Contents

[Introduction 2](#_Toc181449030)

[Background on Kidney Stones 2](#_Toc181449031)

[Significance of Polygenic Risk Scores (PRS) 2](#_Toc181449032)

[Project Goal 3](#_Toc181449033)

[Statistical Considerations 4](#_Toc181449034)

[Project Structure 5](#_Toc181449035)

[Significance of the Capstone Theme 6](#_Toc181449036)

[Literature Review 7](#_Toc181449037)

[Existing PRS Methodologies 7](#_Toc181449038)

[Recent Advances in PRS with Machine Learning 8](#_Toc181449039)

[Deep Learning in PRS: A New Frontier 9](#_Toc181449040)

[Statistical Challenges in PRS for Complex Diseases 10](#_Toc181449041)

[Summary of Literature Review 11](#_Toc181449042)

# Introduction

## Background on Kidney Stones

Kidney stones, or renal calculi, represent a significant clinical challenge, affecting millions worldwide. As small, hard deposits formed from minerals and salts, they develop within the kidneys and can cause severe pain, hematuria, and even urinary tract obstruction. Approximately 10% of the population will experience kidney stones at least once in their lifetime, with recurrence rates as high as 50% within five years for those affected. Kidney stone formation is influenced by various factors, including dietary choices, hydration levels, and genetic predispositions. This complexity underlines the need for predictive tools that can integrate both genetic and environmental data to assess individual risk effectively.

Recent studies have underscored the role of genetics in kidney stone disease (KSD), identifying it as a heritable condition. Twin and family studies have consistently shown a higher prevalence among individuals with a family history of kidney stones, indicating the significant contribution of genetics to KSD risk. Various genetic markers, particularly single nucleotide polymorphisms (SNPs), have been associated with kidney stones. These findings have paved the way for polygenic risk scores (PRS) as a potential tool to estimate an individual’s predisposition to kidney stones based on the cumulative effect of multiple genetic variants.

## Significance of Polygenic Risk Scores (PRS)

Polygenic risk scores have emerged as a powerful approach to quantify genetic susceptibility to complex diseases like kidney stones. PRS aggregates the effects of numerous genetic variants, providing a single numerical score that reflects an individual's inherited risk for a particular condition. Unlike monogenic diseases, where single-gene mutations drive disease risk, complex diseases involve multiple genetic loci with small effect sizes, each contributing incrementally to the overall risk. PRS captures this polygenic architecture by combining the effects of many SNPs, each weighted according to their statistical association with the disease.

The relevance of PRS in clinical practice is gaining traction as it offers a personalized risk assessment tool, helping clinicians identify individuals at high risk for specific diseases. In the context of kidney stones, PRS could inform early intervention strategies, such as dietary recommendations or increased hydration, for individuals with a higher genetic predisposition. Moreover, PRS has the potential to guide resource allocation, targeting preventive care measures toward those most likely to benefit. However, the practical implementation of PRS in clinical settings remains challenging, especially due to limitations in model accuracy, interpretability, and data availability.

## Project Goal

This project aims to develop an accurate, interpretable, and statistically rigorous PRS model for kidney stone prediction, employing deep learning techniques. Traditional PRS models often rely on linear regression or logistic regression frameworks, assuming additive genetic effects across loci. While these models are computationally efficient, they may fail to capture the complex, non-linear interactions among genetic variants that contribute to kidney stone risk. Deep learning models, with their capacity to learn intricate patterns and relationships within large datasets, offer a promising alternative. By leveraging deep learning, this project seeks to improve the predictive power of PRS for kidney stones, while also addressing statistical concerns such as overfitting, model interpretability, and validation.

The primary objective of this capstone project is to apply advanced deep learning methods to enhance PRS prediction for kidney stones. By developing a model that captures both the additive and non-linear genetic effects, this project strives to provide a more accurate risk assessment tool. The model will incorporate genotypic data from individuals with and without kidney stones, aiming to predict disease status based on SNP profiles. Given the importance of statistical rigor in health-related research, this project will prioritize robust evaluation metrics, cross-validation, and comparison with baseline models to ensure reliable performance.

## Statistical Considerations

From a statistical standpoint, developing a PRS model for kidney stones entails several key challenges. First, the non-linear nature of genetic interactions necessitates a model capable of handling complex relationships within the data. Traditional linear models may oversimplify these relationships, leading to suboptimal predictions. In contrast, deep learning models, such as convolutional neural networks (CNNs) or attention-based models, can learn hierarchical features and non-linear interactions, offering a potentially superior approach to PRS.

Second, model interpretability is a critical concern, especially in the context of clinical applications. Deep learning models are often criticized as "black boxes," making it challenging to identify which SNPs contribute most significantly to the risk prediction. This lack of transparency can hinder clinical adoption, as practitioners may be reluctant to rely on models whose underlying mechanisms are not fully understood. To address this, the project will incorporate interpretability techniques, such as SHAP values or attention mechanisms, to provide insights into the SNPs driving risk predictions.

Third, the project will employ rigorous validation strategies to ensure model reliability. Overfitting is a common issue in deep learning, particularly with small or imbalanced datasets. To mitigate this, the model will be trained and validated using cross-validation and regularization techniques. Additionally, statistical performance metrics, including area under the receiver operating characteristic curve (AUC-ROC), sensitivity, and specificity, will be used to assess model accuracy. By comparing the deep learning model with traditional PRS approaches, the project will demonstrate the statistical advantages of using a deep learning framework for PRS in kidney stones.

## Project Structure

The project is structured to integrate statistical analysis with deep learning methodologies, reflecting a cohesive approach to PRS prediction. The initial phase involves data preprocessing, including SNP filtering, missing data handling, and quality control measures, to ensure data integrity. Following this, feature engineering techniques will be applied to select SNPs based on statistical criteria, such as variance and correlation thresholds. The deep learning model will then be implemented, with a focus on capturing complex genetic interactions.

The model will be evaluated using both statistical metrics and clinical relevance, providing a comprehensive assessment of its performance. Statistical tests, such as permutation testing, will be conducted to compare the deep learning model’s performance against baseline models, establishing the significance of any observed improvements. Additionally, interpretability tools will be used to identify the SNPs most predictive of kidney stone risk, providing insights into the genetic factors contributing to the condition.

Finally, the project will conclude with a discussion of the model’s limitations, statistical implications, and potential for clinical translation. By combining deep learning with rigorous statistical methodology, this project aims to set a foundation for future research in PRS and contribute to the growing field of precision medicine.

## Significance of the Capstone Theme

As a statistics capstone project, this work emphasizes the integration of statistical principles with cutting-edge machine learning techniques. Deep learning models, while powerful, can benefit significantly from statistical rigor, particularly in terms of model evaluation, interpretability, and generalizability. By approaching PRS prediction from a statistical perspective, this project not only contributes to kidney stone research but also highlights the critical role of statistics in developing reliable, interpretable, and clinically relevant models.

In summary, this project represents a unique intersection of statistics and machine learning, applying deep learning to a complex, clinically relevant problem in kidney stone prediction. Through a blend of statistical analysis, model interpretability, and rigorous validation, this project aims to push the boundaries of PRS research, offering a novel approach to genetic risk assessment for kidney stones.

# Literature Review

## Existing PRS Methodologies

Polygenic risk scores (PRS) have traditionally been calculated using statistical models that aggregate the effects of numerous single nucleotide polymorphisms (SNPs) across the genome. These models, such as linear regression and logistic regression, have been the cornerstone of PRS calculations, primarily due to their simplicity and interpretability. Early PRS models were based on additive genetic effects, assuming that each SNP contributes independently and additively to disease risk. This assumption, while computationally efficient, is limited by its inability to capture complex interactions between genetic loci, which are increasingly recognized as important factors in multifactorial diseases like kidney stones.

The additive framework often involves weighting each SNP by its effect size, derived from genome-wide association studies (GWAS). PRS models then sum these weighted contributions to produce a single score representing genetic risk. Although widely used, these methods face limitations when applied to diseases with intricate genetic architectures, as they may fail to account for non-linear effects and epistatic interactions. Studies have highlighted that purely additive models might oversimplify the genetic landscape, leading to underperformance in predictive accuracy, particularly in diverse populations where allele frequencies and genetic backgrounds vary.

## Recent Advances in PRS with Machine Learning

To address the limitations of traditional PRS models, researchers have increasingly turned to machine learning methods that allow for more flexible modeling of genetic data. Machine learning techniques, such as random forests, support vector machines, and gradient boosting, have been applied to PRS modeling to capture complex interactions among SNPs. These methods offer enhanced predictive performance compared to traditional approaches by exploring non-linear relationships within the data.

For instance, random forests have been employed in PRS to leverage interactions between SNPs through ensemble learning, aggregating multiple decision trees to make predictions. Gradient boosting, another popular technique, iteratively improves predictive accuracy by adding weak learners to the model. Studies using these machine learning approaches have reported higher predictive accuracy than traditional models, particularly when non-linear effects play a significant role in disease etiology.

However, while machine learning models address some limitations of additive models, they come with their own set of challenges. Machine learning techniques are often more computationally demanding, require larger sample sizes to avoid overfitting, and are sometimes criticized for their lack of interpretability—a crucial aspect in medical research. Understanding which SNPs or interactions contribute most significantly to risk is essential for translating PRS findings into clinical insights, where explainability can directly impact clinical decision-making.

## Deep Learning in PRS: A New Frontier

Deep learning, a subset of machine learning, has garnered attention for its capacity to model complex, non-linear relationships and learn hierarchical features from data. Unlike traditional and machine learning approaches, deep learning models are capable of automatically identifying patterns within large, high-dimensional datasets, making them particularly suited for genetic data, where SNP interactions can be intricate and multi-layered.

In recent years, deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and attention-based models have been explored for PRS applications. CNNs, known for their efficacy in image recognition, can be adapted to genetic data by treating SNPs as spatially arranged features. This allows the model to identify localized patterns and interactions within the data, effectively capturing the non-linear effects that contribute to disease risk. RNNs, on the other hand, are useful for sequential data, making them applicable to ordered SNPs across the genome. Attention mechanisms further enhance model interpretability by focusing on the most relevant features, or SNPs, for each prediction.

A notable advantage of deep learning models in PRS research is their ability to integrate multimodal data, such as combining SNP data with phenotypic information (e.g., age, gender, BMI). By incorporating both genetic and clinical data, deep learning models provide a more comprehensive view of disease risk, potentially offering improved predictive accuracy compared to models based solely on genetic data.

Despite their promise, deep learning models in PRS face several challenges. The complexity of these models requires large datasets for training, which may not always be available in genetic research. Additionally, deep learning models are often criticized for being “black boxes,” as their internal decision-making processes are difficult to interpret. This lack of transparency presents a barrier to clinical adoption, as clinicians may be hesitant to rely on models whose predictions cannot be easily explained.

## Statistical Challenges in PRS for Complex Diseases

The application of PRS to complex diseases like kidney stones brings unique statistical challenges that must be addressed to ensure model reliability and clinical relevance. One of the primary concerns is **overfitting**, particularly when working with high-dimensional genetic data and limited sample sizes. Overfitting occurs when a model learns noise or random fluctuations in the training data rather than the true underlying patterns, leading to poor generalizability. This issue is exacerbated in deep learning models, which have a high capacity to fit data but may struggle to generalize in the absence of adequate regularization techniques or cross-validation.

Another statistical challenge is **data imbalance**, where there may be more controls (healthy individuals) than cases (those with kidney stones). Imbalanced datasets can lead to biased predictions, with models favoring the majority class and thus underperforming in identifying true cases. Techniques such as oversampling the minority class, undersampling the majority class, or using balanced accuracy as an evaluation metric are commonly employed to address this issue.

**Interpretability** is a critical statistical consideration in PRS, as it relates to both the model’s internal structure and the clinical relevance of its predictions. For PRS to be clinically useful, the model should ideally highlight which SNPs or interactions are most influential in determining disease risk. Traditional PRS methods offer interpretability through effect sizes associated with each SNP, while deep learning models often lack this transparency. Recent efforts to improve interpretability in deep learning have focused on techniques like SHAP (SHapley Additive exPlanations) values, which approximate the contribution of each input feature to the model’s prediction. Attention mechanisms, as used in transformer models, also offer a pathway to interpretability by weighting relevant features during prediction.

Finally, **statistical validation** is essential to ensure that the PRS model is robust and generalizable. Cross-validation techniques, such as k-fold or stratified cross-validation, are widely used to assess model performance on different subsets of the data. Statistical performance metrics, including area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, and F1-score, provide insights into model accuracy, balance, and reliability. In PRS research, permutation testing or bootstrapping methods can further validate model significance by comparing performance against random baselines.

## Summary of Literature Review

In summary, traditional statistical methods for PRS have laid a foundation for understanding genetic risk but face limitations in modeling complex diseases like kidney stones. Machine learning approaches have expanded PRS capabilities, offering improved accuracy by capturing non-linear relationships among SNPs. Deep learning, with its advanced architectures and ability to handle high-dimensional data, represents the next frontier for PRS prediction. However, the statistical challenges of interpretability, overfitting, data imbalance, and validation remain significant, underscoring the need for careful methodological considerations. This project aims to address these challenges by leveraging deep learning while adhering to rigorous statistical standards, contributing to the development of a robust, clinically relevant PRS model for kidney stones.

**Methods**

**Data Collection and Preparation**

The dataset used for this study comprises genotypic and phenotypic data, with genotypes obtained from whole-genome sequencing and phenotypic data derived from electronic health records (EHR). The genotypic data includes single nucleotide polymorphisms (SNPs) across the genome for a cohort of individuals, with information on whether each SNP is homozygous reference, heterozygous, or homozygous alternate (encoded as 0/0, 0/1, and 1/1, respectively). Phenotypic data includes patient characteristics and the presence or absence of kidney stones, which serves as the outcome variable.

The initial data preprocessing stage involved loading raw genetic data in VCF (Variant Call Format) into Hail, an open-source framework optimized for large-scale genomic analysis. Hail was instrumental in performing quality control and data cleaning operations on the genotypic data. To ensure data integrity, we conducted filtering steps such as:

* **SNP Quality Control**: SNPs with low call rates or minor allele frequencies below a threshold were excluded to reduce noise and improve the quality of variant selection.
* **Missing Data Imputation**: Hail’s imputation functions were applied to fill in missing genotypes, enabling a complete dataset for downstream analysis.
* **Population Stratification**: Given the genetic diversity in populations, we used principal component analysis (PCA) within Hail to account for population structure. Individuals were assigned population labels based on principal components to control for ancestry-related confounding.

After these preprocessing steps, we annotated the genotypic data with phenotypic attributes. This process was essential for associating each individual’s genetic profile with their clinical status, enabling the development of a personalized polygenic risk score model.

**Hail: A Framework for Scalable Genomic Data Analysis**

Hail played a crucial role in this project, enabling efficient handling of large-scale genomic data through distributed computing. Hail’s MatrixTable data structure, designed for genomic data, allowed us to store variants by rows (keyed by locus and alleles) and individuals by columns (keyed by person\_id). This structure simplified the annotation process, as we could add phenotype information directly to each column.

Using Hail’s annotate\_cols function, we integrated the phenotype data with the genotypic information in ht\_final, the Hail Table containing our filtered and cleaned SNPs. This allowed each individual’s disease status, demographic data, and other clinical features to be accessible in the same data structure. Additionally, Hail’s support for regression analyses and efficient statistical testing enabled us to calculate summary statistics for each SNP, providing a foundation for the deep learning model by identifying SNPs with the highest relevance to kidney stone risk.

**Feature Engineering**

Feature engineering was performed to select relevant SNPs that contribute significantly to kidney stone risk. Initially, we considered a large number of SNPs; however, using all variants could lead to overfitting and increased computational costs in the deep learning model. To mitigate these issues, we applied filtering criteria based on:

* **Minor Allele Frequency (MAF)**: SNPs with very low frequencies were excluded, as their impact on the model’s predictive ability is minimal.
* **Linkage Disequilibrium (LD) Pruning**: Highly correlated SNPs were pruned to reduce redundancy. Using Hail, we calculated LD and removed SNPs with high pairwise correlation, retaining only one representative SNP per LD block.
* **Statistical Significance**: A preliminary regression analysis identified SNPs that were significantly associated with kidney stone status. Only SNPs with p-values below a threshold were included in the final dataset.

The final SNP set was encoded in a matrix format, with each row representing a person and each column representing a SNP, coded as 0, 1, or 2 (for 0/0, 0/1, and 1/1 genotypes, respectively). This feature matrix served as input for the deep learning model.

**Deep Learning Model Architecture**

For polygenic risk score prediction, we selected a deep learning model architecture capable of capturing non-linear interactions between SNPs and providing robust predictions. The final model was a multi-layer neural network (MLNN) with the following structure:

* **Input Layer**: The input layer took in the SNP matrix, representing each individual’s genotype across selected SNPs.
* **Hidden Layers**: Multiple dense (fully connected) hidden layers were included to enable the model to learn complex relationships. Each hidden layer was followed by batch normalization and dropout layers to prevent overfitting.
* **Activation Functions**: Rectified Linear Unit (ReLU) was used as the activation function for all hidden layers, providing non-linearity and enabling the model to capture intricate patterns in the data.
* **Output Layer**: The output layer consisted of a single node with a sigmoid activation function to predict the probability of kidney stone risk.

The model’s hyperparameters, such as the number of hidden layers, the number of neurons per layer, and dropout rates, were optimized through grid search and cross-validation. Additionally, early stopping was implemented to terminate training when performance on the validation set plateaued, preventing overfitting.

**Model Training and Validation**

The dataset was split into training, validation, and test sets in an 80-10-10 ratio. Given the class imbalance between individuals with and without kidney stones, class weights were applied during training to ensure the model did not favor the majority class. Cross-entropy loss was used as the objective function, and Adam optimizer was employed for gradient descent with an adaptive learning rate.

Cross-validation with k-folds was applied to assess the model’s robustness. Performance metrics, including area under the receiver operating characteristic curve (AUC-ROC), sensitivity, specificity, and F1 score, were computed on the validation and test sets. These metrics provided a balanced evaluation of the model’s performance, particularly important in a medical context where both sensitivity and specificity have clinical relevance.

**Statistical Evaluation and Validation**

A crucial aspect of this project was the statistical validation of the model’s predictions. Statistical metrics such as sensitivity, specificity, and AUC-ROC were calculated to provide insight into the model’s accuracy. Additionally, permutation tests were performed to assess the significance of the model’s performance compared to random guessing.

To ensure that the selected SNPs were not spurious, we conducted Bonferroni correction on p-values derived from the preliminary regression analysis. This adjustment helped control the family-wise error rate, ensuring that the SNPs used in the model were genuinely associated with kidney stone risk. The adjusted significance threshold helped in identifying truly predictive SNPs while mitigating the risk of false positives.

**Interpretability and Feature Importance**

Model interpretability was addressed using SHapley Additive exPlanations (SHAP) values, a method that attributes a contribution score to each input feature based on its influence on the model’s output. SHAP values allowed us to identify the SNPs with the highest impact on risk predictions, providing valuable insights into the genetic basis of kidney stones. This interpretability feature is essential for clinical applications, as it helps clinicians understand the genetic factors driving an individual’s risk, potentially guiding personalized prevention strategies.

Additionally, attention mechanisms were explored to highlight the most relevant SNPs for each prediction. This approach allowed the model to weigh specific variants more heavily based on their relevance to the outcome, thereby offering a layer of transparency in decision-making. Combining SHAP values and attention mechanisms provided a dual approach to interpretability, balancing individual SNP impact with broader patterns learned by the model.

**Software and Computational Resources**

All analyses were conducted using Hail, Python, and TensorFlow. Hail was employed for data preprocessing and statistical analysis, Python’s Pandas and NumPy libraries were used for data manipulation, and TensorFlow provided the deep learning framework. The model training was conducted on a high-performance computing cluster to handle the large data volume and computational demands of deep learning. Using distributed computing resources ensured that the model could be trained in a reasonable timeframe, facilitating rapid iteration and refinement of hyperparameters.

**Limitations and Challenges**

Several challenges were encountered in this project. First, the dependency on large genomic datasets required significant computational resources, both in terms of memory and processing power. Additionally, the issue of data imbalance required careful handling to ensure that the model did not produce biased predictions. Another challenge was interpretability; while deep learning models offer high accuracy, their complexity makes them less transparent. To address this, we implemented SHAP values and attention mechanisms, though these methods still leave some aspects of the model’s inner workings opaque.

Finally, the model’s reliance on SNP data means it may not generalize to all populations. Given genetic diversity across populations, the model’s predictive power may vary, necessitating further validation in diverse cohorts.