Early Genetic Disorder Prediction from Diagnosis and Inheritance Patterns

Emma Oo1, Sindhu Bhattarai2, and Dave Friesen3

University of San Diego

Master of Science, Applied Data Science

ADS-503 Summer 2022

June 27, 2022

Author Note

We have no conflicts of interest to disclose.

Correspondence concerning this paper should be addressed to Emma Oo, Sindhu Bhattarai, or Dave Friesen. Email: eoo@sandiego.edu, sbhattarai@sandiego.edu, or dfriesen@sandiego.edu

Abstract

The study of inheritance of genetic diseases from various gene factors and making an informed medical practice with the analytics and early prediction is groundbreaking for human genetic research. The aim of this paper is to study the various gene mutations, symptoms and predict the disorders which can help in early detection of the diseases. Various symptoms of diseases and the inheritance from maternal or paternal gene can contribute to prognostic symbols of various genetic disorder and can contribute on ad hoc clinical attention. Not just the inheritance but other several factors such as eating habits, environmental changes and life style might lead human body to become prone to such diseases. Different studies and approaches are under way to research about such issues. With that cognizance, our study is limited to identifying the pattern of various genome factors and symptoms leading towards different genetic disorders.

*Keywords*: [. . .]

# Table of Contents

List of Figures n

List of Tables n

List of Equations n

Early Genetic Disorder Prediction from Diagnosis and Inheritance Patterns n

Objective n

Method n

Data Collection and Pre-Processing n

Sample Characteristics n

Multivariate Analysis n

Results n

[topic-specific section] n

[topic-specific section] n

[topic-specific section] n

Discussion n

Hypothesis Review n

Strengths, Weaknesses, and Opportunities n

References n

Appendix A – Genetic Disorder Dataset n

Appendix B – Project Code n

# List of Figures

**Figure 1** [figure name] n

# List of Tables

**Table 1** [table name] n

# List of Equations

**Equation 1** [equation name] n

# Early Genetic Disorder Prediction from Diagnosis and Inheritance Patterns

Genetic diseases are a leading cause of childhood mortality (Clark et al., 2018). In the United States alone, the mortality rate is 15% among infants admitted to neonatal, pediatric, and cardiovascular intensive care units (ICUs) with genetic diseases. Worldwide, an estimated 3.3 million children under the age of 5 die yearly from serious congenital disabilities (Zarocostas, 2006). Further, Wojcik et al. note that approximately one-third of infants with genetic diseases die before diagnosis (Wojcik et al., 2018). These statistics and the potential benefits of early treatment emphasize an opportunity to prevent childhood genetic disease mortality through rapid diagnosis, prognosis, and treatment. They also highlight an opportunity to *predict* at-risk children (and inform treatment plans) even earlier than diagnosis through advanced data science techniques.

## Objective

Framed by the above problem statement, this study aims to determine the extent to which genetic disorders may be predicted (classified) from a combination of child diagnosis/context and parent gene trait. Major disorders include Alzheimer’s, cancer, Cystic fibrosis, diabetes, Hemochromatosis, Leber’s hereditary optic neuropathy, Leigh syndrome, Mitochondrial myopathy, and Tay-Sachs. These nine disorders equate to the study’s prediction classifications, and the objective translates to a hypothesis () that these disorders may be predicted (classified) within a reasonable margin of error, with a null hypothesis () of independence or no classification possible (in this use case).

# Method

The study followed a structured data research process, including exploratory data analysis, modeling, and evaluation, beginning with data acquisition, preparation, mining, and representation as described in this section.

## Data Collection and Pre-Processing

Secondary child genetic disorder data was obtained indirectly via Kaggle (Kaggle, 2022). These datasets were initially downloaded as “raw” comma-separated value (CSV) files, imported into a development R file, and preliminarily confirmed for essential data integrity. The following Table 1 summarizes the set of provided data assets:

### Table 1

Genetic Disorder Data Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **File Name** | **Dataset Description** | **Observations** | **Labels** | **Features** |
| train.csv | Training (sample) set of child genetic disorders with related parent gene characteristics and child treatment information | 22,083 | 2 | 43 |
| test.csv | Test set of child data *without* pre-labeled genetic disorder | 9,463 | 0 | 43 |

Based on a preliminary review, train.csv – as the only labeled dataset – was most relevant to this exploratory study and was retained for further analysis and modeling.

## Sample Characteristics

As summarized in Table 1, the full research dataset includes 22,083 unique child genetic disorder observations. After eliminating 3,140 observations with missing labels, 18,493 observations remained. This remaining dataset was then split into 80%/20% training/test components before further pre-processing and ultimately modeling. Note this split was stratified to accommodate imbalance across the nine prediction classifications.

Appendix B itemizes the genetic disorder dataset (train.csv) structure. In summary, the base dataset contains 43 prospective independent variables including 5 nominal-numeric, 2 continuous, 10 discrete (sets), 20 nominal-categorical, and 6 nominal (other) across patient and parent information. A preliminary analysis led to the immediate reduction of 16 of these variables as irrelevant to the study’s objective and hypothesis, leaving 27 for model consideration.

Univariate analysis and pre-processing steps included the following key checkpoints to optimize feature engineering, reduction, and selection:

* Class target and label review (including for imbalance)
* Missing values, distribution/outliers, variances, and other statistical characteristics
* Variable types and naming conventions (for model compatibility)
* Standardization/normalization requirements

Missing values were of key concern across multiple features, and imputation of these values ranged from medians for nominal-numeric and means for continuous, to additive categories for categorical predictors. Note more sophisticated imputation methods including k-NN and MICE were attempted with limited utility given many categorical features.

## Multivariate Analysis

Following univariate analysis and pre-processing, the study reviewed collinearity across the many categorical variables and multi-class target using Cramer’s V as a method to understand the association between nominal categorical values. While this was informative, it did not indicate collinearity issues among predictors, a scenario that may cause instability in certain classification algorithms (e.g., logistic regression). Further, this analysis did not reveal strong “correlation” between this subset of predictors and the target disorder variable, leaving further feature selection to iterative modeling.

# Results

Given a challenging subject area (domain), problem statement, and dataset, the study’s model strategy was to experiment through a series of multinomial classification algorithms including Linear Discriminate Analysis, Multinomial Logistic Regression, Nearest Shrunken Centroids, Random Forest, CART, Bagged Trees, and K-Nearest Neighbors. Restating the study’s objective, the goal of these model iterations was to determine the extent to which genetic disorders may be predicted (classified) from child characteristics and parent gene traits (reference Appendix B). This process was enhanced by evaluating multiple model approaches – the six supervised and one unsupervised algorithm listed above.

As noted above under Multivariate Analysis, the study’s dataset included many (primarily) categorical values. Categorical variables remaining as model features were converted to factors and ultimately “dummy encoded” as c-1 dichotomous variables to optimize their use across model algorithms.

The *caret* R library was used for all models except for Random Forest, which used the *randomForest* library. Final model parameters – where applicable – were ultimately configured to library default values except for Random Forest where weights were applied per target classification priors (frequency distribution) to support this model’s handling of class imbalance.

All models were trained and cross-validated on the stratify-sampled 80% training dataset (as referenced above under Sample Characteristics). Training and validation iterations included the full set[[1]](#footnote-2) of the remaining predictors and a much more selective set of four parental gene predictors. Basic *accuracy* for each of these iterations is summarized in Table 2 below. Note that, while preliminary model vetting focused on basic accuracy, additional classification metrics like recall, precision, F1 scores and others may be derived from code results (reference Appendix B).

### Table 2

Model Training and Validation Accuracy

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Full**  **Train** | **Full**  **Test** | **Gene-Only**  **Train** | **Gene-Only**  **Test** |
| Linear Discriminate Analysis | 0.340 | 0.354 | 0.266 | 0.266 |
| Multinomial Logistic Regression | 0.341 | 0.352 | 0.265 | 0.266 |
| Nearest Shrunken Centroids | 0.280 | 0.274 | 0.263 | 0.262 |
| Random Forest | 0.340 | 0.360 | 0.260 | 0.260 |
| CART | 0.365 | 0.382 | 0.270 | 0.270 |
| Bagged Trees | 0.354 | 0.370 | 0.260 | 0.269 |
| K-Nearest Neighbors | 0.231 | 0.228 | n/a | n/a |

# Discussion

## Hypothesis Review

Table 2 highlights the challenge of predicting (classifying) genetic disorders based on the child and parent characteristics available to this study and, consequently, difficulty in rejecting the study’s null hypothesis () of no classification possible.

## Strengths and Weaknesses of the Present Study

The study contained key strengths and weaknesses which warrant consideration for application or future research.

Strength:

* + The predictive analytics of genetic disorder is a beneficial study since it can help the children with rapid diagnosis and prognosis to some extent. This data set revealed the hidden information such as majority of the children with age seven are diagnosed with Leigh Syndrome. Maternal genes dominate slightly over paternal genes in regards to the diseases. Symptom 5 is the most important unique predictors for all disease.

Weakness

* 1. Computation time is long due to the higher dimensionality per dummy encoding
  2. Higher numbers of classes can introduce the lower accuracy of the models
  3. The models are not recommended for the real-life application usage without any future improvements due to the inconsistent accuracy scores

Future Plans

* + For the future research, more data with small classes should be collected to get more balanced data set. The subset models can be trained using the simpler binary classification analysis such as cancer against all other diseases, cancer versus diabetics, etc. The accuracy is expected to optimize by this strategy.

Address the problem statement:

* + Mention overfitting models
  + Not consistent accuracy by each model due to overfitting and underfitting
  + Highest AUC score, accuracy by \*\*\* model
  + Top ten predictors for all the disease (Symptom 5) revealing that it’s the important unique symptom that diagnose the specific disease.

Reasons Why models are facing these challenges:

* + 1) class imbalance favoring the majority class and ignoring the minor ones
  + 2) not all models inherently support multi-class classification
  + 3) Higher complexity due to the higher number of class (Nine classes in this study)

Suggestions/ Solutions

* + Perform up sampling / down sampling or SMOTE to balance the target classes
  + Train the optimal model with top important predictors to reduce the dimension and noise.
  + Reduce the complexity of the higher number of multiclass classifiers with One-Against-One and One-Against-All (For example, Cancer versus the rest of the disease, cancer versus Diabetics) (Sahare, M., & Gupta, H. (2012)).

# References

Clark, M. M., Stark, Z., Farnaes, L., Tan, T. Y., White, S. M., Dimmock, D., & Kingsmore, S. F. (2018). Meta-analysis of the diagnostic and clinical utility of genome and exome sequencing and chromosomal microarray in children with suspected genetic diseases. NPJ genomic medicine, 3(1), 1-10.

Kaggle Genetic Disorder Prediction Data (2022). \*.csv [Data Set]. Kaggle Inc. <https://www.kaggle.com/datasets/aibuzz/predict-the-genetic-disorders-datasetof-genomes?select=test_genetic_disorders.csv>

Sahare, M., & Gupta, H. (2012). A review of multi-class classification for imbalanced data. International Journal of Advanced Computer Research, 2(3), 160.

Wojcik, M. H., Schwartz, T. S., Yamin, I., Edward, H. L., Genetti, C. A., Towne, M. C., & Agrawal, P. B. (2018). Genetic disorders and mortality in infancy and early childhood: delayed diagnoses and missed opportunities. *Genetics in Medicine*, *20*(11), 1396-1404.

Zarocostas, J. (2006). Serious birth defects kill at least three million children a year. BMJ, 332(7536), 256.

# Appendix A

Genetic Disorder Dataset

|  |  |  |
| --- | --- | --- |
| **Column Name** | **Description** | **Purpose** |
| -Patient.Id | Unique identifier | index |
| \*Patient.Age | Age | feature |
| \*Genes.in.mother.s.side | Gene defect in mother | feature |
| \*Inherited.from.father | Gene defect in father | feature |
| \*Maternal.gene | Gene defect in maternal family side | feature |
| \*Paternal.gene | Gene defect in paternal family side | feature |
| \*Blood.cell.count..mcL. | Blood cell count (per microliter) | feature |
| -Patient.First.Name | First name | feature |
| -Family.Name | Family name | feature |
| -Father.s.name | Father's name | feature |
| \*Mother.s.age | Mother's age | feature |
| \*Father.s.age | Father's age | feature |
| -Institute.Name | Birth medical institute | feature |
| -Location.of.Institute | Location of medical institute | feature |
| -Status | Deceased status | feature |
| \*Respiratory.Rate..breaths.min. | Respiratory rate | feature |
| \*Heart.Rate..rates.min | Heart rate | feature |
| -Test.1-5 | Tests (masked) conducted | feature |
| -Parental.consent | Parent's approval for treatment | feature |
| \*Follow.up | Risk level | feature |
| \*Gender | Gender | feature |
| -Birth.asphyxia | Suffered from birth asphyxia? | feature |
| \*Autopsy.shows.birth.defect..if.applicable. | Autopsy showed birth defect(s)? | feature |
| -Place.of.birth | Place of birth (home or institute) | feature |
| \*Folic.acid.details..peri.conceptional. | Periconceptual folic acid supplementation | feature |
| \*H.O.serious.maternal.illness | Mother consequences from labor/delivery | feature |
| \*H.O.radiation.exposure..x.ray. | Radiation exposure history | feature |
| \*H.O.substance.abuse | Parent history of drug addiction | feature |
| \*Assisted.conception.IVF.ART | Parent treatment used for infertility | feature |
| \*History.of.anomalies.in.previous.pregnancies | Mother anomalies in previous pregnancies | feature |
| \*No..of.previous.abortion | Mother's number of abortions | feature |
| \*Birth.defects | Birth defects? | feature |
| \*White.Blood.cell.count..thousand.per.microliter. | White blood cell count (per microliter) | feature |
| \*Blood.test.result | Blood test results | feature |
| \*Symptom.1-5 | Symptoms (masked) | feature |
| -Genetic.Disorder | Genetic disorder | target |
| \*\*Disorder subclass | Disorder subclass | target |

- = removed feature; \* = full predictor set; \*\* = target

# Appendix B

Project Code

(Reference following page)

1. Reference Appendix B [↑](#footnote-ref-2)