POLITECNICO DI TORINO

BIG Medilytics

CTB center for biomedical technology

Data-Driven Analysis to Improve Oncological Processes in Hospital

Supervisors
Prof. Silvia Anna CHIUSANO
Prof. Ernestina MENASALVAS RUIZ

Candidate
Manuel SCURTI

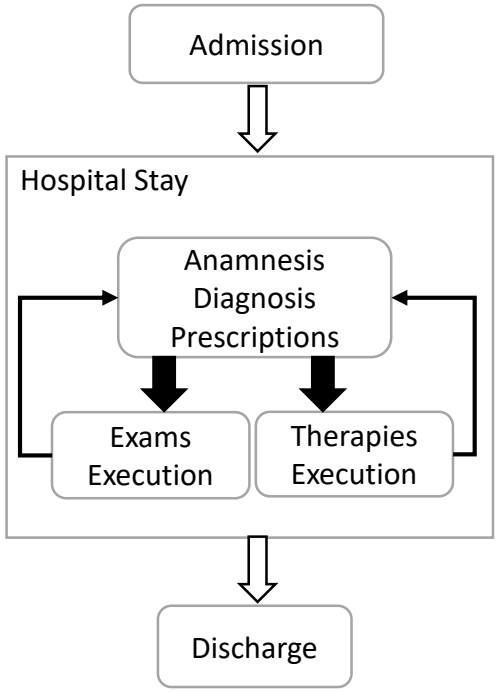
INTRODUCTION

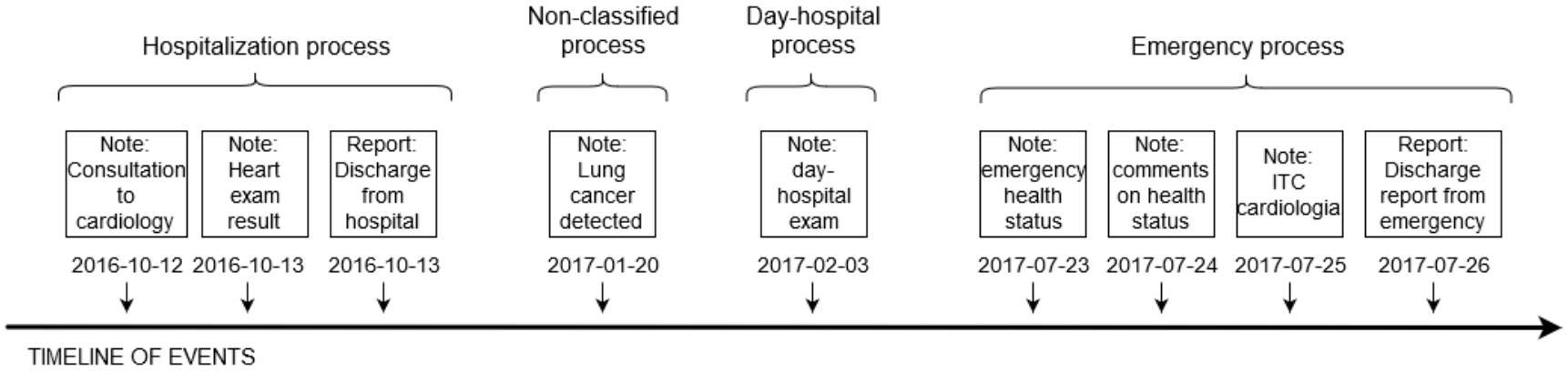
€18.8 billion [2]

- Lung cancer had highest economic cost in EU (15% of overall cancer costs in 2013)
- Low survival
- Sub-optimal management of related oncological processes

2 Fonte: Luengo-Fernandez, R., Leal, J., Gray, A., & Sullivan, R. (2013). *Economic burden of cancer across the European Union: a population-based cost analysis*.

HOSPITAL PROCESSES





Why?

- **Measurements**
  - Length of hospital stay, number of doctors involved per patient, number of unscheduled visits to ER, etc.
- **Improve hospital services**
  - Identify evidences before diagnosis that may lead physicians to clinical suspicion of lung cancer

OBJECTIVE

Enable a data-driven decisional process for clinicians, willing to optimize the management of lung cancer patients by using business intelligence methodologies with the aim of measuring the following KPI:

- KPI 1) Length of hospital stay for oncology patients
- KPI 2) Identification of patients at risk of developing lung cancer

STATE OF THE ART

- H. Baek, M. Cho, S. Kim, H. Hwang, M. Song, and S. Yoo, "**Analysis of length of hospital stay using electronic health records: A statistical and data mining approach**", 2018.

- R. Houthooft, J. Ruyssinck, J. van der Herten, S. Stijven, I. Couckuyt, B. Gadeyne, F. Ongenae, K. Colpaert, J. Decruyenaere, T. Dhaene, et al., "**Predictive modelling of survival and length of stay in critically ill patients using sequential organ failure scores**", 2015.

- J. Zuckerman, M. Ades, L. Mullie, A. Trnkus, J.-F. Morin, Y. Langlois, F. Ma, M. Levental, J. A. Morais, and J. Afilalo, "**Psoas muscle area and length of stay in older adults undergoing cardiac operations**", 2017.

- A. Almashrafi, H. Alsabti, M. Mukaddirov, B. Balan, and P. Aylin, "**Factors associated with prolonged length of stay following cardiac surgery in a major referral hospital in oman: a retrospective observational study**", 2016.

CRISP-DM (Cross-Industry Standard Process for Data Mining)

Fonte: CRISP-DM Website

Provides

- General process model
- Reliability and adaptable to any DM project
- Repeatability of experiments
- Shareable results

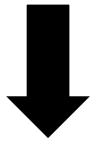
BUSINESS UNDERSTANDING

Situation Assessment

- Available Data: Anonymized EHRs from hospital (2008-2019)
  - Unstructured textual data, no track of processes, Data coming from real use-case

Business Objectives

- KPI 1 – Length of hospital stay of oncology patients
- KPI 2 – Identification of patients at risk of developing lung cancer

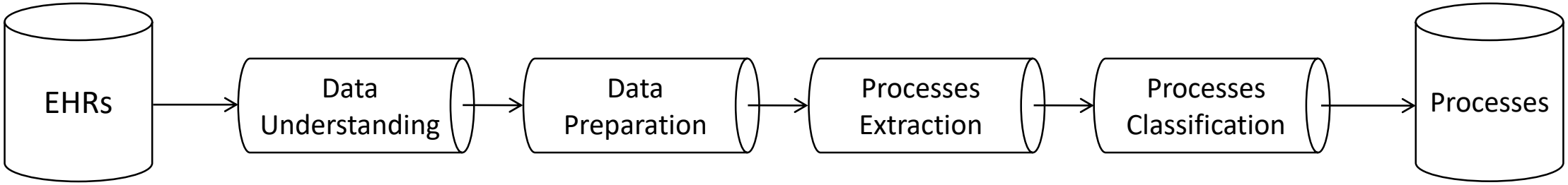


Translation to Data Analysis Goals

Data Analysis Goals

- EHRs cleaning and enrichment
- Grouping of EHRs into processes
- Classification of processes into fixed set of categories
- Analysis of processes to extract measures for KPIs

FRAMEWORK DESIGN AND DEVELOPMENT

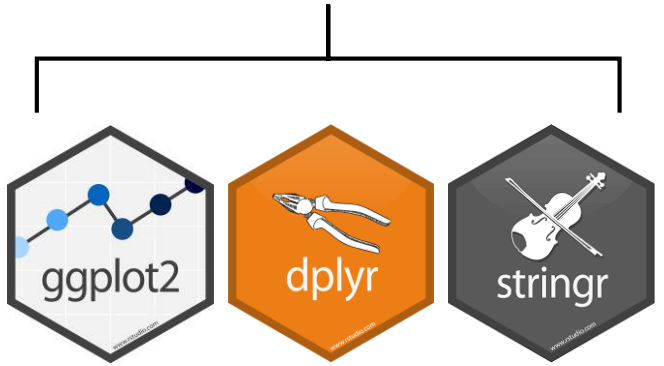


Framework Tools

- R language
- ggplot2, dplyr, stringr

Implementation requirements

- Readable code for any level of expertise
- Reusable



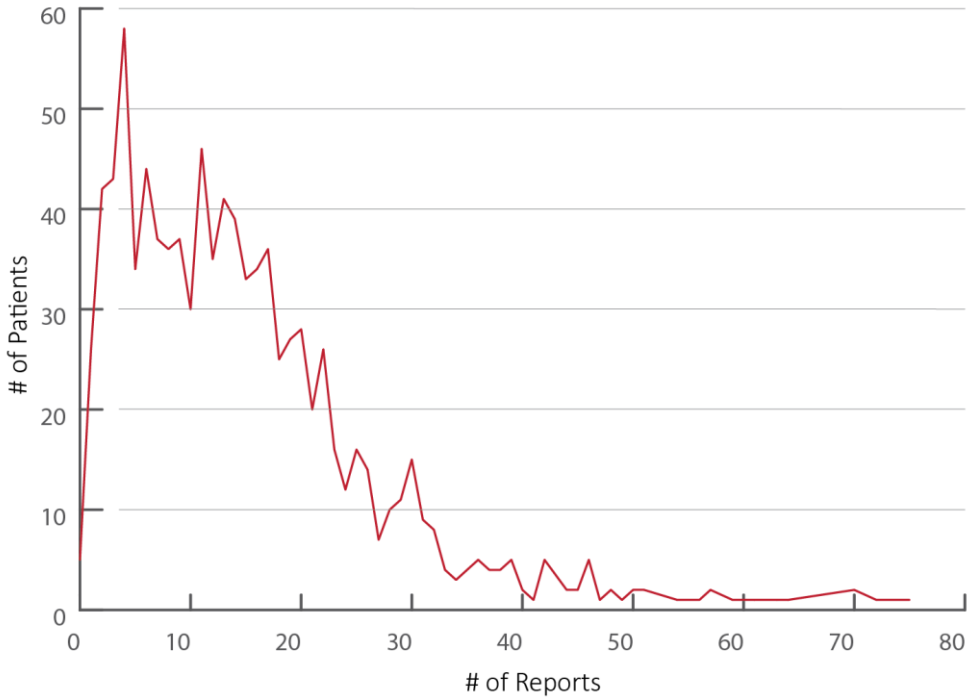
DATA UNDERSTANDING

Table	# Records	# Attributes	Descriptive analysis
Clinical Documents	296003	13	5% reports, 95% notes, unstructured texts
Patients Data	967	122	already structured

Possible part of a report
SERVICIO: Oncología Radioterápica *FECHA INGRESO: 07/03/2012 15:06* *FECHA ALTA: 14/03/2012* *MÉDICO RESPONSABLE INFORME: Dr. XXX (Médico Adjunto). Dra. YYY (Médico Residente)* *Motivo de Ingreso:* *Disnea y fiebre.* *Antecedentes Personales* *NO alergias medicamentosas conocidas.* *- No HTA, no DM, no DL.* *Antecedentes Familiares:* *-*

Sample Text in Clinical Note
Paciente que acude sin cita para solicitar control analítico. *AP: ex ADVP hace 10a. No ttos. F. 30cig/d. No B.* *AF: esposo VHC tratado con respuesta viral sostenida.* *MC: VHC+ conocido hace 10a. Seguida por su médico de primaria. Siempre transas y ECO normales. Solicito analítica con carga viral y genotipo y ECO.*

Distribution of reports per patient



DATA PREPARATION

Data Cleaning

- Low-case letters
- Special characters removal
- Stop-words removal

Data Enrichment

- Hospitalization and discharge dates
- Services names and causes of hospitalization
- **Document classification**

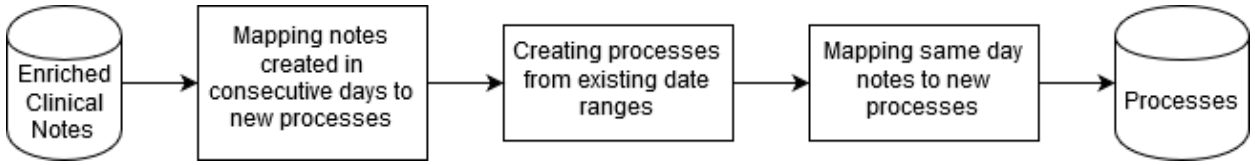
Documents File

EHR	id	creation	ingress	discharge	category	doc_type
34	27	07/07/2016			er admission	em
34	19	08/10/2018			comments hos	hos
34	84	08/07/2016			hos follow-up	hos
34	54	09/08/2019		09/08/2019	death report	nc
34	36	01/03/2017			evolution cex	cons
34	58	15/10/2018	08/10/2018	15/10/2018	discharge report	hos
34	71	01/03/2017			hday cex	day
34	5	09/07/2016			hos therapies	hos
34	3	13/10/2018			hos evolution	hos
34	78	05/08/2019	05/08/2019		admission note	nc

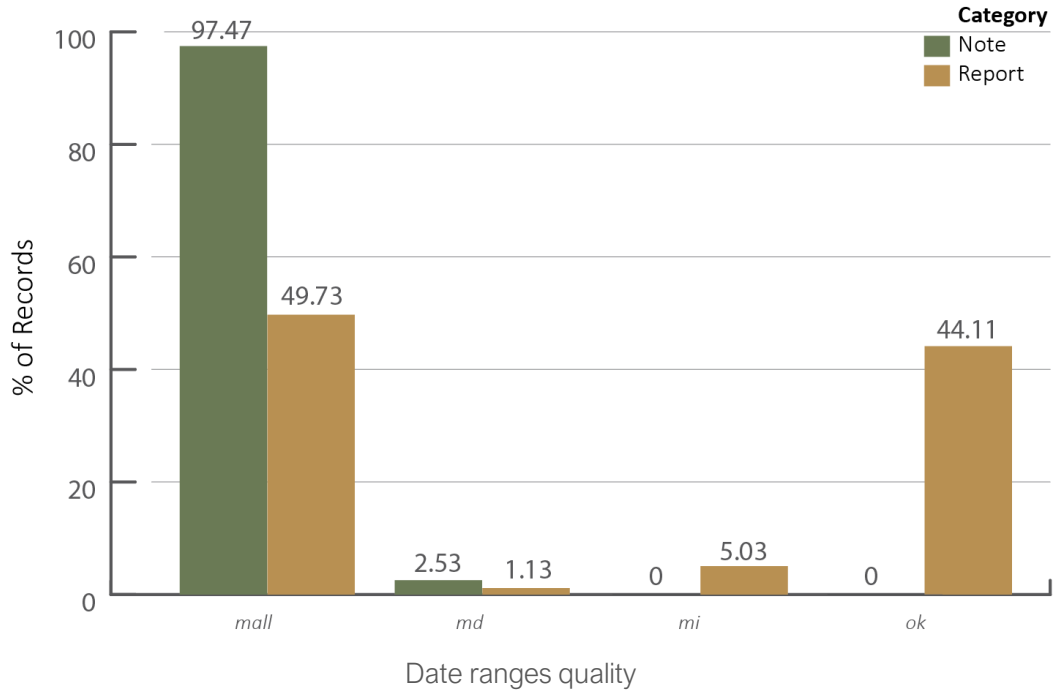
Used for document classification

Report BEFORE	Data Cleaning	Report AFTER
SERVICIO: Oncología Radioterápica FECHA INGRESO: 08/10/2018 FECHA ALTA: 15/10/2018 Motivo de ingreso: Disnea	⇨	servicio oncologia radioterapica fecha ingreso 08/10/2018 fecha alta 15/10/2018 motivo ingreso disnea

PROCESSES EXTRACTION: IDEA



Evaluation of extracted date ranges from texts



<u>Coverage by only using existing date ranges</u>: 65.4%

Assumptions

- Consecutive-days processes
- Single-day processes

PROCESSES EXTRACTION

Documents File

EHR	document	creation	ingress	discharge	category
34	27	07/07/2016			er admission
34	19	08/10/2018			comments hos
34	84	08/07/2016			hos follow-up
34	54	09/08/2019		09/08/2019	death report
34	36	01/03/2017			evolution cex
34	58	15/10/2018	08/10/2018	15/10/2018	discharge report
34	71	01/03/2017			hday cex
34	5	09/07/2016			hos therapies
34	3	13/10/2018			hos evolution
34	78	05/08/2019	05/08/2019		admission note

Processes Table

EHR	pid	ingress	discharge
34	1	08/07/2016	09/07/2016

STEPS

For each patient:

1. Search groups of documents created in consecutive days

PROCESSES EXTRACTION

Documents File

EHR	document	creation	ingress	discharge	category
34	27	07/07/2016			er admission
34	19	08/10/2018			comments hos
34	84	08/07/2016			hos follow-up
34	54	09/08/2019		09/08/2019	death report
34	36	01/03/2017			evolution cex
34	58	15/10/2018	08/10/2018	15/10/2018	discharge report
34	71	01/03/2017			hday cex
34	5	09/07/2016			hos therapies
34	3	13/10/2018			hos evolution
34	78	05/08/2019	05/08/2019		admission note

Processes Table

EHR	pid	ingress	discharge
34	1	07/07/2016	09/07/2016
34	2	08/10/2018	15/10/2018

STEPS

For each patient:

1. Search groups of documents created in consecutive days
2. Use existing date ranges from texts

PROCESSES EXTRACTION

Documents File

EHR	document	creation	ingress	discharge	category
34	27	07/07/2016			er admission
34	19	08/10/2018			comments hos
34	84	08/07/2016			hos follow-up
34	54	09/08/2019		09/08/2019	death report
34	36	01/03/2017			evolution cex
34	58	15/10/2018	08/10/2018	15/10/2018	discharge report
34	71	01/03/2017			hday cex
34	5	09/07/2016			hos therapies
34	3	13/10/2018			hos evolution
34	78	05/08/2019	05/08/2019		admission note

Processes Table

EHR	pid	ingress	discharge
34	1	07/07/2016	09/07/2016
34	2	08/10/2018	15/10/2018
34	3	01/03/2017	01/03/2017

STEPS

For each patient:

1. Search groups of documents created in consecutive days
2. Use existing date ranges
3. Search groups of documents created in the same day

Coverage

- Step 1 – 76.3%
- Step 2 – 78.5%
- Step 3 – 91.8%

Final coverage: 91.8%

CLASSIFYING PROCESSES

Documents File

EHR	document	creation	ingress	discharge	category	doc_type
34	27	07/07/2016			er admission	em
34	19	08/10/2018			comments hos	hos
34	84	08/07/2016			hos follow-up	hos
34	54	09/08/2019		09/08/2019	death report	nc
34	36	01/03/2017			evolution cex	cons
34	58	15/10/2018	08/10/2018	15/10/2018	hos discharge report	hos
34	71	01/03/2017			hday cex	day
34	5	09/07/2016			hos therapies	hos
34	3	13/10/2018			hos evolution	hos
34	78	05/08/2019	05/08/2019		admission note	nc

Processes Table

EHR	pid	ingress	discharge	em	hos	cons	day	nc	category
34	1	07/07/2016	09/07/2016	1	2	0	0	0	urg-hos
34	2	08/10/2018	15/10/2018	0	3	0	0	0	home-hos
34	3	01/03/2017	01/03/2017	0	0	1	1	0	hdia-home

Example features

Target

- Set of manually defined if-else rules for classification
- 12 handcrafted features describing process contents
- Categories defined by clinicians

Results

- 99.1% of classified processes

Drawbacks

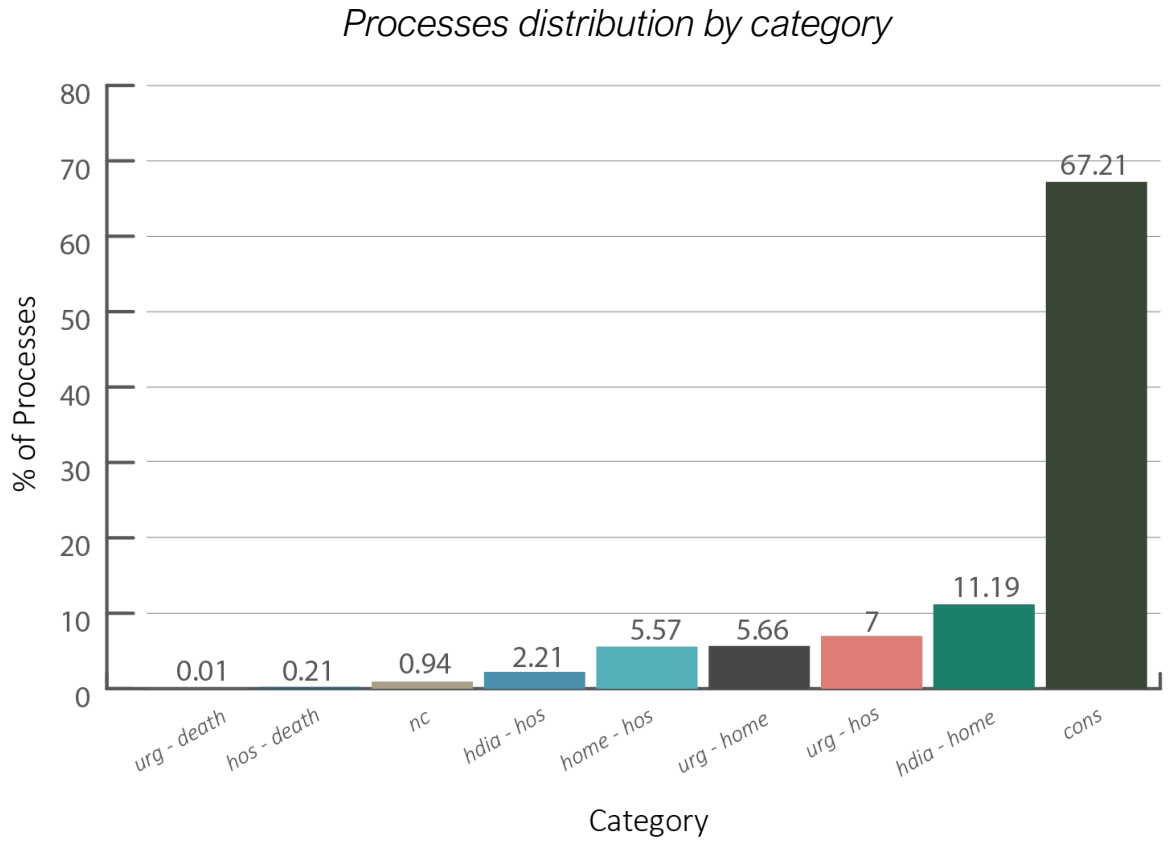
- No exact method to determine the correctness of the classification

RESULTS

- 48849 processes, 81 processes/patient
- 4898 processes describing service visited and cause of hospitalization

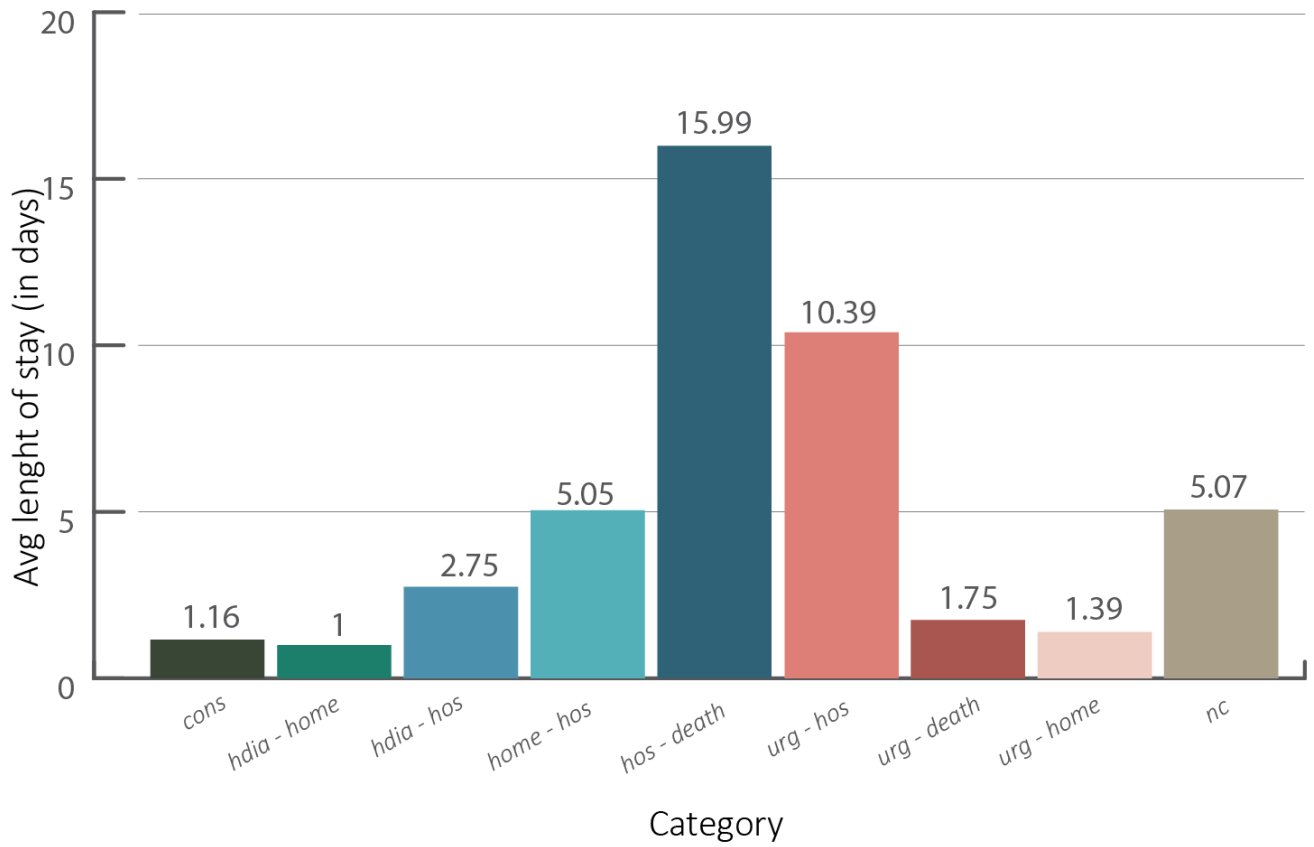
Validation criteria

- Highest coverage
- Process duration similar to expectations



RESULTS: KPI 1

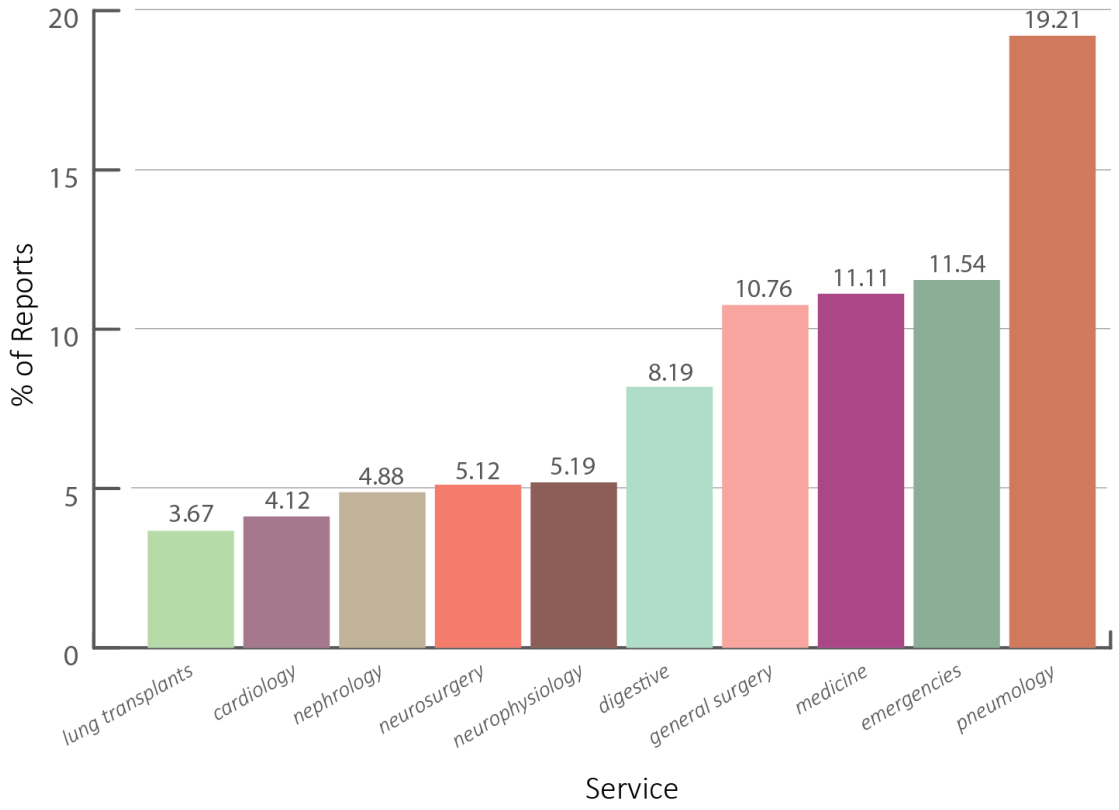
Length of hospital stay by type of process



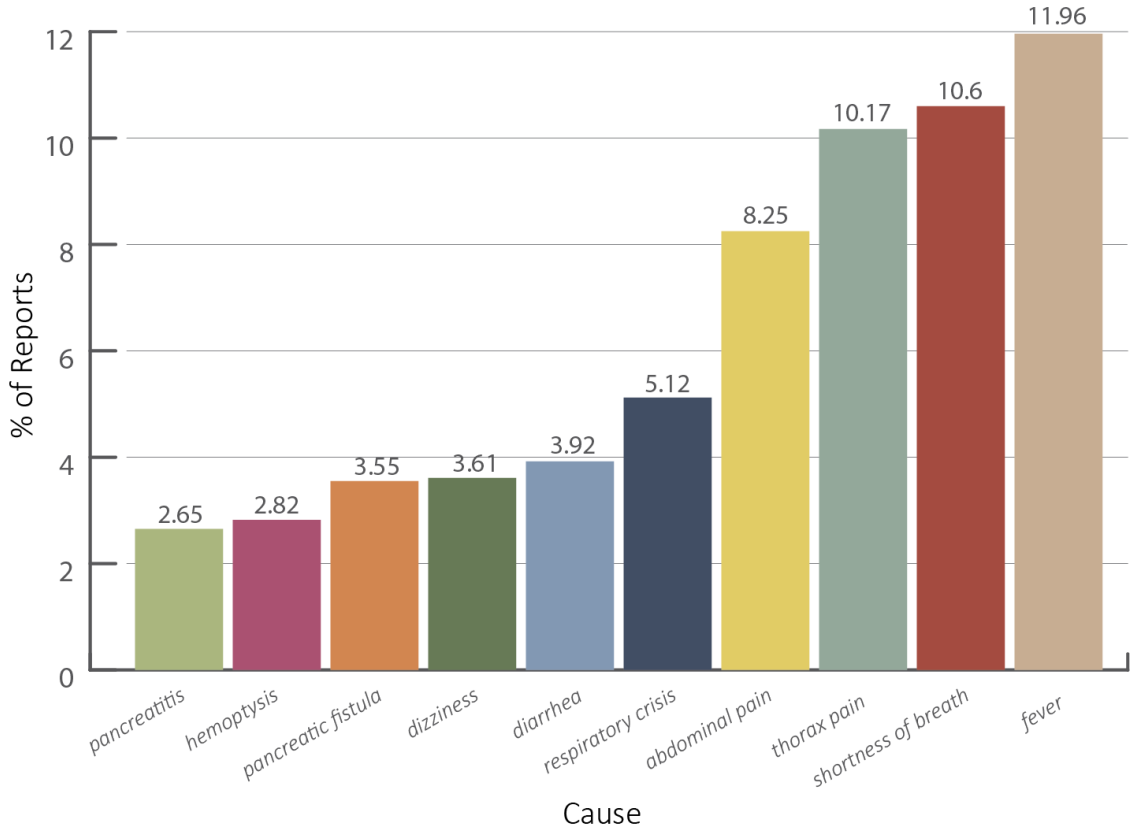
Category	LOS (days)	SD	# of PID
cons	1.16	2.71	32829
hdia-home	1.00	0.00	5467
hdia-hos	2.75	4.05	1079
home-hos	5.05	7.44	2723
hos-death	15.99	20.91	101
urg-hos	10.39	9.77	3421
urg-death	1.75	0.50	4
urg-home	1.39	2.54	2767
nc	5.07	17.55	458

RESULTS: KPI 2

Top 10 most used services before diagnosis



Top 10 most common causes of hospitalization before diagnosis



CONCLUSIONS AND FUTURE WORK

- Find relation between objectives and demographics, habits and comorbidities of patients
- Improve quality of processes by understanding contents of texts
- Extract informations from administrative documents to validate results
- Integrate multiple sources of data to measure other KPIs
- Integrate data from other hospitals to discover common patterns

POLITECNICO DI TORINO

BIG Medilytics

CTB center for biomedical technology

Data-Driven Analysis to Improve Oncological Processes in Hospital

Supervisors
Prof. Silvia Anna CHIUSANO
Prof. Ernestina MENASALVAS RUIZ

Candidate
Manuel SCURTI