



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

Institute of Epidemiology and Preventive Medicine

College of Public Health

National Taiwan University

Master's Thesis

台灣高雄地區結核病預測傳播群聚模型之建立與驗證

Development and Validation of Predictive Models for
Tuberculosis Transmission Clusters in Kaohsiung, Taiwan

陳奕欣

Yi-Hsin Chen

指導教授：林先和 博士

Advisor: Hsien-Ho Lin, MD, ScD

中華民國 114 年 7 月

July, 2025

摘要



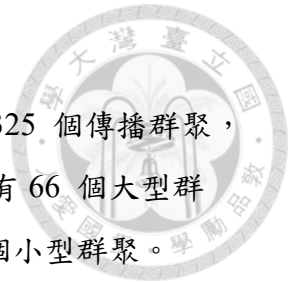
背景

結核病 (Tuberculosis) 是全球最嚴重的公共衛生威脅之一。根據世界衛生組織 (WHO) 估計，2023 年全球新增結核病病例達 1,080 萬例。由於結核病具有較長的潛伏期，僅依賴病患主動就醫常導致診斷延遲與持續傳播。因此，WHO 建議公共衛生體系應主動篩檢潛在個案，並及早介入，以實現「2035 年終結結核病」的全球目標。雖然過去已有多項研究探討導致群聚擴大的風險因子，並建立預測模型，但在台灣，我們缺乏結合分子生物技術與流行病學資料進行預測分析的相關研究。本研究針對台灣高雄地區的結核病傳播群聚進行分析，整合全基因體定序資料與流行病學特徵，以每個群聚中最早通報的前兩位個案資料建構二元分類模型，預測該群聚未來是否會擴展至兩例以上。本研究結果有助於建構早期預警系統，強化結核病防治策略，辨識可能形成大型傳播群聚的高風險因子。

方法

本研究為一項以台灣高雄地區為研究對象，納入 2019 至 2023 年間所有高雄地區通報的結核桿菌陽性個案。透過全基因體定序，分析菌株間的單一核苷酸多型性 (Single Nucleotide Polymorphism, SNP)，並以 SNP 差異不超過 12 個作為同一傳播群聚的判定標準。我們建立結合 WGS 與流行病學特徵的資料集，並以每個群聚中最早通報的前兩例個案為分析單位。若該群聚最終累計病例數達三例以上，定義為「大型群聚」；反之則為「小型群聚」。為排除研究初期可能已存在的盛行群聚對預測的干擾，本研究同時建立三種資料集：包含所有個案的完整資料集，以及僅包含 2019 年後新發群聚的兩個子資料集（分別追蹤兩年與三年）。

分析採用隨機森林、邏輯斯回歸與支援向量機三種模型進行預測建模。資料以 7:3 比例隨機分為訓練與測試集，建模流程包含缺失值填補、上採樣與交叉驗證，並重複執行 2,000 次以提升模型穩定性與可靠性。最後透過變項重要性排序與敏感度分析，評估模型預測效果。所有分析皆使用 R 4.5.0 版本執行。



結果

1,197 例個案經全基因體定序確認為傳播群聚成員，共辨識出 325 個傳播群聚，其中 126 個為大型群聚，199 個小型群聚。而兩年追蹤資料含有 66 個大型群聚，24 個小型群聚；三年追蹤資料含有 48 個大型群聚，24 個小型群聚。我們選用三個預測變數，首兩位通報個案的年齡、通報日期間隔、痰塗片結果。每種模型執行 2,000 次模擬，中位數 AUC 所對應的那次模擬結果，作為模型表現的代表。

使用完整資料集進行分析時，邏輯斯回歸模型表現最佳，其 AUC 為 0.690 (95% CI: 0.601 – 0.776)；SVM 次之，AUC 為 0.685 (95% CI: 0.596 – 0.772)；隨機森林表現較差，AUC 為 0.623 (95% CI: 0.532 – 0.711)。接著分析 2020 年後新發群聚，兩年追蹤期中，邏輯斯回歸模型的 AUC 為 0.643 (95% CI: 0.436 – 0.850)；隨機森林表現近乎隨機猜測，僅 0.507 (95% CI: 0.300 – 0.707)；SVM 則介於兩者之間，AUC 為 0.621 (95% CI: 0.328 – 0.821)。三年追蹤資料集邏輯斯回歸模型的中位 AUC 為 0.673 (95% CI: 0.449 – 0.857)；隨機森林表現同樣不佳，中位 AUC 為 0.541 (95% CI: 0.316 – 0.745)；SVM 中位 AUC 為 0.643 (95% CI: 0.357 – 0.847)。兩年與三年追蹤資料集樣本數小，導致 AUC 信賴區間較寬，預測表現的穩定性相對較低。在三種資料設定中，邏輯斯回歸模型的整體表現最穩定且最佳，SVM 次之，而隨機森林的預測能力最弱。

變項重要性分析顯示，前兩例個案的通報日期間隔為最具預測力的因子。而早期個案的通報間隔短，與平均年齡較輕的群集，日後較可能形成大型群聚。痰塗片結果的預測效果則變異性較大。而敏感度分析顯示，模型的預測力均非隨機猜測。綜合而言，研究結果指出，首兩例個案的通報間隔與年齡是預測群聚擴大的關鍵因子。

結論

結合痰塗片陽性率、平均年齡與通報日期間隔三個因子所建構的模型，具備早期預測大型群聚的潛力。根據本研究結果，超過一半的結核病傳播群聚僅包含兩例病例，顯示許多傳播事件能在早期即被中斷，凸顯了及時隔離病例及基礎流行病學調查在疫情控制中的關鍵角色。總體而言，我們應持續強化基礎防治措施，並對模型判定為高風險的群聚，及早啟動擴大調查。靈活且有針對性的應對策略，有助於優化有限公共衛生資源的運用，加快疫情反應速度，有效遏止疾病傳播。

關鍵字：結核病、傳播群聚、全基因定序、預測模型、機器學習

Abstract



Background.

Tuberculosis (TB) is an important issue around the world. According to the World Health Organization (WHO), there were 10.8 million new TB cases worldwide in 2023.

Due to TB's long latency period, relying solely on patients to seek medical attention often leads to delayed diagnosis and continued transmission. Therefore, the WHO recommends that public health systems proactively identify potential cases and intervene early to achieve the global goal of ending TB by 2035.

. This study investigated TB transmission clusters in Kaohsiung, Taiwan, by integrating whole-genome sequencing (WGS) data with epidemiological characteristics. Using information from the first two reported cases in each cluster, we developed binary classification models to predict whether a cluster was likely to expand beyond two cases. Our findings aim to develop early warning systems, enhance TB control strategies, and help identify risk factors associated with the formation of large transmission clusters.

Methods.

This study included all culture-positive *Mycobacterium tuberculosis* cases reported between 2019 and 2023 in Kaohsiung, Taiwan. Whole-genome sequencing (WGS) was conducted to analyze the isolates' single-nucleotide polymorphisms (SNP). We used a pairwise SNP distance threshold of 12 to define a transmission cluster. The dataset was created by combining WGS data with epidemiological characteristics, using the first two notified cases in each cluster as the unit of analysis. Clusters with three or more total cases were classified as large clusters, while those with fewer than three were designated as small clusters. Three datasets were constructed: A complete dataset that

included all cluster cases, and two subsets focused on incident clusters after 2019, each followed for two and three years, respectively. The two- and three-year follow-up datasets were designed to minimize potential bias from pre-existing clusters at the start of the study period. Three modeling approaches were applied: Random Forest (RF), Simple Logistic Regression (SLR), and Support Vector Machine (SVM). The data were randomly split into training and testing sets in a 70:30 ratio. We repeated the modeling process 2,000 times to ensure robustness and stability. Finally, variable importance calculation and sensitivity analysis were conducted to evaluate model performance. All analyses were performed using R version 4.5.0.

Results.

1,197 cases were included as members of transmission clusters based on whole-genome sequencing (WGS) data. We identified 326 clusters; 126 were classified as large and 199 as small clusters. In the two-year follow-up dataset, 66 large and 24 small clusters were included, while the three-year follow-up dataset comprised 48 large and 24 small clusters.

Three predictors were used in this study: the average age of the first two notified cases in each cluster, the time interval between their notification dates, and their sputum smear results.

In the whole dataset, the logistic regression model showed the best performance, with a median AUC of 0.690 (95% CI: 0.601–0.776), followed by support vector machine (SVM) with a median AUC of 0.685 (95% CI: 0.596–0.772). The random forest model performed the worst, with a median AUC of 0.623 (95% CI: 0.532–0.711). In the two-year follow-up dataset, the logistic regression model again outperformed others, achieving a median AUC of 0.643 (95% CI: 0.436–0.850). Random forest showed near-

random guessing performance with an AUC of 0.507 (95% CI: 0.300–0.707). SVM achieved an AUC of 0.621 (95% CI: 0.328–0.821). For the three-year follow-up dataset, logistic regression maintained the best performance (AUC = 0.673; 95% CI: 0.449–0.857), followed by SVM (AUC = 0.643; 95% CI: 0.357–0.847), while random forest performed poorly (AUC = 0.541; 95% CI: 0.316–0.745).

Feature importance analysis indicated that the notification interval between the first two reported cases was the most influential predictor. Clusters with shorter notification intervals and a younger average age were more likely to evolve into large clusters. The predictive contribution of sputum smear status varied across datasets. Sensitivity analyses further confirmed that the model's performances were not driven by random chance.

In summary, the notification interval and mean age of the first two reported cases were identified as key predictors of cluster expansion.

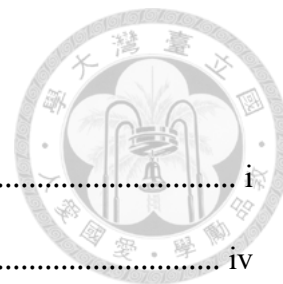
Conclusion.

Our study demonstrated that a predictive model incorporating sputum smear positivity, mean age, and notification date interval showed promise for early identification of large tuberculosis transmission clusters.

The expanded investigations can be initiated earlier for clusters identified as high-risk by the model. Implementing a flexible and targeted response strategy can optimize the allocation of limited public health resources and ultimately reduce the spread of tuberculosis.

Keywords: Tuberculosis; Transmission cluster; Whole-genome sequencing; Predictive model; Machine learning

CONTENTS



中文摘要	i
Abstract.....	iv
Chapter 1 Introduction.....	1
Chapter 2 Methods.....	8
2.1 Data Collection	8
2.2 Whole genome sequencing and bioinformatic analysis.....	8
2.3 Study Design.....	9
2.4 Statistical Analysis.....	11
2.5 Modeling Strategy	12
2.5.1 Framework	12
2.5.2 Random Forest	13
2.5.3 Logistic Regression	14
2.5.4 Support Vector Machine.....	14
2.5.5 Sensitivity Analysis	15
Chapter 3 Results.....	16
3.1 Characteristics of Study Participants	16
3.2 Classification Models Performance.....	19
3.2.1 Evaluation of Model Performance Using ROC Curves	19
3.2.2 Feature Importance and Estimated Effects.....	20
3.3 Sensitivity Analysis	22
Chapter 4 Discussion	23
4.1 Main Findings.....	23
4.2 Temporal Structure of Data	25

4.3	Underperformance of Random Forest models.....	26
4.4	Challenges in Measuring Transmission	26
4.5	Study Limitation.....	27
4.6	Conclusion	28
Reference.....		44



LIST OF FIGURES



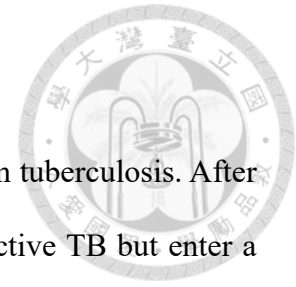
Figure 1	Study flow chart.....	29
Figure 2	Distribution of transmission cluster.....	30
Figure 3	Distribution of transmission cluster onset by year and quarter	31
Figure 4A	Model performance of full-data.....	32
Figure 4B	Model performance of two-year follow-up data.....	32
Figure 4C	Model performance of three-year follow-up data.....	33
Figure 5	Variable importance	34
Figure 6A	Boxplot of beta coefficients for sputum smear status in logistic regression models.....	35
Figure 6B	Boxplot of beta coefficients for mean age and notification date interval in logistic regression models.....	35
Figure 7A	Sensitivity Analysis of full-data.....	36
Figure 7B	Sensitivity Analysis of two-year follow-up data	36
Figure 7C	Sensitivity Analysis of three-year follow-up data	37

LIST OF TABLES



Table 1A	Basic characteristic of full-data (individual-level variables).....	38
Table 1B	Basic characteristic of full-data (cluster-level variables)	39
Table 2A	Basic characteristic of two-year follow-up data (individual-level variables)	40
Table 2B	Basic characteristic of two-year follow-up data (cluster-level variables)	41
Table 3A	Basic characteristic of three-year follow-up data (individual-level variables)	42
Table 3B	Basic characteristic of three-year follow-up data (cluster-level variables) ..	43

Chapter 1 Introduction

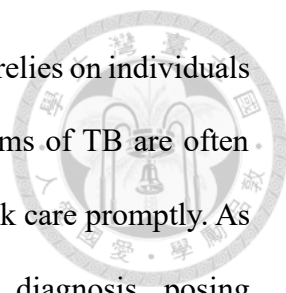


Tuberculosis (TB) is an infectious disease caused by *Mycobacterium tuberculosis*. After the infection, around 90% of people do not immediately develop active TB but enter a latent TB infection (LTBI) stage. While individuals with LTBI are not contagious, they face a lifetime risk of 5–15% of developing active, contagious TB, especially if their immune system becomes weakened. Approximately 5% of infected people develop active TB within two years after infection. TB mainly affects the lungs, but it can also spread to other parts of the body. The disease can transmit through airborne droplets expelled during activities like coughing or speaking with close contacts. Moreover, those nearby, especially household members, are most likely to catch the infection [1].

The World Health Organization's Global TB Report 2024 [2] estimated that in 2023, there were 10.8 million new tuberculosis cases worldwide, with an incidence rate of 134 cases per 100,000 people. The highest burden was seen in South Asia, accounting for 45% of cases, followed by the Western Pacific region with 17%. TB caused approximately 1.25 million deaths, making it the top cause of mortality from a single infectious agent.

In Taiwan, the 2023 Annual TB Control Report [3] identified Taiwan as a moderate TB burden country. Since adopting the Directly Observed Therapy, Short-course (DOTS) strategy approach in 2006, Taiwan has made significant progress in improving treatment compliance and reducing transmission, achieving % DOTS coverage rates 98% for TB cases. There was a steady decline in tuberculosis incidence, from 72.5 per 100,000 population in 2005 to 28.2 per 100,000 in 2023. A total of 6,584 new TB cases were reported in 2023, with a mortality rate of 1.9 per 100,000, primarily affecting individuals aged 65 and older.

To achieve the goal of ending tuberculosis, two intervention strategies are currently



employed to control its transmission. Traditional passive case finding relies on individuals seeking medical care on their initiative. However, the early symptoms of TB are often mild or nonspecific, many patients may not recognize the need to seek care promptly. As a result, transmission may occur within the community before diagnosis, posing significant challenges to TB control. Individuals with latent TB infection who remain undiagnosed may progress to active disease in the future, further spreading the disease. Relying solely on passive case detection is not sufficient to achieve the World Health Organization's (WHO) End TB goal by 2035. WHO recommends implementing systematic screening for active TB among high-risk groups to enhance early detection and interrupt transmission chains [4]. Active case finding enables early identification of latent TB infection, allowing timely preventive therapy that reduces progression to active disease and shortens the transmission period. High-risk groups targeted for screening typically include close contacts of TB patients, immunocompromised individuals, and populations with limited access to healthcare services. Compared to passive case finding, this approach facilitates effective control of TB spread. However, active screening and contact tracing are resource-intensive. It must be prioritized based on risk assessment, feasibility, and cost-effectiveness. Targeted strategies with clearly defined high-risk populations are essential for optimizing public health resources and maximizing the effectiveness of TB control programs.

Establishing an early warning system is crucial for timely responses to TB outbreaks. This requires a clear definition of what constitutes a transmission cluster. Before the widespread adoption of molecular tools, tuberculosis (TB) transmission links were primarily inferred based on clinical information and contact history. Epidemiological investigations can be confused by recall bias, incomplete information, and human error. In contrast, molecular techniques offer a more accurate and objective way to assess the

genetic relatedness between TB isolates, providing deeper insight into transmission patterns and improving outbreak detection.

IS6110-based Restriction Fragment Length Polymorphism (RFLP) analysis emerged as one of the earliest standard tools for distinguishing genetic differences between *Mycobacterium tuberculosis* strains [5]. IS6110 is a DNA sequence specific to the *M. tuberculosis* complex. Its copy number and chromosomal insertion sites vary significantly among strains. They use restriction enzymes to cleave DNA on only one side of the IS6110 element, and the bacterial genome can be fragmented to reflect the number and positions of IS6110 insertions. These fragments are then separated by gel electrophoresis. The resulting IS6110 profiles differ between strains, which serve as reliable molecular fingerprints to assess their transmission relationship.

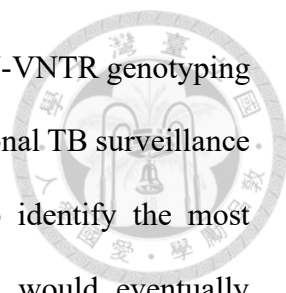
IS6110-RFLP involves time-consuming cultivation of slow-growing bacteria and requires complicated laboratory procedures. Therefore, Researchers developed spoligotyping as an alternative method to address these limitations [6]. Spoligotyping targets the direct repeat (DR) region of the *M. tuberculosis* complex genome, which consists of multiple conserved 36 bp direct repeats interspersed with unique 34–41 bp spacer sequences. PCR amplification of this region allows analysis even from specimens with low bacterial load. The PCR products are hybridized with 43 known spacer oligonucleotide probes. It generates a distinct binary pattern, which can rapidly distinguish strains and identify related cases.

In bacterial chromosomes, certain DNA sequences are arranged in direct repeats at specific genomic locations. These structures are called tandem repeat sequences, with repeat units typically ranging from 2 to 60 base pairs in length. Researchers developed the Variable Number Tandem Repeat (VNTR) analysis method based on this feature. Multiple VNTR loci have been identified across the MTBC genome, termed

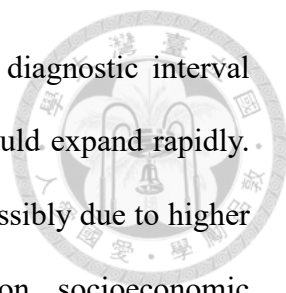
Mycobacterial Interspersed Repetitive Units (MIRUs). The MIRU-VNTR typing method utilizes 5 to 24 specific loci and determines the number of repeats at each locus. Because individual MTBC strains exhibit different repeat numbers at these sites, the combination of these values forms a profile that serves as a basis for strain identification and epidemiological tracking. This method offers more precise information compared to previous methods [7].

With the rapid advancement and decreasing cost of whole genome sequencing (WGS), it has become a core tool in bacterial strain identification and transmission inference. WGS offers the highest resolution for differentiating bacterial strains by comparing single-nucleotide polymorphisms (SNP) across genomes. The SNP distance refers to the number of differing nucleotide positions between two pathogen genomes. A smaller distance indicates greater genetic similarity and suggests a shared transmission chain or common source, whereas a larger distance implies a lower likelihood of relatedness. In molecular epidemiology, SNP thresholds are commonly established to determine whether two isolates belong to the same transmission cluster. A threshold of 12 SNPs is widely accepted as the upper limit for inferring recent transmission. Isolates with more than 12 SNPs of divergence are generally excluded from being part of the same transmission event. However, the appropriate threshold may vary depending on epidemiological settings, strain lineage, and mutation rate. The mutation rate of *Mycobacterium tuberculosis* is estimated at approximately 0.5 SNPs per genome annually. And the divergence rarely exceeds 5 SNPs in three years between pairs [8]. In Taiwan, local research has proposed a threshold of ≤ 5 SNPs to define confirmed transmission events, and ≤ 15 SNPs for possible transmission [12].

Since 2012, research teams in the United States explored factors associated with the expansion or outbreak of tuberculosis (TB) clusters. At that time, whole-genome

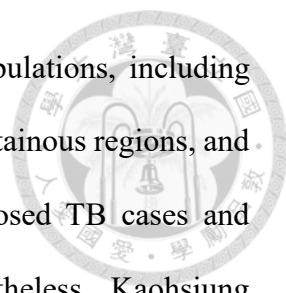


sequencing (WGS) was not widely applied. Spoligotyping and MIRU-VNTR genotyping methods were commonly used to define clusters. They analyzed national TB surveillance data in the United States. A decision tree model was applied to identify the most informative splitting variables that best predicted which clusters would eventually become outbreaks. Clusters had a significantly higher chance of developing into outbreaks when at least one of the first three patients had a socially vulnerable condition, such as homelessness and incarceration, or the diagnosis date interval between the third and the first case was within 5.3 months [13]. Several years later, the research team employed a Cox proportional hazards model to analyze U.S. data from 2009 to 2018. They reported that for each additional month between the first and third case diagnoses, the risk of subsequent cluster expansion decreased by approximately 5%. Moreover, if at least one of the first three cases was over 65, the expansion risk dropped by about 60%. It reflected the reduced transmissibility among elderly individuals [14]. In a subsequent study, they combined the U.S. TB surveillance data with a negative binomial hurdle model to detect clusters experiencing unexpected growth. The most important predictor was the time in months between the first case and the last case reported during the quarter of unexpected growth, underscoring the predictive value of early case accumulation speed for transmission risk [15]. Researchers also conducted a similar study in the Netherlands, which used RFLP genotyping to define clusters. They focused on the characteristics of the first two cases in newly formed clusters to predict whether the cluster would grow to five or more cases within two years. Several factors were significantly associated with cluster expansion, including shorter diagnosis intervals, younger patient age, urban residence, and shared geographic origin. A logistic regression model incorporating these factors demonstrated good predictive performance (AUC=0.79), again highlighting the importance of early case characteristics [16].



Overall, studies based on cluster-level data demonstrated that the diagnostic interval between early cases was a critical predictor of whether a cluster would expand rapidly. Clusters involving younger individuals were more likely to grow, possibly due to higher levels of social interaction or transmission potential. In addition, socioeconomic background (e.g., place of residence, ethnicity) and high-risk behaviors (e.g., smoking, drug use) might further increase the likelihood of cluster expansion. However, most earlier studies relied on lower-resolution genotyping techniques, such as spoligotyping or MIRU-VNTR. With the growing availability of whole-genome sequencing (WGS) in recent years, people began utilizing higher-resolution genomic data to better understand tuberculosis transmission dynamics. Researchers analyzed WGS data from Canada to differentiate early two individuals of small clusters (cluster size = 2) from those in large clusters (cluster size > 2). The study defined potential recent transmission using a threshold of less than 12 SNPs between isolates. Several machine learning and deep learning algorithms were applied. Balanced logistic regression achieved the highest predictive performance (AUC = 0.57), while balanced bagging performed the poorest (AUC = 0.43). All predictor variables were binary individual-level features, including demographic and clinical characteristics, but no cluster-level variables were incorporated [17].

Although whole genome sequencing (WGS) has been increasingly utilized worldwide in tuberculosis (TB) surveillance studies, no research in Taiwan has yet integrated high-resolution genomic data with epidemiological cluster-level characteristics to predict the emergence of large TB clusters. This study investigates TB transmission clusters in Kaohsiung, Taiwan, using data from reported cases between 2019 and 2023. According to Taiwan's national TB surveillance reports, 5,657 new TB cases were reported during these five years [18]. In recent years, Kaohsiung City has actively implemented latent TB



infection screening and treatment programs targeting high-risk populations, including residents of long-term care facilities, individuals in indigenous mountainous regions, and patients with chronic comorbidities. The number of newly diagnosed TB cases and incidence rate have shown a gradual downward trend. Nevertheless, Kaohsiung consistently ranks among the cities with the highest TB incidence in Taiwan. This suggests that additional risk factors may exist, and identifying them could allow health authorities to better target screening and intervention, particularly for individuals with latent or undiagnosed active TB.

By leveraging whole-genome sequencing (WGS) data, we explored potential transmission events with higher resolution. We identified the first two notified cases within each transmission cluster to construct cluster-level features and developed binary classification models to predict the expansion of disease clusters. This study highlights key predictive factors associated with cluster growth through model evaluation and feature importance analysis. Our findings aim to support evidence-based public health surveillance and improve early intervention strategies to control TB transmission.

Chapter 2 Methods



2.1 Data Collection

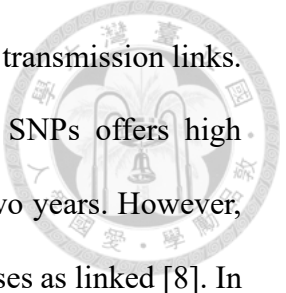
This population-based, prospective study was conducted in Kaohsiung, Taiwan, and included all culture-positive tuberculosis (TB) cases notified between January 2019 to December 2023. Our analysis integrated both public health and laboratory data. For TB cases managed in Kaohsiung, specimens identified as *Mycobacterium tuberculosis* complex (MTBC) were cultured and tested for drug susceptibility in certified laboratories. The resulting isolates were preserved and subsequently subjected to Kaohsiung Medical University Hospital, where pre-processing procedures for genome sequencing (WGS) were conducted.

Public health information was obtained from the routine TB control program. It included epidemiological, clinical, and demographic data, such as contact investigation records, lifestyle assessments, residential addresses, comorbidities, and socioeconomic status.

2.2 Whole genome sequencing and bioinformatic analysis

DNA samples were sequenced at the Laboratory of Genomics and Bioinformatics Service, Texas A&M AgriLife. Sequencing was performed using the Illumina NovaSeq 6000 SPx platform with a read length of 150 base pairs, with a paired-end sequencing approach. The resulting reads were aligned to the *Mycobacterium tuberculosis* H37Rv reference genome. Single nucleotide polymorphisms (SNP) were identified as positions where at least 75 percent of the nucleotides among the aligned reads differed from the reference. SNPs with ambiguous base calls, low coverage, or PPE and PE_PGRS gene families were excluded from further analysis.

Based on the identified SNPs, pairwise SNP distances between isolates were calculated.

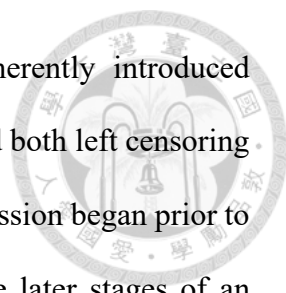


An appropriate SNP distance threshold was required to infer potential transmission links. A study from the Netherlands suggested that a threshold of ≤ 12 SNPs offers high sensitivity for capturing recent transmission events within the past two years. However, higher thresholds may increase the risk of misclassifying unrelated cases as linked [8]. In Taiwan, studies indicated that a threshold of ≤ 5 SNPs is typically applied in low-transmission settings to indicate strong epidemiological links and likely direct transmission. In contrast, a threshold of ≤ 15 SNPs is more commonly used in prolonged outbreaks or high-burden settings to encompass broader temporal and spatial transmission [12]. Since Taiwan is classified as a medium-incidence country and our study spans five years, we selected a 12-SNP threshold as a reasonable cutoff to infer potential transmission events. This threshold is also consistent with widely accepted standards in genomic epidemiology.

Therefore, in this study, if the SNP distance between two isolates was fewer than 12, they were considered to have a potential transmission link and were assigned to the same transmission cluster.

2.3 Study Design

This study aimed to investigate whether the development of large clusters could be predicted by the characteristics of its first two patients. Previous studies based on long-term data observed that more than half of tuberculosis clusters consist of only two cases [16]. Clusters involving only two cases may represent transmission among individuals with close personal contact, such as family members or partners. In contrast, clusters comprising three or more cases are more indicative of widespread community transmission. As a result, we defined clusters with three or more cases as large clusters, while those with two cases were classified as small clusters.



This study utilized data spanning a five-year period, which inherently introduced challenges related to temporal structure. Specifically, we encountered both left censoring and right censoring biases. Left censoring occurred when TB transmission began prior to the start of our observation period. In such cases, we captured the later stages of an ongoing transmission cluster, while the initial transmission events remain unobserved. Clusters that had already expanded before the study period were misclassified as small clusters due to the limited observation window. Conversely, right-censoring posed a challenge due to the ongoing nature of TB transmission. Clusters observed near the end of the study period might still be in their early stages and could develop into larger outbreaks in the future. Misclassifying such clusters as small clusters due to limited observation time introduced bias. However, excluding the earliest and latest data would significantly reduce statistical power because of the small sample size. Therefore, we adopted a compromise strategy. To minimize the impact of censoring, we restricted the analysis to clusters that were more likely to represent newly emerging events. Based on previous literature, approximately five percent of TB infections progressed to active disease within the first two years [1]. A two-years and three-years window was a reasonable time-frame for analysis.

This study evaluated the predictive performance of models using three datasets. The first dataset included all qualified tuberculosis (TB) cases reported during the five-year study period. The other two datasets were restricted to incident clusters, which initial case was reported after 2019. All notified cases within two-year and three-year periods following the initial case in each cluster were included. We excluded clusters for which the observation period extended beyond December 31, 2023. In each dataset, cases within each cluster were chronologically ordered based on their notification dates, and the first two notified cases from each cluster were extracted for subsequent analysis.

2.4 Statistical Analysis

To compare feature differences between individuals in the large and small cluster groups, we performed univariable analyses using chi-square tests for categorical variables and Student's t-test for continuous variables. In cases where the expected cell counts were less than five, the chi-square test was replaced with Fisher's exact test to ensure statistical validity. All analyses excluded missing values (NA) to avoid misleading results.

Cluster-level variables were calculated based on three demographic variables and individual categorical variables.

First, the spatial locations of individual cases were recorded by projected coordinates under the TWD97 TM2 zone 121 coordinate reference system. Each pairwise distance was computed using the Euclidean formula. Every individual in the dataset provided three types of addresses: residential, household, and workplace address. All possible pairwise distance combinations were computed for each cluster, and the minimum distance was taken as the cluster-level geographic distance. We imputed its missing distance value using the average value derived from clusters with the same outcome. Distance was expressed in kilometers. Second, we calculated the average age and age difference between the two cases for each cluster. And all age-related variables were expressed in years. Next, notification date interval in months between the two cases were computed, which were used as a proxy for early cluster growth rate for each cluster. At last, all categorical variables were treated as factor-level variables. For each pair, if both individuals exposed to the specific risk factor, the variable was coded as 2; if only one individual was exposed, it was coded as 1; and if neither was exposed, the variable was coded as 0. The creation process of cluster-level categorical variables was performed after the dataset splitting and imputation to prevent data leakage.

Potential predictors were selected from these variables to develop a predictive model.

2.5 Modeling Strategy

In this study, three modeling algorithms were employed: Random Forest, Simple Logistic Regression, and Support Vector Machine. The following sections are the analytical workflow and model-specific parameter settings applied throughout the analysis.

2.5.1 Framework

After conducting univariable analyses to identify candidate predictors, the modeling process proceeded as follows:

1. Data splitting: A stratified random split based on clusters was performed to divide the data into training and testing sets with a ratio of 70:30.
2. Missing data imputation and cluster-level variable construction: Missing data in both training and testing sets were imputed separately using the imputation method provided by the randomForest package [22]. After imputation, cluster-level categorical variables were constructed based on the imputed data within each set. The data were then analyzed at the cluster level. The training data were upsampled and then used for both hyperparameter tuning and the development of the final model.
3. Hyperparameter Tuning: For the Random Forest and SVM models, five-fold cross-validation was conducted to determine the optimal hyperparameters. Simple Logistic Regression did not involve hyperparameter tuning; the validation process was applied to evaluate model performance on the training data.
4. Model Fitting: Use the full upsampled training dataset and the optimal hyperparameters to construct final models.
5. Model Evaluation: The fitted models were evaluated on the testing dataset to assess their predictive performance.

This entire procedure (steps 1–5) was repeated 2,000 times. In each iteration, data splitting and cross-validation were randomized to enhance model robustness and ensure

stability of the results.

All analyses were performed using R version 4.5.0.



2.5.2 Random Forest

Random Forest is an ensemble learning method that combined the concepts of Bootstrap Aggregation (Bagging) and Decision Trees [23]. Each tree is built by randomly sampling from the training data (bootstrap sampling). And at each split, a random subset of features is selected. This study set the number of trees (ntree) to 500. Each tree was constructed by minimizing the Gini impurity to produce the purest possible splits at each node. Finally, the predictions from all trees were aggregated, and the final classification result was determined by majority voting.

The ranger package [24] was used to construct Random Forest models. Combining with the caret package [25], two hyperparameters were tuned via five-fold cross-validation on the training dataset.

1. mtry: the number of candidate variables randomly selected at each split. It controlled how diverse individual trees were. We considered mtry values of 2, 3, and 4.
2. min.node.size: the minimum number of data points required to form a terminal node (leaf). Smaller values lead to deeper trees and potentially overfitting, while larger values could restrict model complexity and increase bias. Values of 1 through 5 were evaluated.

The combination of “mtry” and “min.node.size” that achieved the highest average classification accuracy across folds was selected as the optimal setting.

The model’s predictive performance was evaluated on the independent testing set. The importance of the permutation-based variable was computed using the out-of-bag (OOB) method. When constructing the final model, each tree was trained on a bootstrap sample

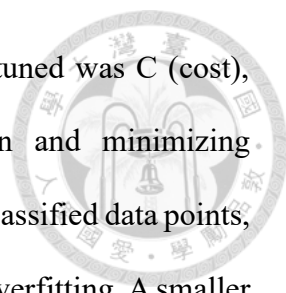
drawn from the training data. Approximately one-third of the observations were excluded from each bootstrap sample. These excluded cases constitute the Out-of-Bag (OOB) set for that tree. The model then generated predictions on the OOB data to compute a baseline classification error rate. To assess the importance of each predictor, the values of that variable in the OOB set were randomly permuted, disrupting any true association with the outcome. The model again predicted outcomes using the permuted OOB data, and the resulting classification error was recalculated. The difference between the permuted and baseline error rates reflected how that predictor contributed to model accuracy. This procedure was repeated across all variables and all trees, and the average increase in error due to permutation was used as the final importance score for each variable.

2.5.3 Logistic Regression

Logistic Regression is a linear model commonly used for binary classification issues. It models the association between independent variables and a binary outcome by applying a sigmoid function and transforms the output to a probability value between 0 and 1 [26]. In our study, we classified observations based on a threshold probability of 0.5. Probability values equal to or greater than 0.5 were predicted as positive class. We applied the base R glm function to construct simple logistic regression models without hyperparameter tuning. The regression coefficients (β coefficients) from each model iteration were recorded.

2.5.4 Support Vector Machine

Support Vector Machine (SVM) aims to identify a hyperplane in the feature space that could optimally separate data points belonging to different classes [27]. In this study, we used a linear kernel without dimensionality expansion. Model construction was



performed using the caret package. The key hyperparameter to be tuned was C (cost), which controlled the trade-off between maximizing the margin and minimizing classification error. A larger C value imposes a higher penalty on misclassified data points, thereby reducing model tolerance to error and increasing the risk of overfitting. A smaller C value increases model tolerance but may result in underfitting. Hyperparameter tuning was conducted via grid search, where C was tested over a range from 0.5 to 10 in increments of 0.5. The value of C that yielded the highest classification accuracy was selected for the final model.

2.5.5 Sensitivity Analysis

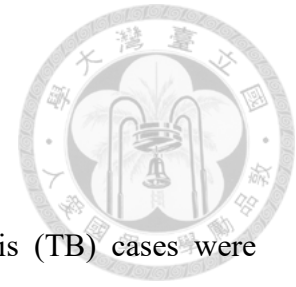
We conducted a sensitivity analysis to evaluate whether the models' predictive performance was better than random guessing. Before modeling, the training dataset's outcome variable was shuffled, disrupting the original associations between predictors and the outcome. The test set was kept unshuffled. The identical modeling procedures were applied to the shuffled training data across 2000 iterations. If the models demonstrated predictive ability beyond random chance, we anticipated that those trained on shuffled outcomes would yield poor performance. The sensitivity analysis was evaluated using the Receiver Operating Characteristic (ROC) curve.

Chapter 3 Results

3.1 Characteristics of Study Participants

From January 1, 2019, to December 31, 2023, 4,918 tuberculosis (TB) cases were reported and managed in Kaohsiung City and included in our dataset. These cases represented approximately 87% of all TB notifications in Kaohsiung during the five-year study period. Among them, whole genome sequencing (WGS) data were available for 4,168 cases. Of these, 1,200 patients were identified as belonging to a transmission cluster. Due to the small sample size, three patients infected with *Mycobacterium tuberculosis* Lineage 3 were excluded from further analysis. As a result, we included 1,197 patients in the final analysis. Three datasets were constructed for model development. The first included all 1,197 patients with five years of follow-up. The other two datasets only included clusters with an onset after 2020. These datasets included 212 and 183 patients, respectively, corresponding to 2-year and 3-year follow-up periods (Figure 1).

From Figure 2, we observed that most clusters were relatively small in size, predominantly consisting of two individuals. Among all clusters, 126 were classified as large clusters, and 199 were small clusters. Clusters with two individuals accounted for approximately 61% (199 out of 325), while clusters with three individuals represented about 18% (57 out of 325). A few clusters exhibited substantially larger sizes, with the largest cluster comprising up to 55 cases. Figure 3 illustrates the onset timing of the first case in each of the 325 clusters. Large and small clusters declined over the course of the study period. 175 clusters had their onset in 2019, whereas only 13 were first identified in 2023. This pattern suggested that some clusters detected early in the observation period (i.e., 2019) might have already existed before the study, indicating a potential left-censoring issue.



Over time, the proportion of large clusters relative to all clusters decreased. Considering the long latency period of tuberculosis and the five years of our dataset, it is possible that some small clusters eventually grew into large clusters beyond our observation window. However, this study could not capture such developments, which represented a right censoring limitation.

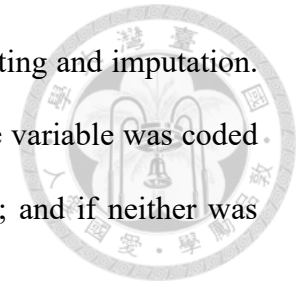
In the five-year full dataset, a total of 325 transmission clusters were identified, including 199 small clusters and 126 large clusters. For analysis, the first two notified cases from each cluster were selected, resulting in 398 individuals from small clusters and 252 from large clusters, for a total of 650 participants (Table 1A). At the individual level, there were no significant differences between the two groups. However, at the cluster level of continuous variables (Table 1B), a significant difference in smear test (p value < 0.05) was observed. On average, large clusters had a younger mean age, about five years younger than those in small clusters. They also had a shorter interval between the notification dates of the first two cases, with a mean difference of about eight months. The two-year follow-up analysis included 180 participants (Table 2A), comprising 24 large clusters and 66 small clusters (Table 2B). Also, a significant difference was observed between the two groups in terms of the notification date interval. The large cluster showed a shorter interval between notification dates, approximately three months shorter than that of the small cluster. Notably, the proportion of smear-positive cases in the large cluster was more than twice that of smear-negative cases. The three-year follow-up dataset (Table 3A) included 144 participants in total, grouped into 24 large clusters and 48 small clusters (Table 3B). The notification date interval between groups was significant, with the large cluster exhibiting notification intervals approximately seven months shorter than those in the small cluster. No additional significant variables were identified through univariate analysis.

Across all three datasets from the Kaohsiung region, the male-to-female ratio within both large and small clusters consistently remained at approximately 7:3. These data also suggested that men were more at risk of contracting TB than women.

We conducted univariable analyses to identify variables suitable for predictive models, using a p-value threshold of <0.05 as the initial selection criterion. Among the three datasets analyzed, the notification date interval consistently emerged as a significant factor distinguishing large and small clusters, making it the most stable and reliable predictor across all settings. However, including only the notification date interval in the model would result in an overly simplistic approach, overlooking other important factors contributing to cluster formation. Previous studies indicated that younger populations tended to have higher transmission potential. In our full dataset, the mean age reached statistical significance. Although this was not replicated in the two- and three-year follow-up datasets, we still observed that the average age in large clusters was slightly lower than that in small clusters, suggesting a potential role in transmission dynamics. Sputum smear-positive patients are known to be more infectious than smear-negative individuals, as supported by existing literature. The sputum smear test did not reach statistical significance in the two-year follow-up dataset ($p \approx 0.08$), but the p-value was close to the threshold. This result implied its borderline significance and possible predictive relevance. Most categorical variables across the three datasets did not achieve statistical significance. Pronounced class imbalance was observed in the two- and three-year follow-up data, which posed challenges for subsequent analysis. In light of these limitations, we excluded other categorical variables and retained the sputum smear test as the only representative categorical predictor.

Based on these findings, we chose three variables for the construction of the predictive model: notification date interval, mean age, and sputum smear test results. The sputum

smear test variable was generated at the cluster level after data splitting and imputation. Specifically, if both individuals in a cluster were smear-positive, the variable was coded as 2; if only one individual was smear-positive, it was coded as 1; and if neither was smear-positive, it was coded as 0.



3.2 Classification Models Performance

3.2.1 Evaluation of Model Performance Using ROC Curves

Figure 4 showed the ROC curves for the three models under different data settings. For each model, we performed 2,000 iterations to estimate the distribution of AUCs. We selected the ROC curve from the iteration whose AUC was closest to the median value, representing the model's median ROC curve.

Figure 4A showed the results using the full dataset. A total of 325 clusters were included in the analysis, with 98 clusters in the testing set, of which 38 were classified as large clusters. In this setting, logistic regression performed best, achieving a median AUC of 0.690 (95% CI: 0.601–0.776), while random forest yielded lower performance, with a median AUC of 0.623 (95% CI: 0.532–0.711). SVM's performance fell between the two; the median AUC value was 0.685 (95% CI: 0.596–0.772). We then examined the two-year follow-up data (Figure 4B). Among the 90 clusters included in the analysis, 27 were allocated to the testing set, with 7 identified as large clusters. In this scenario, logistic regression achieved a median AUC of 0.643 (95% CI: 0.436–0.850), whereas random forest performed poorly, with a median AUC of only 0.507 (95% CI: 0.300–0.707), nearly at the level of random guessing. SVM achieved an AUC of 0.621 (95% CI: 0.328–0.821). The three-year follow-up dataset included 72 clusters, with 21 assigned to the testing set, among which 7 were classified as large clusters. (Figure 4C). Logistic regression showed a median AUC of 0.673 (95% CI: 0.449–0.857), followed by SVM (AUC = 0.643; 95%

CI: 0.357–0.847). Random forest had a median AUC of 0.541 (95% CI: 0.316–0.745), again showing relatively poor performance.

Overall, the two- and three-year follow-up datasets had smaller sample sizes and fewer than 10 positive cases in the testing sets, resulting in wide confidence intervals for the AUC estimates and reduced reliability in performance assessment. Across all settings, logistic regression consistently showed the best performance, followed by SVM, while random forest yielded the weakest results.

3.2.2 Feature Importance and Estimated Effects

We used permutation importance to evaluate variable importance and compared results across the three datasets (Figure 5). The overall patterns of variable importance were consistent across datasets. The notification date interval showed the highest importance, followed by the mean age. Figure 6A and 6B were boxplots of the β coefficients from logistic regression models based on 2,000 simulations. Figure 6A illustrated the estimates for the categorical variable sputum smear test. In the two-year follow-up data (highlighted in pink), the estimated β coefficients exhibited considerable variability, with some simulations exceeding 10, indicating substantial model uncertainty under these conditions. Compared to the reference group (both were smear-negative), the median β coefficient for the group with only one smear-positive individual was 1.16, corresponding to an odds ratio (OR) of approximately 3.19 (95% CI: 0.66–17.9). When both individuals were smear-positive, the median OR was approximately 5.82 (95% CI: 1.18–18.5). In the three-year follow-up data (highlighted in blue), the odds ratio was approximately 1.97 (95% CI: 0.45–18.2) for the group with one smear-positive individual, and 3.13 (95% CI: 0.73–18.7) for the group in which both individuals were smear-positive. Using the full dataset model (highlighted in orange), previously observed positive effects appeared to diminish

and even reverse. The median β for the one smear-positive group was -0.273 , corresponding to an OR of 0.76 (95% CI: 0.42–1.37), while the both smear-positive group had a median β of 0.0527, with an OR near 1.05 (95% CI: 0.59–1.87). Notably, both confidence intervals indicated a lack of statistically significant effect.

Figure 6B examined the impact of two continuous variables, the mean age of the first two notified cases and the interval between their notification dates. In the two-year follow-up data, a 10-year increase in mean age was associated with an approximate 9% decrease in the odds of developing a large cluster (OR = 0.91, 95% CI: 0.66–1.24). The notification date interval showed an inverse relationship, with each additional month corresponding to a significant reduction in the odds of large cluster formation (OR = 0.91, 95% CI: 0.84–0.97). Similarly, in the three-year dataset, both variables demonstrated negative associations. For every 10-year increment in mean age, the odds ratio was about 0.81 (95% CI: 0.54–1.01), suggesting a possible inverse effect. The notification interval maintained a stable negative association, with an OR of 0.90 (95% CI: 0.85–0.94) per additional month. Analysis of the full dataset revealed a stronger inverse relationship. A 10-year increase in the mean age of the first two notified cases corresponded to a reduced likelihood of forming a large cluster (OR = 0.69, 95% CI: 0.59–0.81). Additionally, each month increase in notification date interval was linked to a 4% decrease in the odds of large cluster formation (OR = 0.96, 95% CI: 0.94–0.97).

In summary, the influence of sputum smear positivity on large cluster formation appeared inconsistent. While short-term follow-up data suggested a positive correlation with increased risk, this association was not maintained in the full data analysis and might even reverse. Both mean age and the interval between notification dates consistently showed inverse associations with large cluster formation across all datasets.

3.3 Sensitivity Analysis

Figure 7A to 7C presented the ROC curves and median AUC scores from 2,000 iterations of the three models, each based on random assignment of the response variable. Across all settings, the resulting AUC values were close to 0.5, suggesting that the observed classification performance was not due to chance but captured the true underlying pattern of the data.

Chapter 4 Discussion

4.1 Main Findings

The predictive variables included in this study were sputum smear positivity, mean age, and notification date interval. Among the three datasets used in this study, the two predictors, age and the time interval between notification dates of two cases, showed similar trends in their impact on outcomes.

Rapid accumulation of cases is often observed during tuberculosis outbreaks, which may indicate that the pathogen possesses high transmissibility and virulence. Such patterns can serve as early warning signals for large-scale cluster expansion. Alternatively, when multiple cases appear within a short time frame, it typically suggests that early infections were not promptly detected or controlled, resulting in sustained transmission. The close timing of notification dates may reflect ongoing transmission that has occurred for some time but was simultaneously identified due to public health investigations. Therefore, in some cases, the proximity of notification dates is partly influenced by human factors rather than representing the intrinsic transmission speed of the disease.

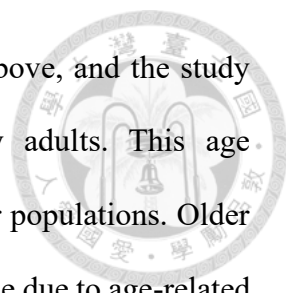
Sputum smear positivity was a key clinical indicator of tuberculosis (TB) infectiousness. The test involved acid-fast staining combined with high-power microscopic examination to detect the presence of acid-fast bacilli in respiratory specimens. The detection threshold typically required approximately 5,000 bacilli per milliliter of sputum, and the results showed a strong correlation with bacterial load [28]. A higher bacillary burden increased the likelihood of a positive smear result, indicating greater infectious potential. Researchers identified sputum smear positivity and the presence of cavitory lesions on chest imaging as two significant clinical markers of high infectiousness. Their study demonstrated that the first two cases with these characteristics were more likely to drive



the expansion of *M. tuberculosis* strain clusters [29]. However, in our analysis, the predictive power of smear positivity was not as strong or statistically significant as that of other variables. Moreover, its direction of association varied across models.

In some cases, smear positivity appeared to have a protective effect, though none of the results reached statistical significance. Several factors may account for this observation. Smear positivity was a categorical variable in our study, and the limited sample size, coupled with high variability across datasets and years, might reduce the stability and precision of the statistical estimates. In Taiwan, smear-positive cases are considered highly infectious and are subject to early public health interventions. According to current TB control policies, clinically suspected and smear-positive cases are prioritized for contact investigation. Furthermore, when a smear-positive case is culture-confirmed as MTBC, all close contacts, regardless of age, are included in latent TB infection (LTBI) screening [30]. These early interventions, including contact tracing, isolation, and initiation of treatment, may effectively reduce the transmission risk, thereby diminishing the observable impact of smear positivity in the statistical models.

Social behavioral factors might be more indicative than clinical characteristics when predicting tuberculosis (TB) transmission. Researchers conducted a study in Lima, Peru, revealing that younger TB patients were more likely to act as sources of transmission due to their greater social contact across diverse age groups. Additionally, TB transmission was more frequent among groups where both the infector and infectee shared smoking or alcohol consumption habits. This highlighted the important role of such social aggregation behaviors in facilitating the spread of tuberculosis [31]. Our study supported these findings. Lower average age among early cases was associated with a higher likelihood of forming large transmission clusters. However, our study might not fully capture the transmission dynamics among younger adults. In Kaohsiung City, 64% of



newly diagnosed TB cases in 2023 were individuals aged 65 and above, and the study population predominantly consisted of middle-aged and elderly adults. This age distribution may limit the generalizability of our findings to younger populations. Older adults are more likely to progress to active TB and transmit the disease due to age-related immune decline and the presence of comorbidities. Common comorbidities in middle-aged and elderly patients may confuse TB symptoms and delay diagnosis. However, their frequent interactions with healthcare systems also provided opportunities for screening and early intervention [32][33]. Controlling tuberculosis in elderly populations is complex and challenging. Healthcare workers need to focus on high-risk social settings like hospitals and long-term care institutions. We also recommend promoting TB education and prevention measures among socially active younger groups. This dual approach could effectively reduce the risk of TB transmission.

4.2 Temporal Structure of Data

We constructed three datasets to compare incident clusters with datasets that did not exclude non-incident clusters and applied the same modeling approach to each. In comparing models among these datasets, we found similar median AUC values in the Logistic Regression and SVM models. This might be due to cluster-level variables that captured composite characteristics of transmission dynamics. The model could capture meaningful trends regardless of the observation time point. Moreover, it is difficult to determine whether a cluster is incident or not in real-world data. Some strains remained latent for long periods before activation, while others spread quickly. In addition, undiagnosed cases are inevitable, leading to missing observations in the dataset. Given these considerations, the models developed in this study appeared to offer generalizability across different temporal contexts and observational windows.

4.3 Underperformance of Random Forest models

This study found that, compared to Logistic Regression (LR) and Support Vector Machine (SVM) with a linear kernel, Random Forest (RF) demonstrated poor predictive performance. This could be attributed to the limited sample size or the feature patterns being more consistent with linear relationships. Random Forest tends to perform better with nonlinear, high-dimensional data and large sample sizes. Conversely, when the sample size is small and the relationship between predictors and outcomes is linear, linear models could outperform Random Forest. Also, complex machine learning models in such situations may lead to overfitting. Simple linear models, such as Logistic Regression, offers a simpler structure and better interpretability, making it a computationally efficient and scalable model suitable for public health applications [34].

4.4 Challenges in Measuring Transmission Event

In a study from China, the delay in the time from symptom onset to first healthcare visit was 15 days. The delay from healthcare visit to diagnosis was 2 days [35]. In contrast, a study conducted in England between 2012 and 2015 reported that approximately one-third of notified cases experienced a diagnostic delay exceeding four months, indicating regional variations in diagnostic delay [36]. This study arranged cases chronologically based on notification dates; however, these dates may not accurately reflect the actual timing of transmission. Therefore, the data used in our analysis may not represent the earliest transmission events.

In this study, we applied a threshold of 12 SNPs to determine genetic relatedness. However, the appropriateness of this threshold needs further validation. As shown in the previous study, even cases with SNP distance below the threshold did not always exhibit epidemiological links upon field investigation. If the threshold is loose, it may include

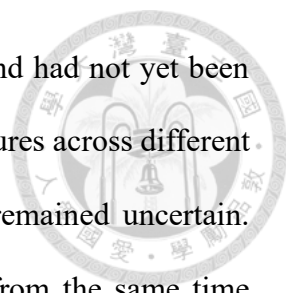
multiple unrelated cases. Genomic information can help, but it cannot fully establish transmission direction. We could only estimate potential relatedness among cases and classify them into the same cluster. Also, the issue of missing data introduced biases in the reconstruction of transmission chains. Although researchers have attempted to infer unsampled cases within clusters using bioinformatics approaches, these methods cannot entirely replace field investigations [37]. Therefore, active and systematic interventions are essential for accurately reconstructing transmission dynamics.

Furthermore, clusters identified via genomic analysis may not correspond with those defined by public health facilities. This discordance can result in missed opportunities for timely interventions. In our dataset, contact history information was frequently missing from epidemiological records. Such gaps posed additional challenges to effective outbreak management and control.

4.5 Study Limitation

This study assumed the availability of whole genome sequencing (WGS) results to support early prediction; nevertheless, current healthcare systems still lack rapid WGS workflows. In reality, bacterial culture and sequencing preparation often took weeks to months. Such an approach is still some distance away from practical implementation.

This study did not incorporate treatment initiation or duration into the analysis. Since TB patients generally experience a significant reduction in infectiousness after two weeks of appropriate therapy, treatment timing may play an important role in limiting cluster expansion. We did not consider the temporal sequence of risk factor exposure. For example, if only one of the first two notified cases is smear-positive, whether it is the first or the second case that tests smear-positive may have important epidemiological implications.



Finally, this study was limited to the Kaohsiung region in Taiwan and had not yet been validated in other areas. Given the variations in disease control measures across different counties in Taiwan, the model's generalizability to other regions remained uncertain. Furthermore, the model was developed and evaluated using data from the same time period, without testing its performance on temporally distinct datasets. The absence of validation data limited our ability to assess the model's robustness and long-term applicability. Future studies should incorporate independent datasets from other settings to strengthen the reliability of the prediction model.

4.6 Conclusion

We found that a model combining sputum smear positivity, mean age, and notification date interval has the potential to predict large clusters at an early stage. Future works could further analyze these risk factors to optimize predictive models and public health intervention strategies. Our findings also revealed that over half of the transmission clusters involved only two cases, indicating that most transmission events were interrupted early, highlighting the critical role of timely case isolation and fundamental epidemiological investigations in outbreak control.

In conclusion, continuous strengthening of basic prevention and control measures is essential. For clusters predicted as high-risk by the model, expanded investigations should be initiated promptly to facilitate real-time monitoring of transmission dynamics. Implementing a flexible and targeted response strategy can enhance the use of limited public health resources, accelerate outbreak response, and effectively curb disease spread.

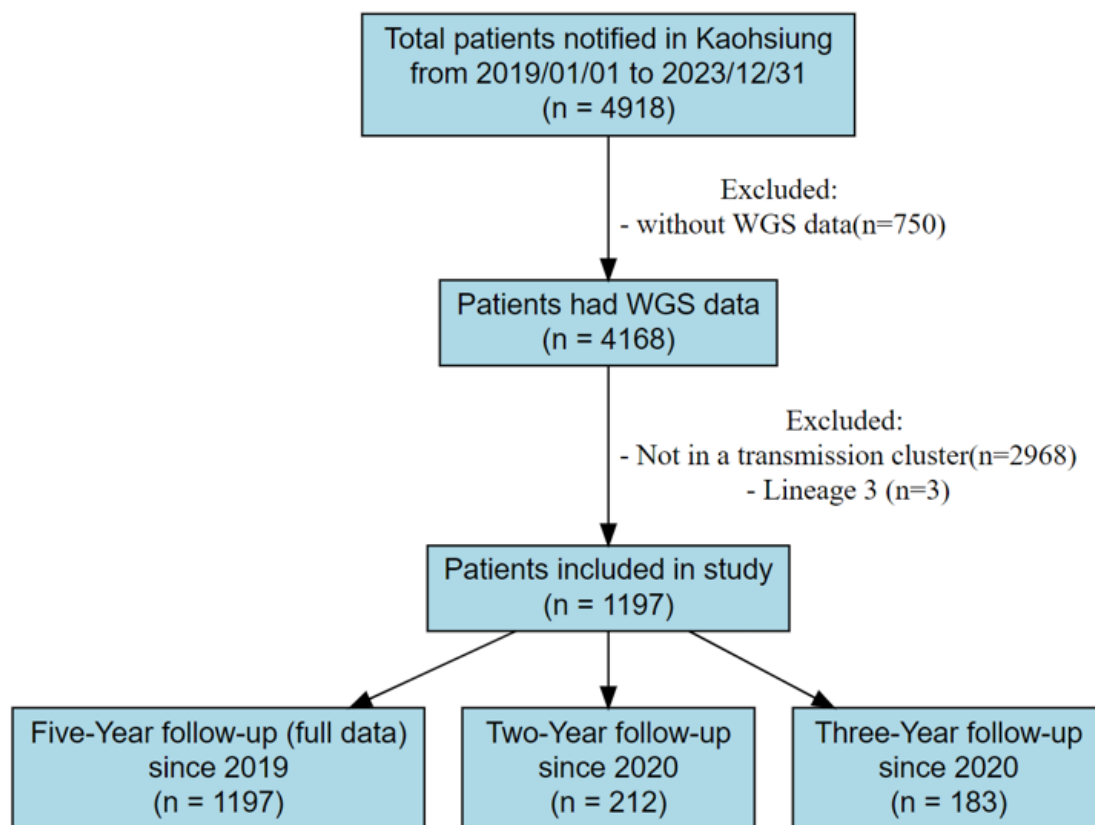


Figure 1. Study flow chart

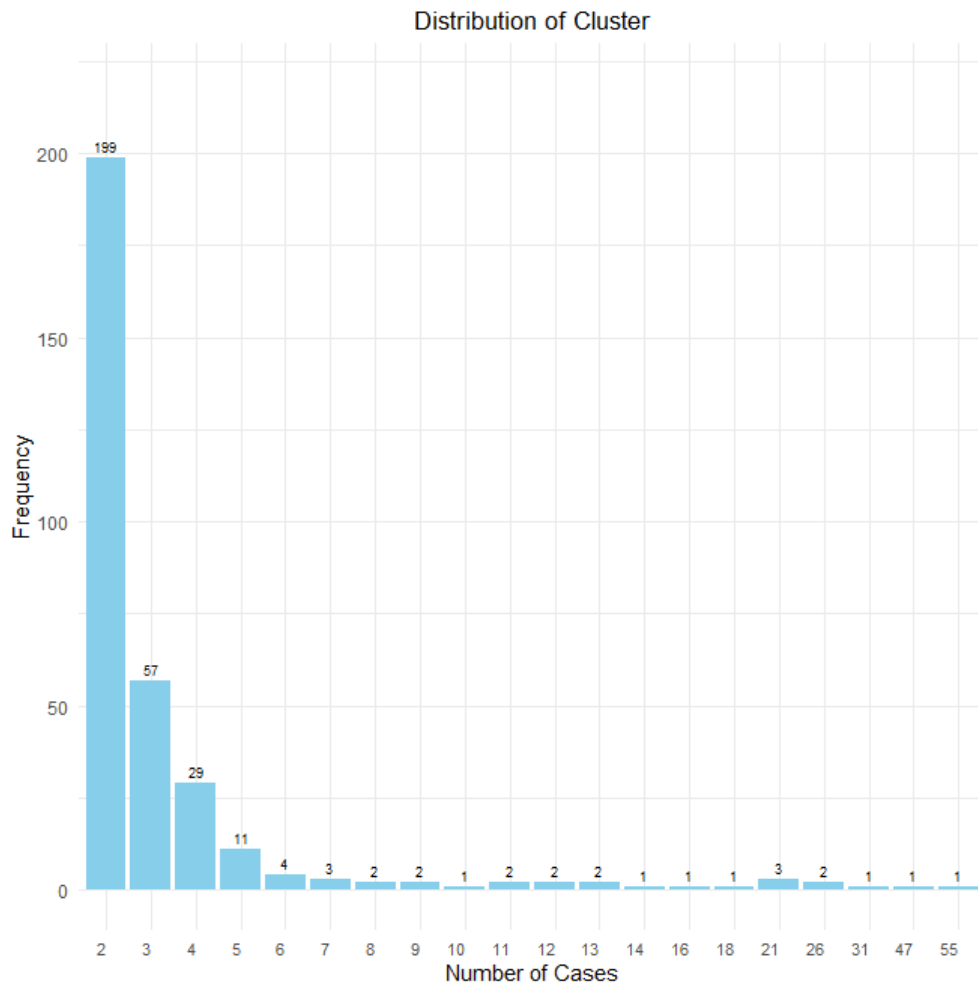


Figure 2. Distribution of transmission cluster

Clusters were defined based on a threshold of single nucleotide polymorphism (SNP) distance fewer than 12.

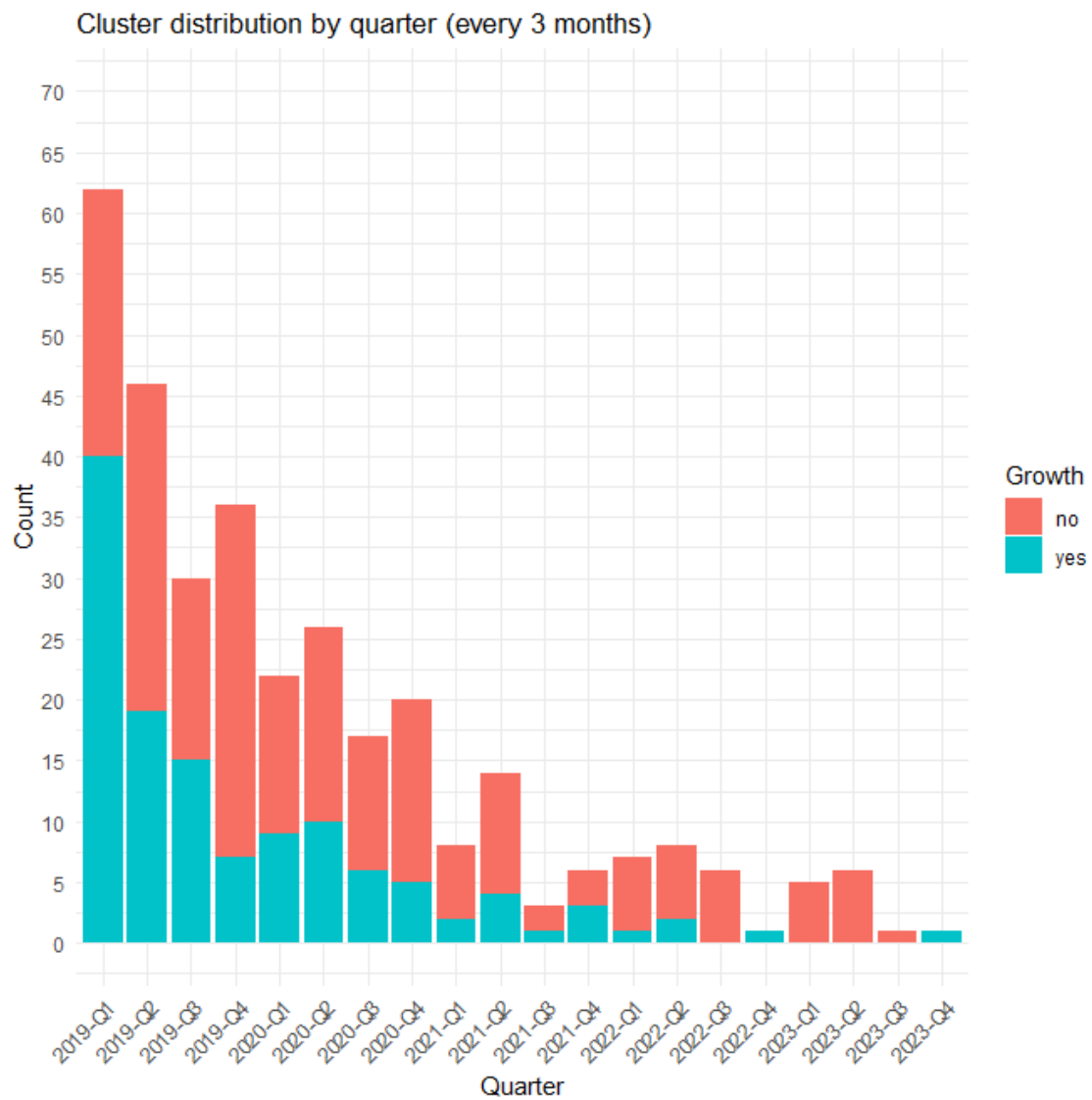


Figure 3. Distribution of transmission cluster onset by year and quarter

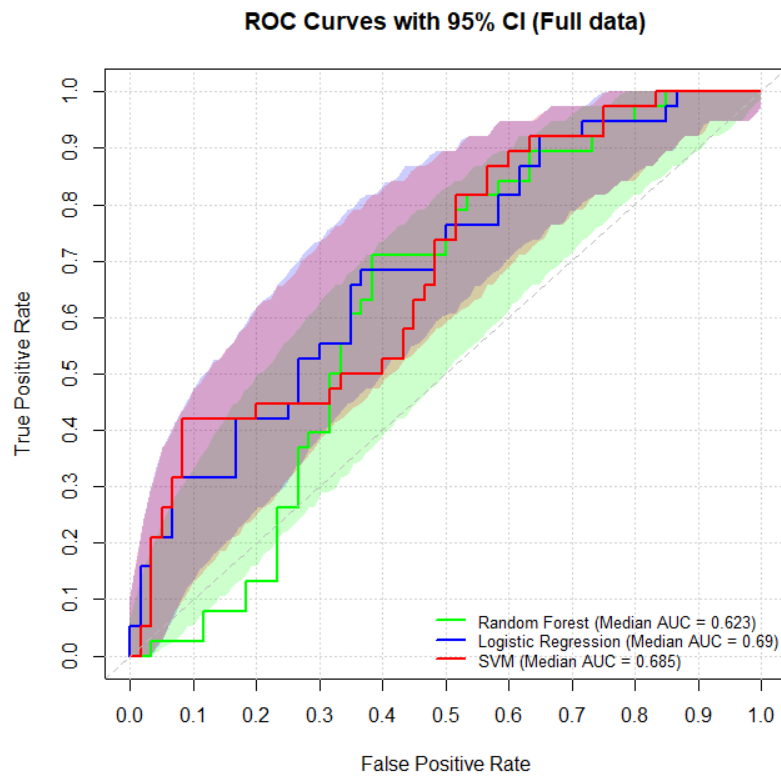


Figure 4A. Model performance of full-data

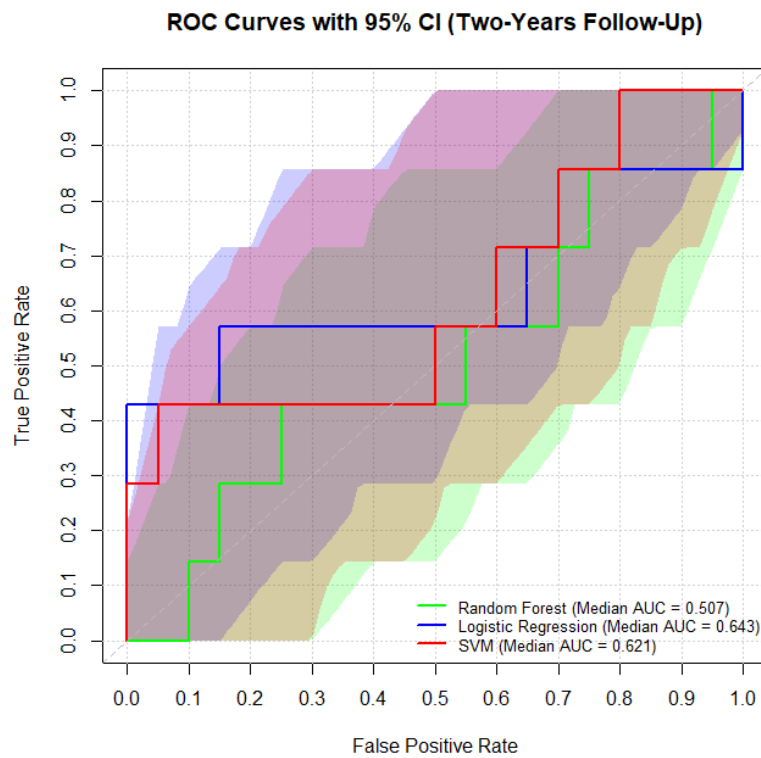


Figure 4B. Model performance of two-year follow-up data

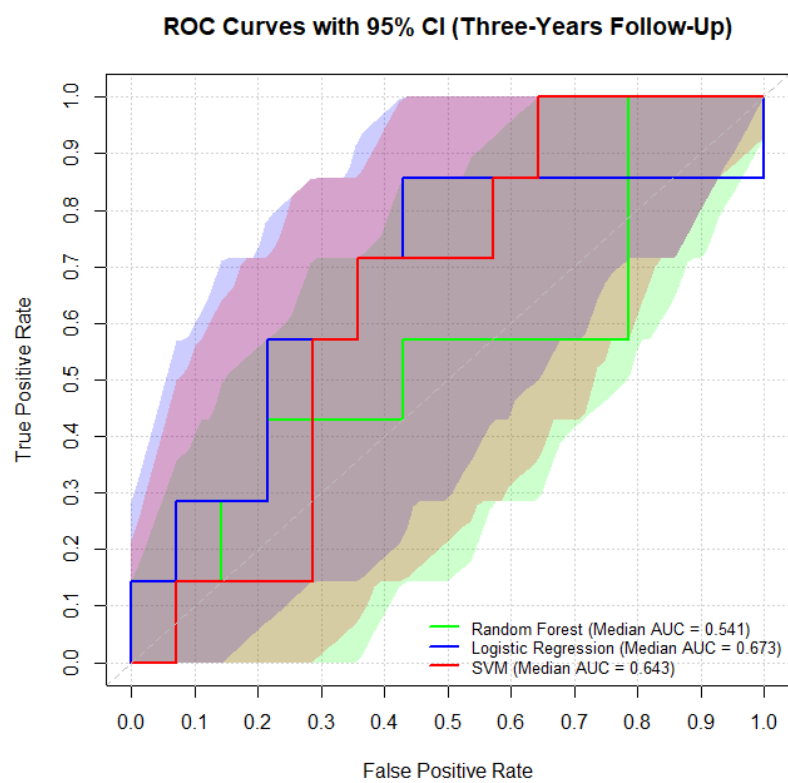


Figure 4C. Model performance of three-year follow-up data

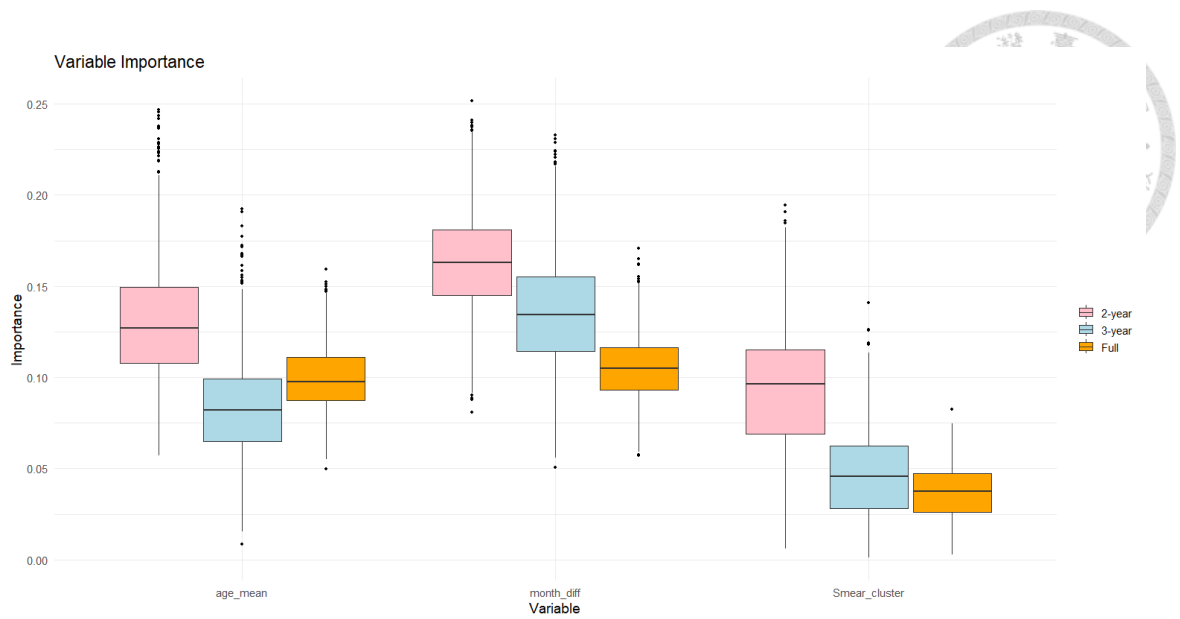


Figure 5. Variable importance

age_mean: mean age of the first two notified cases; month_diff: notification date interval, measured in months; Smear_cluster: categorical cluster-level variable representing the smear status of the first two cases.

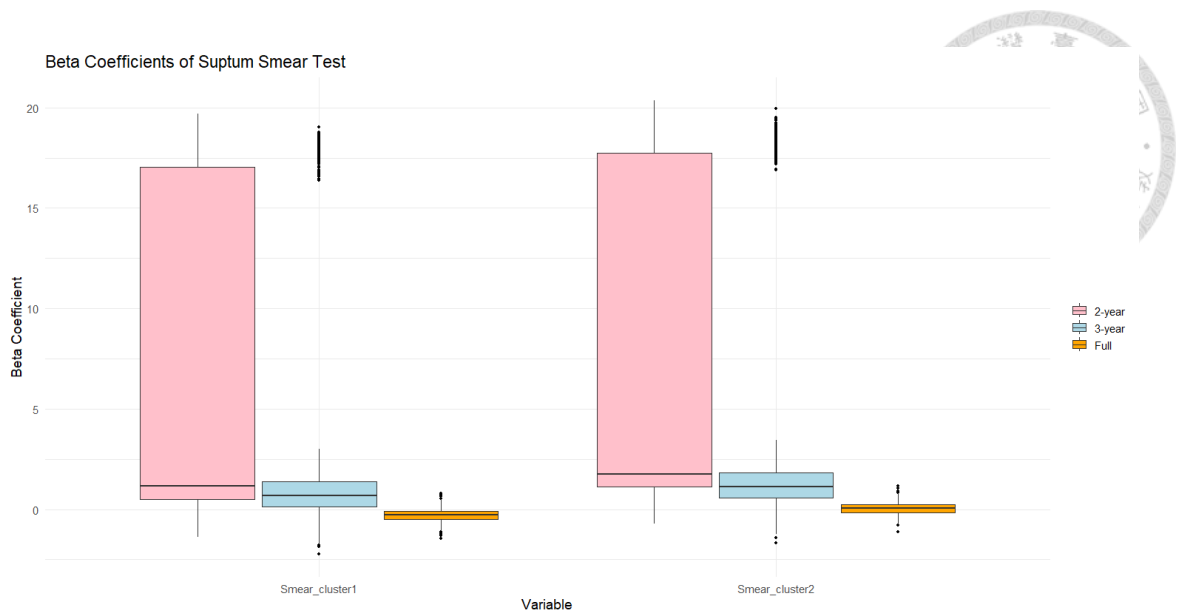


Figure 6A. Boxplot of beta coefficients for sputum smear status in logistic regression models

Smear_cluster1: only one individual was smear-positive; Smear_cluster2: both individuals were smear-positive; Reference group: neither individual was smear-positive.



Figure 6B. Boxplot of beta coefficients for mean age and notification date interval in logistic regression models

age_mean: mean age of the first two notified cases; month_diff: notification date interval.

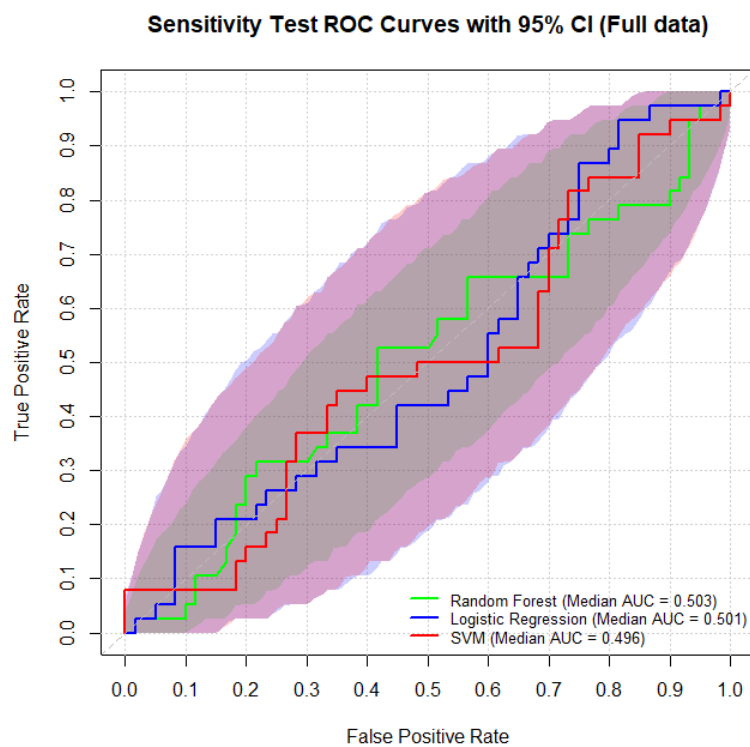


Figure 7A. Sensitivity Analysis of full-data

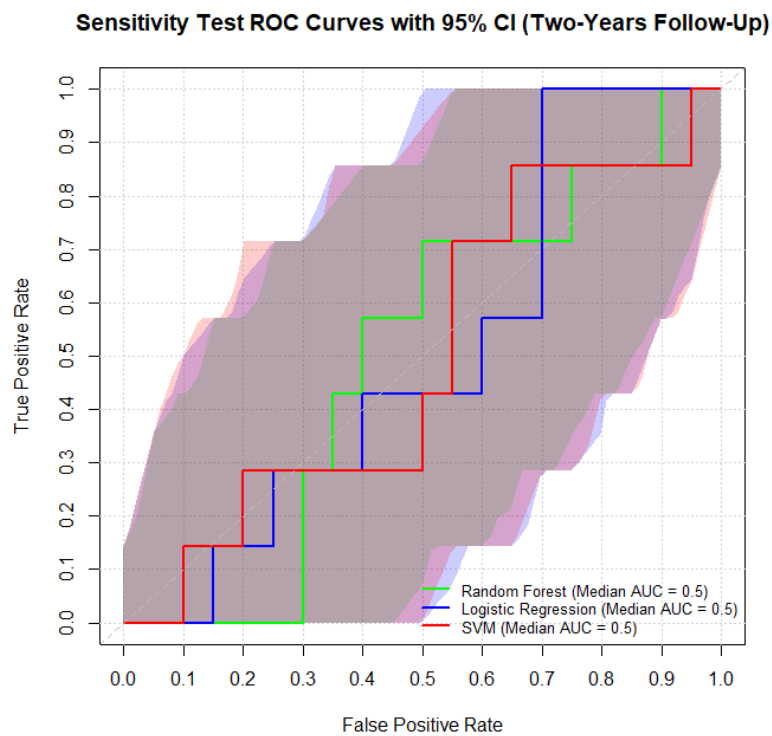


Figure 7B. Sensitivity Analysis of two-year follow-up data

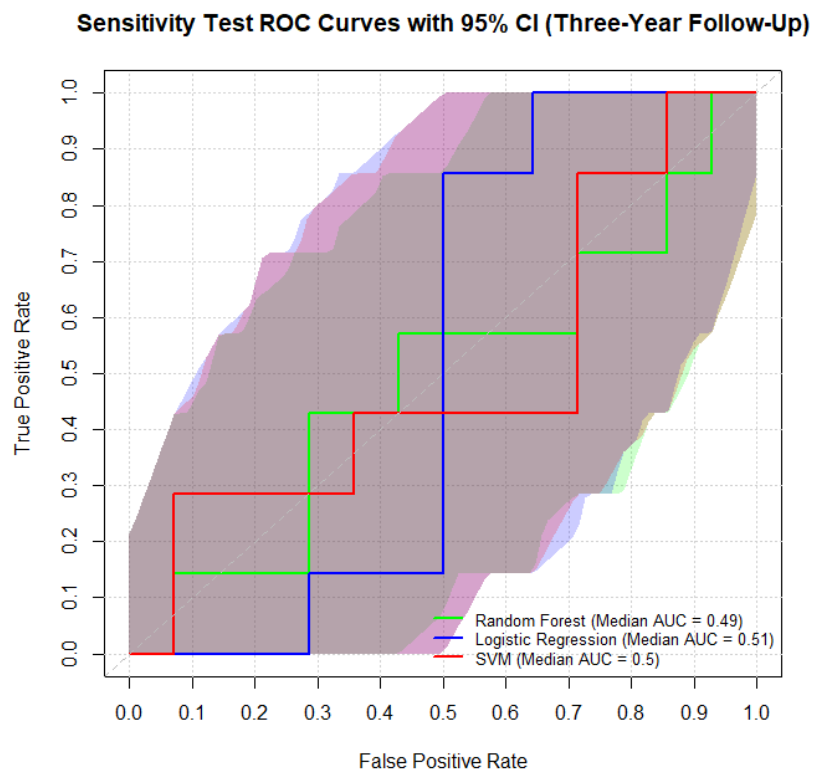


Figure 7C. Sensitivity Analysis of three-year follow-up data

Variables		Cluster size =2 (N=398)	Cluster size ≥3 (N=252)	p value
Sex				
Female		111 (27.89%)	68 (26.98%)	0.895537
Male		287 (72.11%)	183 (72.62%)	
NA		0	1 (0.4%)	
Contact history				
No		305 (76.63%)	170 (67.46%)	1
Yes		46 (11.56%)	26 (10.32%)	
NA		47 (11.81%)	56 (22.22%)	
Chronic Kidney Disease				
No		335 (84.17%)	181 (71.83%)	0.190812
Yes		16 (4.02%)	15 (5.95%)	
NA		47 (11.81%)	56 (22.22%)	
Hemoptysis				
No		333 (83.67%)	183 (72.62%)	0.591351
Yes		18 (4.52%)	13 (5.16%)	
NA		47 (11.81%)	56 (22.22%)	
Cough for two weeks				
No		179 (44.97%)	111 (44.05%)	0.239201
Yes		172 (43.22%)	85 (33.73%)	
NA		47 (11.81%)	56 (22.22%)	
Smoking				
Never		158 (39.7%)	78 (30.95%)	0.259701
Quit		79 (19.85%)	45 (17.86%)	
Smoke		85 (21.36%)	60 (23.81%)	
NA		76 (19.1%)	69 (27.38%)	
Drinking				
Never		193 (48.49%)	94 (37.3%)	0.224435
Quit		61 (15.33%)	39 (15.48%)	
Drink		69 (17.34%)	48 (19.05%)	
NA		75 (18.84%)	71 (28.17%)	
Diabetes mellitus				
No		267 (67.09%)	143 (56.75%)	0.482762
Yes		84 (21.11%)	53 (21.03%)	
NA		47 (11.81%)	56 (22.22%)	
Drug resistant				
Not resistant		331 (83.17%)	214 (84.92%)	0.631366
Resistant to at least one drug		62 (15.58%)	35 (13.89%)	
NA		5 (1.26%)	3 (1.19%)	
Sputum smear test				
Negative		158 (39.7%)	96 (38.1%)	0.643916
Positive		230 (57.79%)	153 (60.71%)	
NA		10 (2.51%)	3 (1.19%)	

Table 1A. Basic characteristic of full-data (individual-level variables)

Variables	Cluster size =2 (N=199)	Cluster size ≥3 (N=126)	p value
Lineage			
Lineage1	76 (38.19%)	41 (32.54%)	0.381534
Lineage2	87 (43.72%)	55 (43.65%)	
Lineage4	36 (18.09%)	30 (23.81%)	
Distance (km)			
Mean (SD)	8.16 (9.13)	9.09 (9.59)	0.385917
Age difference (year)			
Mean (SD)	15.61 (13.44)	17.54 (14.97)	0.240808
Mean age (year)			
Mean (SD)	64.35 (12.73)	58.56 (12.83)	< .001
Notification date interval (month)			
Mean (SD)	18.31 (15.57)	10.30 (11.58)	< .001

Table 1B. Basic characteristic of full-data (cluster-level variables)

Variables	Cluster size =2 (N=132)	Cluster size ≥3 (N=48)	p value
Sex			
Female	35 (26.52%)	13 (27.08%)	1
Male	97 (73.48%)	35 (72.92%)	
Contact history			
No	105 (79.55%)	38 (79.17%)	0.521054
Yes	22 (16.67%)	5 (10.42%)	
NA	5 (3.79%)	5 (10.42%)	
Chronic Kidney Disease			
No	119 (90.15%)	38 (79.17%)	0.318001
Yes	8 (6.06%)	5 (10.42%)	
NA	5 (3.79%)	5 (10.42%)	
Hemoptysis			
No	121 (91.67%)	38 (79.17%)	0.148076
Yes	6 (4.55%)	5 (10.42%)	
NA	5 (3.79%)	5 (10.42%)	
Cough for two weeks			
No	70 (53.03%)	21 (43.75%)	0.591353
Yes	57 (43.18%)	22 (45.83%)	
NA	5 (3.79%)	5 (10.42%)	
Smoking			
Never	56 (42.42%)	21 (43.75%)	0.886843
Quit	30 (22.73%)	9 (18.75%)	
Smoke	35 (26.52%)	12 (25%)	
NA	11 (8.33%)	6 (12.5%)	
Drinking			
Never	68 (51.52%)	22 (45.83%)	0.47341
Quit	22 (16.67%)	5 (10.42%)	
Drink	31 (23.48%)	14 (29.17%)	
NA	11 (8.33%)	7 (14.58%)	
Diabetes mellitus			
No	93 (70.45%)	31 (64.58%)	1
Yes	34 (25.76%)	12 (25%)	
NA	5 (3.79%)	5 (10.42%)	
Drug resistant			
Not resistant	108 (81.82%)	37 (77.08%)	0.730755
Resistant to at least one drug	23 (17.42%)	10 (20.83%)	
NA	1 (0.76%)	1 (2.08%)	
Sputum smear test			
Negative	53 (40.15%)	12 (25%)	0.086348
Positive	76 (57.58%)	35 (72.92%)	
NA	3 (2.27%)	1 (2.08%)	

Table 2A. Basic characteristic of two-year follow-up data (individual-level variables)

Variables	Cluster size =2 (N=66)	Cluster size ≥3 (N=24)	p value
Lineage			
Lineage1	21 (31.82%)	8 (33.33%)	0.975755
Lineage2	27 (40.91%)	10 (41.67%)	
Lineage4	18 (27.27%)	6 (25%)	
Distance (km)			
Mean (SD)	7.74 (6.63)	10.43 (11.40)	0.285254
Age difference (year)			
Mean (SD)	14.24 (12.35)	16.62 (13.38)	0.450531
Mean age (year)			
Mean (SD)	64.55 (14.03)	62.06 (10.42)	0.368681
Notification date interval (month)			
Mean (SD)	10.86 (7.09)	7.01 (6.33)	0.017449

Table 2B. Basic characteristic of two-year follow-up data (cluster-level variables)

Variables		Cluster size =2 (N=96)	Cluster size ≥3 (N=48)	p value
Sex				
	Female	24 (25%)	13 (27.08%)	0.946241
	Male	72 (75%)	35 (72.92%)	
Contact history				
	No	77 (80.21%)	39 (81.25%)	0.409735
	Yes	15 (15.62%)	4 (8.33%)	
	NA	4 (4.17%)	5 (10.42%)	
Chronic Kidney Disease				
	No	87 (90.62%)	40 (83.33%)	0.709289
	Yes	5 (5.21%)	3 (6.25%)	
	NA	4 (4.17%)	5 (10.42%)	
Hemoptysis				
	No	88 (91.67%)	39 (81.25%)	0.265211
	Yes	4 (4.17%)	4 (8.33%)	
	NA	4 (4.17%)	5 (10.42%)	
Cough for two weeks				
	No	52 (54.17%)	25 (52.08%)	1
	Yes	40 (41.67%)	18 (37.5%)	
	NA	4 (4.17%)	5 (10.42%)	
Smoking				
	Never	34 (35.42%)	20 (41.67%)	0.644906
	Quit	26 (27.08%)	12 (25%)	
	Smoke	26 (27.08%)	10 (20.83%)	
	NA	10 (10.42%)	6 (12.5%)	
Drinking				
	Never	44 (45.83%)	22 (45.83%)	0.816122
	Quit	16 (16.67%)	9 (18.75%)	
	Drink	25 (26.04%)	10 (20.83%)	
	NA	11 (11.46%)	7 (14.58%)	
Diabetes mellitus				
	No	65 (67.71%)	33 (68.75%)	0.594546
	Yes	27 (28.12%)	10 (20.83%)	
	NA	4 (4.17%)	5 (10.42%)	
Drug resistant				
	Not resistant	80 (83.33%)	36 (75%)	0.382383
	Resistant to at least one drug	15 (15.62%)	11 (22.92%)	
	NA	1 (1.04%)	1 (2.08%)	
Sputum smear test				
	Negative	40 (41.67%)	14 (29.17%)	0.198354
	Positive	54 (56.25%)	33 (68.75%)	
	NA	2 (2.08%)	1 (2.08%)	

Table 3A. Basic characteristic of three-year follow-up data (individual-level variables)

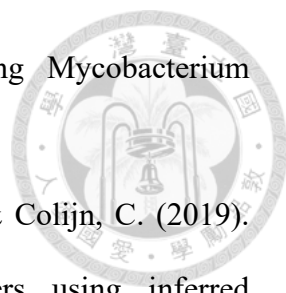
Variables	Cluster size =2 (N=48)	Cluster size ≥3 (N=24)	p value
Lineage			
Lineage1	14 (29.17%)	7 (29.17%)	0.979248
Lineage2	21 (43.75%)	11 (45.83%)	
Lineage4	13 (27.08%)	6 (25%)	
Distance (km)			
Mean (SD)	7.68 (6.39)	9.34 (11.16)	0.505989
Age difference (year)			
Mean (SD)	15.29 (12.75)	15.71 (11.95)	0.892100
Mean age (year)			
Mean (SD)	65.33 (12.86)	61.23 (13.35)	0.219740
Notification date interval (month)			
Mean (SD)	15.68 (11.10)	8.27 (6.26)	< .001

Table 3B. Basic characteristic of three-year follow-up data (cluster-level variables)

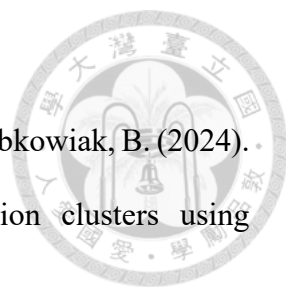
Reference

- 
- [1] Tobin, E. H., & Tristram, D. (2024, December 22). *Tuberculosis overview*. StatPearls - NCBI Bookshelf. <https://www.ncbi.nlm.nih.gov/books/NBK441916/>
- [2] Global tuberculosis report 2024. Geneva: World Health Organization; 2024. Licence: CC BY-NC-SA 3.0 IGO.
- [3] 台灣結核病防治年報 2023，衛生福利部疾病管制署（2023）
- [4] Global Programme on Tuberculosis and Lung Health (GTB). (2015, July 29). Systematic screening for active tuberculosis: an operational guide.
- [5] Thierry, D., Brisson-Noël, A., Vincent-Lévy-Frébault, V., Nguyen, S., Guesdon, J. L., & Gicquel, B. (1990). Characterization of a *Mycobacterium tuberculosis* insertion sequence, IS6110, and its application in diagnosis. *Journal of clinical microbiology*, 28(12), 2668-2673.
- [6] Kamerbeek, J., Schouls, L. E. O., Kolk, A., Van Agterveld, M., Van Soolingen, D., Kuijper, S., ... & van Embden, J. (1997). Simultaneous detection and strain differentiation of *Mycobacterium tuberculosis* for diagnosis and epidemiology. *Journal of clinical microbiology*, 35(4), 907-914.
- [7] Vergnaud, G., & Pourcel, C. (2009). Multiple locus variable number of tandem repeats analysis. In *Molecular epidemiology of microorganisms: methods and protocols* (pp. 141-158). Totowa, NJ: Humana Press.
- [8] 吳俊忠, & 王美嘉. (2020). 醫學分子檢驗 = Textbook of molecular diagnosis in medicine / 吳俊忠[及其他 3 位]主編; 王美嘉[及其他 54 位]著 (第六版). 五南出版.
- [9] Van der Roest, B. R., Bootsma, M. C., Fischer, E. A., Gröschel, M. I., Anthony, R. M., de Zwaan, R., ... & Klinkenberg, D. (2025). Phylodynamic assessment of SNP

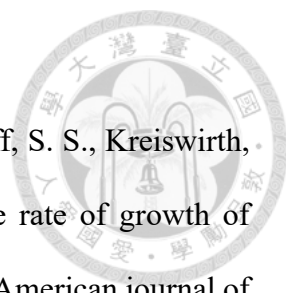
distances from whole genome sequencing for determining *Mycobacterium tuberculosis* transmission. *Scientific Reports*, 15(1), 10694.

- 
- [10] Stimson, J., Gardy, J., Mathema, B., Crudu, V., Cohen, T., & Colijn, C. (2019). Beyond the SNP threshold: identifying outbreak clusters using inferred transmissions. *Molecular biology and evolution*, 36(3), 587-603.
- [11] Walker, T. M., Ip, C. L., Harrell, R. H., Evans, J. T., Kapatai, G., Dedicoat, M. J., ... & Peto, T. E. (2013). Whole-genome sequencing to delineate *Mycobacterium tuberculosis* outbreaks: a retrospective observational study. *The Lancet infectious diseases*, 13(2), 137-146.
- [12] Xiao, Y. X., Chan, T. H., Liu, K. H., & Jou, R. (2024). Define SNP thresholds for delineation of tuberculosis transmissions using whole-genome sequencing. *Microbiology Spectrum*, 12(8), e00418-24.
- [13] Althomsons, S. P., Kammerer, J. S., Shang, N., & Navin, T. R. (2012). Using routinely reported tuberculosis genotyping and surveillance data to predict tuberculosis outbreaks. *PLoS One*, 7(11), e48754.
- [14] Wortham, J. M., Li, R., Althomsons, S. P., Kammerer, S., Haddad, M. B., & Powell, K. M. (2021). Tuberculosis genotype clusters and transmission in the US, 2009–2018. *American journal of preventive medicine*, 61(2), 201-208.
- [15] Althomsons, S. P., Winglee, K., Heilig, C. M., Talarico, S., Silk, B., Wortham, J., ... & Navin, T. R. (2022). Using machine learning techniques and national tuberculosis surveillance data to predict excess growth in genotyped tuberculosis clusters. *American journal of epidemiology*, 191(11), 1936-1943.
- [16] Kik, S. V., Verver, S., van Soolingen, D., de Haas, P. E., Cobelens, F. G., Kremer, K., ... & Borgdorff, M. W. (2008). Tuberculosis outbreaks predicted by characteristics of first patients in a DNA fingerprint cluster. *American journal of*

respiratory and critical care medicine, 178(1), 96-104.

- 
- [17] Gharamaleki, O. G., Colijn, C., Sekirov, I., Johnston, J. C., & Sobkowiak, B. (2024). Early prediction of Mycobacterium tuberculosis transmission clusters using supervised learning models. *Scientific Reports*, 14(1), 27652.
- [18] 台灣結核病防治年報 2019，衛生福利部疾病管制署（2019）
- [19] 台灣結核病防治年報 2020，衛生福利部疾病管制署（2020）
- [20] 台灣結核病防治年報 2021，衛生福利部疾病管制署（2021）
- [21] 台灣結核病防治年報 2022，衛生福利部疾病管制署（2022）
- [22] Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R news*, 2(3), 18-22.
- [23] Ho, T. K. (1995, August). Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition* (Vol. 1, pp. 278-282). IEEE.
- [24] Wright, M. N., & Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of statistical software*, 77, 1-17.
- [25] Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of statistical software*, 28, 1-26.
- [26] Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.
- [27] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- [28] Long, R., Divangahi, M., & Schwartzman, K. (2022). Chapter 2: Transmission and pathogenesis of tuberculosis. *Canadian Journal of Respiratory, Critical Care, and*

Sleep Medicine, 6(sup1), 22-32.

- 
- [29] Driver, C. R., Macaraig, M., McElroy, P. D., Clark, C., Munsiff, S. S., Kreiswirth, B., ... & Zhao, B. (2006). Which patients' factors predict the rate of growth of *Mycobacterium tuberculosis* clusters in an urban community?. *American journal of epidemiology*, 164(1), 21-31.
- [30] 結核病防治工作手冊（第四版），衛生福利部疾病管制署（2023）
- [31] Trevisi, L., Brooks, M. B., Becerra, M. C., Calderón, R. I., Contreras, C. C., Galea, J. T., ... & Huang, C. C. (2024). Who transmits tuberculosis to whom: a cross-sectional analysis of a cohort study in Lima, Peru. *American Journal of Respiratory and Critical Care Medicine*, 210(2), 222-233.
- [32] Seto, J., Wada, T., Suzuki, Y., Ikeda, T., Mizuta, K., Yamamoto, T., & Ahiko, T. (2017). *Mycobacterium tuberculosis* transmission among elderly persons, Yamagata Prefecture, Japan, 2009–2015. *Emerging Infectious Diseases*, 23(3), 448.
- [33] Caraux-Paz, P., Diamantis, S., de Wazières, B., & Gallien, S. (2021). Tuberculosis in the elderly. *Journal of Clinical Medicine*, 10(24), 5888.
- [34] Arshad, M. A., Shahriar, S., & Anjum, K. (2023). The Power Of Simplicity: Why Simple Linear Models Outperform Complex Machine Learning Techniques--Case Of Breast Cancer Diagnosis. *arXiv preprint arXiv:2306.02449*.
- [35] Luo, D., Chen, X., Wang, M., Zhang, M., Li, Y., Chen, S., ... & Chen, B. (2024). Analyzing spatial delays of tuberculosis from surveillance and awareness surveys in Eastern China. *Scientific Reports*, 14(1), 19799.
- [36] Loutet, M. G., Sinclair, C., Whitehead, N., Cosgrove, C., Lalor, M. K., & Thomas, H. L. (2018). Delay from symptom onset to treatment start among tuberculosis patients in England, 2012–2015. *Epidemiology & Infection*, 146(12), 1511-1518.
- [37] Didelot, X., Fraser, C., Gardy, J., & Colijn, C. (2017). Genomic infectious disease

epidemiology in partially sampled and ongoing outbreaks. *Molecular biology and evolution*, 34(4), 997-1007.

