Net-Migration in Relation to Incidence of Cystic Fibrosis in Ireland

MSc Research Project
Research in Computing

Fergal Bell
Student ID: X18119115

School of Computing
National College of Ireland

Supervisor: Dr Catherine Mulwa

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Fergal Bell

Student ID: 18119115

Programme: MSc Data Analytics - Research Project **Year:** 2020

Module:

Supervisor: Dr Catherine Mulwa

Submission Due Date: 17th August 2020

Project Title: Net-Migration in Relation to Incidence of Cystic Fibrosis in Ireland

Word Count: 11,066 **Page Count:** 28

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Fergal Bell

Date: 16th August 2020

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Net-Migration in Relation to Incidence of Cystic Fibrosis in Ireland

Fergal Bell
Student ID: X18119115

Abstract

It is a well-known fact that the occurrence of cystic fibrosis in Ireland is the largest in the world (approximately 27 cases per 100k of population. There is anecdotal evidence that the number of diagnoses of cystic fibrosis per annum is slowing down/decreasing to a high of 48 (2012) to 37 cases diagnosed in 2018. The "curse" of cystic fibrosis is a predominant feature of a terminal illness that has plagued Ireland since time immemorial. Indeed, my youngest daughter suffers from the disease. Hence, if there was a "tool" that could predict and or forecast new diagnoses into the future based on population growth and more importantly -net migration (whereby the nation's "gene pool" is been "diffused") would enhance healthcare planners' decision making process.
A dataset of 10,128 thousand records were with 59 variables used in the research. Various statistical techniques were employed, such as: SPSS, Excel, and RStudio The scope of this paper is to investigate whether net-migration in Ireland has any affect in the number of diagnoses.

1 Introduction

Cystic Fibrosis in Ireland has always been commonplace whereby there is approximately (on average) 40 (range 27-49) new cases diagnosed annually between 2010 and 2018 (2018 Annual Report CF Registry of Ireland, 2018). Ireland has the 'infamous' title of the highest occurrence of CF in the world with 27 PWCF (people with Cystic Fibrosis) per 100,000 of population - calculation based on 1284 PWCF (CF Ireland Annual report, 2018)[1]. There has been numerous papers/articles conducted on this subject over the last two decades from various countries, such as, America looking at socio-demographic background of CF patients (Sawaicki et al, 2007), Sweden Mental Health, (Backström Eriksson, 2020), Ireland Caregiver Burden (Suthoff et al., 2019). However, the effects of net-migration on the population of PWCF in Ireland has not been investigated to model the future diagnoses of PWCF. It is more so applicable to Ireland due to its high number of CF sufferers in proportion to its population.

[1] https://www.cfri.ie/docs/annual_reports/CFRI2018.pdf

Since the annexation of 10 countries in 2004 there has been a large influx of migrants to Ireland resulting in a net-migration average increase of 54.38% from 2006 – 2016 (Population and Migration Estimates April 2019 - CSO - Central Statistics Office, 2019)

1.2 Research Question

RQ: "*Can Net-Migration influence incidence of cystic fibrosis in Ireland and improve predictions based on (Naïve Bayes, Decision Tree Regression, Random Forest, K-Nearest Neighbours, Support Vector Machine, Kernel Support Vector Machine) of cystic fibrosis for future Incidences?".*

This research will be beneficial for stakeholders to implement infrastructure planning, staff levels, medication needs, and overall future strategies.

Sub-RQ1*: "Therefore if there are any influences emanating from net-migration. Can these prediction model(s) improve forecasting diagnoses of cystic fibrosis?"*

Sub-RQ2: *"Based on the proportionalities of PWCF in the migrant's country of origin, is it possible to factor in these measures to enhance the prediction model(s)?"*

To solve the problem and research question, the following objectives were addressed

1.3 Research Objectives: To solve the problem and research question, the following objectives are addressed.

Objective 1: A critical review of literature on Cystic Fibrosis relating to nationalities.
Objective 2: Pre-processing the data (extracting features).
Objective 3: Implementation and evaluation of cystic fibrosis prediction models.
Obj3-1Implementation, evaluation, and results of Naïve Bayes model.
Obj3-2 Implementation, evaluation, and results of Decision Tree Regression model.
Obj3-3 Implementation, evaluation, and results of Random Forest model.
Obj3-4 Implementation, evaluation, and results of K-Nearest Neighbours model.
Obj3-5 Implementation, evaluation, and results of Support Vector Machine model.
Obj3-6 Implementation, evaluation, and results of Kernel SVM model.
Objective 4: Comparison of developed models (objective 3)

2 Literature Review

2.1 Introduction

This literature review aims to investigate what other research has been done in relation to the prevalence of cystic fibrosis over a period of time to what has been done to gain new knowledge, any gaps and furthermore any limitations that has been identified. This section is divided into several sub-sections, such as: (i) Literature review history of Cystic Fibrosis, (ii) Literature review on incidence of cystic fibrosis and Identified Gaps, (iii) Survey of the Prevalence of Cystic Fibrosis in the European Union and how such traits were investigated, (iv) Limitations of research

2.2 Literature Review on the History of Cystic Review

Over the last number of decades there has been numerous articles, theses, journal papers, etc conducted on the prevalence of Cystic Fibrosis in various countries. Some, using measures per capita/per 100k of population. For example, Farrell et al. (2007) used sweat test statistics[2] to find incidences [cystic fibrosis] in relation to births from 2001 to 2003 – which varied from year to year, although the average was 1:1353. As a caveat to the latter Devaney et.al (2002). estimated the incidence as 1:1461[3]. As Farrell et.al (2007). have stated:

> "Ireland's incidence is the highest among Western European nations and is much higher than North American Nations. This is probably attributable to a high CFTR mutation prevalence and consanguinity" [relating to or denoting people descended from the same ancestor] (Concise Oxford English Dictionary, 2006).

As sweat tests were the common indicator of Cystic Fibrosis before July 2011 in Ireland new-born screening also known as 'the heel prick test' is utilized for all newborns in Ireland[4]. This gives a better outcome for all newborns diagnosed with CF from birth[5] in which a high-energy regimen can be started immediately. Delay in diagnosing Cystic Fibrosis has been shown Steinraths, Vallance and Davidson (2008:882) without new -born screening 1:15 CF patients die without being diagnosed. The above has been strengthened by M. Stern et. al. (2014: S45) stating: that follow-up of well-defined groups young patients diagnosed by newborn-screening, gives the best chance of treatment. The above test also has the effect of updating data in real-time providing not only a unique opportunity of quality improvement for CF sufferers it gives a detailed profile of the CF patient, such as: nationality, age, type of CF mutation, etc. M. Stern et. al. (2014: S45)

2.3 Literature Review on incidence of Cystic Fibrosis and any Gaps Identified.

Jackson and Goss stated that information has been monitored via registries over 60 years – see figure 1 - although Ireland's registry has only been in existence since 2001[6].
One of the most comprehensive registries is the U.S. Foundation Patient Registry (CFFPR) it is by far the oldest registry in existence (established in 1966) with over 350 variables captured – table 1, Summary of cystic fibrosis registry information (Jackson, Goss, 2017, p.2) gives an overview of all the known CF registries worldwide, it is estimated that between 70,000 and 100,000 people have CF worldwide, with varying odds of carrying the defective gene based on race and genetics[7] - see tables 1&2.

---

2 https://repository.rcsi.com/articles/Diagnosis_of_cystic_fibrosis_in_the_Republic_of_Ireland_epidemiology_and_costs_/10782233/1

3 https://onlinelibrary.wiley.com/doi/full/10.1034/j.1399-0004.2003.00017.x

4 https://www.cfireland.ie/about-cf/newborn-screening-for-cf

5 https://www.hse.ie/eng/health/child/newbornscreening/newbornbloodspotscreening/information-for-professionals/conditions/cf/

6 https://www.cfri.ie/index.php

7 https://www.healthline.com/health/cystic-fibrosis-facts

Mehta et al. observed CF registries either maintained high coverage of the CF populations in their respective countries, varying from 81-84% in the United States to the highest CF patient coverage in the United Kingdom – 99%. These registries gathered small numbers of highly accurate quantitative demographic data fields or sought to capture many clinically relevant fields – data variables ranged for each registry from 47 [Europe & New Zealand] to greater than 350 [United States]. However, there are gaps in the datasets of most of the registries maintained worldwide – i.e. the coverage of CF patient covered by the registries varies from 81% to 98% due primarily to patients who do not consent to participate in the registry (Knapp et al., 2016, p.1176).
The above data - quality limitation could be mitigated by cross-referencing patients' medical records with the registry depending on variables measured. (Knapp et al., p.1176). However, this method is both time consuming and can be inaccurate.

According to John Hopkins, the risk of certain ethnicities carrying the faulty gene is:

Table 1 Probabilities of carrying faulty CF gene based on ethnicity

Ethnicity	Probability
Caucasians	1 in 29
Hispanics	1 in 46
African Americans	1 in 65
Asians	1 in 90

Note in Ireland 1:19 people carry the faulty gene, which gives a risk of having a child born with cystic fibrosis of 1:1400.

Table 2: The risk of having a child born with cystic fibrosis

Ethnicity	Probability
Caucasians	1 in 2,500 to 3,500
Hispanics	1 in 4,000 to 10, 000
African Americans	1 in 15,000 to 20,000
Asians	1 in 100,000

Figure 1. Countries with a cystic fibrosis registry – source: Jackson, Goss, 2017, p.2

Table 3: Cystic Fibrosis Registries across the globe

Cystic Fibrosis Registries[8]

Country	Registry Name	Year Established	CF Patient Coverage	Data Variables
Australia	Cystic Fibrosis Federation Australia	1998	90-95%	200
Belgium	Belgian Cystic Fibrosis Registry	1998	>90%	200
Brazil	Brazilian Registry of Cystic Fibrosis	2003	82%	
Czech Republic	Cystic Fibrosis Registry of the Czech Republic	2004	>95%	230 approximately
Denmarkc	Danish Cystic Fibrosis Patient Registry	2000	99%	26 core dataset – some variables collected at one of the centres
Europe	European Cystic Fibrosis Society (ECFSPR)b	2008	Varies from country to country 88%	47
France	Registre Francais de la Mucoviscidose	1992	88%	400
Germany	Deutsches Mukoviszidose-Register	1995	>95%	>200
Italy	Italian National Cystic Fibrosis Registry (INCFR)	1988	93%	120
New Zealand	Cystic Fibrosis New Zealand	1968	96%	47
Russian Federation	Russian Cystic Fibrosis Registry	2011		
The Netherlands	Dutch Cystic Fibrosis Registry	2007	98%	150
United Kingdom	Cystic Fibrosis Trust	1996	99%	>250
United States	**Cystic Fibrosis Foundation Patient Registry (CFFPR)**	**1966**	**81-84%**	**>350**

a Countries contributing data to the 2017 ECFSPR Annual Report included: Albania, Armenia, Austria, Belgium, Bulgaria, Croatia, Czech Republic, Denmark, France, Germany, Greece, Hungary, Ireland, Israel, Italy, Latvia, Lithuania, Luxembourg, Republic of Macedonia, Republic of Moldova, The Netherlands, Norway, Poland, Portugal, Romania, Russian Federation, Serbia, Slovakia, Slovenia, Spain, Sweden, Switzerland, Turkey, Ukraine, United Kingdom.
b Various countries have their own registries others have individual contributing centres.

Although, there is a wealth of data available not all datasets are consistent in the amount of data variables each registry accumulates (see table 3). This limitation would affect data analyses depending on which features the analyses is targeting, such as nationality which would have a bearing in this study - i.e. the origin of birth of the CF patient or the origin of the parents nationality is a vital component of this research. Also, date of birth is important –

[8] https://www.ecfs.eu/sites/default/files/general-content-images/working-groups/ecfs-patient-registry/ECFSPR_Report2017_v1.3.pdf

especially if the CF patient was born within the timeframe of 2004 to present [10 countries annexed to the European Union in 2004 – an increase of 74 million persons who are free to travel and work within the European Union – and of more significance to this research the entitlement to live and raise a family in the destination country of choice].[9]

2.4 Prevalence of Cystic Fibrosis in the European Union and how such traits were investigated and Gaps Identified

Cystic fibrosis is most prevalent in Europe, North America, and Australasia (Australia, New Zealand).[10] Urban settings in European countries are adding to the diversity of the gene pool due to non-European populations. This effect is changing the CF inheritance since the CF causing mutations differ from those found in the original European population (Mehta et. 2007). A case in point would be the relatively recent European migration crisis [also known as the refugee crisis] were over one million refugees arrived between 2015 – 2016 - where most of the migrants originated from Syria and Afghanistan[11].

Research has shown that the mean prevalence of CF in Europe is 0.7367 per 10,000 of population (Bobadilla et. al, 2004) – outlier being Republic of Ireland 1 in 19 of the population carrying the defective gene.

However, by using 2017 data[12] it can be shown from registered CF patients collected by European Cystic Fibrosis Society (reported number) that the incidence prevalence has increased slightly from 0.7367 per 10,000 of population to 0.8223 per 10,000 of population[13]. This may be due to the substantial increase in overall European population [2017 population: 511.4 million as compared to 456.5 million - excluding Romania 19.64 million & Bulgaria- 7.102 million (joined EU 2007), Croatia - 4.13 million joined 2013)][14].

2.4.1 Traits Investigated

Mutations: A Review of Incidence and Prevalence and Net-Migration

Worldwide, the ΔF508 allele, which presumably arose from a single origin, ranks as the most common CFTR mutation [Cystic Fibrosis Genetic Analysis Consortium (CFGAC), 1994]. Although ΔF508 is the most prevalent mutation -depending on geographic location – Middle Eastern groups and Jewish populations. (Bobadilla et. al, 2002). There are 36 most common mutations found in populations of European origin[15]. Although the most common mutations

---

9 https://www.cairn-int.info/article-E_POPU_402_0361--the-european-union-at-the-time-of- enlarg.htm

10 http://www.bmbtrj.org/article.asp?issn=2588-

9834; year=2017; volume=1;issue=2;spage=105;epage=112;aulast=Mirtajani

11 https://en.wikipedia.org/wiki/United_Nations_High_Commissioner_for_Refugees

12 https://www.ecfs.eu/ecfspr

13 Method: $\frac{42{,}084}{511.8\times10^{6}} \times 10{,}000 = 0.8223\ to\ 4\ d.p.$

14 https://ec.europa.eu/eurostat/documents/2995521/8102195/3-10072017-AP-EN.pdf/a61ce1ca-1efd-41df-86a2-bb495daabdab

15 https://devyser.com/products/cystic-fibrosis/devyser-cftr/?gclid=Cj0KCQjw-r71BRDuARIsAB7i_QM8eLXy9agz-_sae1IjM9xyUWXGGynjam_1CC9RlDxfhkbD3-pQ-dsaAlCQEALw_wcB#devyser-cftr-uk

number, 5 distinct groups based on geographic location. They are: Fdel 508, G542X, N1303K, G551D, and W1282X (Mateu et.al, 2002). These five mutations are found throughout Europe, although their distribution shows clear geographic patterns: Delta F508 shows a northwest-to-southeast gradient, with a maximum (87.2% of all CF chromosomes) in Denmark and a minimum (21.3%) in Turkey (Gradient of distribution in Europe of the major CF mutation and of its associated haplotype, 1990). G542X is common in Mediterranean countries and is present in most of Europe, being most frequent (16.7%) in the Balearic Islands (Estivill et al. 1997). N1303K is present around the Mediterranean, and it reaches its highest frequency (17.2%) in Tunisia (Estivill et al. 1997). Mutation G551D is common in north-western and central Europe, but it is common in other parts of Europe (Estivill et al. 1997). Finally, mutation W1282X is common in most Mediterranean countries, reaching its highest frequency (36.2%) in Israel (Estivill et al. 1997). In relation to mutation G551D, is most common in the West of Ireland with a 6.9% prevalence rate in Ireland (Cashman et al., 1997).

An interesting finding in the diversity/number of mutations in the EU was in relation to Belgium. According to Bobadilla et al there are 27 mutations present in the Belgian population [the highest rate in the European Union] (Bobadilla, Macek, Fine and Farrell, 2002). This prevalence of so many different mutations in Belgium may be linked to the high ethnic diversity of the Belgian population. In 2007 12.9% of the population were of foreign descent – 6.5% from other EU countries and 6.4% from countries outside the EU. This statistic has increased to 25% of the population in 2012[16]. The above may have some bearing on the increase in mutation variants of CF in the Belgian CF registry report 2016 (Wanyama and Thomas, p.40, 2018)[17]. In this report it showed there was an increase of 27 (Bobadilla et al, 2002, p.580) to 39 mutations. Bobadilla et al have shown not only Belgium, but Spain, Bulgaria, Greece, and Turkey have the most diverse mutational arrays in Europe on average 25 (95% CI, 17.7-32.3) mutations accounted for 84% (95% CI, 77.9-90.1) of the CF alleles[18]. This follows the geographical location of these countries and more importantly the historical movement of peoples in historical times. These countries were used as "gateways" into the European continent either to avoid conflict or for economic reasons. A case in point is the so-called Celtic gene G511D allele associated with populations of a Celtic descent, and is found prominently in Ireland, Brittany in France and interestedly in the former Czechoslovakia – Czech Republic 3.8% (Bobadilla et.al, p.591, 2002). This anomaly can be attributed to the expansion of The Roman Empire whereby most peoples of Celtic origin fled to Ireland, UK, and Northern France. However, there was one tribe (the Boii) stayed north of the Alps and assimilated with the incoming populations (James.S, 1993). This region became to be known as Bohemia which encompasses part of South-East Germany, Northern-Austria, and Czech Republic. One of the potential limitations/gaps identified was the testing process needed to identify all/ most of the mutations of the CF gene that can occur – at the latest count there are 2088 different mutations of this gene[19]. Therefore, the neonatal screening rolled out in all countries needs to have a very-high sensitivity whilst avoiding exclusions of minority populations. This bias can be shown in mutations that do not show up in the usual sweat testing screening. A case in point is the Hispanic population in the United States, where the mutation that causes CF in this cohort 3849+10kbC -T does not react to the traditional sweat test.

---

[16] https://en.wikipedia.org/wiki/Demographics_of_Belgium

[17] https://www.sciensano.be/sites/www.wiv-isp.be/files/report_belgian_cf_registry_2016_en_final_1.pdf

[18] https://www.youtube.com/watch?v=pv3Kj0UjiLE

[19] http://www.genet.sickkids.on.ca/cftr/StatisticsPage.html

2.5 A Review of Incidence and Prevalence of Cystic Fibrosis due to Genetic Drift and Net-migration

2.5.1 Genetic Drift

Genetic drift[20] is the mechanism of evolution in which the frequencies of alleles change/disappear over time in a population due to chance. How these occurrences are triggered come from several originators, as follows:

1. Genetic drift occurs in all population sizes, but its effects are more prominent in small population sizes
2. Genetic drift can have major effects when its population size is reduced, such as natural disasters, major conflict – known as the 'bottle neck effect'
3. Deliberate interference with gene selection – eugenics is a case in point (practiced by the NAZI's during WW2)[21]
4. Founder effect: this effect has more to do with colonisation than catastrophe, whereby, a small group of individuals breaks off from a larger population to find a new colony

In relation to this study genetic drift does not seem to reduce the prevalence of CF alleles, but actually seems to increase slightly the most common allele F508del in Ireland from 72% (Cashman et al., 1995) to 73.8% in 2018[22] and a noticeable increase in the 'Celtic gene' from 6.9% (Cashman et al., 1995) to 8.3% in 2018[23].

2.5.2 Net-Migration

As a follow on from the above comments - genetic drift is related to migration of peoples, but only has a real impact if the population is small. Since the annexation of 10 new countries into the European Union there has been increase of net-migration of 332,400 people up to 2019 – the great majority originated from Poland.[24] It is interesting to note that the prevalence of CF in Ireland is approximately 14 times higher than the prevalence of CF in Poland.

2.6 Literature Review on Prediction of Genetic Disease using Machine Learning

There has been quite a lot of research done on genetic disease prediction via machine learning techniques, for instance Schrodi et.al., 2014 used GRS [Genetic Risk Scores] published on the G.W.A.S. website[25]. Schrodi et al. used a predictive model based on the sum of predisposing genotypes that each individual carrier, either weighted or unweighted by the effect size. For each genome they used SNPS[26]. For instance, Ripatti et al. developed a genetic risk based on

---

20 https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/genetic-drift-founder- bottleneck

21 https://en.wikipedia.org/wiki/Nazi_eugenics

22 https://cfri.ie/annual-reports/

23 https://cfri.ie/annual-reports/

24 https://www.cso.ie/en/releasesandpublications/er/pme/populationandmigrationestimatesapril2019/

25 https://www.genome.gov/genetics-glossary/Genome-Wide-Association-Studies

26 Single nucleotide polymorphisms, frequently called SNPs (pronounced "snips"), are the most common type of genetic variation among people. Each SNP represents a difference in a single DNA building block, called a nucleotide.

13 SNPs associated with coronary heart disease (Ripatti et al., 2010). The formula used to predict genetic risk scores is as follows:

$$GRS = \sum_{i}^{k} w_i R_i$$

Where k is the number of SNPs, is the number of risk alleles at the ith SNP, is the β weights[27] from the GWAS. Schrodi also employed regression methods for constructing prediction models for both dichotomous and quantitative traits. Schrodi, stated that regression models are still commonly used for disease prediction. A cursory glance search of PubMed showed 20 articles published using logistic regression methods in 2019 for a variety of genetically based diseases from Alzheimer's, COPD, Lung cancer to heart disease. Schrodi used 6 different machine learning methods (Schrodi et al., p13, 2014) on a genetic disorder -Inflammatory Arthritis – which can be applied to all genetic-based diseases, such as Cystic Fibrosis. It showed that Naïve Bayes exhibited slightly higher CV accuracy [10-fold Cross-Validation] when compared to other algorithms. Kurgan et al (2005) used neural networks and SVM to predict genotypes associated with Cystic Fibrosis achieving an accuracy of 73.1%. Cattellani et al (2005) use GLM [General Linear Model] to predict incidence of CF via screened neonates in Italy, although no accuracy was given – 100% of the dataset available was used. Capriotti et al (2006) used a novel approach in predicting 'insurgence' of human genetic diseases associated to SNP's (genetic variation) achieving an accuracy of 74%. Although, the variables/features used are not exactly the target variable (Incidence of Cystic Fibrosis) proposed to use in this project – most, if not all the techniques can be implemented in this project.

Table 3: Comparison of Literature in Prediction of Cystic Fibrosis Incidence

Features/Variables Extracted	**Classifiers and Techniques**	**Comparisons of Results (Accuracy)**	**Software Used**	**Authors**
SNP's[28][29], GRS[30]	Logistic Regression, Naïve Bayes, Neural Networks, SVM, Random Forests	77.9% 10-fold cross validation	Weka	Schrodi et al, 2014
Genotype	Neural Networks, SVM	73.1% 10-fold cross validation	MetaSqueezer	Kurgan et al, 2005
Screened Neonates	GLM	NA	R Studio	Castellani et al, 2009

[27] https://www.statisticshowto.com/beta-weight/#:~:text=A%20beta%20weight%20is%20a,is%20a%20single%20predictor%20variable.

[28] A single nucleotide polymorphism, or SNP (pronounced "snip"), is a variation at a single position in a DNA sequence among individuals.

[29] https://www.nature.com/scitable/definition/snp-295/

[30] https://www.nature.com/articles/d42473-019-00270-w

SNP's	Decision Tree	74% accuracy at predicting mutations	BLASTClust	Capriotti et al, 2006

2.7 Conclusion

In reference to the literature reviewed and identified gaps there is a glaring lack of CF incidence projections based on net-migration into Ireland or any other country for that matter to plan future healthcare needs. There is evidence from Italy that new-born screening may have an influence on birth rate due to couples not having children or stopping to have children based on one of their offspring being diagnosed with cystic fibrosis – however, it is a different feature than being investigated in this project. The next chapter will outline the scientific methodology used to develop future projections of Cystic Fibrosis incidence.

This chapter has solved objective 1, in introduction chapter.

3. Scientific Methodology Used

3.1 Introduction

In relation to the substance of the project and the data mining techniques which will be used to garner insights in the overall project KDD [Knowledge Based Discovery] is best suited to modelling this project. Although, Crisp-DM [Cross-industry standard process for data mining] has some features that could be related to the business 'goals' of this project – enhancing healthcare management of clinical needs, so an altered KDD model can be used here with a two-tier structure – client tier - H.S.E and business logic tier.

3.2 Cystic Fibrosis Methodology Used

The cyst Fibrosis methodology approach used (see figure 2) for ethnicity(net-migration) effects on incidence of Cystic Fibrosis in Ireland consisted of the following stages: (i) data selected from annual cystic fibrosis reports from 2008 to 2016, this consisted of manually inputting data from the reports and creating a spreadsheet of 10,189 rows and 56 variables – a total of 570,854 data points, (ii) all data was pre-processed and normalised using statistical techniques, (iii) data was inputted into RStudio, RapidMiner, and SPSS for analysis, (iv) Naive Bayes, SVM, KNN, Neural Networks, Random Forests, decision trees were trained on the data, (v) models were evaluated and interpreted based on accuracy of the models used.

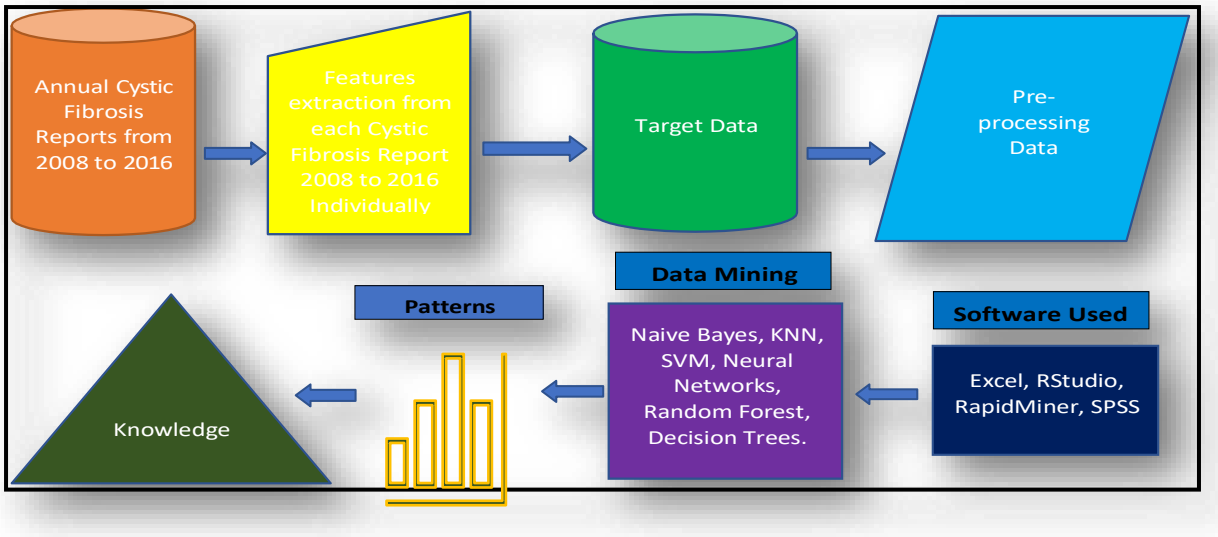

Figure 2: Methodology of Ethnicity Effects on Cystic Fibrosis in Ireland

3.3 Project Design Flow

The project paradigm process is presented in figure 3 of ethnicity/net-migration on CF incidence in Ireland and other influencing factors, feature extraction, cleansing and normalising data, application of various data mining techniques in tier 2. Tier 1 presents the data in visualised form using RStudio, and other software.

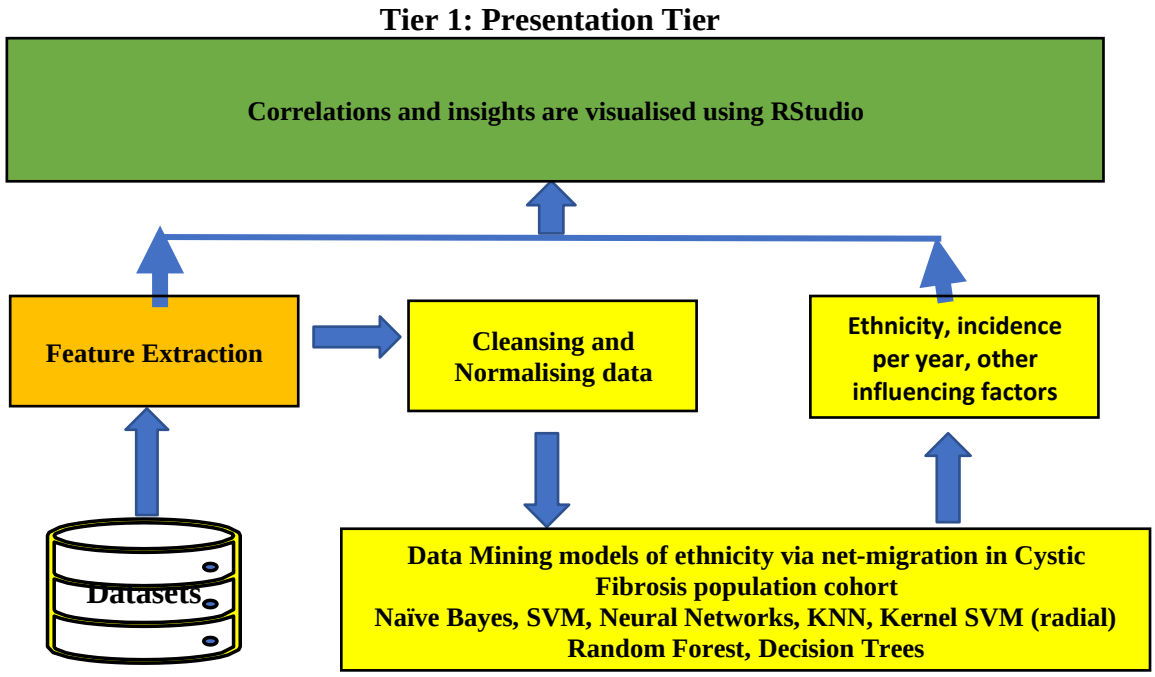

Figure 3: Project Paradigm Process Flow effects of ethnicity/net-migration on CF incidence in Ireland

3.4 Conclusion

A hybrid version of Cystic Fibrosis methodology approach was used in this project both for the research and needs of this project. As can be seen in figure 3 the process flow commenced by extracting data from 10 years of annual reports from the Cystic Fibrosis Registry of Ireland. A two-tier architecture was then used in this project. In chapter 4 implementation, evaluation, and results of the various models will be explored to elicit any factors that influence CF incidence in Ireland primarily focussing on net-migration.

4 Implementation, Evaluation and Results of Cystic Fibrosis Prediction Models

4.1 Introduction

The implementation, evaluation and results of models implemented are explained in this section. All statistical techniques, feature extraction, algorithms, mathematical models are described in detail. The evaluation of the models is based on an accuracy metric via confusion matrices. The best-performing model was chosen at the end of this section by comparing the accuracy of each model.

4.2 Extracting Features from the Dataset

Feature extraction is an important and invaluable method not only to improve accuracy of the models, the "soundness" of the results based on accurate data, excluding data bias via multicollinearity of the independent variables in the dataset. To identify which variables to extract and which variables to exclude a Principal Component Analysis (PCA) was conducted in SPSS version 26. The target variable chosen was Incidence instead of Incidence of CF Yearly due to more accuracy attained using Incidence Rate[31] of CF based on population sizes from 2008 to 2016.

The principal component analysis was run on 59 variables (apart from Severity of Disease – which was a categorical variable unsuitable for PCA analysis) of the dataset, this dataset composed of 10,128 rows of data (597,552 individual datapoints). Inspection of the correlation matrix showed that all variables analysed had at least one correlation greater than 0.3. PCA revealed 7 components greater than 1 and which explained 64.14%, 10.91%, 6.81%, 5.12%, 3.74%, 3.44%, and 3.06% of the total variance, respectively. Visual inspection of the scree plot indicated that 7 components should be retained. The seven-component solution explained 97.21% of the total variance. However, upon more thorough analysis the initial PCA analysis on 59 variables did not produce a KMO Bartlett Test table due to negative eigenvalues[32]. On further analysis the dataset was reduced to 18 variables although this number of variables did create KMO Bartlett Test Table, the Kaiser-Meyer-Olkin values was exceptionally low which would indicate poor factorability between variables. Notwithstanding, by reducing the number of variables from 18 to 9 by process of elimination a Kaiser-Meyer-Olkin of 0.572 was achieved, and Kaiser(1970, 1974) and Bartlett's Test of Sphericity (Bartlett 1954) reached statistical significance, supporting the factorability of the correlation matrix. (Pallant, J, 2016) The second principal component analysis uncovered the presence of two components with eigenvalues exceeding 1, explaining 41.24% and 38.34% of the variance, respectively. After

[31] https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html

[32] https://www.ibm.com/support/pages/factor-does-not-print-kmo-or-bartlett-test-nonpositive-definite-matrices

inspection of the scree plot (see figure 4) it was decided to keep three components for further investigation. The total variance explained by these components was 90.53%

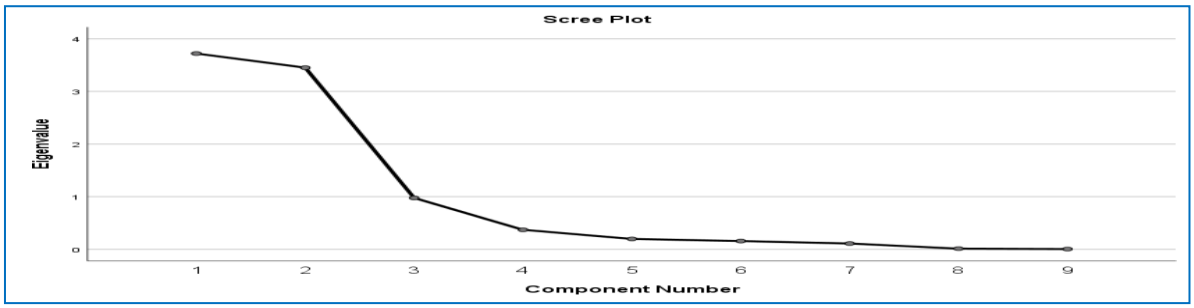

Figure 4: Scree plot of 9 variables of the dataset

Table 2: KMO & Bartlett's Test

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		**.572**
Bartlett's Test of Sphericity	Approx. Chi-Square	**140276.815**
	df	**36**
	Sig.	**.000**

4.3 Multiple and Linear Regression of the Dataset

Factor analysis, multiple regression, linear regression using SPSS and RStudio showed consistent correlations between incidence of Cystic Fibrosis and age groups, especially the age group10 to 14 years of age with a consistent r correlation of 0.810 to 0.870. Figure 5: Feature of Incidence versus Population from 2008 to 2016 (Linear Regression). As can be gleaned from figure 5 linear regression does not give a good model for prediction (blue line) incidence of CF from 2008 to 2016.

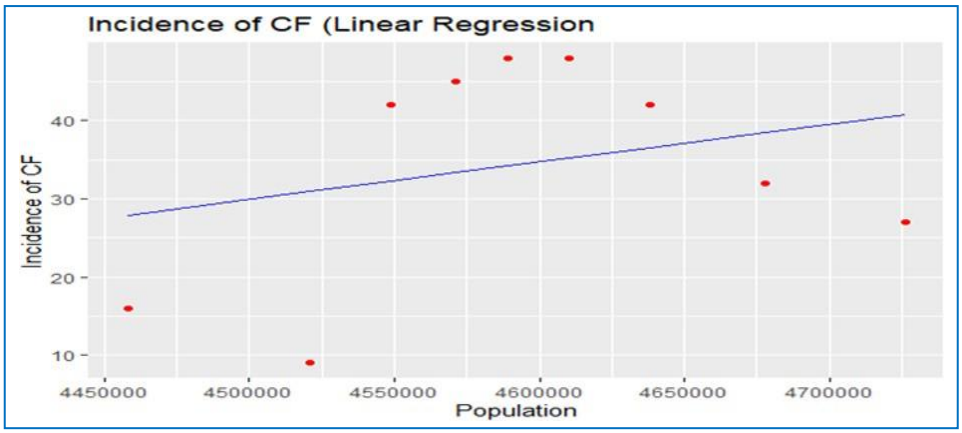

Figure 5: Feature of Incidence versus Population from 2008 to 2016 (Linear Regression)

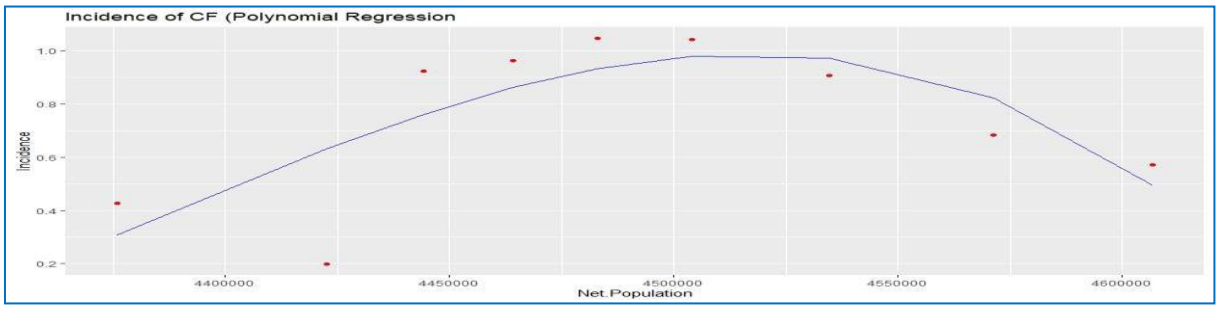

Figure 6: Feature of Incidence versus Population from 2008 to 2016 (Polynomial Regression)

Figure 6 in comparison of using polynomial regression produces an accurate prediction of incidence from 2008 to 2016 apart from an outlier at x = 4,422,600, y = 0.623 which is 0.199 or incidence of 9 diagnoses for 2009. However, the value for 2009 which was reported in the Cystic Fibrosis Registry of Ireland Annual Report[33] for 2009 was misleading/inaccurate, hence the outlier.

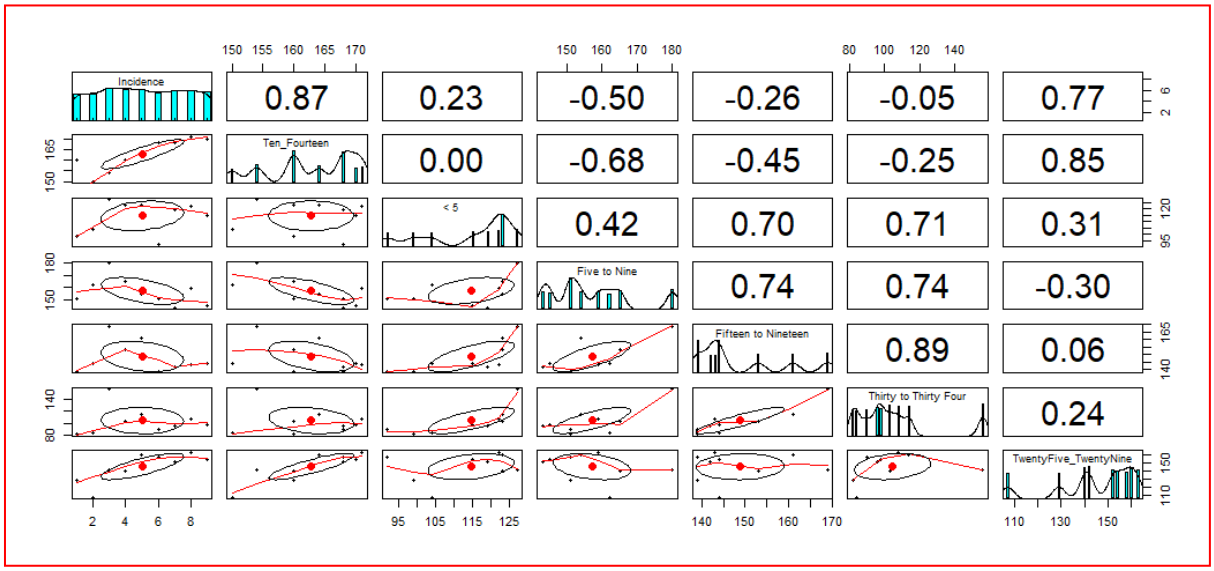

Figure 7**:** Correlation Matrix of Incidence vs age groups ten to fourteen to thirty-four to thirty-nine

Figure 7 demonstrates some interesting correlations between incidence of CF. Again, as stated earlier age 10 to 14 years has a high correlation of 0.87. Age groups less than 5, 15 to 19, and 30 to 34 show low correlations. In respect to age groups 5 to 9 a negative correlation of -0.50 is achieved. Age group 25 to 29 shows a strong correlation of 0.77.

[33] https://cfri.ie/wp-content/uploads/2020/03/CFRI2009.pdf

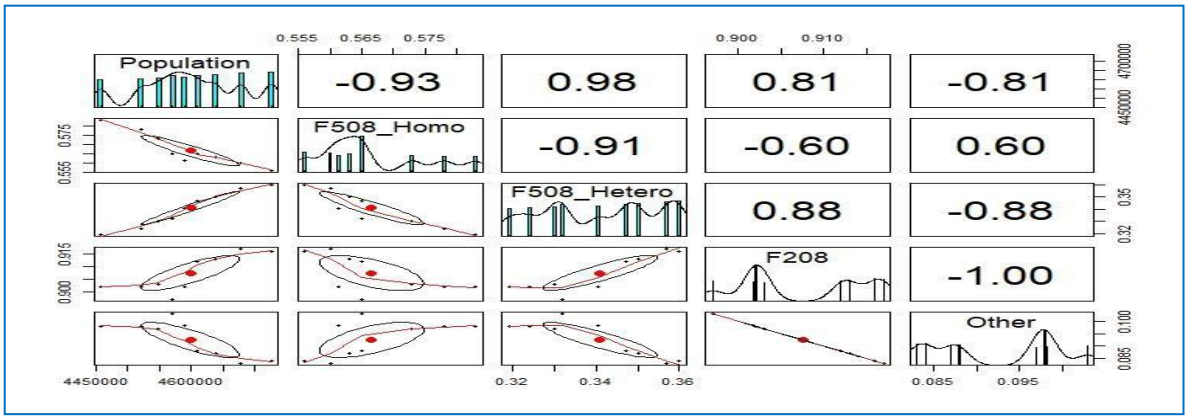

Figure 8**:** Population versus Genotype from 2008 to 2016

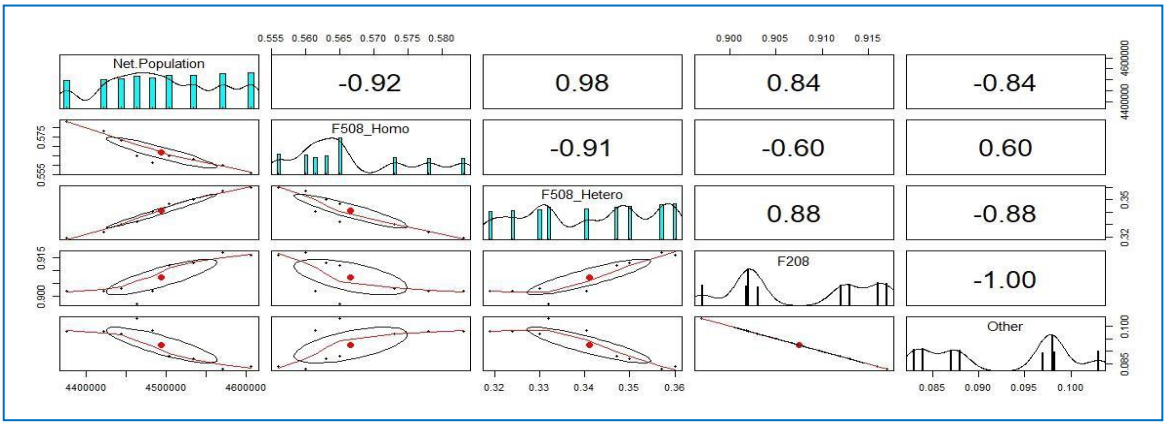

Figure 9: Net-Population versus Genotype from 2008 to 2016 (Excludes Net- Migration)

Correlation matrices figure 8 and 9 compares different frequencies of cystic fibrosis genotypes. It shows slight variations for F508 Homo, F208, and other genotypes between net-population (excluding net-migration) growth and population (including net-migration). This could be a result of mixture of population migration of other ethnicities via genetic drift[34]. These hypotheses will be explored in the discussion section.

4.4 Implementation, Evaluation and Results of Decision Tree Regression

Decision Tree Regression is based on numeric data given a real number outcome. Model was executed using a split ratio of 0.80 i.e., 80% of train data and 20% of test data. Due to high multicollinearity between feature variables which was identified in SPSS it was paramount that the feature variables were chosen carefully, otherwise erroneous results would be the outcome. The following predictors were used (Age Groups) individually (2, 3, 4, 5, 6, 7, 8, 9), genotypes (53, 54, 55); population (59).

Implementation: Decision Tree was implemented using "rpart package"[35] from library in RStudio. The regressor was assigned with the rpart function, which was applied to the training set. The model was applied to age groups: less than 5; 5 to 9; 10 to 14; 15 to 19; 20 to 24; 25 to 29. Also, genotypes were applied to the target variable – incidence.

[34] https://en.wikipedia.org/wiki/Genetic_drift

[35] https://cran.r-project.org/web/packages/rpart/rpart.pdf

Evaluation and Results:
The accuracy achieved by Decision Tree for each age group feature is as follows: less than Five years of age, 88.30%; Five to Nine years of age, 89.83%; Ten to Fourteen years of age, 79.52%; Fifteen to Nineteen years of age, 79.96%; Twenty to Twenty Four years of age, 77.39%; Twenty Five to Twenty Nine, 100%; Thirty to Thirty Four, 100%; Thirty Five to Thirty Nine, 89.83%; Forty to Forty Four, 87.96%; Forty Five to Forty Nine, 89.83%; Greater than Fifty Years of age, 78.28%. Genotypes: F508_Homo, 88.8%; F508_Hetero, 100%; Other, 90.13%. Highest accuracy was achieved for the age groups Twenty-Five to Twenty-Nine, Thirty to Thirty-Four, 100% respectively. Lowest accuracy was for age group Twenty to Twenty-Four, 79.52%. For genotypes: F508_Homo 88.79% was achieved by including the variable "Age", 0.59% improvement on accuracy. By including "Year of Birth", achieved an accuracy of 90.33%, increasing the accuracy by 1.53%. F508_Hetero[36] achieved an accuracy of 100%, such a high accuracy could be explained that the combined heterozygosity of PWCF in Ireland according to the 2016 annual report of CF Registry of Ireland[37] page 12, was 91.6%. Other genotypes; 90.13%. This regressor performed well on identifying the most relevant variables.

4.4 Implementation, Evaluation and Results of Random Forest

Random Forest is a collection of individual decision trees together. Model was executed using a split ratio of 0.80 i.e., 80% of train data and 20% of test data. Due to high multicollinearity between feature variables which was identified in SPSS it was paramount that the feature variables were chosen carefully, otherwise erroneous results would be the outcome. The following predictors were used (Age Groups) individually (2, 3, 4, 5, 6, 7, 8, 9); genotypes (53, 54, 55); population (59)
Implementation: Random Forest was implemented using "randomForest" library in RStudio. The classifier was identified with the function randomForest, which was applied to the training set. The model was applied to age groups using 100 trees: less than 5; 5 to 9; 10 to 14; 15 to 19; 20 to 24; 25 to 29. Also, genotype was applied to the target variable – incidence.
Evaluation and Results:
The accuracy achieved by Random Forest for each age group feature is as follows: less than Five years of age, 88.30%; Five to Nine years of age, 89.83%; Ten to Fourteen years of age, 79.52%; Fifteen to Nineteen years of age, 79.96%; Twenty to Twenty Four years of age, 77.39%; Twenty Five to Twenty Nine, 100%; Thirty to Thirty Four, 100%; Thirty Five to Thirty Nine, 89.83%; Forty to Forty Four, 87.96%; Forty Five to Forty Nine, 89.83%; Greater than Fifty Years of age, 78.28%. Genotypes: F508_Homo, 88.8%; F508_Hetero, 100%; Other, 90.13%. Highest accuracy was achieved for the age groups Twenty-Five to Twenty-Nine, Thirty to Thirty-Four, 100% respectively. Lowest accuracy was for age group Twenty to Twenty-Four, 79.52%. For genotypes: F508_Homo 89.39% was achieved by including the variable "Age", 0.59% improvement on accuracy. By including "Year of Birth", achieved an accuracy of 90.33%, increasing the accuracy by 1.53%. F508_Hetero[38] achieved an accuracy of 100%, such a high accuracy could be explained that the combined heterozygosity of PWCF in Ireland according to the 2016 annual report of CF Registry of Ireland[39] page 12, was 91.6%. On balance this classifier performed well on identifying the most relevant variables.

---

[37] https://cfri.ie/annual-reports/

[38] https://www.google.com/search?q=heterozygous+f508+mutation&rlz=1C1JZAP_enIE782IE782&oq=F508+hetero&aqs=chrome.1.69i57j0l4.10326j0j8&sourceid=chrome&ie=UTF-8

[39] https://cfri.ie/annual-reports/

4.5 Implementation, Evaluation and Results of KNN

KNN is a Euclidean based algorithm which uses the distances between feature variables of nearest neighbours. Model was executed using a split ratio of 0.80 i.e., 80% of train data and 20% of test data. Due to high multicollinearity between feature variables which was identified in SPSS it was paramount that the feature variables were chosen carefully, otherwise erroneous results would be the outcome. The following predictors were used (Age Groups + Age/Year of Birth) (2, 3, 4, 5, 6, 7, 8, 9); genotypes (53, 54, 55).

Implementation: KNN was implemented using "class" from RStudio library. The classifier was identified with the function KNN, which was applied to the training set. The model was applied to age groups with year of birth (too many ties in KNN)[40]: less than five years of age; five to nine years of age; ten to fourteen years of age; fifteen to nineteen years of age; twenty to twenty four years of age; twenty five to twenty nine years of age; thirty to thirty four years of age; thirty five to thirty nine years of age; forty to forty four years of age; forty five to forty nine years of age, and greater than fifty years of age. Also, genotype was applied to the target variable – incidence. Due to the Euclidean nature of KNN feature scaling was applied to the important variables. This technique will stop any one variable dominating another.

Evaluation and Results: The accuracy for each group was determined using the caret package confusion matrix, as follows: less than five years of age + year of birth, 88.40%; Ten to fourteen years of age + year of birth, 89.14%; fifteen to nineteen years of age + year of birth, 78.28%; twenty to twenty four years of age + year of birth, 77.69%; Twenty Five to twenty nine years of age + year of birth, 99.60%; Thirty to thirty four years of age + year of birth, 99.70%; thirty five to thirty nine + year of birth, 88.99%; forty to forty four years of age + year of birth, 87.86%; forty five to forty nine years of age + year of birth, 89.93%; greater than fifty years of age + year of birth, 79.37%. Highest accuracy was achieved for age groups twenty-five to twenty-nine, thirty to thirty-four, 99.60% and 99.70% respectively. Lowest accuracy was for age group twenty to twenty-four, 77.69%. For genotypes: F508_Homo 89.39% was achieved, by including the variable "Age", 89.39% accuracy. By including "Year of Birth", achieved an accuracy of 89.14%, a decrease in accuracy by 1.19% compared to random forest. F508_Hetero[41] achieved an accuracy with Age of 99.75%, and with year of birth 89.14%. Such a high accuracy could be explained that the combined heterozygosity of PWCF in Ireland according to the 2016 annual report of CF Registry of Ireland[42] page 12, was 91.6%. For other genotypes with year of birth and age achieved an accuracy of 89.29% and 89.38% respectively.

4.6 Implementation, Evaluation and Results of Naive Bayes

Naive Bayes is a classification machine learning algorithm based on Bayes Theorem of conditional probabilities. The model was executed using a split ratio of 0.80 i.e., 80% of train data and 20% of test data. Due to high multicollinearity between feature variables which was identified in SPSS it was paramount that the feature variables were chosen carefully, otherwise erroneous results would be the outcome. The following predictors were used (Age Groups + Age/Year of Birth) (2, 3, 4, 5, 6, 7, 8, 9); genotypes (53, 54, 55).

Implementation: Naïve Bayes was executed using "e1071"[43] from RStudio library. The classifier was identified with NaïveBayes classifier which was applied to the training set. The

---

[40] https://www.quora.com/What-is-the-meaning-of-the-error-message-too-many-ties-in-KNN-in-R

[41] https://www.google.com/search?q=heterozygous+f508+mutation&rlz=1C1JZAP_enIE782IE782&oq=F508+hetero&aqs=chrome.1.69i57j0l4.10326j0j8&sourceid=chrome&ie=UTF-8

[42] https://cfri.ie/annual-reports/

[43] https://cran.r-project.org/web/packages/e1071/index.html

model was applied to age groups: less than five years of age; five to nine years of age; ten to fourteen years of age; fifteen to nineteen years of age; twenty to twenty four years of age; twenty five to twenty nine years of age; thirty to thirty four years of age; thirty five to thirty nine years of age; forty to forty four years of age; forty five to forty nine years of age, and greater than fifty years of age. Also, genotype was applied to the target variable – incidence. As Naïve Bayes is not distance based, feature scaling has no effect on the algorithm.

Evaluation and Results: The accuracy for each group was determined using the RStudio caret[44] package confusion matrix, as follows: less than five years of age, 88.30%; five to nine years of age, 89.83%; Ten to fourteen years of age, 79.51%; fifteen to nineteen years of age, 79.96%; twenty to twenty four years of age, 79.96%; Twenty Five to twenty nine years of age, 100%; thirty to thirty four years of age, 100%; thirty five to thirty nine, 100%; forty to forty four years of age, 87.96%; forty five to forty nine years of age, 89.68%; greater than fifty years of age, 78.28%. Highest accuracy was achieved for age groups twenty-five to twenty-nine, thirty to thirty-four, thirty-five to thirty-nine, 100% respectively. For genotypes: F508_Homo 88.80% was achieved, by including the variable "Age", 88.55% accuracy was achieved. By including "Year of Birth", achieved an accuracy of 90.03%, an increase in accuracy of 0.64% compared to KNN. F508_Hetero[45] achieved an accuracy with Age of 100%, and with year of birth 100%. Such a high accuracy could be explained that the combined heterozygosity of PWCF in Ireland according to the 2016 annual report of CF Registry of Ireland[46] page 12, was 91.6%. For other genotypes with year of birth and age achieved an accuracy of 89.88% and 90.42% respectively. On balance Naïve Bayes performed better than previous models.

4.7 Implementation, Evaluation and Results of SVM

SVM (Support Vector Machine) is a supervised classification model using classification algorithms for two group classification problems and regression problems. The model was executed using a split ratio of 0.80 i.e., 80% of train data and 20% of test data. Due to high multicollinearity between feature variables which was identified in SPSS it was paramount that the feature variables were chosen carefully, otherwise erroneous results would be the outcome. The following predictors were used (Age Groups), (2, 3, 4, 5, 6, 7, 8, 9); genotypes (53, 54, 55).

Implementation: SVM was executed using the "e1071" package from the RStudio library. The classifier was applied by a "SVM" classifier on the training set with a linear kernel. The model was applied to age groups: less than five years of age; five to nine years of age; ten to fourteen years of age; fifteen to nineteen years of age; twenty to twenty four years of age; twenty five to twenty nine years of age; thirty to thirty four years of age; thirty five to thirty nine years of age; forty to forty four years of age; forty five to forty nine years of age, and greater than fifty years of age. Also, genotype was applied to the target variable – incidence.

Evaluation and Results: The accuracy for each group was determined using a confusion matrix, as follows: less than five years of age, 88.31%; five to nine years of age, 89.83%; Ten to fourteen years of age, 79.51%; fifteen to nineteen years of age, 79.96%; twenty to twenty four years of age, 77.39%; Twenty Five to twenty nine years of age, 100%; Thirty to thirty four years of age, 100%; thirty five to thirty nine, 89.89%; forty to forty four years of age, 87.95%; forty five to forty nine years of age, 89.68%; greater than fifty years of age, 78.30%. Highest accuracy was achieved for age groups twenty-five to twenty-nine, thirty to thirty-four years of

[44] http://topepo.github.io/caret/index.html

[45] https://www.google.com/search?q=heterozygous+f508+mutation&rlz=1C1JZAP_enIE782IE782&oq=F508+hetero&aqs=chrome.1.69i57j0l4.10326j0j8&sourceid=chrome&ie=UTF-8

[46] https://cfri.ie/annual-reports/

age, 100% respectively. For genotypes: F508_Homo 88.78% was achieved, by including the variable "Year of Birth", 88.55% accuracy was achieved; F508_Hetero, 100% was achieved; Other, 79.96%, by including the variable, "Year of Birth", 84.57% was achieved. The SVM model achieved high accuracy on all the categories explored. Highest accuracy in age groups, was twenty-five to twenty-nine, thirty to thirty-four, 100% respectively; lowest accuracy, twenty to twenty-four, 77.39%.

4.8 Implementation, Evaluation and Results of kernel SVM

Kernel SVM (Support Vector Machine) is a supervised classification model using classification algorithms, and the "kernel trick"[47] for non-linear classification problems and regression problems. The model was executed using a split ratio of 0.80 i.e., 80% of train data and 20% of test data. Due to high multicollinearity between feature variables which was identified in SPSS it was paramount that the feature variables were chosen carefully, otherwise erroneous results would be the outcome. The following predictors were used (Age Groups), (2, 3, 4, 5, 6, 7, 8, 9); genotypes (53, 54, 55).

Implementation: SVM was executed using the "e1071" package from the RStudio library. The classifier was applied by a "SVM" classifier on the training set with a radial kernel. The model was applied to age groups: less than five years of age; five to nine years of age; ten to fourteen years of age; fifteen to nineteen years of age; twenty to twenty four years of age; twenty five to twenty nine years of age; thirty to thirty four years of age; thirty five to thirty nine years of age; forty to forty four years of age; forty five to forty nine years of age, and greater than fifty years of age. Also, genotype was applied to the target variable – incidence.

Evaluation and Results: The accuracy for each group was determined using a confusion matrix, as follows: less than five years of age, 88.31%; five to nine years of age, 89.73%; Ten to fourteen years of age, 79.50%; fifteen to nineteen years of age, 79.98%; twenty to twenty four years of age, 77.37%; Twenty Five to twenty nine years of age, 100%; Thirty to thirty four years of age, 100%; thirty five to thirty nine, 89.89%; forty to forty four years of age, 87.95%; forty five to forty nine years of age, 89.65%; greater than fifty years of age, 78.24%. Highest accuracy was achieved for age groups twenty-five to twenty-nine, thirty to thirty-four years of age, 100% respectively. For genotypes: F508_Homo 88.78% was achieved, by including the variable "Year of Birth", 88.55% accuracy was achieved; F508_Hetero, 100% was achieved; Other, 90.13%, by including the variable, "Year of Birth", 81.20% was achieved. The kernel SVM model achieved high accuracy on all the categories explored. Highest accuracy in age groups, was twenty-five to twenty-nine, thirty to thirty-four, 100% respectively; lowest accuracy, twenty to twenty-four, 77.37%.

Conclusion: all the objectives (chapter 1, sub-section 1.3) have been implemented and the results presented here have solved the research question, sub-RQ1 and sub-RQ2.

4.9 Accuracy of All Models Implemented

Figure 10 illustrates the accuracy of all the models used on the features chosen (age groups and genotypes) by inspection KNN appears to out-perform all other models.

[47] https://en.wikipedia.org/wiki/Kernel_method

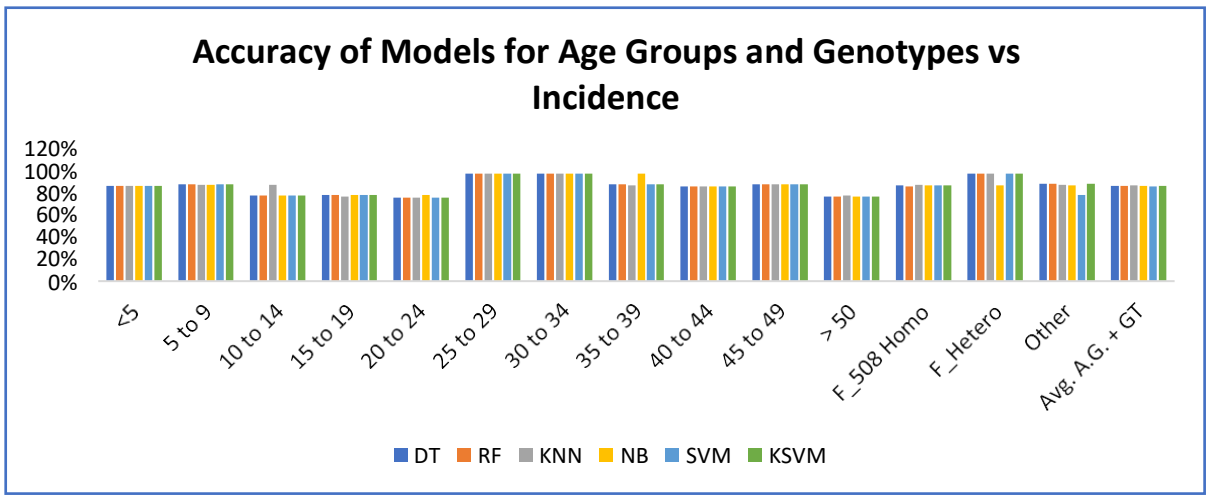

Figure 10: Accuracy of Models for Age Groups and Genotypes vs Incidence of Cystic Fibrosis

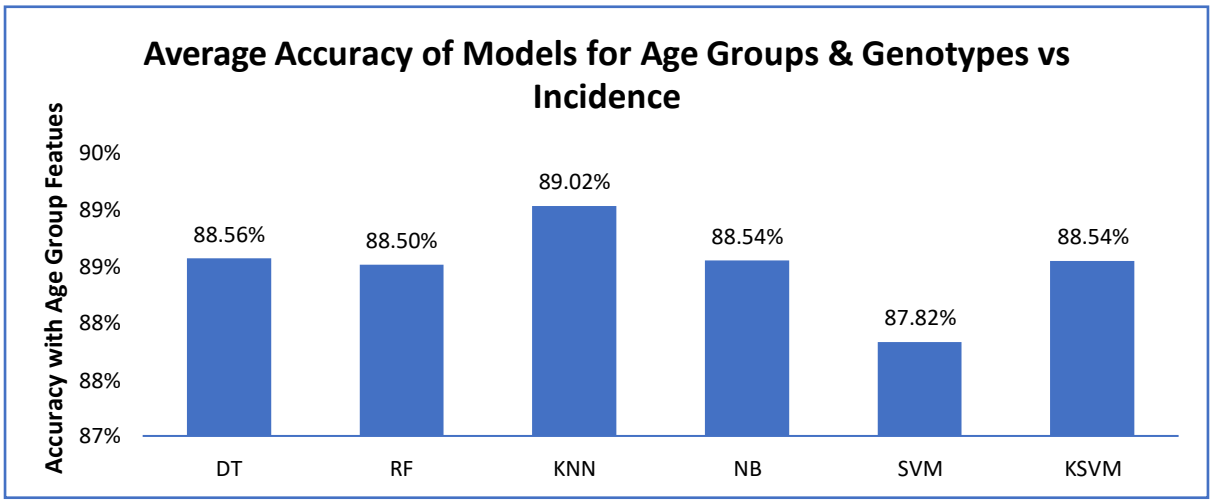

Figure 11: Average accuracy of Models for Age Groups & Genotypes vs Cystic Fibrosis Incidence

Figure 11 concludes that KNN out-performs all other models based on average accuracy for all the chosen features.

5 Discussion

5.1 Limitations

Due to the extreme difficulty in sourcing a dataset in Ireland it was decided to scrape data from Cystic Fibrosis Registry of Ireland annual reports from 2008 to 2016. Some of the variables such as ethnicity had to be randomised due to missing ethnicity variables from some of the reports.

5.2 Outcomes

All the models implemented on the features chosen, appear to show a bearing on incidence of cystic fibrosis. By inspecting figure 6, polynomial regression of the increase in population versus the numbers of incidences of CF from 2008 to 2016 the prediction line follows the datapoints of incidence apart from an outlier for year 2009 (which is explainable by data not updated for that year). However, this does not prove that population increase has an influence

on the research question by itself. By further analysis on age groups, and genotypes showed there was some correlations with incidence. By inspecting figure 7 age group, 10 to 14, showed a high correlation of 0.87, whereas other age groups, such as: less than 5; 5 to 9; 15 to 19; 20 to 24; 25 to 29; 30 to 34; 40 to 44; 45 to 49: 0.23, -.50, -0.26, -0.6, 0.77, -.05, 0.11, and 0.08 respectively.

This result shows that there is a pattern between the "ebb and flow" of the population not only from the time ten countries annexed to the European Union (2004), but farther back in time during recessions in the late eighties and seventies when there was a constant flow of emigration out of Ireland to locate work. Similarly, during the time, ten countries annexed to the European Union there was a large influx of people from these countries, especially Poland, which according to the 2016 census Polish people represent 2.57% - 122,515[48] of the 2016 population in Ireland. Out of all the nationalities that have emigrated to Ireland from 2004 to 2016, people of polish descent are probably more amenable to assimilation due to their common religion (86% stated in the 2016 census as Roman Catholic) although only 4% said they had an Irish partner. Notwithstanding, the prevalence in Poland of cystic fibrosis is only 1.90 per 100,000 as opposed to 26.96 per 100,000 in Ireland – however, if we exclude Non-Irish from the population this prevalence increases to 32.45 per 100,000 of population, so consanguinity (denoting people from the same ancestor) intermingling of populations should in theory be instrumental to incidence/prevalence of any genetic disease. There is also evidence that new-born screening is influential for parents deciding to risk having another child with a genetic disease (Joseph, G et.al, 2016).

6 Conclusion and Future Work

The research undertaken did produce some patterns/correlations especially in relation to age groups whereby these groups did show patterns with net-migration over the period in question 2004 to 2016 and over the last fifty years. There were also some patterns in relation to genotypes via excluding net-migration and including net-migration to the models. Although, minimal changes especially to other genotypes – including net-migration -0.81 r-correlation as opposed to excluding net-population -0.84 r-correlation.

Limitations: Some of the variables, such as: ethnicity, age, were randomised due to inaccessibility to data or missing data. There was a high multicollinearity between 41 variables according to SPSS which were excluded.

Future Work: The research on whether net-migration influences CF incidence can be enhanced/improved on by inputting more variables that may have an influence on the research question or not. These, extra variables were not available to me due to privacy concerns. It is a fact that some of the CF registries in Europe measure over 200 variables that if modelled appropriately could have the potential of forecasting future incidence of CF not only in Ireland, but any of the European countries.

Conclusion: Objective 4 was not fully achieved as there was no research solely dedicated to the solution of RQ1 or other countries which experience high levels of Net-Migration.

7 Acknowledgement

I wish to thank my supervisor, Dr Catherine Mulwa for her support, and guidance throughout this journey. My wife for her support and patience during this time. Dr Abigail Jackson of CF Ireland Registry for her help and advice.

48

https://en.wikipedia.org/wiki/Polish_minority_in_the_Republic_of_Ireland#:~:text=The%20Polish%20minority%20in%20the,according%20to%202016%20census%20figures.

References

Athanazio, R., Silva Filho, L., Vergara, A., Ribeiro, A., Riedi, C., Procianoy, E., Adde, F., Reis, F., Ribeiro, J., Torres, L., Fuccio, M., Epifanio, M., Firmida, M., Damaceno, N., Ludwig-Neto, N., Maróstica, P., Rached, S. and Melo, S., 2017. Brazilian guidelines for the diagnosis and treatment of cystic fibrosis. *Jornal Brasileiro de Pneumologia*, [online] 43(3), pp.219-245. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5687954/> [Accessed 26 April 2020].

Barben, J., Castellani, C., Dankert-Roelse, J., Gartner, S., Kashirskaya, N., Linnane, B., Mayell, S., Munck, A., Sands, D., Sommerburg, O., Pybus, S., Winters, V. and Southern, K., 2017. The expansion and performance of national newborn screening programmes for cystic fibrosis in Europe. *Journal of Cystic Fibrosis*, [online] 16(2), pp.207-213. Available at: <https://reader.elsevier.com/reader/sd/pii/S1569199316306816?token=BA90A776FE8A35FDA661F19AA6056FB990BE467521F7F45C6FA99B5F13508B11E4D3F6B236EF51DC25784886555595B6> [Accessed 3 May 2020].

Barrett, P., 2015. A review of consanguinity in Ireland—estimation of frequency and approaches to mitigate risks. *Irish Journal of Medical Science (1971 -)*, [online] 185(1), pp.17-28. Available at: <https://www.researchgate.net/publication/283079071_A_review_of_consanguinity_in_Ireland-estimation_of_frequency_and_approaches_to_mitigate_risks/link/5b59a73daca272a2d66c28d3/download> [Accessed 14 June 2020].

Belloy, M., Napolioni, V., Han, S., Le Guen, Y. and Greicius, M., 2020. Association of Klotho-VS Heterozygosity With Risk of Alzheimer Disease in Individuals Who Carry APOE4. *JAMA Neurology*,.

Bobadilla, J., Macek, M., Fine, J. and Farrell, P., 2002. *Cystic Fibrosis: A Worldwide Analysis Of CFTR Mutations?Correlation With Incidence Data And Application To Screening*.

Bosch B, e., 2020. *Ethnicity Impacts The Cystic Fibrosis Diagnosis: A Note Of Caution. - Pubmed - NCBI*. [online] Ncbi.nlm.nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/28233695> [Accessed 18 April 2020].

Burgel, P., Bellis, G., Olesen, H., Viviani, L., Zolin, A., Blasi, F. and Elborn, J., 2015. Future trends in cystic fibrosis demography in 34 European countries. *European Respiratory Journal*, [online] 46(1), pp.133-141. Available at: <https://www.scopus.com/record/display.uri?eid=2-s2.0-84937406021&origin=resultslist&sort=plf-f&cite=2-s2.0-84937406021&src=s&imp=t&sid=5edd03fe90dd9c5b86accb05ece5eb1a&sot=cite&sdt=a&sl=0&recordRank=> [Accessed 18 April 2020].

Capriotti, E., Calabrese, R. and Casadio, R., 2006. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, [online] 22(22), pp.2729-2734. Available at: <https://academic.oup.com/bioinformatics/article/22/22/2729/196993> [Accessed 13 June 2020].

Casals, T., Gim&#xfffd;nez, J., Larriba, S., Estivill, X., Ramos, M. and Nunes, V., 1997. High heterogeneity for cystic fibrosis in Spanish families: 75 mutations account for 90% of chromosomes. *Human Genetics*, 101(3), pp.365-370.

Cashman, S., Patino, A., Delgado, M., Byrne, L., Denham, B. and De Arce, M., 1997. *The Irish Cystic Fibrosis Database.*. [online] Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1051780/> [Accessed 4 May 2020].

Cashman, S., Patino, A., Delgado, M., Byrne, L., Denham, B. and De Arce, M., 1995. The Irish cystic fibrosis database. *Journal of Medical Genetics*, 32(12), pp.972-975.
Castellani, C., 2009. *Association Between Carrier Screening And Incidence Of Cystic Fibrosis*.
Cdc.gov. 2020. *Principles Of Epidemiology | Lesson 3 - Section 2*. [online] Available at: <https://www.cdc.gov/csels/dsepd/ss1978/lesson3/section2.html> [Accessed 26 July 2020].
Cff.org. 2016. *The Patient Registry: Where We've Been, Where We're Going*. [online] Available at: <https://www.cff.org/CF-Community-Blog/Posts/2016/The-Patient-Registry-Where-Weve-Been-Where-Were-Going/> [Accessed 18 April 2020].
Cff.org. 2020. *CFTR MUTATION CLASSES*. [online] Available at: <https://www.cff.org/Care/Clinician-Resources/Network-News/August-2017/Know-Your-CFTR-Mutations.pdf> [Accessed 4 May 2020].
Conrad, D. and Bailey, B., 2015. Multidimensional Clinical Phenotyping of an Adult Cystic Fibrosis Patient Population. *PLOS ONE*, 10(3), p.e0122705.
Cso.ie. 2020. *Population And Migration Estimates April 2019 - CSO - Central Statistics Office*. [online] Available at: <https://www.cso.ie/en/releasesandpublications/er/pme/populationandmigrationestimatesapril2019/> [Accessed 24 July 2020].
Devyser. 2020. *CFTR – Detect All Mutations In The CFTR Gene | Devyser*. [online] Available at: <https://devyser.com/products/cystic-fibrosis/devyser-cftr/?gclid=Cj0KCQjw-r71BRDuARIsAB7i_QM8eLXy9agz-_sae1IjM9xyUWXGGynjam_1CC9RlDxfhkbD3-pQ-dsaAlCQEALw_wcB#devyser-cftr-uk> [Accessed 4 May 2020].
Du, B., Zhang, C., Yue, L., Ren, B., Zhao, Q., Li, D., He, Y. and Zhang, W., 2019. Prediction model for the efficacy of folic acid therapy on hyperhomocysteinaemia based on genetic risk score methods. *British Journal of Nutrition*, [online] 122(1), pp.39-46. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
El Tahir, O., de Jonge, R., Ouburg, S., Morré, S. and van Furth, A., 2019. Study protocol: The Dutch 20|30 Postmeningitis study: a cross-sectional follow-up of two historical childhood bacterial meningitis cohorts on long-term outcomes. *BMC Pediatrics*, 19(1).
En.wikipedia.org. 2020. *Demographics Of Belgium*. [online] Available at: <https://en.wikipedia.org/wiki/Demographics_of_Belgium> [Accessed 9 May 2020].
Farrell, P., 2008. The prevalence of cystic fibrosis in the European Union. *Journal of Cystic Fibrosis*, [online] 7(5), pp.450-453. Available at: <https://www.sciencedirect.com/science/article/pii/S1569199308000349> [Accessed 18 April 2020].
Farrell, P., Joffe, S., Foley, L., Canny, G., Mayne, P. and Rosenberg, M., 2007. *Diagnosis Of Cystic Fibrosis In The Republic Of Ireland: Epidemiology And Costs.*. [online] figshare. Available at: <https://repository.rcsi.com/articles/Diagnosis_of_cystic_fibrosis_in_the_Republic_of_Ireland_epidemiology_and_costs_/10782233/1> [Accessed 15 April 2020].
Ferreira, L., Secolin, R., Lopes-Cendes, I., Cabral, N. and França, P., 2019. Association and interaction of genetic variants with occurrence of ischemic stroke among Brazilian patients. *Gene*, [online] 695, pp.84-91. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Foley, L., 2008. *2008 Annual Report CF Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2008.pdf> [Accessed 14 June 2020].
Foley, L., 2020. *2009 Annual Report CF Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2009.pdf> [Accessed 14 June 2020].

Han, Y., Zheng, Q., Tian, Y., Ji, Z. and Ye, H., 2019. Identification of a nine-gene panel as a prognostic indicator for recurrence with muscle-invasive bladder cancer. *Journal of Surgical Oncology*, [online] 119(8), pp.1145-1154. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Han, Y., Zheng, Q., Tian, Y., Ji, Z. and Ye, H., 2019. Identification of a nine-gene panel as a prognostic indicator for recurrence with muscle-invasive bladder cancer. *Journal of Surgical Oncology*, [online] 119(8), pp.1145-1154. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Healthline. 2020. *Fast Facts About Cystic Fibrosis*. [online] Available at: <https://www.healthline.com/health/cystic-fibrosis-facts> [Accessed 19 April 2020].
Hodson, M., Geddes, D. and Bush, A., 2007. *Cystic Fibrosis*. 3rd ed. London: Hodder Arnold, pp.21-43.
https://link.springer.com/article/10.1007/s004390050643. 1997. *High Heterogeneity For Cystic Fibrosis In Spanish Families: 75 Mutations Account For 90% Of Chromosomes*. [online] Available at: <https://link.springer.com/article/10.1007/s004390050643> [Accessed 4 May 2020].
Hu, S., Wang, Y., He, M., Zhang, M., Ding, X. and Shi, B., 2018. Factors associated with the efficacy of intravenous methylprednisolone in moderate-to-severe and active thyroid-associated ophthalmopathy: A single-centre retrospective study. *Clinical Endocrinology*, [online] 90(1), pp.175-183. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Human Genetics, 1990. Gradient of distribution in Europe of the major CF mutation and of its associated haplotype. [online] 85(4), pp.436-445. Available at: <https://pubmed.ncbi.nlm.nih.gov/2210767/> [Accessed 4 May 2020].
Ibm.com. 2020. *FACTOR Does Not Print KMO Or Bartlett Test For Nonpositive Definite Matrices*. [online] Available at: <https://www.ibm.com/support/pages/factor-does-not-print-kmo-or-bartlett-test-nonpositive-definite-matrices> [Accessed 28 July 2020].
Incidence and Prevalence - Everything you need to know. 2016. [video] Directed by F. Wiesbauer. https://www.youtube.com/watch?v=cTp_ONVVrh8: https://www.youtube.com/channel/UC1a4RqaEpMsHja5dSlUE1mg.
Jackson, A. and Goss, C., 2017. *Epidemiology Of CF: How Registries Can Be Used To Advance Our Understanding Of The CF Population*.
Jackson, A., Daly, L., Kelleher, C., Fletcher, G., Harrington, M., Zhou, S. and Fitzpatrick, P., 2010. *Annual Report 2010 Cystic Fibrosis Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2010.pdf> [Accessed 14 June 2020].
Jackson, A., Daly, L., Kelleher, C., Marshall, B., Quinton, H. and Fletcher, G., 2011. *Annual Report 2011 Cystic Fibrosis Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2011.pdf> [Accessed 14 June 2020].
Jackson, D. and Ungar, L., 2020. *2012 Annual Report Cystic Fibrosis Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2012.pdf> [Accessed 14 June 2020].
Jacobsen, L., Larsson, H., Tamura, R., Vehik, K., Clasen, J., Sosenko, J., Hagopian, W., She, J., Steck, A., Rewers, M., Simell, O., Toppari, J., Veijola, R., Ziegler, A., Krischer, J., Akolkar, B. and Haller, M., 2019. Predicting progression to type 1 diabetes from ages 3 to 6 in islet autoantibody positive TEDDY children. *Pediatric Diabetes*, [online] 20(3), pp.263-270. Available at:

<https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
James, S., 1993. *Exploring The World Of The Celts*. London: Thames and Hudson.
Joseph, G., Chen, F., Harris-Wai, J., Puck, J., Young, C. and Koenig, B., 2020. *Parental Views On Expanded Newborn Screening Using Whole-Genome Sequencing*. [online] Available at: <https://pubmed.ncbi.nlm.nih.gov/26729702/> [Accessed 9 August 2020].
Kerr, A., 2005. Understanding genetic disease in a socio-historical context: a case study of cystic fibrosis. *Sociology of Health and Illness*, [online] 27(7), pp.873-896. Available at: <https://onlinelibrary.wiley.com/doi/full/10.1111/j.1467-9566.2005.00462.x> [Accessed 18 April 2020].
Khan Academy. 2020. *Genetic Drift (Article) | Khan Academy*. [online] Available at: <https://www.khanacademy.org/science/biology/her/heredity-and-genetics/a/genetic-drift-founder-bottleneck> [Accessed 10 May 2020].
Kirwan, L., Fletcher, G., Harrington, M., Zhou, S., Jeleniewska, P., Hassan, M., Rubin, J. and Jackson, A., 2018. P267 Pulmonary exacerbations and risk of death amongst patients with cystic fibrosis (CF) with homozygous and heterozygous F508del mutations in Ireland. *Journal of Cystic Fibrosis*, 17, p.S135.
Kluska, A., Kulecka, M., Litwin, T., Dziezyc, K., Balabas, A., Piatkowska, M., Paziewska, A., Dabrowska, M., Mikula, M., Kaminska, D., Wiernicka, A., Socha, P., Czlonkowska, A. and Ostrowski, J., 2018. Whole-exome sequencing identifies novel pathogenic variants across theATP7Bgene and some modifiers of Wilson's disease phenotype. *Liver International*, [online] 39(1), pp.177-186. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Knapp, E., Fink, A., Goss, C., Sewall, A., Ostrenga, J., Dowd, C., Elbert, A., Petren, K. and Marshall, B., 2016. The Cystic Fibrosis Foundation Patient Registry. Design and Methods of a National Observational Disease Registry. *Annals of the American Thoracic Society*, [online] 13(7), pp.1173-1179. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/27078236> [Accessed 18 April 2020].
Kong, S. and Cho, Y., 2019. Identification of female-specific genetic variants for metabolic syndrome and its component traits to improve the prediction of metabolic syndrome in females. *BMC Medical Genetics*, 20(1).
KOSOROK, M., WEI, W. and FARRELL, P., 1996. THE INCIDENCE OF CYSTIC FIBROSIS. *Statistics in Medicine*, [online] 15(5), pp.449-462. Available at: <https://onlinelibrary.wiley.com/doi/abs/10.1002/%28SICI%291097-0258%2819960315%2915%3A5%3C449%3A%3AAID-SIM173%3E3.0.CO%3B2-X> [Accessed 1 May 2020].
Kurgan, L., Cios, K., Sontag, M. and Accurso, F., 2005. *Mining The Cystic Fibrosis Data*. [online] Academia. Available at: <https://www.academia.edu/19381386/Mining_the_cystic_fibrosis_data> [Accessed 13 June 2020].
Lee, C., Cui, Y., Song, J., Li, S., Zhang, F., Wu, M., Li, L., Hu, D. and Chen, H., 2019. Effects of familial hypercholesterolemia-associated genes on the phenotype of premature myocardial infarction. *Lipids in Health and Disease*, [online] 18(1). Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019>.
Li, Y., Peng, C., Zhang, J., Zhu, W., Xu, C., Lin, Y., Fu, X., Tian, Q., Zhang, L., Xiang, Y., Sheng, V. and Deng, H., 2019. Genetic risk factors identified in populations of European descent do not improve the prediction of osteoporotic fracture and bone mineral density in Chinese populations. *Scientific Reports*, [online] 9(1). Available at:

<https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019>.
Liou, T., Adler, F., FitzSimmons, S., Cahill, B., Hibbs, J. and Marshall, B., 2001. Predictive 5-Year Survivorship Model of Cystic Fibrosis. *American Journal of Epidemiology*, [online] 153(4), pp.345-352. Available at: <http://http:academic.oup.com/aje/article/153/4/345/129039> [Accessed 23 March 2020].
LiPuma, J., 2010. The Changing Microbial Epidemiology in Cystic Fibrosis. *Clinical Microbiology Reviews*, [online] 23(2), pp.299-323. Available at: <https://cmr.asm.org/content/cmr/23/2/299.full.pdf> [Accessed 1 May 2020].
Ma, Y., Zhao, S., Li, L., Sun, F., Ye, X., Yuan, F., Jiang, D., Zhou, Z., Zhang, Q., Wan, Y., Zhang, G., Wu, J., Zhang, R., Fang, Y. and Song, H., 2019. A Weighted Genetic Risk Score Using Known Susceptibility Variants to Predict Graves Disease Risk. *The Journal of Clinical Endocrinology & Metabolism*, [online] 104(6), pp.2121-2130. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Mateu, E., Calafell, F., Ramos, M., Casals, T. and Bertranpetit, J., 2002. Can a Place of Origin of the Main Cystic Fibrosis Mutations Be Identified?. *The American Journal of Human Genetics*, [online] 70(1), pp.257-264. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC384895/> [Accessed 4 May 2020].
McCarthy, C., Dimitrov, B., Meurling, I., Gunaratnam, C. and McElvaney, N., 2013. The CF-ABLE Score. *Chest*, [online] 143(5), pp.1358-1364. Available at: <https://www.sciencedirect.com/science/article/abs/pii/S0012369213603359> [Accessed 23 March 2020].
McCormick, J., Mehta, G., Olesen, H., Viviani, L., Macek, M. and Mehta, A., 2010. Comparative demographics of the European cystic fibrosis population: a cross-sectional database analysis. *The Lancet*, [online] 375(9719), pp.1007-1013. Available at: <https://www.sciencedirect.com/science/article/pii/S0140673609621619> [Accessed 18 April 2020].
Mehta G, e., 2010. *Cystic Fibrosis Across Europe: Eurocarecf Analysis Of Demographic Data From 35 Countries. - Pubmed - NCBI*. [online] Ncbi.nlm.nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/21041121> [Accessed 18 April 2020].
Michelle Steinraths, A., 2008. *Delays In Diagnosing Cystic Fibrosis: Can We Find Ways To Diagnose It Earlier?*. [online] PubMed Central (PMC). Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2427000/> [Accessed 18 April 2020].
Monnier, A., 2004. L'Union européenne à l'heure de l'élargissement. *Population*, [online] 59(2), p.361. Available at: <https://www.cairn-int.info/article-E_POPU_402_0361--the-european-union-at-the-time-of-enlarg.htm> [Accessed 26 April 2020].
Onengut-Gumuscu, S., Chen, W., Robertson, C., Bonnie, J., Farber, E., Zhu, Z., Oksenberg, J., Brant, S., Bridges, S., Edberg, J., Kimberly, R., Gregersen, P., Rewers, M., Steck, A., Black, M., Dabelea, D., Pihoker, C., Atkinson, M., Wagenknecht, L., Divers, J., Bell, R., Erlich, H., Concannon, P. and Rich, S., 2019. Type 1 Diabetes Risk in African-Ancestry Participants and Utility of an Ancestry-Specific Genetic Risk Score. *Diabetes Care*, [online] 42(3), pp.406-415. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Padula, M., Leccese, P., Pellizzieri, E., Padula, A., Gilio, M., Carbone, T., Lascaro, N., Tramontano, G., Martelli, G. and D'Angelo, S., 2019. Distribution of rs17482078 and rs27044 ERAP1 polymorphisms in a group of Italian Behçet's syndrome patients: a preliminary case–control study. *Internal and Emergency Medicine*, [online] 14(5), pp.713-718. Available at:

<https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
PB, D., 2020. *Cystic Fibrosis Since 1938. - Pubmed - NCBI*. [online] Ncbi.nlm.nih.gov. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/16126935> [Accessed 1 May 2020].
Ripatti, S., Tikkanen, E., Orho-Melander, M., Havulinna, A., Silander, K., Sharma, A., Guiducci, C., Perola, M., Jula, A., Sinisalo, J., Lokki, M., Nieminen, M., Melander, O., Salomaa, V., Peltonen, L. and Kathiresan, S., 2010. A multilocus genetic risk score for coronary heart disease: case-control and prospective cohort analyses. *The Lancet*, [online] 376(9750), pp.1393-1400. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/20971364> [Accessed 17 May 2020].
Rogers, L., Verlinde, M. and Mias, G., 2019. Gene expression microarray public dataset reanalysis in chronic obstructive pulmonary disease. *PLOS ONE*, 14(11), p.e0224750.
Sawicki, G., Sellers, D., McGuffie, K. and Robinson, W., 2007. Adults with cystic fibrosis report important and unmet needs for disease information. *Journal of Cystic Fibrosis*, [online] 6(6), pp.411-416. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/17452026> [Accessed 18 April 2020].
Schrodi, S., Mukherjee, S., Shan, Y., Tromp, G., Sninsky, J., Callear, A., Carter, T., Ye, Z., Haines, J., Brilliant, M., Crane, P., Smelser, D., Elston, R. and Weeks, D., 2014. Genetic-based prediction of disease traits: prediction is very difficult, especially about the futureâ€. *Frontiers in Genetics*, [online] 5. Available at: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4040440/> [Accessed 24 March 2020].
Schwarz, C. and Hartl, D., 2015. Cystic fibrosis in Europe: patients live longer but are we ready?. *European Respiratory Journal*, 46(1), pp.11-12.
Serra, G., Koukopoulos, A., De Chiara, L., Koukopoulos, A., Sani, G., Tondo, L., Girardi, P., Reginaldi, D. and Baldessarini, R., 2018. Early clinical predictors of long-term morbidity in major depressive disorder. *Early Intervention in Psychiatry*, [online] 13(4), pp.999-1002. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Shou, D., Harrington, M., Jeleniewska, P., Kirwin, D. and Jackson, D., 2016. *Annual Reports -*. [online] Cfri.ie. Available at: <https://cfri.ie/annual-reports/> [Accessed 14 June 2020].
Shteinberg, M., Downey, D., Beattie, D., McCaughan, J., Reid, A., Stein, N. and Elborn, J., 2017. Lung function and disease severity in cystic fibrosis patients heterozygous for p.Arg117His. *ERJ Open Research*, [online] 3(1), pp.00056-2016. Available at: <https://openres.ersjournals.com/content/erjor/3/1/00056-2016.full.pdf> [Accessed 31 May 2020].
Sisters, A., 2018. *Alleles And Genes*. [video] Available at: <https://www.youtube.com/watch?v=pv3Kj0UjiLE> [Accessed 9 May 2020].
Smyth, A., Bell, S., Bojcin, S., Byron, M., Duff, A. and Flume, P., 2014. [online] Cysticfibrosisjournal.com. Available at: <https://www.cysticfibrosisjournal.com/article/S1569-1993(14)00085-X/pdf> [Accessed 18 April 2020].
Stern, M., Bertrand, D., Bignamini, E., Corey, M., Dembski, B., Goss, C., Pressler, T., Rault, G., Viviani, L., Elborn, J. and Castellani, C., 2014. European Cystic Fibrosis Society Standards of Care: Quality Management in cystic fibrosis. *Journal of Cystic Fibrosis*, 13, pp.S43-S59.
Taylor-Robinson, D., Archangelidi, O., Carr, S., Cosgriff, R., Gunn, E., Keogh, R., MacDougall, A., Newsome, S., Schlüter, D., Stanojevic, S. and Bilton, D., 2020. *Data Resource Profile: The UK Cystic Fibrosis Registry*.

Valls, J., Cambray, S., Pérez-Guallar, C., Bozic, M., Bermúdez-López, M., Fernández, E., Betriu, À., Rodríguez, I. and Valdivielso, J., 2019. Association of Candidate Gene Polymorphisms With Chronic Kidney Disease: Results of a Case-Control Analysis in the Nefrona Cohort. *Frontiers in Genetics*, [online] 10. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Valls, J., Cambray, S., Pérez-Guallar, C., Bozic, M., Bermúdez-López, M., Fernández, E., Betriu, À., Rodríguez, I. and Valdivielso, J., 2019. Association of Candidate Gene Polymorphisms With Chronic Kidney Disease: Results of a Case-Control Analysis in the Nefrona Cohort. *Frontiers in Genetics*, [online] 10. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Wang, L., You, Z., Chen, X., Li, Y., Dong, Y., Li, L. and Zheng, K., 2019. LMTRDA: Using logistic model tree to predict MiRNA-disease associations by fusing multi-source information of sequences and similarities. *PLOS Computational Biology*, [online] 15(3), p.e1006865. Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Wanyama, S. and Thomas, M., 2018. *Annual Report Belgian Cystic Fibrosis Registry 2016*. [online] Sciensano.be. Available at: <https://www.sciensano.be/sites/www.wiv-isp.be/files/report_belgian_cf_registry_2016_en_final_1.pdf> [Accessed 9 May 2020].
Yang, Q., Khoury, M., Botto, L., Friedman, J. and Flanders, W., 2003. Improving the Prediction of Complex Diseases by Testing for Multiple Disease-Susceptibility Genes. *The American Journal of Human Genetics*, 72(3), pp.636-649.
Zhang, W., Dang, S., Zhang, G., He, H. and Wen, X., 2019. Genetic polymorphisms of IL-10, IL-18 and IL12B are associated with risk of non-small cell lung cancer in a Chinese Han population. *International Immunopharmacology*, 77, p.105938.
Zhang, X., Ni, Y., Liu, Y., Zhang, L., Zhang, M., Fang, X., Yang, Z., Wang, Q., Li, H., Xia, Y. and Zhu, Y., 2019. Screening of noise-induced hearing loss (NIHL)-associated SNPs and the assessment of its genetic susceptibility. *Environmental Health*, [online] 18(1). Available at: <https://www.ncbi.nlm.nih.gov/pubmed/?term=logistic+regression+models+used+in+associated+with+genetic+disease+prediction+2019> [Accessed 17 May 2020].
Zhou, D., Harrington, M. and Jackson, D., 2013. *2013 Annual Report Cystic Fibrosis Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2013.pdf> [Accessed 14 June 2020].
Zhou, D., Harrington, M. and Jackson, D., 2014. *2014 Annual Report Cystic Fibrosis Registry Of Ireland*. [online] Cfri.ie. Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2014.pdf> [Accessed 14 June 2020].
Zhou, D., Harrington, M., Jeleniewska, M., Rice, D., Babu, M., Kirwin, D. and Jackson, D., 2018. *Annual Reports* -. [online] Cfri.ie. Available at: <https://cfri.ie/annual-reports/> [Accessed 10 May 2020].
Zhou, D., Harrington, M., Jeleniewska, P., Kirwin, D. and Jackson, D., 2015. *2015 Annual Report Cystic Fibrosis Of Ireland*. [online] Available at: <https://cfri.ie/wp-content/uploads/2020/03/CFRI2015.pdf> [Accessed 14 June 2020].

Appendix

Feedback from Examiners

Hi **Fergal,**

Your MSCDA project provisional result is **83%**:

COMBINED COMMENTS FROM BOTH EXAMINERS:

Project Specification: Real life solution. Objectives have been clearly specified and are creative and appropriate. Objectives have been fully achieved or surpassed

Literature Review: Critical application and critique of relevant theory and concepts. Evidence of breadth and depth of literature reviewed. Recommend removing full stops in sub-sections e.g sub-section 2.3

Artefact/Product Development: Alternative methodologies have been fully considered and the chosen methodology fully justified. The application has been rigorously carried out.

Evaluation: Project have commercialization or further research potential based on outputted results which enhance Cystic Fibrosis research field in Ireland. Rigorous and creative analysis of findings. Demonstrated the ability to synthesise data collected and relevant theory. Insightful conclusions which appreciate limitations and implications of the study.

Configuration manual, document presentation/structure, and referencing: Excellent description and structure to reproduce environments. Excellent presentation and structure with rigorous referencing. Correct grammar and spelling.

Viva: Excellent viva video presentation. Brilliant response to questions asked.

*** Please note, you are required to double check your technical report to make sure there are no spelling errors, all tables and figures are referenced in paragraphs they appear and all references and correct; and then submit the final version before** Mon 28th September 2020