A machine learning model for predicting fetal Haemoglobin levels in sickle cell disease patients

Konstantinos Oikonomou[1][a], Kathleen Steinhöfel[1][b] and Stephan Menzel[2][c]
[1]*Department of Informatics, King's College, Strand, London, United Kingdom*
[2]*Department of Molecular Haematology, King's College, Strand, London, United Kingdom*
{k1896365, kathleen.steinhofel, stephan.menzel}@kcl.ac.uk

Keywords: Sickle Cell Disease, Fetal Haemoglobin, Machine Learning Prediction Model

Abstract: Sickle cell disease is one of the commonest genetic diseases and is defined as a decrease in hemoglobin concentration in the blood. The main known factor that can alleviate the disease is the persistence of fetal Haemoglobin (HbF) and thus the aim of our research is to build a model to predict the HbF% of patients based on the 3 regulating genes of the disease (BCL11A, Xmm1-HBG2, HBS1L-MYB). A machine-learning approach is employed in order to improve the accuracy of the model, with various algorithms of that type being explored. At the end, the K-Nearest Neighbors algorithm is chosen and an initial version of it is implemented and tested. Finally, the algorithm is optimized enabling our optimized model to predict the HbF% of a patient with 87.25% accuracy, a major improvement over the existing alternative that has a mean error of 336.33%. Furthermore, 93.45% of our predictions have a sheer error that is less than 0.5 and all these facts reinforce the strength of our model as a quick and accurate estimation tool for small and medium-sized clinical trials, where fast HbF% predictions can help adjust for genetic background variability that obscures test outcomes.

1 INTRODUCTION

Sickle cell disease is one of the commonest genetic diseases with the levels of fetal Haemoglobin (HbF) being one of the most important indications of its severity. However, high levels of HbF have been proven clinically beneficial for patients and are associated with longer survival and lower pain rates per Platt (1991 & 1994) and Palkari (2018). The persistence of HbF in patients beyond the first year of life and throughout adulthood is the main known factor that can alleviate sickle cell disease. Our main objective is the development of an accurate model to predict the levels of a patient's HbF% and thus the severity of the sickle cell disease. A model of this kind could be especially powerful in smaller and medium-sized clinical trials and help geneticists explain human variation in health and disease through underlying genetic variability.

The development of a prediction model for various aspects of sickle cell disease has been a topic for several scientific researches. Adams et al. (1992) try to predict the chances of a stroke for patients of the disease, using transcranial ultrasonography, while Miller et al. (2000) focus on predicting the disease's adverse outcomes on children. Furthermore, Steinberg (2005) and Rees (2010) seeks to better define the phenotypes and genotype – phenotype relationships regarding the disease, Razak et al.(2018) try the same for β-Thalassemia Patients and Gil et al. (2004) present a pain prediction model for African American adults affected by it. However, there is little research focused on the development of an accurate prediction model specifically for the HbF% of a patient, as the majority of previous work in this particular area focuses on the general connection of gene regulation with sickle cell. Bae et al. (2012) confirm the effect of genes BCL11A and HBS1L-MYB on sickle cell anemia in African Americans while Makani et al. (2011) and Darshana

[a] https://orcid.org/0000-0001-7544-3429
[b] https://orcid.org/0000-0002-9533-4649
[c] https://orcid.org/0000-0002-1590-9108

et al.(2020) investigate the same for Tanzanian, British and Sri Lankan patients. A similar model as the one we are trying to develop was proposed by Gardner et al. (2018) and its 22% accuracy will be used as a comparison benchmark but it suggests a linear regression connection between the γ-chain genes and HbF. This is a major disadvantage in our opinion as it fails to take into account the effect of the γ-gene activation and the creation of young blood cells that also contain fetal haemoglobin (F-Cells) and both will be incorporated in our proposed approach.

In this paper, the goal is to create a prediction model for sickle cell disease patients' HbF%, optimize it and improve on the accuracy of the already published model from Gardner et al. (2018). At first, the aforementioned model is implemented to be used as a comparison metric and an accuracy target for our model to beat. Then several algorithms are tested in an attempt to choose the most appropriate one as base for our prediction model. Their results and characteristics are compared and the K-Nearest Neighbors algorithm is selected amongst them as the one indicating the highest potential accuracy. The K-Nearest Neighbors algorithm is then customised and optimised for our problem and the final implementation of our algorithm is tested against the data set. Finally, our optimised model is compared to the model proposed by Gardner et al. (2018) in terms of accuracy between the predicted and actual values of HbF%.

This paper is organised as follows. Some background information regarding sickle cell anemia and fetal Haemoglobin are discussed in Section 2. Section 3 contains a detailed description of our approach both regarding the theoretical model and the development of the algorithm. On the other hand, the actual initial and final implementations are discussed in Section 4, along with an overview of the chosen algorithm. The accuracy results of our model are presented in Section 5, accompanied by a comparison with a previously implemented model and information regarding the model's optimisation and cross validation. Finally, Section 6 includes the conclusions of our research and proposals for future work on this topic. From this point on, the abbreviations HbF for Fetal Haemoglobin and F-Cells for young blood cells containing HbF will be used for brevity's sake.

2 BACKGROUND

Red blood cells (or erythrocytes) are the most common type of blood cell. Their main function is oxygen binding and transport to the tissues as well as the return of carbon dioxide from the peripheral tissues to the lung. In order to achieve this exchange, the erythrocytes contain a specific protein called haemoglobin. Haemoglobin is a metalloprotein and an erythrocyte contains about 640 million haemoglobin molecules. Haemoglobin types found in adults are HbA (α2β2), HbA2 (α2δ2) and HbF (α2γ2). However, the primary haemoglobin during the fetal development period (2^{nd}-3^{rd} trimester of pregnancy) is HbF (α2γ2).

2.1 Polymorphism & Haemoglobin regulating genes

The expression of haemoglobin genes is regulated by factors in the same gene locus (acting in cis) or those encoded in other regions (acting in trans, i.e. transcription factors). Either type of regulatory factors can be affected by DNA sequence polymorphisms, leading to changes in haemoglobin synthesis. Some of those polymorphisms can lead to persistence of γ-globin and therefore HbF in adults. The presence of HbF is critical to the phenotype of sickle cell disease and because of this, the effort to induce HbF synthesis has begun as a therapeutic approach to these diseases as Makani et al. (2011) describe. The 3 polymorphic loci that affect the expression of haemoglobin γ-chains are *BCL11A*, *Xmm1-HBG2* and *HBS1L-MYB.* Their most representative variants are summarized in the following table.

Variant	Gene
rs1427407	BCL11A
rs654816	BCL11A
rs66650371	HBS1L-MYB
rs7482144	Xmn1-HBG2

Table 1. Representative variants.

The representative variants of Table 1. can have values 0, 1 or 2 and will be used as inputs for our prediction model.

2.2 Sickle Cell Anemia

Anemia is defined as a decrease in hemoglobin concentration in the blood. Although normal values vary across laboratories, the indicative values for the

diagnosis of anemia could be less than 13.5g / dl in adult men and less than 11.5g / dl in adult women.
Sickle cell anemia/disease is a genetic hemoglobinopathy which is inherited with an autosomal recessive mode of inheritance. It occurs in different clinical phenotypes, with a common background in the inheritance of the mutant (sickle cell) β-globin gene. This disease is one of the commonest genetic diseases with over 300,000 annual births worldwide, of which about 70% occur in sub-Saharan Africa, where most of the affected children die before the age of 5. In the UK and other Western countries, the disease is present mostly through the African diaspora and shows significant clinical diversity.

3 OUR APPROACH

A visualization of our theoretical model is the following:

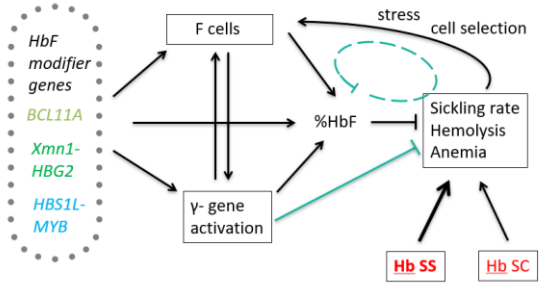

Figure 1. Model Visualization.

Our approach takes into account both the young blood cells that contain fetal haemoglobin (F cells) as well as the γ-gene activation that affect the HbF% of a patient. These 2 factors can independently affect the concentration of HbF on a patient but are in turn affected by the values of the HbF modifier genes, expressed through their representative variants. The arrows on Figure 1. represent these relationships that our model will try to explore.

Furthermore, the interaction between the F cells, the HbF% and the sickling rate forms a negative feedback loop. This happens because high concentration of HbF leads to reduced haemolysis, which is the rupturing of red blood cells. This reduction causes less need for the creation of new blood cells that would contain fetal haemoglobin and would therefore be categorized as FCells. However, the number of FCells affects the percentage of HbF and thus, a decrease in FCells leads to a decline of HbF% which can subsequently cause increased haemolysis. The aforementioned interaction has been taken into account in our approach and is a key factor for the accuracy of our model. Finally, the straight line that connects the HbF modifier genes with the %HbF represents the linear regression proposed by Gardner et al. (2018). It was implemented as a comparison tool for accuracy but is not part of our model.

3.1 Machine learning approach

A pure mathematical approach to our model was dismissed since the early development stages, as our theoretical model is very complicated and involves many unknown relationships. The efforts to try and seek the functions that represent these relationships would be extremely time consuming and possibly not realistic, as our data suggest that most of these functions could be potentially discontinuous in multiple parts. Therefore, a machine learning approach was chosen instead with a supervised algorithm being the first choice.

3.1.1 Training/Test set

The data set included 465 patients with complete data regarding the values of the representative variants, their HB and HbF levels. A fairly common strategy employed in supervised machine learning applications is to split the data set in 2 subcategories. More specifically, the training set consists of pairs of an input and an output vector. The model is then adjusted based on the results of the model's predictions against the actual output vectors that are used as targets. On the contrary, the test set's outputs are calculated by the trained model and can be used to measure its accuracy. In our case, 296 cases were used as the training set while the remaining 169 cases were the evaluation/test set. Various test runs occurred and in each, the size of the training and test sets was constant but the cases that formed the 2 sets were randomly selected in order to cross validate the model.

4 IMPLEMENTATION

4.1 Previous model accuracy

The model proposed by Gardner et al. (2018) was implemented and tested against the data set. This model is based on linear regression in order to predict the HbF% of a patient. The formula that was proposed to calculate the HbF% was

$$e^{1.89+0.14*G1+0.3*G2+0.13*G3+0.1*G4} \quad (1)$$

where G1, G2, G3, G4 are the values of the representative gene variants summarized on Table 1. An initial observation is that due to the nature of the formula, the percentage levels that are closer to the edge of the range could potentially be misrepresented. The exponent is a first-degree polynomial function and this is bound to affect the accuracy of the model. The results confirm our initial reservations regarding its accuracy, as the model has a mean error of 336.33 % over the 465 samples. This evidence further supports our initial assumption that linear regression is not the optimal method to support our model as it fails to take into account the effect of the γ-gene activation and the young F-Cells.

4.2 K-Nearest Neighbors algorithm implementation

The K-Nearest Neighbors algorithm per Keller et al. (1985) is a supervised machine learning algorithm that can be used to solve both classification and regression problems. Our problem requires a regression approach, as our algorithm's output is the percentage of HbF% of a patient which is a real and positive number. The K-Nearest Neighbors algorithm is based on the assumption that similar things exist in close proximity. It calculates the Euclidian distance

$$d(\mathbf{p},\mathbf{q}) = \sqrt{(p_1-q_1)^2+(p_2-q_2)^2+(p_3-q_3)^2}. \quad (2)$$

when trying to predict the value of a new input. In our case, the 3 dimensions are the γ-gene activation, the FCells and the HbF%. The algorithm was implemented in Java but a high-level description of it in pseudocode is on Figure 2. that follows

```
Load the data
For each example in the test set
    For each example in the training set
        Calculate the distance between the 2 examples
        Add the distance and the index of the training set example to an Array
    Sort the Array and pick the first K elements
    Find the mean of the corresponding training set values and add as predicted value
Compare the predicted values to the actual values
```

Figure 2. K-Nearest Neighbors pseudocode

The value K represents the number of "Neighbors" that will be used as comparison in our predictions. The algorithm was tested for various values of K, in order to find the optimal one, as explored in more detail on Section 5.2. In general, we can expect that as we decrease the value of K our predictions become more unstable but if we increase it disproportionally, we will begin to witness more errors in our predictions.

4.3 Model data processing

The final model was implemented in Java and employed an object-oriented programming approach. This allowed for a separate implementation of our prediction algorithm and accuracy test and led to easier manipulation of the data and cleaner maintenance and reusability for future versions of the model. The main class is called "*Case*" and contains instances of the other classes as well as useful data fields for the operations of the program. Figure. 3 illustrates the steps of the model regarding *Case*, along with its class diagram. The cells highlighted yellow represent the data fields or instances of classes that are different from the previous step.

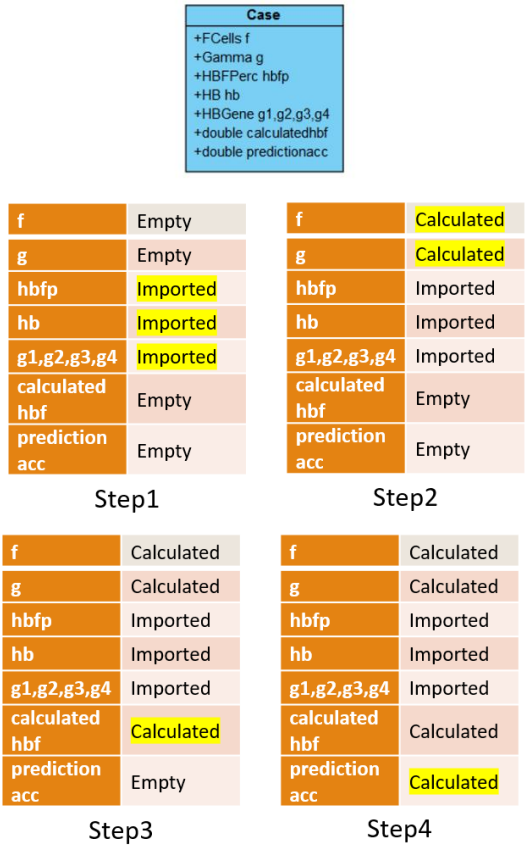

Figure 3. Case diagram & execution steps

At first, the data set is imported and the values for *hbfp* (HbF%), *hb* and *g1*, g2, g3, g4 (representative variants) are populated. The values of *Gamma* and *Fcells* are initially empty, as they are calculated based on the values of the representative gene variables and

the Hb levels of the patient during the next step. Each gene affects those values differently, through its representative variants. The application of the K-Nearest Neighbors algorithm follows, with a training set of 296 entries and a test set of the remaining 169 entries. Section 5 contains all the information regarding the choice of the number K, as well as data on test runs that were conducted with different sized training/test sets. The model produces a prediction for each entry of the test set and its value is stored in the respective instance of *Case* in the *calculatedhbf* data field. Finally, the accuracy of the predicted value is calculated and stored in the *predictionacc* data field. The mean accuracy of all predictions on the test set is calculated and the program terminates after printing that overall value for the entirety of the model.

4.4 Alternative algorithms comparison

Various alternative algorithms were implemented as base of our model and their accuracy was tested against the K-Nearest Neighbors algorithm. This was an especially important comparison as the base algorithm could potentially greatly affect the accuracy of the model. In order to have an initial metric of their accuracy, their standard deviation for our training set was compared and summarized in the following table.

Potential Algorithm	Standard Deviation
Decision Tree	7.07
Gradient Boosted Trees	5.03
Linear Regression	5.52
Nearest Neighbors	4.57
Neural Network	4.92
Random Forest	4.96
Gaussian Process	4.73

Table 2. Standard deviation of potential algorithms.

The comparison of the standard deviation presented on Table 2. is a first indicator that the Nearest Neighbors algorithm should be used as a base for our model, since it has the lowest value amongst the ones tested. In order to have a more correct comparison, the unoptimized version of the Nearest Neighbors algorithm was used, as the rest of the algorithms were also unoptimized at this stage. Standard deviation does not guarantee that the chosen algorithm will be the optimal, but graphs such as those in Figure 4. were also taken into account in the decision-making process, to ensure the best possible choice.

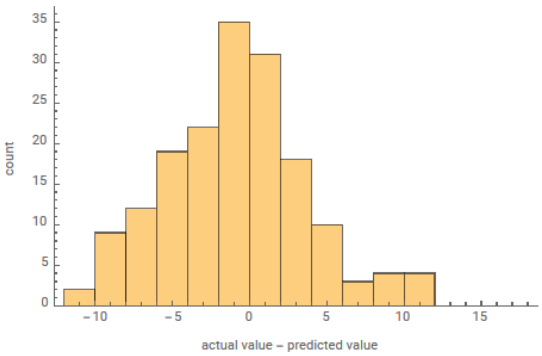

Figure 4. Actual vs Predicted output values count.

The perpendicular axis on Figure 2. represents the difference between the actual and the predicted value of HbF% on a patient while the horizontal axis the count of cases with that difference. Even in this unoptimized state the Nearest Neighbors algorithm shows promising results, with many predictions coming really close to the corresponding actual values. All the aforementioned observations led us to the conclusion that the Nearest Neighbors algorithm is the optimal algorithm to be used as base for our model.

5 RESULTS & EVALUATION

5.1 Accuracy

The mean accuracy achieved with the use of 296 entries as a training set and 169 as test set is **87.25 %.** This is an excellent result that proves the validity of our theoretical model and our assumption that both the F-Cells and the γ-gene activation greatly affect the percentage of HbF of a patient. Furthermore, the mean error of our model's predictions is just 12.75% and compares favourably to the model proposed by Gardner et al. (2018), which has a mean error of 336.33 % and an accuracy of 22%. Moreover, only 11 cases have a difference greater than 0.5% between the predicted and real HbF%. That means that 93.45% of our predictions have a sheer error that is less than 0.5. All these facts allow us to be extra confident that our model is safe to use, as there are no cases for which predictions are extremely wrong and potentially dangerous. There are 3 graphs presented below, each representing a different implementation, in order to better visualize the accuracy of the respective model.

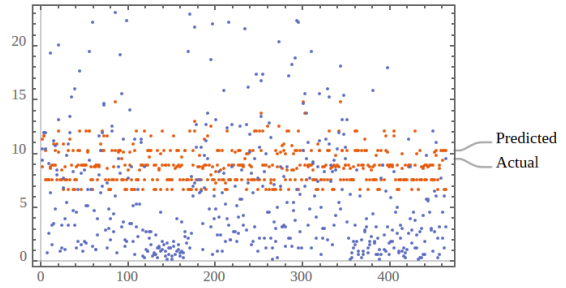

Figure 5. Linear regression.

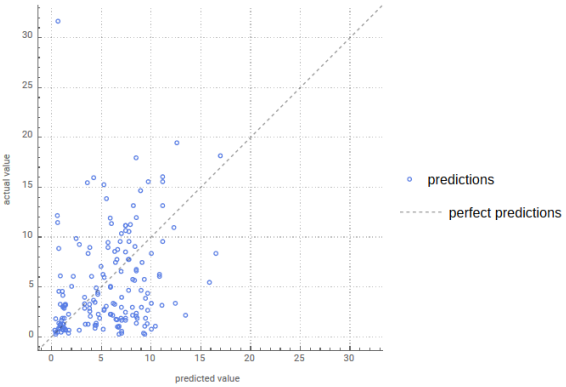

Figure 6. Initial K-Nearest Neighbors implementation.

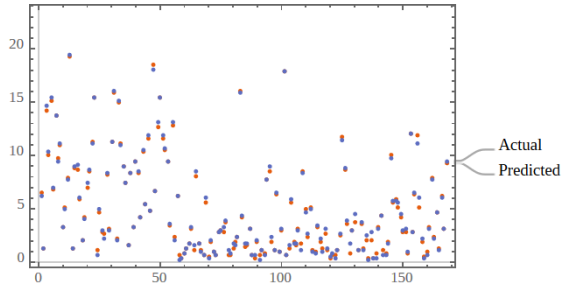

Figure 7. Optimised Nearest Neighbors implementation.

Figure 5. represents the linear regression model proposed in previous work. The horizontal axis is the index number of the patient case while the perpendicular represents the HbF% values. Red dots mark the prediction values generated by the model while blue ones show the actual HbF% values. It is obvious that this approach lacks accuracy and in many cases the difference between the predicted and the actual value is quite pronounced. Our initial implementation of the K-Nearest algorithm can be seen in Figure 6. The perpendicular axis contains the actual HbF% values while the horizontal represents the predicted ones. If a prediction is perfect it occupies a spot on the diagonal dotted line. Although far from perfect, even this initial implementation shows more promise than Figure 5. and has already several cases where the predicted value is prefect or nearly perfect. This indicates that the chosen algorithm has significant potential as base for our model. The results from our optimised model can be seen on Figure 7, with the same color-coding and axes as before. The very satisfactory accuracy of our model is apparent on the aforementioned figure as the blue and red dots belonging to the same pair are very close or even non-distinguishable from one another and the different pairs can be easily pinpointed. It is an obvious improvement over our unoptimized model as most of the cases where there was a sizeable difference between the predicted and the actual value have been eliminated and the overall mean error has been therefore greatly reduced.

5.2 K-Value evaluation

The Nearest Neighbors algorithm was tested for a range of values for the variable K in order to decide on the optimal one. This value represents the number of “Neighbors” that should be taken into account by the algorithm in determining the value of its prediction. The accuracy results for the different values of K are summarized on the following table

K Value	Mean Accuracy
1	85.35
2	86.29
3	86.39
4	87.25
5	86.43
6	85.87
7	84.84
8	84.64
9	84.36

Table 3. K Value accuracy

As mentioned before, there is an optimal value for K, which in our case is 4, while all other values that are higher or lower have decreased accuracy in their predictions. Despite the fact that 4 is an even number, there is no obstacle to choosing it, as our problem is a regression one and thus there are no tiebreakers than need resolving.

5.3 Training set size & cross validation

The algorithm was executed with varying amount of entries for the training and the test set. It was observed that increasing the number of entries for the training set resulted in improved accuracy for the predictions of the algorithm. This implies that our model can further improve its accuracy by processing a larger data set than the one provided. Still, our data set of

465 patients proved more than enough to achieve 87.25% accuracy in our predictions.

Moreover, all tests were conducted multiple times with random sampling for the training set, in order to cross validate the model and ensure the validity of its accuracy. All accuracy values presented in this paper are the mean accuracy of the respective test runs.

5.4 Execution time

The execution time of our model ranged between 1500 – 1750 ms which is a very acceptable duration for a program of this size. Increasing the data set size might affect this but the code was developed with that in mind. Typical operations that significantly affect the execution time, such as large array searches, were avoided and necessary logic was added to parts of the code that could potentially be problematic in that regard.

6 CONCLUSIONS

The objective of our research was to develop a prediction model in order to find the HbF% of patients suffering from sickle cell disease. Our aim was to explain a large proportion of HbF variability in sickle cell patients through common genetic variants and to build a polygenic score that can be calculated to predict HbF levels and, to a certain degree, disease severity.

The inputs of the model are the representative variants of 4 genes regulating the aforementioned sickle cell disease. After consecutive implementations and testing, our final model was based on the Nearest Neighbors algorithm and utilized data from 465 patients to make a regression prediction regarding their HbF%. With a test set of 169 cases, our model achieved 87.25 % accuracy in its predictions, a significant improvement from the currently implemented model by Gardner et al. (2018).

6.1 Model applications

A major advantage of our model is that the high accuracy of its results allows it to be used as a quick estimation tool for small and medium-sized clinical trials. Another advantage is the small number of inputs that it requires to predict the HbF%. This makes our model easy to implement, as those inputs (representative variants, Hb levels) can be made readily available with common blood samples from the patients. Further beneficiaries of such a system will be researchers and clinicians conducting clinical or drug trials, in Africa, the UK, and elsewhere, where this score will help to adjust for genetic background variability that obscures test outcomes.

6.2 Future work

The fact that the accuracy of our model increased as the size of the test set was incremented can motivate us to further expand our data set in the future, in order to provide an even more accurate prediction model. Furthermore, the possibility of more genes affecting the HbF% of a patient can be explored by enhancing our model in order to accommodate these extra inputs. Finally, if the size of the training/test sets increases notably, methods and techniques can be explored and implemented in order to improve the execution time of the model, if it increases disproportionately in that case.

REFERENCES

Platt, O. S., Brambilla, D. J., Rosse, W. F., Milner, P. F., Castro, O., Steinberg, M. H., & Klug, P. P. (1994). Mortality in sickle cell disease--life expectancy and risk factors for early death. *New England Journal of Medicine*, *330*(23), 1639-1644.

Platt, O. S., Thorington, B. D., Brambilla, D. J., Milner, P. F., Rosse, W. F., Vichinsky, E., & Kinney, T. R. (1991). Pain in sickle cell disease: rates and risk factors. *New England Journal of Medicine*, *325*(1), 11-16.

Paikari, A., & Sheehan, V. A. (2018). Fetal haemoglobin induction in sickle cell disease. British journal of haematology, 180(2), 189-200.

Adams, R., McKie, V., Nichols, F., Carl, E., Zhang, D. L., McKie, K., ... & Hess, D. (1992). The use of transcranial ultrasonography to predict stroke in sickle cell disease. *New England Journal of Medicine*, *326*(9), 605-610.

Miller, S. T., Sleeper, L. A., Pegelow, C. H., Enos, L. E., Wang, W. C., Weiner, S. J., ... & Kinney, T. R. (2000). Prediction of adverse outcomes in children with sickle cell disease. *New England Journal of Medicine*, *342*(2), 83-89.

Steinberg, M. H. (2005). Predicting clinical severity in sickle cell anaemia. *British journal of haematology*, *129*(4), 465-481.

Rees, D. C., Williams, T. N., & Gladwin, M. T. (2010). Sickle-cell disease. The Lancet, 376(9757), 2018-2031.

β-Thalassemia Patients

Gil, K. M., Carson, J. W., Porter, L. S., Scipio, C., Bediako, S. M., & Orringer, E. (2004). Daily mood and stress predict pain, health care use, and work activity in

African American adults with sickle-cell disease. *Health Psychology*, *23*(3), 267.

Bae, H. T., Baldwin, C. T., Sebastiani, P., Telen, M. J., Ashley-Koch, A., Garrett, M., ... & Bhatnagar, P. (2012). Meta-analysis of 2040 sickle cell anemia patients: BCL11A and HBS1L-MYB are the major modifiers of HbF in African Americans. *Blood*, *120*(9), 1961-1962.

Makani, J., Menzel, S., Nkya, S., Cox, S. E., Drasar, E., Soka, D., ... & Fegan, G. (2011). Genetics of fetal hemoglobin in Tanzanian and British patients with sickle cell anemia. *Blood*, *117*(4), 1390-1392.

Darshana, T., Bandara, D., Nawarathne, U., de Silva, U., Costa, Y., Pushpakumara, K., ... & Wijayawardena, M. (2020). Sickle cell disease in Sri Lanka: Clinical and molecular basis and the unanswered questions about disease severeity.

Gardner, K., Fulford, T., Silver, N., Rooks, H., Angelis, N., Allman, M., ... & Rees, D. C. (2018). g (HbF): a genetic model of fetal hemoglobin in sickle cell disease. *Blood advances*, *2*(3), 235-239.

Keller, J. M., Gray, M. R., & Givens, J. A. (1985). A fuzzy k-nearest neighbor algorithm. *IEEE transactions on systems, man, and cybernetics*, (4), 580-585.