Chapter 1

Introduction

1.1 Background

Ischemic stroke is the most common type of stroke and one of the leading causes of morbidity and mortality worldwide. It is centered on the interruption of blood flow to the brain, leading to oxygen deprivation and tissue death (infarction), responsible for approximately 87% of all stroke cases (Lui et al., 2025). This condition is affected by modifiable and non-modifiable risk factors, ranging from advanced age and biological sex to lifestyle choices such as diet and smoking. Not only that, genetics also becomes a major factor in stroke susceptibility (Boehme et al., 2017). There are some cases, although very rare, caused by single-gene mutations, however, the majority of ischemic strokes are polygenic in nature. This means that it is affected by a complex web of hundreds or thousands of genes, which affects pathways such as vascular integrity, blood pressure regulation, and coagulation. However, a significant limitation of current clinical risk assessments is their reliance on observable physiological markers (Oyovwi et al., 2025). These factors explain the state of a patient's health but fail to account for the hidden baseline genetic burden that exists regardless of lifestyle, leaving a gap in our ability to predict and prevent cerebrovascular diseases before they occur.

In recent years, Polygenic Risk Scores (PRS) have risen as a powerful tool to quantify this genetic susceptibility by aggregating the effects of thousands of Single Nucleotide Polymorphisms (SNPs) into a single metric (Schwarzerova et al., 2024). However, a major limitation in the clinical translation of modern genetics is ancestry bias. The majority of Genome-Wide Association Studies (GWAS) have historically been conducted on populations of European descent and because of the genetic architecture, including SNP allele frequencies and linkage disequilibrium patterns that varies significantly across populations, applying European-derived genetic weights to Asian populations systematically underestimates risk and yields poor predictive accuracy (Ciochetti et al., 2023). Other than that, a critical computational limitation also exists. In order to utilize cross-ancestry genetic

weights on a specific population, it requires a highly standardized bioinformatics pipeline that is capable of harmonizing various global summary statistics with local patient data.

While large-scale consortia have developed ancestry-specific summary statistics for various regions, Southeast Asia still remains underrepresented, particularly the Indonesian population. They possessed a highly complex and uniquely mixed genetic history, shaped by both Austronesian and Asian influences (Fachrul et al., 2026). Currently, there is no published and standardized computational framework for PRS calculation optimized for the Indonesian population. Furthermore, it remains unclear in which ancestral representatives provide the most accurate risk stratification for this group. To address these gaps, this study aims to develop a standardized cross-ancestry bioinformatics pipeline for PRS calculation in an Indonesian context. Furthermore, this study will test the pipeline's operational feasibility and evaluate preliminary predictive trends through a matched pilot cohort, systematically comparing a baseline clinical model against six distinct ancestry-specific and trans-ancestry PRS models.

1.2 Objective

The objective of this study is to develop a standardized bioinformatics pipeline for cross-ancestry Polygenic Risk Scores (PRS) and to evaluate its computational feasibility and preliminary predictive trends for ischemic stroke in a matched Indonesian pilot cohort.

1.3 Hypothesis

The standardized bioinformatics pipeline can successfully harmonize global GWAS data with a localized Indonesian dataset, and a PRS-integrated logistic regression model is computationally feasible to deploy. In addition, the PRS integrated model should outperform the clinical baseline model.