amedes

DATA BUSINESS CHALLENGE

Group 3

A TEAM OF 4 **DATA SCIENTISTS** TO HELP YOU CREATE VALUE FROM YOUR DATA

Aicha BOKBOT
Data-scientist
Aicha.bokbot@hec.edu
06 63 93 41 64

Leon LEITAO
Data-scientist
leon.leitao@hec.edu
07 54 59 29 51

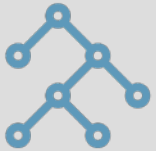
Pragya SINGH
Data-scientist
Pragya.singh@hec.edu
07 50 91 21 71

Corentin SENE
Data-scientist
corentin.sene@hec.edu
06 51 26 77 76

PROJECT PRESENTATION & OBJECTIVES

DATA ANALYSIS

OUR METHODOLOGY



MODELS & RESULTS

SUGGESTIONS, LIMITATIONS & NEXT STEPS

DATA SOLUTIONS TO AID MEDICAL PROFESSIONALS

KEY OBJECTIVES AND EXPECTED BENEFICTS

PROJECT OBJECTIVES

- Extract and **structure** the **valuable information** from treatment journals
- Create a model that **identifies patients** who may suffer from a disease

BUSINESS IMPACT

- Develop early, **accurate diagnoses**, which lead to quicker treatment and mitigate the long-term damage caused by the disease
- **Reduce risk** of misdiagnoses

OUR APPROACH

- Develop **Personalize solutions** for each disease

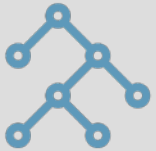
DATA AVAILABLE

PATIENT INFO | ANAMNESTIC DATA | DIAGNOSIS | LAB RESULTS

PROJECT PRESENTATION & OBJECTIVES

DATA ANALYSIS

OUR METHODOLOGY



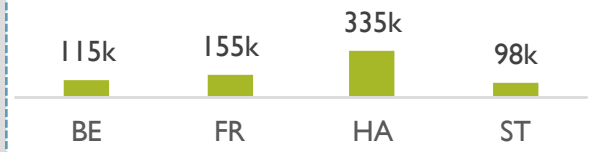
MODELS & RESULTS

SUGGESTIONS, LIMITATIONS & NEXT STEPS

NUMEROUS DATA ON PATIENTS AND YET LIMITED DATA ON SOME DISEASES

OVERVIEW OF THE DATA

702 258 patients



on which we have a **lot of data**

on average

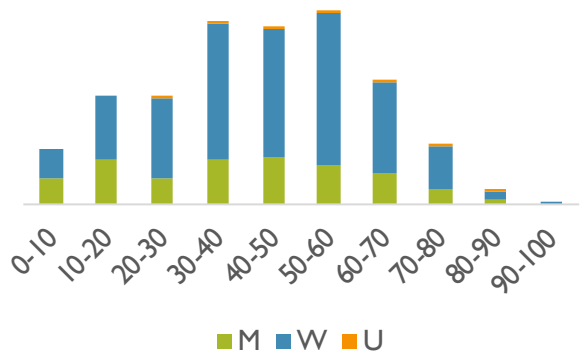
12 appointments
in **4** years
100 rows per patient

data from **1985** to **2020** on

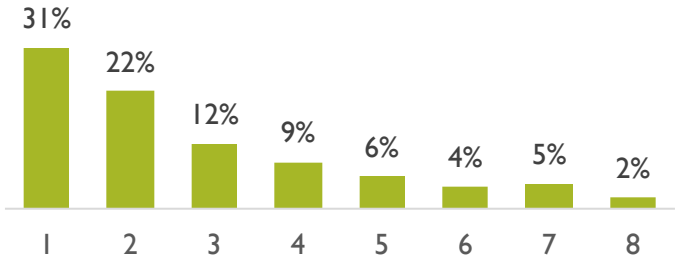
c. **6000** diseases

75% of them concern
less than **60** patients

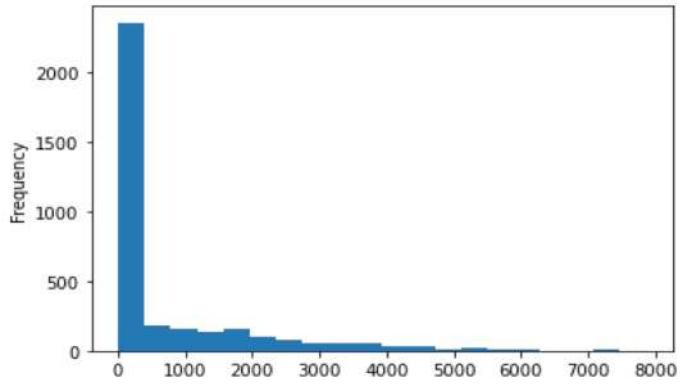
Age distribution by sex



Number of distinct diseases among diagnosed patients



Distribution of time between diagnosis and first visit (in days)

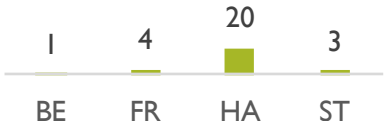


A **HETEROGENEOUS** NUMBER OF PATIENTS FOR EACH DISEASE

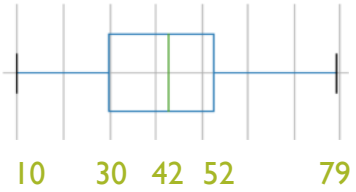
OVERVIEW OF THE 4 DISEASES

GAUCHER DISEASE – E75.22

28 patients

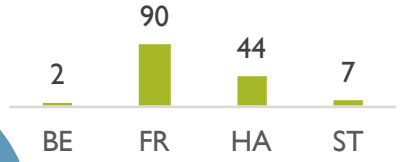


53% women
diagnosed at 43 y-o on avg

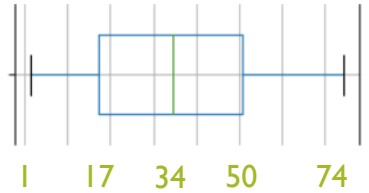


FAMILIAL HYPERCHOLESTEROLEMIA – E78.01

143 patients



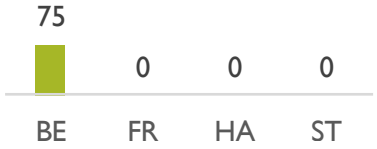
70% women
diagnosed at 34 y-o on avg



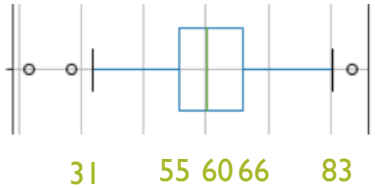
DISEASES

CHYLOMICRONEMIE – E78.3

75 patients

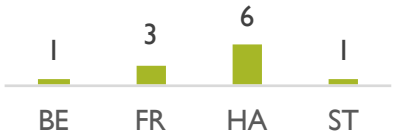


65% men
diagnosed at 60 y-o on avg

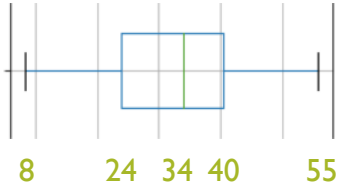


BETA OXYDATION DEFFECT – E71.3

11 patients



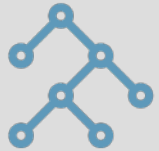
65% men
diagnosed at 31 y-o on avg



PROJECT PRESENTATION & OBJECTIVES

DATA ANALYSIS

OUR METHODOLOGY



MODELS & RESULTS

SUGGESTIONS, LIMITATIONS & NEXT STEPS

THE TEXT COLUMN COMPLETES PROVIDES **VALUABLE INFORMATION** TO PROCESS

FEATURE EXTRACTION

PATIENT_HASH	ZENTRUM_ID	PATIENT_ID	PAT_GEBDATUM	PAT_GESCHLECHT	DATUM	TYP	TYP_EXT	TEXT	ICD10	SICHERHEIT
145858	FRA01	150256	07.07.07	W	27.11.17	Y		GLU=71; HS=2.9; GPT=15; GOT=32; GGT=9; AP=259;...	NaN	NaN
145858	FRA01	150256	07.07.07	W	27.11.17	A		Jessica wird uns zur Beurteilung der Körperhöh...	NaN	NaN
145858	FRA01	150256	07.07.07	W	15.11.18	Y		GLU=107 +; HS=3.6; GPT=14; GOT=33; GGT=10; AP=...	NaN	NaN
145858	FRA01	150256	07.07.07	W	10.01.19	*		Hypercholesterinämie	E78.0	G

Age

Average age of the patient (between the first and last visit)

Sex

Sex of the patient

Test results

Results of a selection of relevant tests

Symptoms

Does the patient show any relevant symptoms?

Co-morbidity

Does the patient have any other relevant diseases?

METHODOLOGY BASED ON THE SEARCH FOR **SYMPTOMS**, **TESTS** AND **CO-MORBIDITY**

OUR METHODOLOGY

Preprocessing

- **DATUM:** convert to Datetime (e.g."10.06.50", "10.06.90", "31.12.20")
- **PAT_GESCHLECHT**: convert to Datetime (e.g. "27.19.19")
- **ICD10**: treat missing values
- **SICHERHEIT**: treat missing values
- **TEXT**: interpret test results (from MCV=99.2 + to MCV = High)

For each of the four diseases:

1 – Research

Research on the disease to identify:

1. Symptoms: what are the common symptoms?
2. Tests: which tests enable to diagnose the disease?

2 – Pattern recognition in the data

Analyze the data of people diagnosed with the disease to identity:

1. Symptoms: what are the most recurring symptoms?
2. Tests: which are the most recurring tests with abnormal results?
3. Co-morbidity: which other other diseases do the patients also have?

3 – Feature creation

Build the features

1. Symptom
2. Test results
3. Co-morbidity

COMBINING **RESEARCH** AND **DATA**

EXAMPLE : GAUCHER DISEASE

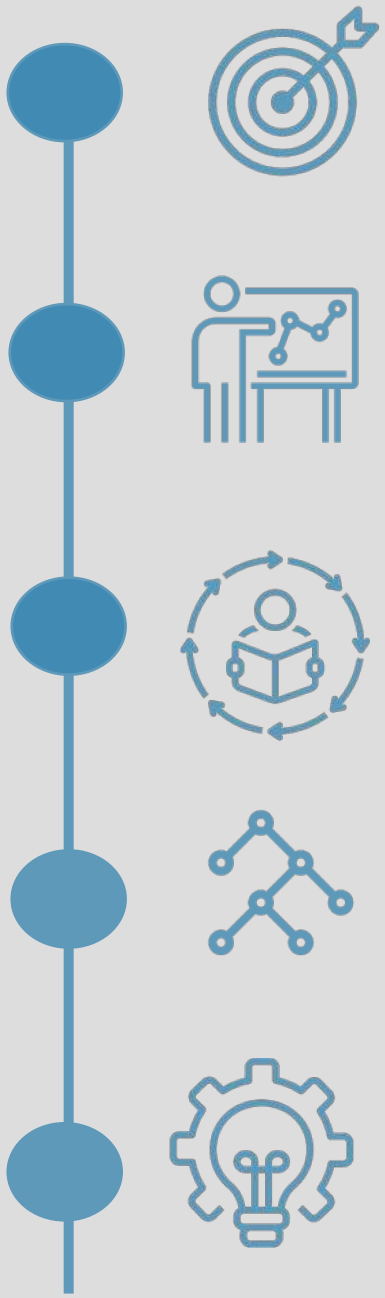
1 – Research

- **Reasons for referral:**
  - Splenomegaly
  - Hepatosplenomegaly
  - Bone Involvement
  - Chololithiasis
  - Thrombocytopenia
  - Pancytopenia
  - Leucopenia
  - Anemia
  - Member of patient family
- **Diagnosis**
  - Enzyme test called Beta-glucosidase leukocyte (BGL) test

2 – Pattern recognition in the data

- **Symptoms**
  - Fatigue
  - Bone pain
  - Splenomagaly
  - Thrombocytopenia
- **Co-Morbidity**
  - E55.9: Vitamine D defficiency
  - I10.90: Hypertension
  - D69.61: Thrombocytopenia
  - G93.3: Fatigue Syndrom
  - R16.1: Splenomegaly
- **Relevant tests**
  - High Osteocalcin (68%)
  - High Kappe Free Light chains (48%)
  - High Albumin (36%)
  - Low Thrombocytes (36%)
  - High Ferritin (32%)
  - Low Transferrin saturation (32%)
  - High DPD (32%)
  - High GGT (32%)
  - Low MCH (32%)
  - Low Hematocrit (32%)

Source: "Gaucher disease: Biochemical and molecular findings in 141 patients diagnosed in Greece", Molecular Genetics and Metabolism Reports, Volume 24, 2020



PROJECT PRESENTATION & OBJECTIVES

DATA ANALYSIS

OUR METHODOLOGY

MODELS & RESULTS

SUGGESTIONS, LIMITATIONS & NEXT STEPS

ADABOOST ALGORITHM ALLOWS US TO OBTAIN A 68% RECALL

MODELING METHODOLOGY

Model intuition

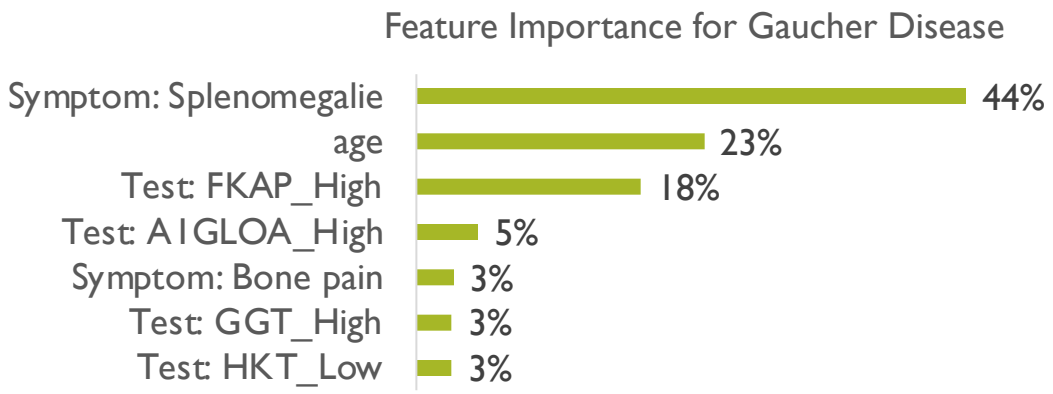
Patient ID	Sex	Age	Test results	Symptoms	Co-morbidity	diagnosed
	…	…	…	…	…	1
	…	…	…	…	…	1
	…	…	…	…	…	0
	…	…	…	…	…	0
	…	…	…	…	…	0

Best performing model

AdaBoost Classifier

Scores after hyper-parameter tuning and resampling :

- Classifies correctly 68% of positive patients
- Classifies correctly 99% of negative patients
- Accuracy: 96%



ADABOOST ALGORITHM ALLOWS US TO OBTAIN A **91% ACCURACY** ON THE DETECTION OF NEGATIVE PATIENTS

MODELING METHODOLOGY
Example on Gaucher Disease

Dataset definition

Prevalence of Gaucher Disease: 1/40000
We have 32 patients with GD : we need a sample of size 1.28 million to match the prevalence!

How to define the dataset on which to run and evaluate the model? How many non-GD patients to pick?

Proportion of patients with Gaucher Disease

- Berlin : 0.017 %
- Frankfurt : 0.066 %
- Hamburg : 0.319 %
- Stuttgart : 0.373 %

minimum = 0.017 % / # diagnosed = 32
=> 32 / 0.017 % = 188 000

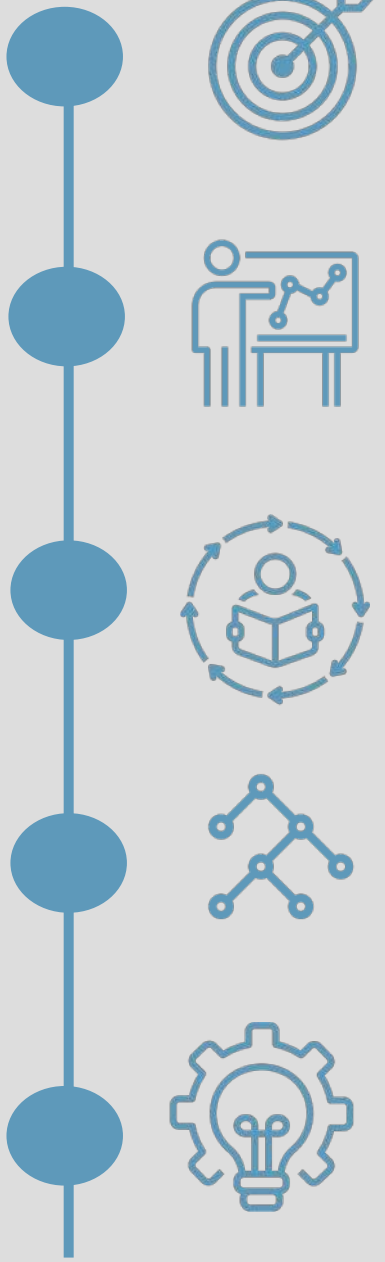
Model Results with size 1880

Best performing models

```
Logistic Regression 0.989
      Positives:  56.00000000000001 % misclassifed      18 / 32
      Negatives:  0.0 % misclassifed      3 / 1798
Decision Tree 0.981
      Positives:  47.0 % misclassifed      15 / 32
      Negatives:  1.0 % misclassifed      18 / 1798
Neural Net 0.986
      Positives:  59.0 % misclassifed      19 / 32
      Negatives:  0.0 % misclassifed      6 / 1798
AdaBoost 0.986
      Positives:  50.0 % misclassifed      16 / 32
      Negatives:  1.0 % misclassifed      9 / 1798
```

After Hyperparameter Tuning and Resampling (SMOTE and Edited Nearest Neighbors Undersampling)

```
AdaBoost
      Positives:  19.0 % misclassifed      6 / 32
      Negatives:  9.0 % misclassifed      167 / 1798
```



PROJECT PRESENTATION & OBJECTIVES

DATA ANALYSIS

OUR METHODOLOGY

MODELS & RESULTS

SUGGESTIONS, LIMITATIONS & NEXT STEPS

MORE **ACCURATE DETECTION** THROUGH MORE DETAILED DATA PROCESSING

NEXT STEPS

Extract more information from the anamnestic data using natural language processing techniques

Improve the model performance by enriching data with external datasets

Build a more global model that can detect any disease

OUR SUGGESTIONS

Data quality

- Reducing missing values (columns ICD10, SICHERHEIT)
- Date formatting (year with 4 digits)

Standardization

- Laboratory test codes
- Diseases names

Additional information

- Subcategories in doctor notes (symptom, diagnosis, treatment)
- Specialty of doctor (GP, specialist)

A MODEL HAS BEEN DEVELOPED FOR **EACH OF THE DISEASES**

EXECUTIVE SUMMARY

KEY OBJECTIVE	Generate **business value** from Amedes' treatment journals using **machine learning**
OUR APPROACH	Develop **personalized solutions** for each disease based on **research** and **pattern recognition** from the data
FEATURES	Create features based on **tests**, **symptoms** and **co-morbidity**
INTERPRETABILITY	Relevant **interpretabiliity** as we based our appoach on disease characteristics
RESULTS	**Algorithms** that allow us to **detect** diseases with a **relevant accuracy**

APPENDIX

COMBINNIG **RESEARCH** AND **DATA**

EXAMPLE : β-oxidation defect

1 – Research

- **Reasons for referral:**
  - Adrenoleukodystrophy
  - Adrenomyeloneuropathy
  - Member of patient family

2 – Pattern recognition in the data

- **Symptoms**
  - Fatigue
  - Abdominal pain
  - Adrenal insufficiency
  - Irritability
- **Relevant tests**
  - High SHGB Protein
  - High Thyroxine
  - Low Red blood cell level
  - Low Uric acid
- **Co-Morbidity**
  - E27.1 : Primary adrenocortical insufficiency
  - E06.3 : Autoimmune thyroiditis
  - G40.9 : Epilepsy
  - M62.89 : Other specified disorders of muscle
  - G40.6 : Grand mal seizures

COMBINNIG **RESEARCH** AND **DATA**

EXAMPLE : Hypercholesteramie

1 – Research

- **Reasons for referral:**
  - Chest pain
  - Family History
  - Member of patient family

2 – Pattern recognition in the data

Symptoms:

- Cholesterol deposits in the eyelids
- Chest pain
- Sudden stroke-like symptoms

- **Relevant tests**
  - LDL Tests
  - GGT Tests
  - MGV Tests