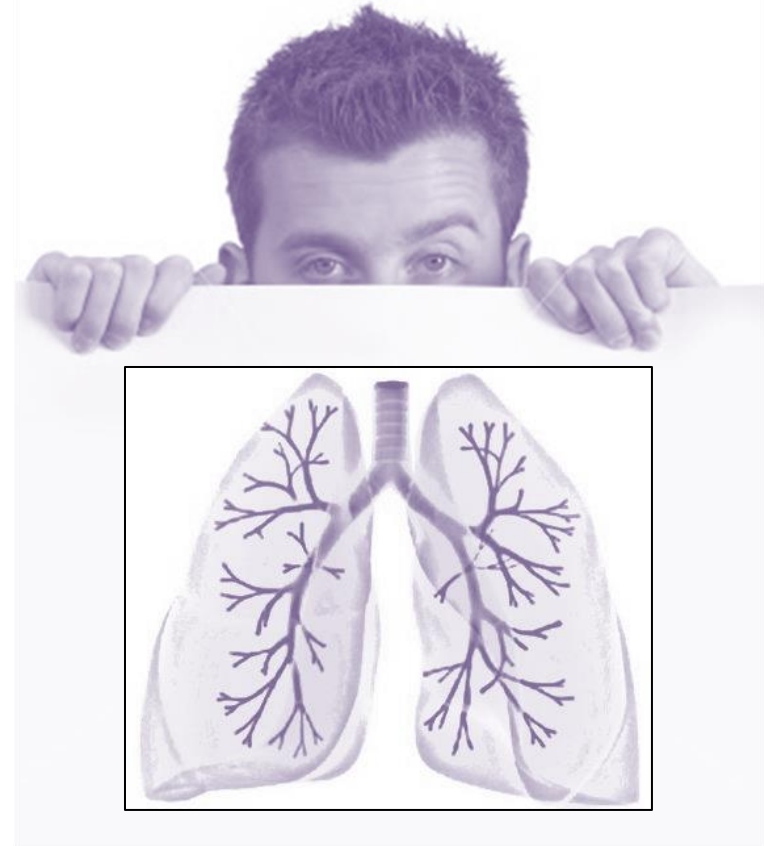


# Cracking the Code for Cystic Fibrosis using Machine Learning

Mihaela van der Schaar  
Ahmed Alaa

Cystic  
Fibrosis Trust



The  
Alan Turing  
Institute

# Overview

**Section A: Vision**

**Section B: CF Registry Data Analysis**

**Section C: Research Agenda**

**Section D: Preliminary Results**

This presentation provides a comprehensive research agenda and preliminary results for the project: **“Personalized risk scoring and monitoring for cystic fibrosis patients”**, a partnership between **Alan Turing Institute** and the **UK CF Trust**

# Section A: Vision

**Alan Turing Institute:** Mission and Vision

**Partnership with the UK CF Trust**



# Alan Turing Institute: Mission and Vision



- **Making great leaps in data science research in order to change the world for the better.**
- **Data-centric healthcare is one of the main areas of research interest in Alan Turing Institute.**

“ This combination of data science techniques and human decision making is an excellent example of augmented intelligence. This opens the way to personalised intelligent medicine, which is set to have a transformative effect on healthcare ”

**Sir. Alan Wilson**  
CEO of Alan Turing Institute

# Alan Turing Institute: Mission and Vision

## Vision

**Changing the way medicine  
is done...!**

...by providing clinicians  
with actionable intelligence  
extracted from data using  
machine learning.



Watch Mihaela's  
Turing lecture!

## Mission

**Developing prognostic tools  
for personalized medicine:**

Personalized risk  
assessment,  
Personalized treatment  
planning

Read Mihaela's publications!

<http://medianetlab.ee.ucla.edu/MedAdvance>

**Prof. Mihaela van der Schaar**

Faculty Fellow, Alan Turing Institute  
MAN Professor, University of Oxford

# This Presentation...

- **The objective of this presentation is to:**
  - Present to the collaborators at the Trust our understanding of the **data** and the **clinical set up (Section B)**
  - Propose a detailed **research agenda** including the research questions we are planning to answer, in addition to our action plan and timeline **(Section C)**
  - Present some preliminary results for the potentiality of our methods applied to the CF registry data **(Section D)**

# Section B: Analysis of the CF Registry Data

The Structure of the Data

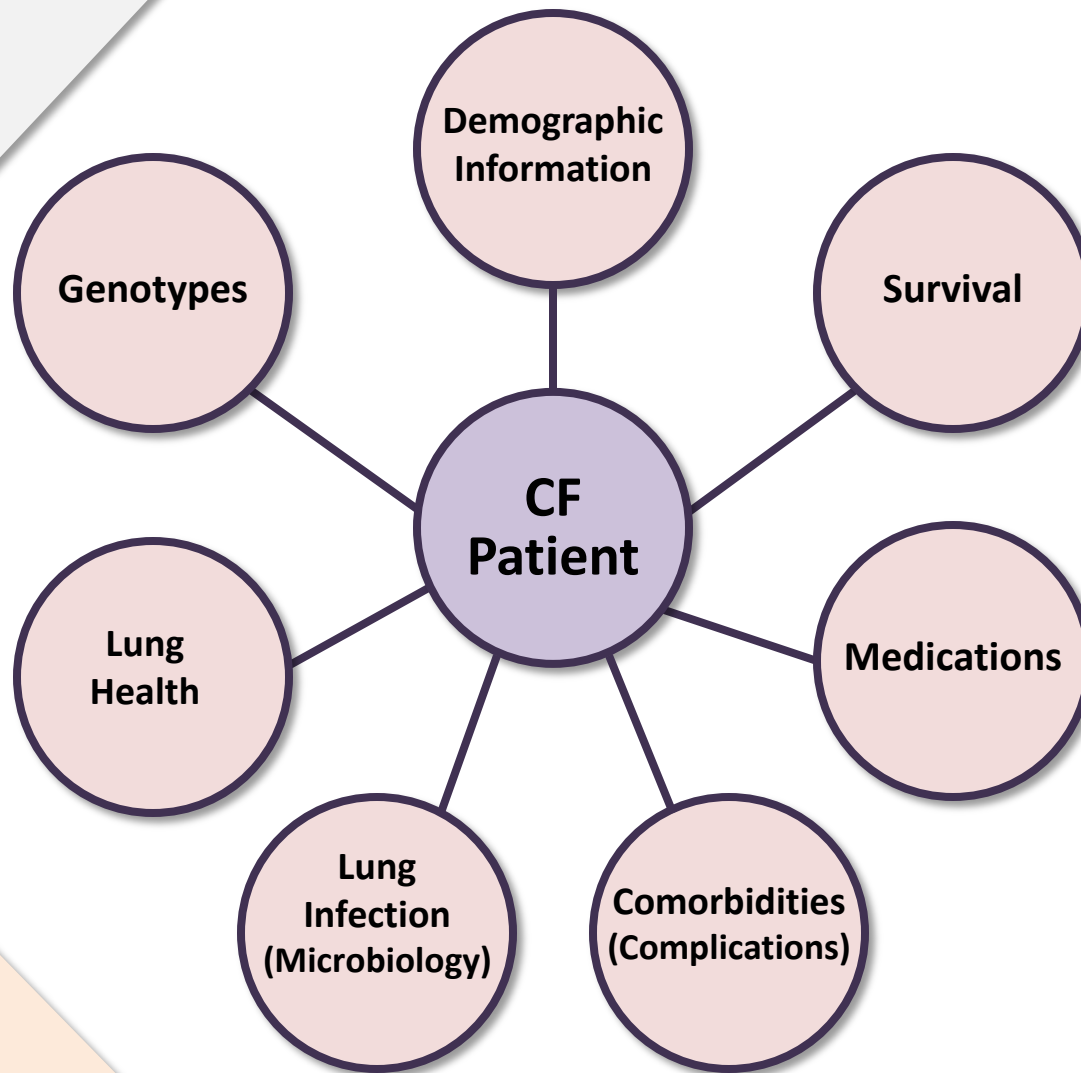
Detailed Data Analysis

Data-induced Hypotheses



# The Structure of the Data (I)

“ Every CF patient in the UK registry is linked with **7** different types of information ”

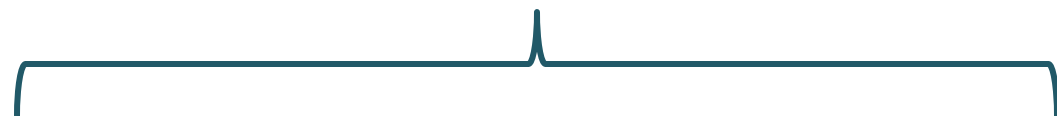




# The Structure of the Data (II)

The CF registry data spans the years **2008** to **2015**.

Repeated assessments of Lung function



IV Antibiotic  
Hospitalization

Genotypes  
Demographics

Bacterial  
Infection

Physiotherapy  
Complication

Bacterial  
Infection

2008

2009

2010

2011

2012

2013

2014

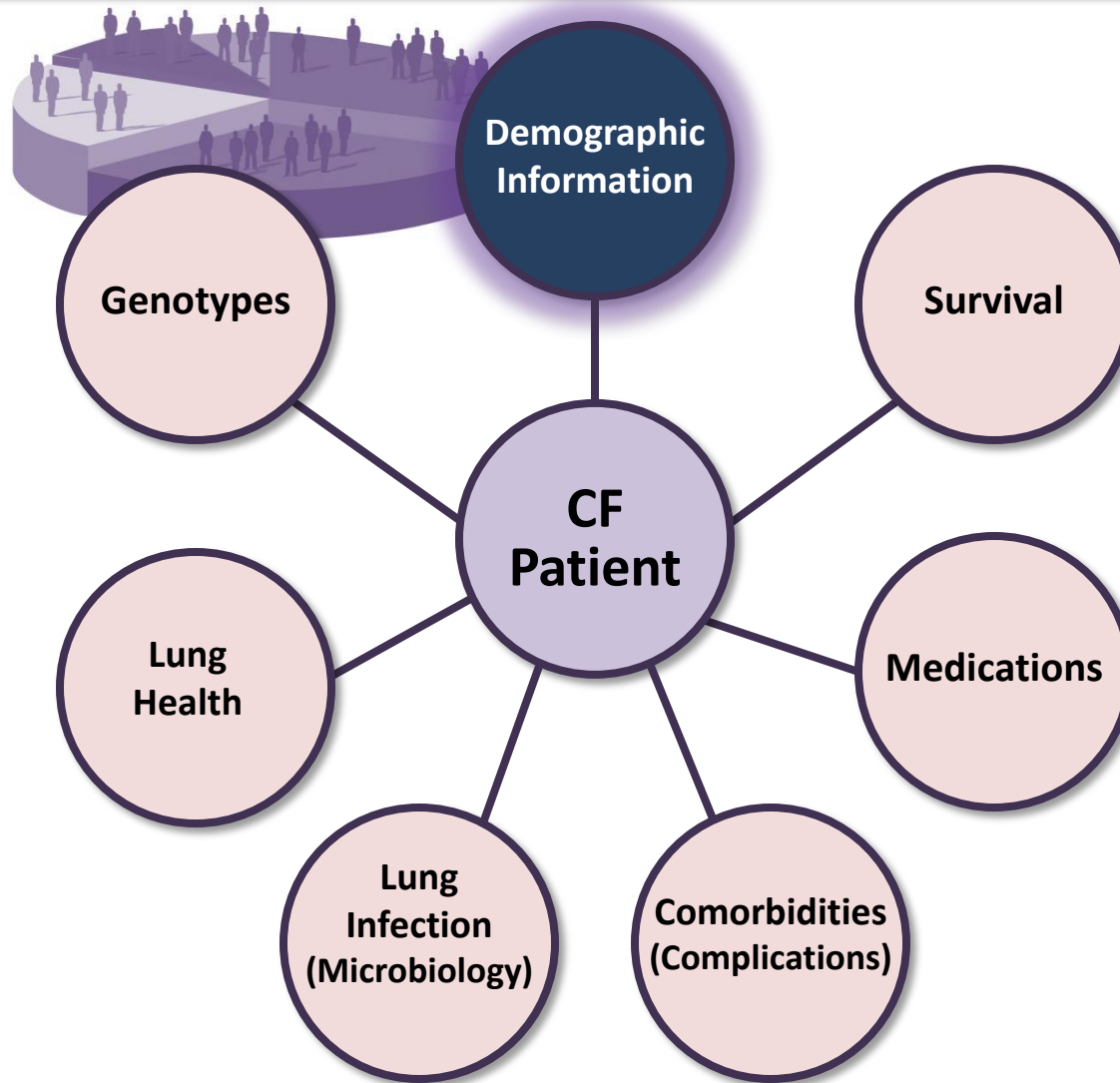
2015

An exemplary trajectory for a CF patient.

“ Every CF patient in the UK registry is linked with **7** different types of information ”

“ Every patient’s **temporal trajectory** is formed via **annual follow-ups** ”

# Data Analysis: Demographic Information



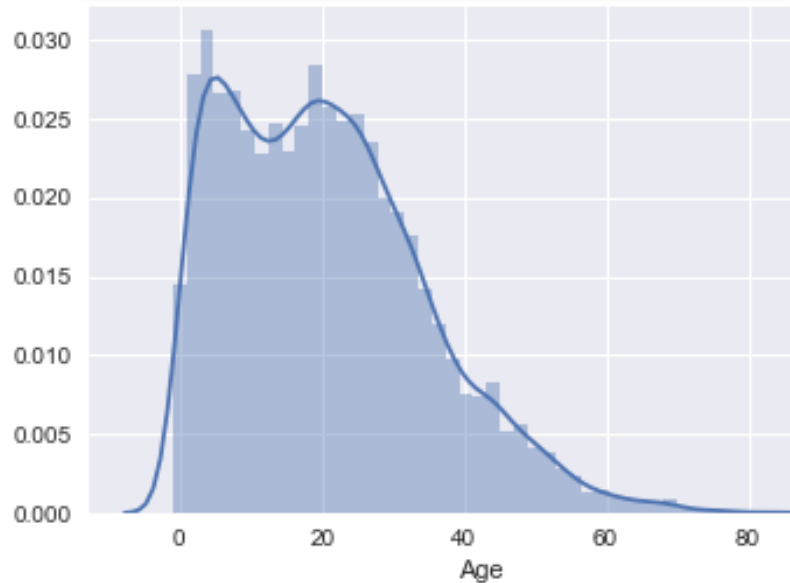
# Demographic Information: Age

- **Average age in 2015: male (21.1 years), female (20.1 years).**
- Male patients are **1 year older** on average.

**n = 9,587**

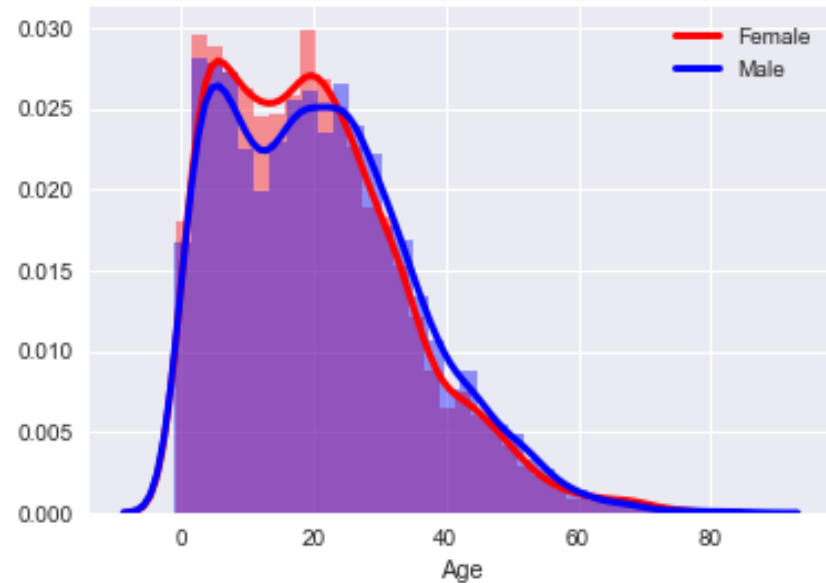
Average age = **20.63 years**

Median age = **19.16 years**



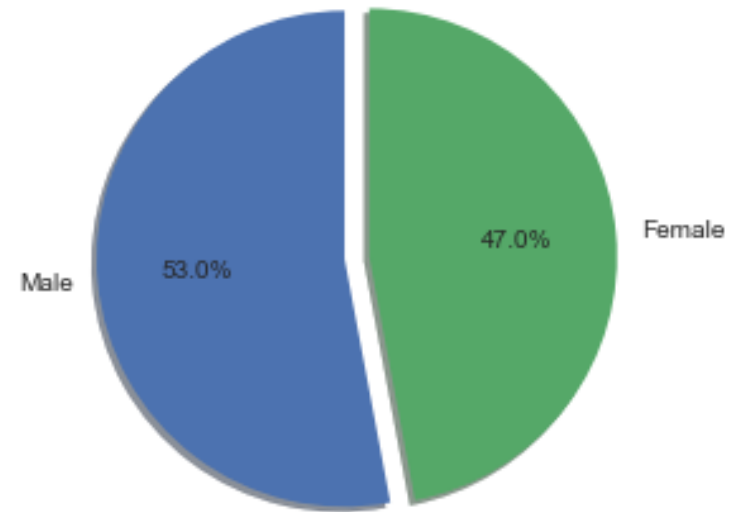
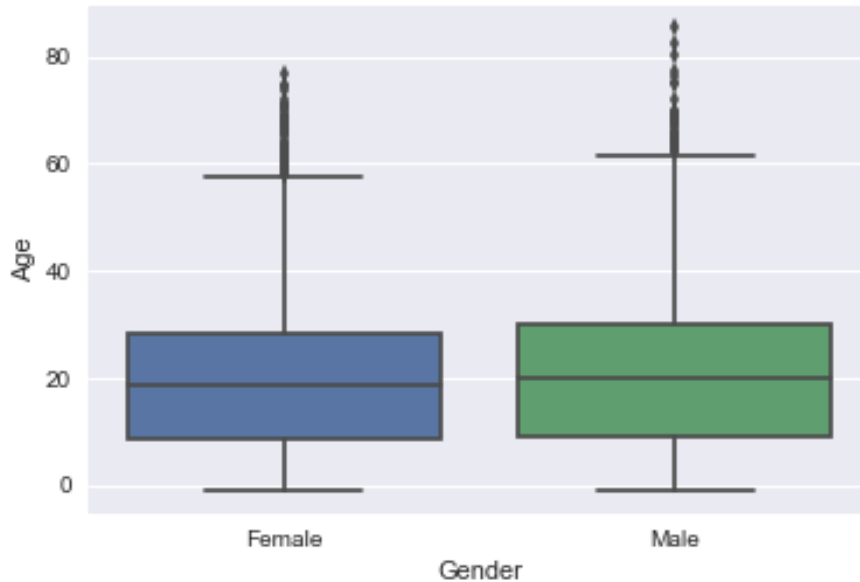
**Welch's t-test**

$t = 3.48732$ ,  $p\text{-value} = 0.00049$



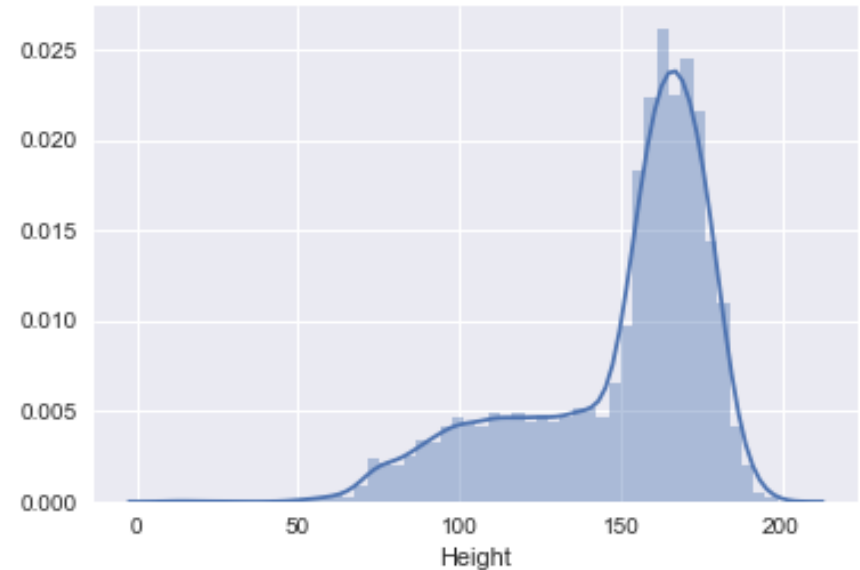
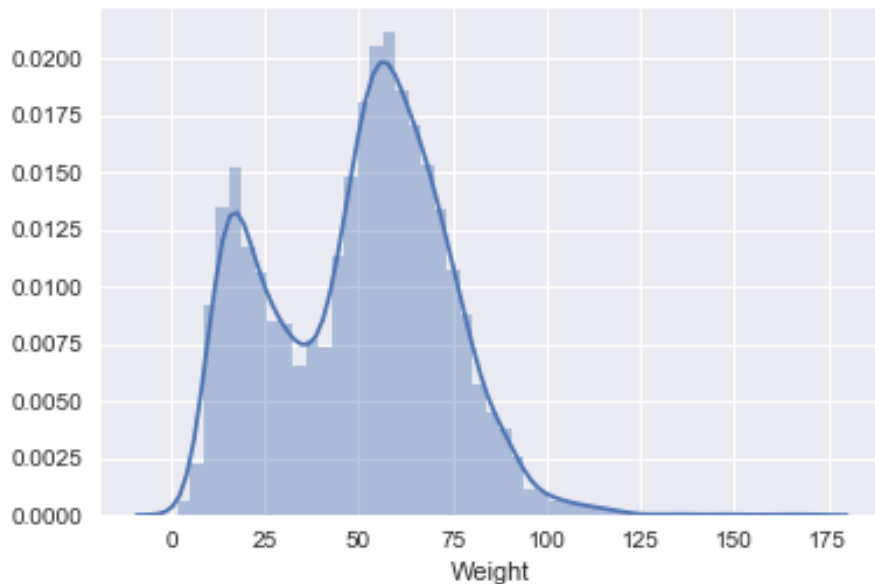
# Demographic Information: Gender

- **Gender distribution (2015):** male (53%), female (47%).



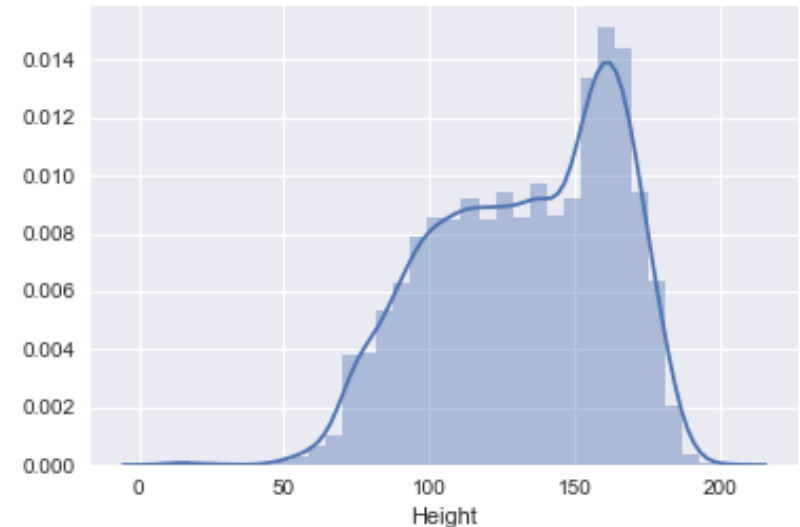
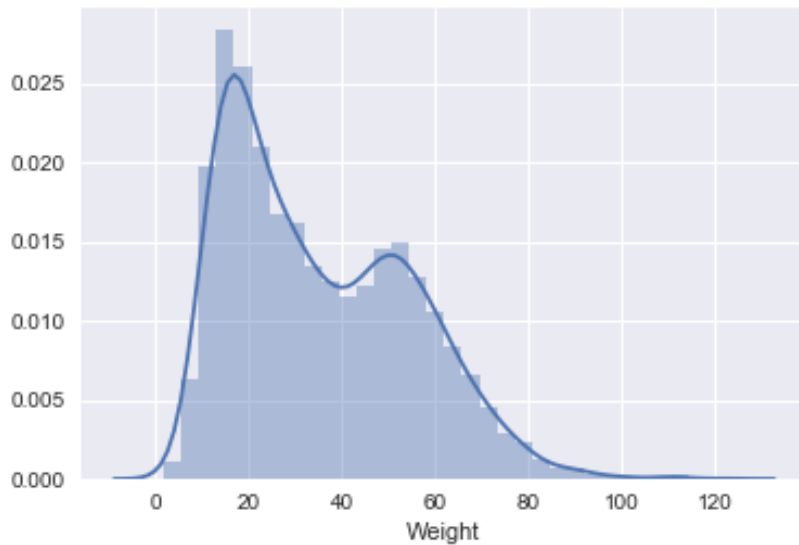
# Demographic Information: Weight and Height (I)

- Distribution of **weights** and **heights** of all patients in the UK CF registry in 2015.



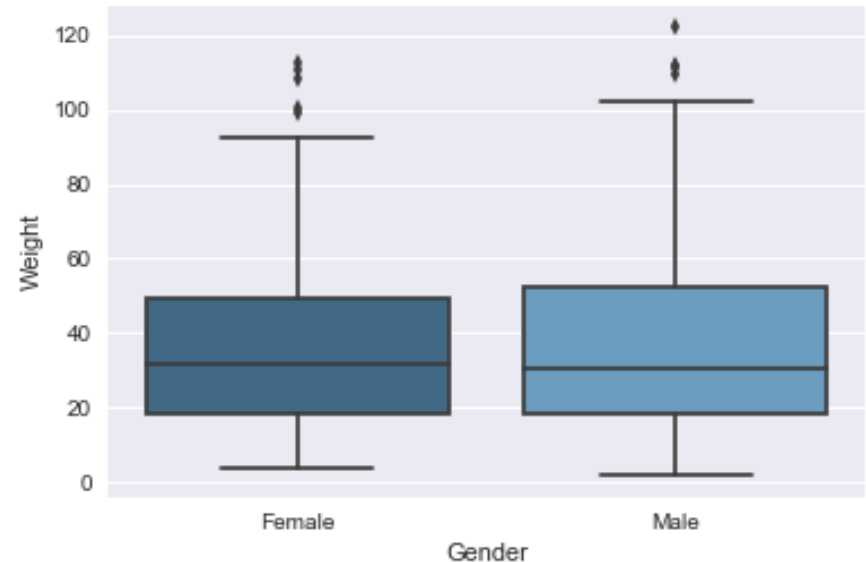
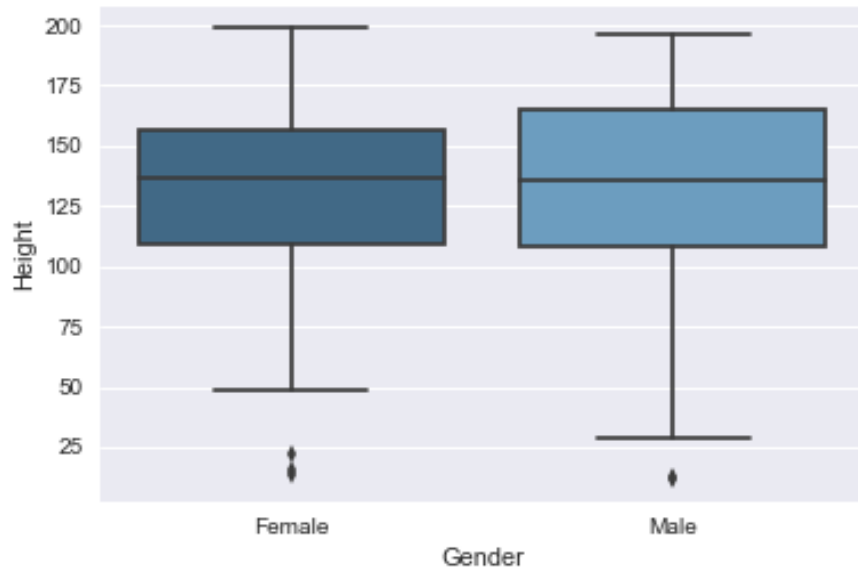
# Demographic Information: Weight and Height (II)

- Distribution of **weights** and **heights** for children and young people (< 20 years, n = 4,481) in 2015.



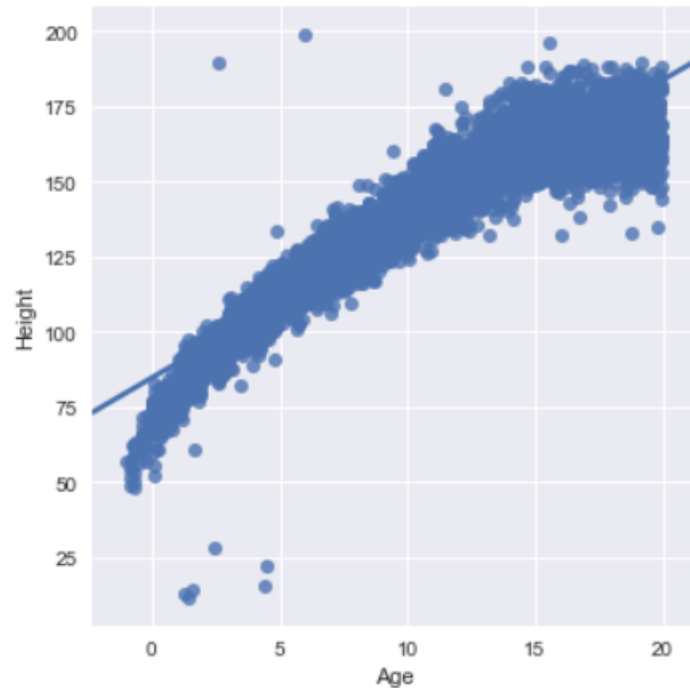
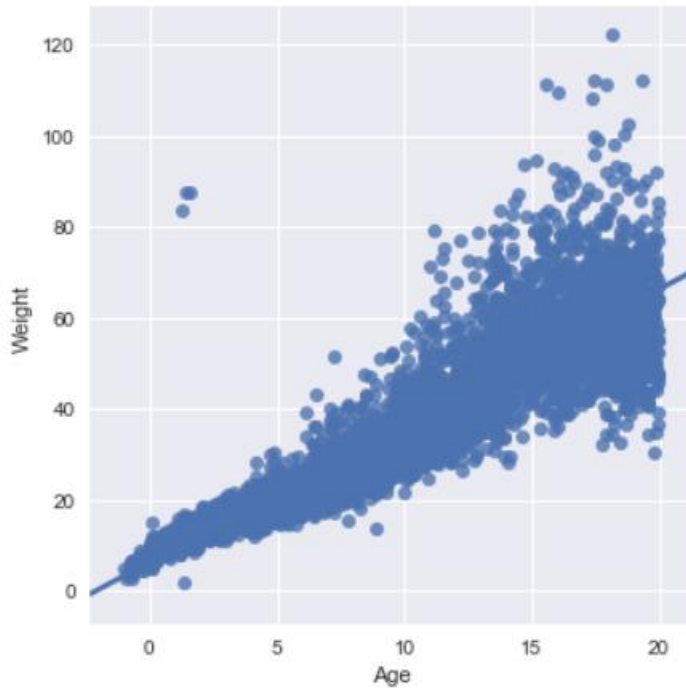
# Demographic Information: Weight and Height (III)

- Boxplots for **weights** and **heights** for children and young people (< 20 years, n = 4,481) stratified by gender.



# Demographic Information: Weight and Height (IV)

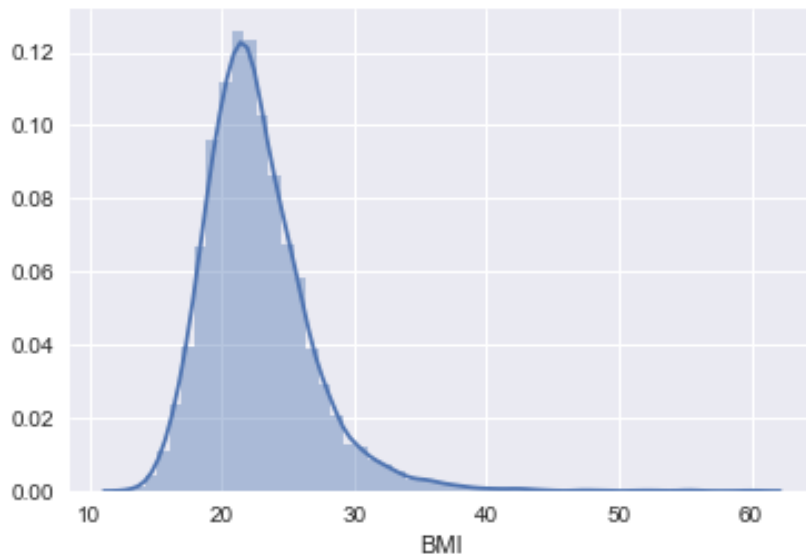
- Growth trajectory (weight and height) for **CF patients**.





# Demographic Information: Body Mass Index (I)

- Distribution of the **BMI** for patients in the registry (2015)

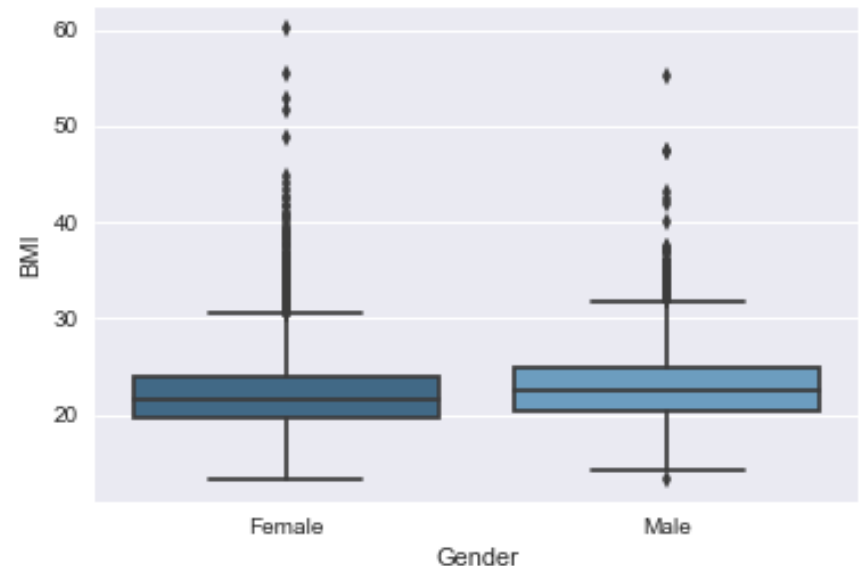


**n = 9,587**

Average BMI = **22.67 kg/m<sup>2</sup>**

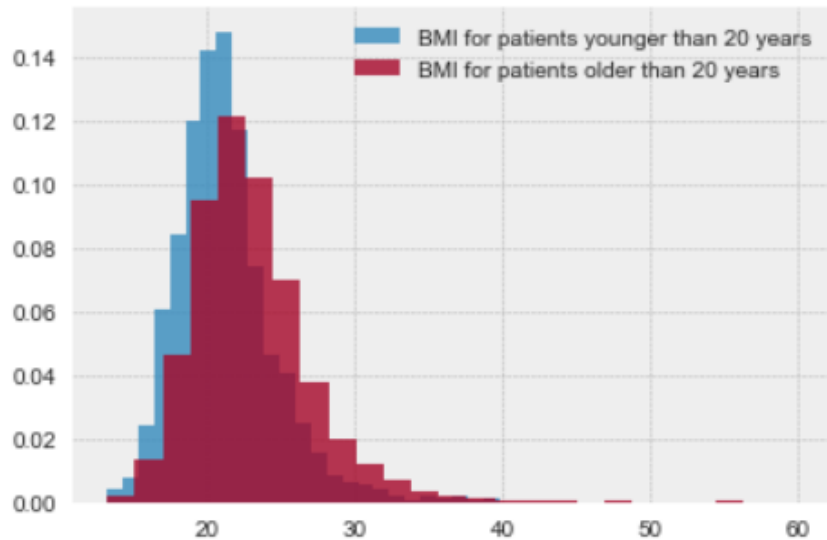
Median BMI = **22.08 kg/m<sup>2</sup>**

- Boxplots for the **BMI** of CF patients stratified by gender

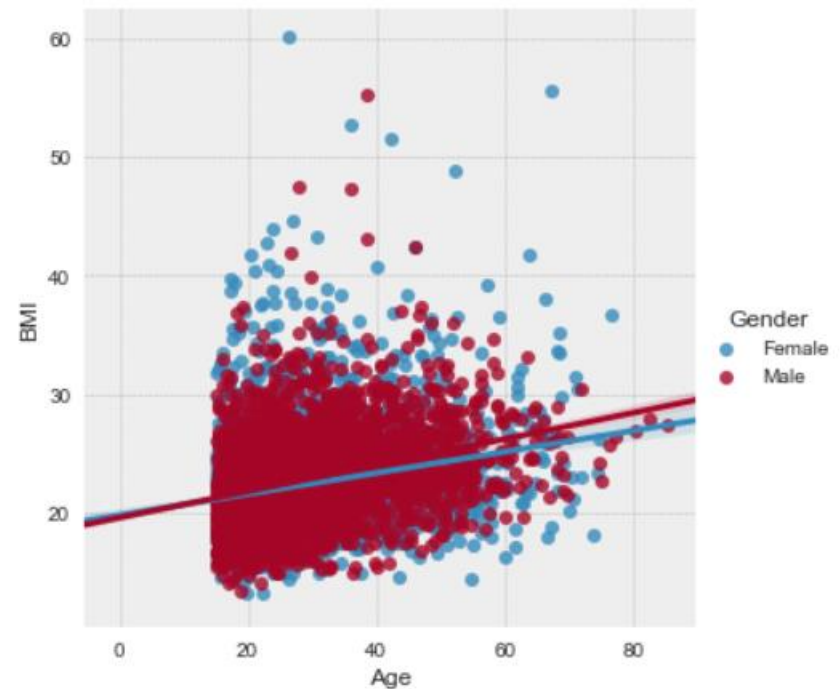


# Demographic Information: Body Mass Index (II)

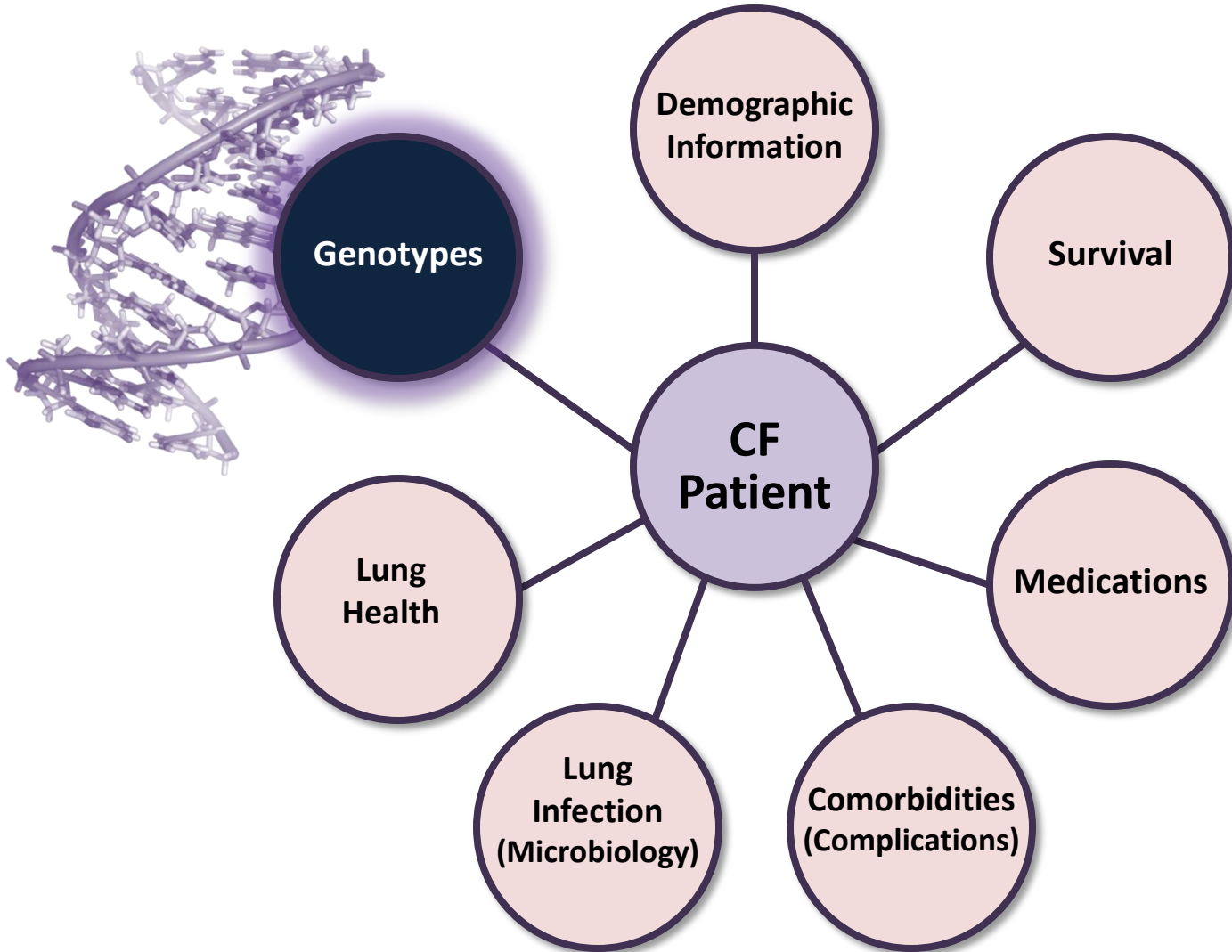
- Distribution of the **BMI** for the adults and young patient groups



- **BMI** trajectories stratified by gender



# Genotyping and Genetic Mutations

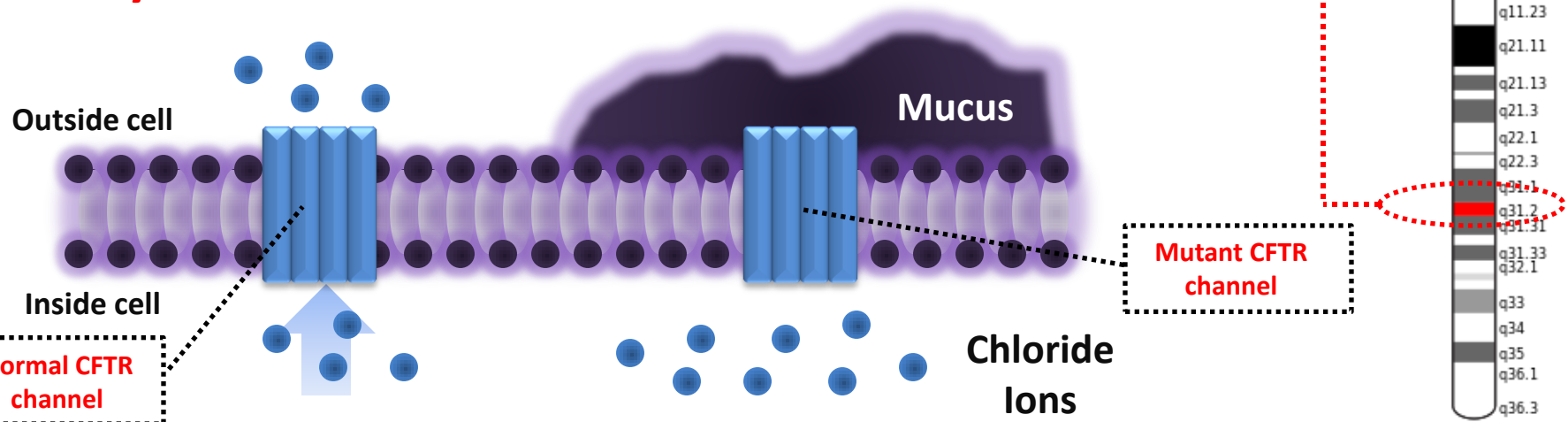


# Cystic Fibrosis: An Epidemiological Perspective

- **Cystic fibrosis** is the most common life-limiting **autosomal recessive disease** among Caucasians. [Tobias, 2011]
- **Incidence:**
  - ❑ **UK:** 1/2,500 live births and **one in every 25** people is a carrier.
  - ❑ **USA:** **one in every 30** people is a carrier.
  - ❑ Much less prevalent in people with African and Asian descent.
- **For reasons that remain unclear:** males tend to have a longer life expectancy than females. [Rosenfeld et. al, 1997], [Coakley, 2008]
- CF is caused by the malfunctioning of the **Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)**.

# Cystic Fibrosis Transmembrane Conductance Regulator (CFTR)

- The gene encoding the **CFTR** protein is on **chromosome 7** (position **q31.2**).
- **Mutations** of the **CFTR** gene may affect the functionality of the chloride ion channel. This can lead to dysregulation of epithelial fluid transport in the lung, pancreas and other organs, resulting in **cystic fibrosis**.



# Mutations of the CFTR Gene

- **Mutations of the CFTR** gene may affect the functionality of the chloride ion channel. This can lead to dysregulation of epithelial fluid transport in the lung, pancreas and other organs, resulting in **cystic fibrosis**.
- **2,019 mutations** can cause CF. [CFMDB Statistics, 2014]

Most common mutations among Caucasians [Araújo et. al, 2005]

**ΔF508**

**G542X**

**G551D**

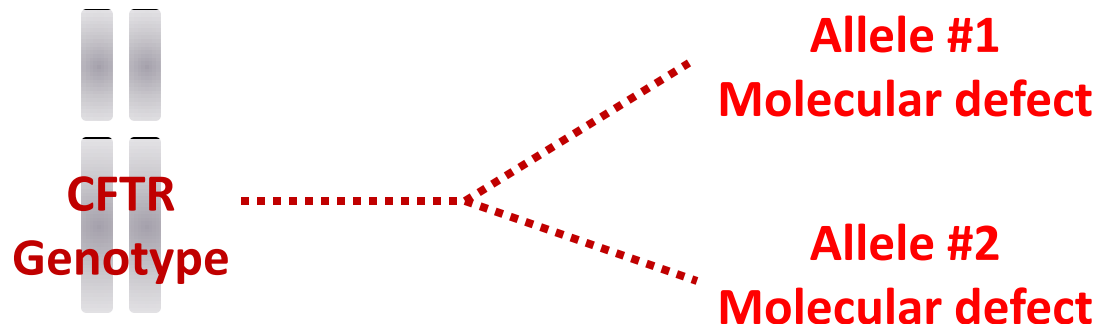
**N1303K**

**W1282X**

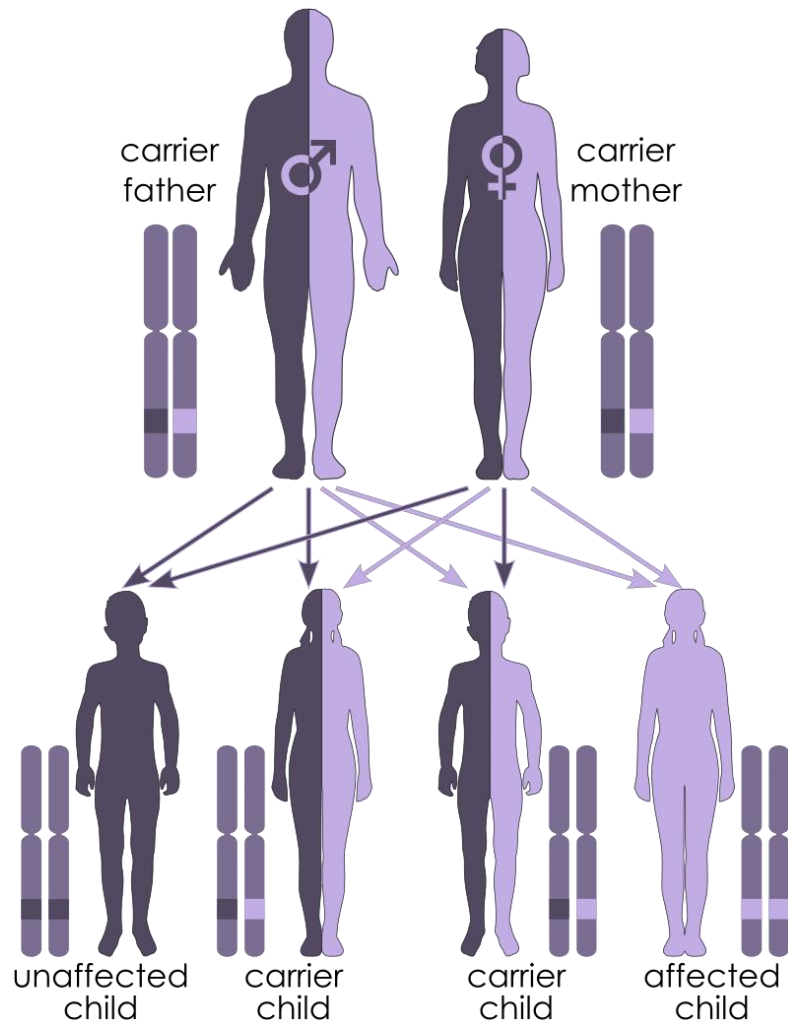
The most common  
mutation

# Genotypes and Genetic Mutations (I)

- **Genotypes** control the CF phenotypic characteristics.
- Genotypes reveal which **mutations** of the CF genes caused CF for a particular patient.
- **Every CF patient has two mutations of the gene for CFTR:** one on each allele (one inherited from the mother and one from the father).



# Genotypes and Genetic Mutations (II)



CF follows a simple **Mendelian** (autosomal recessive) inheritance model.

A CF patient is **homozygous** if both mutations are the same.

A CF patient is **heterozygous** if she/he has two different mutations.



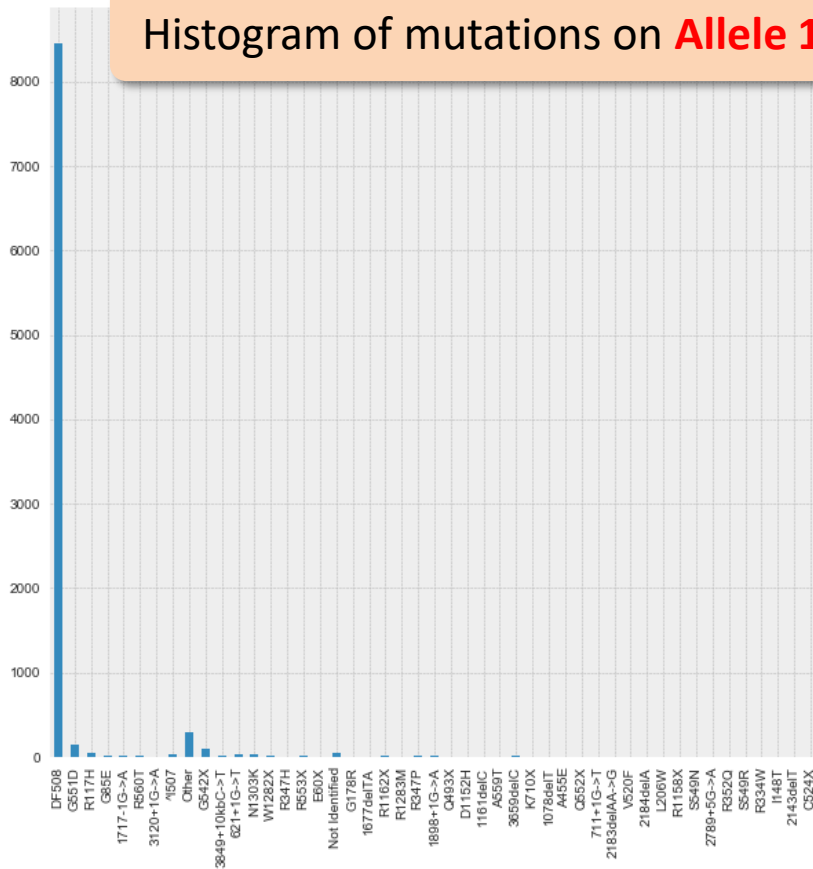
# Genetic Mutations Data Analysis (I)

- A total of **9,401** Cystic Fibrosis patients in the registry are genotyped (98.05%).
- A total of **8,507** patients had **ΔF508** mutations (90.49%):
  - ❑ Homozygous **ΔF508** mutations: 4,728 patients (50.29%)
  - ❑ Heterozygous **ΔF508** mutations: 3,779 patients (40.19%)
- Among the **2,019** mutations that are known to cause CF, only **66** mutations were frequent in the registry.

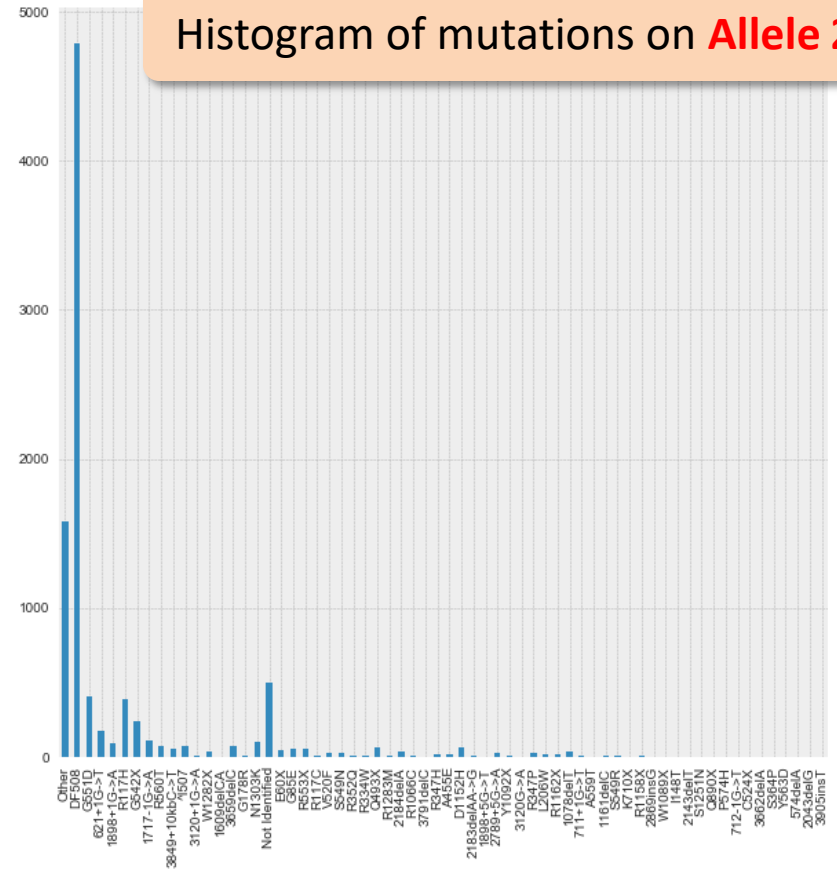
# Genetic Mutations Data Analysis (II)

- A total of **8,507** patients had  **$\Delta F508$**  mutations (**90.49%**).

Histogram of mutations on **Allele 1**



Histogram of mutations on **Allele 2**



# Genetic Mutations Data Analysis (III)

- Genetic mutation counts (in both Alleles) for CF patients in the registry. (counts are for the 20 most frequent mutations.)

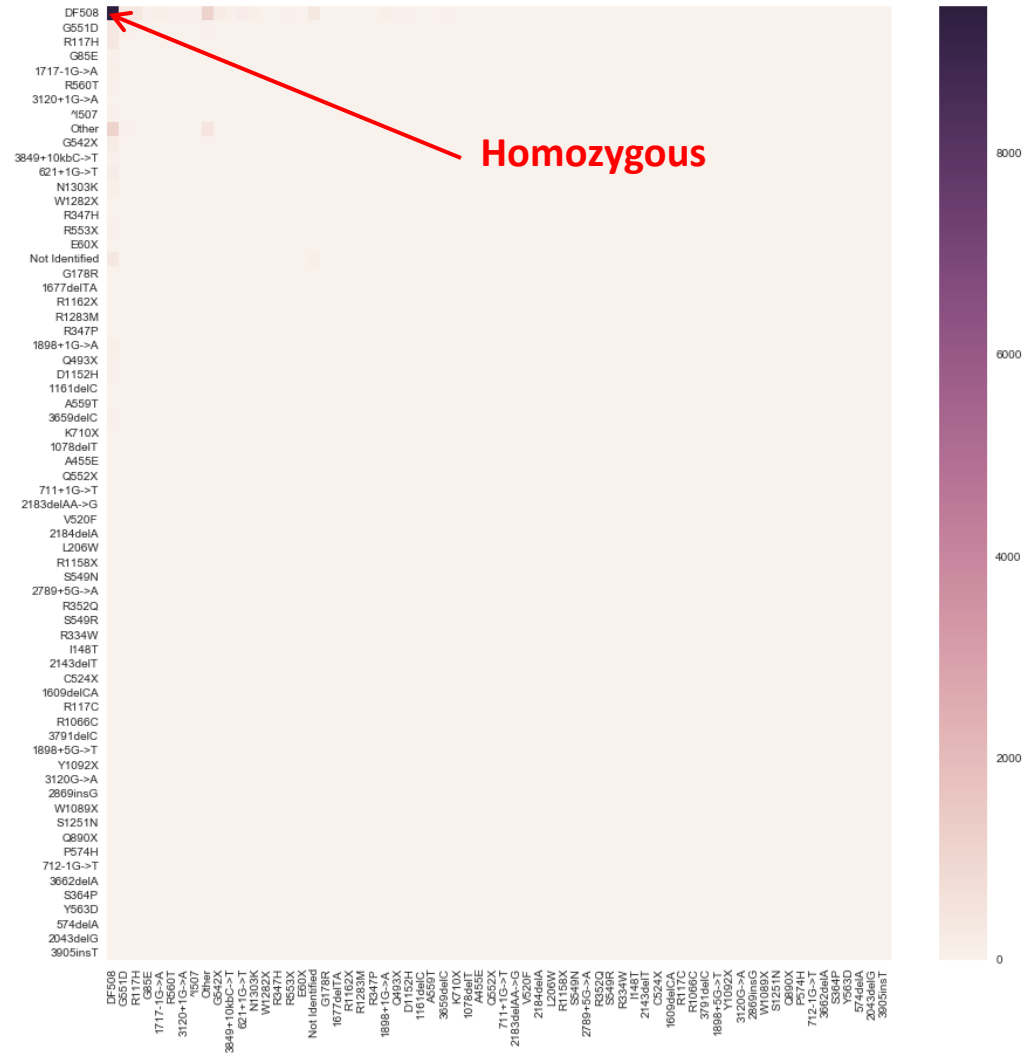
| Mutation             | Count  | Mutation                | Count |
|----------------------|--------|-------------------------|-------|
| <b>ΔF508</b>         | 13,235 | <b>3659delC</b>         | 85    |
| <b>G551D</b>         | 552    | <b>3849+10kbC-&gt;T</b> | 80    |
| <b>R117H</b>         | 438    | <b>R553X</b>            | 79    |
| <b>G542X</b>         | 329    | <b>D1152H</b>           | 75    |
| <b>621+1G-&gt;T</b>  | 213    | <b>G853</b>             | 73    |
| <b>N1303K</b>        | 135    | <b>Q493X</b>            | 69    |
| <b>1717-1G-&gt;A</b> | 119    | <b>E60X</b>             | 56    |
| <b>1898+1G-&gt;A</b> | 112    | <b>W1282X</b>           | 52    |
| <b>ΔI507</b>         | 102    | <b>1078delT</b>         | 46    |
| <b>R560T</b>         | 89     | <b>2184delA</b>         | 35    |

# Genetic Mutations Data Analysis (IV)

Co-occurrence counts for mutations on alleles 1 and 2.

Most common mutation pairs

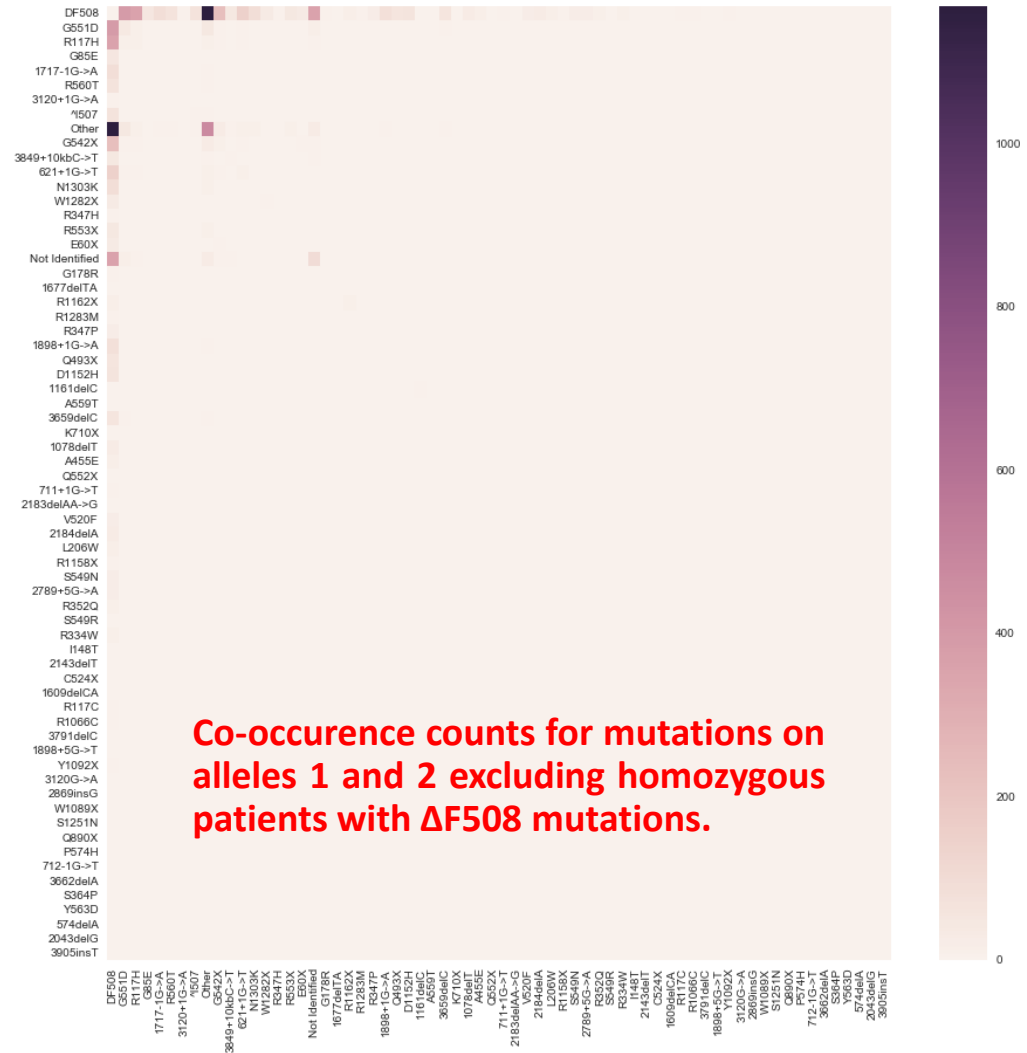
| Allele 1      | Allele 2      | Count |
|---------------|---------------|-------|
| $\Delta F508$ | $\Delta F508$ | 4,728 |
| $\Delta F508$ | G551D         | 392   |
| $\Delta F508$ | R117H         | 359   |
| $\Delta F508$ | G542X         | 226   |
| $\Delta F508$ | 621+1G->T     | 148   |
| $\Delta F508$ | 1717-1G->A    | 119   |



# Genetic Mutations Data Analysis (V)

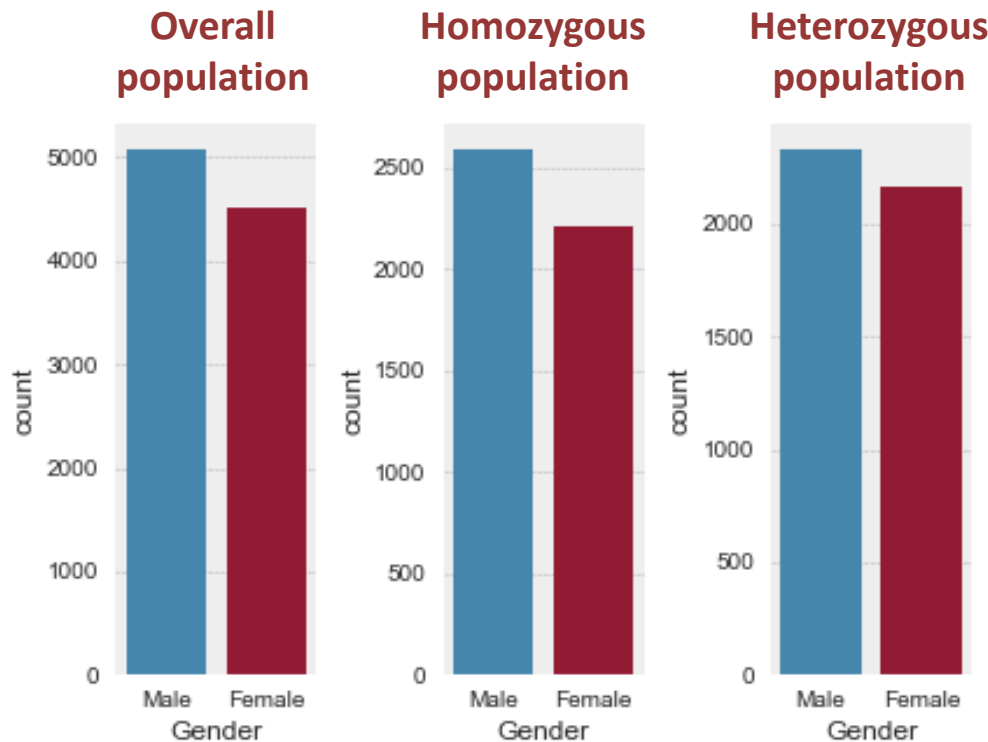
## Most common mutation pairs

| Allele 1      | Allele 2      | Count |
|---------------|---------------|-------|
| $\Delta F508$ | $\Delta F508$ | 4,728 |
| $\Delta F508$ | G551D         | 392   |
| $\Delta F508$ | R117H         | 359   |
| $\Delta F508$ | G542X         | 226   |
| $\Delta F508$ | 621+1G->T     | 148   |
| $\Delta F508$ | 1717-1G->A    | 119   |



# Genetic Mutations Data Analysis (VI)

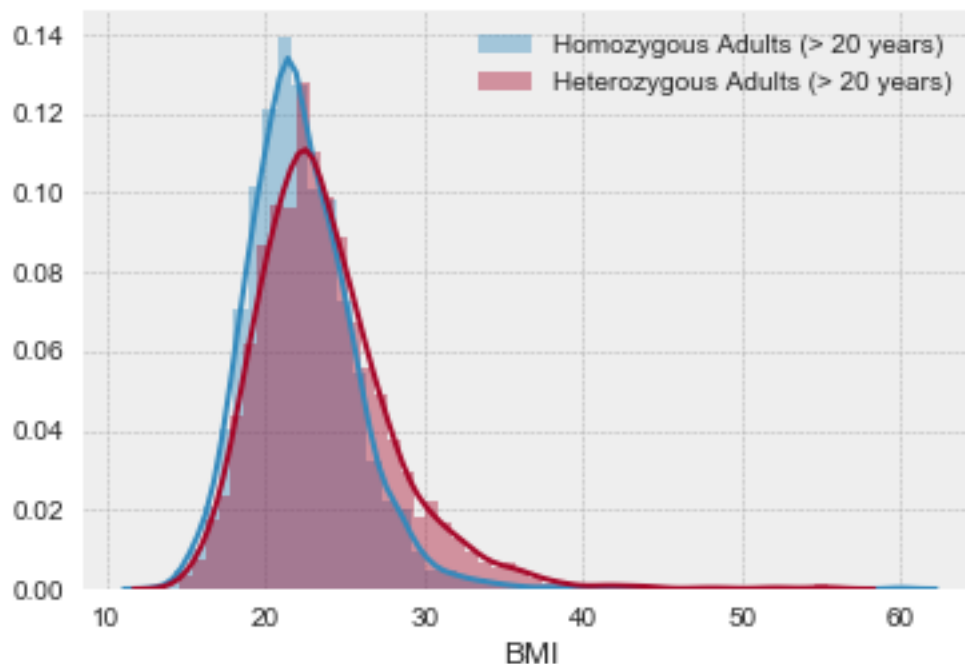
- Among the **9,587** CF patients registered in **2015**:
  - ❑ **4,801** are known to be **homozygous**
  - ❑ **4,500** are known to be **heterozygous**



The females' share in the heterozygous population is significantly larger than their share in the homozygous population

# Genetic Mutations Data Analysis (VII)

- There is a statistically significant difference in the average BMI of adults (> 20 years) in the **homozygous** and **heterozygous** populations. (**p-value** < 0.0001 for a **Welch test**.)



Average BMI for adults in the homozygous population =

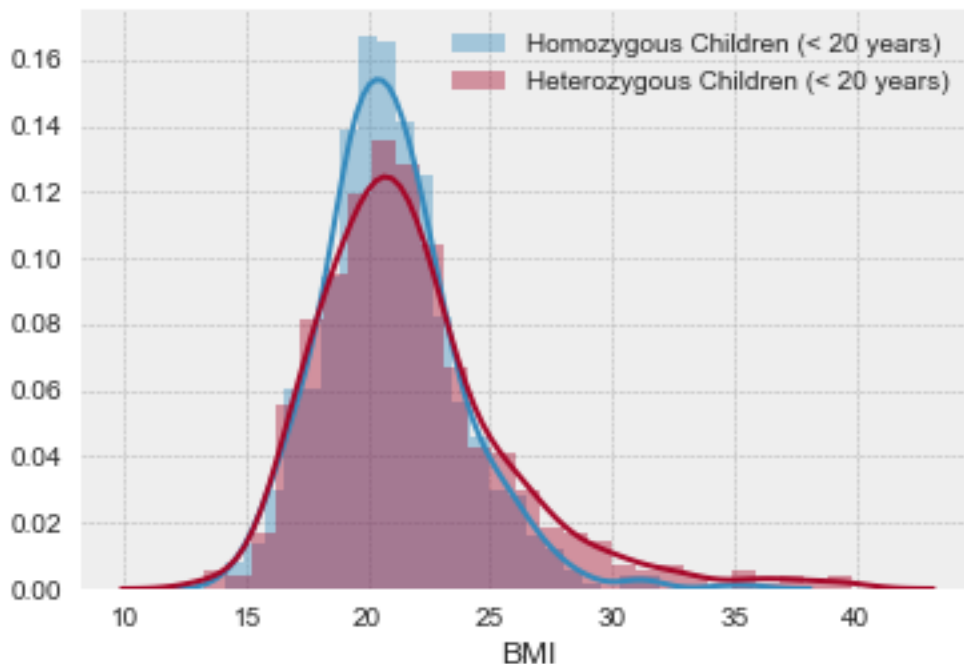
**22.24 kg/m<sup>2</sup>**

Average BMI for adults in the heterozygous population =

**23.74 kg/m<sup>2</sup>**

# Genetic Mutations Data Analysis (VIII)

- There is a statistically significant difference in the average BMI of children (< 20 years) in the **homozygous** and **heterozygous** populations. (**p-value** < 0.0001 for a **Welch test**.)



Average BMI for children in the homozygous population =

**20.92 kg/m<sup>2</sup>**

Average BMI for children in the heterozygous population =

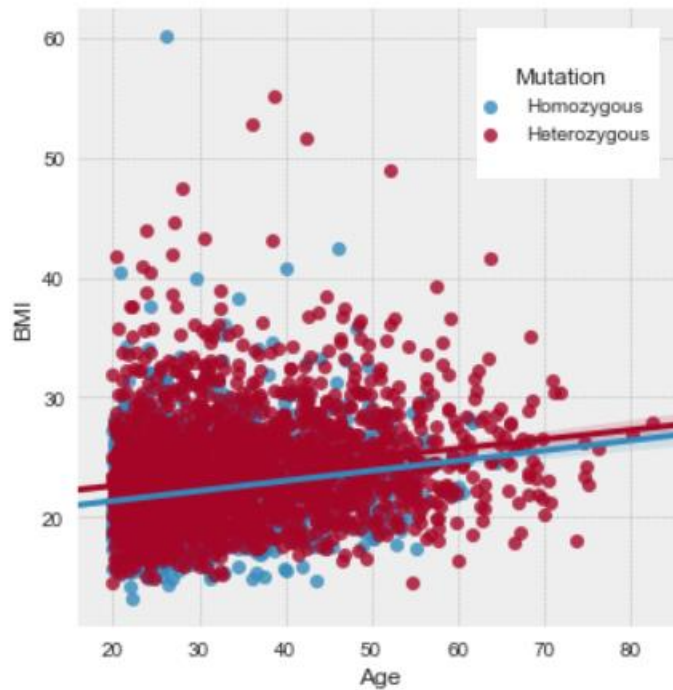
**21.72 kg/m<sup>2</sup>**



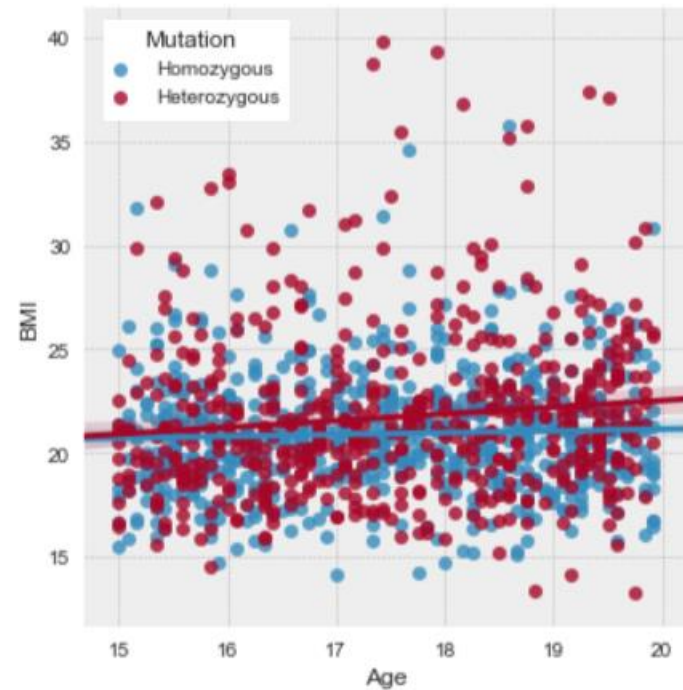
# Genetic Mutations Data Analysis (IX)

- The growth (BMI) trajectories for **heterozygous** CF patients are faster than those of **homozygous** CF patients.

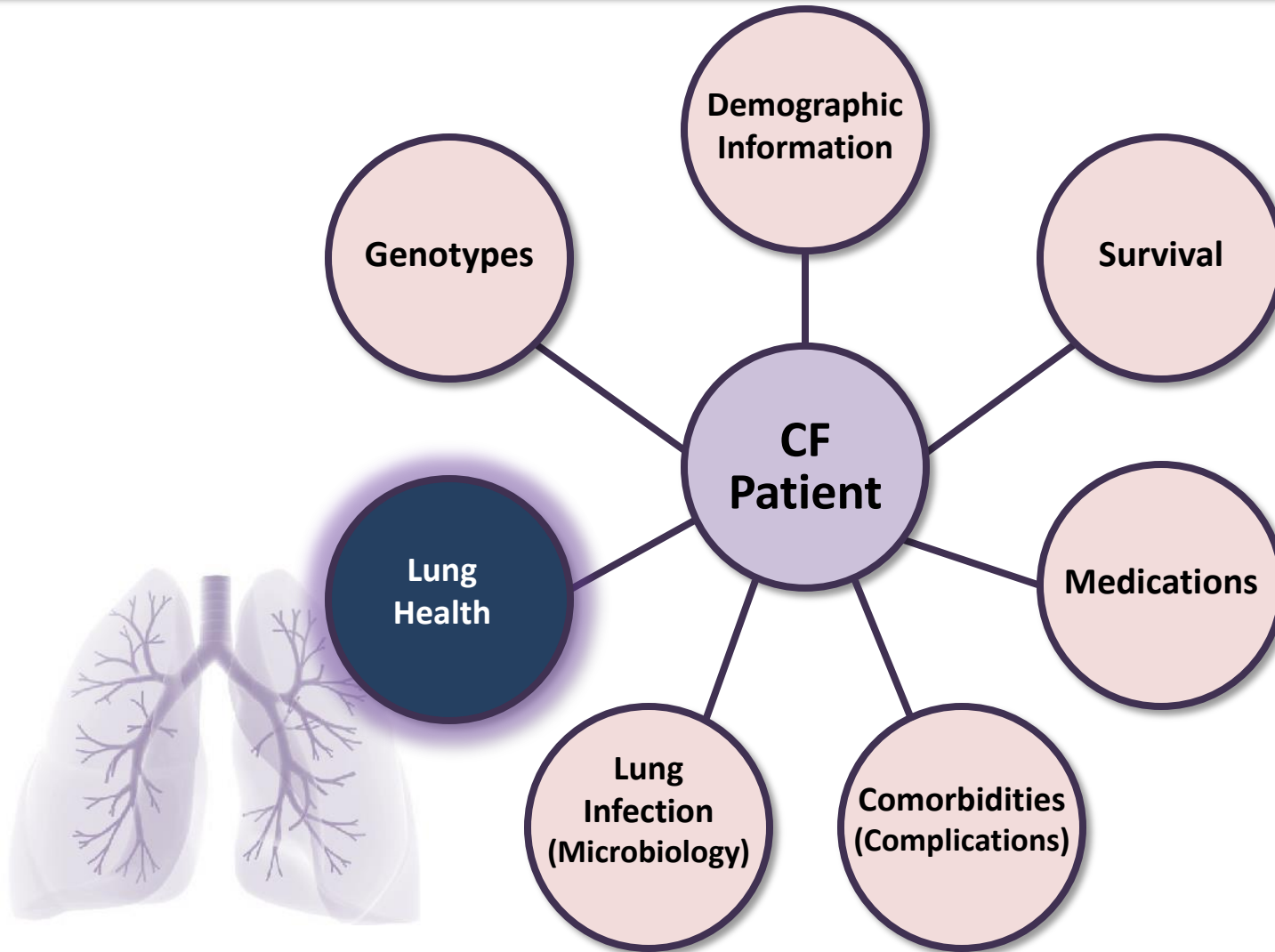
Adult CF Patients  
( $\geq 20$  years)



Children and Young CF Patients  
( $< 20$  years)



# Data Analysis: Lung Health

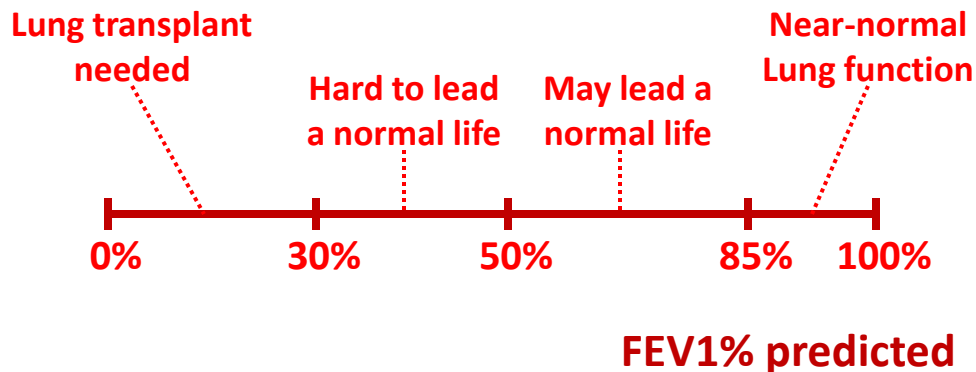


# FEV1 as a Measure of Lung Function

- Condition of the lungs is measured using **FEV1** (Forced Expiratory Volume of air in the first second of an exhaled breath).
- **FEV1% predicted** is based on the **FEV1** expected for a person without CF of the same age, gender, height, and ethnicity.
  - ❑ **FEV1% predicted of 50%** means the CF patient breathes out **half** the volume of air as a comparable person without CF.
  - ❑ **FEV1% predicted** is calculated using the **Global Lung function Initiative equation (GLI)**.

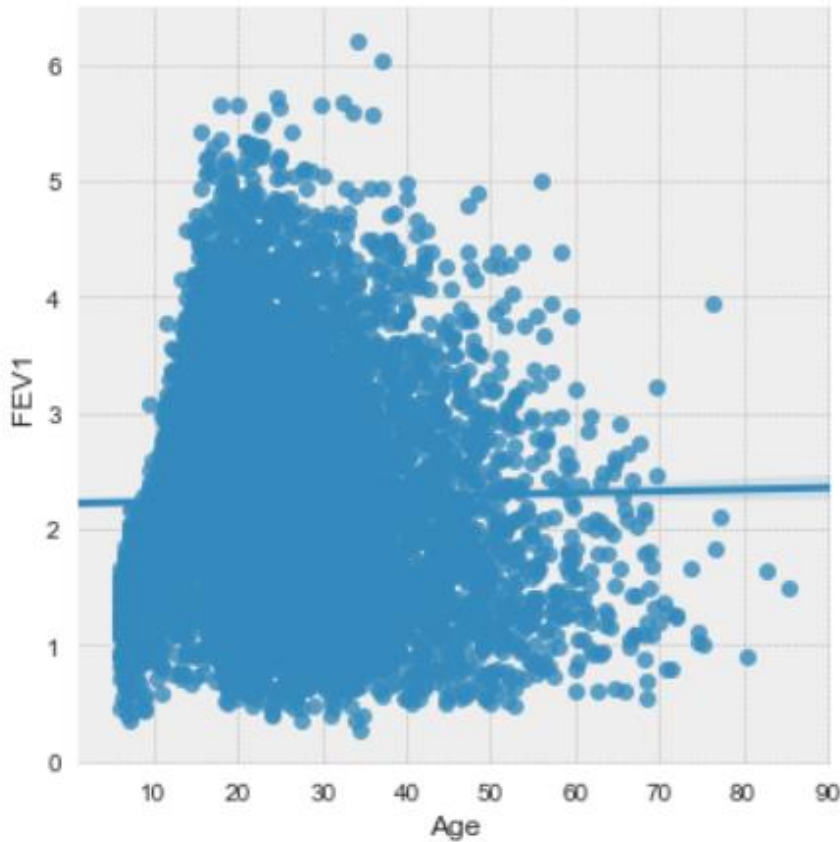
## Target of Care

Maintaining an **FEV1% predicted** of **85%** or higher!

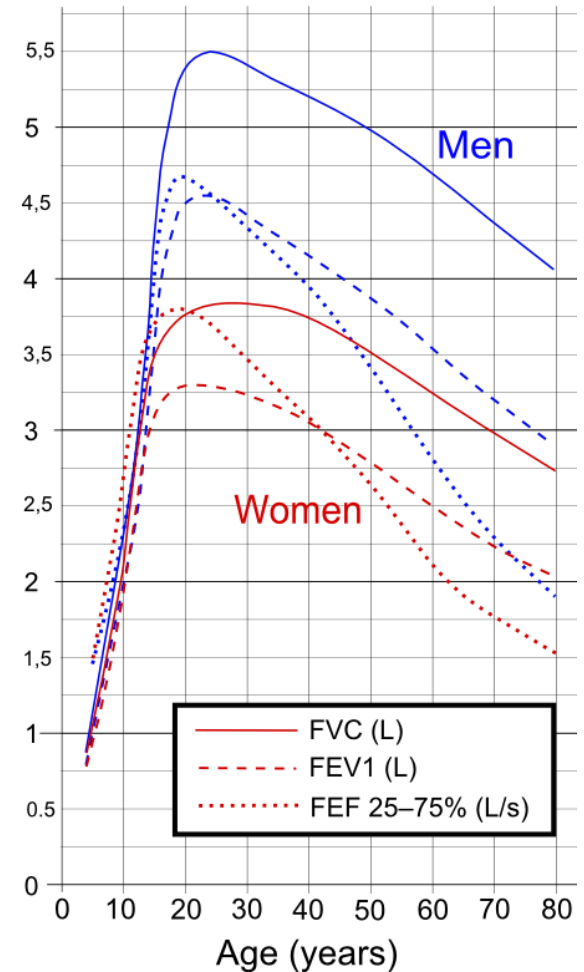


# FEV1 Data Analysis (I)

- Scatterplot for the **raw FEV1 values** for patients aged 6 years and over.



Normal values for FVC, FEV1 and FEV 25-75%



# FEV1 Data Analysis (II)

- General trend:** FEV1 % predicted deteriorates with age.

Current status of care

What is the fraction of patients for whom the **target FEV1** is met?

**36.80%** (2,849 patients)

What are the **experiences/needs** of current CF patients?

5.42%

Need a Lung transplant

15.03%

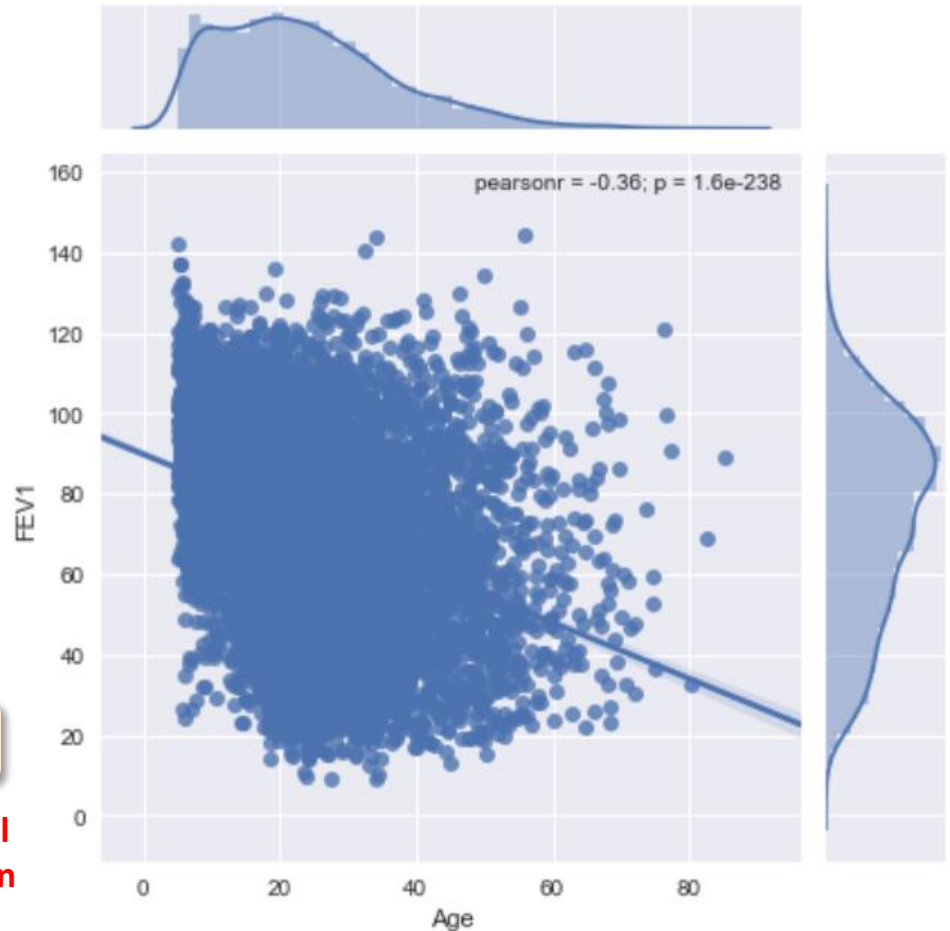
Hard to lead a normal life

42.73%

May lead a normal life

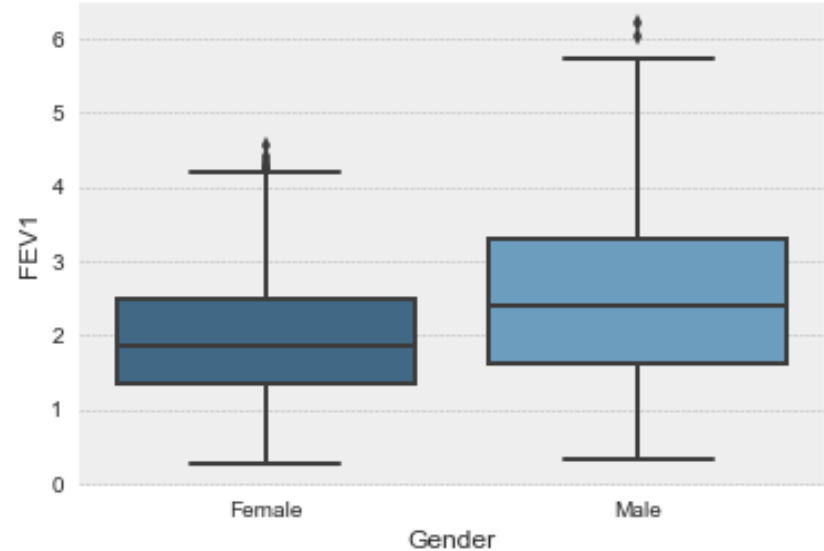
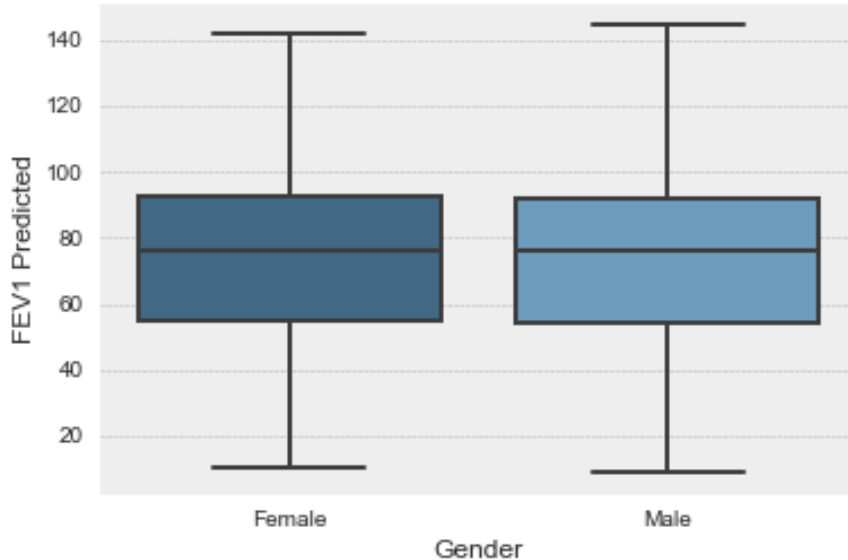
36.80%

Near-normal Lung function



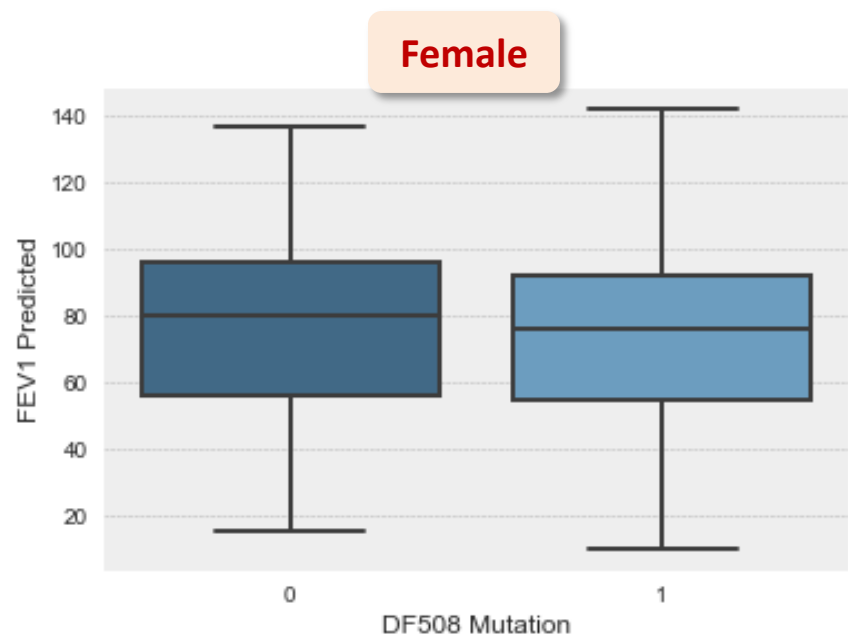
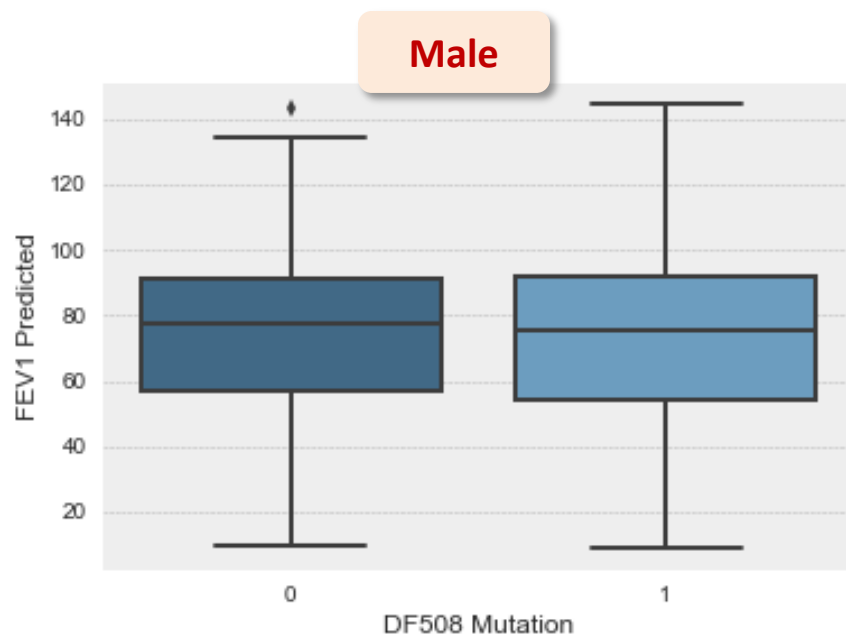
# FEV1 Data Analysis (III)

- Raw FEV1 and FEV1 % predicted stratified by **gender**.
- **No evidence** that either genders experience a better FEV1 outcomes.
  - Longer male survival is inexplicable via FEV1 markers alone.



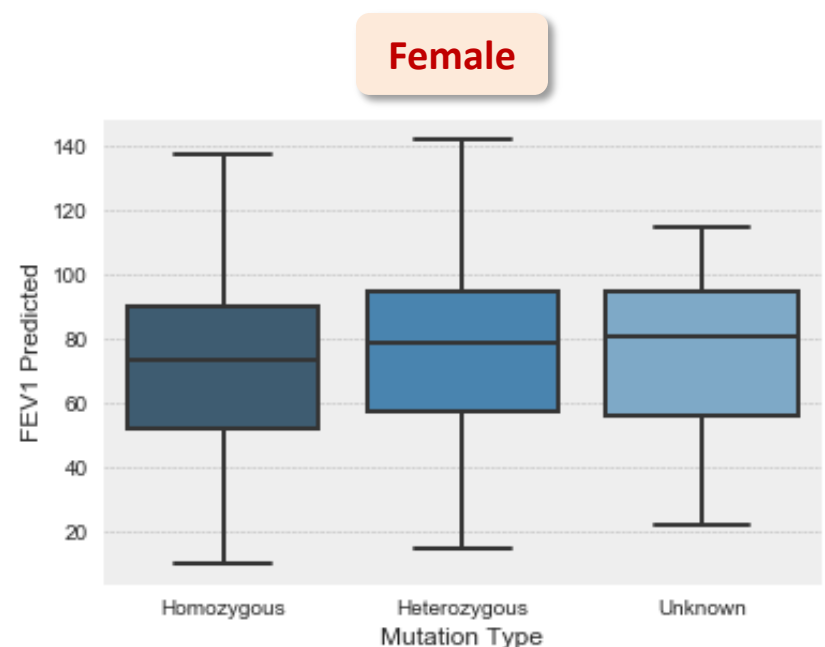
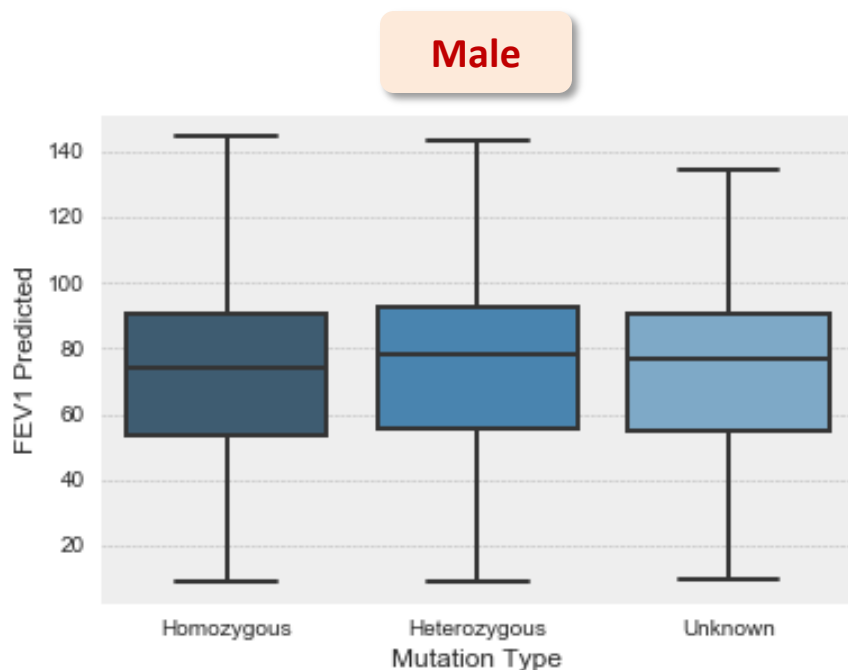
# FEV1 Data Analysis (IV)

- FEV1 % predicted stratified by gender and  $\Delta F508$  mutation.
  - **No evidence** that existence of  $\Delta F508$  mutation is relevant for lung function in males.
  - **Thin evidence** that  $\Delta F508$  mutation leads to worse outcomes for females.
  - Longer male survival may have a genetic explanation.



# FEV1 Data Analysis (V)

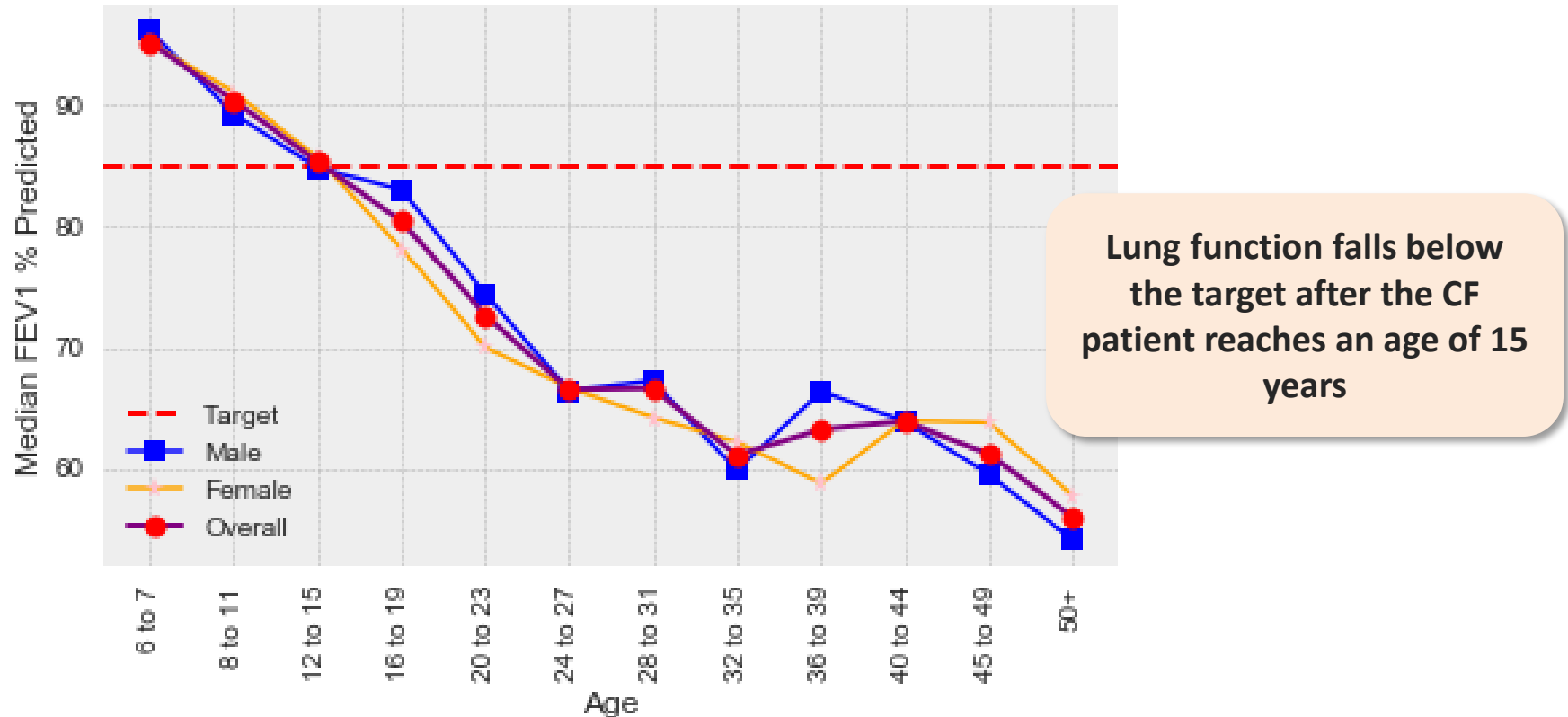
- FEV1 % predicted for homozygous and heterozygous patient groups stratified by gender.
  - **Thin evidence** that **heterozygous** mutations are advantageous for both male and female CF patient groups.



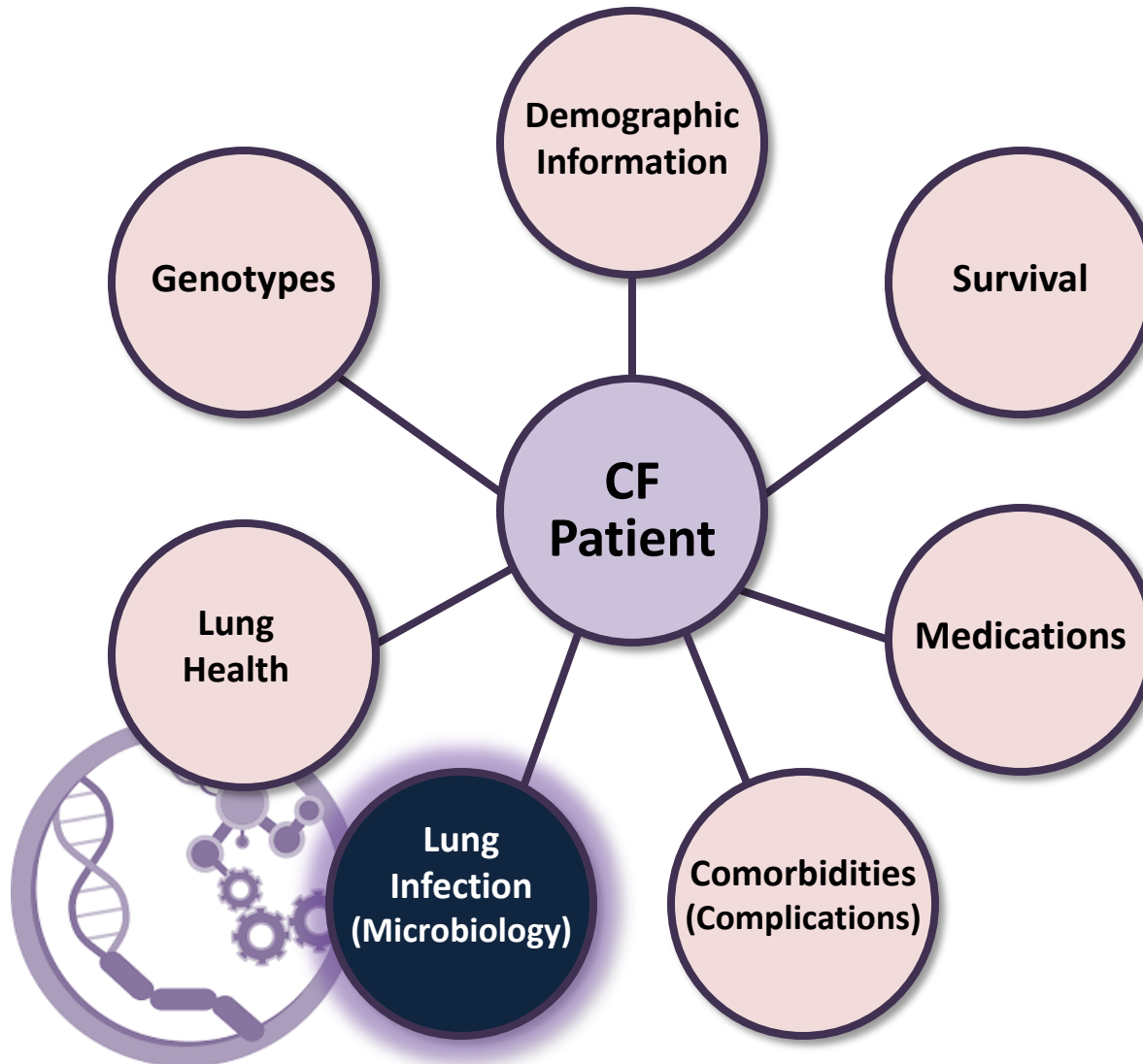


# FEV1 Data Analysis (VI)

- FEV1 % predicted among CF patients aged 6 years and over. (patients who had lung transplants are excluded, **n= 7,689.**)



# Data Analysis: Microbiology



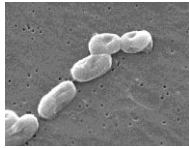
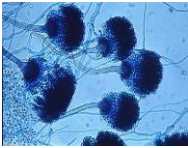

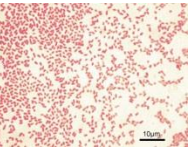





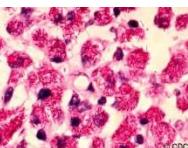
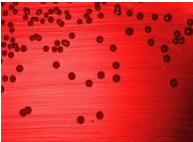

# Data Analysis: Microbiology (Lung Infections)

- CF patients are susceptible to **bacterial** and **fungal infections** which can reduce lung function.
- CF patients receive regular courses of **intravenous antibiotics**, usually delivered in hospital.
- A large proportion of patients with CF succumb to respiratory failure brought on by **chronic bacterial infection!**

Interplay between Genetic and Microbiological Data

Lyczak et. al, "Lung Infections Associated with Cystic Fibrosis," *Clinical Microbiology Reviews*, 2002

# List of Prevalent Lung Infections

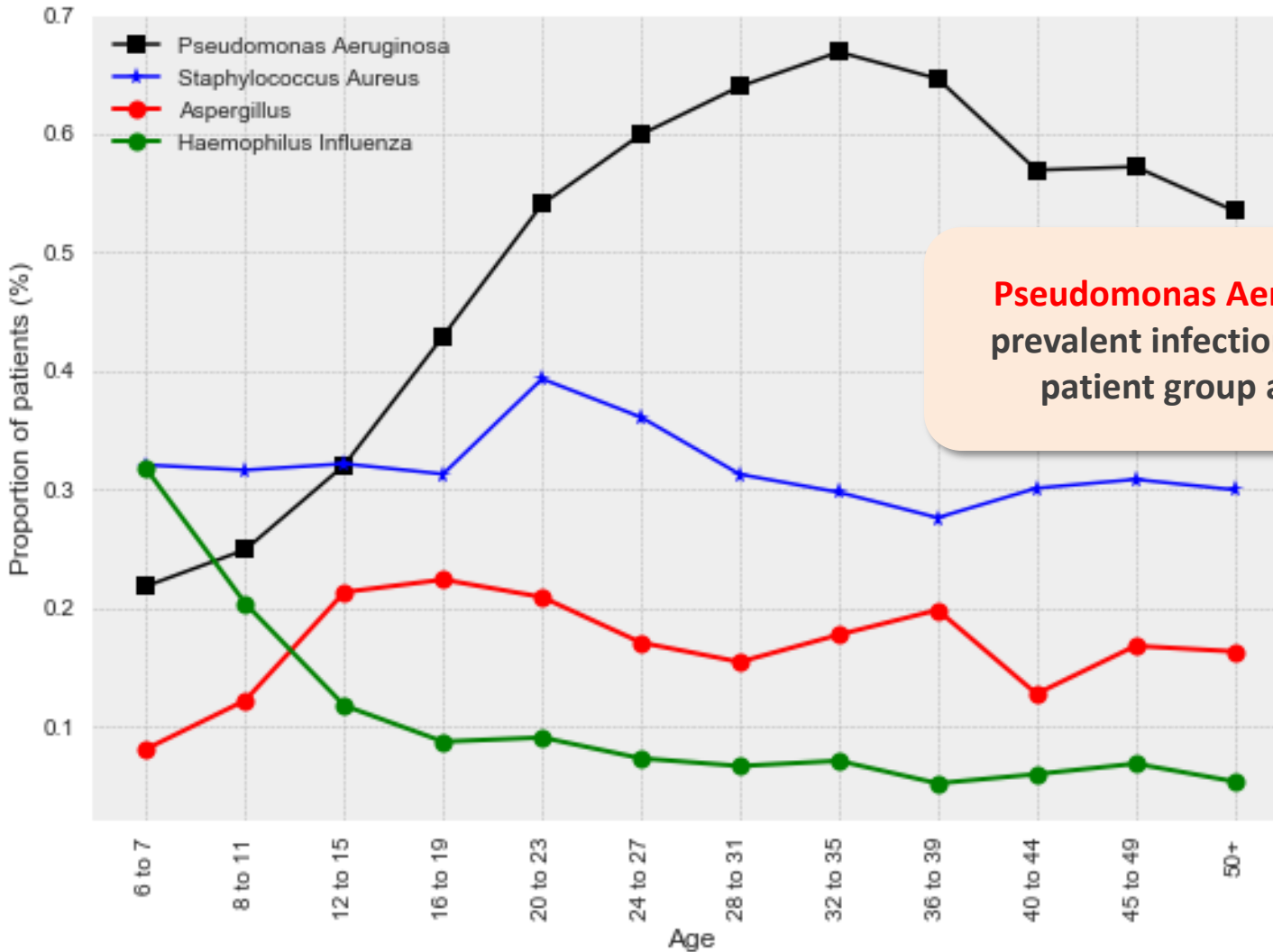
| Lung Infection  |   | Lung Infection                           |   |
|---|---|--|---|
| <b>Burkholderia Cepacia</b>                               |    | <b>Aspergillus</b>                       |    |
| <b>Pseudomonas Aeruginosa</b>                             |    | <b>Gram-Negative</b>                     |    |
| <b>Xanthomonas</b>  |    | <b>E.coli</b>                            |    |
| <b>Staphylococcus Aureus</b>                              |    | <b>Klebsiella Pneumoniae</b>             |    |
| <b>Methicillin-Resistant Staphylococcus Aureus (MRSA)</b> |  | <b>Nontuberculous Mycobacteria (NTM)</b> |   |
| <b>Haemophilus Influenza</b>                              |  | <b>Burkholderia Multivorans</b>          |  |

# Prevalence of Lung Infections

- Proportions of patients with different lung infections in **2015**.

| Lung Infection  |        | Lung Infection                           |               |
|---|--------|--|---------------|
| <b>Burkholderia Cepacia</b>                               | 3.56%  | <b>Aspergillus</b>                       | 14.99%        |
| <b>Pseudomonas Aeruginosa</b>                             | 44.05% | <b>Gram-Negative</b>                     | 1.47%         |
| <b>Xanthomonas</b>  | 5.94%  | <b>E.coli</b>                            | 2.11%         |
| <b>Staphylococcus Aureus</b>                              | 30.43% | <b>Klebsiella Pneumoniae</b>             | 1.88%         |
| <b>Methicillin-Resistant Staphylococcus Aureus (MRSA)</b> | 2.57%  | <b>Nontuberculous Mycobacteria (NTM)</b> | Insignificant |
| <b>Haemophilus Influenza</b>                              | 13.49% | <b>Burkholderia Multivorans</b>          | 1.84%         |

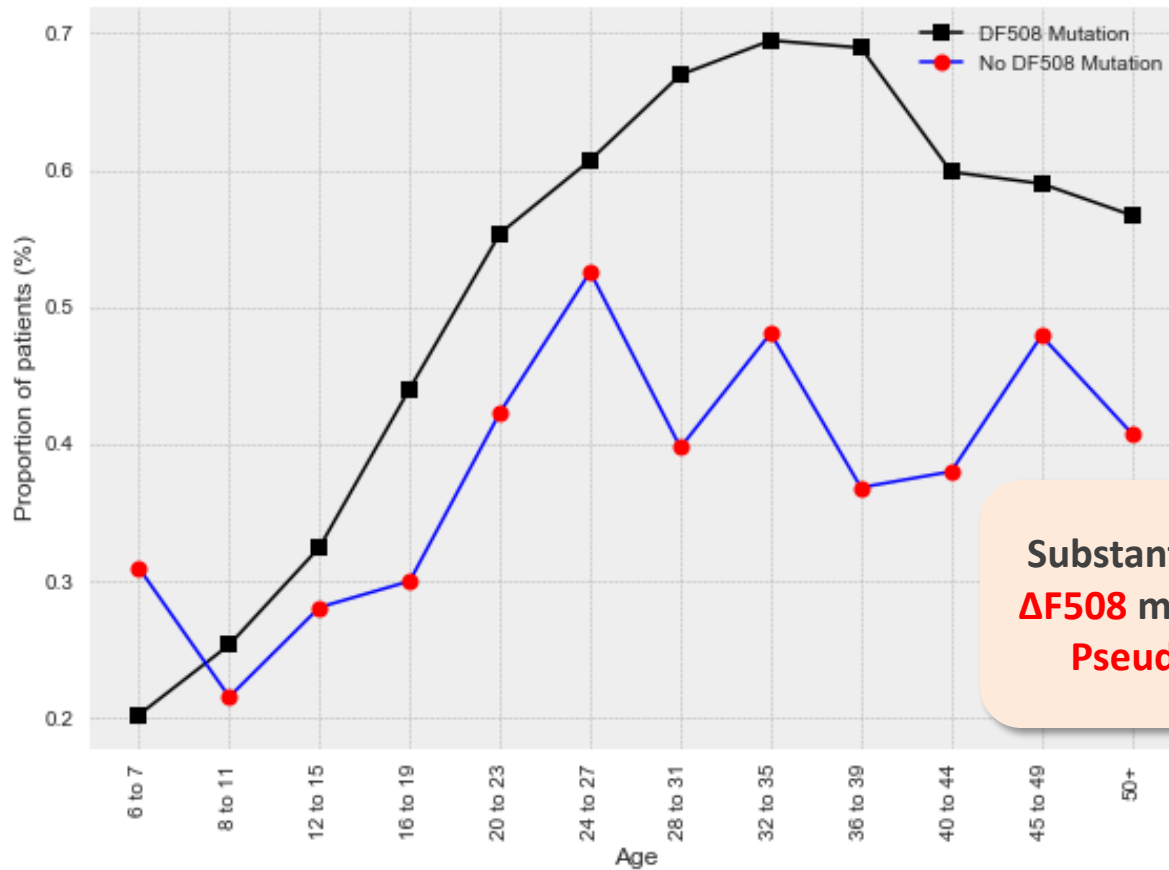
# Lung Infections Over Time



**Pseudomonas Aeruginosa** is the most prevalent infection and it peaks in the patient group aging 32-35 years

# Lung Infections and Genetic Mutations (I)

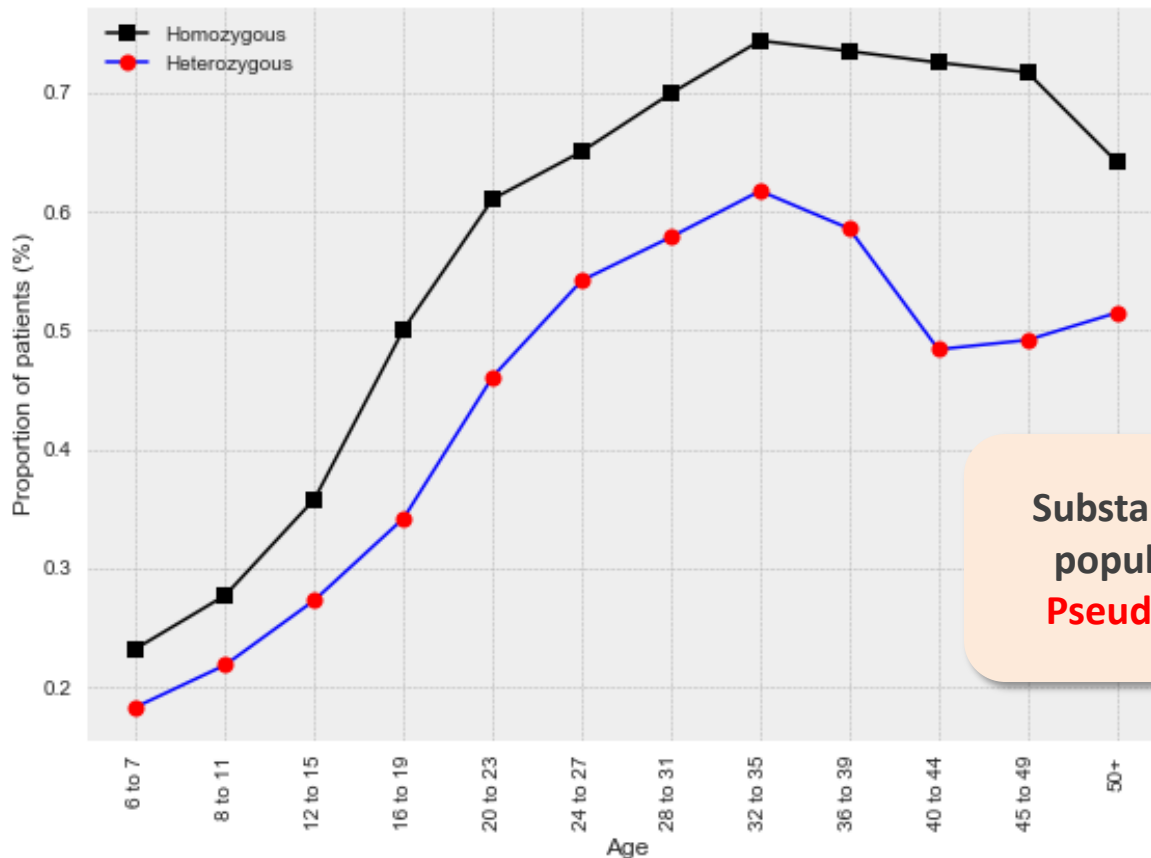
- Proportion of CF patients with **Pseudomonas Aeruginosa** stratified by the existence of a  **$\Delta F508$**  mutation.



Substantial evidence that patients with a  **$\Delta F508$**  mutation are more susceptible to a **Pseudomonas Aeruginosa** infection.

# Lung Infections and Genetic Mutations (II)

- Proportion of CF patients with **Pseudomonas Aeruginosa** in **homozygous** and **heterozygous** populations.

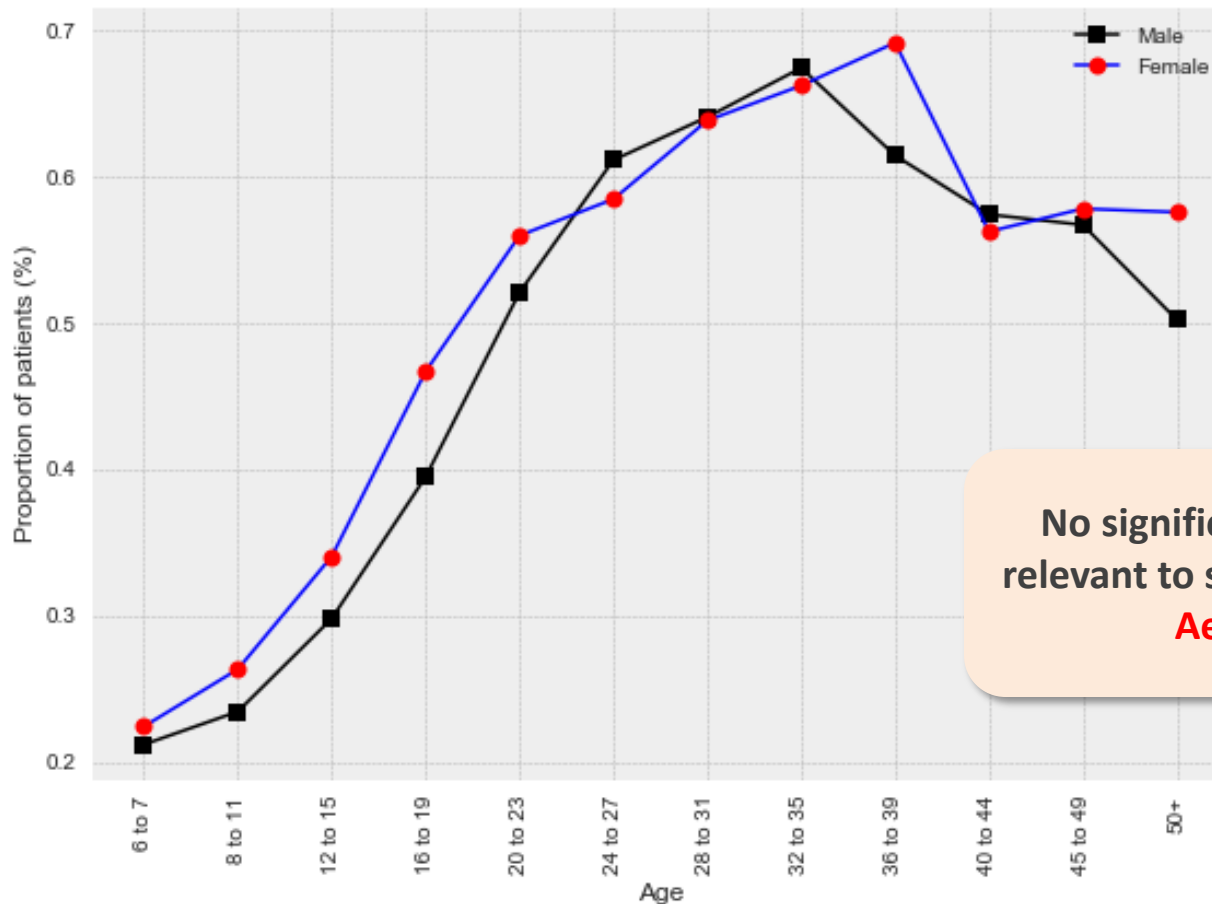


Substantial evidence that **homozygous** populations are more susceptible to **Pseudomonas Aeruginosa** infections.



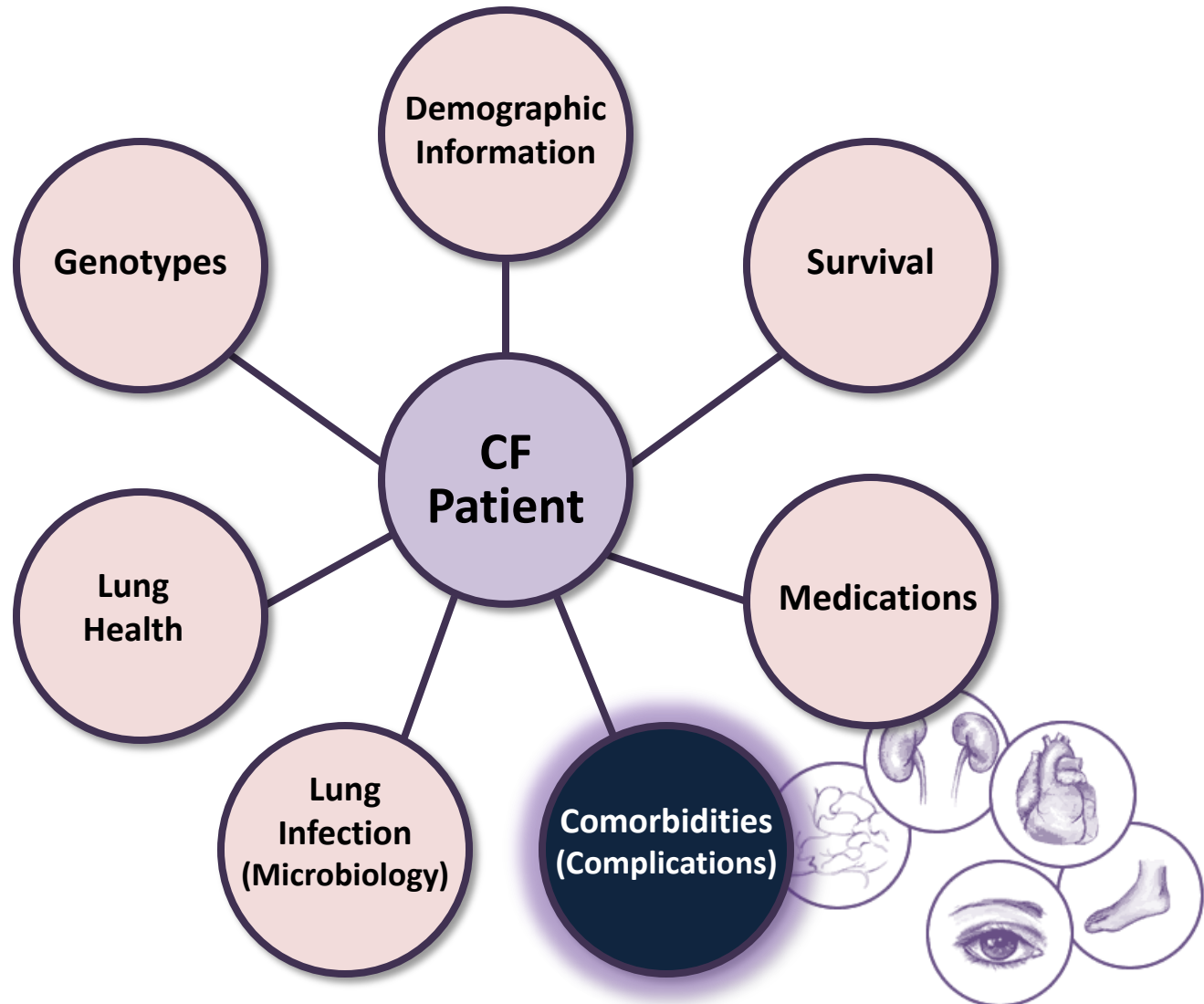
# Lung Infections and Genetic Mutations (III)

- Proportion of CF patients with **Pseudomonas Aeruginosa** in stratified by **gender**.



No significant evidence that **gender** is relevant to susceptibility to **Pseudomonas Aeruginosa** infections.

# Data Analysis: Comorbidities

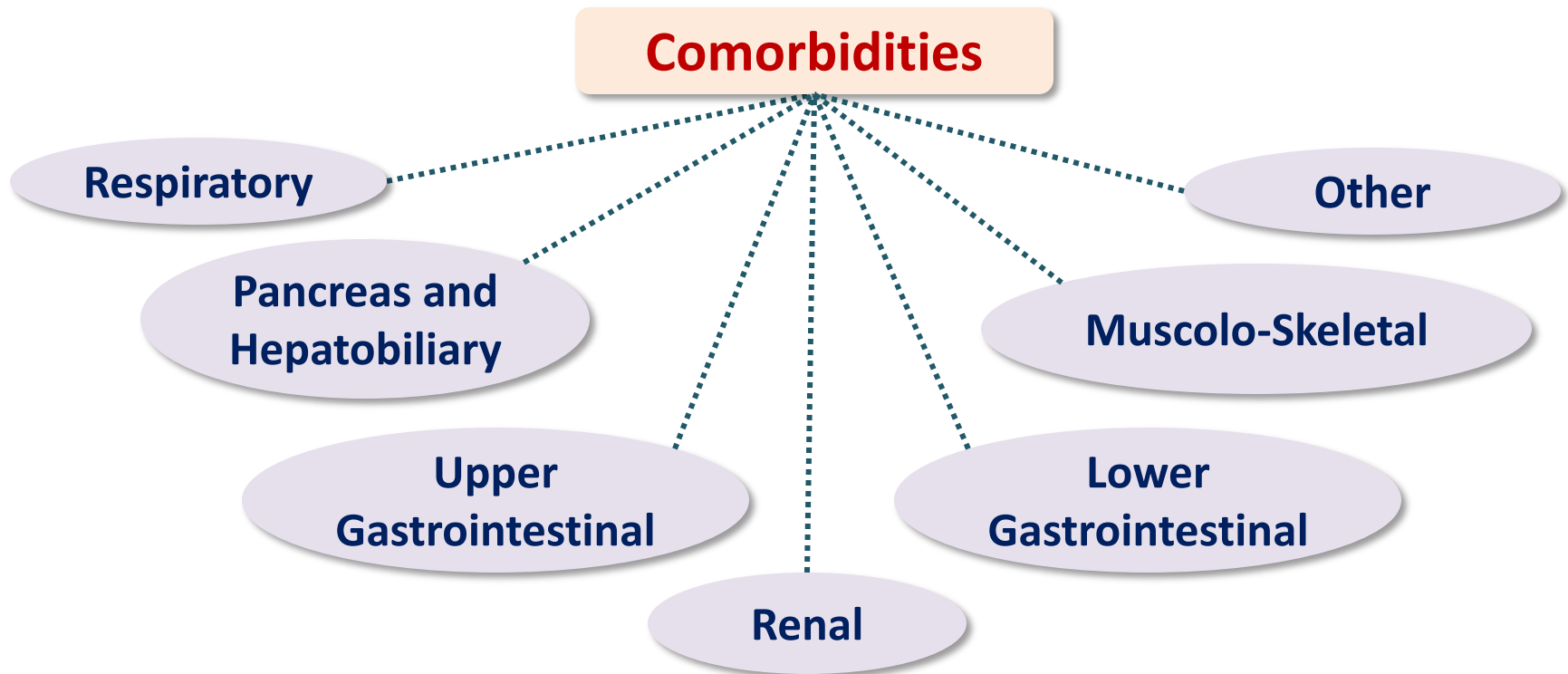


# Data Analysis: Comorbidities

- Improvement in the age profile of CF patients in the last two decades: more CF patients are now **adults**!
- **Comorbidities are more likely in adult CF patients:** pulmonary disease, CF-related diabetes, renal disease, metabolic bone disease, cancers, etc.
- **CF-related diabetes (CFRD)** is common in adults because CF affects the pancreatic sufficiency.



# Broad Categories of Comorbidities Prevalent in CF Patients



# Incidences of Comorbidities (2015)

## Comorbidities

Respiratory

Pancreas and  
Hepatobiliary

Upper  
Gastrointestinal

Renal

Nasal Polyps Requiring Surgery

2.3%

Sinus disease

9.8%

Asthma

14.4%

ABPA

10.9%

Haemoptysis

7.9%

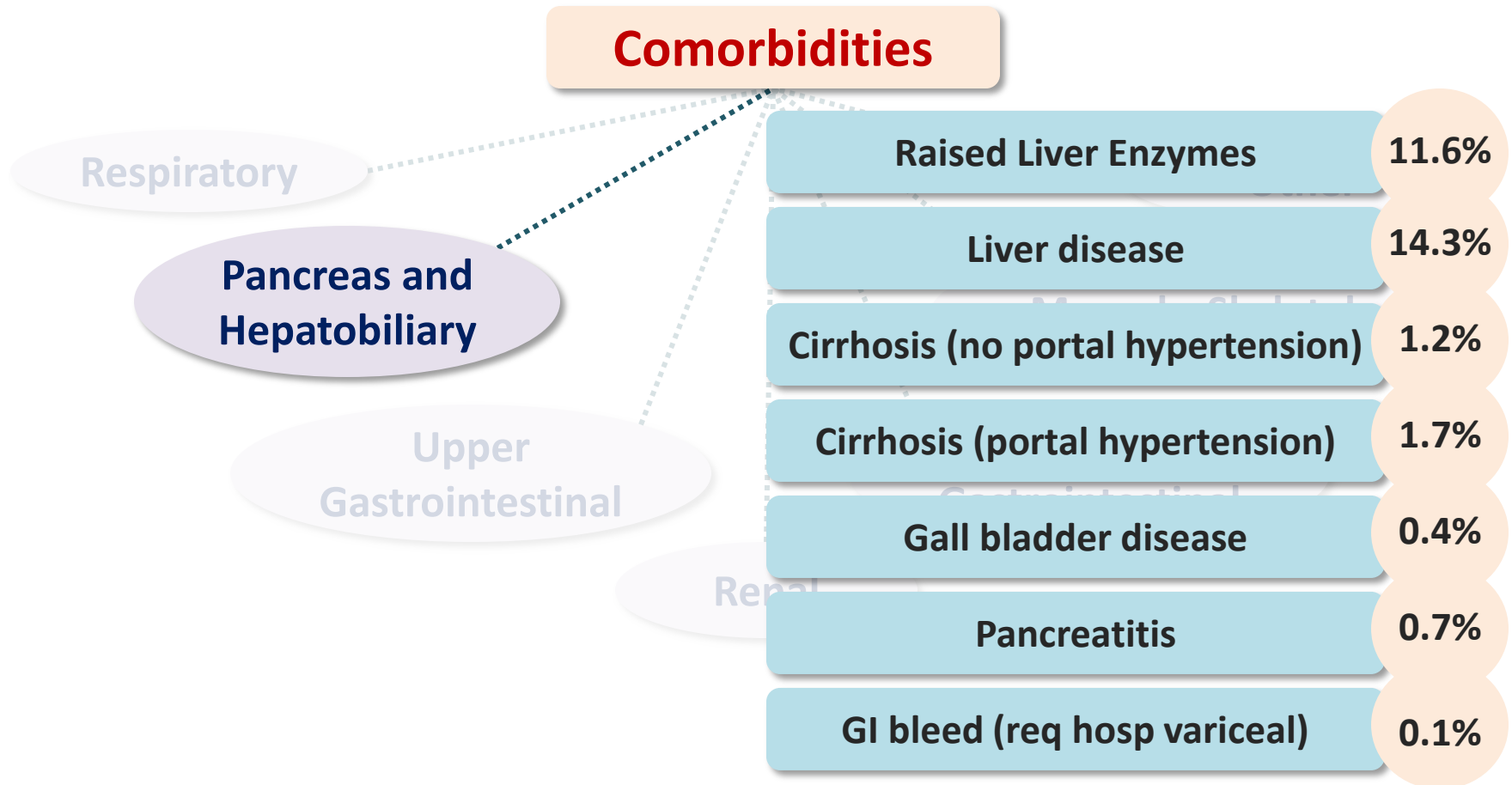
Pneumothorax

0.6%

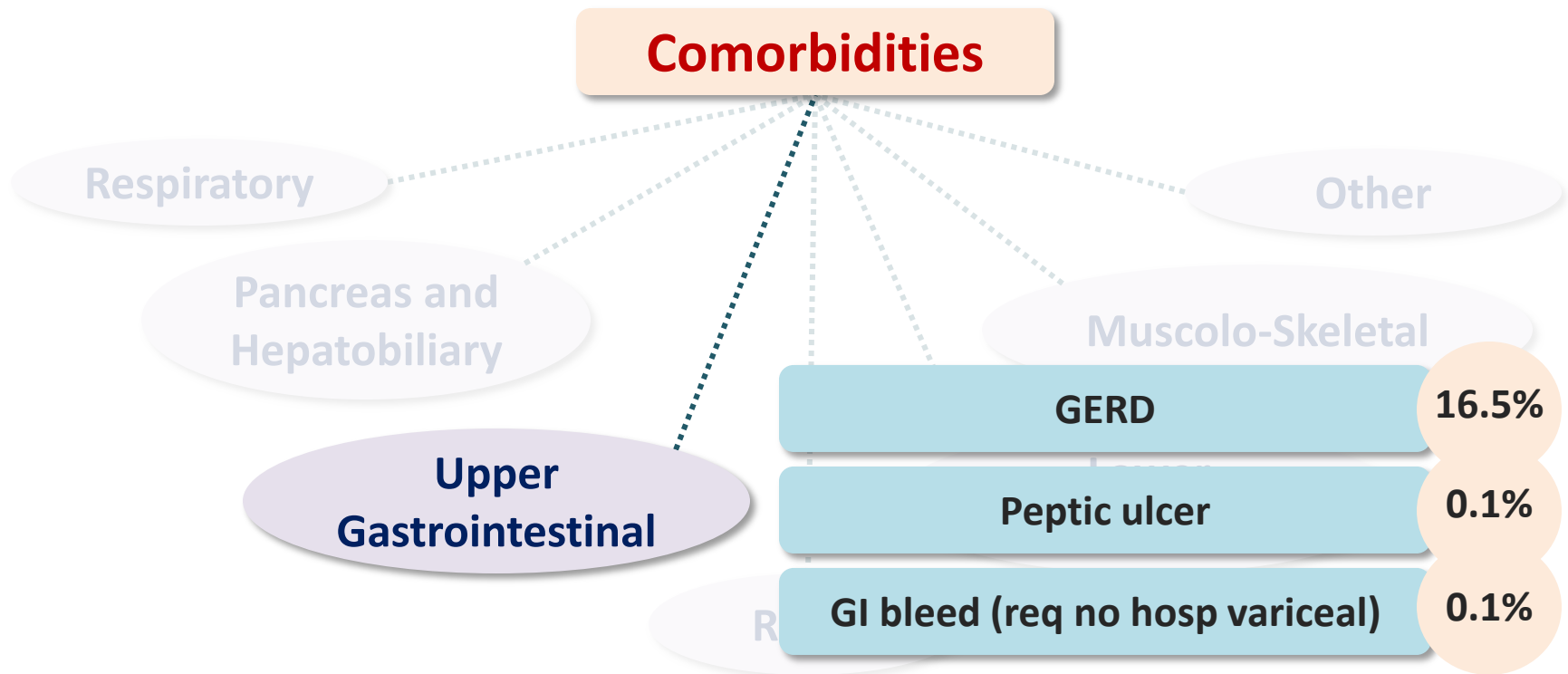
Nontuberculous mycobacteria

5.6%

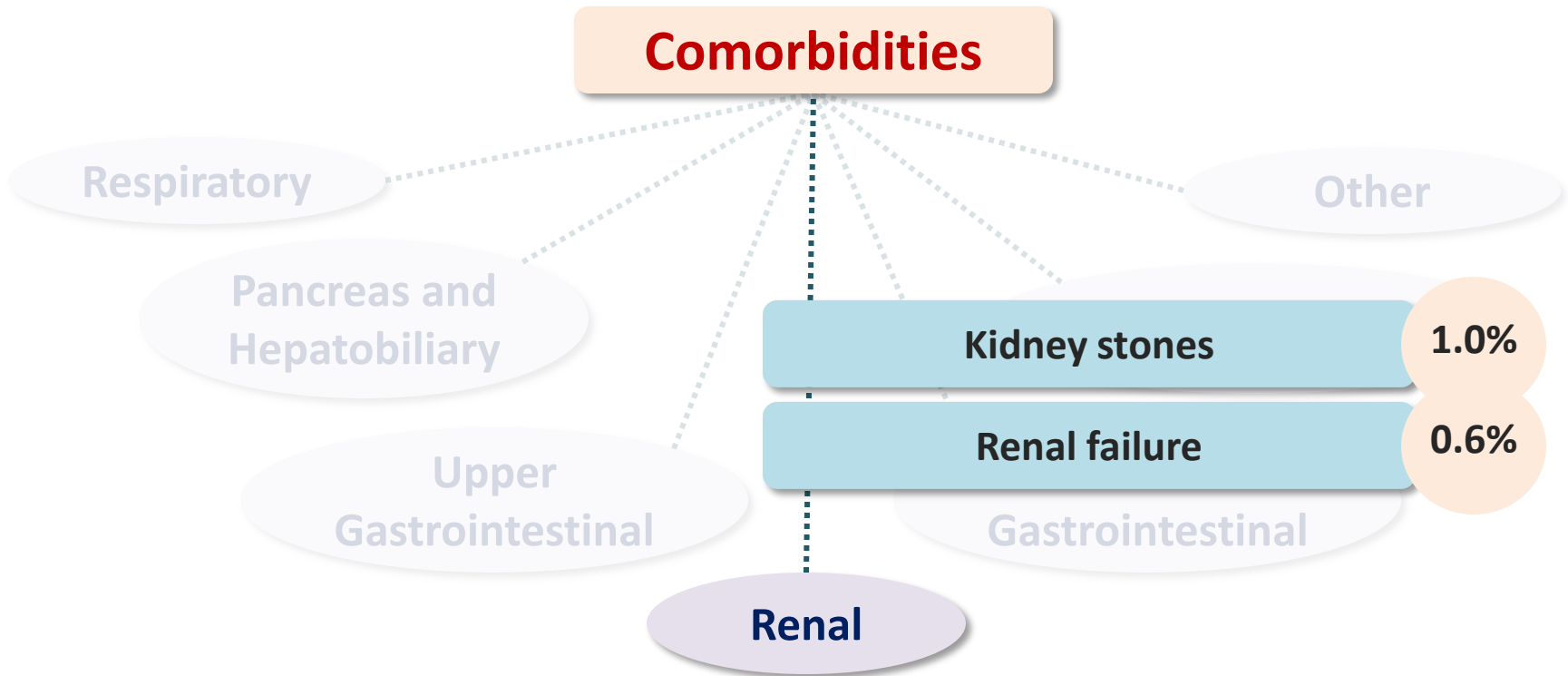
# Incidences of Comorbidities (2015)



# Incidences of Comorbidities (2015)

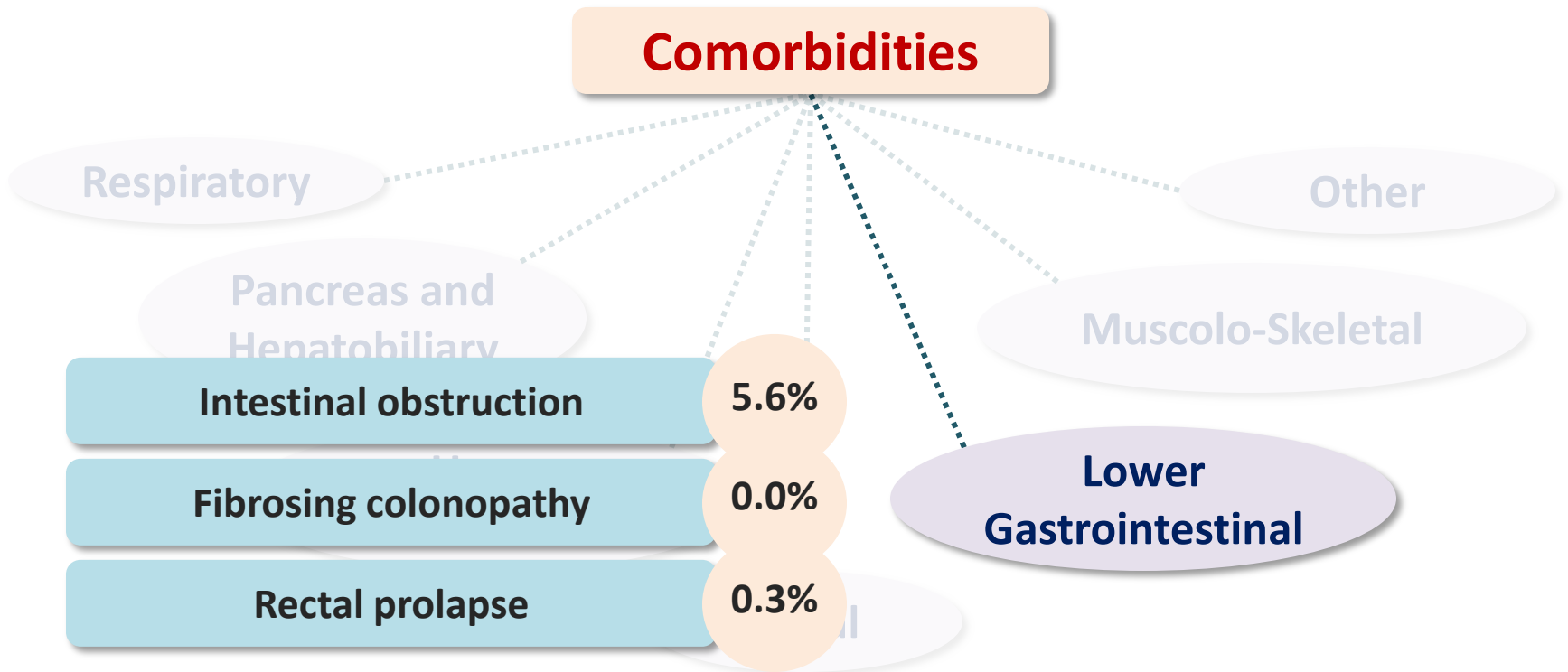


# Incidences of Comorbidities (2015)

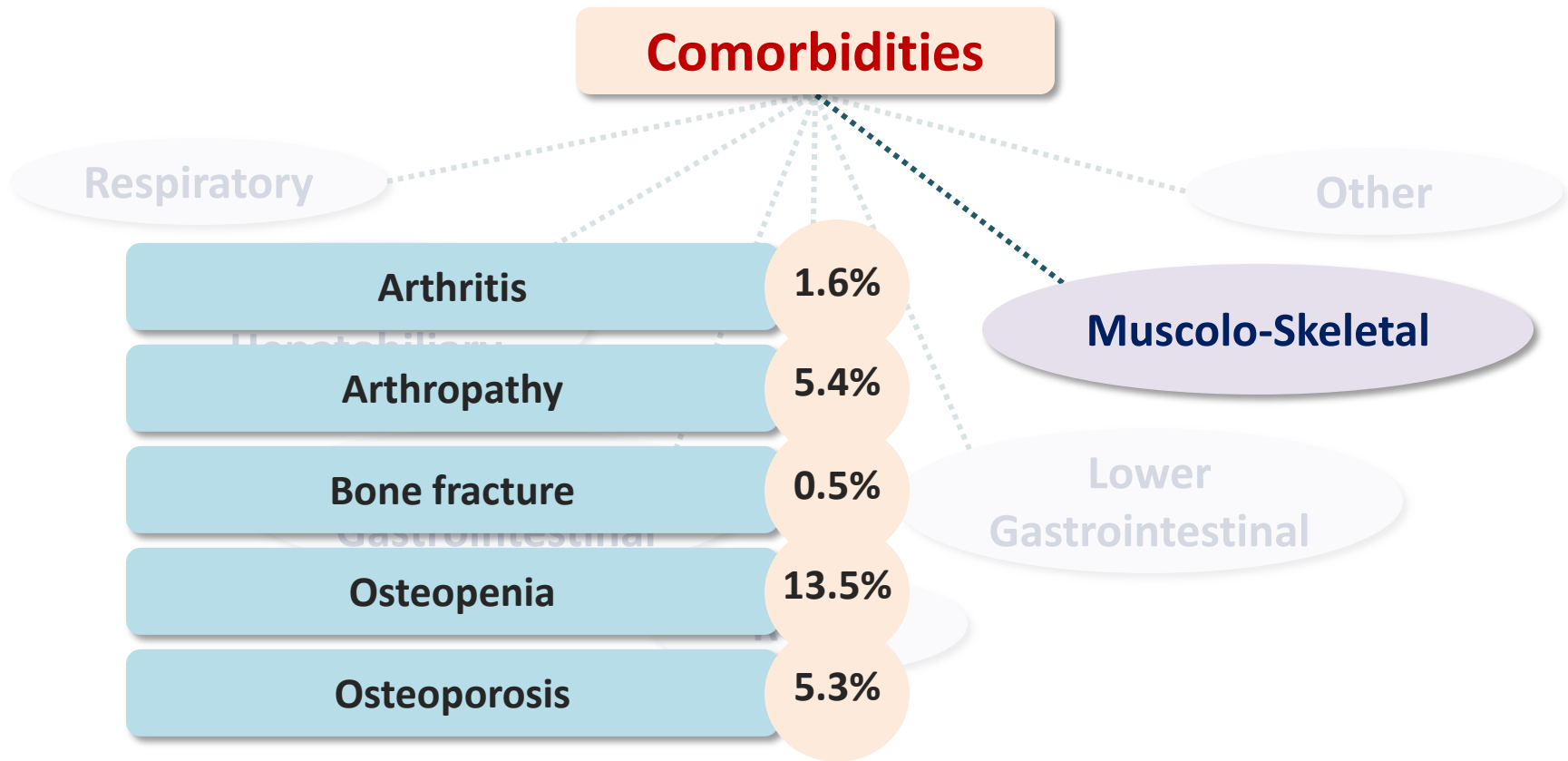




# Incidences of Comorbidities (2015)

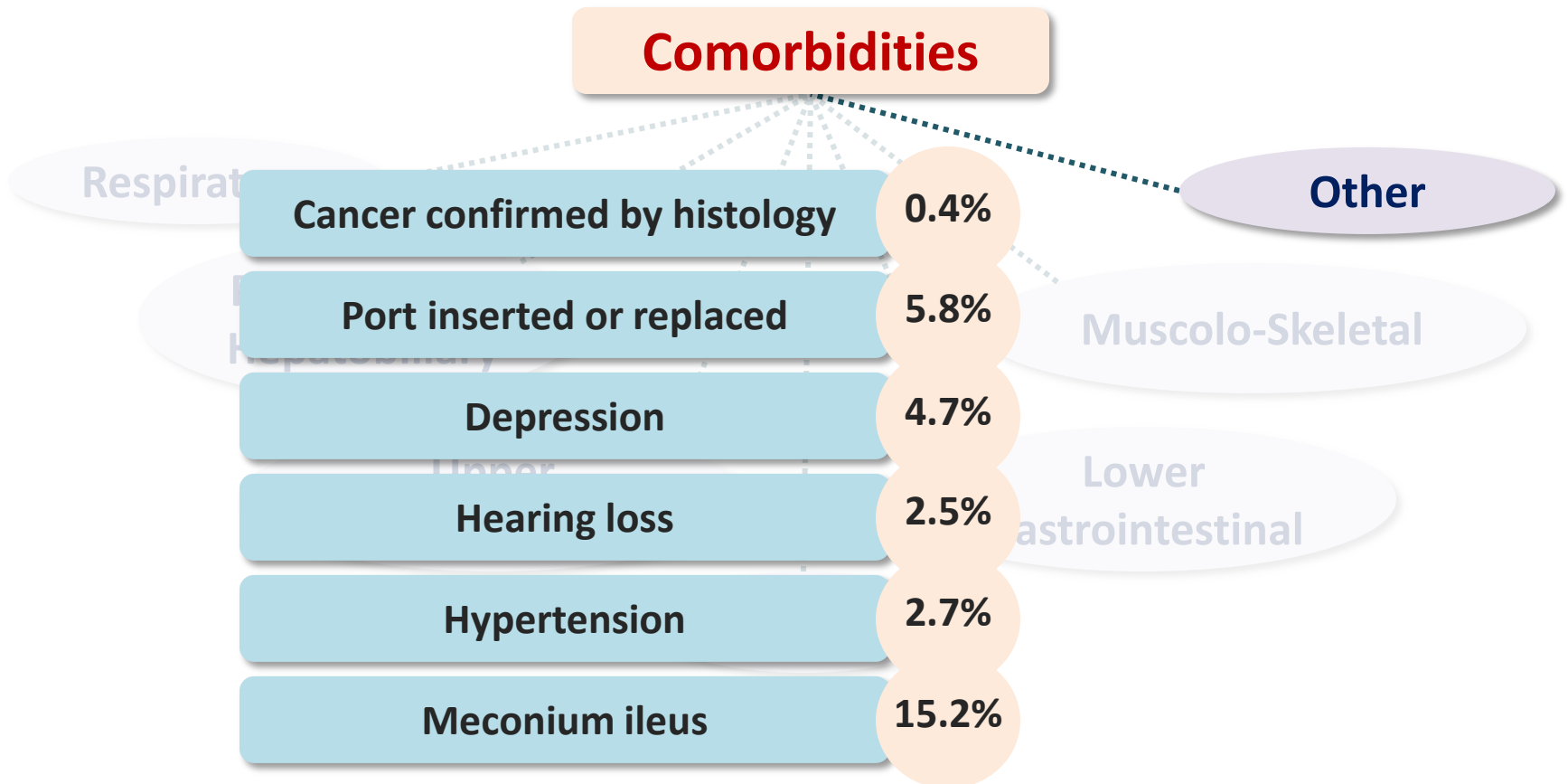


# Incidences of Comorbidities (2015)



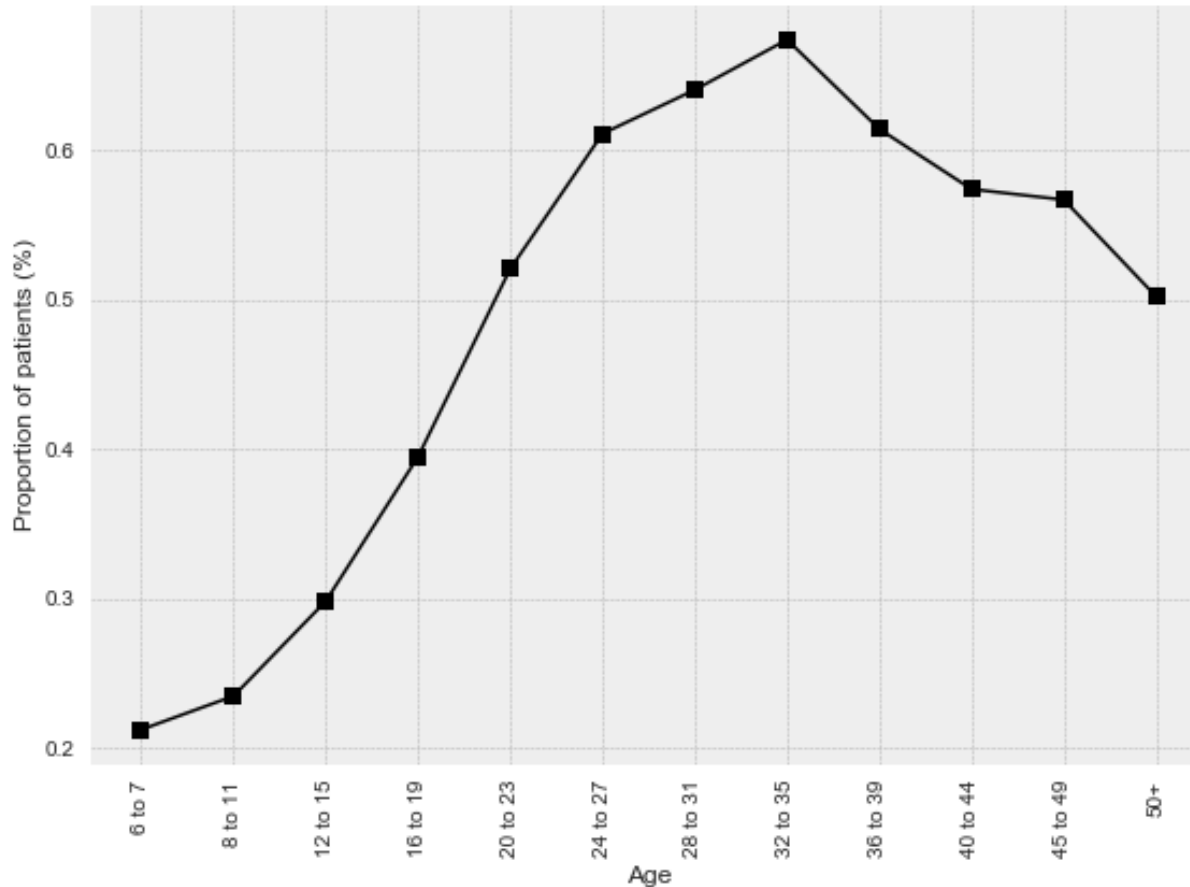
# Incidences of Comorbidities (2015)

## Comorbidities



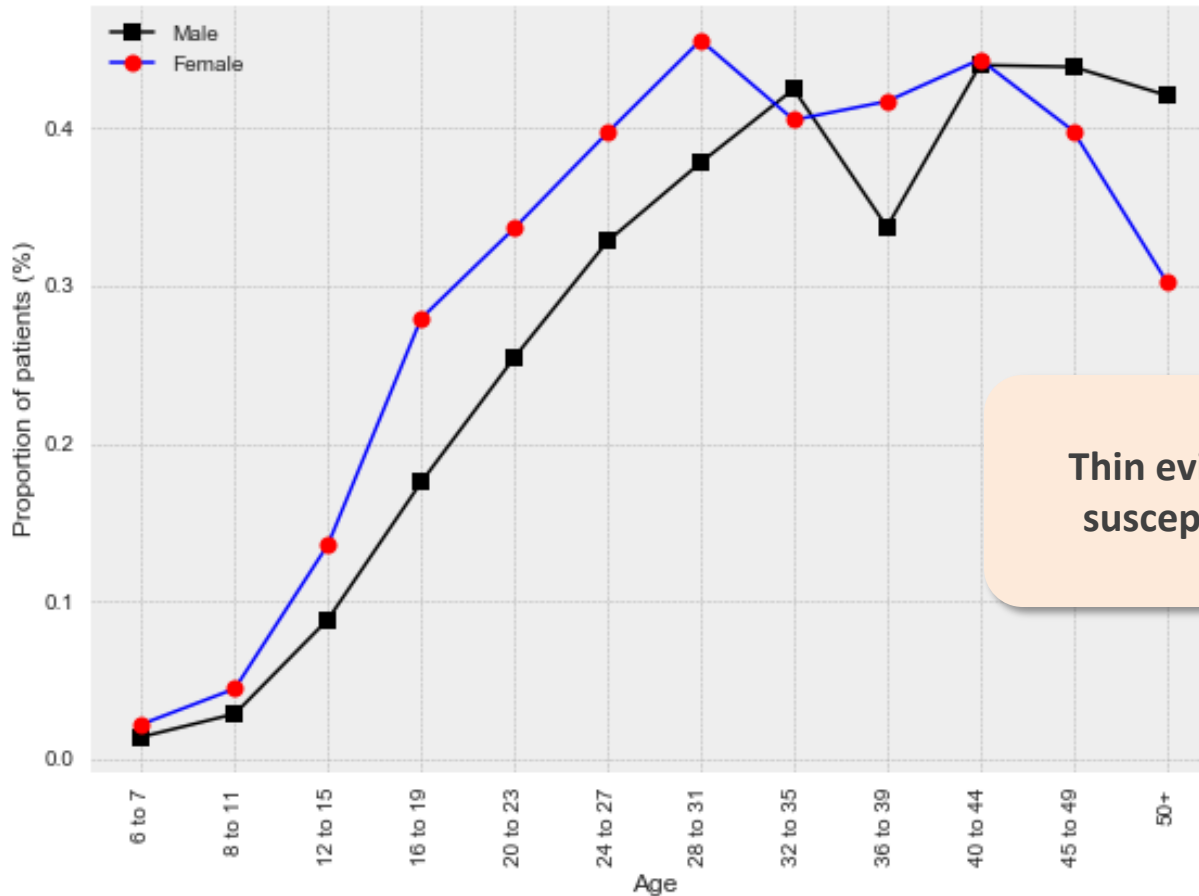
# Prevalence of CF-related Diabetes

- Incidences of **CFRD** peak in the patient group aged **32-35 years**.



# Prevalence of CF-related Diabetes

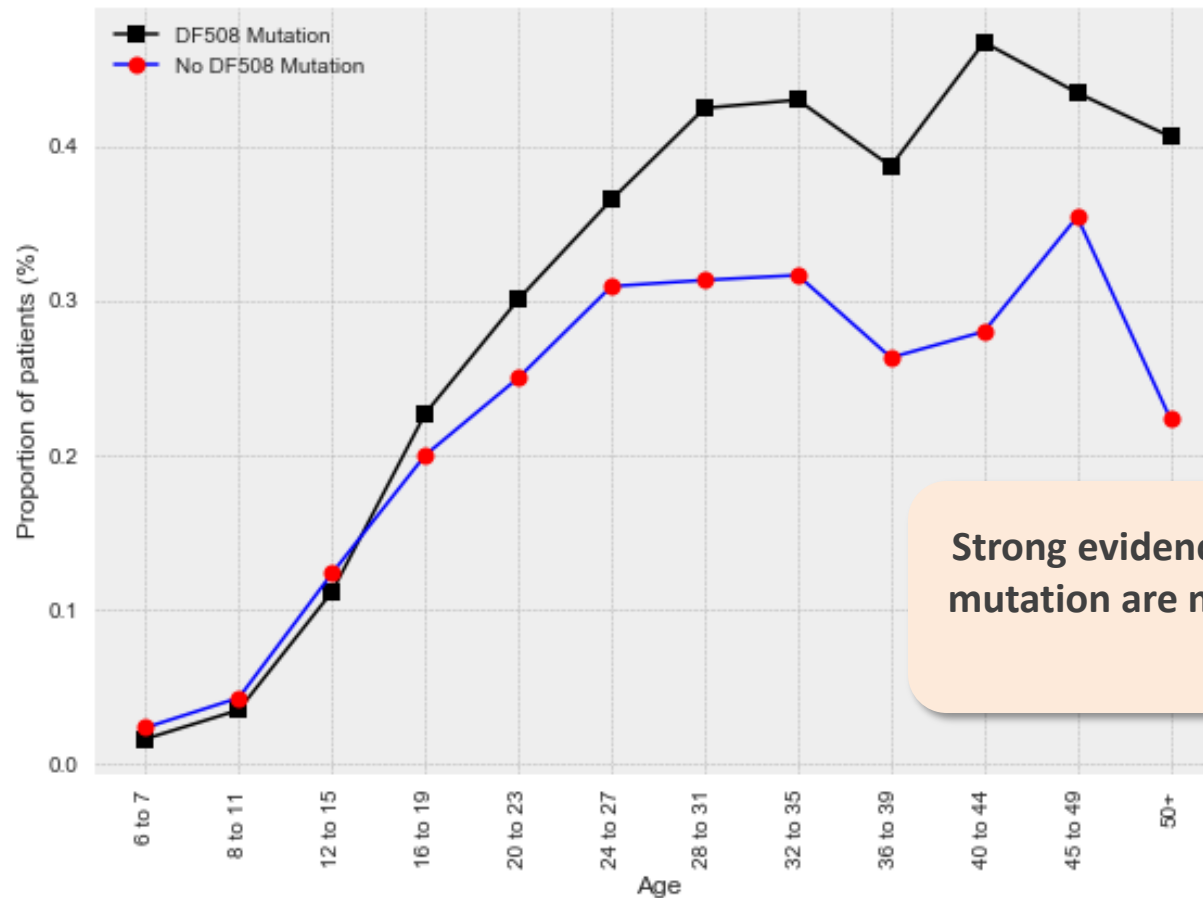
- Incidences of **CFRD** over time stratified by **gender**.



Thin evidence that **females** are more susceptible to **CFRD** at earlier ages.

# Prevalence of CF-related Diabetes

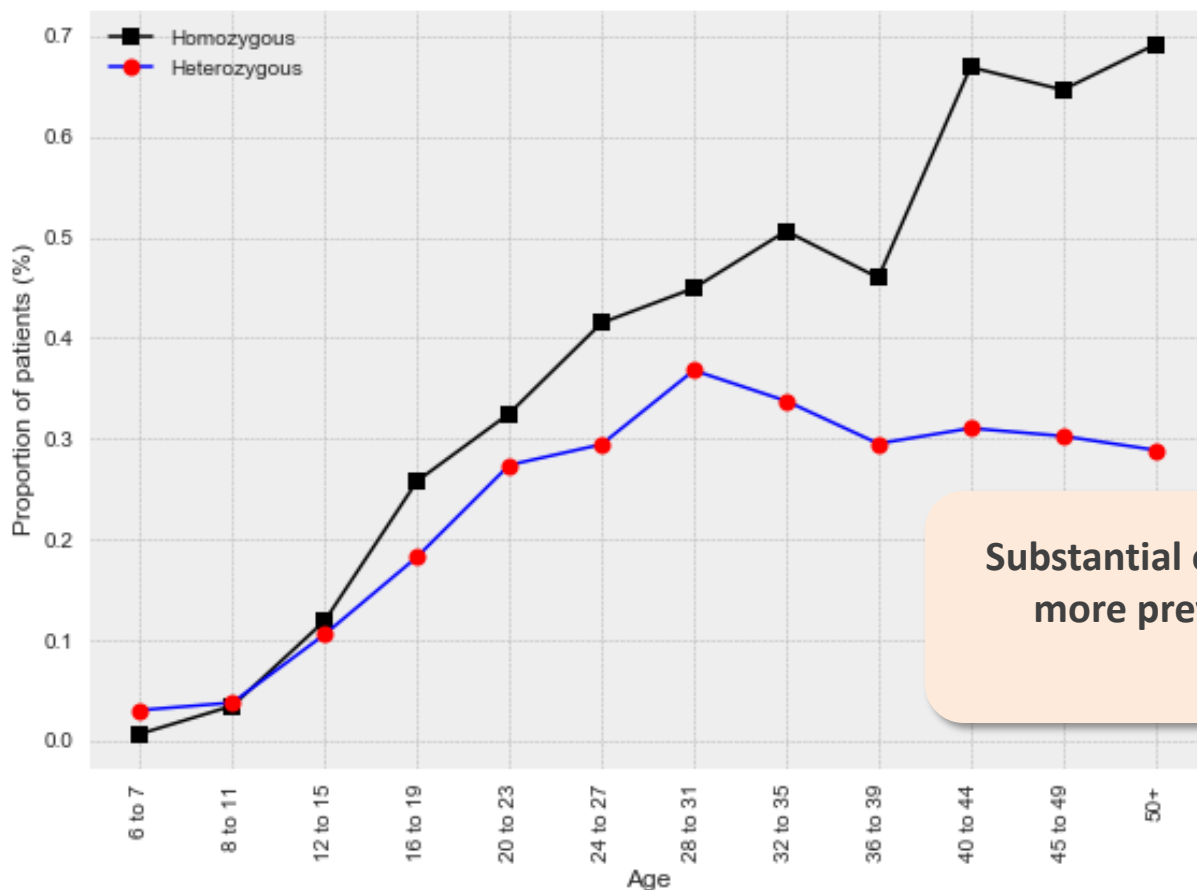
- Incidences of **CFRD** over time stratified by the existence of a  **$\Delta F508$**  mutation.



Strong evidence that patients with  $\Delta F508$  mutation are more likely to develop **CFRD** later in life.

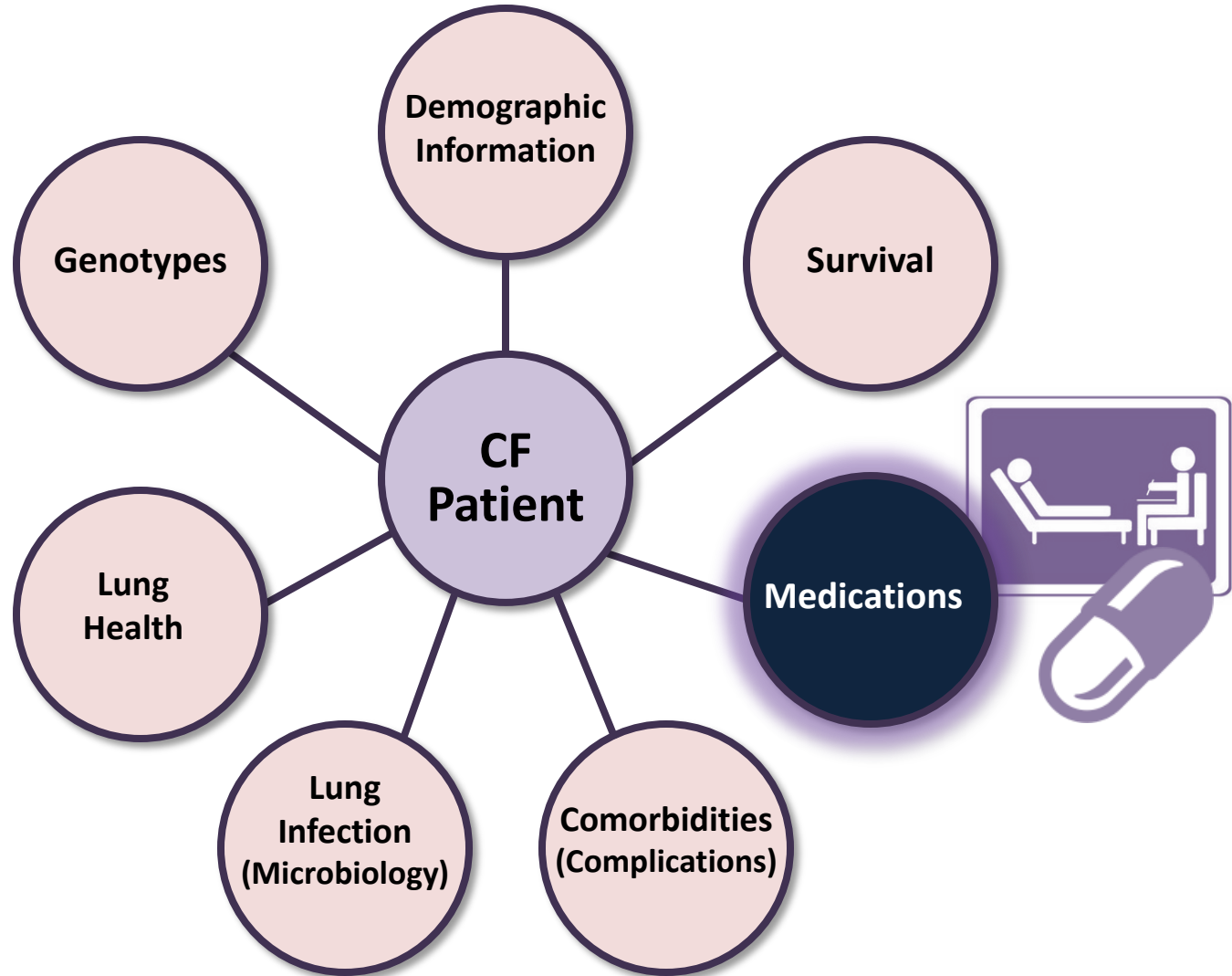
# Prevalence of CF-related Diabetes

- Incidences of **CFRD** over time in homozygous and heterozygous populations.



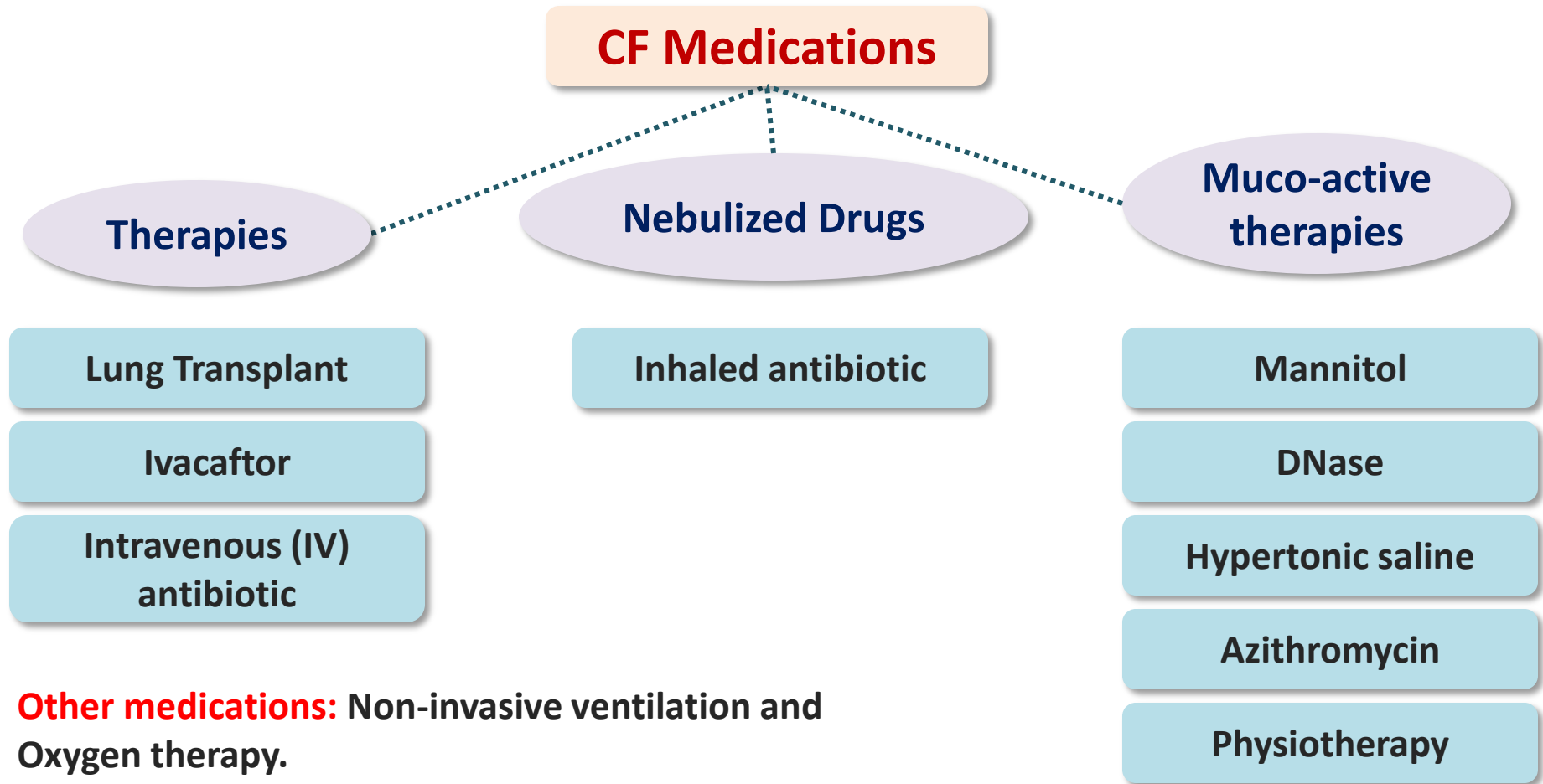
Substantial evidence that **CFRD** is much more prevalent in **homozygous** CF patients.

# Data Analysis: Medications





# Data Analysis: Medications



# Therapies

## Therapies

### Lung Transplant

**Double lung transplant**  
(Bilateral Sequential Lung Transplant)

Becomes necessary when lung function (FEV1 % predicted) declines

**Number of transplants in 2015 = 48**

### Ivacaftor

**Prescribed for patients with mutations: G551D, G187R, S549N, S549R, G551S, G1244E, S1251N, S1255P, or G139D.**

**Expensive**

**Number of patients on Ivacaftor in the UK = 439**

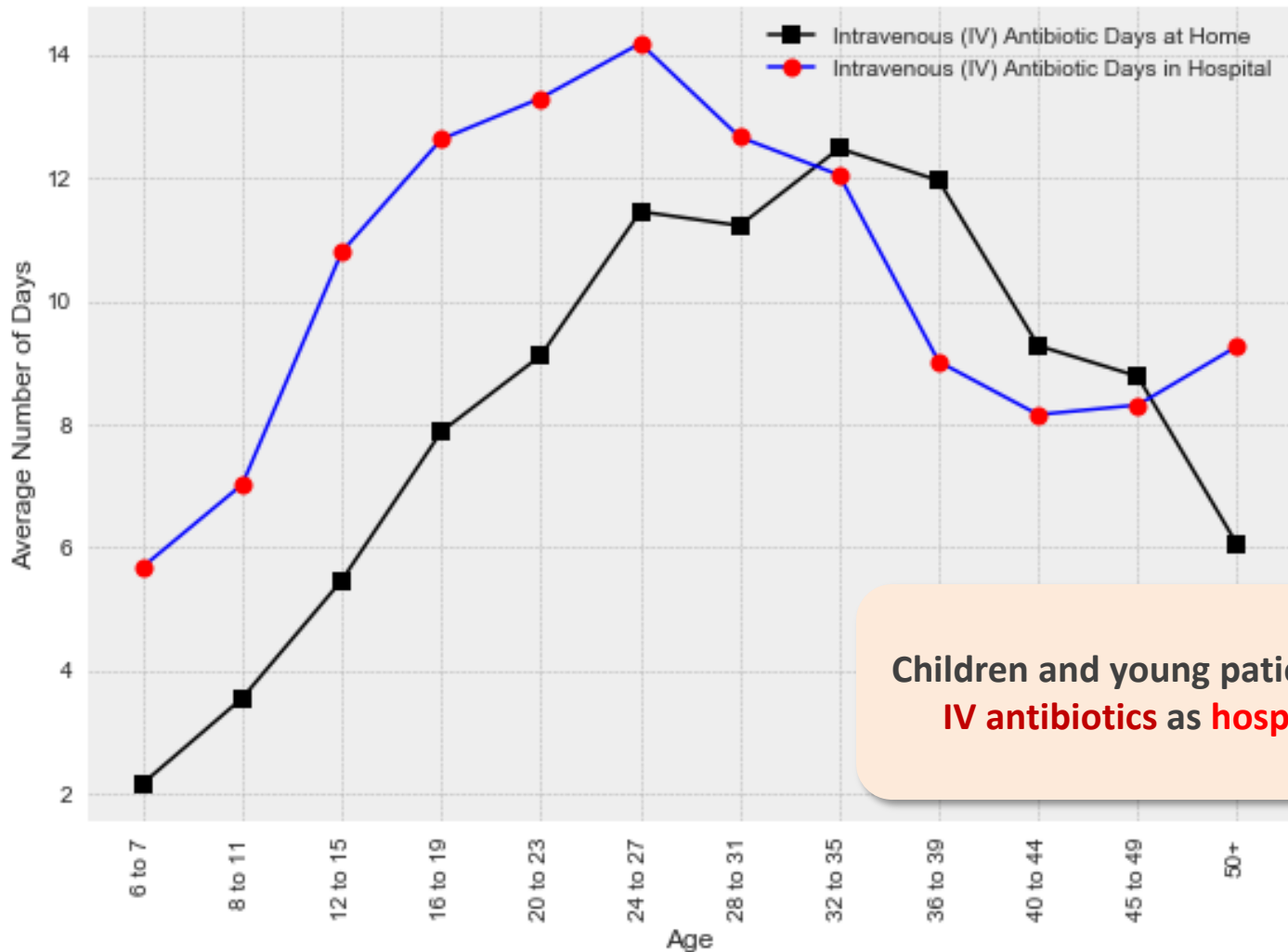
### Intravenous (IV) antibiotic

**Prescribed for patients with infections**

Given to the patient through their **veins**

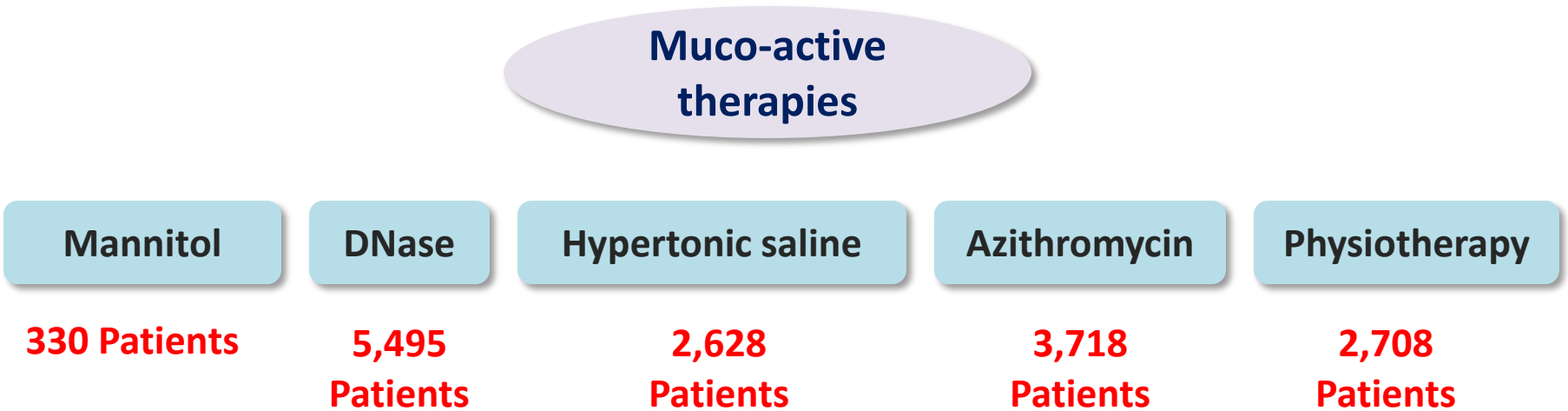
Treatment can take a number of days and might take place as a hospital inpatient, or at home.

# Intravenous (IV) antibiotic Hospitalization Time



Children and young patients often receive **IV antibiotics** as **hospital inpatients**.

# Muco-active therapies



# List of Data-induced Hypotheses

- Our data analysis led to the following hypotheses regarding the interaction between CF genetic, microbiological and phenotypic variables:

**Males** survive longer than **females**.

Patients with a **ΔF508** mutation are more susceptible to a **Pseudomonas Aeruginosa** infection.

Patients with **ΔF508** mutation are more likely to develop **CFRD** later in life.

**Homozygous** patients are more susceptible to **Pseudomonas Aeruginosa** infections.

**CFRD** is more prevalent in **homozygous** CF patients.

Our ultimate goal is to use machine learning to model the entire patient's trajectory and automatically capture all the manifestations above!

# Section C: Research Agenda

Overview of Previous Works

Research Objectives

Research Plan



# Prognostic Models Developed in Previous Works

| Study  | Objective   |
|--|---|
| Szczesniak et. al,<br><i>Am J Respir. Crit. Care Med</i> , <b>2017</b> | Phenotyping of rapid pulmonary decline              |
| Nkam et. al, <i>J. Cystic Fibrosis</i> , <b>2017</b>                   | Prognostic score of 3-year death or lung transplant |
| McCarthy et. al, <i>CHEST</i> , <b>2013</b>                            | Prognostic score of CF outcomes                     |
| George et. al, <i>BMJ</i> , <b>2011</b>                                | Evaluating survival of CF patients                  |
| Liou et. al, <i>J. Cystic Fibrosis</i> , <b>2010</b>                   | Characterizing FEV1 trajectories                    |
| Liou et. al,<br><i>Am J Respir. Crit. Care Med</i> , <b>2005</b>       | Impact of lung transplant on CF patient survival    |

# Data used in Previous Works

| Study  | Data Source  | Sample Size |
|--|--|-------------|
| Szczesniak et. al,<br><i>Am J Respir. Crit. Care Med</i> , <b>2017</b> | US Registry (CFFPR)                                    | 18,387      |
| Nkam et. al, <i>J. Cystic Fibrosis</i> , <b>2017</b>                   | French CF Registry                                     | 8,000       |
| McCarthy et. al, <i>CHEST</i> , <b>2013</b>                            | Irish CF Registry                                      | 370         |
| George et. al, <i>BMJ</i> , <b>2011</b>                                | Royal Brompton Hospital                                | 276         |
| Liou et. al, <i>J. Cystic Fibrosis</i> , <b>2010</b>                   | ESCF (encounter-based longitudinal multi-center study) | 20,644      |
| Liou et. al,<br><i>Am J Respir. Crit. Care Med</i> , <b>2005</b>       | US Registry (CFFPR)                                    | 33,415      |



# Covariates and Risk Factors used in Previous Studies

| Study   | Covariates   |
|---|--|
| Szczesniak et. al, <i>Am J Respir. Crit. Care Med</i> , <b>2017</b> | Gender, $\Delta$ F508 copies, age, age at diagnosis, FEV1(% predicted), BMI, pancreatic enzyme use, Infections (MRSA, Pa, B. cepacia, ABPA, NTM, Stenotrophomonas), CFRD, Lower SES.   |
| Nkam et. al, <i>J. Cystic Fibrosis</i> , <b>2017</b>                | Gender, CFTR genotype, Airway colonization, Comorbidities, FEV1(% predicted), FVC(% predicted), Age, Weight, Height, BMI, IV antibiotics usage, Days of hospitalization, Non-invasive ventilation, Azithromycin, Oxygen therapy, Oral corticosteroids, Inhaled therapies |
| McCarthy et. al, <i>CHEST</i> , <b>2013</b>                         | Age at first FEV1, Gender, $\Delta$ F508 homozygous, BMI   |
| George et. al, <i>BMJ</i> , <b>2011</b>                             | Gender, BMI, Pancreatic insufficiency, Chronic Pseudomonas aeruginosa infection, CF-related diabetes, Recombinant Human DNase, Nebulised antibiotics, Oxygen Therapy , Exacerbation  |
| Liou et. al, <i>Am J Respir. Crit. Care Med</i> , <b>2005</b>       | Age, Acute exacerbations, Arthropathy, Diabetes, FEV <sub>1</sub> , Gender, Pancreatic Insufficiency, Weight, Staphylococcus Aureus  |



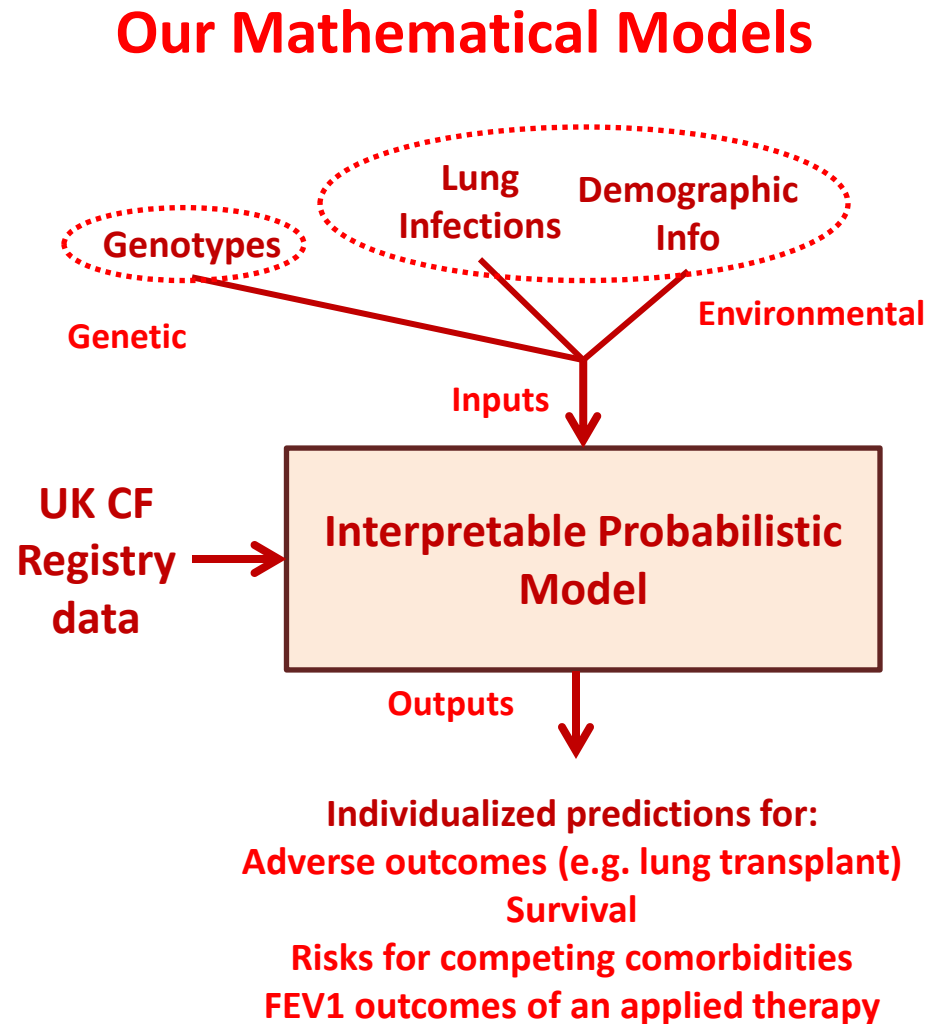
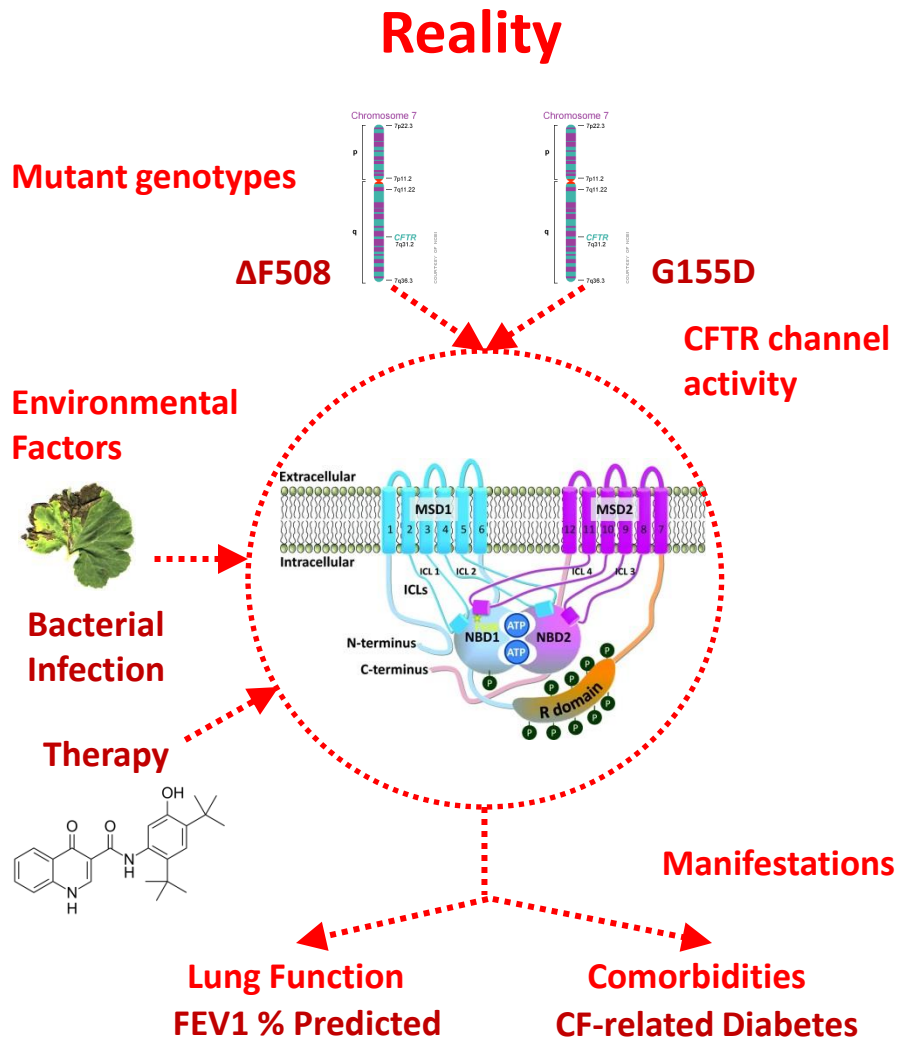
# Our Research Objectives

Our ultimate objective is to learn highly-granular, data-driven **temporal phenotypic expression models** that describe the relation between a CF patient's **individual traits** (genetic and environmental factors) and manifestations of **survival, lung function, comorbidities** and **responses to treatments**.

**Our models will provide clinicians with actionable intelligence that would help:**

- ❑ **Assess a patient's individualized risk to competing adverse outcomes, including CF-related complications and comorbidities.**
- ❑ **Understand the CF phenotypic expressions and its complex interaction with genetic and microbiological information.**
- ❑ **Construct individualized treatment plans that select the right treatment at the right time for a particular patient based on her individual traits.**

# High-level Conception of our Models



# Research Plan

Our models will provide clinicians with actionable intelligence that would help:

- ❑ Assess a patient's **individualized risk to competing adverse** outcomes, including CF-related complications and comorbidities.

**Milestone 1: Individualized Risk Scoring**

- ❑ Understand the CF **phenotypic expressions** and its complex interaction with genetic and microbiological information.

**Milestone 2: Temporal Phenotyping**

- ❑ Construct **individualized treatment plans** that select the right treatment at the right time for a particular patient based on her individual traits.

**Milestone 3: Individualized Treatment Planning**

# Milestone 1: Individualized Risk Scoring

## Limitations of the current **prognostic scores** (such as **CF-ABLE**):

- ❑ Quantifies the risk of a single adverse outcome at a single time horizon
- ❑ Coarse, one-size-fits-all prediction rule
- ❑ No principled mathematical model, fails to scale when more variables become available

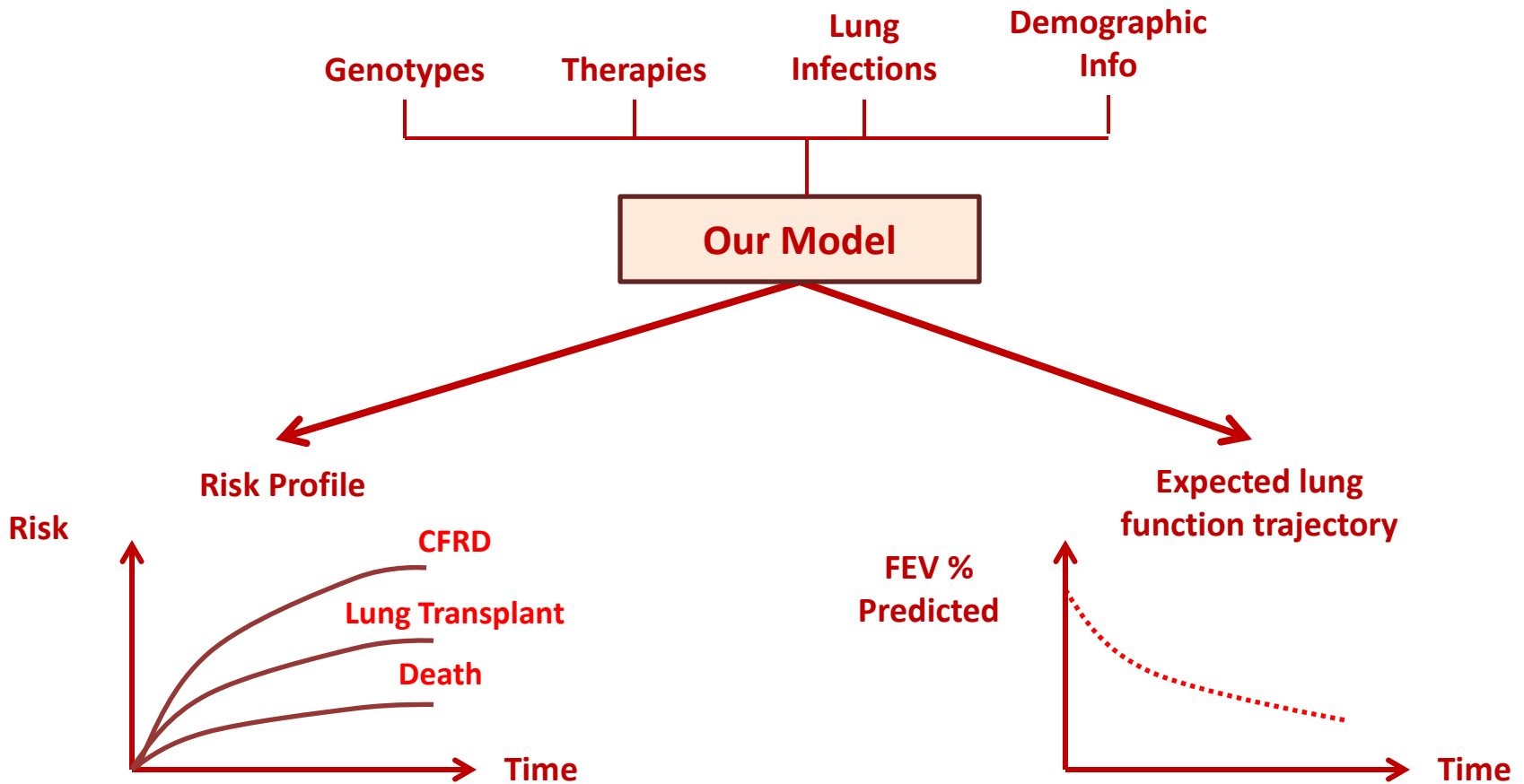
## Our data-driven model will be able to:

- ❑ Forecast a full **lung function profile** (FEV1 % predicted trajectory)
- ❑ Provide a full **risk profile** at arbitrary time horizons. A risk profile accounts for all **competing adverse events**: death, lung transplant, CF-related diabetes, respiratory, pancreatic and renal complications.
- ❑ Tailor all predictions to the patient's demographic, environmental, microbiological and genetic traits.

**Machine learning tools used:** Deep multi-task probabilistic models.

# Milestone 1: Individualized Risk Scoring

Depiction for the inputs and outputs of our model:



## Milestone 2: Temporal Phenotyping

### Limitations of the current **phenotypic expressions**:

- ❑ Limited to static manifestations (e.g. eventual manifestation of pancreatic insufficiency)
- ❑ Poor understanding of the interaction between classes of genetic mutation and microbiological infections (essential for treatment planning)

### Our data-driven **temporal phenotypic expression** will be able to:

- ❑ Describe CF manifestation as a **temporal trajectory** of lung function
- ❑ Incorporate **all comorbidities** in the CF manifestation
- ❑ Capture interactions between genetic and **microbiological factors**.

**Machine learning tools used:** Unsupervised functional clustering.



# Milestone 2: Temporal Phenotyping

Depiction for our envisioned **temporal phenotypic expressions**:

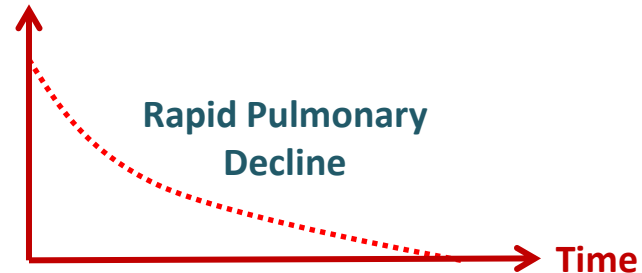
Patient Classes defined in terms of genetic mutations, demographic and microbiological factors



Phenotypic Expression



FEV % Predicted



Prevalence of CFRD

Susceptible to Pseudomonas Auregonis infections

Treatment planning will not be only targeted to the phenotype, but also to the time horizon within a phenotypic expression!

## Milestone 3: Individualized Treatment Planning

**Our data-driven model for counterfactual inference will be able to:**

- ❑ Infer the individualized benefit of **Ivacaftor** in terms of **4-6 months** improvement in **FEV % predicted**.
- ❑ Use the phenotypic expressions constructed in the previous milestone to design a **phenotype-specific treatment plans** that decides which antibiotics/therapies should each patient take at every point of time.

**Machine learning tools used:** Bayesian nonparametric models for causal inference.

# Tentative Timeline



**December 2017**

**Milestone 1: Individualized Risk Scoring**

**February 2018**

**Milestone 2: Temporal Phenotyping**

**June 2018**

**Milestone 3: Individualized Treatment Planning**

# Section D: Preliminary Results

Risk Scoring via Automated Prognostic Model Learning



# Objectives

## The goal of this section is to:

- ❑ Use our **Automated prognostic model construction algorithm** for predicting **3-year outcomes** for CF patients by applying **machine learning** to the **CF registry data**.
- ❑ Compare the predictive power of machine learning with that of the CF-ABLE score and the FEV1 biomarker.
- ❑ Demonstrate the utility of using our methods for individualized risk scoring and illustrate the nature of contributions that machine learning can offer in the CF healthcare setting.

# Risk Factors

We have included the following genetic, microbiological and therapeutic information as risk factors in our analysis. (44 risk factors)

|                  |                       |                        |                         |
|------------------|-----------------------|------------------------|-------------------------|
| Age              | Best FEV1 % Predicted | Xanthomonas            | IV Anti. Days Home      |
| Gender           | B. Cepacia            | B. Multivorans         | IV Anti. Days Hosp.     |
| Height           | P. Aeruginosa         | CF-related Diabetes    | Dornase Alpha           |
| Weight           | MRSA                  | ABPA                   | Tobi Solution           |
| BMI              | NTM                   | Depression             | Chronic Oral Antibiotic |
| Smoking          | H. Influenza          | Intestinal Obstruct.   | Hypersaline             |
| Homozygous       | E. Coli               | Cirrhosis              | Inh. Bronchodilators    |
| DF508 Mutation   | Aspergillus           | Cancer                 | Promixin                |
| FEV1             | K. Pneumoniae         | GERD                   | Oxygen Therapy          |
| FEV1 % Predicted | Gram-negative         | Liver Disease          | Non-Invasive Vent.      |
| Best FEV1        | Staphylococcus Aureus | Chronic Staphylococcus | Lab Liver Enzymes       |

# The Cohort

- ❑ We extracted a cohort of patients who were enrolled in the registry in **2012** and obtained their 3-year outcomes from the **2015** registry.
- ❑ **Adverse outcomes are defined as:** death or lung-transplant in 3 years.
- ❑ We excluded all patients who have had a lung transplant by 2012 from the study since for those the definition of the adverse outcome does not apply.

## **IMPORTANT (need to discuss with collaborators in the Trust)**

- Explicit information on individual patient deaths are not available in the registry
- We assumed that patients who disappear from the registry in 2015 are dead
- This may not be true as it could be that they were not enrolled in the registry as their information for this year was not complete

# Clinical Scores (I)

We compared the predictive power of machine learning with three prognostic approaches:

- **The CF-ABLE score:** designed to predict mortality and lung transplant endpoints using a simple rule for mapping the patient's clinical features to a risk score.

McCarthy et. al, "The CF-ABLE score: a novel clinical prediction rule for prognosis in patients with cystic fibrosis," CHEST, 2013.

## **Computation of the CF-ABLE score [0-7]:**

Score = (3.5 points if FEV1 < 52%) + (1.5 points for exacerbations) + (2 points if age < 24 years and BMI < 20.1 kg/m<sup>2</sup>)



## Clinical Scores (II)

- **The CF-ABLE-UK score:** a modification of the CF-ABLE score that replaces exacerbations with days spent on intravenous antibiotics

Dimitrov et. al, "CF-ABLE-UK score: Modification and validation of a clinical prediction rule for prognosis in cystic fibrosis on data from UK CF registry," European Respiratory Journal, 2015.

### **Computation of the CF-ABLE-UK score [0-7]:**

Score = (3.5 points if FEV1 < 52%) + (1.5 points for usage of IV antibiotics) + (2 points if age < 24 years and BMI < 20.1 kg/m<sup>2</sup>)

- **Predictions based solely on FEV1**

# Results

Our Automated Prognostic Model construction algorithm searches for a prognostic model in a space of **72** machine learning models.

|   |                      |
|---|----------------------|
| <b>AUCROC of CF-ABLE</b>  | <b>0.6692</b>        |
| <b>AUCROC of CF-ABLE-UK</b>   | <b>0.6590</b>        |
| <b>AUCROC of FEV1</b>   | <b>0.6711</b>        |
| <b>AUCROC of Automated Prognostic Model Construction<br/>(Fast ICA + Gradient boosting)</b> | <b>0.7766 ± 0.08</b> |

# Performance of ML Pipelines Searched by the Automated Prognostic Model Construction Algorithm (I)

## No preprocessing, 5-fold CV

| Benchmark                       | AUCROC               | Benchmark                    | AUCROC               |
|---------------------------------|----------------------|------------------------------|----------------------|
| <b>No Prep. + Logistic Reg.</b> | <b>0.7760 ± 0.10</b> | No Prep. + Linear SVM        | 0.7160 ± 0.18        |
| No Prep. + SGD Perceptron       | 0.6943 ± 0.14        | No Prep. + Random Forest     | 0.7366 ± 0.07        |
| No Prep. + kNN                  | 0.6481 ± 0.05        | No Prep. + Extra Trees       | 0.7381 ± 0.06        |
| No Prep. + Decision Tree        | 0.5865 ± 0.03        | No Prep. + AdaBoost          | 0.7567 ± 0.10        |
| No Prep. + Kernel SVM           | 0.6409 ± 0.09        | No Prep. + Bagging           | 0.7075 ± 0.08        |
| No Prep. + Gauss. Naïve Bayes   | 0.7165 ± 0.15        | No Prep. + Gradient Boosting | 0.7751 ± 0.09        |
| No Prep. + Bern. Naïve Bayes    | 0.6693 ± 0.13        | <b>No Prep. + XGBoost</b>    | <b>0.7760 ± 0.10</b> |
| No Prep. + LDA                  | 0.7689 ± 0.10        | No Prep. + MLP (2 layers)    | 0.7569 ± 0.07        |
| No Prep. + Passive Aggressive   | 0.7254 ± 0.12        | No Prep. + MLP (3 layers)    | 0.7612 ± 0.07        |

# Performance of ML Pipelines Searched by the Automated Prognostic Model Construction Algorithm (II)

## PCA (25 components), 5-fold CV

| Benchmark                | AUCROC        | Benchmark               | AUCROC               |
|--------------------------|---------------|-------------------------|----------------------|
| PCA + Logistic Reg.      | 0.7653 ± 0.10 | PCA + Linear SVM        | 0.6067 ± 0.24        |
| PCA + SGD Perceptron     | 0.6417 ± 0.25 | PCA + Random Forest     | 0.7086 ± 0.09        |
| PCA + kNN                | 0.6492 ± 0.05 | PCA + Extra Trees       | 0.7237 ± 0.07        |
| PCA + Decision Tree      | 0.5608 ± 0.04 | PCA + AdaBoost          | 0.7351 ± 0.05        |
| PCA + Kernel SVM         | 0.6307 ± 0.11 | PCA + Bagging           | 0.6938 ± 0.08        |
| PCA + Gauss. Naïve Bayes | 0.7504 ± 0.11 | PCA + Gradient Boosting | 0.7626 ± 0.08        |
| PCA + Bern. Naïve Bayes  | 0.7023 ± 0.10 | <b>PCA + XGBoost</b>    | <b>0.7691 ± 0.08</b> |
| PCA + LDA                | 0.7586 ± 0.10 | PCA + MLP (2 layers)    | 0.6904 ± 0.07        |
| PCA + Passive Aggressive | 0.4785 ± 0.39 | PCA + MLP (3 layers)    | 0.6807 ± 0.10        |

# Performance of ML Pipelines Searched by the Automated Prognostic Model Construction Algorithm (III)

## Sparse PCA (25 components), 5-fold CV

| Benchmark                 | AUCROC        | Benchmark                    | AUCROC               |
|---------------------------|---------------|------------------------------|----------------------|
| SPCA + Logistic Reg.      | 0.7259 ± 0.15 | SPCA + Linear SVM            | 0.7444 ± 0.12        |
| SPCA + SGD Perceptron     | 0.7378 ± 0.13 | SPCA + Random Forest         | 0.7292 ± 0.08        |
| SPCA + kNN                | 0.6520 ± 0.05 | SPCA + Extra Trees           | 0.7266 ± 0.08        |
| SPCA + Decision Tree      | 0.5755 ± 0.03 | SPCA + AdaBoost              | 0.7354 ± 0.05        |
| SPCA + Kernel SVM         | 0.6833 ± 0.07 | SPCA + Bagging               | 0.6985 ± 0.06        |
| SPCA + Gauss. Naïve Bayes | 0.7319 ± 0.15 | SPCA + Gradient Boosting     | 0.7628 ± 0.10        |
| SPCA + Bern. Naïve Bayes  | 0.7201 ± 0.13 | SPCA + XGBoost               | 0.7708 ± 0.09        |
| SPCA + LDA                | 0.7587 ± 0.10 | <b>SPCA + MLP (2 layers)</b> | <b>0.7711 ± 0.09</b> |
| SPCA + Passive Aggressive | 0.7395 ± 0.14 | SPCA + MLP (3 layers)        | 0.7553 ± 0.10        |

# Performance of ML Pipelines Searched by the Automated Prognostic Model Construction Algorithm (IV)

## Fast ICA (25 component), 5-fold CV

| Benchmark                | AUCROC        | Benchmark                      | AUCROC               |
|--------------------------|---------------|--------------------------------|----------------------|
| ICA + Logistic Reg.      | 0.7760 ± 0.10 | ICA + Linear SVM               | 0.7395 ± 0.07        |
| ICA + SGD Perceptron     | 0.6097 ± 0.13 | ICA + Random Forest            | 0.7447 ± 0.06        |
| ICA + kNN                | 0.6481 ± 0.05 | ICA + Extra Trees              | 0.7286 ± 0.06        |
| ICA + Decision Tree      | 0.5941 ± 0.02 | ICA + AdaBoost                 | 0.7567 ± 0.09        |
| ICA + Kernel SVM         | 0.6408 ± 0.09 | ICA + Bagging                  | 0.7005 ± 0.08        |
| ICA + Gauss. Naïve Bayes | 0.7165 ± 0.15 | <b>ICA + Gradient Boosting</b> | <b>0.7766 ± 0.08</b> |
| ICA + Bern. Naïve Bayes  | 0.6693 ± 0.14 | ICA + XGBoost                  | 0.7760 ± 0.09        |
| ICA + LDA                | 0.7689 ± 0.10 | ICA + MLP (2 layers)           | 0.7569 ± 0.07        |
| ICA + Passive Aggressive | 0.6939 ± 0.11 | ICA + MLP (3 layers)           | 0.7612 ± 0.07        |