J Am Med Inform Assoc* 2019;**26**:1466-1477.

48. Do SN *et al.* Predictive validity of the quick Sequential Organ Failure Assessment (qSOFA) score for the mortality in patients with sepsis in Vietnamese intensive care units *PLoS One* 2022;**17**:e0275739.

49. Fan C *et al.* A new prediction model for acute kidney injury in patients with sepsis *Ann Palliat Med* 2021;**10**:1772-1778.

50. Asif H *et al.* GWAS significance thresholds for deep phenotyping studies can depend upon minor allele frequencies and sample size *Mol Psychiatry* 2021;**26**:2048-2055.

51. Choi SW *et al.* PRSice-2: Polygenic Risk Score software for biobank-scale data *Gigascience* 2019;**8**.

52. Page MJ *et al.* The Role of Lipopolysaccharide-Induced Cell Signalling in Chronic Inflammation *Chronic Stress (Thousand Oaks)* 2022;**6**:24705470221076390.

53. Vacchelli E *et al.* The ambiguous role of FPR1 in immunity and inflammation *Oncoimmunology* 2020;**9**:1760061.

54. Zhang Q *et al.* Circulating mitochondrial DAMPs cause inflammatory responses to injury *Nature* 2010;**464**:104-107.

55. Panda D *et al.* Triad of human cellular proteins, IRF2, FAM111A, and RFC3, restrict replication of orthopoxvirus SPI-1 host-range mutants *Proc Natl Acad Sci U S A* 2017;**114**:3720-3725.

56. Hanna S *et al.* StarD13 is a tumor suppressor in breast cancer that regulates cell motility and invasion *Int J Oncol* 2014;**44**:1499-1511.

doi:10.6342/NTU202301905

57. Ismail V *et al.* Identification and functional evaluation of GRIA1 missense and truncation variants in individuals with ID: An emerging neurodevelopmental syndrome *Am J Hum Genet* 2022;**109**:1217-1241.

doi:10.6342/NTU202301905

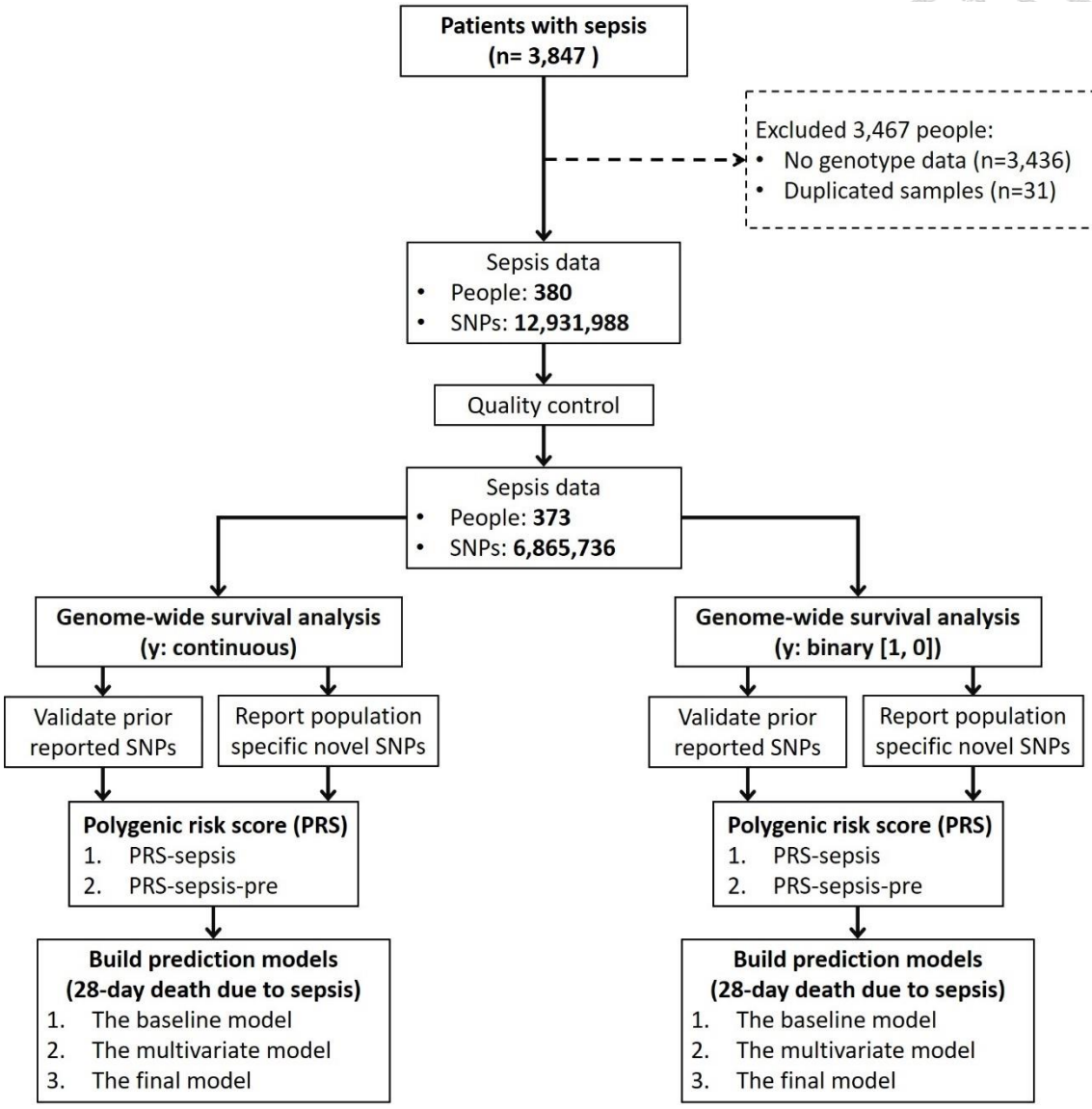


Figure 1. Study flow chart

doi:10.6342/NTU202301905

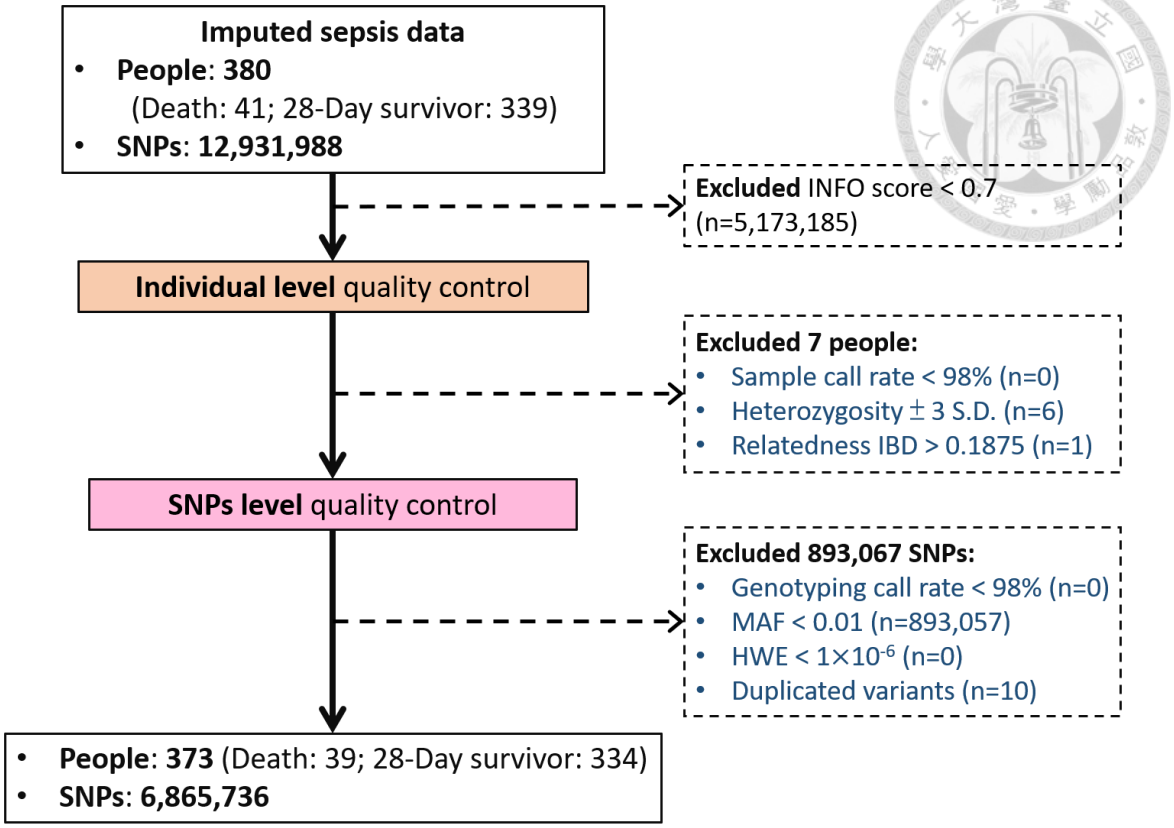


Figure 2. Quality control steps

doi:10.6342/NTU202301905

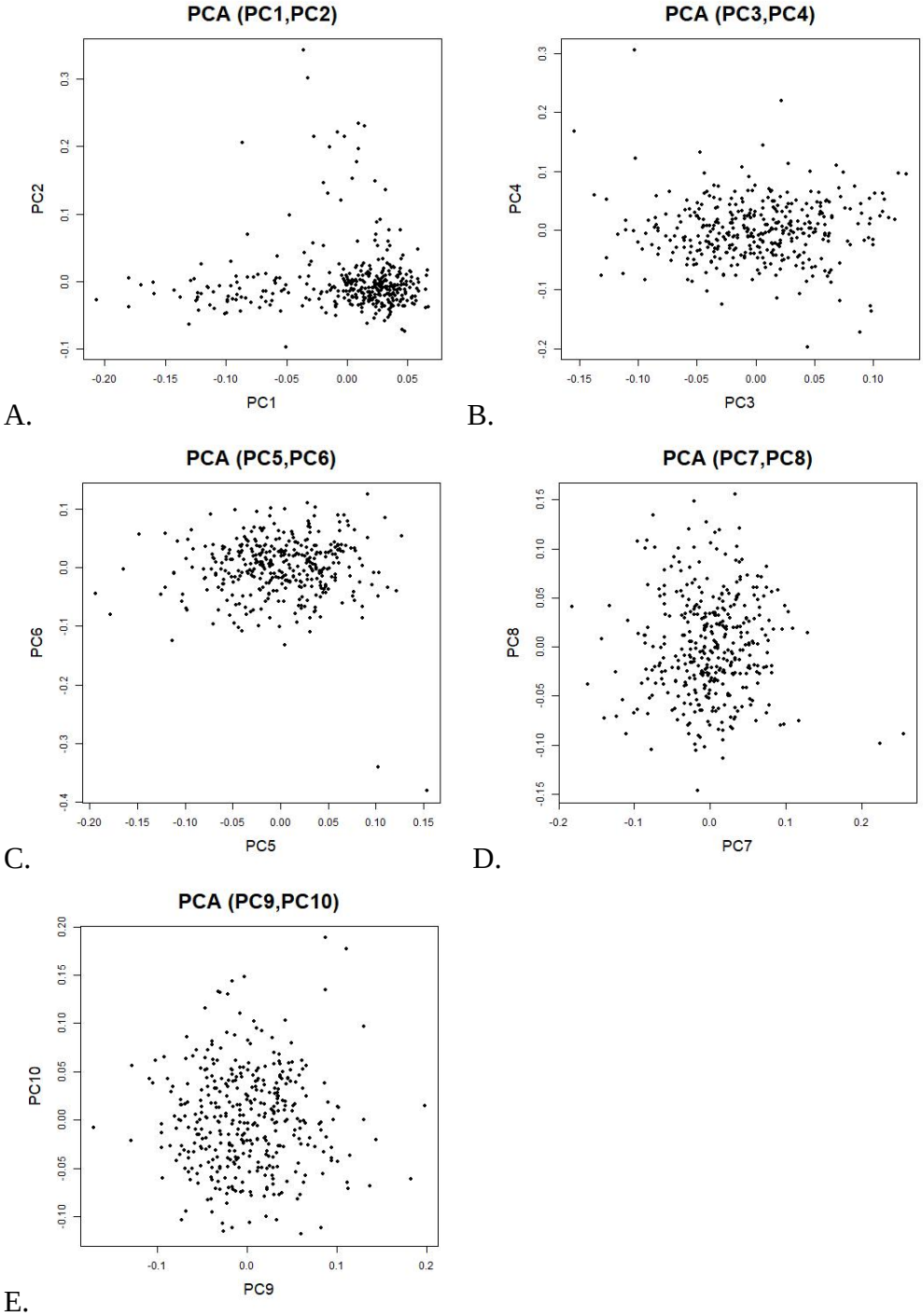


Figure 3. PCA plot for the first 10 principal components (PCs)

doi:10.6342/NTU202301905

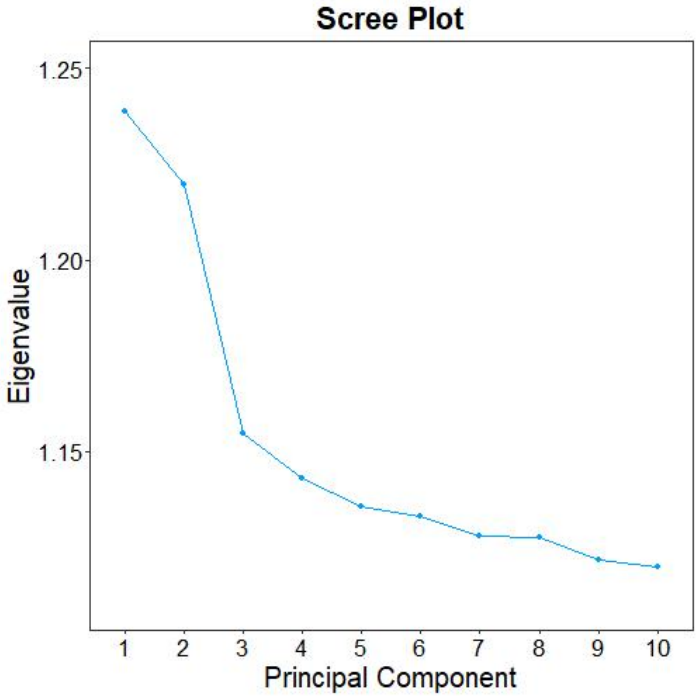


Figure 4. Scree plot for the eigenvalue of the first 10 principal components

doi:10.6342/NTU202301905

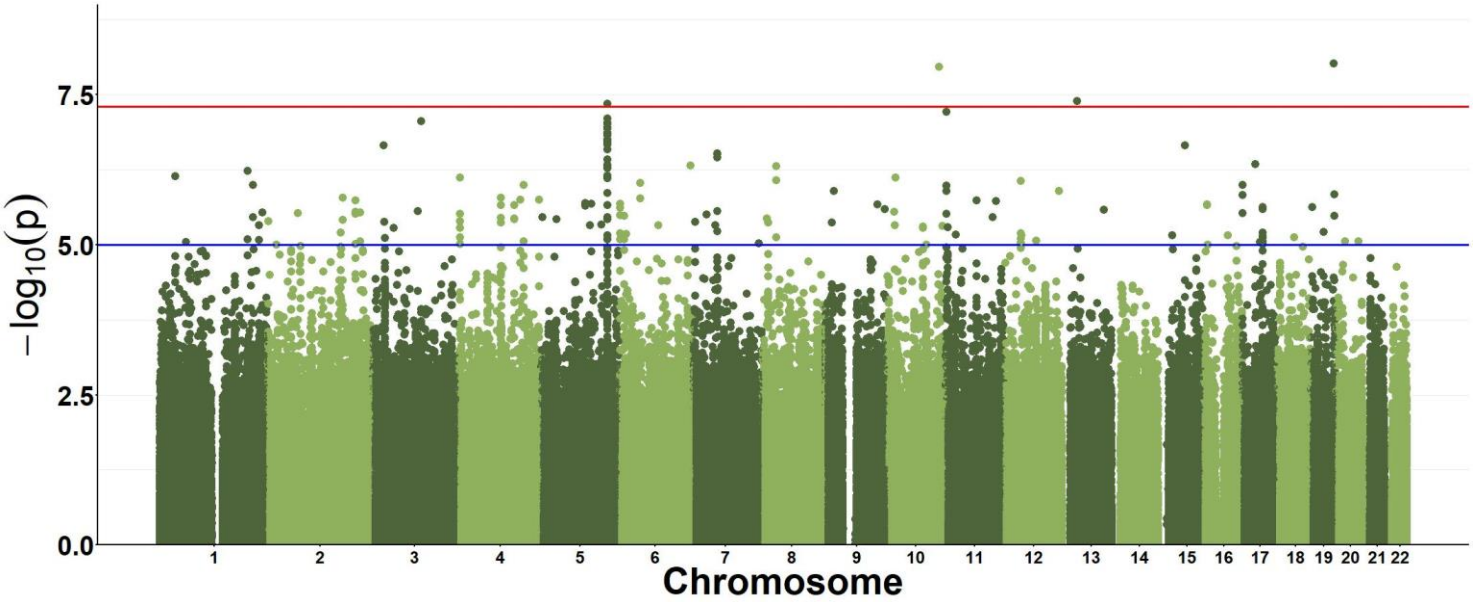


Figure 5. Manhattan plot of GWAS result using Cox regression

doi:10.6342/NTU202301905

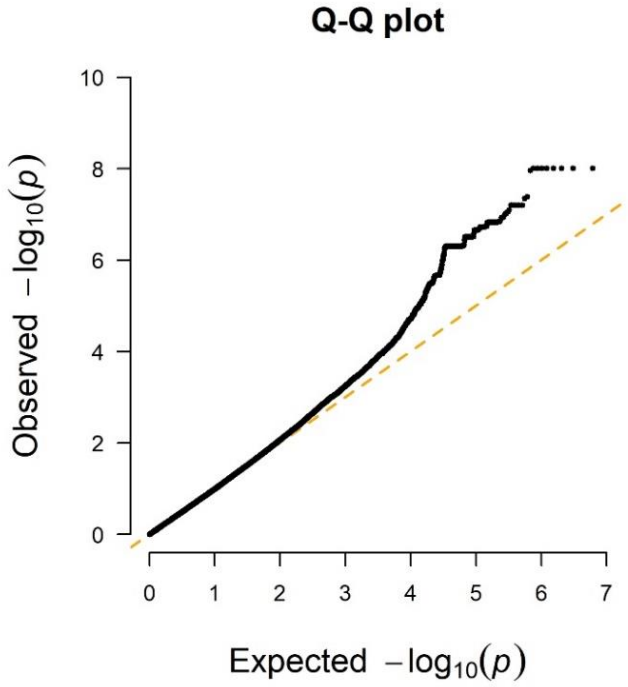


Figure 6. Q-Q plot of GWAS result using Cox regression

doi:10.6342/NTU202301905

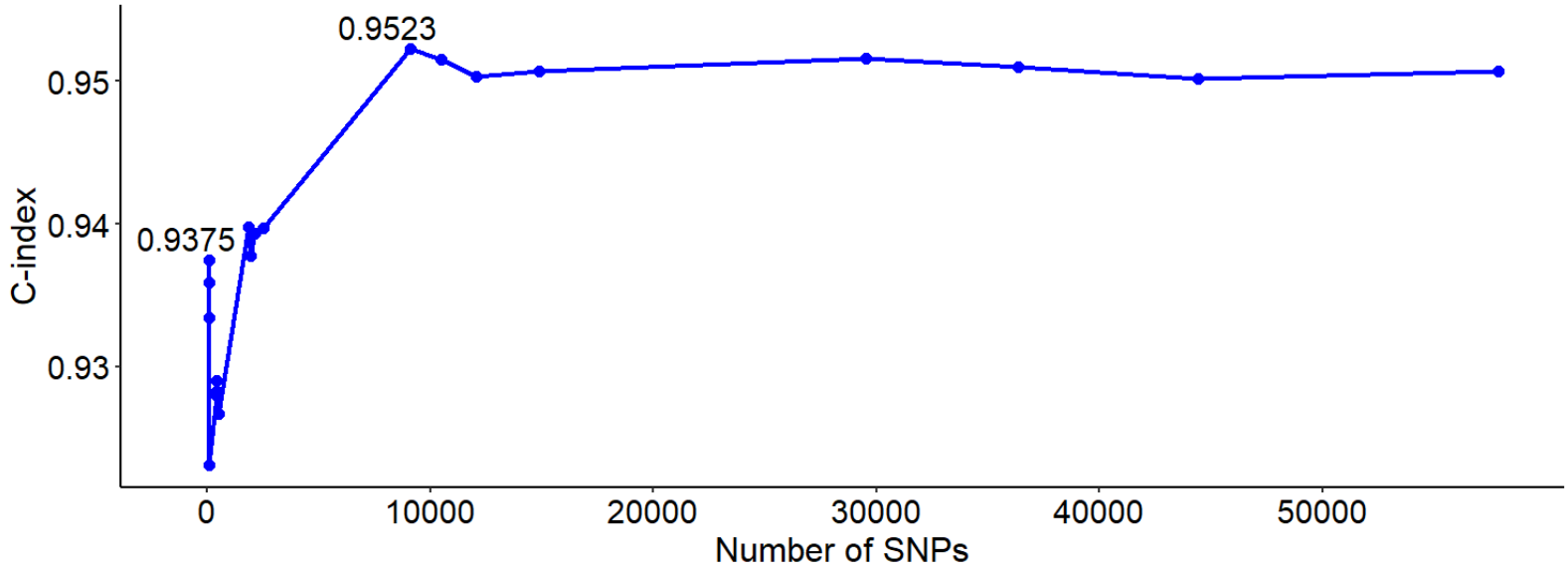


Figure 7. The performance of PRS-sepsis using C + T method in Cox regression model

doi:10.6342/NTU202301905

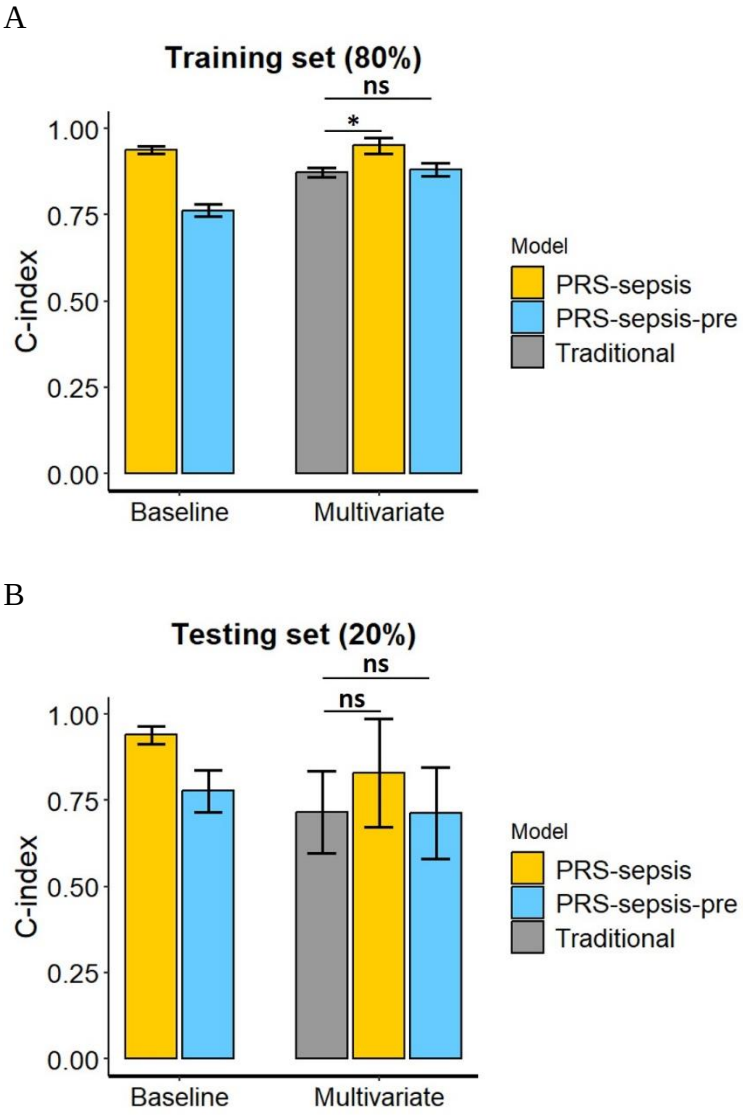
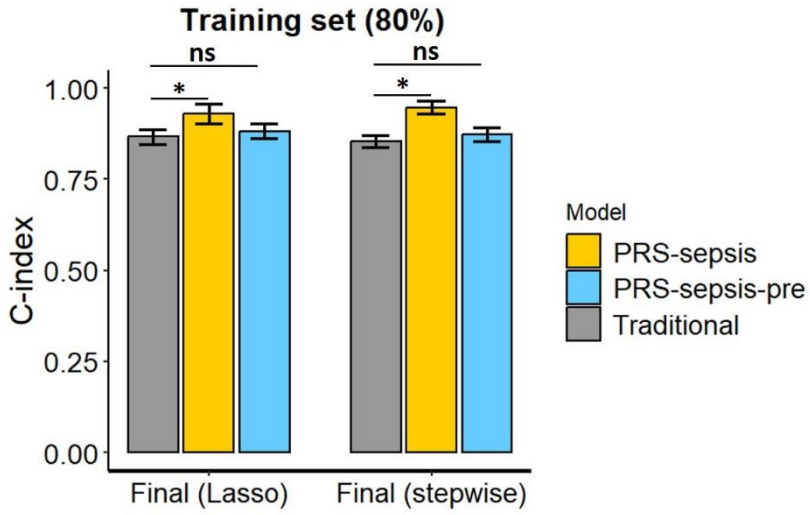


Figure 8. Comparison of the C-index of the baseline models and the multivariate models

doi:10.6342/NTU202301905

A



B

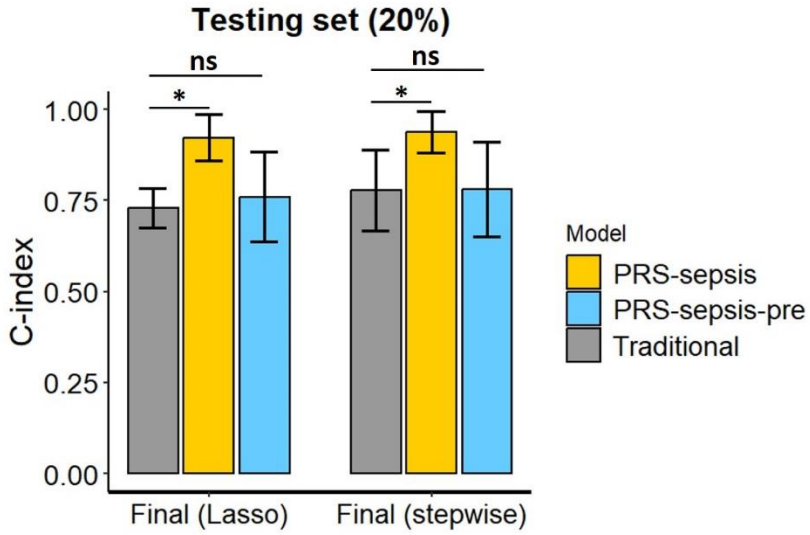


Figure 9. Comparison of the C-index of the final models with Lasso and stepwise selection methods

doi:10.6342/NTU202301905

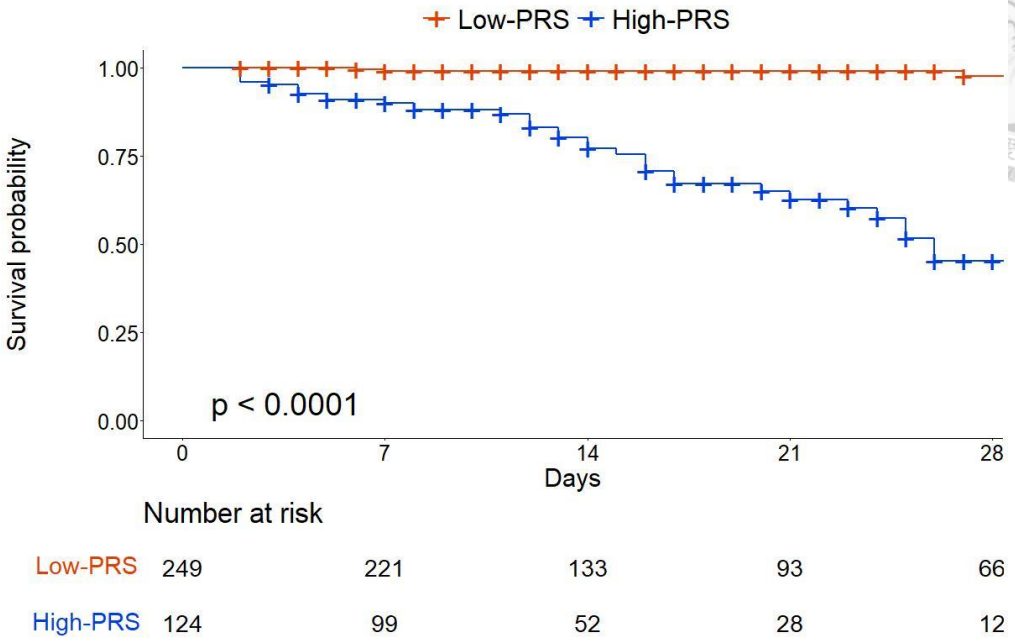


Figure 10. Kaplan–Meier (K-M) plot of PRS-sepsis groups

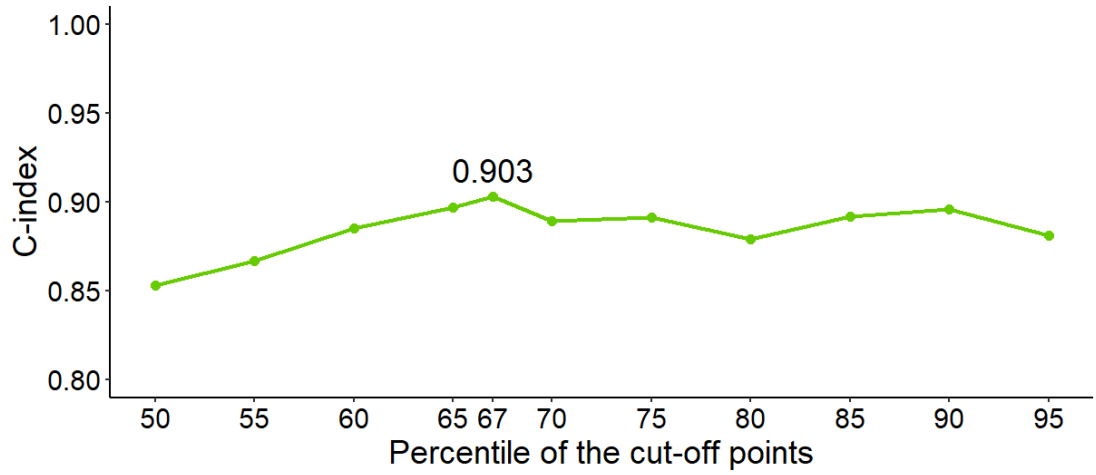


Figure 11. The predictive power of PRS-sepsis using the different cut-off points

doi:10.6342/NTU202301905

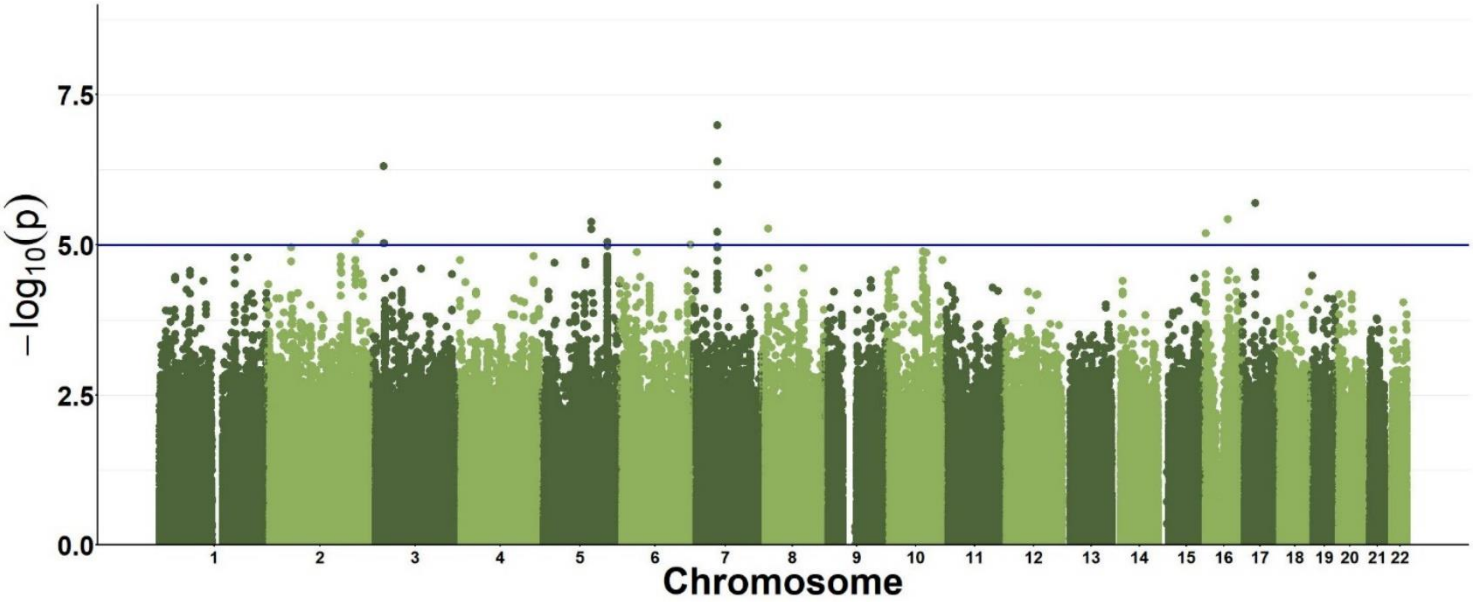


Figure 12. Manhattan plot of GWAS result using logistic regression

doi:10.6342/NTU202301905

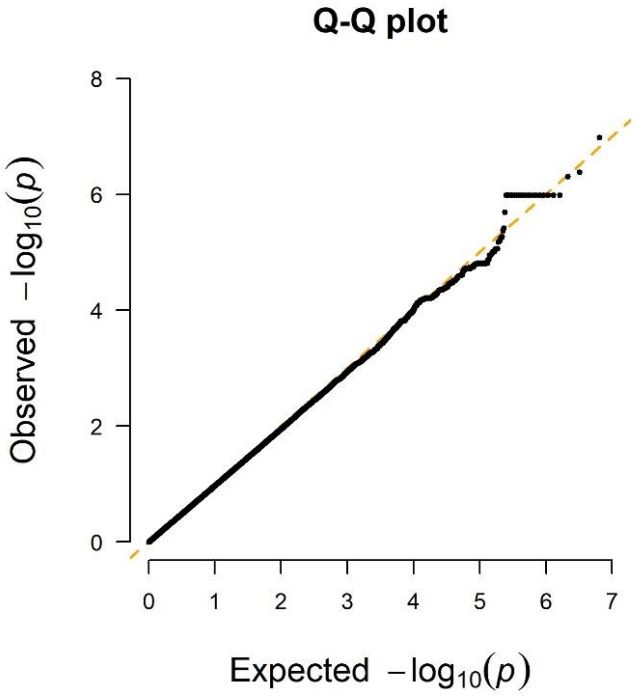


Figure 13. Q-Q plot of GWAS result using logistic regression

doi:10.6342/NTU202301905

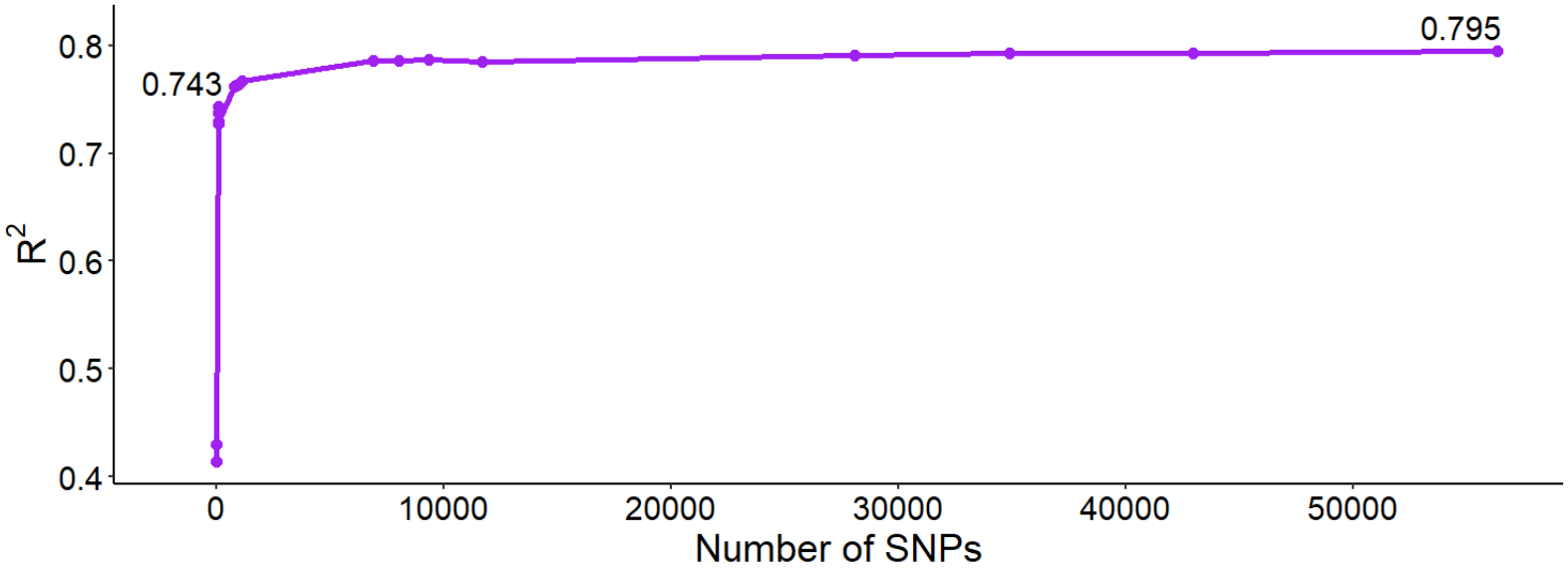


Figure 14. The performance of PRS-sepsis using C + T method in logistic regression model

doi:10.6342/NTU202301905

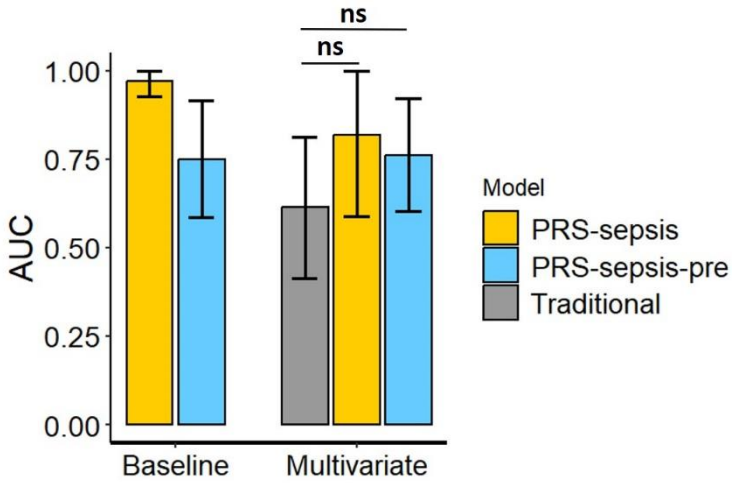


Figure 15. Comparison of the AUC of the baseline models and the multivariate models

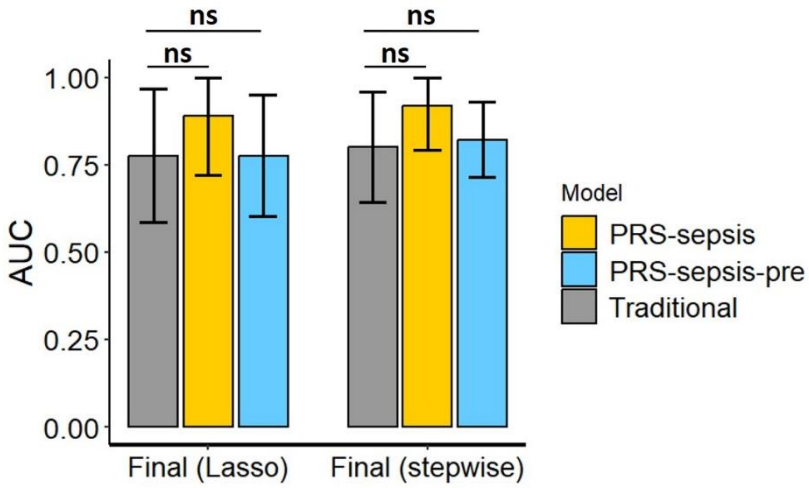


Figure 16. Comparison of the AUC of the final models with Lasso and stepwise selection methods

doi:10.6342/NTU202301905

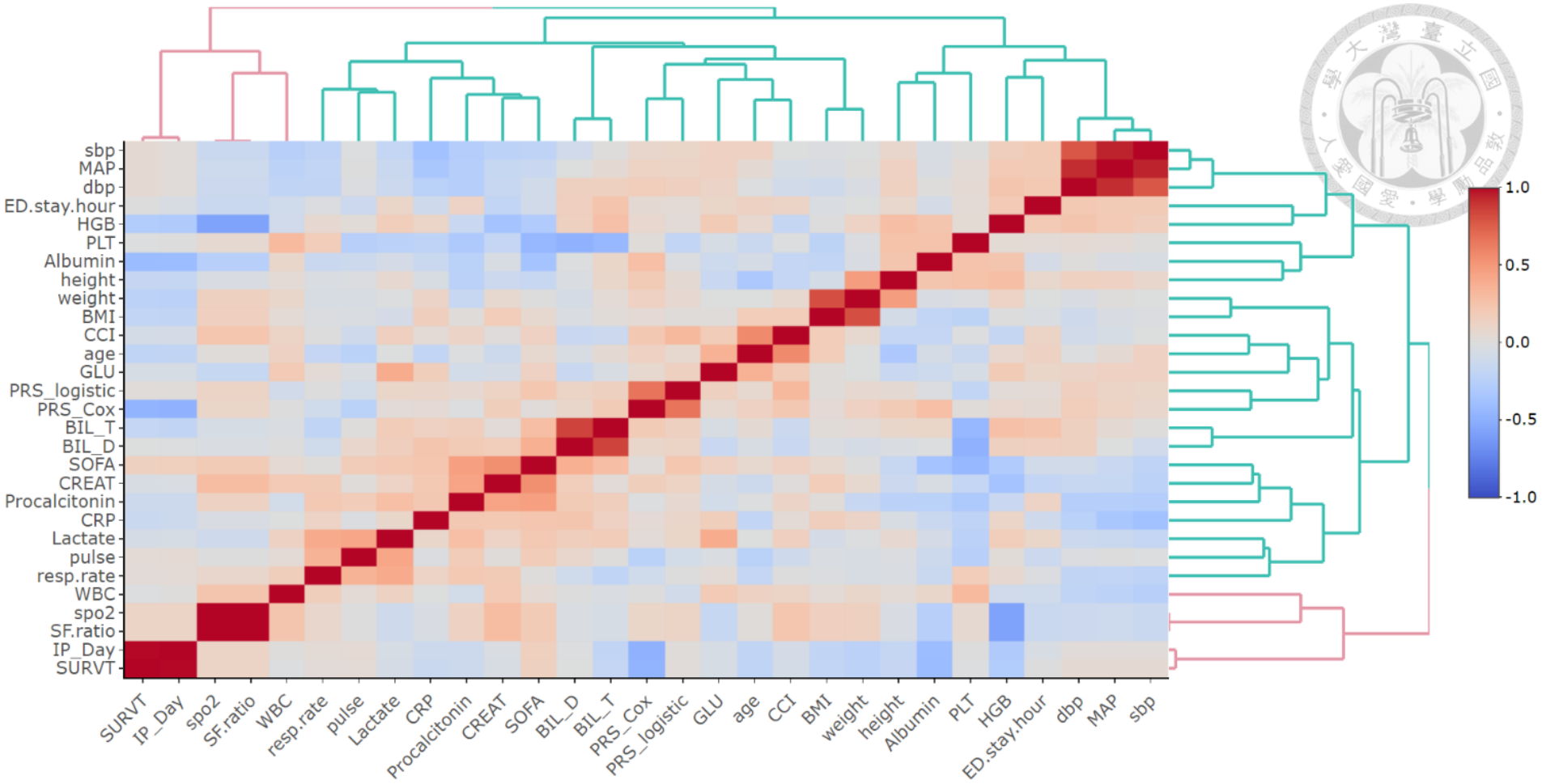


Figure 17. Correlation matrix

doi:10.6342/NTU202301905

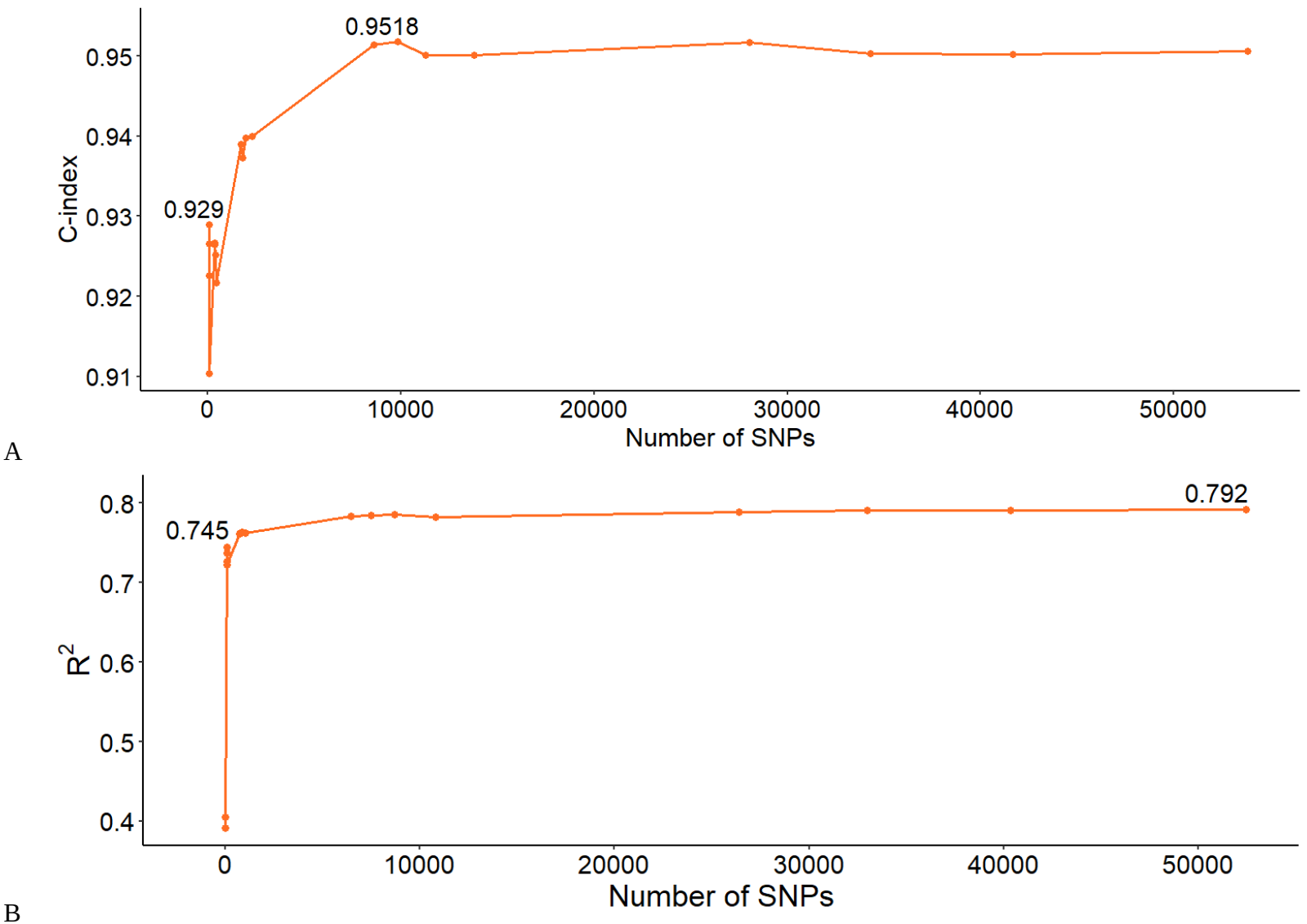


Figure 18. The performances of PRS-sepsis in models were evaluated using PRSice-2 software (C + T method)

doi:10.6342/NTU202301905

Table 1.

Characteristics significantly associated with 28-day mortality

Clinical characteristics	Death (n=39)	28-day survivor (n=334)	p-value
Demographic characteristics			
Sex (Male, n (%))	33 (84.62)	188 (56.29)	0.001
Age (mean (SD))	69.05 (13.52)	64.05 (14.8)	0.06
Height (mean (SD))	165.88 (8.44)	161.58 (8.14)	0.01
BMI (mean (SD))	22.92 (4.81)	25.01 (4.69)	0.02
BMI class (n (%))			0.05
Underweight	5 (17.24)	19 (6.38)	
Normal	15 (51.72)	117 (39.26)	
Overweight	3 (10.34)	66 (22.15)	
Obese	6 (20.69)	96 (32.21)	
No. of days of hospital stay (mean (SD))	11.92 (8.16)	19.96 (19.52)	0.01
Septic shock (n (%))	24 (61.54)	128 (38.32)	0.01
Score system (mean (SD))			
Charlson comorbidity index	6.82 (2.48)	4.84 (3.09)	<0.0001
SOFA score	10.97 (3.28)	8.61 (3.04)	0.0002
Comorbidity (n (%))			
Metastatic solid tumor	14 (35.90)	58 (17.37)	0.01
Laboratory data (mean (SD))			
WBC	10431.03 (5736.35)	12925.77 (7529.7)	0.03
Hemoglobin	10.19 (2.61)	11.75 (2.41)	0.001
Creatinine	3.22 (3.43)	2.22 (2.30)	0.02
Direct bilirubin	1.17 (1.90)	0.55 (1.13)	0.05
Indirect bilirubin	2.23 (3.59)	1.06 (1.76)	0.04
Albumin	2.83 (0.49)	3.08 (0.61)	0.02
Medications (n (%))			
Dopamine	6 (15.38)	21 (6.29)	0.05
Norepinephrine	24 (61.54)	121 (36.23)	0.003
Procedures (n (%))			
Endotracheal intubation	13 (33.33)	49 (14.67)	0.01
Hemodialysis	10 (25.64)	14 (4.19)	<0.0001
Transfusion	30 (76.92)	176 (52.69)	0.01
Main Infection site (n (%))			
Respiratory system	15 (38.46)	71 (21.26)	0.03
Genitourinary system	9 (23.08)	153 (45.81)	0.01

* P-values for continuous variables were computed using either the Mann-Whitney U test or the two-sample t-test. Categorical variables were analyzed using the Chi-square test or Fisher's exact test to calculate the corresponding p-values.

doi:10.6342/NTU202301905

Table 2.

SNPs associated with survival within the 28-day follow-up period ($p < 5\times10^{-8}$) were analyzed using Cox regression

SNP	CHR	Position	Minor Allele	Major Allele	MAF	p-value	HR	Gene	Alternate allele frequency					
									TWB	1KG EAS	1KG EUR	1KG AFR	1KG AMR	1KG SAS
rs296175	5	152615506	**A**	G	0.138	4.48×10^{-8}	4.445	*LINC01470/GRIA1*	0.137	0.143	0.068	0.327	0.078	0.060
rs138138121	10	120074044	**C**	G	0.016	1.09×10^{-8}	26.5	*FAM204A*	0.017	0.039	0.0	0.0	0.0	0.0
rs117040844	13	34334202	**G**	T	0.013	4.12×10^{-8}	19.04	*STARD13/RFC3*	0.023	0.020	0.0	0.0	0.0	0.0
rs374290727	19	52248488	**T**	A	0.011	9.63×10^{-9}	14.461	*FPR1*	0.005	0.012	0.0	0.0	0.0	0.0

SNP: single nucleotide polymorphism, CHR: chromosome number, MAF: minor allele frequency, HR: hazard ratio, TWB: Taiwan biobank, IKG: 1000 genome, EAS: East Asians, EUR: Europeans, AFR: Africans, AMR: Americans, SAS: South Asians, **Bold type**: Alternate allele

doi:10.6342/NTU202301905

Table 3.

The performance of the polygenic risk score in Cox prediction model using clumping and thresholding method (C+T) parameters

			PRS-based prediction model*				
r^2	p-value	number of SNPs	C-index	se (C-index)	HR	95% CI	p-value
0.2	1×10^{-5}	83	0.9359	0.0199	5.1753	(3.7344 – 7.1723)	$< 2\times10^{-16}$
	1×10^{-4}	393	0.9280	0.0249	4.8213	(3.5353 – 6.5750)	
	1×10^{-3}	1825	0.9398	0.0203	4.6321	(3.4115 – 6.2894)	
	0.01	9109	0.9523	0.0174	4.5175	(3.3091 – 6.1670)	
	0.05	29528	0.9516	0.0168	4.2791	(3.1559 – 5.8021)	
0.4	**1×10^{-5}**	**86**	**0.9375**	**0.0193**	**4.9488**	**(3.6248 – 6.7565)**	**$< 2\times10^{-16}$**
	1×10^{-4}	412	0.9290	0.0246	4.7893	(3.5195 – 6.5171)	
	1×10^{-3}	1953	0.9378	0.0203	4.5717	(3.3796 – 6.1845)	
	0.01	10474	0.9515	0.0174	4.4811	(3.2873 – 6.1084)	
	0.05	36352	0.9510	0.0171	4.2372	(3.1316 – 5.7330)	
0.6	1×10^{-5}	92	0.9334	0.0203	4.7622	(3.5215 – 6.4399)	$< 2\times10^{-16}$
	1×10^{-4}	436	0.9283	0.0247	4.7614	(3.5042 – 6.4697)	
	1×10^{-3}	2151	0.9393	0.0206	4.5571	(3.3743 – 6.1545)	
	0.01	12085	0.9503	0.0180	4.4404	(3.2625 – 6.0436)	
	0.05	44437	0.9502	0.0177	4.2056	(3.1124 – 5.6828)	
0.8	1×10^{-5}	102	0.9231	0.0227	4.0579	(3.1219 – 5.2744)	$< 2\times10^{-16}$
	1×10^{-4}	497	0.9267	0.0245	4.6395	(3.4452 – 6.2478)	
	1×10^{-3}	2519	0.9397	0.0207	4.5451	(3.3696 – 6.1307)	
	0.01	14887	0.9507	0.0177	4.4129	(3.2465 – 5.9983)	
	0.05	57896	0.9507	0.0174	4.2015	(3.1083 – 5.6791)	

doi:10.6342/NTU202301905

Table 4.

Univariate Cox regression of clinical variables on 28-day mortality.

Clinical characteristics	HR (95% CI)	p-value
Demographic characteristics		
Sex (Female)	0.23 (0.097 – 0.551)	0.001
Age	1.02 (1.001 – 1.048)	0.0383
Height	1.07 (1.020 – 1.115)	0.005
BMI	0.92 (0.844 – 0.996)	0.0401
Charlson comorbidity index	1.17 (1.062 – 1.293)	0.0015
SOFA score	1.15 (1.033 – 1.278)	0.0106
Comorbidity		
Hemiplegia Paraplegia	2.64 (1.025 – 6.786)	0.0443
Metastatic solid tumor	2.07 (1.077 – 3.993)	0.0292
Laboratory data		
Hemoglobin	0.81 (0.714 – 0.915)	0.0008
Indirect bilirubin	1.11 (1.031 – 1.196)	0.0057
Procedures		
Hemodialysis	3.86 (1.871 – 7.963)	0.0003
Main Infection site		
Other / undetermined origin	0.42 (0.175 – 0.997)	0.0492

BMI: body mass index, SOFA score: sequential organ failure assessment score

doi:10.6342/NTU202301905

Table 5.

Five-fold cross validation results of the baseline and multivariate prediction models using Cox regression

		Training		Testing	
Model	Model parameters	C-index	se (C-index)	C-index	se (C-index)
PRS-sepsis					
Baseline Model	PRS-sepsis + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.9389	0.0210	0.9402	0.1169
Multivariate Model	PRS-sepsis + Clinical variables† + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.9513	0.0208	0.8304	0.1394
PRS-sepsis-pre					
Baseline Model	PRS-sepsis-pre + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.7631	0.0494	0.7772	0.1155
Multivariate Model	PRS-sepsis-pre + Clinical variables† + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.8813	0.0372	0.7134	0.1377
Traditional					
Multivariate Model	Clinical variables† + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.8711	0.0396	0.7088	0.1390

† The clinical variables exhibited significant associations with survival time in the univariate Cox regression analysis

doi:10.6342/NTU202301905

Table 6.

Five-fold cross validation results of the final prediction models using Cox regression

		Training		Testing	
Model	Model parameters	C-index	se (C-index)	C-index	se (C-index)
PRS-sepsis					
Final Model (Lasso)†	PRS-sepsis + CCI + Hemodialysis + Indirect bilirubin + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.9254	0.0320	0.8811	0.1275
Final Model (stepwise)	PRS-sepsis + Metastatic solid tumor +HGB + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.9467	0.0234	0.9298	0.1428
PRS-sepsis-pre					
Final Model (Lasso)	PRS-sepsis-pre + Height + BMI + CCI + SOFA + HGB + Indirect bilirubin + Metastatic solid tumor + Hemodialysis + Other site infection + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.8826	0.0367	0.7597	0.1393
Final Model (stepwise)	PRS-sepsis-pre + HGB + Hemodialysis + Indirect bilirubin + BMI + Other site infection + Height + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.8729	0.0393	0.7808	0.1393
Traditional					
Final Model (Lasso)	Height + BMI + CCI + SOFA + HGB + Indirect bilirubin + Metastatic solid tumor + Hemodialysis + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.8669	0.0389	0.7293	0.1400

doi:10.6342/NTU202301905

Final Model (stepwise)	HGB + Hemodialysis + Indirect bilirubin + BMI + Other site infection + Age + Sex + $\sum_{k=1}^{3} PC_k$	0.8537	0.0479	0.7775	0.1403

BMI: body mass index, CCI: Charlson comorbidity index, SOFA: sequential organ failure assessment score, HGB: Hemoglobin

† Lasso and stepwise were the methods for selecting the clinical variables

doi:10.6342/NTU202301905

Table 7.
The hazard ratio (HR) and 95% CI of PRS in Cox prediction models

	PRS	
Model	HR (95% CI)	p-value
PRS-sepsis		
Baseline Model	4.95 (3.62 – 6.76)	**7.77×10^{-24}***
Multivariate Model	4.07 (2.63 – 6.30)	**2.75×10^{-10}***
Final Model (Lasso)	4.04 (2.78 – 5.86)	**2.59×10^{-13}***
Final Model (stepwise)	3.96 (2.71 – 5.78)	**9.23×10^{-13}***
PRS-sepsis-pre		
Baseline Model	28.82 (2.28 – 364.25)	**0.0094***
Multivariate Model	9.00 (0.30 – 268.93)	0.2050
Final Model (Lasso)	9.11 (0.32 – 259.54)	0.1960
Final Model (stepwise)	6.93 (0.34 – 142.25)	0.2091

* p-value < 0.05

doi:10.6342/NTU202301905

Table 8.

The performance of PRS-sepsis groups using different cut-off points

Percentile (cut-off point)	Low-PRS/High-PRS†	C-index of model	se (C-index)	The mortality rate		
				Low-PRS group	High-PRS group	p-value
95	354/19	0.881	0.029	0.06	1	$< 2\times10^{-16}$
90	335/38	0.896	0.027	0.03	0.76	$< 2\times10^{-16}$
85	317/56	0.892	0.029	0.03	0.55	$< 2\times10^{-16}$
80	298/75	0.879	0.031	0.02	0.43	$< 2\times10^{-16}$
75	279/94	0.891	0.027	0.02	0.36	$< 2\times10^{-16}$
70	261/112	0.889	0.029	0.02	0.31	$< 2\times10^{-16}$
67	**249/124**	**0.903**	**0.020**	**0.01**	**0.29**	**$< 2\times10^{-16}$**
65	242/131	0.897	0.020	0.01	0.27	$< 2\times10^{-16}$
60	223/150	0.885	0.023	0.01	0.24	1×10^{-15}
55	205/168	0.867	0.025	0.01	0.21	3×10^{-13}
50	186/187	0.853	0.026	0.02	0.19	4×10^{-11}

† The number of the sepsis patients respectively in the low-PRS and high-PRS groups

doi:10.6342/NTU202301905

Table 9.

Spearman's rank-order correlation between continuous variables and PRS variables constructed by the Cox regression

	PRS-sepsis		PRS-sepsis-pre	
Variable	r	p-value	r	p-value
Age	-0.057	0.271	-0.010	0.845
Height	-0.019	0.729	0.071	0.199
Weight	-0.065	0.213	0.042	0.428
BMI	-0.027	0.623	0.041	0.458
Respiration rates (> 20 times/min or ≤ 20 times/min)	-0.03	0.565	0.008	0.874
SBP	0.046	0.371	0.003	0.953
DBP	0.032	0.536	0.030	0.565
Pulse	-0.151	**0.003***	-0.016	0.76
Mean arterial pressure	0.046	0.377	0.017	0.748
No. of hours of ED stay	-0.04	0.445	-0.007	0.89
No. of days of hospital stay	-0.31	**<0.0001***	-0.052	0.315
SF ratio	-0.01	0.842	-0.022	0.679
Charlson comorbidity index	0.077	0.136	0.004	0.94
WBC	-0.072	0.169	-0.096	0.066
Hemoglobin	-0.183	**<0.0001***	-0.081	0.117
Platelet	-0.067	0.196	-0.126	**0.015***
Creatinine	-0.036	0.488	0.087	0.094
Direct bilirubin	0.034	0.628	0.026	0.706
Indirect bilirubin	-0.002	0.964	0.074	0.176
Glucose	-0.031	0.592	-0.037	0.529
Lactate	0.008	0.884	-0.072	0.183
Oxygen saturation	-0.01	0.842	-0.022	0.679
Albumin	0.075	0.227	-0.097	0.118
CRP	-0.098	0.140	0.002	0.972
Procalcitonin	-0.01	0.899	0.055	0.472
SOFA score	-0.041	0.472	0.135	**0.017***

BMI: body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, ED: emergency department, SF: SpO_2/FiO_2, WBC: white blood cells, CRP: C-reactive protein, SOFA score: sequential organ failure assessment score

* $p < 0.05$

doi:10.6342/NTU202301905

Table 10.

SNPs associated ($p < 1\times10^{-5}$) with 28-day mortality were analyzed using logistic regression

SNP	CHR	Position	Minor Allele	Major Allele	MAF	p-value	OR	Gene	Alternate AF					
									TWB	1KG EAS	1KG EUR	1KG AFR	1KG AMR	1KG SAS
rs71425909	2	202846390	**C**	T	0.071	8.73×10^{-6}	5.386	.	0.099	0.086	0.035	0.153	0.048	0.146
rs141849589	2	214775203	**A**	G	0.024	6.63×10^{-6}	13.410	*SPAG16*	0.035	0.021	0.030	0.001	0.022	0.037
rs192451106	3	25992974	**G**	C	0.024	4.94×10^{-7}	16.980	.	0.018	0.016	0.0	0.0	0.0	0.0
rs148193677	5	116386431	**C**	A	0.031	4.21×10^{-6}	9.292	.	0.057	0.051	0.0	0.0	0.001	0.001
rs296175	5	152615506	**A**	G	0.138	8.92×10^{-6}	3.925	.	0.137	0.143	0.068	0.327	0.078	0.06
rs1220760	7	54675770	**G**	A	0.035	1.03×10^{-7}	12.500	.	0.033	0.029	0.292	0.028	0.182	0.144
rs55737846	7	54682213	**G**	A	0.157	6.13×10^{-6}	3.940	.	0.159	0.156	0.203	0.071	0.159	0.103
rs10091661	8	13715245	**G**	A	0.046	5.37×10^{-6}	7.577	.	0.077	0.093	0.149	0.18	0.212	0.082
rs12932511	16	57939298	**T**	C	0.162	3.77×10^{-6}	3.948	*RP11-420N3.2*	0.471	0.476	0.030	0.025	0.199	0.026
rs78944872	16	6266045	**T**	C	0.481	6.45×10^{-6}	3.765	*CNGB1*	0.168	0.173	0.379	0.236	0.378	0.526
rs78432526	17	30976365	**G**	A	0.020	2.02×10^{-6}	17.210	*RP11-220C2.1*	0.027	0.016	0.001	0.0	0.056	0.0

SNP: single nucleotide polymorphism, CHR: chromosome number, MAF: minor allele frequency, HR: hazard ratio, TWB: Taiwan biobank, IKG: 1000 genome, EAS: East Asians, EUR: Europeans, AFR: Africans, AMR: Americans, SAS: South Asians, **Bold type**: Alternate allele

doi:10.6342/NTU202301905

Table 11.

The performance of the PRS in logistic prediction model using clumping and thresholding method (C+T) parameters

			PRS-based prediction model		
r^2	p-value	number of SNPs	R^2	OR (95% CI)	p-value
0.2	1×10^{-5}	10	0.430	5.909 (3.544 – 9.854)	9.77×10^{-12}
	1×10^{-4}	**94**	**0.743**	**40.245 (9.893 – 163.720)**	**2.45×10^{-7}**
	1×10^{-3}	818	0.762	52.175 (5.109– 532.886)	0.0009
	0.01	6904	0.786	6971.254 (3.381 – 1.44×10^{7})	0.0230
	0.05	28080	0.791	40549.886 (1.158 – 1.42×10^{9})	0.0469
0.4	1×10^{-5}	10	0.43	5.909 (3.544 – 9.854)	9.77×10^{-12}
	1×10^{-4}	98	0.727	27.196 (8.545 – 86.561**)**	2.45×10^{-7}
	1×10^{-3}	880	0.763	53.725 (5.248 – 550.032)	0.0008
	0.01	8014	0.786	8165.588 (2.830 – 2.36×10^{7})	0.0267
	0.05	34910	0.793	102767.834 (0.896 – 1.18×10^{10})	0.0522
0.6	1×10^{-5}	10	0.43	5.909 (3.544 – 9.854)	9.77×10^{-12}
	1×10^{-4}	103	0.737	32.064 (9.090 – 113.099**)**	9.67×10^{-8}
	1×10^{-3}	963	0.764	55.136 (5.064 – 600.349)	0.0010
	0.01	9348	0.787	9569.825 (2.772 – 3.30×10^{7})	0.0274
	0.05	42972	0.793	89228.608 (1.110 – 7.17×10^{9})	0.0480
0.8	1×10^{-5}	11	0.414	5.709 (3.487 – 9.347)	4.35×10^{-12}
	1×10^{-4}	117	0.729	26.895 (8.474 – 85.359**)**	2.31×10^{-8}
	1×10^{-3}	1148	0.767	58.890 (4.927 – 703.913)	0.0013
	0.01	11702	0.785	2646.937 (3.225 – 2.17×10^{6})	0.0213
	0.05	56359	0.795	114045.394 (1.356 – 9.59×10^{9})	0.0442

doi:10.6342/NTU202301905

Table 12.

Univariate logistic regression of clinical variables on 28-day mortality.

Clinical characteristics	OR (95% CI)	p-value
Demographic characteristics		
Sex (Female)	0.23 (0.086 – 0.536)	0.0015
Age	1.03 (1.001 – 1.051)	0.0463
Height	1.07 (1.018 – 1.123)	0.0081
Weight	0.97 (0.944 – 0.998)	0.045
BMI	0.9 (0.820 – 0.983)	0.0236
BMI class		
Underweight	Reference	-
Normal	0.49 (0.166 – 1.634)	0.2091
Overweight	0.17 (0.033 – 0.767)	0.0235
Obese	0.24 (0.065 – 0.897)	0.0283
Score system		
Charlson comorbidity index	1.23 (1.102 – 1.372)	0.0002
SOFA score	1.25 (1.120 – 1.407)	<0.0001
Septic shock	2.58 (1.315 – 5.193)	0.0066
Comorbidity		
Metastatic solid tumor	2.66 (1.279 – 5.377)	0.007
Medications		
Dopamine	2.71 (0.941 – 6.843)	0.0452
Norepinephrine	2.82 (1.437 – 5.685)	0.003
Laboratory data		
WBC	0.99994 (0.99988 – 0.99999)	0.0464
Hemoglobin	0.77 (0.671 – 0.886)	0.0003
Creatinine	1.13 (1.012 – 1.253)	0.0208
Indirect bilirubin	1.17 (1.044 – 1.338)	0.0085
Procedures		
Endotracheal intubation	2.91 (1.366 – 5.963)	0.0042
Brain CT	2.09 (1.048 – 4.418)	0.0422
Hemodialysis	7.88 (3.15 – 19.255)	<0.0001
Transfusion	2.99 (1.431 – 6.869)	0.0056
Main Infection site		
Respiratory system	2.32 (1.134 – 4.607)	0.0181
Genitourinary system	0.35 (0.155 – 0.742)	0.0088

BMI: body mass index, SOFA score: sequential organ failure assessment score, WBC: white blood cells, CT: computed tomography

doi:10.6342/NTU202301905

Table 13.

Five-fold cross validation results of the baseline and multivariate prediction models using logistic regression

Model	Model parameters	AUC (95% CI)
PRS-sepsis		
Baseline Model	PRS-sepsis + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.971 (0.919 – 1.000)
Multivariate Model	PRS-sepsis + Clinical variables† + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.818 (0.622 – 0.978)
PRS-sepsis-pre		
Baseline Model	PRS-sepsis-pre + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.750 (0.590 – 0.911)
Multivariate Model	PRS-sepsis-pre + Clinical variables† + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.761 (0.547 – 0.946)
Traditional		
Multivariate Model	Clinical variables† + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.613 (0.256 – 0.930)

† The clinical variables were significantly associated with sepsis mortality in the univariate logistic regression analysis

doi:10.6342/NTU202301905

Table 14.

Five-fold cross validation results of the final prediction models using logistic regression

Model	Model parameters	AUC (95% CI)
PRS-sepsis		
Final Model (Lasso)†	PRS-sepsis + CCI + Metastatic solid tumor +HGB + Hemodialysis + Transfusion + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.891 (0.745 – 0.998)
Final Model (stepwise)	PRS-sepsis + HGB + Hemodialysis + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.918 (0.785 – 1.000)
PRS-sepsis-pre		
Final Model (Lasso)	PRS-sepsis-pre + Height + BMI + CCI + SOFA + Metastatic solid tumor + HGB + Indirect bilirubin + Hemodialysis + WBC + Creatinine + Endotracheal intubation + CT + Genitourinary system infection + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.776 (0.573 – 0.943)
Final Model (stepwise)	PRS-sepsis-pre + HGB + Hemodialysis + Genitourinary system infection + BIL_T + Metastatic solid tumor + height + WBC + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.822 (0.611 – 0.995)
Traditional		
Final Model (Lasso)	Height + BMI + CCI + SOFA + Metastatic solid tumor + HGB + Indirect bilirubin + Hemodialysis + WBC + Creatinine + Endotracheal intubation + CT + Genitourinary system infection + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.776 (0.562 – 0.959)

doi:10.6342/NTU202301905

Final Model (stepwise)	HGB + Hemodialysis + Endotracheal intubation + Indirect bilirubin + Genitourinary system infection + Height + WBC + Age + Sex + $\sum_{k=1}^{5} PC_k$	0.801 (0.591 – 0.983)

BMI: body mass index, CCI: Charlson comorbidity index, SOFA: sequential organ failure assessment score, HGB: Hemoglobin, WBC: white blood cells, CT: computed tomography

† Lasso and stepwise were the methods for selecting the clinical variables

doi:10.6342/NTU202301905

Table 15.

The odds ratio (OR) and 95% CI of PRS in logistic prediction models

	PRS	
Model	OR (95% CI)	p-value
With PRS-sepsis		
Baseline Model	40.25 (9.89 – 163.72)	**2.45×10^{-7}***
Multivariate Model	236.65 (2.96 – 18913.49)	**0.0145***
Final Model (Lasso)	53.98 (8.97 – 324.61)	**1.32×10^{-5}***
Final Model (stepwise)	49.78 (8.88 – 279.14)	**8.92×10^{-6}***
With PRS-sepsis-pre		
Baseline Model	2.41 (1.63 – 3.58)	**1.22×10^{-5}***
Multivariate Model	3.68 (1.55 – 8.78)	**0.0033***
Final Model (Lasso)	3.63 (1.57 – 8.38)	**0.0026***
Final Model (stepwise)	2.77 (1.49 – 5.18)	**0.0014***

* p-value < 0.05

doi:10.6342/NTU202301905

Table 16.

Spearman's rank-order correlation between continuous variables and PRS variables constructed by the logistic regression

	PRS-sepsis		PRS-sepsis-pre	
Variable	r	p-value	r	p-value
Age	0.016	0.751	0.003	0.954
Height	-0.054	0.332	0.071	0.202
Weight	-0.12	**0.022***	0.008	0.879
BMI	-0.085	0.125	-0.003	0.952
Respiration rates (> 20 times/min or ≤ 20 times/min)	0.073	0.159	0.012	0.811
SBP	0.018	0.736	0.009	0.861
DBP	0.043	0.404	0.009	0.869
Pulse	-0.029	0.579	-0.027	0.599
Mean arterial pressure	0.044	0.402	0.01	0.844
No. of hours stay in ED	-0.055	0.29	0.005	0.926
No. of days stay in hospital	0.007	0.893	0.029	0.58
SF ratio	0.042	0.42	-0.002	0.976
Charlson comorbidity index	0.114	**0.028***	0.052	0.317
WBC	-0.117	**0.024***	-0.082	0.115
HGB	-0.168	**0.001***	-0.089	0.088
Platelet	-0.107	**0.038***	-0.066	0.202
Creatinine	0.016	0.754	0.073	0.164
Direct bilirubin	0.055	0.426	0.02	0.773
Indirect bilirubin	0.006	0.911	0.052	0.337
Glucose	0.151	**0.009***	0.026	0.659
Lactate	0.097	0.074	-0.021	0.703
Oxygen saturation	0.042	0.42	-0.002	0.976
Albumin	-0.132	**0.033***	-0.134	**0.030***
CRP	0.015	0.82	0.129	0.053
Procalcitonin	0.096	0.211	0.137	0.073
SOFA score	0.141	**0.013***	0.192	**0.001***

BMI: body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, ED: emergency department, SF: SpO_2/FiO_2, WBC: white blood cells, CRP: C-reactive protein, SOFA score: sequential organ failure assessment score

* $p < 0.05$

doi:10.6342/NTU202301905

Table 17.

The performance of PRS-sepsis in Cox regression model via PRSice-2 software

			PRS-based prediction model*				
r^2	p-value	number of SNPs	C-index	se (C-index)	HR	95% CI	p-value
0.2	**1×10^{-5}**	**78**	**0.9290**	**0.0217**	**4.8936**	**3.5688 – 6.7103**	**$< 2\times10^{-6}$**
	1.1×10^{-5}	83	0.9262	0.0231	4.8770	3.5612 – 6.6789	
	1.2×10^{-5}	92	0.9265	0.0227	4.7335	3.4911 – 6.4179	
	1.3×10^{-5}	99	0.9258	0.0228	4.5667	3.4062 – 6.1224	
0.4	1×10^{-5}	80	0.9266	0.0223	4.7404	3.4940 – 6.4314	$< 2\times10^{-6}$
	1.1×10^{-5}	85	0.9247	0.0234	4.7555	3.5000 – 6.4614	
	1.2×10^{-5}	94	0.9251	0.0230	4.6370	3.4420 – 6.2469	
0.6	1×10^{-5}	84	0.9226	0.0239	4.6658	3.4602 – 6.2912	$< 2\times10^{-6}$
	1.1×10^{-5}	90	0.9226	0.0241	4.6629	3.4559– 6.2913	
	1.2×10^{-5}	99	0.9231	0.0238	4.5845	3.4173 – 6.1505	
0.8	1×10^{-5}	93	0.9104	0.0270	3.9856	3.0726 – 5.1699	$< 2\times10^{-6}$
	1.1×10^{-5}	99	0.9127	0.0266	3.9909	3.0725 – 5.1838	

doi:10.6342/NTU202301905

Table 18.

The performance of PRS-sepsis in logistic regression model via PRSice-2 software

			PRS-based prediction model		
r^2	p-value	number of SNPs	R^2	OR (95% CI)	p-value
0.2	1×10^{-5}	9	0.405	5.303 (3.2965 – 8.5308)	6.09×10^{-12}
	2×10^{-5}	19	0.545	10.772 (5.3517 –21.6823)	2.74×10^{-11}
	3×10^{-5}	22	0.568	10.612 (5.3231 – 21.1563)	1.95×10^{-11}
	4×10^{-5}	30	0.603	13.190 (6.6825 – 31.7905)	6.48×10^{-11}
	5×10^{-5}	40	0.659	23.056 (7.8209 – 67.9684)	1.28×10^{-8}
	6×10^{-5}	48	0.697	34.439 (9.646 – 122.9573)	5.02×10^{-8}
	7×10^{-5}	58	0.705	33.502 (9.5022 – 118.1155)	4.71×10^{-8}
	8×10^{-5}	66	0.719	38.237 (9.9073 – 147.5745)	1.24×10^{-7}
	9×10^{-5}	74	0.730	38.441 (10.2561 – 144.0779)	6.19×10^{-8}
	1×10^{-4}	82	0.745	50.662 (11.1637 – 229.9073)	3.65×10^{-7}
	1.1×10^{-4}	**88**	**0.760**	**59.981 (11.3294 – 317.5551)**	**1.47×10^{-6}**
	1.2×10^{-4}	99	0.752	43.118 (10.55281 – 176.1774)	1.60×10^{-7}
0.4	1×10^{-5}	9	0.405	5.303 (3.2965 – 8.5308)	6.09×10^{-12}
	2×10^{-5}	20	0.528	9.700 (5.0383 – 18.6753)	1.06×10^{-11}
	3×10^{-5}	23	0.554	9.823 (5.1208 – 18.8427)	6.23×10^{-12}
	4×10^{-5}	32	0.595	12.235 (5.8421 – 25.6228)	3.14×10^{-11}
	5×10^{-5}	42	0.652	19.240 (7.3884 – 50.1041)	1.40×10^{-9}
	6×10^{-5}	52	0.679	21.923 (7.8939 – 60.8851)	3.13×10^{-9}
	7×10^{-5}	62	0.687	22.042 (7.973 –60.9344)	2.50×10^{-9}
	8×10^{-5}	70	0.701	23.667 (8.1236 – 68.9489)	6.65×10^{-9}
	9×10^{-5}	78	0.712	25.223 (8.522 – 74.6545)	5.54×10^{-9}

doi:10.6342/NTU202301905

r^2	p-value	number of SNPs	PRS-based prediction model		
			R^2	OR (95% CI)	p-value
0.4	1×10^{-4}	86	0.727	31.110 (9.2404 – 104.7417)	2.86×10^{-8}
	1.1×10^{-4}	93	0.743	35.078 (9.6127 – 128.0049)	7.19×10^{-8}
0.6	1×10^{-5}	9	0.405	5.303 (3.2965 – 8.5308)	6.09×10^{-12}
	2×10^{-5}	21	0.527	10.015 (5.1544 – 19.4598)	1.06×10^{-11}
	3×10^{-5}	26	0.559	10.482 (5.3948 – 20.3657)	4.11×10^{-12}
	4×10^{-5}	35	0.601	12.336 (5.9598 – 25.5344)	1.30×10^{-11}
	5×10^{-5}	45	0.655	18.680 (7.3564 – 47.4332)	7.41×10^{-10}
	6×10^{-5}	56	0.697	28.685 (8.9821 – 91.6045)	1.47×10^{-8}
	7×10^{-5}	66	0.705	29.223 (9.0924 – 93.9251)	1.46×10^{-8}
	8×10^{-5}	74	0.715	29.525 (9.0216 – 96.6274)	2.19×10^{-8}
	9×10^{-5}	82	0.724	29.838 (9.2528 – 96.2200)	1.31×10^{-8}
	1×10^{-4}	90	0.737	35.966 (9.8129 – 131.8245)	6.45×10^{-8}
	1.1×10^{-4}	97	0.752	40.703 (10.2113 – 162.248)	1.49×10^{-7}
0.8	1×10^{-5}	10	0.391	5.224 (3.2716 – 8.3408)	4.37×10^{-12}
	2×10^{-5}	24	0.510	9.239 (4.9069 – 17.3943)	5.70×10^{-12}
	3×10^{-5}	30	0.528	9.156 (4.9724 – 16.8592)	1.17×10^{-12}
	4×10^{-5}	40	0.573	11.465 (5.7526 – 22.8511)	4.14×10^{-12}
	5×10^{-5}	50	0.628	14.792 (6.6938 – 32.6863)	2.75×10^{-11}
	6×10^{-5}	62	0.672	20.275 (7.8513 – 52.3577)	5.06×10^{-10}
	7×10^{-5}	72	0.684	21.556 (7.9494 – 58.4522)	1.61×10^{-9}
	8×10^{-5}	81	0.697	23.22 (8.1672 – 66.0180)	3.65×10^{-9}
	9×10^{-5}	91	0.708	24.862 (8.5876 – 71.9789)	3.13×10^{-9}

doi:10.6342/NTU202301905

Table 19.

The significant SNPs ($p < 1\times10^{-5}$) reported from the previous studies

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p-value	Study
1	82383883	rs146730869	na	na	na	*LPHN2*	4.48	5.30×10^{-6}	Rautanen et al, 2015[16]
2	134086271	rs10928450	na	na	na	*NCKAP5*	0.63	9.00×10^{-6}	
2	134091649	rs13392963	na	na	na	*NCKAP5*	0.63	1.60×10^{-5}	
2	201260365	rs78318430	na	na	na	*SPATS2L*	2.96	2.00×10^{-5}	
2	201308486	rs893357	na	na	na	*SPATS2L*	2.70	3.40×10^{-6}	
4	91619360	rs72661871	na	na	na	*CCSER1*	2.87	7.80×10^{-6}	
4	91671916	rs72661895	na	na	na	*CCSER1*	2.92	6.40×10^{-6}	
5	108402140	rs4957796	na	na	na	*FER*	0.52	9.70×10^{-8}	
5	108406299	rs975056	na	na	na	*FER*	0.56	3.30×10^{-7}	
5	108417332	rs62375529	na	na	na	*FER*	0.56	7.50×10^{-7}	
5	134514728	rs553438	na	na	na	Intergenic	0.63	2.20×10^{-5}	
5	134532593	rs639405	na	na	na	Intergenic	0.62	8.00×10^{-6}	
6	103810003	rs79423885	na	na	na	Intergenic	2.05	8.10×10^{-6}	
6	103813490	rs77054842	na	na	na	Intergenic	2.06	7.30×10^{-6}	
7	83635586	rs4732529	na	na	na	*SEMA3A*	1.78	1.60×10^{-6}	
7	83670129	rs76881522	na	na	na	*SEMA3A*	1.77	6.20×10^{-6}	
7	123248738	rs35947027	na	na	na	*ASB15*	1.54	7.60×10^{-6}	
8	19554995	rs12114790	na	na	na	*CSGALNACT1*	1.43	6.40×10^{-5}	
8	19574243	rs10095344	na	na	na	*CSGALNACT1*	1.50	1.00×10^{-5}	

doi:10.6342/NTU202301905

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p	Study
13	38969457	rs17057959	na	na	na	Intergenic	1.78	1.50×10^{-6}	Rautanen et al, 2015[16]
13	39102184	rs9566343	na	na	na	Intergenic	1.71	1.40×10^{-6}	
2	49177601	rs7591064	na	na	na	Intergenic	0.67	9.50×10^{-6}	
2	133396179	rs17324515	na	na	na	*LYPD1*	0.74	3.50×10^{-5}	
2	133426183	rs2709532	na	na	na	*LYPD1*	0.71	3.80×10^{-6}	
3	134815187	rs78690211	na	na	na	*EPHB1*	1.99	1.30×10^{-6}	
3	157311299	rs9876830	na	na	na	*C3orf55*	1.42	7.30×10^{-6}	
5	113387846	rs114618137	na	na	na	*lincRNA*	1.80	2.70×10^{-6}	
6	163602261	rs942635	na	na	na	*PACRG*	1.54	6.90×10^{-6}	
6	163603305	rs2763993	na	na	na	*PACRG*	1.49	4.60×10^{-5}	
7	83635586	rs4732529	na	na	na	*SEMA3A*	1.59	1.80×10^{-6}	
7	83670129	rs76881522	na	na	na	*SEMA3A*	1.56	1.40×10^{-5}	
10	95820702	rs112692056	na	na	na	*PLCE1*	2.20	3.30×10^{-6}	
13	27422997	rs74438932	na	na	na	Intergenic	1.90	1.10×10^{-6}	
17	67665781	rs6501341	na	na	na	*AC003051.1*	1.98	3.50×10^{-6}	
18	51427254	rs117914209	na	na	na	Intergenic	2.96	6.50×10^{-6}	
21	33704100	rs2096460	na	na	na	*URB1*	0.61	8.10×10^{-6}	
1	68916123	rs382422	C	G	0.22	*RPE65/DEPDC1*	2.10	3.21×10^{-6}	Scherag et al, 2016 [38]
3	11217691	rs58764888	A	T	0.02	*HRH1*	13.3	6.70×10^{-7}	
3	37853059	rs72862231	A	T	0.05	*ITGA9/ITGA9-AS1*	4.40	1.73×10^{-6}	
3	188004948	rs150062338	T	C	0.01	*LPP*	38.6	2.32×10^{-7}	
3	194027568	rs10933728	G	A	0.03	*LINC00887*	7.00	5.62×10^{-6}	

doi:10.6342/NTU202301905

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p	Study
4	856102	rs115550031	A	G	0.02	*GAK*	13.8	2.45×10^{-6}	Scherag et al, 2016 [38]
5	117409248	rs62369989	G	T	0.26	*LOC102467224*	2.10	7.98×10^{-6}	
6	33000554	rs115036193	T	C	0.01	*HLA-DOA/HLA-DPA1*	16.2	2.21×10^{-6}	
9	80020874	rs117983287	A	C	0.01	*VPS13A*	18.2	8.16×10^{-8}	
12	23661042	rs150811371	A	G	0.08	*ETNK1/SOX5*	3.40	2.93×10^{-6}	
13	27621985	rs945177	A	G	0.02	*GPR12/USP12*	14.7	1.31×10^{-6}	
13	69899506	rs9529561	G	A	0.08	*LINC00550/KLHL1*	3.90	3.34×10^{-7}	
16	84885777	rs2641697	G	C	0.36	*CRISPLD2*	2.00	5.99×10^{-6}	
17	14257083	rs7211184	C	G	0.72	*HS3ST3B1/CDRT7*	2.00	9.43×10^{-6}	
3	145685067	rs16857698	G	A	0.014	.	4.51	1.75×10^{-9}	Rosier et al, 2021[17]
3	145701146	rs5029231	T	C	0.019	.	3.99	1.37×10^{-8}	
3	145709314	rs6763296	C	T	0.018	.	4.67	2.55×10^{-9}	
3	145752473	rs16857836	T	G	0.014	.	5.20	5.51×10^{-10}	
8	143994806	rs4544	C	T	0.01	*CYP11B2*	6.87	8.86×10^{-9}	
8	144001245	rs11991278	T	C	0.01	*CYP11B2*	6.87	8.48×10^{-9}	
8	144007939	rs6981918	A	C	0.01	*CYP11B2*	6.85	8.74×10^{-9}	
9	86846933	rs956727	G	A	0.009	*SLC28A3*	17.4	3.22×10^{-8}	
12	112927208	rs7974468	A	G	0.013	*PTPNN11*	3.34	1.60×10^{-8}	
12	119712137	rs10849640	A	G	0.116	.	2.25	3.22×10^{-8}	
12	119712137	rs10849641	T	C	0.115	.	2.27	2.65×10^{-8}	
12	119725314	rs10849642	T	C	0.117	.	2.25	4.04×10^{-8}	
3	50556581	rs12491812	T	C	0.011	*CACNA2D2*	7.32	4.18×10^{-11}	

doi:10.6342/NTU202301905

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p	Study
3	50645158	rs2239753	C	T	0.011	*CISH*	7.02	2.80×10^{-11}	Rosier et al, 2021[17]
3	50645413	rs2239752	T	C	0.011	*CISH*	5.62	5.43×10^{-11}	
3	50647888	rs2239751	C	A	0.011	*CISH*	5.62	5.21×10^{-11}	
3	50651395	rs743753	T	C	0.011	*MAPKAPK3*	5.62	5.21×10^{-11}	
3	50668532	rs616689	A	G	0.014	*MAPKAPK3*	5.79	1.87×10^{-10}	
3	50685642	rs9879397	A	G	0.012	*MAPKAPK3*	4.86	8.79×10^{-9}	
3	50686517	rs2170840	C	A	0.014	*MAPKAPK3*	5.78	1.87×10^{-10}	
3	50698155	rs12492982	T	C	0.011	*MAPKAPK3*	7.32	4.18×10^{-11}	
3	50721892	rs2035484	G	A	0.011	*DOCK3*	5.62	5.21×10^{-10}	
3	50751643	rs17051403	A	C	0.011	*DOCK3*	5.62	5.21×10^{-10}	
3	65229760	rs17072628	A	G	0.012	.	4.96	8.25×10^{-9}	
8	89929277	rs7840669	G	A	0.015	.	3.77	2.38×10^{-8}	
12	79993704	rs7953683	T	C	0.024	*PAWR*	2.07	3.07×10^{-8}	
17	51544776	rs1502522	G	A	0.029	.	2.64	2.57×10^{-8}	
17	51560869	rs1393467	C	T	0.029	.	2.64	2.57×10^{-8}	
3	36047297	rs114581095	T	C	0.016	*ARPP21/STAC*	1.79	7.84×10^{-6}	Hernandez-Beeftink et al, 2022[22]
3	195775144	rs114658749	T	C	0.062	*SDHAP1/TFRC*	1.44	5.56×10^{-6}	
5	84399852	rs76805442	A	G	0.066	*EDIL3/NBPF22P*	1.65	3.39×10^{-6}	
7	25189764	rs73285901	A	G	0.032	*C7orf31*	1.37	8.77×10^{-5}	
7	92730745	rs34896991	T	C	0.018	*SAMD9*	1.64	4.92×10^{-8}	
11	26983813	rs146257041	G	A	0.015	*SLC5A12/FIBIN*	4.59	3.00×10^{-9}	
13	61367197	rs138347802	G	A	0.014	*LINC00378/MIR3169*	2.57	4.44×10^{-8}	

doi:10.6342/NTU202301905

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p	Study
13	93552176	rs113925942	C	T	0.022	*GPC5/LINC00363*	2.21	1.34×10^{-6}	Hernandez-Beeftink et al, 2022[22]
14	58109460	rs183364907	T	A	0.011	*SLC35F4*	2.02	1.41×10^{-6}	
17	14555200	rs146077854	A	G	0.01	*HS3ST3B1/LOC101928475*	7.25	4.65×10^{-7}	
21	39902441	rs1573332	A	T	0.345	*HS3ST3B1/LOC101928475*	1.33	1.14×10^{-5}	
1	201711585	rs114078858	T	C	0.017	*NAV1*	1.003	1.44×10^{-6}	Neale et al, 2018 [39]
2	59364237	rs75601073	G	A	0.021	.	1.003	9.01×10^{-7}	
2	30637041	rs114965692	C	T	0.059	.	1.001	5.29×10^{-7}	
3	21939427	rs259492	A	T	0.298	*ZNF385D*	1.001	3.86×10^{-6}	
4	80633295	rs13146131	C	G	0.141	.	1.001	8.35×10^{-6}	
4	80710324	rs34497103	A	G	0.155	.	1.001	4.08×10^{-6}	
4	80730535	rs36066750	C	G	0.156	.	1.001	4.95×10^{-6}	
4	80762635	rs7690505	C	T	0.155	*PCAT4*	1.001	1.64×10^{-6}	
4	80791075	rs35981701	T	C	0.151	.	1.001	2.85×10^{-6}	
4	80781873	rs4690159	C	G	0.156	*PCAT4*	1.001	2.97×10^{-6}	
4	156514879	rs1466515	G	A	0.103	.	0.999	2.47×10^{-6}	
4	171783881	rs114349669	T	A	0.014	.	1.002	8.79×10^{-10}	
5	87634748	rs17000198	T	C	0.016	*TMEM161B-AS1*	1.001	3.73×10^{-9}	
6	33049123	rs61406966	A	G	0.012	*HLA-DPB1*	1.003	5.13×10^{-6}	
6	33072598	rs16868789	G	A	0.014	.	1.003	5.23×10^{-6}	
6	36671178	rs74648178	A	G	0.014	*RAB44*	1.004	1.90×10^{-7}	
6	36689198	rs75667310	A	T	0.015	*RAB44*	1.004	3.05×10^{-9}	
6	160147058	rs41267779	C	G	0.052	*SOD2*	1.002	8.15×10^{-7}	

doi:10.6342/NTU202301905

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p	Study
6	160224203	rs56130537	G	A	0.05	*PNLDC1*	1.002	1.51×10^{-6}	Neale et al, 2018 [39]
7	78164937	rs79086515	G	A	0.015	*DSCC1*	1.003	1.24×10^{-6}	
8	41321196	rs72646737	C	G	0.172	.	0.999	6.04×10^{-6}	
8	80766054	rs4444770	A	G	0.154	*LOC101927040*	1.001	1.38×10^{-6}	
8	84157112	rs79500616	A	G	0.042	.	0.998	7.80×10^{-6}	
8	120857083	rs71532474	G	A	0.06	.	1.002	1.16×10^{-7}	
9	100479810	rs78009740	T	G	0.024	*PTCSC2*	1.002	7.39×10^{-6}	
9	117057928	rs74307801	T	G	0.059	*COL27A1*	0.998	2.89×10^{-6}	
10	35900880	rs78934109	T	C	0.03	.	1.002	9.13×10^{-6}	
11	111162890	rs4356266	T	A	0.295	.	1.001	5.73×10^{-6}	
11	131511485	rs148092072	T	A	0.018	*NTM*	1.003	6.39×10^{-6}	
12	54609469	rs77032716	G	A	0.016	.	1.003	9.73×10^{-6}	
12	91431056	rs7136342	T	C	0.176	.	0.999	5.70×10^{-6}	
12	130493235	rs61939273	A	G	0.111	.	1.001	1.67×10^{-6}	
13	49435399	rs78535155	G	A	0.014	.	1.003	3.37×10^{-6}	
14	54609469	rs185052458	T	C	0.01	.	1.003	9.26×10^{-6}	
14	81043621	rs75678135	C	T	0.011	*CEP128*	1.004	8.60×10^{-6}	
15	80532134	rs72740098	G	A	0.192	*XND1*	1.001	6.96×10^{-6}	
17	31170397	rs78219843	A	G	0.021	.	1.001	3.77×10^{-8}	
20	7580820	rs61939273	A	G	0.014	.	1.003	1.82×10^{-6}	
21	28901549	rs242323	G	A	0.147	.	1.001	3.81×10^{-6}	
22	17664296	rs28535971	G	C	0.345	*ADA2*	1.001	3.75×10^{-6}	

doi:10.6342/NTU202301905

CHR	Position	rsID	Minor allele	Major allele	MAF	Gene	OR†	p	Study
6	98104575	rs9489328	T	G	0.1	*AL589740.1*	0.1	1.05×10^{-10}	D'Urso et al, 2020[21]
5	142470334	rs11167801	T	C	0.09	*ARHGAP26*	0.13	8.77×10^{-8}	
13	111145073	rs368584	G	C	0.39	*COL4A2*	2.56	2.55×10^{-7}	
4	58088682	rs7698838	C	T	0.07	Intergenic	0.15	6.11×10^{-7}	
14	92882826	rs17128291	G	A	0.18	Intergenic	0.22	7.48×10^{-7}	

na: There was no information in the previous study

† In Hernandez-Beeftink et al. study, hazard ratio (HR) was used to represent the effect size of SNPs

CHR: chromosome number, MAF: minor allele frequency, OR: odds ratio

doi:10.6342/NTU202301905