



ADVANCED DATA MINING APPROACH

Prediction of costs associated with the hospital management of HIV patients in France

INTRODUCTION

Data mining is not new and has been efficiently employed in others domains (*Bank, Cybernetics, Marketing, Energy, etc.*)¹. The amount of data collected in **medical information systems** is tremendous, but much data remains unused because of their complexity². This still increasing volume of data requires new analytical approaches that are efficient, sensitive and better than classical statistics to handle Big Data¹.

Figure 1
METHODOLOGICAL DIVERGENCES BETWEEN DATA MINING AND STATISTICS

DATA MINING	CRITERIA	STATISTICS
Explanatory approach, discover the unknown with no preconception	PARADIGM	(Un)-validate predefined hypothesis, risk of experimenter bias
Unrivalled quality of predictions and explanations, using cross-validations. Use extrapolation for missing data	MAIN FEATURE	Precise measurement of uncertainty, comparison of 2 populations, confidence assessment of measures
High adaptability and reusability in other domains	SECOND FEATURE	Mathematical definition of statistical tests ensures significant results
Decision Tree, Neural Network, Bayesian, Genetic Algorithm, Support Vector Machine	EXAMPLES OF TECHNIQUES	Principal Component Analysis, p-value testing, multivariate & univariate regression
Capable of dealing with millions of data ("Big data")	VOLUME	Suitable for middle size datasets (< 1 million observations)
Data mining selects and tells you the non-linear combinations of many variables that best explain the value of the target variable.	TYPICAL RESULTS	Logistic regression tells you that X% of the cost is explained by having this or this comorbidity, PCA tells you if variables are redundant.

OBJECTIVE

The main objective of this study is to assess the capability of a state-of-the-art data mining technique to be applied on a healthcare database in order to **better understand drivers of healthcare expenditure and the management of diseases**.
We utilized the approach in order to explain the patient features that are the main drivers of cost associated with HIV patients' hospital management. A secondary objective is to assess how reproducible such an approach is on other medical databases with potentially different patient features.

METHODS

KEY FIGURES

Study period: January 2013 - December 2014
Number of HIV patients: 30,294
Number of stays: 70,180
Total annual cost: 180 million €
Mean annual cost per patient : 5,981 € (+/- 10,661)
Nb of patient profiles found using Data Mining: 165

DATA EXTRACTION

The PMSI-MSO (*French Medical Information System - Medicine, Surgery, Obstetric units*) database was used to extract all hospital stays in 2013 with at least one of the following HIV ICD-10 (*International Classification of Diseases, 10th revision*) codes as principal diagnosis, related diagnosis or significantly associated diagnosis: **B20***, **B21***, **B22***, **B23***, **B24***.

1-YEAR FOLLOW-UP OF PATIENTS & CLASSIFICATIONS OF STAYS

For each patient hospitalized once in 2013 with an HIV code, a 12 months follow-up was performed to capture any further hospital stay. The same algorithm and medical review as for the inclusion stay were conducted on these additional stays, in order to exclude stays not related to HIV. All stays were classified by HIV clinician experts into **10 groups of comorbidities** and **5 groups of opportunistic infections (OI)** using ICD-10 codes (*figure 2*).

HIV HOSPITALIZATION COST

The cost of stays were evaluated from a **NHS perspective**. The PMSI database provides many details about stays' cost as it was specifically designed for economic evaluation of hospital activities.

OUR DATA MINING ALGORITHM

Data were analyzed with a supervised-learning data mining technique, an Enhanced Decision Tree algorithm based on Breiman' CART model³ (*figure 3*).

Figure 2
INPUT VARIABLES FOR THE ANALYSIS OF PATIENT PROFILES WITH DATA MINING

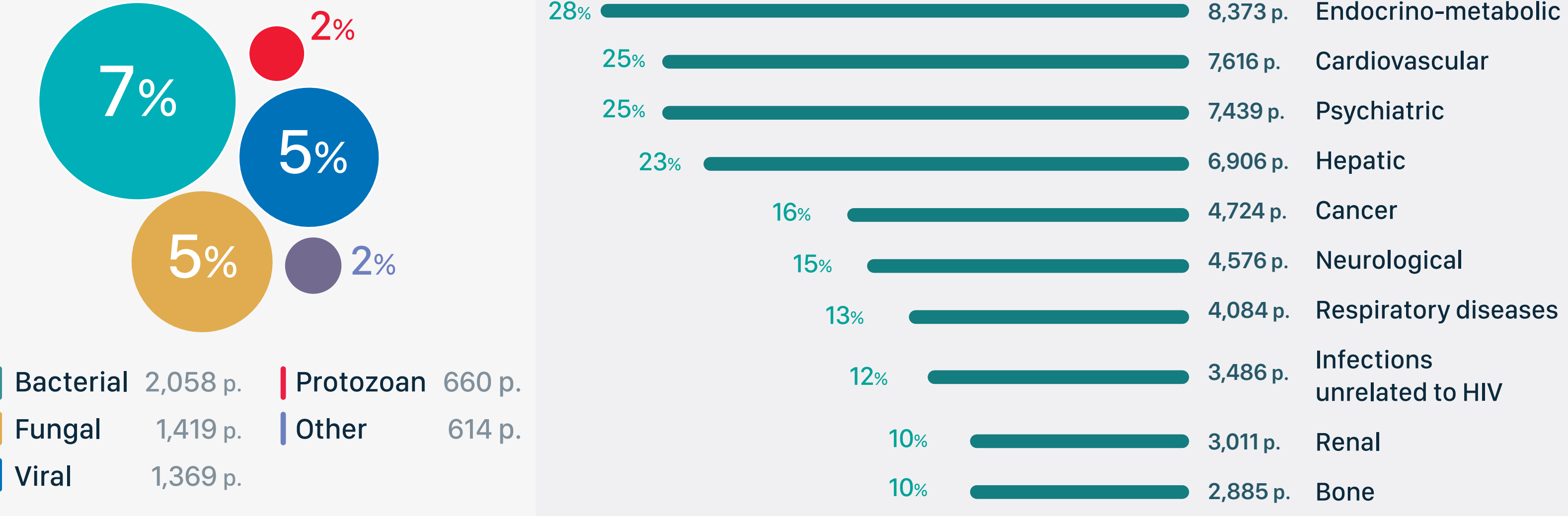
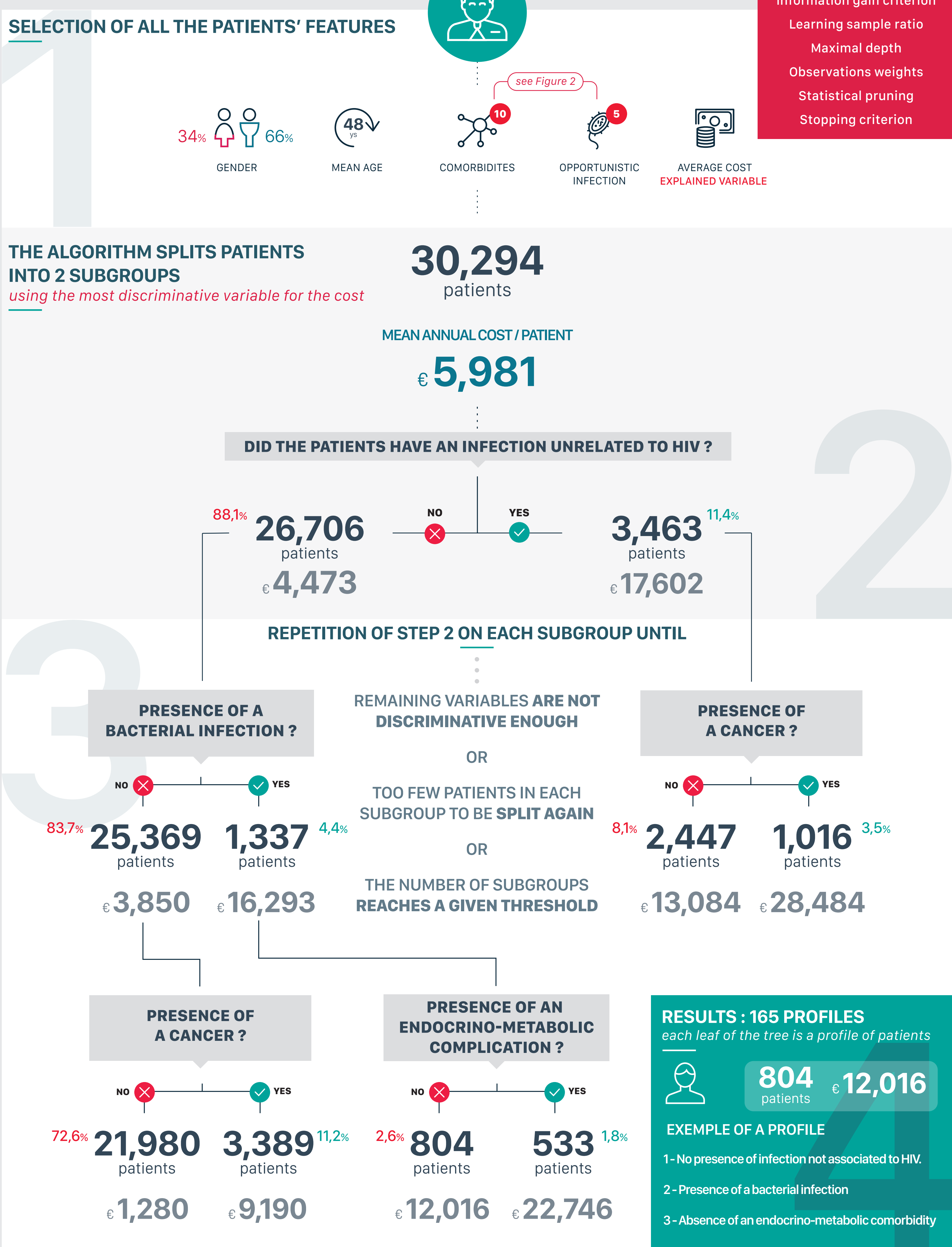


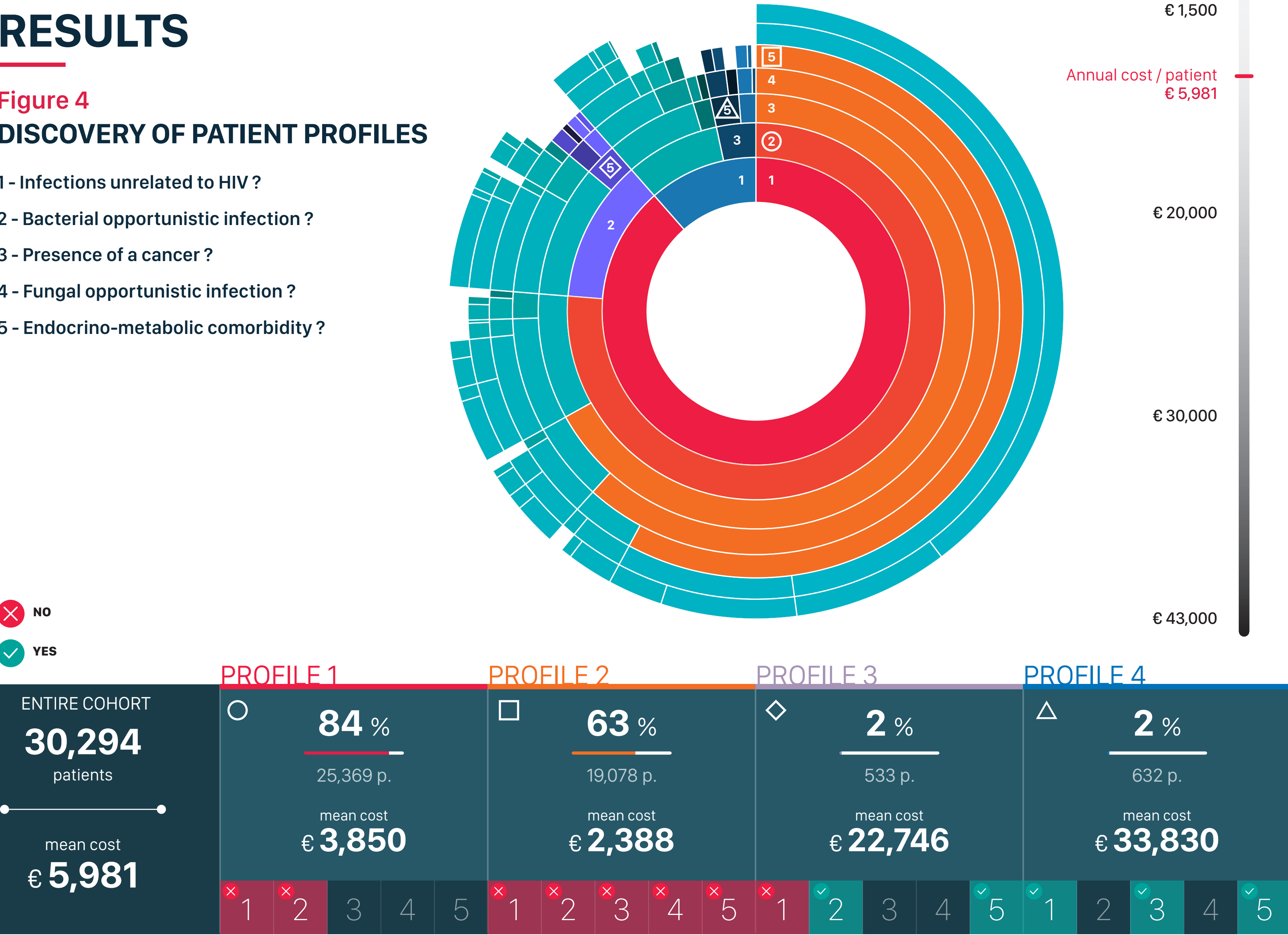
Figure 3
THE 4 STEPS OF THE DATA MINING APPROACH



RESULTS

Figure 4
DISCOVERY OF PATIENT PROFILES

- 1 - Infections unrelated to HIV ?
- 2 - Bacterial opportunistic infection ?
- 3 - Presence of a cancer ?
- 4 - Fungal opportunistic infection ?
- 5 - Endocrino-metabolic comorbidity ?



CONCLUSION

The present study is a pilot that successfully demonstrates how a data mining technique, from the field of Artificial Intelligence, can help us better understand Hospital costs for HIV patients: our data mining algorithm identified **specific patient profiles which explain the differentiating cost drivers in HIV inpatient care**. Classical statistical approaches would struggle to provide such detailed profiles with numerous combinations of variables.

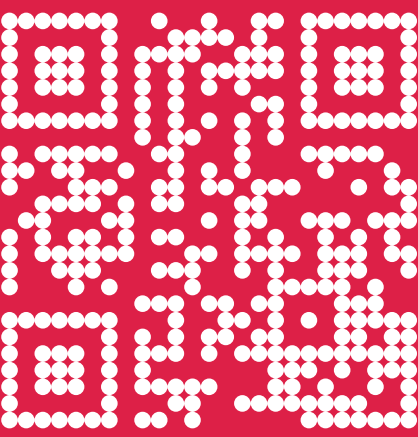
1 - This approach could work with other data sources, especially with more clinical and laboratory data (e.g.: viral load).

It can also explain drivers for any available target variables (*cost, death, patients virologically suppressed or patients dropping out of care*).

2 - Such collaborative efforts between health care professionals and engineers can lead to enhanced uses of health data that can provide new answers and thus improve disease management
Example: Preventing a bacterial infection in newly diagnosed HIV patients with Endocrino-Metabolic comorbidities, is much more cost reducing than preventing a fungal infection.

3 - Based on results of Data Mining analyses, clinicians will know the medical interventions to focus on and prioritize.

Example: Results show that newly diagnosed HIV patients are more likely to get bacterial infections if they already have neurological and renal comorbidities (=predictive capabilities).



REFERENCES : ¹Adem Karahoca, Dilek Karahoca and Mert Sanver (2012). Survey of Data Mining and Applications (Review from 1996 to Now), Data Mining Applications in Engineering and Medicine, Associate Prof. Adem Karahoca (Ed.), InTech, DOI: 10.5772/48803.
²Herland, Matthew, Taghi M Khoshgoftaar, et Randall Wald. «A review of data mining using big data in health informatics.» Journal Of Big Data, 2014: 1--35.
³Breiman, L., Friedman, J.H., Olshen, R.A., and Stone, C.J. (1984) "Classification and Regression Trees", Wadsworth, Belmont, CA. Republished by CRC Press