Salut/GSS Gestió de Serveis Sanitaris

Universitat de Lleida

Machine Learning Approaches for Comprehensive Analysis of Population Cancer Registry Data

Ph.D Defense

Authored by
Dídac Florensa Cazorla

Supervised by
Pere Godoy Garcia
Francesc Solsona Tehàs
Jordi Mateo Fornés

Industrial supervision by
Miquel Mesas Julió

19th April 2023

Salut/GSS Gestió de Serveis Sanitaris

Contents

Universitat de Lleida

1. Motivation
2. Introduction
3. Hypotheses
4. Objectives
5. Methodology (6 papers)
6. Conclusions and future research

Motivation

Main barriers

- Continuously medical recording and massive data
- Isolation between databases
- Not use of the medical information

Add value to the data

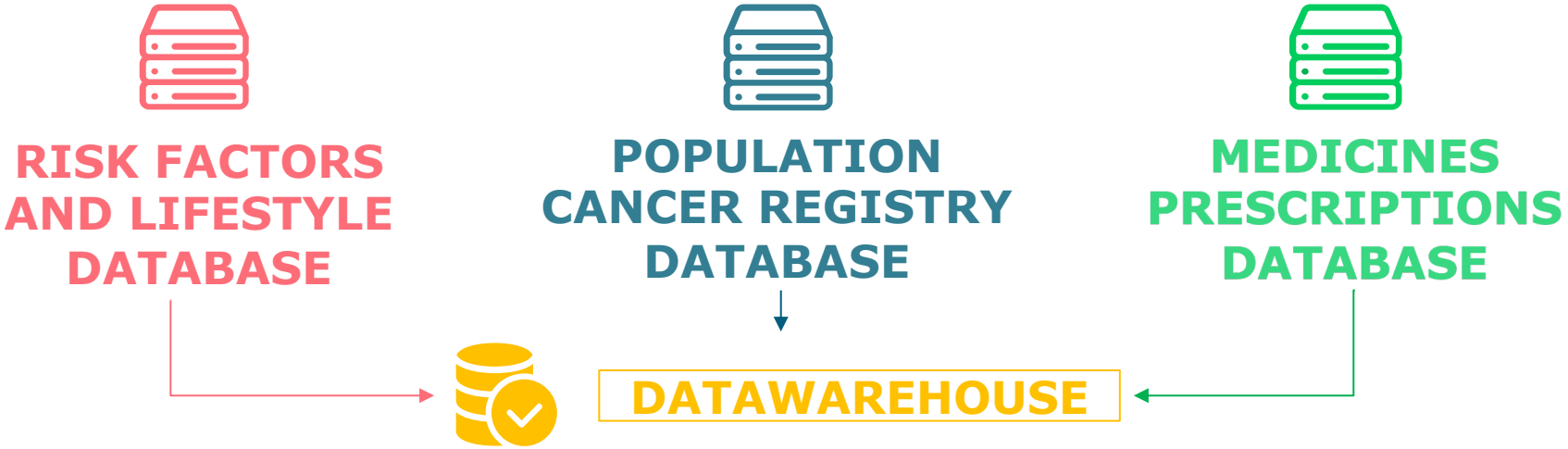
Motivation

RISK FACTORS AND LIFESTYLE DATABASE

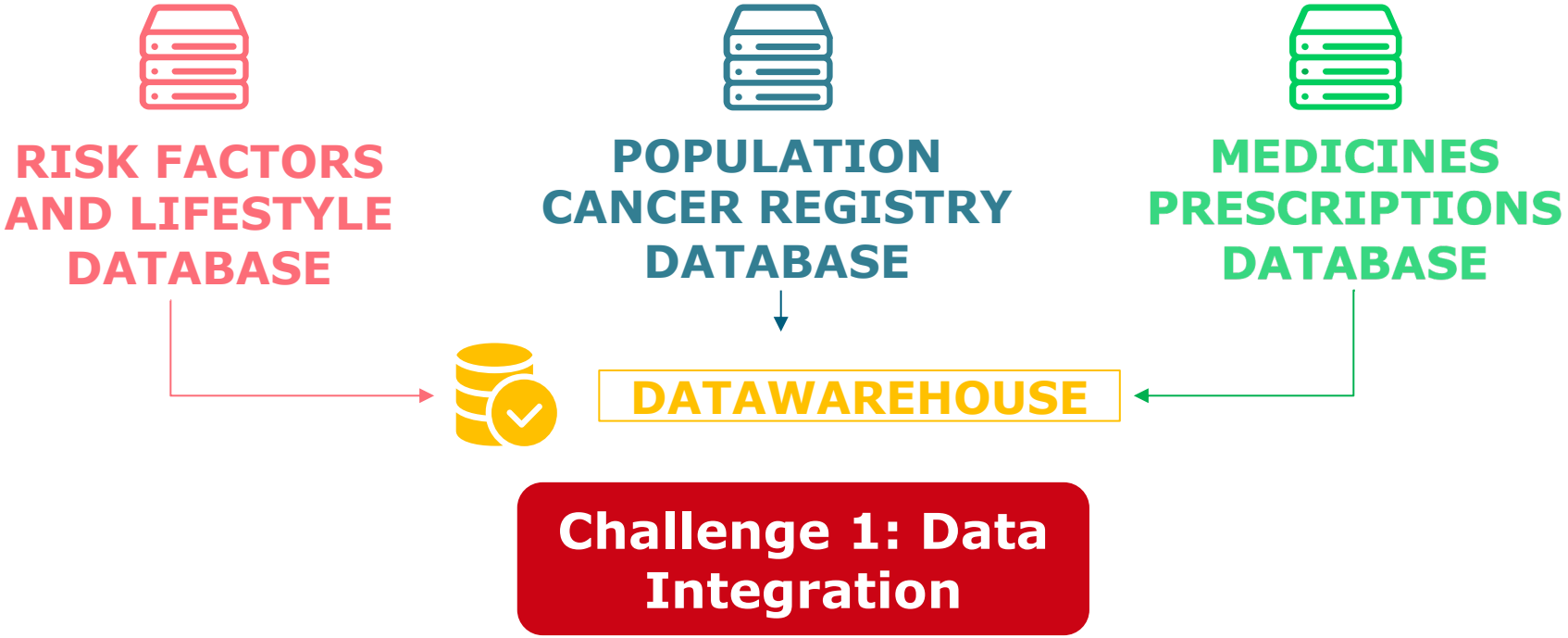
POPULATION CANCER REGISTRY DATABASE

MEDICINES PRESCRIPTIONS DATABASE

Motivation


RISK FACTORS AND LIFESTYLE DATABASE
POPULATION CANCER REGISTRY DATABASE
MEDICINES PRESCRIPTIONS DATABASE
DATAWAREHOUSE

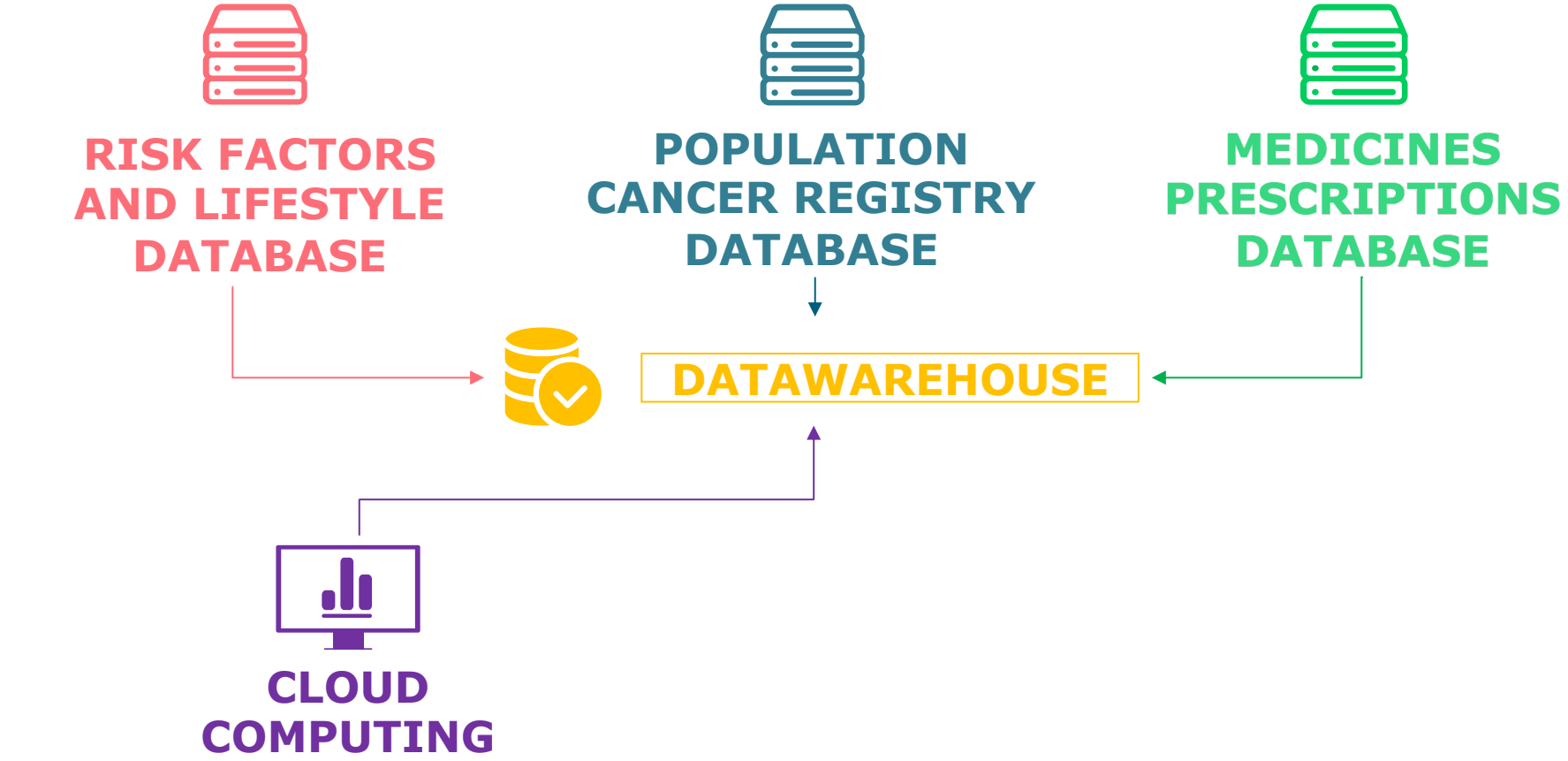
Motivation


RISK FACTORS AND LIFESTYLE DATABASE
POPULATION CANCER REGISTRY DATABASE
MEDICINES PRESCRIPTIONS DATABASE
DATAWAREHOUSE
Challenge 1: Data Integration

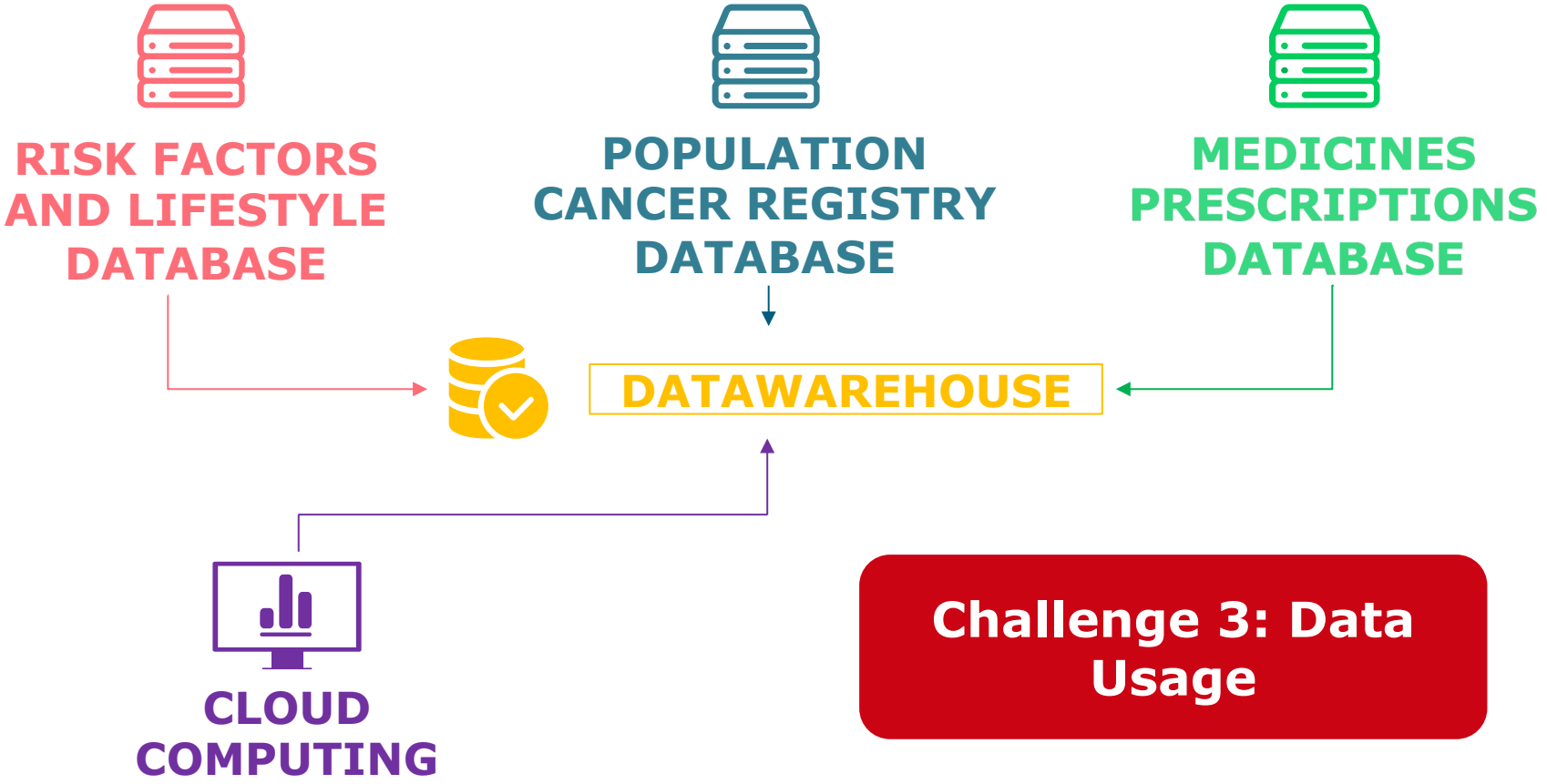
Motivation

Challenge 2: Data Accessibility

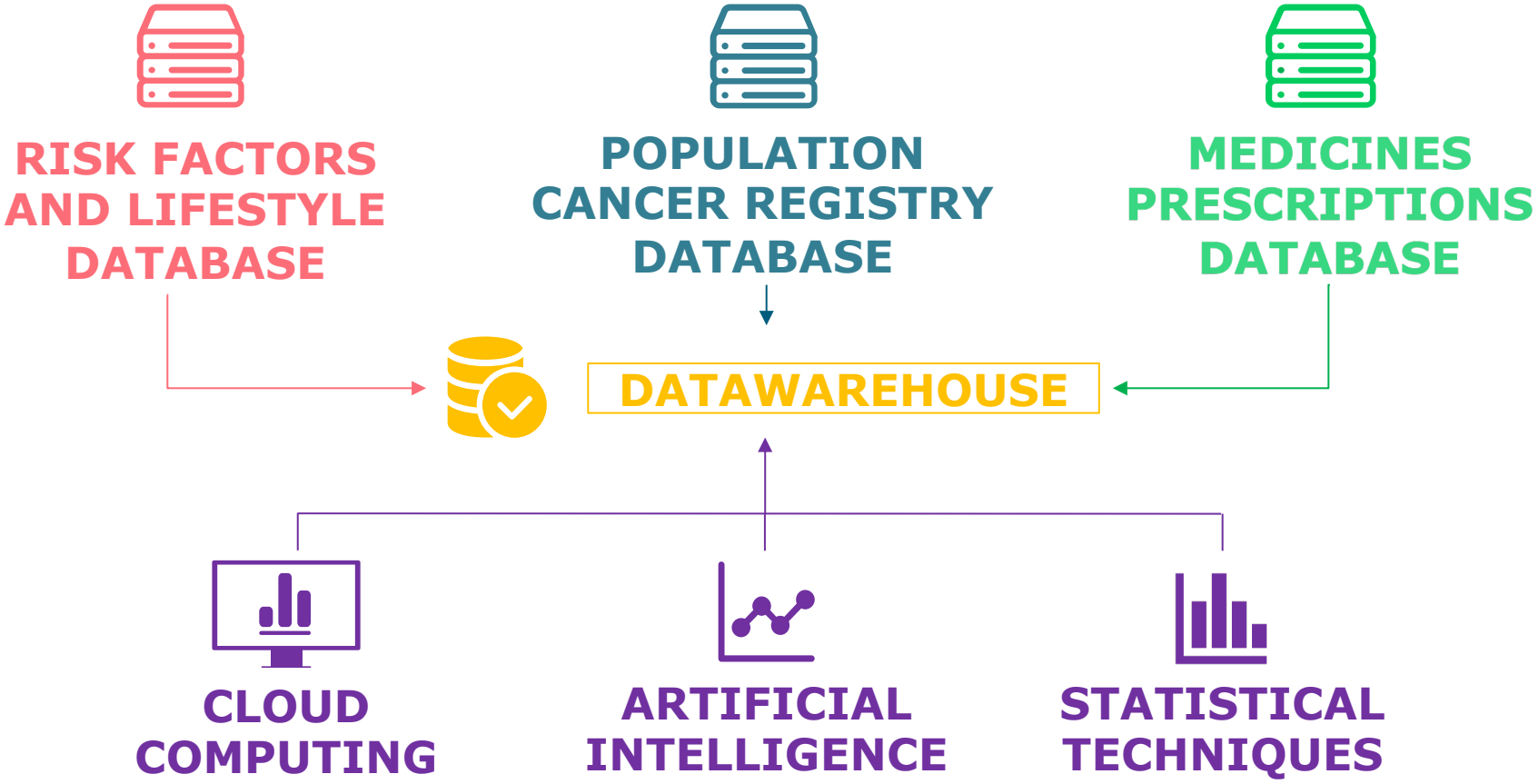
Motivation


RISK FACTORS AND LIFESTYLE DATABASE
POPULATION CANCER REGISTRY DATABASE
MEDICINES PRESCRIPTIONS DATABASE
DATAWAREHOUSE
CLOUD COMPUTING

Motivation

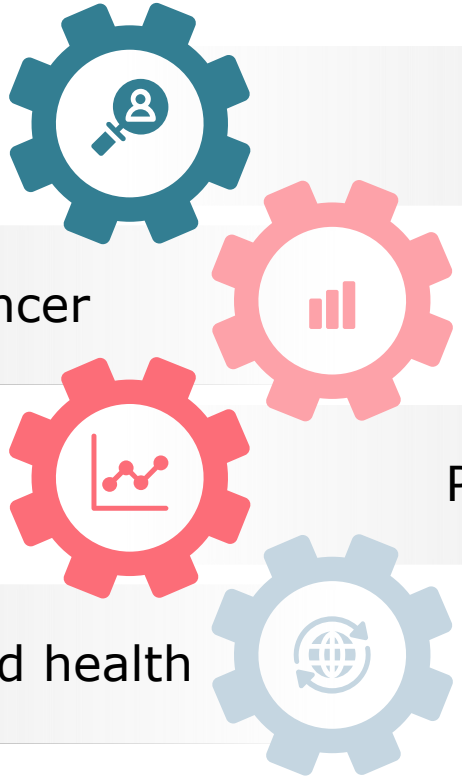

RISK FACTORS AND LIFESTYLE DATABASE
POPULATION CANCER REGISTRY DATABASE
MEDICINES PRESCRIPTIONS DATABASE
DATAWAREHOUSE
CLOUD COMPUTING
Challenge 3: Data Usage

Motivation


RISK FACTORS AND LIFESTYLE DATABASE
POPULATION CANCER REGISTRY DATABASE
MEDICINES PRESCRIPTIONS DATABASE
DATAWAREHOUSE
CLOUD COMPUTING
ARTIFICIAL INTELLIGENCE
STATISTICAL TECHNIQUES

Introduction

Population-based Cancer Registry (PBCR)



Cancer incidence
Determining cancer trends among various populations

Control
Planning and evaluating cancer control efforts

Health resources
Prioritizing health resource allocations

Research
Clinical, epidemiological and health service research

Introduction
PBCR in Lleida

The registry was established in 2017 to retrospectively register new cases from 2012

This system allows to study the characteristic of Lleida. This population differs from other regions

The system is registering 90% of cases

Introduction
Risk factors

Smoking

Overweight

Heavy alcohol use

Certain infections

Exposure to chemical

Family history

Introduction
Risk factors

Smoking ✓

Overweight ✓

Heavy alcohol use ✓

Certain infections ✗

Exposure to chemical ✗

Family history ✗

Introduction
Pharmacology medicines

· Previous studies demonstrated the relationship between some medicines and cancer[1].

· There is also prior evidence for a protective effect of some drugs. One such case is aspirin[2].

· Aspirin is prescribed for preventing recurrent cardiovascular events and for relieving symptoms of rheumatoid arthritis[3].

[1]Friedman GD *et al*. Screening pharmaceuticals for possible carcinogenic effects: Initial positive results for drugs not previously screened. Cancer Causes Control. 2009;20(10):1821–35.
[2]Rothwell P.M. *et al*. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomized trials. The Lancet. 2010; 376 (9754); 1714-1750.
[3]Bibins-Domingo. *et al*. Aspirin Use for the Primary Prevention of Cardiovascular Disease and Colorectal Cancer: U.S. Preventive Services Task Force Recommendation Statement. Annals of Internal Medicine. 2016.

Introduction
Artificial Intelligence

Artificial Intelligence (AI)

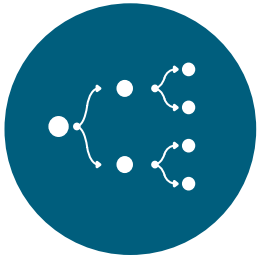
Digital computer's ability to understand and perform tasks associated with intelligent beings.

AI applications

1. Natural Language Process
2. Computer vision
3. Robotics
4. **Machine Learning**

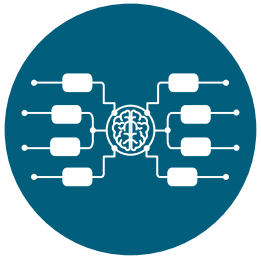
Introduction
Machine Learning

Machine Learning



Supervised learning

It involves training a model to learn a function that maps input data to the correct output labels



Unsupervised learning

Type of machine learning where the model is not given any labelled training data and is instead asked to learn patterns

Hypotheses

Cloud applications enable monitoring and analysis of cancer data for public health surveillance.

Machine learning models are capable of accurately associating cancer risk based on previous risk factors and reveal previously unknown patterns and associations.

DATAWAREHOUSE

There is a correlation between geographical area of exposure and the risk of developing specific types of cancer.

Risk factors such as smoking or heavy alcohol drinking are associated with secondary primary cancer risk.

Potential benefits of aspirin for cancer prevention. Especially for colorectal cancer prevention.

Objectives

1 3 4 5 6 To extract, integrate and assess external databases such as lifestyle and medicines prescriptions.

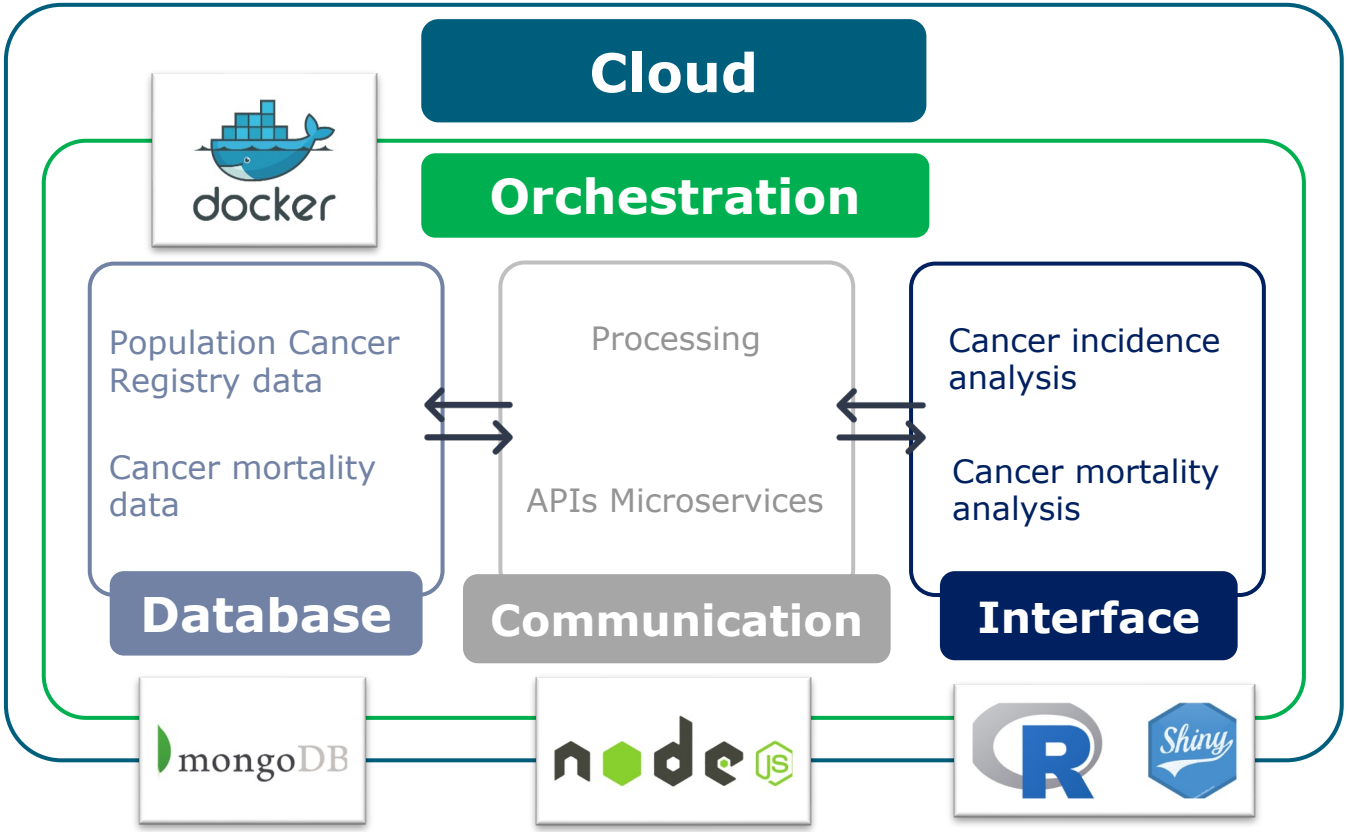
1 To develop a cloud platform to analyze cancer incidence.

2 3 To search associations between geographical area, smoking and socio-demographic information by unsupervised learning algorithms.

4 To analyze associations between overweight, smoking and heavy alcohol use and Secondary Primary Cancer.

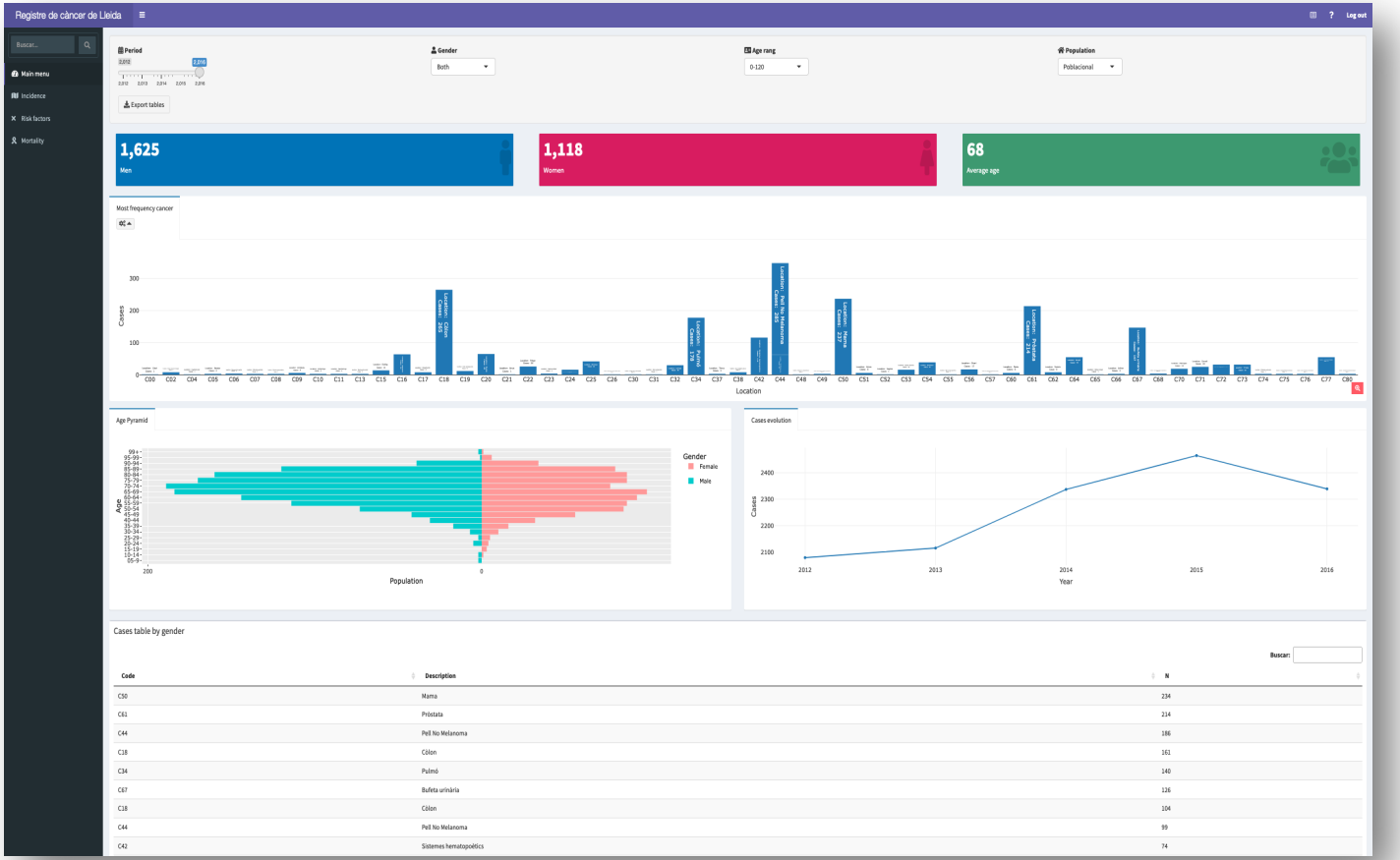
5 6 To analyze the association between aspirin on some cancers.

Paper I
Context

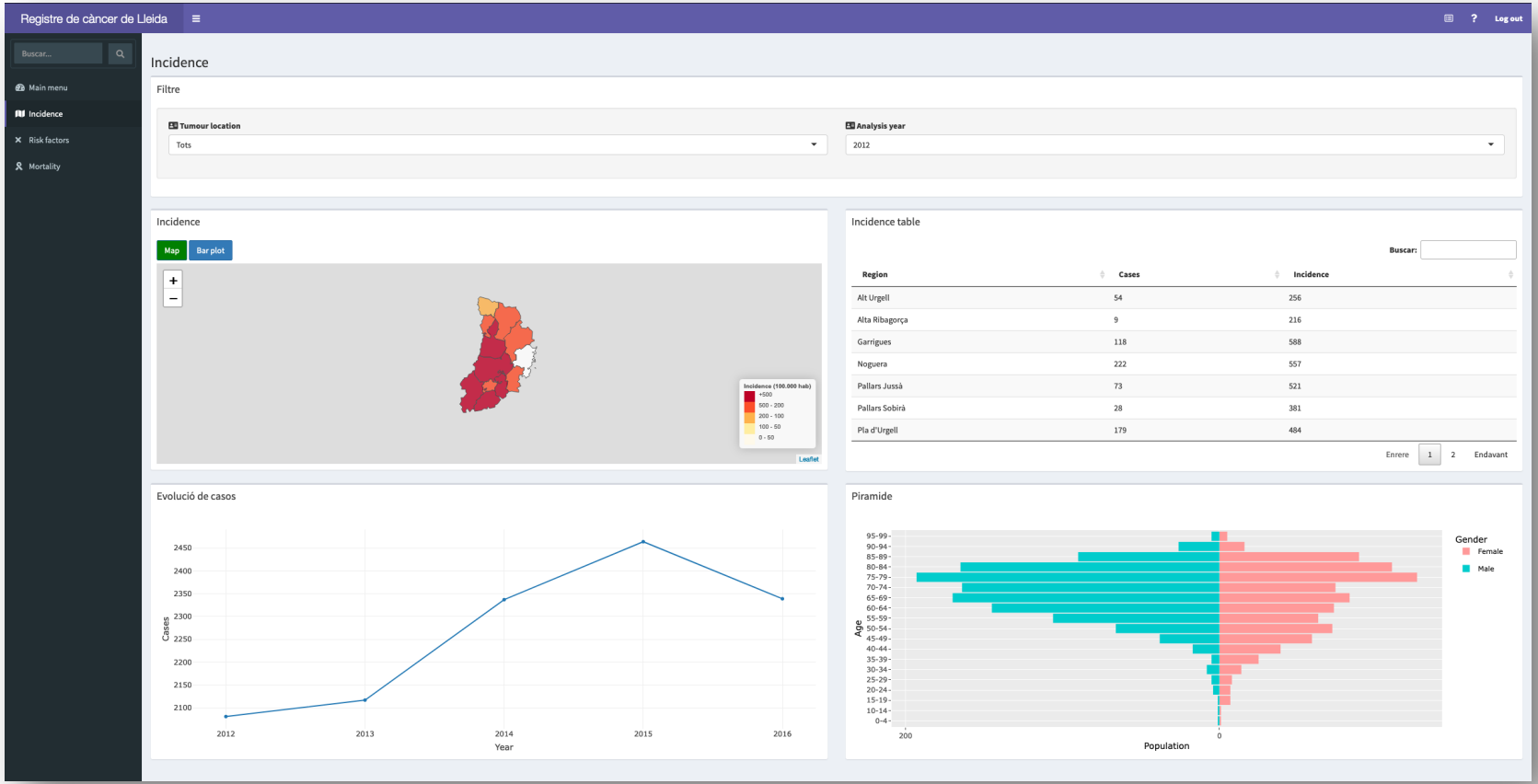


Paper I

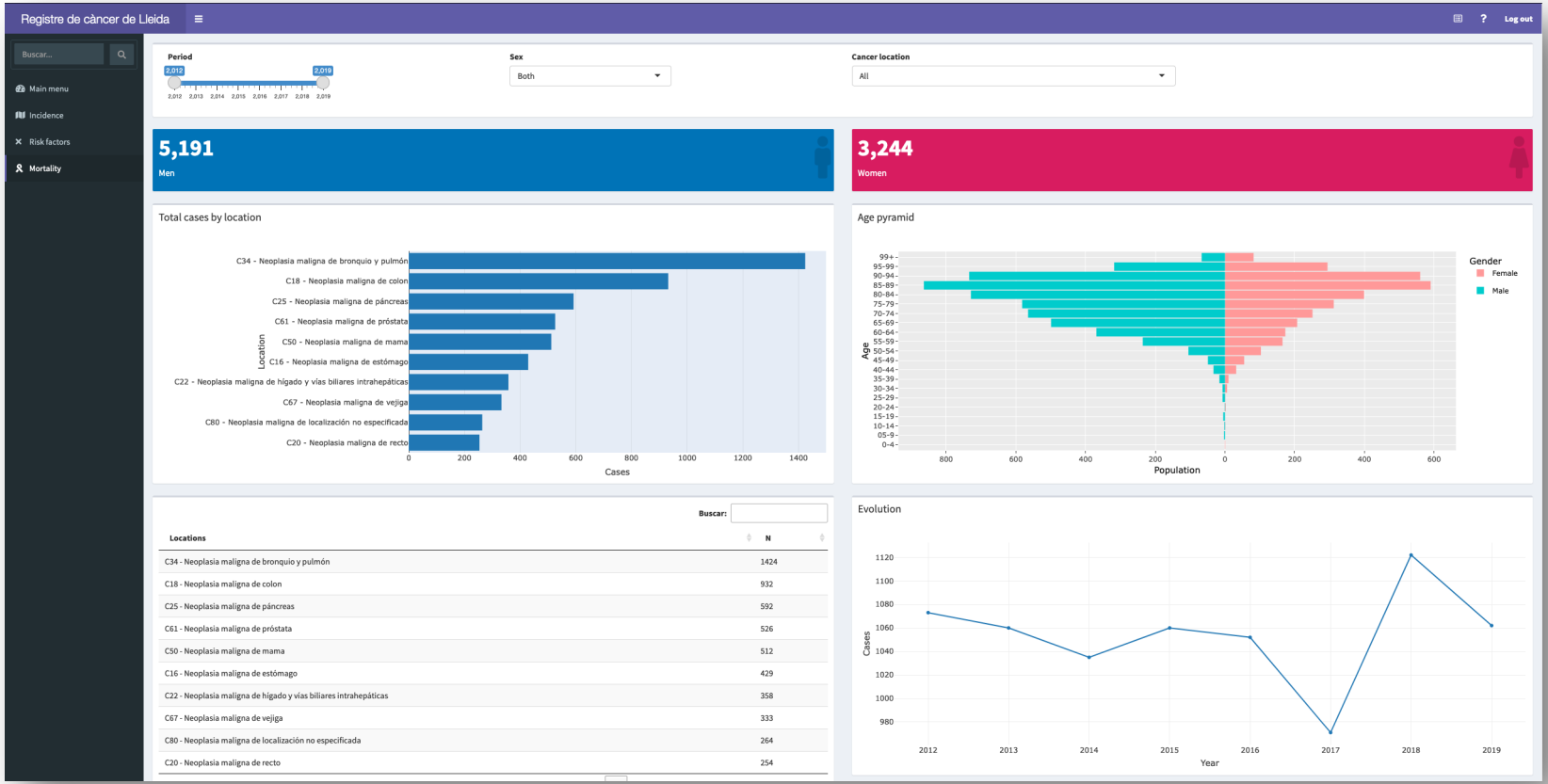
Cancer incidence view



Paper I
Cancer incidence view

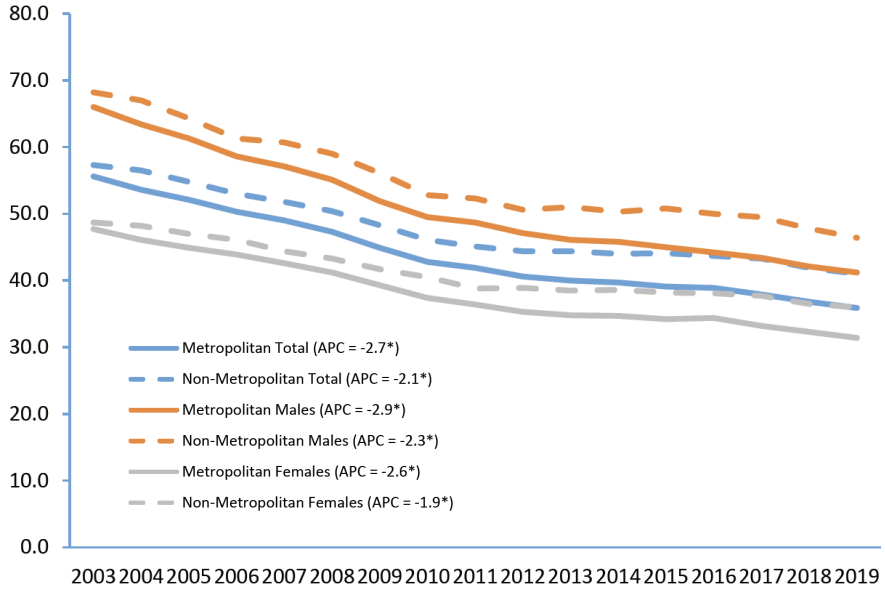


Paper I
Cancer mortality



Paper II
Context

- Rural and urban areas differ in cancer incidence rates[1]
- In Lleida, approximately half of the population lives in rural areas
- Multiple Correspondence Analysis as a technique to explore associations between cancer incidence and socio-demographic information.



1 Centers for Disease Control and Prevention. *Colorectal Cancer Incidence, United States—2003–2019.* USCS Data Brief, no. 33. Atlanta, GA: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2023.

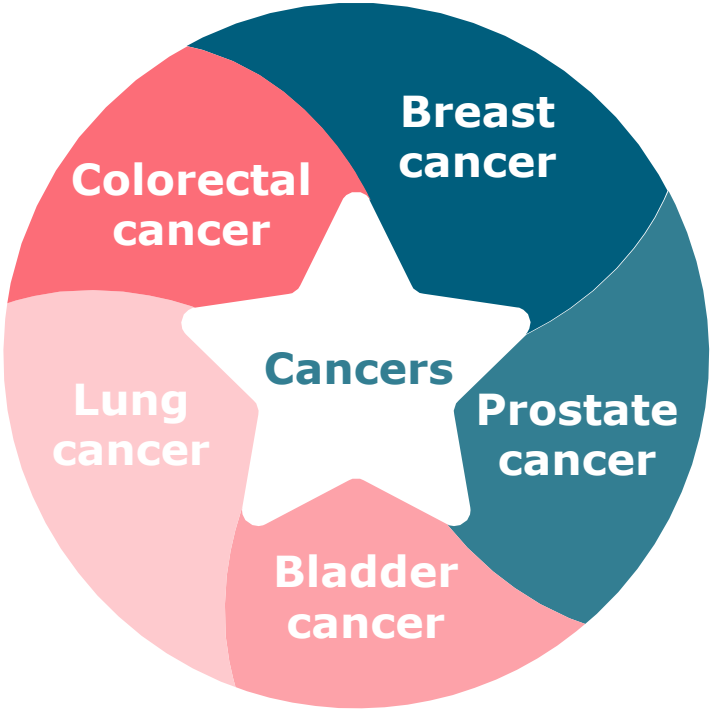
Paper II
Methodology

Multiple Correspondence Analysis (MCA)

1. To explore and visualize information contained on individuals described by categorical variables

2. The **contribution** enables us to consider how much influence a category has in determining to the entire set of the active category.

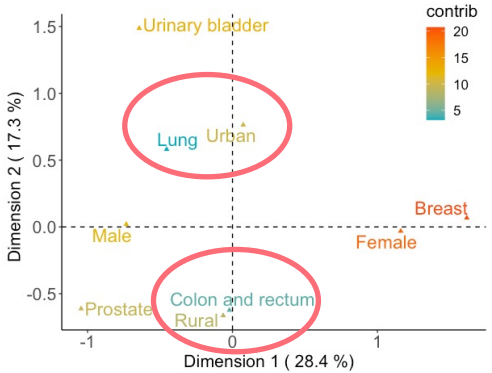
3. To evaluate the relationships between population, age and gender for each cancer.



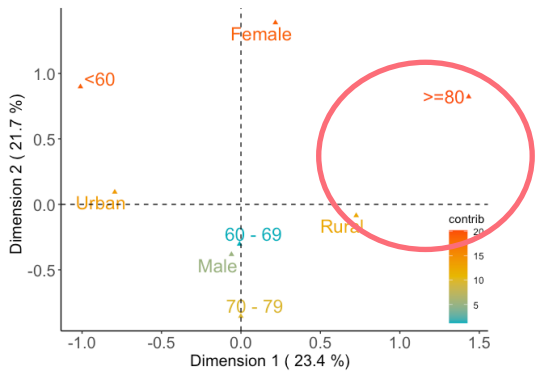
Salut/GSS Gestió de Serveis Sanitaris

Paper II
MCA Associations

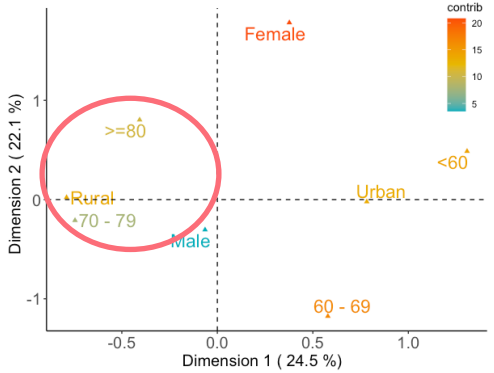
Universitat de Lleida



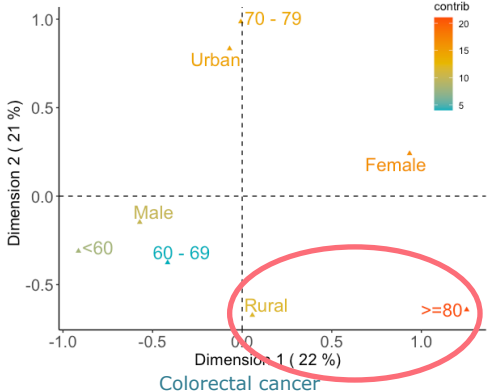
All cancers



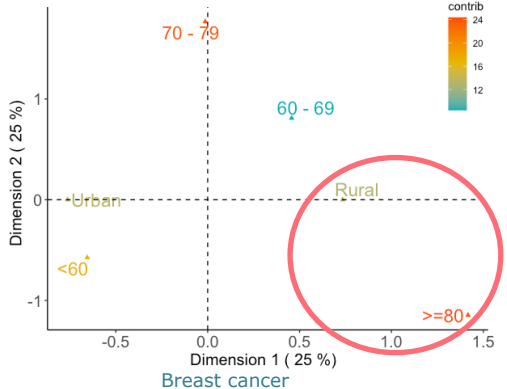
Lung cancer



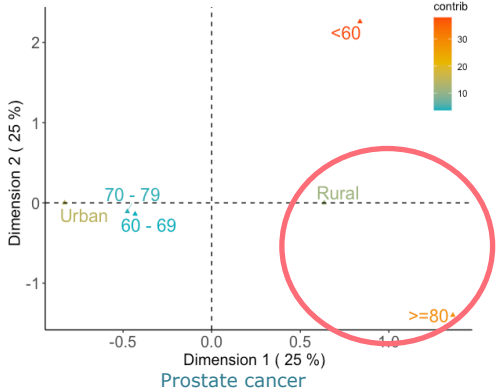
Bladder cancer



Colorectal cancer



Breast cancer



Prostate cancer

Paper III
Context

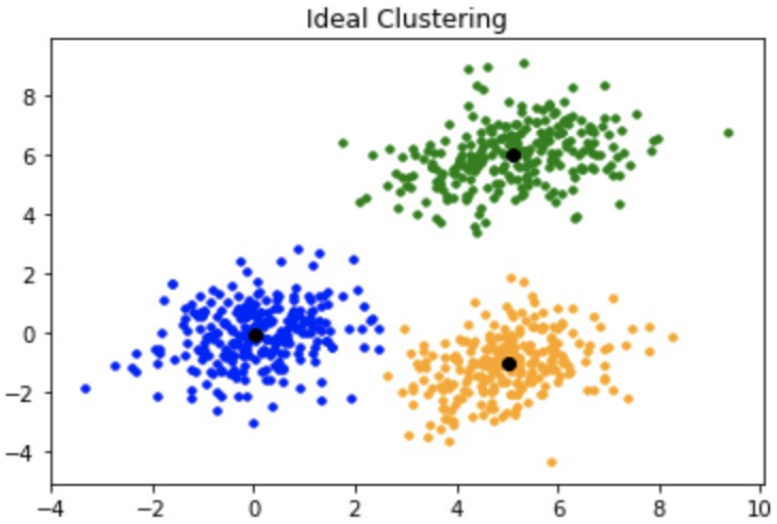
Colorectal cancer

- Colorectal cancer is the highest incidence cancer in Lleida region for both genders

- Overweight, smoking or demographic information can be associated with the risk of colorectal cancer[1]

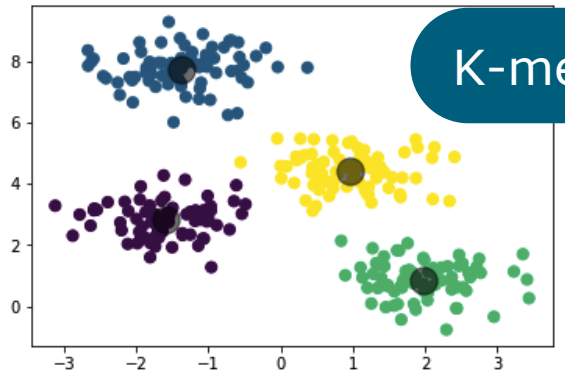
Non-supervised algorithms

MCA to detect relations among large datasets and K-means to identify cluster of patients



[1]Safiri *et al*. The global, regional, and national burden of colorectal cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet Gastroenterology and Hepatology. 2019; 4(12).

Paper III
MCA and K-means



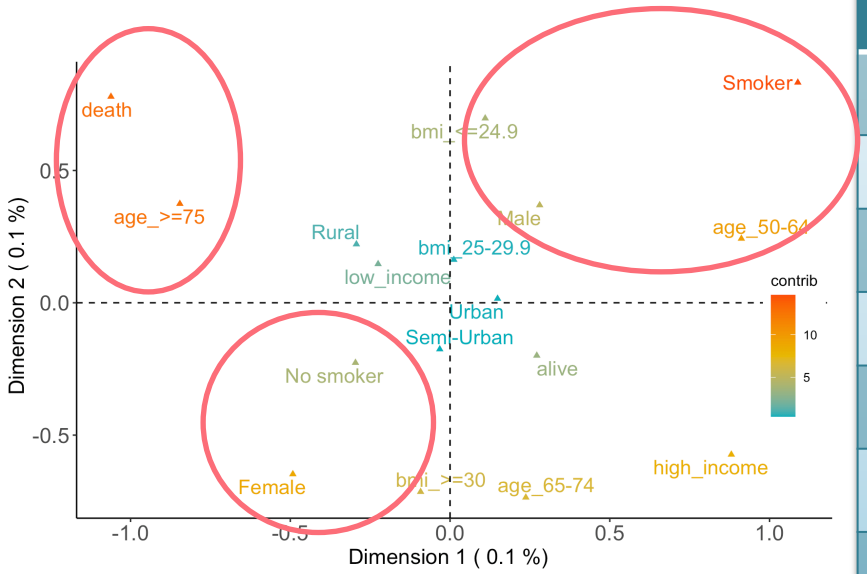
K-means

- Un-supervised learning algorithm used in data-mining and pattern recognition.
- The algorithm partitions the data set into K pre-defined distinct non-overlapping clusters where each data point belongs to only one group.

MCA

Combining MCA and K-means to explore association and patterns among large dataset. In this case, lifestyle and colorectal cancer

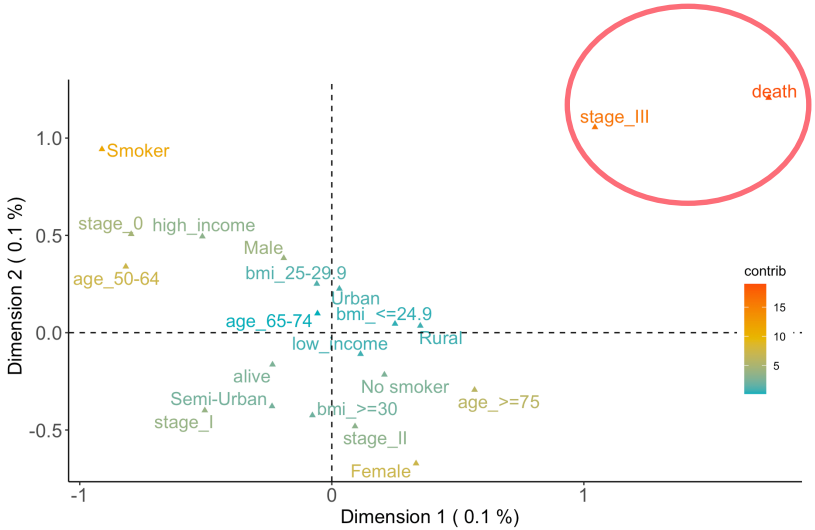
Paper III
Results



Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Urban***	Rural*	**Semi-urban****	Urban***	Semi-urban**
Age >75	Age 50-64	**Age >75**	Age 65-74	Age 65-74
Low income	High income	**Low income**	Low income	Low income
Male	Female	**Male**	Male	Female
Non-smoker	Non-smoker	**Non-smoker**	Smoker	Non-smoker
Overweight	Normal weight	**Obesity**	Normal weight	Overweight
Alive	Alive	**Death**	Alive	Alive

*< 2,000 inhabitants: **rural**
>2,000 and <10,000 inhabitants: **semi-urban
*** > 10,000 inhabitants: **urban**

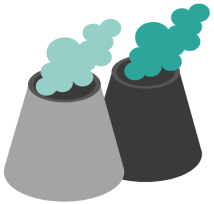
Paper III
Results



Cluster1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Urban***	Semi-urban**	Semi-urban**	Urban***	**Semi-urban****
Age 65-74	Age >75	Age 50-64	Age 65-74	**Age >75**
High income	Low income	Low income	Low income	**Low income**
Male	Male	Male	Male	**Male**
Non-smoker	Non-smoker	Non-smoker	Non-smoker	**Non-smoker**
Obesity	Obesity	Overweight	Obesity	**Overweight**
Alive	Alive	Alive	Alive	**Death**
Stage II	Stage II	Stage 0	Stage III	**Stage III**

*< 2,000 inhabitants: **rural**
>2,000 and <10,000 inhabitants: **semi-urban
*** > 10,000 inhabitants: **urban**

Paper IV
Context



1. Cancer survival trends are generally increasing. One medical consequence is an increased of subsequent diagnosis with another cancer.

2. Such risk factors as obesity, smoking or heavy alcohol use could be determinant in developing a subsequent primary cancer (SPC).

3. To analyze the association between smoking and heavy drinking and the risk of SPC in Lleida.

Paper IV
Methodology

Study population

· Cancer cases diagnosed between 2012 and 2016

· Those first primary cancer diagnosed in 2016 was excluded

· Patients aged 50 or more

Risk factors

· **Body Mass Index (BMI):** 18.5–24.9 normal weight, 25-29.9 overweight and >30 obese.

· **Smoking:** at least 5 years previous cancer diagnosis

· **Heavy drinking:** >40 grams/day in men and >24 grams/day in women for 1 or more years

Statistics

· Cox proportional hazard model

· Hazard ratios and 95% Confidence Intervals

Person-years at risk

- Date of first cancer diagnoses
- Date of SPC diagnoses or date of death
- 31th December 2016

Salut/ Gestió de Serveis Sanitaris

Paper IV
Results

Universitat de Lleida

	Total		SPCs[a]			
	(py)	%	n	% (n/py) * 100	Crude HR	95% CI
Gender						
Female	4,349	42.6	69	1.6	Ref. group	-
Male	6,208	58.4	165	2.7	1.7	1.3 - 2.2
Age						
50-59	2,195	20.8	30	1.4	Ref. group	-
60-69	3,415	32.3	79	2.3	1.7	1.1 - 2.6
70-79	3,416	32.4	102	3.0	2.1	1.5 - 3.3
80-	1,531	14.5	24	1.6	1.1	0.6 - 2.0
Cancer type						
Type 3	323	3.1	4	1.2	Ref. group	-
Type 2	2,584	23.4	29	1.1	0.9	0.3 - 2.6
Type 1	2,119	21.2	66	3.0	2.4	0.9 - 6.7
Type 0	5,524	52.3	135	2.4	2.0	0.7 - 5.4

	Total		SPCs[a]			
	(py)	%	n	% (n/py) * 100	Crude HR	95% CI
Body mass index						
Normal weight	2,781	26.3	63	2.3	Ref. group	-
Overweight	4,616	43.7	108	2.3	1.0	0.7 - 1.4
Obese	3,160	29.9	63	2.0	0.9	0.6 - 1.3
Smoking						
No	7,462	70.7	146	2.0	Ref. group	-
Yes	3,095	29.3	88	2.8	1.5	1.1 - 1.9
Diabetes						
No	10,162	96.3	224	2.2	Ref. group	-
Yes	395	3.7	10	2.5	1.2	0.6 - 2.2
Heavy drinking						
No	10,404	98.6	225	2.2	Ref. group	-
Yes	153	1.4	9	5.9	2.7	1.4 - 5.4

Paper IV
Results

	Hazard ratio	95% CI [a]
Female	1.0	Ref. group
Male	1.4	1.1 - 1.9
Age 50-59	1.0	Ref. group
Age 60-69	1.6	1.1 - 2.5
Age 70-79	2.2	1.5 - 3.4
Age 80-	1.2	0.7 - 2.0
Smoking	1.3	1.0 - 1.7
Heavy drinking	2.4	1.3 - 4.8

Paper IV
Results

	Males		Females	
	Adjusted HR[a]	95% CI[b]	Adjusted HR	95% CI
Age 50-59	1.0	Ref. group	1.0	Ref. group
Age 60-69	1.6	1.0 - 2.7	1.8	0.8 - 3.9
Age 70-79	2.0	1.3 - 3.4	2.6	1.2 - 5.7
Age 80-	0.5	0.2 - 1.2	3.0	1.3 - 7.1
Diabetes	1.4	0.7 - 2.8	-	-
Smoking	1.2	1.0 - 1.6	1.8	0.8 - 3.7
Heavy drinking	2.3	1.1 - 4.7	3.2	0.4 - 23.6

Paper V
Context

- Approximately 1.8 million new colorectal cancer cases were diagnosed worldwide.
- Between 350 and 400 new cases in Lleida were diagnosed each year.
- Some studies estimate that 30-50 % of colorectal cases could be avoided[1].
- Aspirin has long been known to prevent cardiovascular and cerebrovascular.

1GBD 2019 Cancer Risk Factors Collaborators. The global burden of cancer attributable to risk factors, 2010-19: a systematic analysis for the Global Burden of Disease Study 2019. Lancet. 2022 Aug 20;400(10352):563-591.

Paper V
Methodology

Study population and aspirin use

· Colorectal cancer cases diagnosed between 2012 and 2016

· Inhabitants from Lleida region aged > 50 years

· Aspirin exposition during, at least, 5 years.

· Aspirin use > 75 mg/daily

Risk factors

· **Body Mass Index (BMI):** 18.5–24.9 normal weight, 25-29.9 overweight and >30 obese.

· **Smoking:** at least 5 years previous cancer diagnosis

· **Heavy drinking:** >40 grams/day in men and >24 grams/day in women for 1 or more years

Statistics

· Cox proportional hazard model

· Hazard ratios and 95% Confidence Intervals

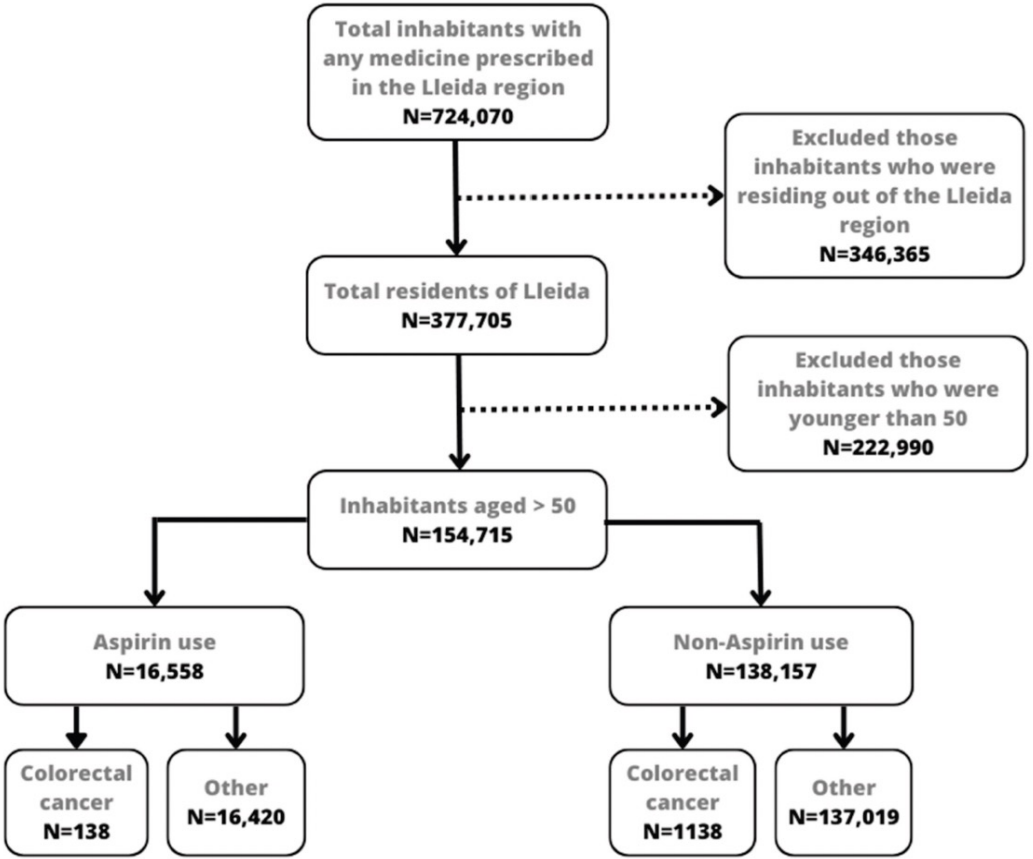
Person-years at risk

1st January 2007

Date of CRC diagnose or date of death

31th December 2016

Paper V
Workflow



	Total		n	% (n/p-y)	Crude HR [1]	95% CI
	Person-Year (p-y)	%				
Gender						
Female	639,455	53.1	485	0.8	1.0	Ref. group
Male	563,716	46.9	791	1.4	1.9	1.6–2.1
Age						
(50–59)	393,275	32.7	303	0.8	1.0	Ref. group
(60–69)	297,538	24.7	426	1.4	1.8	1.6–2.1
(70–79)	215,272	17.9	349	1.6	2.0	1.9–2.6
(80–89)	147,817	12.3	184	1.2	1.6	1.3–1.9
(90–)	149,269	12.4	14	0.1	0.1	0.1–0.2
Aspirin						
Non-use	1,068,470	88.8	1138	1.2	1.0	Ref. group
Use	134,701	11.2	138	1.0	0.9	0.8–1.1
Body mass index						
Normal weight	350.994	29.2	169	0.5	1.0	Ref. Group
Overweight	404.905	33.7	504	1.2	2.5	2.2–3.1
Obesity	447,272	37.2	603	1.3	2.7	2.3–3.3
Risky drinking						
No	1,177,736	97.9	1220	1.0	1.0	Ref. Group
Yes	25,435	2.1	56	2.2	2.1	1.6–2.7
Smoking						
No	1,094,891	91.0	1056	1.0	1.0	Ref. Group
Yes	108,280	9.0	220	2.0	2.0	1.8–2.4

[1] Hazard ratio.

	Adjusted Hazard Ratio (aHR); 95% CI [1]	*p*-Value
Female	-	Ref. Group
Male	1.8 (1.6–2.1)	<0.001
(50–59)	-	Ref. Group
(60–69)	1.8 (1.6–2.1)	<0.001
(70–79)	2.3 (1.9–2.7)	<0.001
(80–89)	2.2 (1.8–2.6)	0.007
(90-)	0.2 (0.1–0.3)	<0.001
Aspirin use	0.7 (0.6–0.8)	0.006
Normal weight	-	Ref. Group
Overweight	1.4 (1.2–1.7)	<0.001
Obesity	1.5 (1.3–1.8)	<0.001
Risky drinking	1.6 (1.2–2.0)	0.006
Smoking	1.4 (1.3–1.7)	<0.001

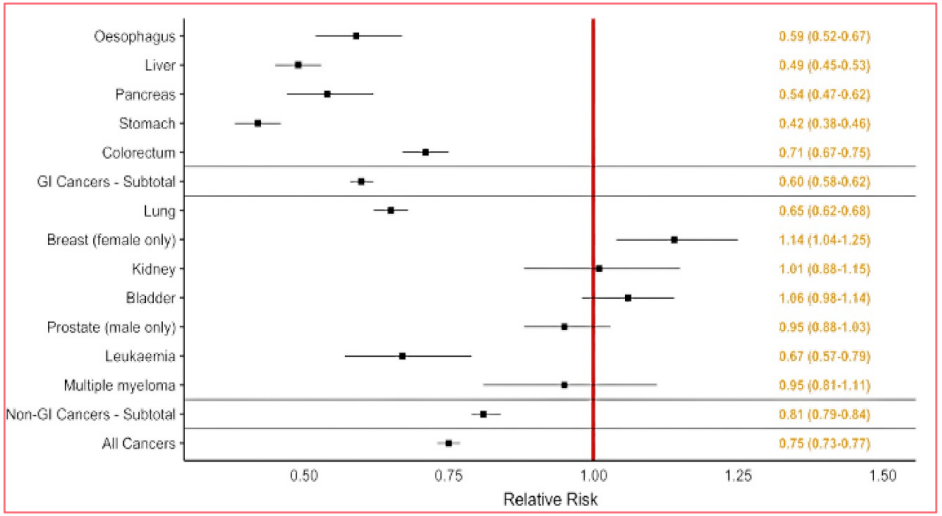
[1] Confidence interval.

	Men		Women	
	Adjusted Hazard Ratio (aHR); 95% CI [1]	*p*-Value	Adjusted Hazard Ratio (aHR); 95% CI [1]	*p*-Value
(50–59)	-	Ref. Group	-	Ref. Group [2]
(60–69)	1.9 (1.7–2.3)	<0.001	1.7 (1.3–2.2)	<0.001
(70–79)	2.3 (1.9–2.8)	<0.001	2.3 (1.7–2.9)	<0.001
(80–89)	2.1 (1.6–2.7)	<0.001	2.2 (1.7–3.0)	<0.001
(90–)	0.2 (0.1–0.4)	<0.001	0.2 (0.1–0.5)	<0.001
Aspirin use	0.7 (0.6–0.9)	0.005	0.6 (0.4–0.8)	0.005
Normal weight	-	Ref. Group [2]	-	Ref. Group [2]
Overweight	1.5 (1.2–2.0)	<0.001	1.2 (0.9–1.6)	0.1
Obesity	1.6 (1.3–2.1)	<0.001	1.4 (1.2–1.9)	0.004
Risky drinking	1.6 (1.2–2.1)	0.001	1.2 (0.4–3.7)	0.7
Smoking	1.5 (1.3–1.7)	<0.001	1.4 (0.9–2.2)	0.1

[1] Confidence interval. [2] Reference group.

Paper VI
Context

- Nowadays, cancers such as pancreatic or lung are the highest mortal.
- Some studies suggest that aspirin decreased the risk of some types of cancer[1].
- Exist controversies against the protective effect of aspirin on some cancers.



[1]Tsoi K *et al*. Long-tearm use of low-dose aspirin for cancer prevention: A 10-year population cohort study in Hong Kong. International Journal of Cancer. 2019; 145 (267).

Paper VI
Methodology

Study population and aspirin use

· **Colorectal, oesophageal, stomach, liver, pancreatic, lung, bladder, lymphoma, leukaemia, prostate and breast** cancer cases diagnosed between 2012 and 2016.

· Inhabitants from Lleida region aged > 50 years

· Aspirin exposition during, at least, 5 years.

· Aspirin use > 75 mg/daily

Risk factors

· **Body Mass Index (BMI):** 18.5–24.9 normal weight, 25-29.9 overweight and >30 obese.

· **Smoking:** at least 5 years previous cancer diagnosis

· **Heavy drinking:** >40 grams/day in men and >24 grams/day in women for 1 or more years

Statistics

· Cox proportional hazard model

· Hazard ratios and 95% CI

· Relative risks to estimate aspirin association

Person-years at risk

1st January 2007

Date of cancer diagnose or date of death

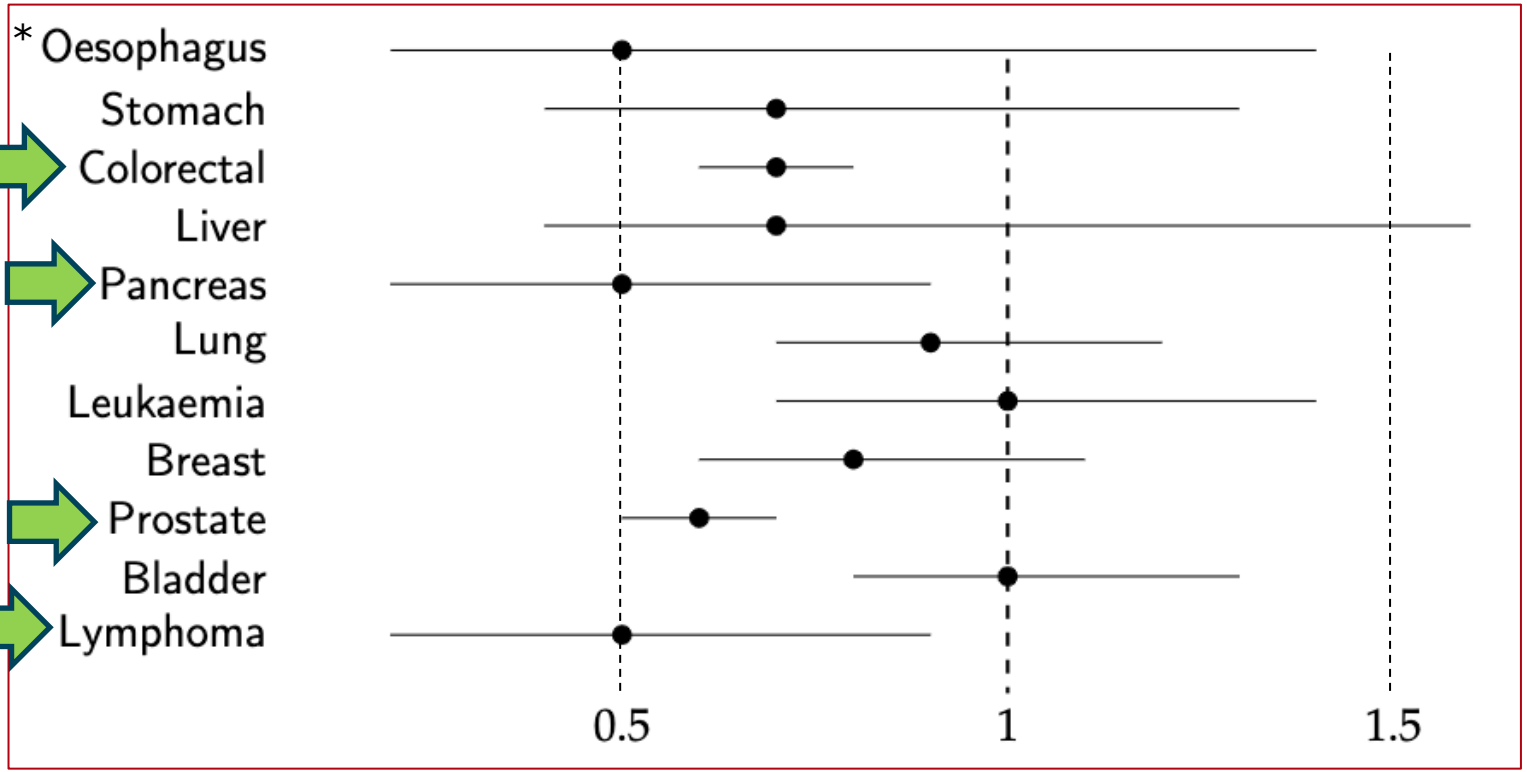
31th December 2016

Paper VI
Results

Relative risk

	Cancer Incidence			
Cancers	Total n= 154,715 (%)	Aspirin group n = 16,598 (%)	Non-aspirin group n = 138,117 (%)	Relative risk (95% CI [a])
Oesophagus	32 (0.02)	3 (0.02)	29 (0.03)	0.86 (0.26 - 2.82)
Stomach	135 (0.09)	18 (0.11)	117 (0.08)	1.28 (0.78 – 2.1)
Colorectal	1,276 (0.82)	138 (0.83)	1,138 (0.82)	1.01 (0.8 – 1.2)
Liver	60 (0.04)	7 (0.04)	53 (0.04)	1.09 (0.5 – 2.42)
Pancreas	97 (0.06)	8 (0.05)	89 (0.06)	0.63 (0.31 – 1.31)
Lung and bronchus	426 (0.28)	61 (0.37)	365 (0.26)	1.39 (1.06 – 1.82)
Leukaemia	328 (0.21)	56 (0.34)	272 (0.20)	1.71 (1.28 – 2.28)
Breast (female only)	737 (0.48)	56 (0.34)	681 (0.49)	0.77 (0.58 – 1.02)
Prostate (male only)	916 (0.59)	113 (0.78)	803 (0.58)	1.0 (0.82 – 1.23)
Bladder	567 (0.37)	100 (0.60)	467 (0.34)	1.78 (1.43 – 2.21)
Lymphoma	131 (0.08)	8 (0.05)	123 (0.09)	0.54 (0.26 – 1.11)

Paper VI
Results



*A Cox regression was calculated for each cancer

Limitations

- MCA and K-means are used for exploring large databases without pre-existing hypotheses. However, in cancer epidemiology is typically hypotheses-driven.
- The shorter follow-up did not allow for more Secondary Primary Cancers to be observed. A longer observation period would improve the quality of the sample.
- Aspirin can be purchased directly from pharmacies without a doctor's prescription. The results are consistent with previous literature.
- Risk factors data comes from clinical records and may be underreported.

Conclusions

- **The cloud applications offer accessibility and variability that are crucial to add qualified knowledge from data to medical experts.**

Conclusions

- The cloud applications offer accessibility and variability that are crucial to add qualified knowledge from data to medical experts.

- **Multiple Correspondence Analysis (MCA) is an ideal algorithm to analyze associations between cancer and some factors.**

-

Conclusions

- The cloud applications offer accessibility and variability that are crucial to add qualified knowledge from data to medical experts.

- Multiple Correspondence Analysis (MCA) is an ideal algorithm to analyze associations between cancer and some factors.

- **MCA and K-means is a perfect alliance for detecting patterns and associations of cancer from information about patients and risk factors.**

Conclusions

- **The significance of PBCR for monitoring and analyzing cancer and their potential research when integrated with other databases, such as risk factors or medication exposures.**

Conclusions

- The significance of PBCR for monitoring and analyzing cancer and their potential research when integrated with other databases, such as risk factors or medication exposures.

- **Smoking and heavy alcohol use increase the risk of SPC during the first follow-up years, especially among men.**

-

Conclusions

- The significance of PBCR for monitoring and analyzing cancer and their potential research when integrated with other databases, such as risk factors or medication exposures.

- Smoking and heavy alcohol use increase the risk of SPC during the first follow-up years, especially among men.

- **Aspirin use decreases the risk of colorectal cancer and overweight, smoking and heavy drinking increase this risk. Aspirin also decreases the risk of pancreatic, prostate cancer and lymphoma.**

Future work

Data warehouse

Add external databases

Future work

Data warehouse

Add external databases

Data Analysis

Extend similar studies with other types of cancer

Future work

Data warehouse

Add external databases

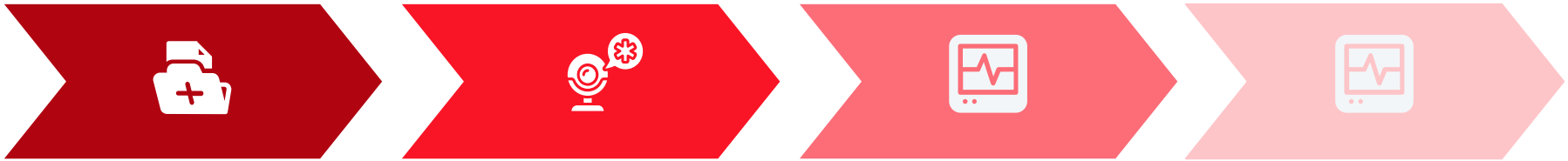
Data Analysis

Extend similar studies with other types of cancer

Cancer prediction

Train and deploy supervised algorithms to predict cancer

Future work



Data warehouse

Add external databases

Data Analysis

Extend similar studies with other types of cancer

Cancer prediction

Train and deploy supervised algorithms to predict cancer

Decision Support System

Integration to other population cancer registries

Future work

In eHealth, quality is more important than quantity. This drives the innovation in the healthcare industry.

Salut/GSS Gestió de Serveis Sanitaris

Universitat de Lleida

Machine Learning Approaches for Comprehensive Analysis of Population Cancer Registry Data

Ph.D Defense

Authored by
Dídac Florensa Cazorla

Supervised by
Pere Godoy Garcia
Francesc Solsona Tehàs
Jordi Mateo Fornés

Industrial supervision by
Miquel Mesas Julió

19th April 2023