國立臺灣大學公共衛生學院環境與職業健康科學研究所

碩士論文

Institute of Environmental and Occupational Health Sciences

College of Public Health

National Taiwan University

Master Thesis

以機器學習方法改善新生兒異位性皮膚炎風險之預測

Applying machine learning methods to improve risk prediction of atopic dermatitis in newborns

林義哲

Yi-Che Lin

指導教授：郭育良教授

Advisor: Yue-Liang Leon Guo, M.D., M.P.H., PH.D.

中華民國111年8月

August 2022

doi:10.6342/NTU202202317

國立臺灣大學碩士學位論文
口試委員會審定書

論文中文題目：

以機器學習方法改善新生兒異位性皮膚炎風險之預測

論文英文題目：

Applying machine learning methods to improve risk prediction of atopic dermatitis in newborns

本論文係林義哲君（學號 R08852041）在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 111 年 7 月 21 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：郭育良（簽名）

（指導教授）

doi:10.6342/NTU202202317

國立臺灣大學碩士學位論文
口試委員會審定書

論文中文題目：

以機器學習方法改善新生兒異位性皮膚炎風險之預測

論文英文題目：

Applying machine learning methods to improve risk prediction of atopic dermatitis in newborns

本論文係林義哲君（學號 R08852041）在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 111 年 7 月 21 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）

（指導教授）

doi:10.6342/NTU202202317

國立臺灣大學碩士學位論文
口試委員會審定書

論文中文題目：

以機器學習方法改善新生兒異位性皮膚炎風險之預測

論文英文題目：

Applying machine learning methods to improve risk prediction of atopic dermatitis in newborns

本論文係林義哲君（學號 R08852041）在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 111 年 7 月 21 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）

（指導教授）

蘇大成

doi:10.6342/NTU202202317

國立臺灣大學（碩）博士學位論文
口試委員會審定書

論文中文題目：

以機器學習方法改善新生兒異位性皮膚炎風險之預測

論文英文題目：

Applying machine learning methods to improve risk prediction of atopic dermatitis in newborns

本論文係林義哲君（學號R08852041）在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國111年7月21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）

（指導教授）

_____ _____

_____ _____

_____ _____

_____ 徐證男

doi:10.6342/NTU202202317

國立臺灣大學（碩）博士學位論文
口試委員會審定書

論文中文題目：

以機器學習方法改善新生兒異位性皮膚炎風險之預測

論文英文題目：

Applying machine learning methods to improve risk prediction of atopic dermatitis in newborns

本論文係林義哲君（學號 R08852041）在國立臺灣大學環境與職業健康科學研究所完成之碩士學位論文，於民國 111 年 7 月 21 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）

（指導教授）

張紘宣

doi:10.6342/NTU202202317

誌謝

在修讀研究所碩士學位的道路上，我遇到許多的貴人，在這些貴人的鼎力協助下，我的研究工作及本論文才得以順利完成，在此我要向他們表達我最深切的感謝。

首先我要感謝我的指導教授郭育良老師，感謝老師給我寶貴的機會加入這項跨領域、跨團隊的研究，讓我得以一窺機器學習在臨床醫學應用上的潛力與挑戰，非常感謝老師在學習的路上的鞭策與鼓勵，老師深具啟發性的建議、高標準的要求以及對於細節的重視，是我在學術路上的典範，未來若我有幸能有些許成就，實在要歸功於老師的栽培與教誨。

接著我要感謝三位資訊工程的教授賴飛羆老師、徐讚昇老師、張紘睿老師，三位機器學習領域的專家給予本研究許多建議與指導，尤其在研究方法最核心的部分，不斷探索機器學習方法的各種可能性並且設計新的演算法，讓本研究得以有突破性的研究結果，沒有幾位老師的指導，本研究不可能完成，在研究過程中，我深刻學習到老師們對於研究的熱忱、執行力與創意，令我欽佩且感謝。

再來我要感謝林欣儀及呂欣儒兩位研究夥伴，在研究過程中提供寶貴的技術支援，設計 Python 程式碼，雖然我是資訊工程的門外漢且程式語言的程度非常淺，欣儀與欣儒仍花時間用 Colab 分享程式碼，並且給我許多指導、解答我的問題，讓我有充分的機會練習與學習程式語言，對此我非常感激，祝福兩位未來能在專長的領域發光發熱。

我也要感謝同為郭老師指導學生、也可說是兩位最重要的戰友黃宇丞與黃尹岑，感謝宇丞花很多時間和我討論研究設計與方法，在機器學習的基本知識、應用、R 及 Python 程式碼上也給我許多指導，讓我快速入門並到後來有能力自學，祝福宇丞畢業後的職涯順利；感謝尹岑在兩年碩士班期間的陪伴及各種即時的提醒，讓因為工作繁忙而經常忘東忘西的我，能夠準時完成碩士班各項進度，在我的學業及研究遭遇挫折時，給我支援及鼓勵，祝福尹岑在碩士畢業後的生涯一切順遂。

doi:10.6342/NTU202202317

我還要感謝郭老師研究團隊的所有研究夥伴，在每週四的專題研究討論會上，給予本研究諸多指教與建議，讓我能持續反思尚可改善之處，在討論的過程中教學相長，也讓英文能力進步。另外，我也要感謝環職所所辦的助教，在繁忙的日常中不厭其煩的提醒我各項重要日程，在課程安排與行政事務上也給予我許多協助，少了他們的悉心叮嚀，此研究絕無可能準時完成。

最後，我要感謝我的家人，在就讀碩士班的過程中支持並鼓勵我，讓我可以無後顧之憂的在學術的道路上前行，我非常感謝他們。

doi:10.6342/NTU202202317

中文摘要

研究背景：異位性皮膚炎是最常見的皮膚疾病，許多患者在幼兒時期即已罹病，2009 年一篇研究使用邏輯斯迴歸，建立新生兒 6 個月大時罹患異位性皮膚炎的風險預測模型，近年來機器學習演算法快速發展，廣泛應用於臨床醫學，有潛力改善 2009 年研究建立的異位性皮膚炎預測模型。

研究目的：使用機器學習演算法，以先天及後天的風險因子建立 6 個月大新生兒異位性皮膚炎的風險預測模型，並與邏輯斯回歸模型比較。

研究方法：研究使用與 2009 年研究相同的資料集，即「台灣世代研究」資料庫，此資料庫抽樣收集台灣 88 個鄉鎮於 2005 年出生的新生兒資料，於新生兒 6 個月大時進行首次調查。本研究先移除遺漏值，並以性別將資料分開，再以 80%：20% 的比例將原資料集切成訓練集與測試集。在機器學習模型部分，預測變數使用 19 個特徵，首先依照臨床上合理的切點將特徵離散化，並新定義倆性別的風險組，分為極低、低、高、極高四組，接著對訓練集進行 100 次「隨機特徵集選取、風險重標籤」以創造出新訓練集，使用新訓練集訓練 XGBoost 模型，並使用測試集以「5 組驗證」的方式驗證模型，透過窮舉搜索的方式調整參數，找出預測各風險組的最佳模型，再定義「二模型混合預測」與「三模型混合預測」規則，採用三模型混合預測作為機器學習模型預測結果；在邏輯斯迴歸模型部分，使用與 2009 年相同的 8 個特徵訓練邏輯斯迴歸模型，並使用測試集以「5 組驗證」的方式驗證模型。兩模型最終以混淆矩陣呈現，以對角線和、均方根誤差、加權誤差等作為模型表現的指標。

研究結果：本研究最終使用的資料集包含 20235 名新生兒（9607 名女性，占 47%），女性異位性皮膚炎比例約 6%，男性異位性皮膚炎比例約 8%，女性機器學習三模型混合預測準確率為：低風險組 0.953、高風險組 0.753、極高風險組 0.706，混淆矩陣對角線和 2.412，均方根誤差 0.533，加權誤差 0.302，女性邏輯斯迴歸模

doi:10.6342/NTU202202317

型預測準確率為：低風險組 0.958、高風險組 0.734、極高風險組 0.644，混淆矩陣對角線和 2.337，均方根誤差 0.580，加權誤差 0.370。男性機器學習三模型混合預測準確率為：低風險組 0.963、高風險組 0.811、極高風險組 0.816，混淆矩陣對角線和 2.590，均方根誤差 0.394，加權誤差 0.175，男性邏輯斯迴歸模型預測準確率為：低風險組 0.936、高風險組 0.772、極高風險組 0.821，混淆矩陣對角線和 2.529，均方根誤差 0.412，加權誤差 0.227。

結論：本研究將機器學習方法應用於一個具全國代表性的出生世代資料集，為 6 個月大的新生兒建立異位性皮膚炎風險預測模型，研究顯示機器學習模型比過去的邏輯斯迴歸模型表現更佳，有效提高預測準確率，可以協助臨床醫師預測新生兒罹患異位性皮膚炎的風險並採取預防措施。

關鍵字：異位性皮膚炎、新生兒、風險預測模型、機器學習、XGBoost

doi:10.6342/NTU202202317

Abstract

Background: Atopic dermatitis (AD) is the most common skin disorder and many patients develop symptoms early. A risk prediction model of AD in 6-month-old newborns was established in 2009 using logistic regression (LR). Recently, machine learning (ML) methods keep gaining popularity and have been applied in various clinical settings. Whether ML can outperform LR remains inconclusive.

Objective: To apply ML methods to set up AD risk prediction model among 6-month-old newborns based on hereditary and environmental risk factors, and to compare performance between ML model and LR model.

Methods and Participants: Taiwan Birth Cohort Study (TBCS) was used in this study, same as the study in 2009. Babies born in 2005 in 88 townships in Taiwan were sampled and the first follow-up interview took place when the babies were 6 months old. Data with missing values were removed. The data were stratified based on gender and were split to a train set and a test set in 80-20 ratio. Nineteen features (risk factor) were included in the ML model. Feature discretization, 100 rounds of random feature set selection and AD risk level relabeling were performed sequentially to create a new train set. The ML model was trained on the new train set and was validated by 5-run validation on the test set. Through exhaustive grid search of parameters, the best model of each risk level was identified. We assigned prediction rules of 2-model and 3-model mixed

doi:10.6342/NTU202202317

prediction. The 3-model mixed prediction was the final ML model. The LR model was set up using the same 8 features as the study in 2009 and was validated by 5-run validation on the test set. Standardized confusion matrix was used to summarize the final prediction results of two models. Sum of diagonals, RMSE and weighted error were calculated to compare performance between ML and LR.

Results: A total of 20235 newborns (9607 female [47%]) were analyzed. The AD percentage was about 6% in female and about 8% in male. The prediction accuracy of ML model of female was 0.953, 0.753, and 0.706 in low, high and very high risk group, respectively and the sum of diagonals, RMSE and weighted error were 2.412, 0.533 and 0.302, respectively. The prediction accuracy of LR model of female was 0.958, 0.734 and 0.644 in low, high and very high risk group, respectively and the sum of diagonals, RMSE and weighted error were 2.337, 0.580 and 0.370, respectively. The prediction accuracy of ML model of male was 0.963, 0.811, and 0.816 in low, high and very high risk group, respectively and the sum of diagonals, RMSE and weighted error were 2.590, 0.394 and 0.175, respectively. The prediction accuracy of LR model of male was 0.936, 0.772 and 0.821 in low, high and very high risk group, respectively and the sum of diagonals, RMSE and weighted error were 2.529, 0.412 and 0.227, respectively. Overall, compared to the LR model, the ML model of female had 3.2% higher sum of diagonals, 8.1% lower RMSE and 18.4% lower weighted error. Compared to the LR model, the ML model of male had

doi:10.6342/NTU202202317

2.4% higher sum of diagonals, 4.4% lower RMSE and 23% lower weighted error.

Conclusions: In this study, a novel ML approach combining with XGBoost was applied on a national representative birth cohort to set up AD risk prediction models in 6-month-old newborns. For both genders, the ML model had better overall performance than the LR model. Our ML model can help clinicians stratify newborns into different risk levels with high accuracy and help clinicians design preventive strategies based on the risk.

Keywords: atopic dermatitis, newborn, risk prediction model, machine learning, XGBoost

doi:10.6342/NTU202202317

目錄

doi:10.6342/NTU202202317

doi:10.6342/NTU202202317

圖表目錄

doi:10.6342/NTU202202317

Chapter 1. Introduction

1.1 Background and motivation

Atopic dermatitis (AD) is one of the most common skin disorders in both the developed and developing world. AD is notorious for its early onset, relapsing nature, severely itchy skin lesions and accompanying complications, which lead to substantial medical costs and economic burden for the whole society. Although various hereditary and environmental risk factors for AD have been identified, there has not been an optimal AD prediction model for newborns. As the old saying goes, prevention is better than cure. We can put the old saying into practice through a risk prediction model based on risk factors. The model could help clinicians and parents design individualized prevention strategies for those newborns at risk.

Clinical prediction models usually aim to predict binary outcomes such as survival status or disease status, and thus, logistic regression (LR) has commonly been used to set up many models. Machine learning (ML) offers an alternative choice to accomplish the task and has been prosperously developed lately. Various new algorithms were proposed to solve real-world problems, to make prediction and to assist decision-making, not only in the commercial world but also in clinical scenarios.

Previously, a study applied logistic regression (LR) model to set up an AD prediction model for 6-month-old newborns using Taiwan Birth Cohort Study (TBCS)

doi:10.6342/NTU202202317

[1]. The model was a typical regression model with preselection of important variables from the dataset based on experts' knowledge. Despite its easy interpretability, its area under receiver-operating characteristics curve (AUC) is suboptimal, indicating that its discrimination ability is not satisfactory. There is still room for improvement. ML, solving classification problem through algorithms different from LR, has the potential to improve AD risk prediction.

1.2 Research aims

The study aims to apply ML method to set up model which could separate newborns with AD from those healthy individuals at the age of 6 months, based on hereditary and environmental risk factors. We examined whether ML could improve prediction performance of the LR model proposed by the previous study.

doi:10.6342/NTU202202317

Chapter 2. Literature Review

2.1 Epidemiology and impacts of atopic dermatitis

Atopic dermatitis (AD) is a common inflammatory skin disease, characterized by relapsing eczematous rash [2]. AD typically occurs early in the childhood [3, 4]. One study showed that among those children who were diagnosed with AD, nearly half had already developed symptoms within the first six months of their lives [5]. AD prevalence in children has increased for several decades around the globe [6]. The lifetime prevalence in children under 8 years of age was 20%-40% in both the developed and developing countries [7]. The figure in Taiwan showed a similar rising trend. The lifetime prevalence of AD in 6 to 15-year-olds rose from less than 2% before 1995, to 3.35% in 2002, and reached 6.7% during the period of 2000-2007 [8, 9].

AD has a substantial impact on many aspects of life. At individual level, AD does harm to the physical and psychological health of the patients and further impair their interpersonal relationship and social functioning. As children are the most susceptible to AD, the quality of life of their caregivers is also inevitably affected [10, 11]. From a social point of view, annual costs per patient were approximately USD 1000 to 6000 in Asia-Pacific countries, USD 2000 in the United States, Cdn$1200 in Canada [12, 13]. In Taiwan, the costs are more than USD 1000 per patient per year [14]. Therefore, development of preventive strategies could be critical to save the society from

doi:10.6342/NTU202202317

considerable economic burden. To address this issue, a risk prediction model is the first step.

2.2 Risk factors for atopic dermatitis

AD is a disease with multiple etiologies, in which biological, psychological and socioeconomic factors all play a role. The exact pathophysiological mechanism of AD has not been fully understood. There is a complex interaction between potential allergens in the surroundings and impaired host immune system. Various risk factors for atopic dermatitis have been investigated. Recently, a systematic review and meta-analysis summarized epidemiological studies examining risk factors for AD in Asia [15]. The study classified risk factors into modifiable and non-modifiable. Modifiable factors are generally easier to be changed or improved, such as lifestyle, environmental exposures and medications. Non-modifiable factors refer to factors which are typically inherited or unchangeable by nature, including demographic data, medical condition and socioeconomic status of parents. Although practically, preventive strategies can only target modifiable factors, the prediction model should include both modifiable and non-modifiable factors since both could affect development of AD.

2.3 Previously established AD prediction model using LR

Wen and colleagues applied logistic regression (LR) with backward stepwise

doi:10.6342/NTU202202317

variable selection on TBCS, which was a national representative birth cohort consisting of more than 20,000 newborns, to set up a risk prediction model for AD in 6-month-old newborns [1]. In their study, Wen and colleagues examined the association between AD and various hereditary and environmental risk factors. Family history of atopy, higher maternal education, presence of fungus on walls, exposure to indoor painting and renovation during pregnancy were identified to be significantly associated with increased risk of AD and were included in their final LR model. Based on these risk factors, individualized AD risk can be calculated by the LR model. A 5-run validation test was performed to validate the model and the validation results confirmed the prediction ability of the model. However, misestimate of AD risk occurred when the LR model made prediction for high-risk female babies. Thus, there is room for improvement.

2.4 Machine learning

Lately, machine-learning (ML) methods continued gaining attention and have been applied in various fields of research. such as 6-month mortality of patients with malignancy[16], risk of hypoxemia during general anesthesia [17] and long-term prognosis of ischemic stroke patients [18]. Risk prediction models provide valuable assistance in clinical decision-making and have already become a crucial part in modern clinical medicine. Traditionally, regression models make a great contribution in deriving

doi:10.6342/NTU202202317

many of those models. Logistic regression is a common choice for the task since many of the models aim to predict binary outcomes. [19, 20].

ML algorithms analyze data in a different way than traditional regression models do. Regression models are based on statistics theories. To make regression models robust enough, certain assumptions are required and certain rules should be met. For example, independence of errors, low multicollearity between variables and lack of outliers in data should be checked, and sample size should be much greater than number of variables [21]. Human intervention based on subject knowledge is heavily needed in the variable selection process, in which researchers have to decide which variables should be removed from or be kept in the model so as to prevent multicollearity problem. Furthermore, nonlinear relationship could sometimes exist between an independent variable and the dependent variable and interaction between some independent variables could also exist. It is time-consuming and difficult for researchers to test all potential nonlinear relationship and interaction between variables. On the other hand, ML models solve classification task based on theories different from regression models. ML models can learn directly from data without specifying too many rules beforehand [22]. It is argued that ML methods could outperform traditional regression due to its capability of dealing with much greater amounts of explanatory variables and variable intercorrelation by automatically testing all possible variable combinations through

doi:10.6342/NTU202202317

method of exhaustion. However, a recent systematic review concluded that whether ML models definitely have better performance still remains a subject of debate [19].

XGBoost, one of the widely used ML algorithms, is a tree-based ensemble learning method known for its high predictive accuracy, learning efficiency and capability of analyzing nonlinear relations. With its algorithmic innovation, it has offered solutions to some of the challenging problems in the ML competition and it has gained enormous popularity lately [23]. XGBoost has been shown effective in predicting clinical outcomes in various medical fields [24-26]. The objectives of this study are to apply ML approach and XGBoost algorithm on the same cohort of TBCS to establish a prediction model for AD risk prediction in 6-month-old newborns and to compare the performance of the XGBoost model and the LR model.

doi:10.6342/NTU202202317

Chapter 3. Materials and Methods

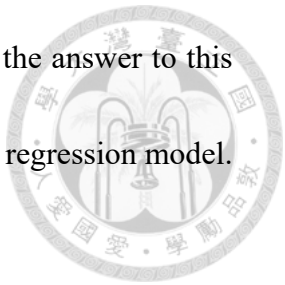
3.1 Study cohort recruitment

The study cohort was derived from the Taiwan Birth Cohort Study (TBCS). The details of the TBCS recruitment were described elsewhere [27]. Initiated in the year 2003 by the Taiwan Health Promotion Administration, TBCS aimed to examine early development of children and risk factors potentially affecting children health. The babies who were born between 1st Jan 2005 and 31st Dec 2005 were eligible for enrollment. The TBCS employed stratified systematic random sampling method to recruit the cohort. First, 369 townships of Taiwan were divided into 12 strata based on urbanization levels and total fertility rate. Eighty-eight townships were randomly selected from the 12 strata. From those townships, eligible population were sampled by the method of probability proportional to size. In the year 2005, there were 206,741 live births in total in Taiwan. In the end, the TBCS recruited 24,200 mother-infant pairs, which accounted for approximately one-eighth of the newborns in that year.

The first home interview was conducted by trained interviewers when the babies were 6 months old. Serial follow-up interviews took place when the babies were 18 months, 3 years and 5.5 years old. The mothers were the prioritized interview targets. If a baby's mother was unavailable for the interview, the main caregiver would be the alternative. "Whether the baby had ever been diagnosed with AD by a physician" was

doi:10.6342/NTU202202317

one of the questions in the interview. In the study of Wen *et al.* [1], the answer to this question, a binary variable, was the dependent variable of the logistic regression model.



3.2 Data preprocessing

The data with missing values were removed. We did not impute missing value. Due to difference of disease prevalence between both genders, we aimed to set up model for girls and boys separately. Therefore, the data were stratified by gender at the beginning of the analysis. All the following steps were performed in both genders separately.

We randomly split the data into a train set and a test set in 80:20 ratio. Next, we prepared the features for the analysis. In total, there were 19 features, including demographics, medical history of parents and baby and environmental exposures. Feature discretization was carried out. The categorical features were kept categorical and the continuous features were transformed into categorical based on clinically appropriate cut-off values. Table 1 showed the features included in the final analysis.

3.3 Defining a new variable, “AD risk level”

The dependent variable of the logistic regression model by Wen et al. [1] was a binary variable and recorded whether a 6-month-old newborn had ever been diagnosed with AD. In the current study, we defined a new dependent variable with four levels, the “AD risk level”, to replace the original dependent variable. Our prediction target shifted

doi:10.6342/NTU202202317

from binary disease outcomes to multiple risk levels. The new dependent variable had four risk levels, including very low, low, high and very high. The three cut-off values separating the four levels were derived from the AD percentage of the original data, including half of the AD percentage, the AD percentage and twice of the AD percentage. In the original dataset, the AD percentage of boys and girls were 8.07% and 5.79%, respectively. These two numbers were rounded to the nearest whole number, and the cut-off values were 4%, 8% and 16% for the boys, 3%, 6% and 12% for the girls.

3.4 Random feature set selection and AD risk level relabeling

The complete steps of our novel ML approach are shown in Figure 1. We began our ML analysis with random feature set selection and AD risk level relabeling for the study participants in the train set before training the prediction models. We randomly selected some features out of the 19 features to form a random feature set. The process began with an empty feature set. One feature was randomly selected and added in the feature set. Next, the study participants who had the same value combination of the selected feature(s) were grouped together. For example, if maternal history of asthma, paternal history of AD, breastfeeding duration and molds on walls at home were selected, all participants whose mother had asthma, whose father had AD, who had never been breastfed and whose home was free from molds would be put into one group. Similarly,

doi:10.6342/NTU202202317

all participants whose mother had asthma, whose father had AD, who had never been breastfed but whose home suffered from moldy walls would be put into another group.

We set criteria to keep at least a certain percentage of groups in optimal size to prevent potential bias due to small group size. The group size was kept above X for at least Y% of the boys, and above Z for at least W% of the girls. After a new feature was selected and added in the feature set, we examined the criteria. If the criteria were met, the newly selected feature stayed in the feature set. If the criteria were not met, the newly selected feature was removed from the feature set. The feature was selected one at a time until the criteria cannot be kept. The parameter X, Y, Z and W were adjusted through an exhaustive grid search process in a later step to identify the best model of each risk level. The algorithm and criteria of random feature set selection was shown in Figure 2.

After all participants were grouped based on value combination of selected features, actual AD percentage of each group was calculated and compared to the pre-defined cut-off values of the "AD risk levels". Based on the comparison, the new dependent variable of each study participant was imputed with one of the four risk levels. For example, if actual AD percentage of a boy group was 11%, the new dependent variable, "AD risk level", of all boys of this group would be imputed with "high risk" because 11% was between the cut-off values, 8% and 16%. Through this method, the study participants of the train set were relabeled with an AD risk level.

doi:10.6342/NTU202202317

The random feature set selection and AD risk level relabeling were repeated 100 rounds on the train set. The results of every round were all kept. After 100 rounds, each study participant in the train set was turned from one single data into 100 data and the size of the train set became 100 times larger. The size of the test set remained unchanged.

3.5 XGBoost model training and 5-run validation

After data preparation, we applied XGBoost algorithm to train the prediction model based on the newly relabeled train set. The prediction target was the multi-level dependent variable, AD risk level. The analysis was conducted using Python and the XGBoost Python package version 1.1.1. Three important hyperparameters of XGBoost algorithm, max_depth, n_estimators and max_delta_step, were set equal to 6, 3 and 5, respectively. The rest of the hyperparameters were kept as default.

We performed 5-run validation to test the trained XGBoost model on the test set. The test set was randomly divided into 5 partitions. The trained XGBoost model was applied on each partition to predict AD risk level for the participants. In each partition, the participants were classified into 4 predicted risk groups based on the predicted results. The actual AD percentage in each predicted risk group was calculated and compared with the predicted risk level. We performed the 5-run validation for 10,000 rounds. Each round resulted in 4 predicted risk groups in the 5 partitions, so 20 comparison results were yielded through each round. The comparison results between predicted risk and

doi:10.6342/NTU202202317

actual risk of the 10,000 rounds of 5-run validation were summed up in a confusion matrix, as shown in Table 2. A standardized confusion matrix can be derived by having each grid divided by the sum of all grids of a confusion matrix. Take Table 2 as an example, a standardized confusion matrix can be derived by having every X_{ij} divided by the sum of all X_{ij}.

Wen and colleagues used 3 risk levels to validate the performance of the LR model, which were < 8%, 8%-16%, and ≥ 16% for the boys and < 6%, 6%-12%, ≥ 12% for the girls, respectively [1]. The risk level < 8% for the boys and < 6% for the girls corresponded to the very low risk level and the low risk level of our XGBoost model. The two higher risk levels of the LR model corresponded to the high risk level and the very high risk level of our XGBoost model. To allow comparison between XGBoost and LR model, we combined the predicted results of very low risk level and the low risk level into one risk level, the low risk level. Consequently, only three risk levels remained in the standardized confusion matrix, including low, high and very high risk levels.

3.6 Mixed XGBoost prediction

To improve prediction performance of XGBoost, we propose a mixed XGBoost prediction method by combining the best XGBoost model of each risk levels. We repeatedly performed 100 rounds of random feature selection and AD risk relabeling, XGBoost model training and 10,000 rounds of 5-run validation with exhaustive grid

doi:10.6342/NTU202202317

search of all possible combination of X, Y, Z and W. The best model of each risk level was identified based on the results of 5-run validation. As shown in Table 2, the XGBoost model with the highest accuracy in the grid X_{11} was defined as the best model of low risk level and was named Model 1. The model with the highest accuracy in the grid X_{22} was defined as the best model of high risk and was named Model 2. The model with the highest accuracy in the grid X_{33} was defined as the best model of very high risk level and was named Model 3.

Next, we defined the prediction rules for the 2-model mixed XGBoost prediction. Table 3 summarized the prediction rules of the 3 possible model pairs consisting of Model 1 and Model 2, Model 1 and Model 3, Model 2 and Model 3. Following the rules, a participant's risk level separately predicted by two models was integrated into one mixed predicted outcome. Finally, the 3-model mixed XGBoost prediction was performed on the test set through majority voting among the predicted outcomes of the three 2-model mixed prediction. The final results of 3-model mixed prediction were demonstrated in a standardized confusion matrix.

3.7 Logistic regression (LR) model training, 5-run validation and model comparison

We applied LR on the preprocessed but unenlarged train set to train the LR model. The features included here were the same as those used to set up the LR model of Wen

doi:10.6342/NTU202202317

and colleagues' work [1], including parental history of AD and allergic rhinitis, fungus on walls, painting furniture during pregnancy, home renovation during pregnancy and maternal education. The trained LR model was tested on the test set through 10,000 rounds of 5-run validation, same as XGBoost model. The ranges of risk levels of the 5-run validation here were < 8%, 8%-16%, and ≥ 16% for the boys and > 6%, 6%-12%, ≥ 12% for the girls, respectively, which were the same as the aforementioned risk levels of XGBoost model and the same as those in Wen and colleagues' work. The prediction of the LR model was demonstrated in a standardized confusion matrix.

To compare the prediction performance between the 3-model mixed XGBoost prediction and the LR model, we calculated the sum of diagonals, root mean squared error (RMSE) and weighted error based on the results summarized in the confusion matrix. Take Table 2 as an example. The sum of diagonals and the RMSE were calculated as below:

$$\text{Sum of diagonals} = \frac{\sum x_{ij,}\ i=j}{\sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij}}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{3}\sum_{j=1}^{3}[x_{ij}\times(i-j)^2]}{\sum_{i=1}^{3}\sum_{j=1}^{3} x_{ij}}}$$

Besides the sum of diagonals and the RMSE, we calculated weighted error to further compare the models. As shown in Table 4, we assigned the costs for different

doi:10.6342/NTU202202317

scenarios of inaccurate predictions. The cost of an accurate prediction was assigned to be 0, and the costs of all possible inaccurate predictions were assigned as Table 4 demonstrated. Overall, underestimation results in higher costs than overestimation of risk. The cost is 10 when newborns at very high risk of AD are predicted to be at low risk. The costs are 3 when newborns at high risk and very high risk are predicted to be at low risk and high risk, respectively. The cost is 2 when newborns at low risk are overestimated to be at very high risk. The costs are 1 when newborns with low risk and high risk are overestimated to be at high and very high risk, respectively. Take Table 2 as an example. The costs in Table 4 served as the weight as the weighted error was calculated as below:

$$\text{Weighted error} = \frac{\sum_{i=1}^{3}\sum_{j=1}^{3}(x_{ij}\times w_{ij})}{\sum_{i=1}^{3}\sum_{j=1}^{3}x_{ij}}$$

3.8 Statistical analysis

All the above analytical steps were performed by Python. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated to evaluate univariate association between the features and AD. The significance level was 0.05.

doi:10.6342/NTU202202317

Chapter 4. Results

4.1 Study cohort recruitment

The steps of study cohort recruitment and data preprocessing were shown in Figure 3. In total, 24,200 mother-infant pairs were recruited in the TBCS. Among them, 2,952 lost follow-ups due to refusal of being interviewed, wrong addresses, moving abroad, absence after multiple interview visits, death of the baby and other reasons. Multiparity samples were also excluded. Among the remaining 20,687 subjects, 452 (2.18%) had missing values in some features and were excluded. In total, we included 20,235 participants in the final analysis.

4.2 Characteristics of study population

The demographic characteristics of the female and male newborns were shown in Table 5 and Table 6. The data included in the final analysis consisted of 9,607 girls and 10,628 boys. Among them, 557 girls and 858 boys were diagnosed with AD. The AD prevalence was 5.80% in female babies and 8.07% in male babies. Overall, the demographic characteristics were relatively similar between both genders. Approximately one-third of the girls and the boys were delivered by cesarean section, and about half of the babies were the first child of the mothers. In average, the breastfeeding duration was 2.4 months in girls and 2.1 months in boys. At childbirth, the

doi:10.6342/NTU202202317

mean maternal age was 30 years and the mean paternal age was 33 years among both genders. Parents with longer than 12 years of education accounted for 45% of the participants. About 1.5% of the mothers of both genders and 1.2% of the fathers of both genders had history of asthma. Nearly 36%, 2.6% and 4% of the mother-infant pairs were exposed to wall molds, indoor painting and indoor renovation in their residences during pregnancy, respectively.

The unadjusted ORs of risk factors for AD were also shown in Table 5 and Table 6. In both genders, higher parental educational level, parental history of AD and allergic rhinitis, maternal history of asthma, longer duration of breastfeeding, older maternal age at childbirth, molds on walls and indoor renovation during pregnancy were associated with higher risk of AD, while older maternal menarche age was associated with lower risk of AD. Longer gestational length and higher birth weight were positively associated with AD risk in girls. Indoor painting during pregnancy was positively associated with AD risk in boys.

4.3 Model comparison

The results of the XGBoost model (3-model mixed prediction) and the LR model on the test sets of both genders were shown in Table 7. For the girls, the accuracy of XGBoost model in the low risk, high risk and very high risk level was 0.953, 0.753 and 0.706, respectively, and the accuracy of LR model was 0.958, 0.734 and 0.644,

doi:10.6342/NTU202202317

respectively. Compared to the LR model, the XGBoost model has higher sum of diagonals, lower RMSE and lower weighted error. The sum of diagonals of the XGBoost model and the LR model were 2.412 and 2.337, respectively. The RMSE of the XGBoost model and the LR model were 0.533 and 0.580, respectively. The weighted error of the XGBoost model and the LR model were 0.302 and 0.370, respectively.

For the boys, the accuracy of XGBoost model in the low risk, high risk and very high risk level was 0.963, 0.811 and 0.816, respectively, and the accuracy of LR model was 0.936, 0.772 and 0.821, respectively. Compared to the LR model, the XGBoost model has higher sum of diagonals, lower RMSE and lower weighted error. The sum of diagonals of the XGBoost model and the LR model were 2.590 and 2.529, respectively. The RMSE of the XGBoost model and the LR model were 0.394 and 0.412, respectively. The weighted error of the XGBoost model and the LR model were 0.175 and 0.277, respectively.

doi:10.6342/NTU202202317

Chapter 5. Discussion

In this study, machine learning method was applied on a national representative birth cohort to set up AD risk prediction models for 6-month-old newborns. For both genders, the XGBoost model had better overall performance than the LR model in terms of prediction accuracy, sum of diagonals, RMSE and weighted error. The XGBoost model had higher accuracy than the LR model in predicting girls at high and very high risk of AD, while the accuracy of both models in predicting girls at low risk of AD was similar. For the girls, the XGBoost resulted in 3.2% increase of sum of diagonals, 8.1% decrease of RMSE and 18.4% decrease of weighted error compared to the LR model. The XGBoost model had higher accuracy than the LR model in predicting boys at low and high risk of AD, while the accuracy of both models in predicting boys at very high risk of AD was comparable. For the boys, the XGBoost resulted in 2.4% increase of sum of diagonals, 4.4% decrease of RMSE and 23% decrease of weighted error compared to the LR model. A recent systematic review argued that whether ML methods outperform regression remained inconclusive [19]. Our results demonstrated that XGBoost could have better performance than LR.

The improvement achieved by the XGBoost model can be attributed to the inclusion of 11 additional features. The LR model in our study and in Wen and colleagues' study [1] was set up only based on 8 risk factors, including parental history

doi:10.6342/NTU202202317

of AD and allergic rhinitis, maternal education, fungus on walls, painting furniture and renovation in the house during pregnancy, while the XGBoost model were established based on 19 risk factors. The 11 extra features included birth order, gestational age, birth weight, delivery mode, parental age at childbirth, parental history of asthma, maternal menarche age, breastfeeding duration, paternal education.

The association of these 11 extra features and risk of AD has been examined in previous studies. Lower parity and higher gestational age were shown associated with higher risk of AD [28, 29]. In a meta-analysis, very preterm babies were associated with decreased risk for AD, and moderate preterm babies were at comparable risk for AD compared to full term babies [30]. Newborns with higher birth weight were at higher risk for AD and low birth weight could be a protective factor [31-33]. Compared to vaginal delivery, Cesarean section did not increase risk of developing AD in newborns [34-36]. One study using a national representative survey showed that younger paternal age was a risk factor for AD in children. The researchers argued that there could be confounding factors affecting this association such as richer economic status or smaller family size [37]. Another study reported unadjusted association between AD and maternal age at childbirth [38]. A recent systematic review and meta-analysis demonstrated that parental history of asthma was associated with increased risk of AD in children [39]. One study analyzing a Finnish birth cohort found that children whose

doi:10.6342/NTU202202317

mothers' first menstrual cycle began earlier than the age of 15 were at lower risk of AD [40]. However, another study using a birth cohort of the UK did not find a significant association between AD risk in children and maternal age of menarche. Thus, further investigation of this association is warranted. A study using a Japanese nationwide birth cohort showed that breastfeeding duration was positively associated with risk of AD in children up to the age of 3 years. Newborns who were breastfed exclusively were at significantly higher risk of AD compared to those who were fed with formula milk exclusively [38]. However, a recent meta-analysis demonstrated that breastfeeding acted as a protective factor against AD in children with parental history of atopy and breastfeeding could be a risk factor for AD in children without parental history of atopy[41]. High parental education levels were correlated with increased risk of AD [2]. In Wen and colleagues' study, parental education levels were significant risk factors for AD [1]. However, in that study, the final LR model which adopted backward variable selection strategy only included maternal education level. Paternal education level was left out possibly due to high collinearity with maternal education level. In our XGBoost model, parental education level was kept.

Another reason the XGBoost model could improve prediction of AD in newborns was the methodological innovation. Our novel ML approach comprised the feature discretization, random feature set selection, risk level relabeling, XGBoost model

doi:10.6342/NTU202202317

training, 5-run validation, identification of best model for each risk level through exhaustive grid search, 2-model mixed prediction and 3-model mixed prediction. The identification of best model for each risk level through exhaustive grid search of optimal combination of X, Y, Z and W was a key step. It allowed the 2-model mixed prediction and the subsequent 3-model mixed prediction to be carried out. Human intervention played a critical role in the rule assignment of the multiple-model mixed prediction. The rules of the 2-model mixed prediction were defined based on clinically reasonable rationale. The weight of error which was used to calculate weighted error was also assigned based on clinical experience.

There are several limitations in this study. First, the test set we used to validate our models was derived from the TBCS data. We did not use an independent external dataset due to lack of such data. To enhance the robustness of the validation, we performed 5-run validation for 10,000 times. Second, the AD status of the study participants was self-reported. Whether or not the newborns were diagnosed with AD by a physician was not validated by medical records or other databases. Recall bias could occur. Last, the TBCS did not collect biological samples such as urine, blood or skin tissue. Therefore, such data was not available for model establish. Although the XGBoost model was set up solely based on epidemiological and socioeconomic risk factors, it achieved good prediction performance with high prediction accuracy, low RMSE and low weighted

doi:10.6342/NTU202202317

error.

5.1 Strengths

Our novel ML approach has several strengths and can assist in clinical scenarios in several ways. First, in preventive medicine, a model stratifying risk levels can be more usable than a model performing binary decision. Through our ML approach, we transformed binary AD outcomes into multi-level AD risk levels, and the binary classification task was transformed into a multi-level classification task. Our XGBoost model does not predict whether a newborn would have AD within 6 months of age. Instead, our XGBoost model put a newborn into a certain AD risk level. With the assistance of our model, clinicians can classify newborns into different AD risk levels and design different prevention plans. Second, our ML method can prevent underestimation of AD risk and can solve the imbalanced data problem. For a newborn whose risk is merely lower than the risk threshold of having AD, a model only predicting binary outcomes would predict that this individual would not have AD. This leads to underestimation of risk, and consequently, clinicians and caregivers may not take sufficient preventive measures to protect this newborn. Risk underestimation can even be worsened by the problem of imbalance data. When majority and minority outcomes of a dataset are highly imbalanced, it is common to yield a model that tends to predict majority outcomes. In the TBCS data, without AD is the majority outcome, which

doi:10.6342/NTU202202317

accounted for 92% of the boys and 94% of the girls. A model trained on this imbalanced data can easily predict a newborn to be without AD. This imbalanced prediction tendency is an additional aggravating factor for risk underestimation.

5.2 Limitations

There are several limitations in this study. First, the test set we used to validate our models was derived from the TBCS data. We did not use an independent external dataset due to lack of such data. To enhance the robustness of the validation, we performed 5-run validation for 10,000 times. Second, the AD status of the study participants was self-reported. Whether or not the newborns were diagnosed with AD by a physician was not validated by medical records or other databases. Recall bias could occur. Last, the TBCS did not collect biological samples such as urine, blood or skin tissue. Therefore, such data was not available for model establish. Although the XGBoost model was set up solely based on epidemiological and socioeconomic risk factors, it achieved good prediction performance with high prediction accuracy, low RMSE and low weighted error.

doi:10.6342/NTU202202317

Chapter 6. Conclusion

In this study, a combination of a novel ML approach and XGBoost algorithm was applied on a national representative birth cohort to set up a prediction model for risk of AD in 6-month-old newborns. Our XGBoost model improved AD prediction compared to the previously established LR model and can help clinicians classify newborns into risk levels of AD and take timely preventive measures. The incidence of many other diseases such as malignancies can be much lower than AD and the datasets of such diseases can be much more imbalanced than this study. Our ML approach provides a promising solution to solve the imbalanced data problem for such disease.

doi:10.6342/NTU202202317

Reference

1. Wen, H.-J., et al., *Predicting risk for early infantile atopic dermatitis by hereditary and environmental factors.* British Journal of Dermatology, 2009. **161**(5): p. 1166-1172.
2. Weidinger, S., et al., *Atopic dermatitis.* Nat Rev Dis Primers, 2018. **4**(1): p. 1.
3. Hwang, C.Y., et al., *Prevalence of atopic dermatitis, allergic rhinitis and asthma in Taiwan: a national study 2000 to 2007.* Acta Derm Venereol, 2010. **90**(6): p. 589-94.
4. Illi, S., et al., *The natural course of atopic dermatitis from birth to age 7 years and the association with asthma.* J Allergy Clin Immunol, 2004. **113**(5): p. 925-31.
5. Kay, J., et al., *The prevalence of childhood atopic eczema in a general population.* Journal of the American Academy of Dermatology, 1994. **30**(1): p. 35-39.
6. Deckers, I.A.G., et al., *Investigating International Time Trends in the Incidence and Prevalence of Atopic Eczema 1990–2010: A Systematic Review of Epidemiological Studies.* PLOS ONE, 2012. **7**(7): p. e39803.
7. Odhiambo, J.A., et al., *Global variations in prevalence of eczema symptoms in children from ISAAC Phase Three.* J Allergy Clin Immunol, 2009. **124**(6): p. 1251-8.e23.
8. Liao, P.F., et al., *Prevalence of childhood allergic diseases in central Taiwan over the past 15 years.* Pediatr Neonatol, 2009. **50**(1): p. 18-25.
9. Chu, C.-Y., et al., *Taiwanese Dermatological Association consensus for the management of atopic dermatitis.* Dermatologica Sinica, 2015. **33**(4): p. 220-230.
10. Drucker, A.M., et al., *The Burden of Atopic Dermatitis: Summary of a Report for the National Eczema Association.* Journal of Investigative Dermatology, 2017. **137**(1): p. 26-30.
11. Carroll, C.L., et al., *The Burden of Atopic Dermatitis: Impact on the Patient, Family, and Society.* Pediatric Dermatology, 2005. **22**(3): p. 192-199.
12. Barbeau, M. and H.L. Bpharm, *Burden of Atopic dermatitis in Canada.* International Journal of Dermatology, 2006. **45**(1): p. 31-36.
13. Lee, B.W. and P.R. Detzel, *Treatment of Childhood Atopic Dermatitis and Economic Burden of Illness in Asia Pacific Countries.* Annals of Nutrition and Metabolism, 2015. **66(suppl 1)**(Suppl. 1): p. 18-24.
14. Tsai, T.-F., et al., *Burden of atopic dermatitis in Asia.* The Journal of

doi:10.6342/NTU202202317

Dermatology, 2019. **46**(10): p. 825-834.
15. Ng, Y.T. and F.T. Chew, *A systematic review and meta-analysis of risk factors associated with atopic dermatitis in Asia.* The World Allergy Organization journal, 2020. **13**(11): p. 100477-100477.
16. Parikh, R.B., et al., *Machine Learning Approaches to Predict 6-Month Mortality Among Patients With Cancer.* JAMA Network Open, 2019. **2**(10): p. e1915997-e1915997.
17. Lundberg, S.M., et al., *Explainable machine-learning predictions for the prevention of hypoxaemia during surgery.* Nat Biomed Eng, 2018. **2**(10): p. 749-760.
18. Heo, J., et al., *Machine Learning-Based Model for Prediction of Outcomes in Acute Stroke.* Stroke, 2019. **50**(5): p. 1263-1265.
19. Christodoulou, E., et al., *A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models.* Journal of Clinical Epidemiology, 2019. **110**: p. 12-22.
20. Pavlou, M., et al., *How to develop a more accurate risk prediction model when there are few events.* BMJ : British Medical Journal, 2015. **351**: p. h3868.
21. Stoltzfus, J.C., *Logistic regression: a brief primer.* Acad Emerg Med, 2011. **18**(10): p. 1099-104.
22. Rajkomar, A., J. Dean, and I. Kohane, *Machine Learning in Medicine.* New England Journal of Medicine, 2019. **380**(14): p. 1347-1358.
23. Chen, T. and C. Guestrin, *XGBoost: A Scalable Tree Boosting System*. 2016. 785-794.
24. Wang, L., et al., *Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model.* Healthcare, 2020. **8**(3): p. 247.
25. Chang, W., et al., *A Machine-Learning-Based Prediction Method for Hypertension Outcomes Based on Medical Data.* Diagnostics (Basel, Switzerland), 2019. **9**(4): p. 178.
26. Taninaga, J., et al., *Prediction of future gastric cancer risk using a machine learning algorithm and comprehensive medical check-up data: A case-control study.* Scientific Reports, 2019. **9**(1): p. 12384.
27. Lung, F.-W., et al., *Developing and refining the Taiwan Birth Cohort Study (TBCS): Five years of experience.* Research in Developmental Disabilities, 2011. **32**(6): p. 2697-2703.
28. Moore, M.M., et al., *Perinatal predictors of atopic dermatitis occurring in the first six months of life.* Pediatrics, 2004. **113**(3 Pt 1): p. 468-474.
29. Olesen, A.B., et al., *Atopic dermatitis and birth factors: historical follow up by record linkage.* BMJ (Clinical research ed.), 1997. **314**(7086): p. 1003-1008.

doi:10.6342/NTU202202317

30. Zhu, T., et al., *Association of very preterm birth with decreased risk of eczema: A systematic review and meta-analysis.* Journal of the American Academy of Dermatology, 2018. **78**(6): p. 1142-1148.e8.
31. Panduru, M., et al., *Birth weight and atopic dermatitis: systematic review and meta-analyis.* Acta Dermatovenerol Croat, 2014. **22**(2): p. 91-6.
32. Wooldridge, A.L., et al., *Relationship between birth weight or fetal growth rate and postnatal allergy: A systematic review.* J Allergy Clin Immunol, 2019. **144**(6): p. 1703-1713.
33. Parazzini, F., et al., *Perinatal factors and the risk of atopic dermatitis: a cohort study.* Pediatr Allergy Immunol, 2014. **25**(1): p. 43-50.
34. Skajaa, N., et al., *Cesarean delivery and risk of atopic dermatitis.* Allergy, 2020. **75**(5): p. 1229-1231.
35. Richards, M., et al., *Caesarean delivery and the risk of atopic dermatitis in children.* Clin Exp Allergy, 2020. **50**(7): p. 805-814.
36. Bager, P., J. Wohlfahrt, and T. Westergaard, *Caesarean delivery and risk of atopy and allergic disease: meta-analyses.* Clin Exp Allergy, 2008. **38**(4): p. 634-42.
37. Park, S., I.C. Hwang, and H. Ahn, *Parental age at birth and the risk for atopic dermatitis.* Australasian Journal of Dermatology, 2019.
38. Ito, J. and T. Fujiwara, *Breastfeeding and risk of atopic dermatitis up to the age 42 months: a birth cohort study in Japan.* Ann Epidemiol, 2014. **24**(4): p. 267-72.
39. Ravn, N.H., et al., *How does parental history of atopic disease predict the risk of atopic dermatitis in a child? A systematic review and meta-analysis.* Journal of Allergy and Clinical Immunology, 2020. **145**(4): p. 1182-1193.
40. Xu, B., et al., *Maternal age at menarche and atopy among offspring at the age of 31 years.* Thorax, 2000. **55**(8): p. 691-3.
41. Lin, B., et al., *Breastfeeding and Atopic Dermatitis Risk: A Systematic Review and Meta-Analysis of Prospective Cohort Studies.* Dermatology, 2020. **236**(4): p. 345-360.

doi:10.6342/NTU202202317

Tables

doi:10.6342/NTU202202317

TABLE 1. FEATURES INCLUDED IN THE FINAL ANALYSIS

Features	**Cut-off values for discretization**
Gender	2 levels: male, female
Gestational week	3 levels: < 38, ≥ 38 and < 41, ≥ 41
Birth weight (grams)	3 levels: < 2500, ≥ 2500 and < 4000, ≥ 4000
Parity	3 levels: 1, 2, > 2
Cesarean section	2 levels: yes, no
Maternal age at childbirth (years)	3 levels: < 26, ≥ 26 and < 33, ≥ 33
Paternal age at childbirth (years)	3 levels: < 26, ≥ 26 and < 33, ≥ 33
Maternal asthma history	2 levels: yes, no
Paternal asthma history	2 levels: yes, no
Maternal AR history	2 levels: yes, no
Paternal AR history	2 levels: yes, no
Maternal AD history	2 levels: yes, no
Paternal AD history	2 levels: yes, no
Maternal first menarche (years)	3 levels: ≤ 11, > 11 and ≤ 15, 15
Parental education (years)	3 levels: both ≤ 12, one > 12 and one ≤ 12, both > 12
Breastfeeding duration (months)	3 levels: 0, 1, > 1
Molds on walls at home	2 levels: yes, no
Furniture painting during pregnancy	2 levels: yes, no
House renovation during pregnancy	2 levels: yes, no

Abbreviation: AR, allergic rhinitis; AD, atopic dermatitis

doi:10.6342/NTU202202317

TABLE 2. A PROTOTYPICAL CONFUSION MATRIX

	Actual AD risk level		
	Low	High	Very high
Predicted AD risk level			
Low	x_{11}	x_{12}	x_{13}
High	x_{21}	x_{22}	x_{23}
Very high	x_{31}	x_{32}	x_{33}

doi:10.6342/NTU202202317

TABLE 3. RULES OF 2-MODEL MIXED PREDICTION

AD risk level predicted by combining Model 1 and Model 2		AD risk level predicted by Model 2		
		Low	High	Very high
AD risk level predicted by Model 1	Low	Low	High	Low
	High	High	High	Very high
	Very high	Very high	High	Very high

AD risk level predicted by combining Model 1 and Model 3		AD risk level predicted by Model 3		
		Low	High	Very high
AD risk level predicted by Model 1	Low	Low	Low	Very high
	High	High	High	Very high
	Very high	Very high	High	Very high

AD risk level predicted by combining Model 2 and Model 3		AD risk level predicted by Model 3		
		Low	High	Very high
AD risk level predicted by Model 2	Low	Low	Low	Very high
	High	High	High	Very high
	Very high	Low	Very high	Very high

doi:10.6342/NTU202202317

TABLE 4. WEIGHT OF ERRORS

	Actual AD risk level		
	Low	High	Very high
Predicted AD risk level			
Low	0 ($\boldsymbol{w_{11}}$)	3 ($\boldsymbol{w_{12}}$)	10 ($\boldsymbol{w_{13}}$)
High	1 ($\boldsymbol{w_{21}}$)	0 ($\boldsymbol{w_{22}}$)	3 ($\boldsymbol{w_{23}}$)
Very high	2 ($\boldsymbol{w_{31}}$)	1 ($\boldsymbol{w_{32}}$)	0 ($\boldsymbol{w_{33}}$)

doi:10.6342/NTU202202317

TABLE 5. DEMOGRAPHIC CHARACTERISTICS OF THE 6-MONTH-OLD FEMALE NEWBORNS AND UNADJUSTED OR FOR AD

Categorical variables	**No. of subjects**	**No. with AD (%)**	**Unadjusted OR (95% CI)**
Total	9607	557	
Delivery method			
Cesarean section	3303	181 (5.5)	Reference
Natural delivery	6574	376 (5.7)	0.96 (0.80-1.15)
Birth order			
1st child	4924	304 (6.2)	Reference
2nd child	3730	202 (5.4)	0.87 (0.72-1.04)
≥ 3rd child	953	51 (5.4)	0.86 (0.63-1.16)
Maternal education level			
≤ 12 years	5259	233 (4.4)	Reference
> 12 years	4348	324 (7.5)	1.74 (1.46-2.07) ***
Paternal education level			
≤ 12 years	5187	222 (4.3)	Reference
> 12 years	4420	335 (7.6)	1.83 (1.54-2.19) ***
Maternal asthma history			
No	9457	539 (5.7)	Reference
Yes	150	18 (12.0)	2.26 (1.32-3.62) **
Paternal asthma history			
No	9503	546 (5.7)	Reference
Yes	104	11 (10.6)	1.94 (0.98-3.49)
Maternal AR history			
No	8175	436 (5.3)	Reference
Yes	1432	121 (8.4)	1.64 (1.32-2.01) ***
Paternal AR history			
No	8041	408 (5.1)	Reference
Yes	1566	149 (9.5)	1.97 (1.61-2.39) ***
Maternal AD history			
No	9303	502 (5.4)	Reference
Yes	304	55 (18.1)	3.87 (2.83-5.22) ***
Paternal AD history			
No	9320	511 (5.5)	Reference
Yes	287	46 (16.0)	3.29 (2.34-4.52) ***
Molds on walls at home			
No	6123	319 (5.2)	Reference
Yes	3484	238 (6.8)	1.33 (1.12-1.59) **
Painting during pregnancy			
No	9358	537(5.7)	Reference
Yes	249	20 (8.0)	1.43 (0.87-2.23)
Renovation during pregnancy			
No	9219	520 (5.6)	Reference
Yes	388	37 (9.5)	1.76 (1.22-2.47) **
Continuous variables		**Mean (SD)**	**Unadjusted OR (95% CI)**
Gestational length (weeks)		38.7 (1.29)	1.07 (1.01-1.14) *
Birth weight (gram)		3126.6 (400.8)	1.00 (1.00-1.00) **
Breastfeeding duration (months)		2.4 (2.3)	1.08 (1.04-1.12) ***
Maternal age at childbirth (years)		30.3 (4.4)	1.04 (1.03-1.06) ***
Paternal age at childbirth (years)		33.3 (4.8)	1.00 (0.99-1.02)
Mother menarche age (years)		13.2 (1.3)	0.91 (0.86-0.97) **

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis

*P < 0.05, **P < 0.01, ***P < 0.001

doi:10.6342/NTU202202317

TABLE 6. DEMOGRAPHIC CHARACTERISTICS OF 6-MONTH-OLD MALE NEWBORNS AND UNADJUSTED OR FOR AD

Categorical variables	**No. of subjects**	**No. with AD (%)**	**Unadjusted OR (95% CI)**
Total	10628	858	
Delivery method			
Cesarean section	3597	273 (7.6)	Reference
Natural delivery	7031	585 (8.3)	1.11 (0.95-1.28)
Birth order			
1st child	5384	456 (8.5)	Reference
2nd child	4041	330 (8.2)	0.96 (0.83-1.11)
≥ 3rd child	1203	72 (5.9)	0.69 (0.53-0.88) **
Maternal education level			
≤ 12 years	5754	333 (5.8)	Reference
> 12 years	4847	525 (10.8)	1.97 (1.70-2.27) ***
Paternal education level			
≤ 12 years	5675	328 (5.8)	Reference
> 12 years	4953	530 (10.7)	1.95 (1.69-2.26) ***
Maternal asthma history			
No	10473	836 (8.0)	Reference
Yes	155	22 (14.2)	1.91 (1.18-2.95) **
Paternal asthma history			
No	10493	842 (8.0)	Reference
Yes	135	16 (11.9)	1.54 (0.88-2.53)
Maternal AR history			
No	8949	655 (7.3)	Reference
Yes	1679	203 (12.1)	1.74 (1.47-2.05) ***
Paternal AR history			
No	8909	648 (7.3)	Reference
Yes	1719	210 (12.2)	1.77 (1.50-2.09) ***
Maternal AD history			
No	10274	771 (7.5)	Reference
Yes	354	87 (24.6)	4.02 (3.10-5.15) ***
Paternal AD history			
No	10323	778 (7.5)	Reference
Yes	305	80 (26.2)	4.36 (3.33-5.66) ***
Molds on walls at home			
No	6855	498 (7.3)	Reference
Yes	3773	360 (9.5)	1.35 (1.17-1.55) ***
Painting during pregnancy			
No	10349	818 (7.9)	Reference
Yes	279	40 (14.3)	1.95 (1.37-2.71) ***
Renovation during pregnancy			
No	10218	810 (7.9)	Reference
Yes	410	48 (11.7)	1.54 (1.12-2.08) **
Continuous variables		**Mean (SD)**	**Unadjusted OR (95% CI)**
Gestational length (weeks)		38.4 (1.45)	1.02 (0.97-1.06)
Birth weight (gram)		3196.3 (427.4)	1.00 (0.99-1.00)
Breastfeeding duration (months)		2.1 (2.2)	1.06 (1.03-1.09) ***
Maternal age at childbirth (years)		30.3 (4.4)	1.04 (1.02-1.06) ***
Paternal age at childbirth (years)		33.4 (4.8)	1.00 (0.99-1.02)
Mother menarche age (years)		13.3 (1.3)	0.93 (0.88-0.97) **

Abbreviation: AD, atopic dermatitis; AR, allergic rhinitis

*P < 0.05, **P < 0.01, ***P < 0.001

doi:10.6342/NTU202202317

TABLE 7. RESULTS OF THE FINAL 3-MODEL MIXED PREDICTION AND THE LOGISTIC REGRESSION MODEL ON THE TEST SETS OF BOTH GENDERS.

Girls

XGBoost				Logistic regression			
	Actual AD risk level				Actual AD risk level		
Predicted AD risk level	Low (eP < 6%)	High (eP = 6%-12%)	Very high (eP ≥ 12%)	Predicted AD risk level	Low (eP < 6%)	High (eP = 6%-12%)	Very high (eP ≥ 12%)
Low (eP < 6%)	0.953	0.047	0.000	Low (eP < 6%)	0.958	0.042	0.000
High (eP = 6%-12%)	0.182	0.753	0.065	High (eP = 6%-12%)	0.148	0.734	0.118
Very high (eP ≥ 12%)	0.095	0.199	0.706	Very high (eP ≥ 12%)	0.127	0.229	0.644
Sum of diagonals =		2.412		Sum of diagonals =		2.337	
RMSE =		0.533		RMSE =		0.580	
Weighted error =		0.302		Weighted error =		0.370	

Boys

XGBoost				Logistic regression			
	Actual AD risk level				Actual AD risk level		
Predicted AD risk level	Low (eP < 8%)	High (eP=8%-16%)	Very high (eP ≥ 16%)	Predicted AD risk level	Low (eP < 8%)	High (eP=8%-16%)	Very high (eP ≥ 16%)
Low (eP < 8%)	0.963	0.037	0.000	Low (eP < 8%)	0.936	0.064	0.000
High (eP = 8%-16%)	0.182	0.811	0.007	High (eP = 8%-16%)	0.198	0.772	0.030
Very high (eP ≥ 16%)	0.027	0.157	0.816	Very high (eP ≥ 16%)	0.021	0.158	0.821
Sum of diagonals =		2.590		Sum of diagonals =		2.529	
RMSE =		0.394		RMSE =		0.412	
Weighted error =		0.175		Weighted error =		0.227	

eP, expected probability.

doi:10.6342/NTU202202317

Figures

doi:10.6342/NTU202202317

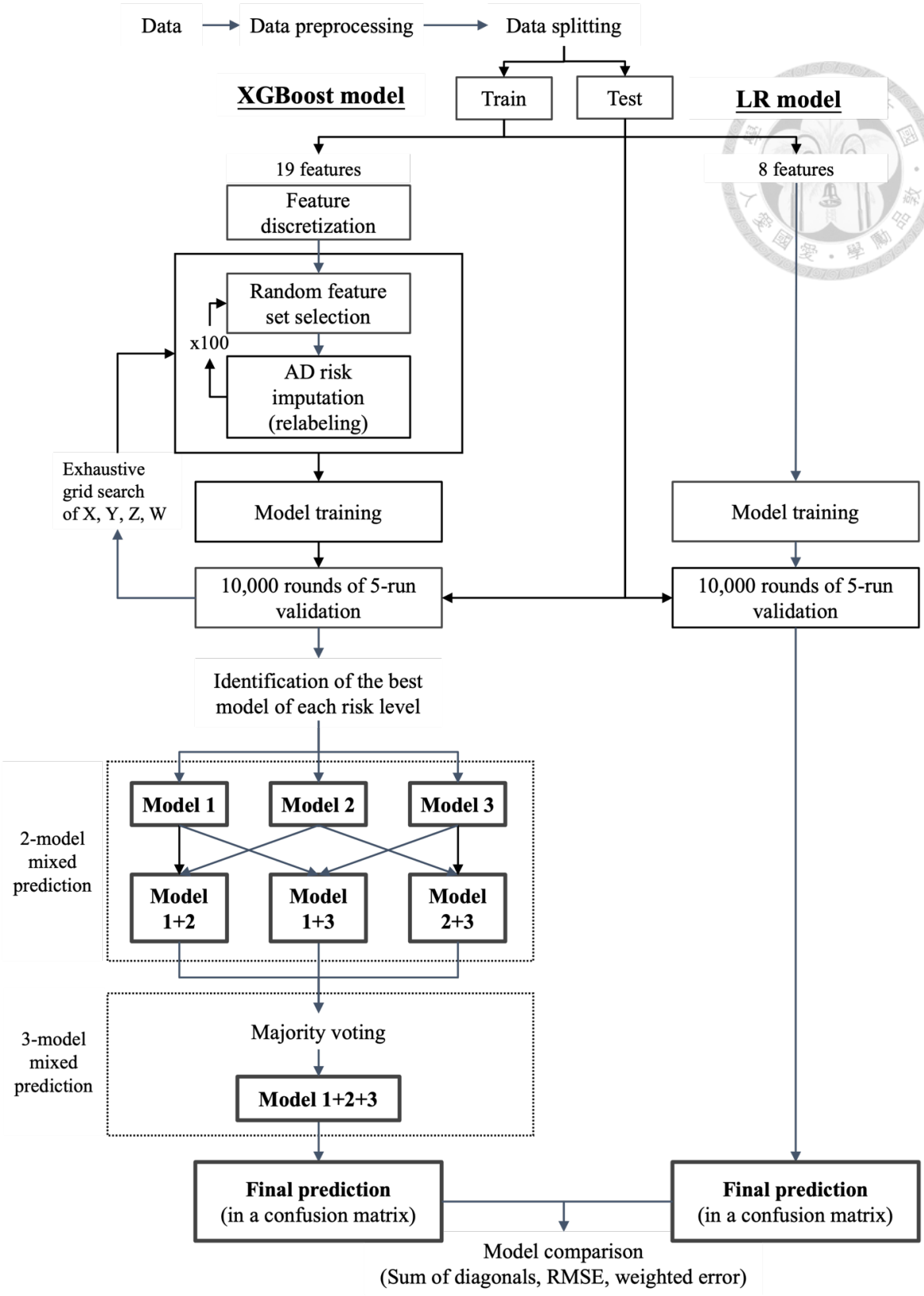


FIGURE 1. STEPS OF ANALYSIS.

Data was preprocessed and split into a train set (80%) and a test set (20%). The XGBoost model was set up through random feature set selection, AD risk imputation (relabeling), model training, 5-run validation, 2-model and 3-model mixed prediction. The LR model was set up through model training and 5-run validation. The performance of the XGBoost and LR model was compared based on sum of diagonals, RMSE and weighted error.

doi:10.6342/NTU202202317

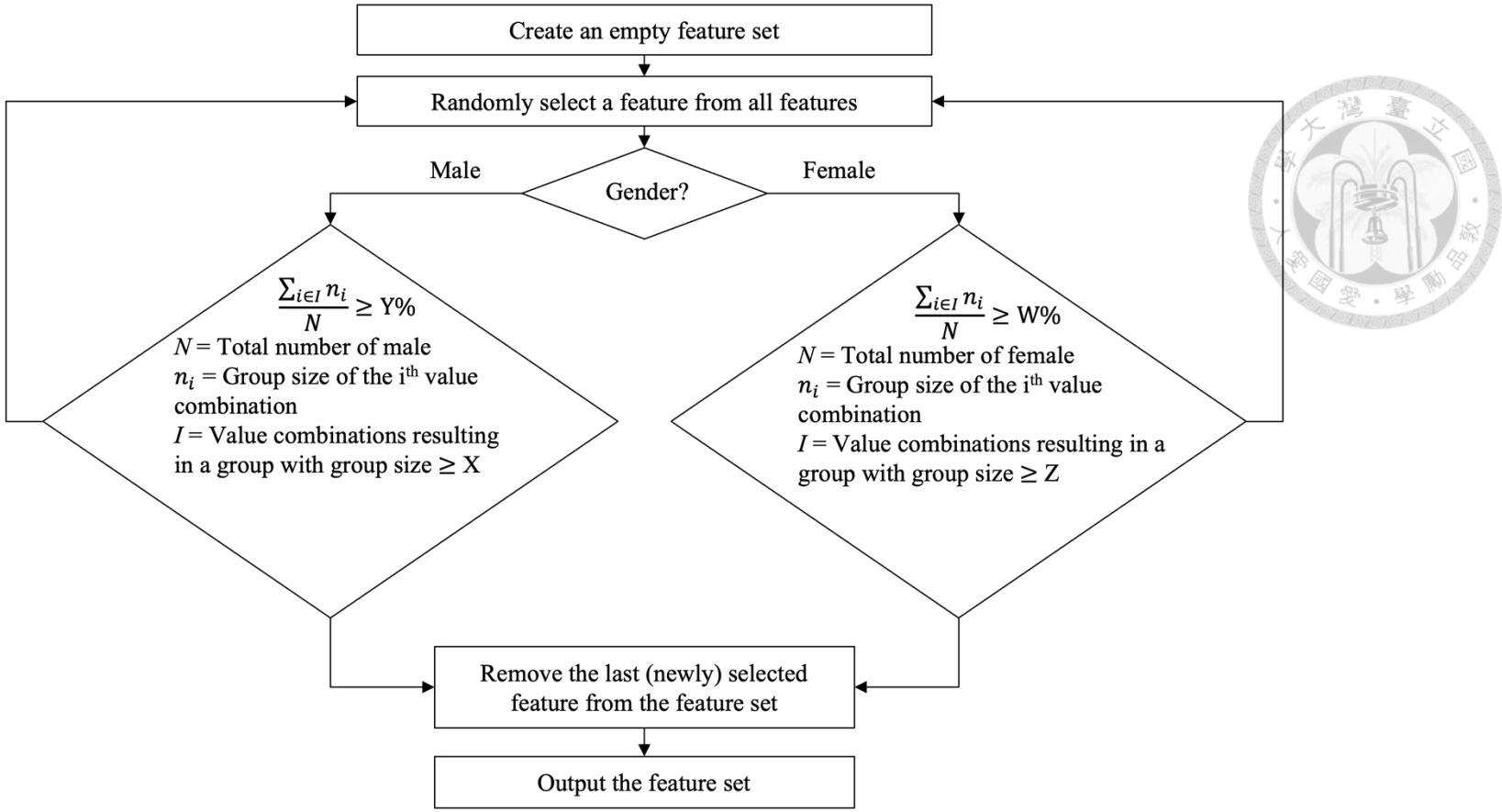


FIGURE 2. ALGORITHM AND CRITERIA OF RANDOM FEATURE SET SELECTION.

The random feature set selection process began with an empty feature set. One feature was randomly selected and added in the feature set. Next, the study participants who had the same value combination of the selected feature(s) were grouped together. We set the above criteria to keep at least a certain percentage of groups in optimal size. The group size was kept above X for at least Y% of the boys, and above Z for at least W% of the girls. After a new feature was selected and added in the feature set, we examined the criteria. If the criteria were met, the newly selected feature stayed in the feature set. If the criteria were not met, the newly selected feature was removed from the feature set. The feature was selected one at a time until the above criteria cannot be kept. The parameter X, Y, Z and W were adjusted through a grid search process in a later step to identify the best model of each risk level.

doi:10.6342/NTU202202317

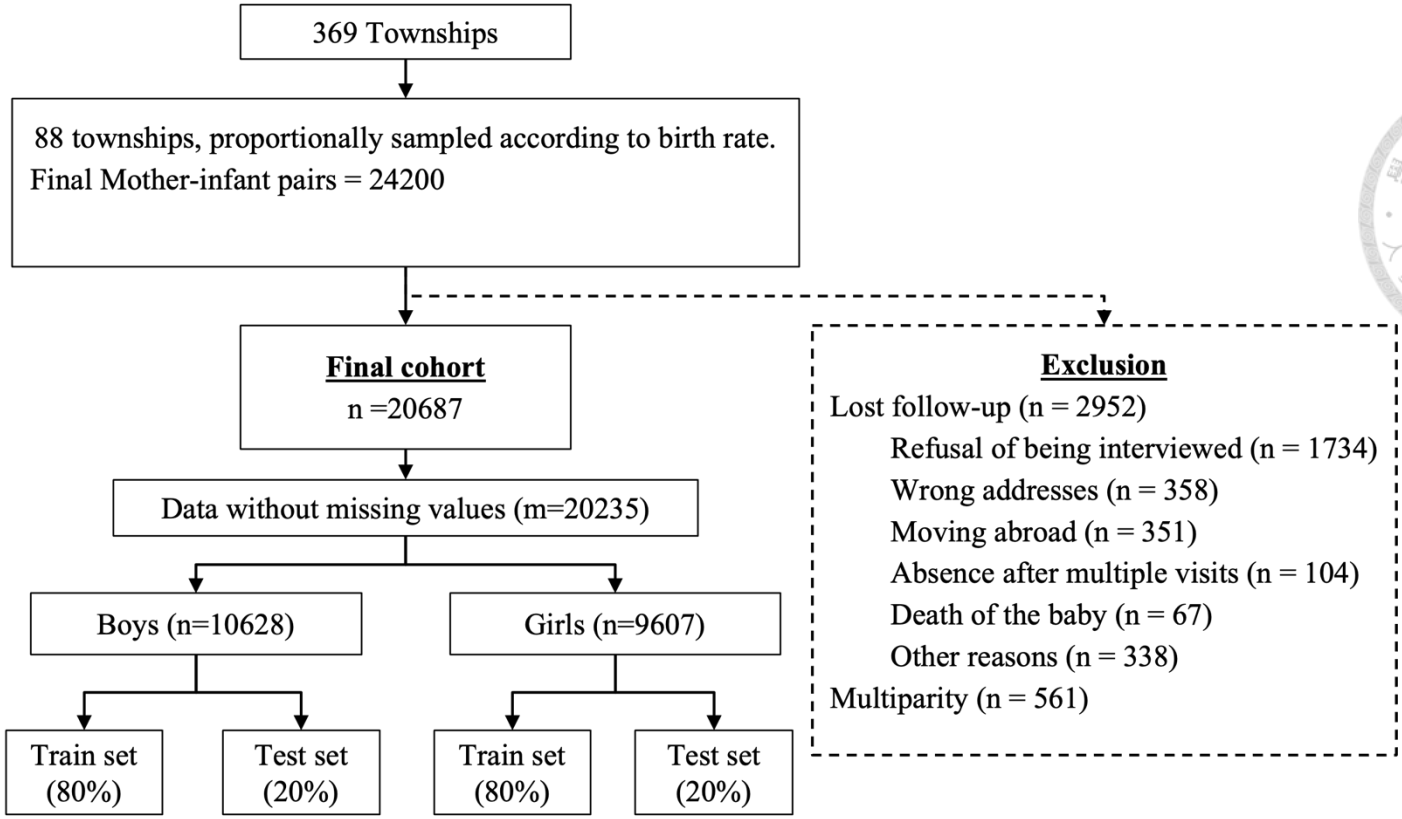


FIGURE 3. STUDY COHORT ENROLLMENT, DATA PREPROCESSING AND SPLITTING.

doi:10.6342/NTU202202317