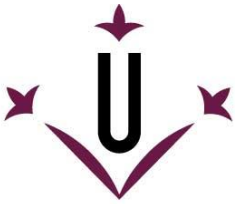
Universitat de Lleida

Machine Learning Approaches for Comprehensive Analysis of Population Cancer Registry Data

Dídac Florensa Cazorla



Universitat de Lleida

TESI DOCTORAL

Machine Learning Approaches for Comprehensive Analysis of Population Cancer Registry Data

Dídac Florensa Cazorla

Memòria presentada per optar al grau de Doctor per la Universitat de Lleida
Programa de Doctorat en Enginyeria Informàtica i Tecnologies de la Informació.

Director/a
Pere Godoy Garcia
Francesc Solsona Tehàs
Jordi Mateo Fornés

Supervisor/a industrial
Miquel Mesas Julió

Tutor/a
Francesc Solsona Tehàs

2023

It always seems impossible until it's done.

Nelson Mandela

To you as a reader.

ABSTRACT

Background: Population-based cancer registries are crucial for controlling and studying cancer incidence, mortality, and survival. These systems focus on collecting new cancer cases and analyzing their impact in a specific region. In addition, exploring external information sources to complement the data registry allows for the identification of search patterns and correlations in each specific region. This thesis focuses on integrating some databases, such as risk factors and prescription medicines, cloud computing and artificial intelligence (AI). This is a term that has been present in the business and social sectors over the last few years. The capacity of artificial intelligence to learn, simulating the human brain, has permitted the automation of the process and decreased the time required. This technique is characterised by an algorithm training process to then decide what the machine has learned. Artificial intelligence algorithms have opened new ways for the analysis, detection, prediction and pattern search of cancer which are explored in this thesis.

Methods: Cloud computing and decision support system were used to implement a web-platform to show the cancer incidence in a specific region (Lleida). Non-supervised machine learning algorithms were used as a tool for detecting patterns of cancer by the different lifestyles of cancer patients. The *Multiple Correspondence Analysis* algorithm was trained to detect cancer patterns for the most frequent cancers. *K-means* was implemented for cancer detection among cases of colorectal cancer. Next, epidemiological studies were employed to ensure the validity of the external databases by assessing the association of some risk factors with the occurrence of a second cancer. Additionally, the protective effect of aspirin against certain types of cancer was analyzed, while accounting for relevant risk factors.

Results: This thesis explores cloud computing methods and artificial intelligence algorithms for pattern detection. It also explores in depth how some risk factors increase the risk of cancer and other functionalities, such as estimates the risk of developing secondary primary cancer. Finally, this thesis explores how to export this knowledge to society through cutting-edge technology. The main outcomes of this thesis highlight a cloud application to assist population cancer registries in analysing cancer incidence and mortality and the use of machine learning algorithms to detect patterns and associations of the factors that may increase the risk of cancer. In this exploration, we have discovered a strong association

between colorectal cancer and individuals living in rural populations. However, lung cancer is more common among those living in urban areas. On the other hand, our analysis of the risk of developing secondary primary cancer revealed that certain risk factors, such as smoking and heavy alcohol use, significantly increase the likelihood of developing such cancers. Finally, the results show the protective effect of aspirin against some tumours, taking into account such risk factors as smoking or heavy alcohol use or excess weight.

Conclusions: This thesis integrates risk factor and medication prescription databases and employs a cloud-based decision support system (DSS) that utilizes population-based cancer registry data to assess the current state of cancer in Lleida. We also analyse and implement various machine learning algorithms, especially non-supervised. The outcomes provide solid evidence that these non-supervised algorithms can search for patterns among cancer patients. In addition, they also help to detect possible associations, which is interesting for the health sector. In a health context, this thesis demonstrates an association between smoking and heavy alcohol use with the risk of second primary cancer, especially among men. It also corroborated that aspirin use decreases the risk of some specific cancers, taking risk factors into account. The results obtained in this thesis are an essential seed to continue exploring other methods and algorithms, with a high potential to become a reference in the use of artificial intelligence in the epidemiological cancer sector.

RESUM

Introducció: Els registres poblacions de càncer són crucials pel control i l'estudi de la incidència, la mortalitat i la supervivència de càncer. Aquests sistemes es focalitzen en recollir nous casos de càncer i analitzar el seu impacte en una regió específica. A més, l'exploració de fonts d'informació externes per complementar el registre permet realitzar un pas més i identificar patrons i correlacions en cada regió específica. En aquesta tesi, ens focalitzem en la integració de diferents bases de dades com els factors de risc i les prescripcions de medicaments, els sistemes basats en el núvol i l'ús de la intel·ligència artificial. Aquest últim és un terme que durant els últims anys ha estat més present en diferents sectors empresarials i socials. La capacitat de la intel·ligència artificial en aprendre, simulant la ment humana, ha permès automatitzar processos disminuint el temps emprat. Aquesta tècnica es caracteritza per un procés d'entrenament d'algoritmes perquè a posterior puguin prendre decisions en funció del que han après. Els algoritmes d'intel·ligència artificial han obert noves maneres d'anàlisi, detecció, predicció i cerca de patrons en càncer.

Mètodes: La computació en el núvol i els sistemes d'ajuda a la presa de decisions han estat utilitzats per implementar una plataforma web per tal d'informar de la situació epidemiològica del càncer en una regió específica (Lleida). Algoritmes d'aprenentatge automàtic no supervisats s'han usat com a instrument per la detecció de patrons de càncer a partir dels diferents estils de vida dels pacients. L'algoritme *Multiple Correspondence Analysis* s'ha entrenat per la detecció de patrons en els càncers més incidents. El *K-means* s'ha implementat per tal de detectar patrons en el càncer colorectal. A continuació, es van realitzar estudis epidemiològics per assegurar la validesa de les bases de dades externes mitjançant l'estudi de l'associació d'alguns factors de risc i el desenvolupament d'un segon càncer primari. Per altra banda, s'ha analitzat l'efecte protector de l'aspirina vers alguns càncers tenint en compte certs factors de risc.

Resultats: Aquesta tesi explora mètodes basats en el núvol i algoritmes d'intel·ligència artificial per la detecció de patrons i la predicció en càncer. També explora en profunditat alguns factors de risc que incrementen el risc de càncer i també altres factors que actuen com a protectors, com per exemple, el risc de presentar un segon càncer. Finalment, s'ha explorat com exportar i lliurar aquest coneixement a la societat mitjançant tecnologia puntera. Els principals resultats d'aquesta tesi ressalten una aplicació al núvol per assistir a registres

poblacionals de càncer sobre l'anàlisi de la incidència i la mortalitat d'aquesta malaltia i l'ús d'algoritmes d'aprenentatge automàtic per la detecció de patrons i l'associació de factors que poden incrementar el risc de càncer. En aquest estudi, s'observa que alguns càncers es relacionen més amb la població que viu en àmbits rurals tal com és el colorectal. En canvi, altres com el pulmó es van relacionar amb habitants que viuen en zones urbanes. D'altra banda, l'anàlisi sobre el risc de presentar un segon càncer primari revela que certs factors de risc, com són el consum de tabac i alcohol, incrementen el risc de presentar aquest tipus de càncers. Finalment, els resultats mostren que l'aspirina té un efecte protector vers a alguns càncers tenint en compte factors de risc habituals com el consum de tabac, alcohol o el sobrepès.

Conclusions: En aquesta tesi s'ha integrat la base de dades de factors de risc i de medicaments prescrits i s'ha implementat un sistema de suport a la presa de decisions en el núvol basat en les dades del registre poblacional de càncer. També s'han analitzat i implementat diversos algoritmes d'aprenentatge automàtic, tant supervisats com no supervisats. Els resultats obtinguts proporcionen evidències sòlides que aquests algoritmes no supervisats poden ser utilitzats per la cerca de patrons en pacients en càncer. A part, ajuden a la detecció de possibles associacions, interessants pel sector mèdic, que usant altres mètodes no es podrien detectar. En un context més mèdic, s'ha demostrat que existeix una associació entre el consum de tabac i alcohol i el risc de patir un segon càncer primari, especialment entre els homes. També s'ha corroborat que el consum d'aspirina prolongat actua com a protector vers alguns càncers específics tot i presentar altres factors de risc. Tots els resultats assolits en aquesta tesi són una llavor poderosa per continuar explorant altres mètodes i altres algoritmes, amb un gran potencial per convertir-se en una referència en l'ús de la intel·ligència artificial en aquest àmbit més epidemiològic del càncer.

RESUMEN

Introducción: Los registros poblacionales de cáncer son cruciales para el control y el estudio de la incidencia, la mortalidad y la supervivencia de cáncer. Estos sistemas se focalizan en recoger nuevos casos de cáncer y analizar su impacto en una región específica. Además, la exploración de fuentes de información externas para complementar el registro permite ir un paso más e identificar patrones y correlaciones en cada región específica. En esta tesis nos focalizamos en la integración de varias bases de datos como los factores de riesgo y los medicamentos prescritos, los sistemas basados en la nube y el uso de la inteligencia artificial. Este último es un término que durante los últimos años ha estado presente en diferentes sectores empresariales y sociales. La capacidad de la inteligencia artificial en aprender, simulando la mente humana, ha permitido automatizar procesos disminuyendo el tiempo utilizado. Esta técnica se caracteriza por un proceso de entrenamiento de algoritmos para que después puedan tomar decisiones en función de lo que han aprendido. Los algoritmos de inteligencia artificial han abierto nuevas maneras de análisis, detección, predicción y búsqueda de patrones en cáncer.

Métodos: La computación en la nube y los sistemas de ayudas a la toma de decisiones han estado utilizados para implementar una plataforma web que permite informar de la situación epidemiológica del cáncer en una región específica (Lleida). Algoritmos de aprendizaje automático no supervisados se han utilizado como instrumento para la detección de patrones de cáncer a partir de los diferentes estilos de vida de los pacientes con cáncer. El algoritmo *Multiple Correspondence Analysis* se ha entrenado para la detección de patrones de los cánceres más incidentes. El *K-means* se ha implementado para poder detectar patrones en el cáncer colorrectal. A continuación, se han diseñado estudios epidemiológicos para asegurar la validez de las bases de datos externas a través del estudio de la asociación de algunos factores de riesgo con la aparición de un segundo cáncer primario. Por otro lado, se ha analizado el efector protector de la aspirina frente algunos cánceres, teniendo en cuenta ciertos factores de riesgo.

Resultados: Esta tesis explora métodos basados en la nube y algoritmos de inteligencia artificial para la detección de patrones y la predicción del cáncer. También analiza en profundidad ciertos factores de riesgo que aumentan la posibilidad de padecer cáncer, así como otros factores que actúan como protectores, como por ejemplo, el riesgo de presentar un segundo cáncer. Por último, se explora

cómo exportar y difundir el conocimiento a la sociedad mediante tecnología de vanguardia. Los principales resultados de esta tesis resaltan una aplicación a la nube para asistir a registros poblacionales de cáncer sobre el análisis de la incidencia y la mortalidad de esta malatía y el uso de algoritmos de aprendizaje automático para la detección de patrones y la asociación de factores que puede incrementar el riesgo de cáncer. En este estudio se observó que algunos cánceres se asociaron en estilos de vida propios de las poblaciones rurales, tal como es el colorrectal. En cambio, algunos como el cáncer pulmón se asociaron con estilos de zonas urbanas. Por otro lado, el análisis sobre el riesgo de presentar un segundo cáncer primario revela que ciertos factores de riesgo como son el consumo de tabaco y alcohol incrementan el riesgo de presentar estos tipos de cánceres. Finalmente, los resultados muestran que la aspirina tiene un efecto protector frente algunos cánceres, teniendo en cuenta factores de riesgo habituales como son el consumo de tabaco, alcohol o sobrepeso.

Conclusiones: En esta tesis se han integrado la base de datos de factores de riesgo y de medicamentos prescritos y se ha implementado un sistema de soporte a la toma de decisiones en la nube basado en los datos del registro poblacional de cáncer. También se han analizado e implementado varios algoritmos de aprendizaje automático, tanto supervisados como no supervisados. Los resultados obtenidos proporcionan evidencias sólidas que estos algoritmos no supervisados pueden ser utilizados para la búsqueda de patrones en pacientes con cáncer. También, detectan posibles asociaciones, interesantes para el sector médico, que utilizando otros métodos no se podrían detectar. En un contexto más médico, se ha demostrado que existe una asociación entre el consumo de tabaco y alcohol y el riesgo de padecer un segundo cáncer primario, especialmente entre los hombres. También se ha corroborado que el consumo prolongado de aspirina actúa como protector frente algunos cánceres específicos, aunque se presenten factores de riesgo. Todos los resultados obtenidos en esta tesis son una semilla poderosa para continuar explorando otros métodos y otros algoritmos con un gran potencial para convertirse en una referencia en el uso de la inteligencia artificial en este ámbito más epidemiológico del cáncer.

ACKNOWLEDGEMENTS

Life is like riding a bicycle. To keep your balance, you must keep moving.

— Albert Einstein

Doing this PhD has been an amazing experience and a great opportunity to grow as a person. It has been a long 3 years period full of difficulties and problems to overcome. A lot of times I thought If all this work was worth it. Nevertheless, my answer always was the same, If you love scabies, they don't hurt. I am very happy that I have had a chance to complete it. It would not have been possible without all those people who helped me along this journey throughout the unknown.

First of all, I would like to thank my supervisors, Prof. Pere Godoy, Prof. Francesc Solsona and Dr. Jordi Mateo, whose constant attention and care pushed me unhurriedly and steadily throughout this adventure. Also, Mr. Miquel Mesas to be a fantastic business supervisor, who has been helping and pushing to the success. Moreover, I would like to thank Mr. Ramon Piñol for trusting me since the beginning and for having always a clear idea of how to mitigate and avoid the issues.

I want to acknowledge Dr. Leonardo Espinosa-Leal and Dr. Giacomo Spigler for being excellent hosts and supervisors during my doctoral stays at the Arcada University of Applied Science and Tilburg University respectively, and for providing valuable comments and suggestions on improving my work.

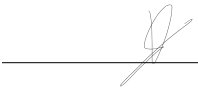
I would like to thank all the past and current members of the Department of Computer Science and Industrial Engineering, at the University of Lleida. In particular, I thank Jordi Vilaplana and Josep Maria Rius. I also want to thank my past and current co-workers, Lluis Mas i Radu Spaimoc for being such a wonderful bunch of people. I have had a great time with them. I also want to mention Bibiana for being the best clerk a department can hope for.

Furthermore, I would like to thank the managers of the Gestió Serveis Sanitaris and Catalan Health Service in Lleida for trusting me and allowing me to do this thesis. In addition, I would also like to thank the members of the Population Cancer Registry, the SAP-Argos department and Information System department at Santa Maria University Hospital.

I would also like to thank my friends and family for the support they have provided me with throughout my life, especially my family, my parents Josep

and Dolors, my brothers Sergi and Anna, and also my grandparents, uncles and cousin who exposed me to all sorts of opportunities as I grew up, and have always been incredibly encouraging of everything I wanted to do.

Lleida, Catalonia, April 16, 2023



Dídac Florensa

CONTENTS

1

INTRODUCTION AND SCOPE OF THE RESEARCH

Many of life's failures are people who did not realize how close they were to success when they gave up.
— Thomas Edison

This thesis presents a collection of artificial intelligence algorithms, statistical techniques and a web platform to tackle problems related to population-based cancer registries (PBCR). These include data accessibility, data usage and data automation. It also presents the importance of integrating different external databases, such as risk factors or registers of exposure to medicines to analyze possible associations in the Lleida region. This work presents solutions ranging from providing services and displaying data information to exploring novel methods for relating cancer factors.

The public health authorities promote population-based cancer registries to collect new cancer cases and analyze cancer incidence, mortality and survival in specific regions. Cancer has been associated with lifestyles and other risk factors. The Lleida region presents particularities that differ from other areas of Catalonia, Spain. Therefore, the Lleida register has been linked with external databases, such as medical history records to obtain risk factors and medicine prescriptions to analyze possible associations.

Cloud platforms open up opportunities for designing new **decision support systems (DSS)** to permit the transfer of knowledge from specific sectors to society. These services aim to store and process the raw data, delivering the knowledge to end-users. In this way, especially in cancer, we offer access to data information and permit the cancer situation in a specific region to be known.

Artificial intelligence (AI) simulates human intelligence processes by machines, especially computer systems. One branch of AI is **machine learning (ML)**, which uses data and algorithms to imitate how humans learn. These new methods permit interpretation of the data and add value to society, taking decisions based on it. ML systems work by absorbing large amounts of trained data, analyzing it for correlations and patterns, and using these outcomes to make predictions about future states. In this way, ML can provide practical solutions to analyze related factors faster. In addition, these methods can be integrated as a tool in DSS platforms to transfer data knowledge to society.

There are many ML applications related to cancer, from detection to treatment. However, this thesis focuses on applying ML algorithms to investigating the relationship between cancer and lifestyles. Therefore, these techniques enable some lifestyles that may play a crucial role in cancer to be detected and analyzed. In consequence, it is essential to guarantee the quality of the data and the results obtained.

The proposals in this thesis are applied to integrate external databases in order to enhance the use of PBCR data, present the cancer situation in Lleida, and detect, understand and predict how lifestyle affects cancer using ML algorithms on a population-based cancer registry. The study performed in this thesis highlights the excellent results of combining PBCR with external databases such as risk factors and prescription medicines, the benefits of AI and ML in detecting cancer patterns prior to disease detection and other data information and implementing a cloud platform to analyze the cancer situation.

1.1 CONTEXT

Cancer is one of the most important health problems that public administrations focus on understanding, and population-based cancer registries are the key. One of their primary functions is to collect information on all new cancer cases that appear in a well-defined population. Storing this information permits the analysis of cancer incidence, mortality and survival. In addition, combining the PBCR dataset with external factors, such as risk factors or exposure to medicines, allows the exploration of possible associations. Consequently, new opportunities arise where technology, such as web platforms and artificial intelligence, can help analyze and predict the incidence of cancer.

Figure 1.1 shows an overall scheme of the work involved in this thesis, from registering and integrating the information with external databases to applying machine learning algorithms and analytical statistics. The recording, merging, cleaning and analysis of the data was the key to the significant outcomes obtained in the studies. A web platform looks at the cancer situation based on cloud computing as a solution to access the data. The resulting cloud platform enables the analysis and conditions of cancer in our region to be seen.

The exploration and application of algorithms are carried out in such a web platform to find associations and patterns among cancer patients. In this way, we found solutions to transform data into information or knowledge using machine learning algorithms. The PBCR data was merged with the primary and hospital clinical records to assess the exposure to risk factors prior to cancer

diagnosis. In this thesis, we focus on exploiting data in an increasingly complex and heterogeneous way.

A cohort study was designed and implemented to analyze the relationship between risk factors and second primary cancer (SPC). Observing the good results obtained by previous studies about risk factors, we added new information to analyze the potential of PBCR and explore relationships between the exposure to medicines prior to a cancer diagnosis. To find associations, firstly, we corroborated the protective effect of aspirin against colorectal cancer in Lleida, highlighting the importance of exploring external datasets. Then, an exploration was made of the effect of aspirin against some cancers, taking the risk factors into account.

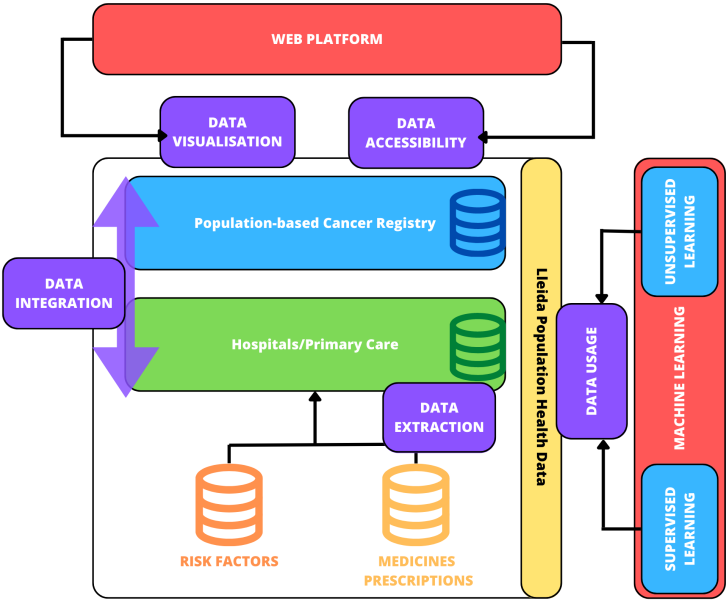


FIGURE 1.1: Overview of the external databases combined with the PBCR and the technology solutions designed and used.

1.2 POPULATION-BASED CANCER REGISTRIES

Nowadays, cancer is one of the diseases with the highest incidence in the world and the second cause of death globally. Lung, prostate, colorectal, stomach and liver cancer are the most common types in men, while breast, colorectal, lung, cervical and thyroid cancer are the most common among women [1]. In 2020,

the number of diagnosed breast cancer cases globally was 2.26 million; cases of lung cancer rose to 2.21 million; there were 1.93 million colorectal cancers, 1.41 million cases of prostate cancer, and 1.09 million cases of gastric cancer. Approximately 50% of cancer cases present a bad prognosis, and for some, such as lung cancer, the mortality is higher [1]. These statistics are a challenge for public health authorities, and researchers putting great efforts into understanding and finding a prevention for cancer.

Cancer morbidity and mortality are increasing worldwide despite the development of new treatments, research studies and trials. The global number of cancer patients (incidence rate) is expected to increase over the coming years due to negative lifestyles, demographic changes related to population aging and growth [2].

Knowing the rate of incidence of cancer is essential for public health surveillance [3]. This rate approximates the average risk of developing cancer, allowing geographic comparisons of the risk of the disease in different populations. This calculation requires a population-based cancer registry (PBCR) to record, store and organise all the cancer cases in a reference region. A continuous systematic process does this collection, storage, analysis, interpretation and reporting of data on the occurrence and characteristics of cancer cases [4]. Therefore, the aim of PBCRs is a continuous and systematic collection, storage and analysis to estimate cancer incidence and mortality.

Over the last decades, there has been exponential growth in population-based cancer registries (PBCRs). The first volume of Cancer Incidence in Five Continents (CI5), published in 1966, contained information from 32 registries in 29 countries, whereas the latest volume, published in 2021, included 343 PBCR in 65 countries.

Several data sources are integrated into PBCRs, including hospitals, death certificates, and laboratory services. Moreover, PBCRs follow international procedures, ensuring their data's high quality and reliability. This last point is vital to ensure that PBCRs worldwide follow the same rules to register cancer cases. These goals are accomplished by performing exhaustive validity checks [5], mainly manual registration.

PBCRs have commonly been used in epidemiological research. Thus, they have a crucial role in providing extensive information about tumour histology, stage at diagnosis, place and nature of the treatment, and survival [6]. Descriptive studies use the registry database to examine differences in incidence, survival, and the prevalence of risk factors or comorbidities (obesity, smoking, or diabetes) across populations and their context (such as variables associated with time, place, sex, ethnicity, and social status) [7, 8]. Risk factors are characteristics

or exposures that increase the likelihood of developing a disease or condition. Comorbidities are the presence of two or more chronic medical conditions in a patient simultaneously.

The data of population-based cancer registries are the basis for estimating the cancer burden and its trends over time and are crucial in the planning and evaluation of cancer control programmes in the area of registration [9].

Population-based registry [10] is designed to:

- To determine cancer patterns among various populations or sub-populations.
- To monitor cancer trends over time.
- To guide the planning and evaluation of cancer control efforts.
- To help prioritise health resource allocations.
- To advance clinical, epidemiological and health service research.

The PBCRs aim to identify and register all cancer cases diagnosed in a specific population or region exhaustively and continuously. In Catalonia (Spain), two population cancer registries, Tarragona Cancer Registry and Girona Cancer Registry have enabled the incidence for Catalonia to be estimated. In addition, they have demonstrated that there are differences between regions, which should be investigated [11, 12].

FIGURE 1.2: Provinces of Catalonia (Spain).

In the province of Lleida, except for Lleida city, the population is distributed in cities and towns with less than 15,000 inhabitants (considered rural areas in some cases). In this context, different lifestyles, risk factors and specific work activities can be traduced to specific incidences of certain cancers that can only be detected and analysed through a population-based cancer registry.

1.2.1 *Lleida Population-based Cancer Registry*

The Population-based Cancer Registry in Lleida was established in 2017 to retrospectively register and validate new cancer cases from 2012. The registry was implemented to analyze the incidence and risk factors of cancer in the Lleida region. As mentioned in the previous section, the main goal is to register all new cases diagnosed among the population of our region. Currently, the available period for analysis and study is 2012-2017. This system allows us to study the unique characteristics of Lleida. With half of the population living in rural areas, the exposure and risk factors may differ from other regions.

This is a new registry that is growing and improving the quality of the data collected. Initially, the population coverage for new cases was around 80%, but currently, the registry covers between 90% and 95% of cases. The registry is on track to be consolidated as a population cancer registry by the REDECAN (Red Española de Registros de Cáncer) [13]. In fact, the registry is awaiting approval from the members of REDECAN to be included. The B appendix chapter shows the rates and percentage of coverage from the PBCR of Lleida (tables B.1 - B.10). A higher tendency to greater coverage is observed. The tables B.11, B.12 and B.13 show an analysis of the cancer mortality registered in our region.

1.2.2 *Data Collection*

Cancer registries are responsible for gathering a variety of data, including information about the demographics of patients, the characteristics of tumours (cancers), treatment plans and patient outcomes. They also handle the storage and management of this data.

The process of collecting cancer data begins by identifying individuals who have been diagnosed with cancer or received treatment for cancer in various settings, including hospitals, outpatient clinics, radiology departments, doctors' surgeries, laboratories, surgical centres or from other providers (such as pharmacists) who diagnose or treat cancer patients.

The data collected by cancer registries can be classified into several categories:

- Patient demographics: Patient demographic information identifies the cancer patient. It includes the patient's name, age, gender, race, ethnicity and birthplace.

- Tumor (cancer) characteristics: Tumour characteristics describe the tumour cell type(s), biological and clinical aspects of the malignancy (such as the body organ where the cancer started), and now genomic information on

the tumour (such as specific biomarkers that might predict outcomes or responses to specific therapies).

- Stage of disease: The cancer stage describes the extent of the disease, such as how far the cancer has spread. This information informs whether a cancer has been diagnosed early or late and what treatment plans should be considered.
- Treatment: Treatment information records the various options used to treat the patient, such as surgery, radiation therapy, chemotherapy, hormone therapy, and immunotherapy.
- Outcomes: Outcomes information consists of patient's vital status, cause of death, and survival time.

The Lleida population-based cancer registry reports that the majority of cancer cases in the region are found in the Arnau de Vilanova University Hospital and Santa Maria University Hospital, which are the leading hospitals in the area. To identify potential cancer cases, the registry initially relied on hospital and pathological anatomy records as their primary sources of information. This information is prepared for a software, ASEDAT [14], which is used to validate all potential cancer cases. The software automatically validates and confirms some cases based on previously established rules. However, the remaining cases were manually reviewed and validated by doctors and nurses from the registry. During this review process, the professionals collected sociodemographic and tumor information, such as the incidence date, morphology, and pathological staging. The incidence date is defined initially as the first histology date, and if this does not exist, the hospital admission date. Another point that should be taken into account when validating is the possibility of multiple primary tumors. Cancer cases were identified and validated based on the guidelines established by the International Association of Cancer Registries, the International Agency for Research on Cancer (IARC), and the European Network of Cancer Registries [15]. The IARC establishes rules to register multiple tumors which are summarized as follows: (1) the existence of two or more primary tumors does not depend on the time at which they have been diagnosed, (2) a primary tumor is one that originates in a primary location or tissue and is not an extension, recurrence, or metastasis, (3) primary cancers that originate in the same organ (or tissue) will be considered a single tumor. Some groups of topographical codes are considered as a single organ for the purpose of defining multiple tumors (see in annex B, supplementary table B.14). Finally, once the review process was complete, the PBCR saved the patients from the Lleida region into their database from the ASEDAT database.

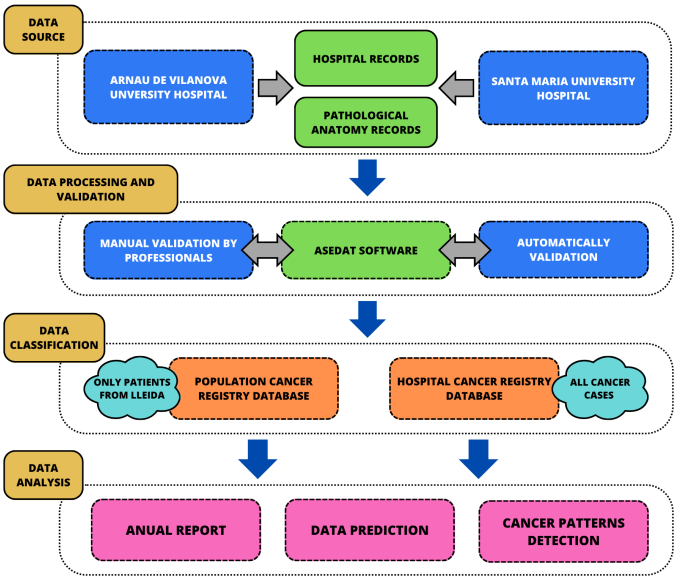


FIGURE 1.3: Workflow of data collection and data validation in the Population-based cancer registry in Lleida.

Once the cancer cases are registered and validated correctly for an extended period, it permits the patients' analysis and follow-up. In some cases, they may be diagnosed with another new cancer, named Secondary Primary Cancer (SPC). This thesis also researched the relationship between SPCs and risk factors and their prediction, exploring machine learning algorithms using the data stored in the PBCR.

1.2.3 *Secondary primary cancer*

A secondary primary cancer is a new primary cancer that develops in someone who has previously had cancer. Second primary cancers can occur months or years after the initial (primary) cancer is diagnosed and treated. Some cancer treatments, such as chemotherapy and radiation therapy, may increase the risk of a second primary cancer. Inheriting specific gene mutations and exposure to certain cancer-causing substances, such as smoking, may also increase the risk of a second primary cancer.

Cancer survival trends are generally increasing, even for some of the more mortal ones such as liver, pancreas and lung cancer [16]. Long-term survivors face

physical, psychosocial, medical, behavioural, and socioeconomic consequences due to cancer and its treatment [17]. One medical consequence is an increased likelihood of subsequent diagnosis with another cancer [18]. Cancer survivors might be especially prone to developing new cancers for various reasons. These include common etiologic risk factors with the primary cancer (i.e., environmental exposure, genetics, lifestyle choices) and the after-effects of the cancer treatment [19]. Such risk factors as obesity, smoking or heavy alcohol use could be determinant in developing a subsequent primary cancer (SPC) [20, 21]. Some risk factors were stronger related to specific cancers, such as smoking with cancer of the larynx or obesity with the stomach cancer [22].

The literature concludes the significant association between cancer and such risk factors as smoking [20]. This thesis focused on analysing the association between a first primary cancer (FPC) or SPC and risk factors. This is a challenging problem in this domain because part of this data is not stored in the PBCR. Therefore, we need to integrate it with this external database.

1.3 CANCER FACTORS

1.3.1 *Risk factors*

Exposure to risk factors plays an essential role in the biology and burden of many cancer types. Approximately 40% of cancer are associated with some risk factor [23]. There are many different risk factors for cancer, some of which may be controlled [24]. The following list highlight the main critical associated factors:

- Smoking: tobacco use is the leading cause of preventable death worldwide, and it is a significant risk factor for many types of cancer, including lung, throat, and bladder cancer.
- Unhealthy diet and lack of physical activity: a diet that is high in processed and red meats, and low in fruits and vegetables, can increase the risk of cancer. Similarly, a lack of physical activity has been linked to an increased risk of certain types of cancer.
- Heavy alcohol use: drinking alcohol has been linked to an increased risk of certain types of cancer, including breast, colon, and liver cancer.
- Exposure to radiation: exposure to high radiation levels, either through medical procedures or environmental sources, can increase cancer risk.

- Certain infections: some infections, such as human papillomavirus (HPV), VIH, and hepatitis B and C, can increase the risk of certain types of cancer.
- Exposure to certain chemicals and substances: some chemicals and substances, such as certain pesticides and asbestos, have been linked to an increased risk of cancer.
- Family history: having a family history of cancer can increase the risk of developing certain types of cancer.
- Certain inherited genetic mutations: specific inherited genetic mutations can increase the risk of developing cancer.

In this thesis we focus on the main preventable risk factors such as smoking, unhealthy diet and heavy alcohol use. These factors increase the risk of developing cancer significantly [25, 26]. Smoking is mainly related with lung cancer and many other types [27–29]. There is also strong evidence that risky alcohol use causes cancer. In addition to liver cancer [30], such others as oesophageal, gastric or colorectal cancer has been associated with alcohol [31, 32]. And, excess weight is also associated with an increased risk of cancer [33].

The residents of the Lleida region present lifestyles, risk factors and work activity which may increase the incidence of certain types of cancer. Nearly half the population of Lleida province live in rural and semi-urban areas. As a consequence, their lifestyle may be different from that of the more urban populations in other Catalan provinces [34, 35]. Thus, they may face different risk factors and socioeconomic status (SES).

In order to conduct our research, it was necessary to investigate additional external sources, such as hospital records and primary care databases, especially the latter. To access the required information for each patient, we used a personal identification code called the CIP (Personal Identification Code). The CIP is only available in Catalonia and was created by the Catalan Health Service to identify each person when they come into contact with the healthcare system (hospitals, primary care, and pharmacies) [36]. We used the CIP to retrieve the height and weight measurements of each patient that were recorded prior to their cancer diagnosis in order to determine their body mass index (BMI). With these parameters, we calculated the BMI using the formula

$$BMI = (weight(kg))/height(m)^2$$

. For smoking, we consulted the primary care database to determine whether each patient has been registered with smoking, using the Z72.0 code from the International Classification of Diseases (ICD-10). A similar process was followed

to determine alcohol consumption, using the F10.9 code (alcohol use) for this purpose. Once these patients were identified, we obtained data on the number of grams they consumed per day prior to their cancer diagnosis. The information about these factors is all extracted before the first primary cancer diagnoses. Therefore, for each patient, we obtain the first cancer incidence and extract their BMI, smoking, and alcohol consumption. In the case of the latter two factors, if a diagnosis exists, we evaluated whether the exposure was for at least five years.

This thesis has made several significant contributions demonstrating the potential for merging external sources with the population-based cancer registry and adding studies about cancer in Lleida and the association between risk factors.

1.3.2 *Pharmacology medicines*

The long exposure to some medicines could be considered a risk factor for some pathologies due to their secondary effects [37], despite differences between the literature. Some previous studies demonstrated the relation between some medicines and cancer [38]. However, there is also prior evidence for a protective effect of some drugs. One such case is aspirin, but less research has been done on other medicines [39].

1.3.2.1 *Acetylsalicylic acid*

Aspirin belongs to a group of drugs called salicylates. Its action is to stop the production of certain natural substances that cause fever, pain, inflammation and blood clots. Aspirin has long been known to prevent cardiovascular and cerebrovascular diseases [40]. Daily administration of a low dose of aspirin has been proven beneficial for preventing recurrent cardiovascular events [41]. It is also prescribed to relieve symptoms of rheumatoid arthritis, osteoarthritis, systemic lupus erythematosus, and certain other rheumatologic disorders (disorders in which the immune system attacks parts of the body).

Recently, many studies have shown that long-term use of aspirin can significantly reduce the risk of cancer [42]. Specifically, aspirin consumption has been strongly related with a protective effect against colorectal cancer (CRC) [43]. The exact mechanism by which aspirin exerts this effect is started to investigate and understand but it is thought to be related to its anti-inflammatory properties. A recent study concluded that the preventive effect of aspirin has been attributed to the inhibition of cy-clooxygenase (COX), the enzyme responsible for the synthesis of prostaglandins [44].

Several large studies have shown that regular use of aspirin is associated with a lower risk of developing colorectal cancer and other types of cancer, such as breast, pancreas, and prostate cancer [39, 42]. However, it is important to note that aspirin can also have side effects, including an increased risk of bleeding, and it is not appropriate for everyone. It is important to speak to a healthcare specialist before taking any new medications.

The results gathered in this thesis have been made possible by using pharmacy prescriptions, highlighting the importance of the population-based cancer registry. This service permits the register and control of cancer cases, which can be used to analyse other sources, such as exposure to medicines. The results presented in the paper 5 and 6 show the association between cancer and aspirin use among Lleida population.

1.4 CLOUD BASED DECISION SUPPORT SYSTEM

Decision support systems (DSS) are information system that offers solutions to help a user in the decision-making process. These systems integrate the data belonging to different heterogeneous sources. Secondly, they are presented as usable and attractive, allowing a fluent and simple user interaction. Thirdly, they can generate descriptive analytics to extract knowledge from the data. Finally, a DSS must have realistic, credible and successful pilot cases. Figure 1.4 presents a cloud architecture of a DSS presented in this thesis.

Decision support systems are similar to traditional client/server web applications. Both must handle a variety of data sources, whether stored locally or remotely, and both require a client to interact with users, gather information, process requests, and execute operations. It is beneficial to use restful APIs (Application Programming Interface) and the JSON (JavaScript Object Notation) format to organise and communicate between the different components of an intelligent cloud architecture, including the server, client, data sources, and workers.

A cloud-based DSS leverages the benefits of cloud technology to address the traditional limitations of DSSs. Hosting the DSS on the cloud makes it accessible to a wider audience and offers elasticity and scalability, allowing virtual resources to be created and eliminated as needed. It ensures that data storage can effectively scale to meet various requirements.

1.4.1 *Cloud platform*

Let us start defining the main concepts of cloud computing. Mainly, there are three kinds of cloud services:

- Infrastructure as a service (IAAS): clients pay-as-you-go access to storage, networking, servers and other computing resources in the cloud.
- Platform as a service (PAAS): offers access to a cloud-based environment in which users can build and deliver applications.
- Software as a service (SAAS): provides software and data through the Internet. Users subscribe to the software and access it via the web or APIs.

This thesis focuses on SAAS implementation to understand the cancer situation. Software-as-a-service (SAAS) removes the need for a high-powered computer with a dedicated environment to run tasks. All that is required is a terminal with an Internet connection. With a web browser, end-users can access virtually unlimited computing power from any device, including handheld devices, desktop computers, or laptops.

The implemented platform is a solution for analysing the incidence of cancer, risk factors and mortality in the Lleida area also for data accessibility and transparency. In addition, the system has been designed to be deployed to other PBCRs.

Once the cancer situation in Lleida has been analysed, the thesis focuses on how artificial intelligence algorithms can detect associations and cancer patterns. These algorithms are explained in the next section.

1.5 ARTIFICIAL INTELLIGENCE

Artificial intelligence (AI) is changing the concept of data. It permits value to be obtained from large data stores and automates any process, thus using less time. There are many definitions for AI. In the context of this thesis, we could describe it as a digital computer's ability to understand and perform tasks associated with intelligent beings. AI represents many complex algorithms that depend on the situation, the problem and the solution we want to achieve. AI includes techniques such as Natural Language Process (NLP), Computer Vision, Machine Learning (ML), Robotics, Deep Learning, Robotics and Fuzzy Logic. In this thesis, our system is based on a specific aspect of AI. The research focuses on using ML, which permits a machine to learn to do an activity or task without human supervision.

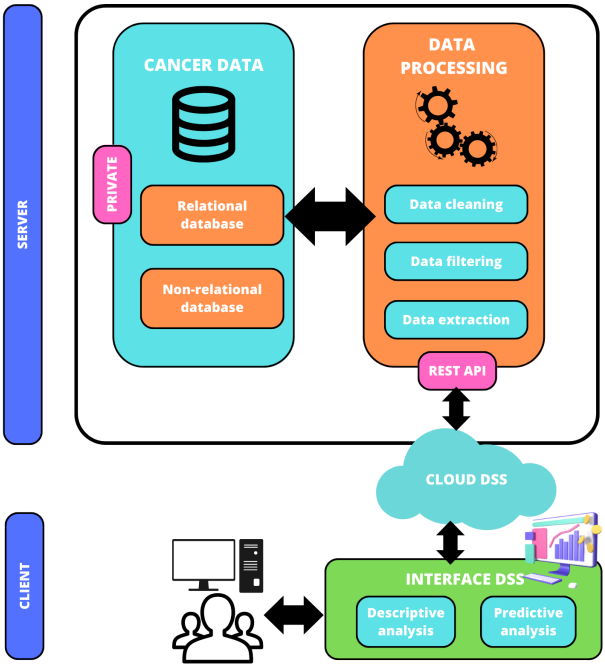


FIGURE 1.4: Cloud architecture of a DSS presented in this thesis.

1.5.1 *Machine Learning algorithms*

As defined in the previous section, ML is a branch of AI that uses data and algorithms to imitate how humans learn. Its algorithms learn from previous information to act without human supervision. Through statistical models, algorithms are trained to make classifications or predictions. In terms of classification, the programmes learn, from the given specifications or observation, to classify in classes or groups. These algorithms, through previous characteristics, permit the element in the class to be detected and classified. For example, given certain characteristics, they can classify a dog or cat. The prediction is about fitting a shape that is as close to the data as possible. As the name indicates, given a historical dataset, it permits a future outcome to be predicted.

ML is a branch of AI that learns from the data to take decisions and regarding the output that we want, we will choose supervised, unsupervised or reinforcement learning. Main branches of ML (see figure 1.5 to explore different ML applications):

- Unsupervised learning
- Supervised learning
- Reinforcement learning

This thesis uses unsupervised and supervised learning for the detection and prediction of patterns respectively. It also highlights the importance of using this data to train these algorithms and the value of the outcomes they add.

1.5.2 *Unsupervised-learning*

Unsupervised learning is a type of machine learning where the model is not given any labelled training data and is instead asked to learn patterns or relationships in the data on its own. The goal of unsupervised learning is often to discover the underlying structure in the data or learn more about it by identifying common patterns.

Some common types of unsupervised learning algorithms include:

- **Clustering algorithms**: These algorithms group similar data points together into clusters.
- **Dimensionality reduction algorithms**: These algorithms aim to reduce the number of dimensions or features in the data while preserving as much important information as possible.

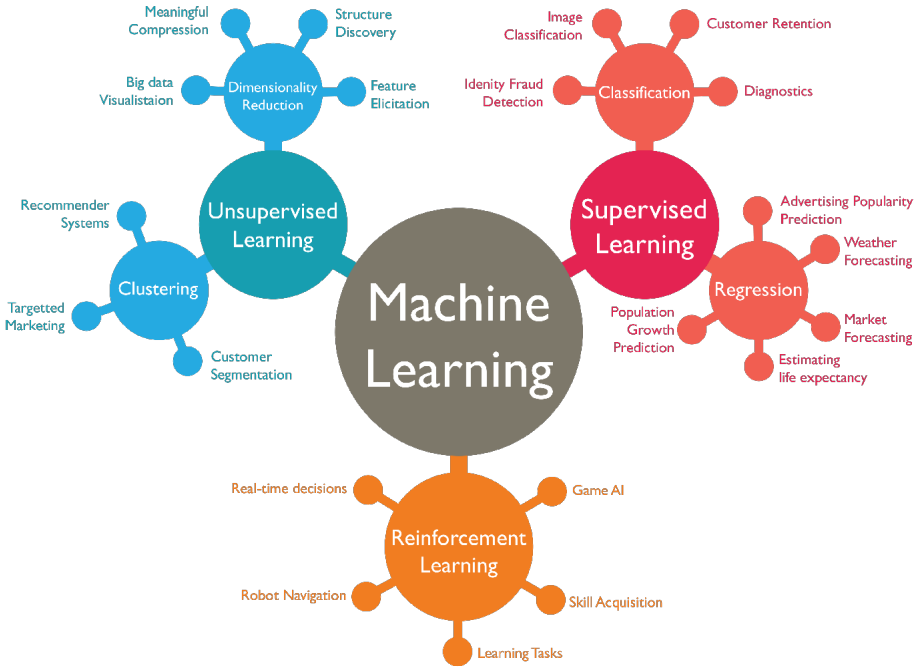


FIGURE 1.5: Machine-learning branches and some applications

- **Generative models**: These algorithms learn to generate new data points similar to the ones in the training set.

Unsupervised learning can be helpful for such tasks as anomaly detection, where the goal is to identify unusual data points that do not fit the pattern of most of the data. This thesis focuses on applying clustering and dimensionality reduction algorithms as pattern detection. More specifically, in Multiple Correspondence Analysis and K-means. This is explained below.

1.5.2.1 *Multiple Correspondence Analysis*

Multiple Correspondence Analysis (MCA) is an extension of Correspondence Analysis (CA). MCA is an unsupervised learning algorithm for visualising the patterns in extensive tables and for multi-dimensional categorical data [45]. This method can describe, explore, summarise and visualise information on individuals described by categorical variables within a data table [46]. Unlike CA, MCA can deal with more than one categorical variable. This is the main advantage of the MCA technique. In our case, MCA was first used to evaluate the relationships between the features. It was then used to assess the associations between socio-demographic information for each cancer.

The factors produced are interpreted with the help of various statistical coefficients, which complemented each other to provide a better interpretation. The most common and important are inertia, the eigenvalue and the contribution and factorial coordinates. Inertia measures the dispersion of the set of computed distances between points. Analogously, in Principal Component Analysis (PCA), inertia corresponds to the explained variance of dimensions. The eigenvalue allows the inertia that a specific category produces to be quantified. The contribution enables us to consider how much influence a category has in determining a certain percentage relative to the entire set of the active category. The percentage coordinates (x- and y-axis) of the graph enable the category points in a graph to be represented and established. In MCA, the distance between two or more categories of different variables can be interpreted as the associations and correlations between these. If two categories present high coordinates and are close in space, they tend to be directly associated. If two categories present high coordinates but are distant from each other (e.g. they have opposite signs), they tend to be inversely associated. If two categories present the same coordinate sign, they can be related to each other [47, 48].

1.5.2.2 *K-means*

K-means [49] is a non-supervised learning algorithm used in data-mining and pattern recognition. The algorithm partitions the data set into K pre-defined distinct non-overlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intra-cluster data points as similar as possible while keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneity (similarity) between the data points within the same cluster.

The K-means algorithm is composed of the following steps: 1) It places K points in the space represented by the patients who are being clustered. 2) It assigns each patient to the group that has the closest centroid. 3) When all patients have been assigned, it recalculates the positions of the K centroids. Steps 2 and 3 are repeated until the centroids no longer move. This produces a separation of the patients into homogenous groups while maximising heterogeneity across groups. The optimal number of clusters was obtained by the elbow method [50].

To the best of our knowledge, this well-studied and traditional method has not been fully explored to detect patterns of cancer. In this thesis, we prove the importance and the benefits of using it for cancer.

1.5.3 *Supervised-learning*

In supervised learning, a model is trained using labelled data that include both input and corresponding correct output labels. The model uses this input/output mapping to make predictions and the objective is to accurately predict the output for new, unseen data. This type of machine learning involves training a model on labelled data.

Supervised learning involves training a model to learn a function that maps input data to the correct output labels. The model is provided with a set of labelled training examples and adjusts its internal parameters to minimise the error between the predicted output and the correct output. During the training process, the model is presented with input data and the corresponding correct output labels, and it uses this information to improve its ability to make accurate predictions on new, unseen data.

There are various types of supervised learning, including classification, regression, and structured prediction. In classification, the goal is to predict a categorical label. In regression, the target is to predict a continuous value, such as the price of a house or the yield of a crop. In structured prediction, the aim

is to predict a structured output, such as a sequence of words or a tree-like structure.

It is worth noting that supervised learning is often contrasted with unsupervised learning, a type of machine learning in which the model is not provided with correct output labels and must discover patterns in the data on its own.

1.5.3.1 *Supervised-learning models*

Various classification algorithms can be used to perform predictions for classification purposes. Table 1.1 summarises the classification models used in this thesis including those which use supervised and semi-supervised classification training.

1.5.4 *Performance metrics*

In a classification or prediction task, there are four possible outcomes:

- If the sample is positive and is correctly classified as positive, it is considered a *true positive* (TP).
- If the sample is positive but is incorrectly classified as negative, it is considered a *false negative* (FN).
- If the sample is negative and is correctly classified as negative, it is considered a *true negative* (TN).
- If the sample is negative but is incorrectly classified as positive, it is considered a *false positive* (FP).

Based on that, the Accuracy, Sensitivity (also known as recall), Specificity, Precision and F-score metrics [58] are the most relevant metrics used to evaluate the performance of the classification models. The AUC is also useful.

- **Accuracy** (Eq. 1.1). Ratio between the correctly classified samples.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.1}$$

- **Sensitivity** (Eq. 1.2). Proportion of correctly classified positive samples compared to the total number of positive samples.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1.2}$$

Model	Definition
RF: Random Forest	RF classifier is a combination of tree predictors. Each decision tree performed the classification independently and RF computed each tree predictor classification as one "vote". The majority of the votes computed by all of the tree predictors decided the overall RF prediction [51].
DT: Decision Tree	DT is a flowchart with a tree structure, where each internal node represents a test or decision that needs to be made, while the branches represent the possible outcomes of that decision. Leaf nodes represent the final outcome or prediction [52].
NN: Neuronal Networks	NNs are networks that utilize complex mathematical models for data processing. A neural network connects simple nodes, also known as neurons or units. And layers of such nodes forms a network of nodes. An array of algorithms are used to identify and recognise relationships in data sets [53].
BNB: Bernoulli Naive Bayes	BNB supports categorical features and it models each as conforming to a Multinomial Distribution [54].
GNB: Gaussian Naive Bayes	GNB supports continuous valued features and it models each as conforming to a Gaussian (normal) distribution [54].
LDA: Linear Discriminant Analysis	LDA estimates the mean and variance in the training set and computed the covariance matrix to capture the covariance between the groups to make predictions by estimating the probability that the test set belonged to each of the groups [55].
LR: Logistic Regression	LR uses a generalised linear model for binomial distributions. A logit link function is used to model the probability of "success". The purpose of the logit link is to take a linear combination of the covariate values and convert those values into a probability scale [56].
SVM: Support Vector Machines.	SVM is a powerful, kernel-based classification paradigm. Support vector machines (SVMs) are particular linear classifiers which are based on the margin maximisation principle. They perform structural risk minimisation, which improves the complexity of the classifier with the aim of achieving excellent generalisation performance. The SVM accomplishes the classification task by constructing, in a higher dimensional space, the hyperplane that optimally separates the data into two categories [57].

TABLE 1.1: Supervised-learning models

- **Specificity** (Eq. 1.3). Proportion of correctly classified negative samples compared to the total number of negative samples.

$$Specificity = \frac{TN}{TN + FP} \tag{1.3}$$

- **Balanced accuracy** (Eq. 1.4). Arithmetic mean of sensitivity and specificity

$$Balanced\ accuracy = \frac{Sensitivity + Specificity}{2} \tag{1.4}$$

- **AUC** (Eq. 1.5). The Receiver operating characteristics (ROC) curve is a two-dimensional graph in which Sensitivity is plotted on the y-axis and $1 - Specificity$ is plotted on the x-axis. The points of the curve are obtained by sweeping the classification threshold from the most positive classification value to the most negative. The AUC score is a scalar value that measures the area under the ROC curve and is always bounded between 0 to 1.

$$AUC = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} 1_{p_i > p_j}, \tag{1.5}$$

  where i runs over all m samples with true label positive, and j runs over all n samples with true label negative; p_i and p_j denote the probability score assigned by the classifier to sample *i* and *j*, respectively.

1.6 RELATED WORK AND CONTRIBUTIONS

1.6.1 *Risk factors and lifestyle*

Risk factors can be crucial when discussing cancer incidence because approximately 40% of cases are related to some factors [23]. Smoking, risky drinking or bad diet are one of the most preventable factors associated with cancer. Smoking and alcohol use have recently been shown to increase the risk of some cancers by the UK Biobank study [59]. And same happens with excess weight, which also is associated with cancer risk [33].

Recently some studies have concluded that the incidence of some cancers depends on the geographical area [60–62]. And another study concludes which cancers are more related to rural than urban areas, and the contrary [63]. Potential explanations for lower overall incidence rates in rural areas compared with urban zones were given. These include smoking (more prevalent in urban areas) and

exposure to environmental pollutants. And in rural areas, the patients are typically older, less educated, poorer and have less access to such healthcare services as early-cancer detection. Whitney E Zahnd *et al.* presented a report about rural-urban differences in cancer incidence and trends in the United States [64].

Related to secondary primary cancer, some studies have demonstrated an increase of this incidence related to effectiveness of the new treatments [16]. In addition, the factors cited before also are related to. Therefore, they play a crucial role in terms of present a new cancer after the previous one. Risky drinking and smoking have been related with the risk of second cancer [20, 21]. In section 4.4, relations between SPC and risk factors are discussed for the Lleida population. Also, section 4.6 an association between some cancers and risk factors.

This thesis contributes to integrating the external database, such as lifestyles and PBCR data. Merging these databases permits the analysis of the association between cancer and risk factors and how they increase the incidence of this disease. The implications are analysed for the first and the second primary cancer.

1.6.1.1 *Acetylsalicylic acid*

Above, some factors that increase cancer risk are presented. In this case, we cite studies about the protective effect of aspirin against cancer. Some previous studies have corroborated that aspirin protects against colorectal cancer [39]. The long-term effects of aspirin on colorectal cancer outcomes using trials of aspirin were corroborated [65]. Many studies have demonstrated this association even though the effects of risk factors and aspirin use have yet to be analysed [66].

Due to the association between colorectal cancer and aspirin as a protector, other cancers have been explored. One of the cancers for which the literature has demonstrated an association with the long-term use of aspirin as a protector is oesophageal [67]. Other studies have investigated the effect of aspirin on a range of cancers. Hwang *et al.* concluded that a low-dose of aspirin was associated with a significantly lower risk of hepatocellular carcinoma [68]. Another cancer with a lower risk related to aspirin use was pancreatic cancer [69]. In the case of lung and bronchial cancer, the combination of aspirin and metformin also had independent protective associations [70]. Finally, the same conclusions were obtained for breast cancer, in which the use of aspirin was related to a lower risk [71].

There is some controversy in other cancer associations, as Tsoi *et al.* showed [42]. Their study analysed the relation between aspirin use and some cancers. Their conclusions did not suggest the same results in all cancers. Figure 1.6 shows the effect of low-dose aspirin consumption against some cancers.

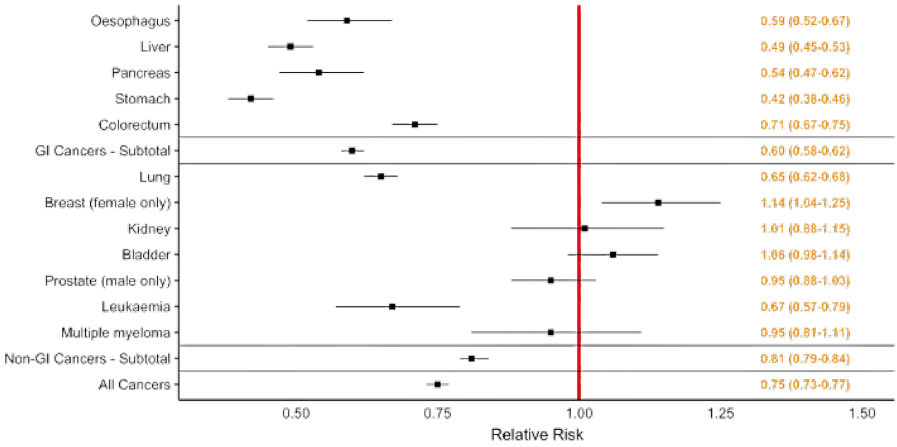


FIGURE 1.6: Effect of low-dose of aspirin against some cancers. Outcomes obtained from the study done by Tsoi *et al.* [42].

The main contribution in this field is the integration of prescription medicines into the PBCR data. This merging corroborates that long-use aspirin decreased the risk of colorectal, liver and pancreatic cancers, taking into account the risk factors. In section 4.6, the associations between some cancers and aspirin use are presented, taking into account the role of risk factors for the Lleida population.

1.6.2 *Cloud platforms*

Cloud platforms based on DSS help the user in the decision-making process. Recent literature has focused on this kind of platform, even though the use of these tools for cancer prediction is scarce.

New applications have been designed to facilitate the analysis of data sets. Some studies have suggested that the latest technologies can extract information and value from the data rapidly and obtain the results instantly in different contexts. Miller *et al.* designed and implemented an application to generate anatomical visualisations of cancer lesions [72]. They concluded that data visualisations of clinical tumour characteristics could help to understand the natural history of malignancies. Therefore, this interactive data visualisation app was designed to enable the analysis of the tumour characteristics. Another Rshiny application related to cancer data was published by Zhang *et al.* [73]. The researchers designed a platform to analyse cell line responses to an anti-cancer drug. They concluded that it helped researchers understand the response

of tumour cell lines to 15 therapeutic agents. Finally, a similar platform was implemented by Xia *et al.* to visualise cancer risk factors and mortality [74].They shared a data warehouse and Rshiny app to improve their understanding of spatial and temporal trends across the population served by the University of Kansas Cancer Center.

The main contribution to this area is integrating the PBCR data in a web tool as a service of the society comfortably and intuitively to analyse the cancer incidence in the Lleida region.

1.6.3 *Machine Learning*

The emergence of technological advances, such as artificial intelligence and data analytics during recent years has generated radical changes in the health sector. In this context, Machine Learning has become a relevant instrument for understanding the consequences of and help the diagnosis of some diseases, in particular, cancer, one of the diseases with the highest incidence.

Machine Learning is often used to analyse and predict the prognosis in cancer patients [75]. Some applications focus on cancer survival by the tumour information, age at diagnosis and histology. And most of them focus on specific cancer. Usually, on the highest incidence cancers such as lung or breast cancer [76, 77]. And another tumour that has been analysed to be predictable is prostate [78]. In this study, the authors showed that the computer-based decision-support systems that use machine learning (ML) have the potential to transform medicine by performing complex tasks that specialists currently carry out. These systems can improve diagnostic accuracy, increase efficiency in streamlining clinical workflow, reducing human resource costs, and enhancing treatment choices. These characteristics could be particularly useful in managing prostate cancer, with increasing applications in diagnostic imaging, surgical interventions, skills training and assessment, digital pathology and genomics.

And also, some studies demonstrated that the use of risk factors and the combination of interacting genetic variants predict specific cancer, in this case, for breast and oral cancer [79, 80].

1.6.3.1 *Unsupervised learning models*

MCA is an unsupervised learning model used in many applications as a dimensionality reduction algorithm. In this case, it was used to associate some factors and search for patterns between them. This is the novelty for cancer epidemiology in this thesis.

To understand the application of MCA, we initially based on a study [81] about healthy ageing. Next, we also based on another study [82] which concluded that bad driving and crashes could be affected by differences between urban and rural areas, traffic volume, driver age and more. Another starting point was the study done by White *et al.* [83] which evaluated the relationship between air pollution, PM components and the risk of breast cancer in a United States-wide prospective cohort using a clustering technique. In addition, a previous study used MCA to analyse the prognosis in surgery for low rectal cancer [84].

The combination of many unsupervised algorithms, such as MCA and K-means, was also a novelty in this area, although some previous studies had been done in other fields. A starting point for this combination was the study presented in [85]. This used the combination of MCA and K-means to ascertain multimorbidity patterns. It concluded that these techniques could help to identify these patterns. Another study our work was based on is the one presented in [86]. It studied the trends in incidence of cancers associated with being overweight and obese. Another study [87] analyzed the possible relation between obesity and colorectal cancer. These articles studied the impact of the risk factors on colorectal cancer but did not use the MCA technique or K-means algorithm to explore associations between these and their impact. Another study used K-means to search patterns in colorectal patients. Even though, its main aim was to detect emotion regulation patterns and personal resilience [88]. Therefore, many prior studies have used MCA or K-means, although the novelty of this thesis is how we have applied these techniques to link types of risk factor, SES, tumor stage and patients' characteristics as we do.

This thesis contributes to implementing and training of unsupervised algorithms to detect cancer patterns. Furthermore, we suggest different ways to develop and adapt them to similar fields.

1.6.3.2 *Supervised learning models*

Supervised learning algorithms allow researchers to build models that predict or classify different classes. In this thesis, this enables the risk of cancer to be predicted depending on risk factors and socio-demographic information. Related to that, several studies have been carried out to predict cancer risk in general. Brockmoeller *et al.* used deep learning to identify risk factors for lymph node metastasis in colorectal cancer [89]. Another study used machine learning to construct a model to predict cancer under normal and tumour conditions with genetic information [90]. Most of the publications related to predictive analytics, such as Zhang *et. al* [91], focus on identifying the risk factors for predicting secondary cancer, but are not used to build tools to transfer this knowledge to

practical situations. However, many efforts have been made to associate the risk of an SPC with the FPC and risk factors [92].

The present work does not include the results of the experiment as it is currently in the initial stages. Despite this, the preliminary findings are promising and indicate a new avenue for further research.

1.7 PUBLICATIONS

The following publications have been derived from the work on this thesis.

1.7.1 *Journal publications*

1. D. Florensa, P.Godoy, J. Mateo, F. Solsona, T. Pedrol, M. Mesas and R. Piñol. The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables and Cancer Incidence. IEEE Journal of Biomedical and Health Informatics, 2021, 25 (9), 3659-3667. doi:10.1109/JBHI.2021.3073605.

2. D. Florensa, J. Mateo, F. Solsona, T. Pedrol, M. Mesas, R. Piñol and P.Godoy. Use of Multiple Correspondence Analysis and K-means to Explore Associations Between Risk Factors and Likelihood of Colorectal Cancer: Cross-sectional Study. Journal of Medical Internet Research, 2022, 24 (7), e29056. doi:10.2196/29056.

3. D. Florensa, J. Mateo, S. López, A. Torres, F. Solsona and P.Godoy. Exploring Cancer Incidence, Risk Factors and Mortality in the Lleida Region: An Interactive open-source R Shiny Application for cancer data analysis. JMIR Cancer, Accepted, 2023.

4. D. Florensa, J. Mateo, F. Solsona, C.Miret, M. Mesas, R. Piñol and P.Godoy. Association of smoking and heavy alcohol use with risk of Subsequent Primary Cancer. Cancers, Under Review, 2023.

5. D. Florensa, J. Mateo, F. Solsona, L. Galván, R. Piñol, M. Mesas, L. Espinosa-Leal and P.Godoy. Acetylsalicylic acid effect in colorectal cancer taking into account the role of tobacco, alcohol and excess weight. International Journal of Environmental Research and Public Health, 2023, 20 (5), 4104. doi: 10.3390/ijerph20054104

6. D. Florensa, J. Mateo, F. Solsona, L. Galván, R. Piñol, M. Mesas, L. Espinosa-Leal and P.Godoy. Low-dose of acetylsalicylic acid for cancer

prevention taking into account risk factors: A retrospective cohort study. Annals of Epidemiology, Under Review, 2023.

1.7.2 *Conference publications and attendance*

1. D. Florensa, J. Mateo, F. Solsona, C. Miret, S. Godoy and P. Godoy. Asociación de consumo de tabaco y alcohol con el riesgo de un segundo cáncer primario (2022). XL Reunión Anual de la Sociedad Española de Epidemiologia (SEE). Gac Sanit; 2022; 36:269

2. D. Florensa, J. Mateo, F. Solsona, C. Miret, S. Godoy and P. Godoy. Uso de aspirina prolongado, factores de riesgo y asociación con el cáncer de colon y recto (2022). XL Reunión Anual de la Sociedad Española de Epidemiologia (SEE). Gac Sanit; 2022; 36:336

3. D. Florensa, J. Mateo, F. Solsona, P. Godoy and L. Espinosa-Leal. Predicting the Colorectal Cancer mortality in the Region of Lleida, Spain: A Machine Learning Study (2022). The 12th International Conference on Extreme Learning Machine (ELM2022).

1.7.3 *Other publications*

1. D. Florensa, T. Pedrol, I. Modol, X. Farre, A. Salud, J. Mateo and P.Godoy. El Registre poblacional de càncer a Lleida en zones urbanes i rurals. Resultats de l'any 2014. Butlletí Epidemiològic de Catalunya, 2019, 40 (12), 252-264.

2. D. Florensa, J. Mateo, F. Solsona, C. Miret, S. Godoy and P. Godoy. Severity of COVID-19 cases in the months of predominance of the Alpha and Delta variants. Scientific Reports, 2022, 12 (1), 15456. doi: 10.1038/s41598-022-19125-4

3. D. Florensa, J. Mateo, F. Solsona, P. Godoy and L. Espinosa-Leal. On the Intensive Care Unit Admission During the COVID-19 Pandemic in the Region of Lleida, Spain: A Machine Learning Study (2021). Proceedings of ELM 2021. ELM 2021. Proceedings in Adaptation, Learning and Optimization, 16, 92–103. doi: 10.1007/978-3-031-21678-7_9

1.8 DOCTORAL STAYS

1.8.1 *Arcada University of Applied Science - Helsinki*

During the course of the present thesis, I spent a four-month doctoral stay at the Arcada University of Applied Science in Helsinki, where I was involved in a project analysing drug exposure and cancer risk. My supervisor was Dr. Leonardo Espinosa-Leal from the Department of Big Data Analytics. This stay was supported by the HPC-Europa3 programme, which permits researchers access to world-class HPC systems for academic and industrial researchers.

My work consisted of designing and implementing software to automate the analysis of possible associations between cancer and drugs. Next, different machine learning algorithms were trained to build models capable of predicting the cancer risk from exposure to medicines.

The main objective of my work was to develop the models and integrate the data information into a client application and permit access to it for researchers. The system was based on a non-relational database done with the MongoDB and an API service to extract the information and prepare the cases to be analysed. Then, the results are shown and presented to the user through a web-based dashboard.

Appendix A shows the partial work done in this stay.

1.8.2 *Tilburg University*

During the course of the present thesis, I spent a three-month doctoral stay at Tilburg University in Tilburg. There, I was involved in a project about using the Decision Tree algorithm as a clustering technique using cancer datasets. My supervisor was Dr. Giacomo Spigler from the Department of Cognitive Science and Artificial Intelligence.

My work consisted of creating new software to understand and process Decision Tree nodes. Next, this software was adapted to change the interpretation results of this algorithm as a cluster algorithm. Finally, the different configurations were tested to obtain the best performance and corroborate the results of the previous literature.

The main objective of my work was to develop a library that allows the Decision Tree algorithm to be adapted like a clustering technique. This new software was also offered to other researchers. Then, the aim was to discover patterns that relate risk factors, comorbidities and cancer risk.

2

HYPOTHESES

The implementation of new technological solutions can help automate the use and accessibility of Population-based Cancer Registry (PBCR) data. Furthermore, these solutions can integrate external databases to enhance the PBCR's capabilities. Artificial intelligence can correlate cancer risks and patterns, while cloud solutions provide fast access to a wide range of analyzed information.

1. Cloud applications enable real-time monitoring and analysis of cancer data for public health surveillance and outbreak response. Cloud platforms facilitate the rapid sharing and analysis of epidemiological cancer data. This can help identify areas or populations at increased risk of cancer and inform targeted interventions and prevention strategies.

2. Machine learning models are capable of accurately associating cancer risk based on previous comorbidities and risk factors, and can reveal previously unknown patterns and associations between these factors. By analyzing large datasets of medical records, machine learning algorithms may identify subtle relationships between patient characteristics and cancer risk, leading to more accurate risk assessments.

3. There is a correlation between the geographical area of exposure and the risk of developing specific types of cancer. The incidence of certain types of cancer may be influenced by factors related to the location of exposure, such as environmental pollutants, lifestyle factors, or genetic predisposition.

4. Studies have shown that cancer patients who smoke or consume alcohol are at a higher risk of developing secondary primary cancers, which are new cancers that arise in a different part of the body from the original cancer.

5. The potential benefits of aspirin use for colorectal cancer prevention are promising, and the decision to use aspirin should be made on an individual basis after careful consideration of the potential risks and benefits. Aspirin use over a prolonged period of time has been shown to reduce the risk of colorectal cancer by 20%.

6. Many observational studies and clinical trials have suggested that aspirin use may reduce the risk of developing certain types of cancer, over an extended period of time in patients who are aged > 50 years.

3

RESEARCH OBJECTIVES

The primary objective of this thesis is to enhance the accessibility and utilization of the Population-Based Cancer Registry (PCBR) by utilizing artificial intelligence models and statistical methods. The research aims to investigate the application of machine learning algorithms and statistical techniques for analyzing cancer incidence, risk, and patterns in Lleida. This involves extracting and integrating data on risk factors and prescription medications from various databases. To address this objective, a set of specific objectives was identified:

1. To analyze the cancer incidence in a specific region (Lleida) by cloud architecture based on DSS and offer data accessibility to society.
2. To extract and integrate factors such as body mass index, alcohol and smoking, and a database of prescription medicines from the hospitals and primary care with the PBCR database.
3. To assess the effectiveness of integrating external databases, including risk factors and prescription medications, with the PBCR database.
4. To analyze the association of possible risk factors such as overweight/obesity, smoking or alcohol use and secondary primary cancer by statistical methods.
5. To search for association between geographical area, smoking and body mass index and specifics cancers by unsupervised learning algorithms.
6. To analyze the association between the effect of acetylsalicylic acid on some cancers, highlighting the important role of population-based cancer registries.

4

METHODOLOGY

This chapter presents the studies conducted and presented in this thesis. As explained in the previous chapter, the thesis focuses on the use of PBCR data and its integration with external databases using new technological solutions, particularly in the Lleida region. The main goal of the registry is to analyze the incidence, mortality, and morbidity of cancer in a specific territory to identify its unique characteristics. The data used in the presented studies includes socio-demographic information such as age, date of birth, date of death, and population type (urban or rural), tumor information such as location and staging, risk factors (obtained from external databases), and pharmacological information (also obtained from external databases). Other variables are also included, and they are described in detail in the respective studies.

This information is saved and stored on servers that comply with protection rules. Access to server resources is limited only to authorized personnel who require it for their job duties. Access controls may involve the use of passwords, two-factor authentication, and other security measures to prevent unauthorized access to sensitive data. Regular server maintenance and updates are also essential to maintain security, including applying security patches, updating software and hardware, and performing regular security audits to identify and address potential vulnerabilities. Data is encrypted both when in transit and at rest. Encryption involves converting data into a form that can only be read with a decryption key. This ensures that even if data is intercepted by unauthorized parties, it cannot be read or used without the decryption key.

All data have been anonymized in order to protect the privacy and confidentiality of patients. Approval for this thesis and the use of this information has been granted by the reference ethics committee, specifically the Committee of Ethics and Clinical Research of Lleida - CEIC. As the present study is a retrospective investigation, and the patients were blinded to the investigators, written informed consent is not required, in accordance with the Clinical Investigation Ethical Committee. The extraction of data has been performed by professionals who did not directly participate in the subsequent analysis.

4.1 PAPER I: DSS PLATFORM TO ASSIST CANCER ANALYSIS

Authors: *Dídac Florensa, Jordi Mateo, Sergi López, Anna Torres, Francesc Solsona and Pere Godoy*
Journal: JMIR Cancer
Status: Accepted
Keywords: *R Shiny, Cloud computing, Microservices, Docker, Cancer incidence, Risk factors, Cancer mortality*

Exploring Cancer Incidence, Risk Factors and Mortality in the Lleida Region: An interactive open-source R Shiny application for cancer data analysis

Abstract: The cancer incidence rate is essential in public health surveillance. The analysis of this information allows the authorities to know the cancer situation in their regions. This study aimed to present the design and implementation of a shiny app to assist cancer registers in conducting rapid descriptive and predictive analytics in a user-friendly, intuitive, portable and scalable way. Moreover, we want to describe the design and implementation roadmap to inspire other population registers to exploit their datasets and develop similar tools and models. The first step is to consolidate the data into the population register cancer database. This data is cross-validated by ASEDAT software, checked later, and reviewed by experts. Next, we developed an online tool to visualise data and generate reports to assist decision-making under the R Shiny framework. Currently, the app can generate descriptive analytics using population variables, such as age, sex and cancer type; cancer incidence in region-level geographical heatmaps; line plots to visualise temporal trends and typical risk factor plots. The results provide a successful case study where the tool was applied to the cancer register of the Lleida region. The study illustrates how researchers and cancer registers can use the app to analyse cancer databases. Furthermore, the results highlight the analytics related to risk factors and second tumours. This paper aims to show a successful methodology for exploiting the data in the population cancers register and propose guidelines for other similar records to develop similar tools. We intend to inspire other entities to build an app that can help decision-making and also make data more accessible and transparent for the community of users.

Original Paper

Exploring Cancer Incidence, Risk Factors, and Mortality in the Lleida Region: Interactive, Open-source R Shiny Application for Cancer Data Analysis

Didac Florensa[1,2,3], PhD; Jordi Mateo-Fornes[1], PhD; Sergi Lopez Sorribes[1], MD; Anna Torres Tuca[1], MD; Francesc Solsona[1], Prof Dr; Pere Godoy[2,3,4], Prof Dr

[1]Department of Computer Engineering, University of Lleida, Lleida, Spain

[2]Population-based Cancer Registry, Santa Maria University Hospital, Lleida, Spain

[3]Field Epidemiology Unit, Lleida Biomedical Research Institute, Lleida, Spain

[4]CIBER Epidemiology and Public Health (CIBERESP), Health Institute Carlos III, Madrid, Spain

Corresponding Author:
Didac Florensa, PhD
Department of Computer Engineering
University of Lleida
C/ Jaume II, 69
Lleida, 25002
Spain
Phone: 34 603534021
Email: didac.florensa@gencat.cat

Abstract

Background: The cancer incidence rate is essential to public health surveillance. The analysis of this information allows authorities to know the cancer situation in their regions, especially to determine cancer patterns, monitor cancer trends, and help prioritize the allocation of health resource.

Objective: This study aimed to present the design and implementation of an R Shiny application to assist cancer registries conduct rapid descriptive and predictive analytics in a user-friendly, intuitive, portable, and scalable way. Moreover, we wanted to describe the design and implementation road map to inspire other population registries to exploit their data sets and develop similar tools and models.

Methods: The first step was to consolidate the data into the population registry cancer database. These data were cross validated by ASEDAT software, checked later, and reviewed by experts. Next, we developed an online tool to visualize the data and generate reports to assist decision-making under the R Shiny framework. Currently, the application can generate descriptive analytics using population variables, such as age, sex, and cancer type; cancer incidence in region-level geographical heat maps; line plots to visualize temporal trends; and typical risk factor plots. The application also showed descriptive plots about cancer mortality in the Lleida region. This web platform was built as a microservices cloud platform. The web back end consists of an application programming interface and a database, which NodeJS and MongoDB have implemented. All these parts were encapsulated and deployed by Docker and Docker Compose.

Results: The results provide a successful case study in which the tool was applied to the cancer registry of the Lleida region. The study illustrates how researchers and cancer registries can use the application to analyze cancer databases. Furthermore, the results highlight the analytics related to risk factors, second tumors, and cancer mortality. The application shows the incidence and evolution of each cancer during a specific period for gender, age groups, and cancer location, among other functionalities. The risk factors view permitted us to detect that approximately 60% of cancer patients were diagnosed with excess weight at diagnosis. Regarding mortality, the application showed that lung cancer registered the highest number of deaths for both genders. Breast cancer was the lethal cancer in women. Finally, a customization guide was included as a result of this implementation to deploy the architecture presented.

Conclusions: This paper aimed to document a successful methodology for exploiting the data in population cancer registries and propose guidelines for other similar records to develop similar tools. We intend to inspire other entities to build an application that can help decision-making and make data more accessible and transparent for the community of users.

(JMIR Cancer 2023;9:e44695) doi: 10.2196/44695

KEYWORDS

R Shiny; cloud computing; microservices; Docker; decision support system; cancer incidence; cancer risk factors, cancer mortality

Introduction

Cancer morbidity and mortality are increasing worldwide despite the development of new prevention strategies and screening programs. This increase can be attributed to several factors, including population growth, aging, and changes in lifestyle and environmental factors. The authors of [1] estimated that the global number of cancer patients (incidence rate) will increase over the coming years due to negative lifestyle and demographic changes related to population aging and growth.

The cancer incidence rate is essential for public health surveillance [2]. The incidence rate approximates the average risk of developing cancer, allowing geographic comparisons of the disease risk in different populations. This calculation requires a population-based cancer registry (PBCR) to record, store, and organize all the cancer cases in a reference region. This is achieved by a continuous process of systematic collection, storage, analysis, interpretation, and reporting of data on the occurrence and characteristics of cancer cases [3].

Over recent decades, there has been an exponential growth in PBCRs. The first volume of the Cancer Incidence in Five Continents (CI5), published in 1966, contained information from 32 registries in 29 countries, whereas the latest volume, published in 2021, included information from 343 PBCR in 65 countries.

Several data sources are integrated into PBCRs, including hospitals, death certificates, and laboratory services. Moreover, PBCRs follow international procedures, ensuring high-quality and reliable data. These goals are accomplished by performing exhaustive (automatic and manual) validity checks [4].

PBCRs are commonly used in epidemiological research. Thus, they have a crucial role in providing extensive information about tumor histology, stage at diagnosis, place and nature of the treatment, and survival [5]. Descriptive studies use registry databases to examine differences in incidence, survival, and prevalence of risk factors or comorbidities (obesity, tobacco consumption, or diabetes) across populations and their context (such as variables associated with time, place, sex, ethnicity, and social status) [6,7].

The data sets and databases stored in PBCRs grow year on year. Data visualization is essential for exploring and communicating findings in medical research, especially in epidemiological surveillance. Hence, there is an intrinsic need for rapid raw data visualization. The current situation and context (historical data) can be understood by navigating among descriptive analyses, and, before executing time-consuming predictive or prescriptive models, it is essential to generate alarms and accurate predictions or discover hidden trends or patterns.

Previous literature has described the research of the implementation of web platforms to analyze data information related to cancer. Petrov and Alexeyenko [8] implemented an application to explore molecular features and responses to anticancer drugs. Deng et al [9] presented another web application implemented on R Shiny that permitted the analysis of molecular cancer gene data sets. The user can analyze outcomes from individual genes and cancer entities. A similar application was designed by Yang et al [10]. It also analyzed and provided information on cancer gene isoform expression. Finally, another application about cancer genes was presented by Dwivedi et al [11]. In this case, it was used to perform a survival analysis on single-cell RNA sequencing data. A study by van de Water et al [12] presented a web-based tool to inform patients about esophagogastric cancer treatment options and their outcomes. These kinds of web applications can also be linked to a trained prediction tool, as demonstrated by Xu et al [13]. They developed a sexually transmitted infection prediction tool. Therefore, the literature has focused on cancer genes, cancer treatments, or other diseases, but few applications are based on epidemiological cancer data. In addition, our system is entirely adaptable to other PBCRs.

Currently, PBCRs expend resources and time to extract, analyze, and present the data to gain insight into the incidence, mortality, and survival rates for cancer. Moreover, these insights are generated manually.

One approach to solving this limitation is to develop a generic platform based on microservices for PBCRs capable of generating interactive plots, tables, and statistics to determine the epidemiological cancer situation. To address this challenge, in this paper, we propose a platform capable of (1) navigation across time and feature-based data, (2) plotting aggregated and disaggregated data on demand, and (3) automatic integration of new data.

The core activities of the PBCR have expanded beyond the provision of data to perform epidemiological research or the provision of cancer reports and statistics for a region. The data in PBCRs are the basis for estimating the cancer burden and its trends over time and are crucial in the scheduling and evaluation of cancer control programs in the registration area. One of the simplest ways of tackling this problem is to use segregated information to convince authorities about which population segments need more or different attention. For instance, geographical heat maps can be used to spot differences across urban or rural areas, while age pyramids can highlight age group differences. This can help authorities to invest and generate personalized prevention campaigns.

In summary, in this article, we propose a seed to develop this platform. The main contributions are the presentation of a successful case study for Lleida PBCR and guidelines to evolve these into a reference that can be adopted by the community. The platform was designed to be differentiated by end user. One end user is the PBCR professional who analyzes the incidence of cancer in a specific region and makes decisions to research

or prevent cancer. Another end user is the nonprofessional user who wants to know the cancer situation in his or her area.

The paper is structured as follows. The next section presents the methodology involved in designing and implementing the web platform. The Results section describes the different views implemented in this application and how the customization works. The presented data visualizations are related to cancer incidence, risk factors, and mortality. Finally, the results are discussed in the Discussion section, which also includes our conclusions.

Methods

The application is based on the model-view-controller pattern. For the visual part, we used the open-source programming language R [14] in conjunction with RStudio [15], an open-source integrated desktop environment for R. The database was created by MongoDB [16], an open-source, nonrelational database, and based on document store database, where documents are grouped into collections according to their structure. To communicate these systems and obtain the information, we implemented an application programming interface (API). Finally, to encapsulate this system and facilitate the deployment, we ran it into Docker containers that Docker Compose orchestrated [17]. Docker permits encapsulating and deploying the execution of applications in packages. All these technologies are free of charge. The deployment and code are available to download in this GitHub repository [18].

Workflow

Until the implementation of this application, PBCR professionals were manually extracting the data on demand. Once the cases were received, they cleaned and prepared the tables and plots to analyze them. Finally, they added these results to a formal report sent to public health officials.

However, once the application has been deployed, the professionals can automatically present the data to public health officials. The data extraction and cleaning steps are done by an extract, transform, and load system deployed in a server; therefore, they do not need to spend time preparing the data. In addition, the application permits real-time comparison of cancer cases between the previous years. The following subsections show how the web application has been designed and implemented.

Front-end Service

The front end was implemented using the Shiny [19] package from the R programming language, making it easy to build interactive web applications. Shiny allows R users to create interactive web applications without extensive knowledge of web design. It also permits standalone applications to be hosted on a web page and extends the application with CSS themes, html widgets, and Javascript actions.

All the plots were made using the plotly library [20], which is defined as an interactive, open-source, browser-based graphing library. It contains over 30 types of plots, including scientific charts, statistical charts, 3D graphs, and more. The tables were made using DataTable [21], defined as a plug-in for the jQuery Javascript library, which enabled the building of interactive and flexible tables. The map was made with the GeoJSON package [22]. It is a format for encoding a variety of geographic data structures and uses a geographic coordinate reference system. It also permits a specific zone and highlighted part of this map to be represented by a palette of colors.

Back-end Service

The back end consisted of an API and a database for the web application. Both these services were encapsulated using the Docker system, which permits scalability to other infrastructures. The API established the communication between the database and the view. This system was implemented by NodeJS [23], which can be described as an open-source environment based on the JavaScript programming language. This technology has increased exponentially over the last few years because it is based on asynchronous tasks, which permit executing calls without the need to wait for a response from the previous one. In addition, this uses a single threaded model with an event loop and is based on JSON format. The database implementation was based on a nonrelational database using the MongoDB system [16,24]. It saves the information through documents that are grouped into collections. This database permits large volumes of constantly changing structured, semistructured, and unstructured data. Nonrelational databases are designed by dynamic schemes to insert data without a specific structure as the relational databases specify. Therefore, it makes it easy to make significant changes to applications in real time without service interruptions.

Docker and Docker Compose

The front-end and back-end technologies were encapsulated into Docker containers. Docker is a platform designed to build, share, and run modern applications into containers [17] where the applications are virtualized and executed. The main purpose of these containers is to implement some processes and applications separately to take advantage of the infrastructure simultaneously. The way Docker is designed is to give a quick and lightweight environment where code can run efficiently. Docker contains 4 main internal components: Docker client and server, Docker images, Docker registries, and Docker containers [25].

These containers were defined using Docker Compose, which orchestrated all of them. It composes a set of components, each of which is an image and a set of options that specify what the component should have. It uses a configuration file where the user selects the parameters, and when it is executed, it runs the needed processes to build the Docker container. The user can reuse the same image for different components, and these images will be managed in other containers once instantiated [26].

Data

The case data were extracted from the official Cancer Population Registry in Lleida and the Mortality Registry of Catalonia. Experts from the cancer registry previously validated these cases to ensure the validity of the tumor. In the case of mortality, the included individuals were those patients who died from cancer in the Lleida region. The cancer patients were complemented with their risk factors, extracted from the clinical history records

at the time of diagnosis. This information permitted us to build the databases and show them in the visual part.

The database was structured into 3 collections: Patients, Tumors, and Mortality. The Patients collection included sociodemographic information and risk factors; the Tumors collection included such information as the diagnosis and the kind of tumor. Finally, the Mortality collection registered sociodemographic information and cause of death (tumor list). Table 1 specifies the variables in each collection.

Table 1. Database collections and their variables.

Variables	Specification
Patients	
sex	Gender (man/woman)
data_naix	Date of birth (date)
postal_code	Postal code of city residence (number)
postal_desc	Name of city residence (characters)
comarca	Specific region in Lleida (characters)
comarca_desc	Specific region description in Lleida (characters)
alcoholism	Alcohol consumption (yes/no)
diabetes	Diabetes diagnosed (yes/no)
smoking	Smoking consumption (yes/no)
bmi	Body mass index (number)
Tumors	
data_inc_pobl	Diagnoses date (date)
ltum	Tumor location (characters)
ltum_desc	Tumor location description (characters)
morf	Tumor morphology (characters)
morf_descr	Tumor morphology description (characters)
metode_dx	Diagnostic method (number)
metode_dx_descr	Diagnostic method description (characters)
Mortality	
data_naix	Date of birth (date)
data_def	Date of death (date)
cause10	Death cause (characters)
cause10_desc	Death cause description (characters)
sex	Gender (man/woman)
comarca	Specific region in Lleida (characters)
comarca_desc	Specific region description in Lleida (characters)
yeard	Year of death (number)

Ethical Considerations

All data were anonymized to protect patient privacy and confidentiality. The study was part of the public health response to the impact of cancer on the society. It was approved by the Committee of Ethics and Clinical Research of Lleida (CEIC). As it was a retrospective cohort study and the patients were blinded to the investigators, no written informed consent was necessary according to the CEIC. All methods were carried out in accordance with relevant guidelines and regulations.

Results

This web application consisted of an intuitive analytical web platform for rapid analysis of the population cancer registry data set, containing incidence, mortality, and risk factors related to tumor information. The application shows the incidence and evolution of each cancer during a specific period for gender and age groups. It also permits knowledge of the situation of all the cancers in a particular period and subregion in Lleida. The application also summarizes patients' risk factors detected in the cancer registry and shows results about cancer mortality.

These plots enable the number of cases to be analyzed for each year, filtered by tumor location, gender, and age group.

Cancer Incidence

The web application was designed as a web browser–based dashboard (see Figure 1) to show the information according to what the user specifies in the filters. The users can filter by years between 2012 and 2016, gender, age group, and population. This last filter can show only residents of Lleida or all cases diagnosed in the reference hospitals. Below the input filters, 3 boxes show the numbers of men and women and the average age of the patients. If the user decides to filter by men, the women box will be hidden, and the average age box will be calculated only for men. Next, the bar plot represents the number of cases diagnosed by the tumor location. The pyramid age plot helps the user analyze which age group registered the most diagnosed cases among men and women. These plots can be recalculated for all the filter inputs. Next to the pyramid age plot, the display shows the evolution of the incidence for the available years, and it allows analysis of the change in men, women, or a specific age group, depending on the chosen filters. At the end, a table with the number of diagnosed cases by tumor location is displayed and can be updated using all the filters.

Figure 2 shows a view for analyzing the incidence in the Lleida region. Specifically, it permits observation of diagnosed cases by year and cancer for specific subregions in Lleida, as the filter header represents. The view is also designed as a dashboard to enable user interaction. First, a heat map of the Lleida region is implemented. It shows the cancer incidence (per 100,000 habitants) for each area, where the color represents the incidence value. The view also offers analysis of this incidence in a bar plot (see the blue button in the map box). On the right, it shows a table with the number of cases and incidence for each area represented in the map information. These 2 elements are updated by year and the kind of cancer the user chooses in the filter. Below them, there is an evolution plot of the number of cancer cases registered. This plot is only recalculated when the user chooses a different cancer, and the year filter does not affect it. Finally, the age pyramid plot is represented, and it can be calculated by cancer and year.

Figure 1. Main menu of the web application.

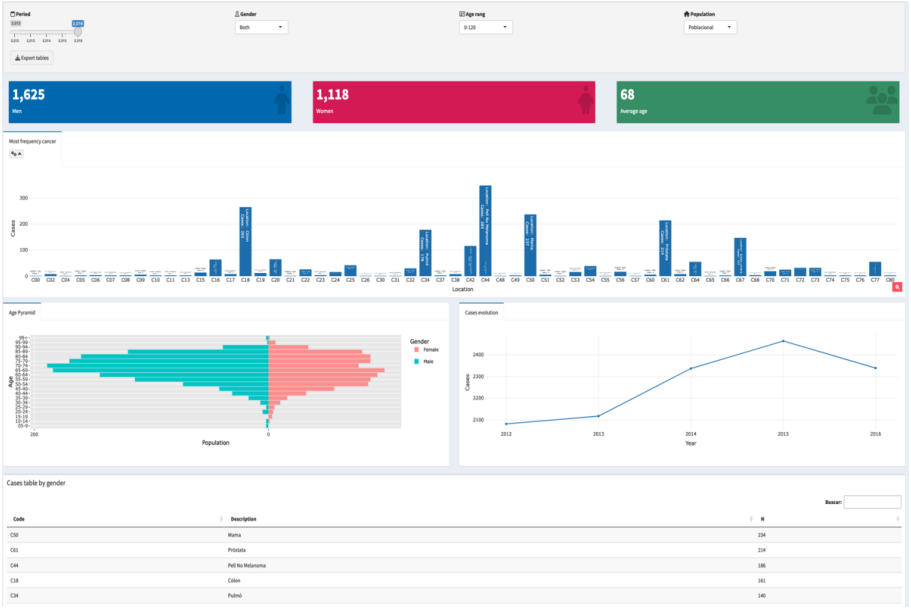
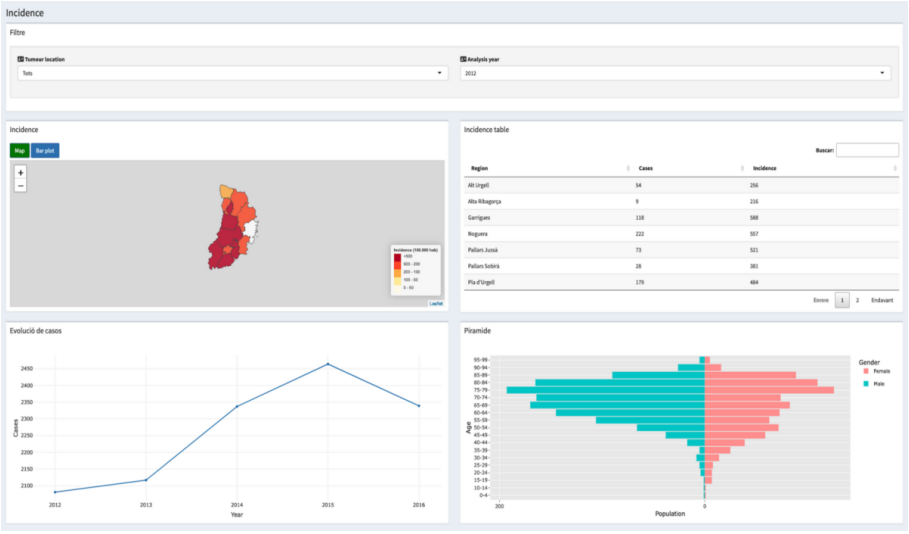


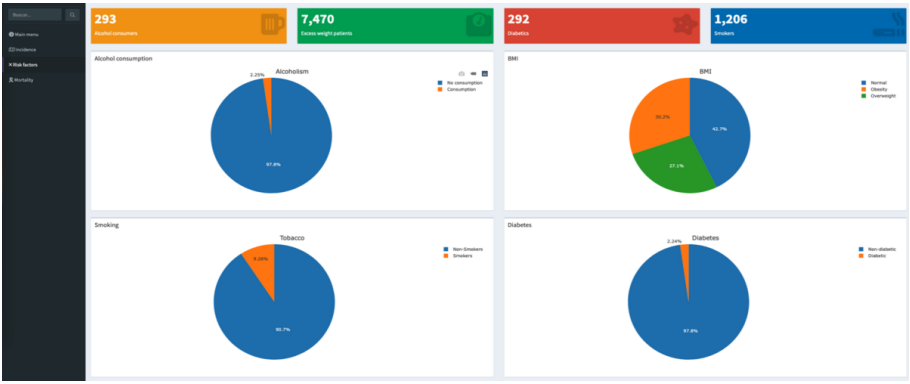
Figure 2. Specific incidence view.



Cancer Risk Factors

This view permits the risk factors' impact on cancer patients to be analyzed. Figure 3 shows 4 value boxes with the number of cases for each risk factor. First, it shows the number of patients exposed to alcohol consumption before a cancer diagnosis. Next, the number of patients with excess weight (overweight or obese) and the number of patients diagnosed with diabetes before tumor registration are shown. Finally, the number of smokers among all those who were registered is shown. Below the value box, 4 pie charts were designed to compare the exposure to these risk factors. First, alcohol risk was represented, and only 2.2% (293/13,030) of the patients were exposed. On the right, body mass index was defined; overweight affected 27.1% (3532/13,030) of the patients, and obesity affected 30.2% (3938/13,030) of the patients. At the bottom, smoking was reported for 9.3% (1212/13,030) of patients, and diabetes was reported for 2.2% (292/13,030) of patients.

Figure 3. Risk factors view.



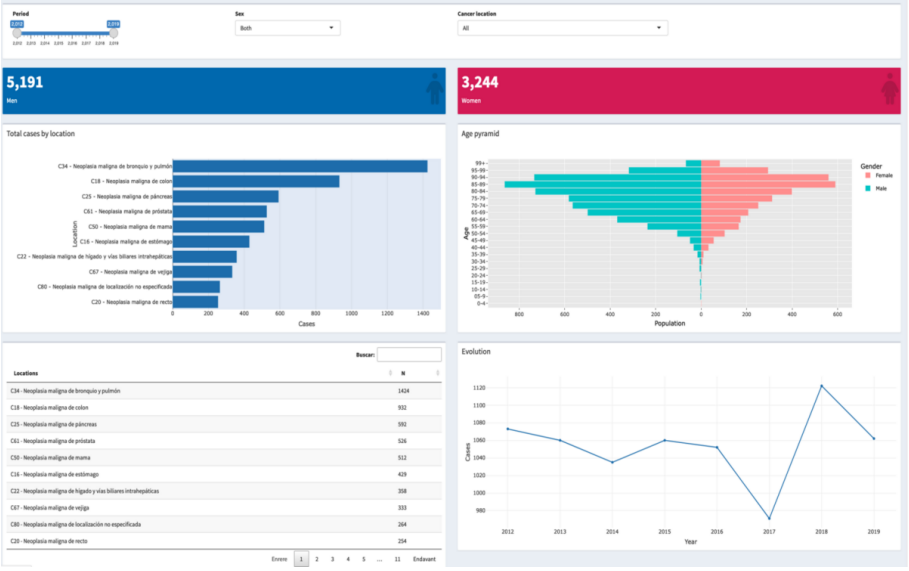
Cancer Mortality

The last implemented view shows an analysis of Lleida residents affected by tumors. In this case, the observed years were between 2012 and 2019 because the Mortality Register of Catalonia was already available for this time. Therefore, as Figure 4 shows, the filter box enables filtering by a period of years or by only 1 year. It permits showing the information by only men or women and by specific tumor location. Below the filter box, the user sees 2 value boxes representing the number of men and women who passed away among the chosen years

and by tumor location. When a specific gender is selected, the other is hidden, making visible the value box chosen in the filter.

This view also contains 4 figures, 3 plots, and 1 table. At the top left, there is a horizontal bar plot representing the 10 tumors with the most cases of mortality. It is recalculated by the period and gender chosen; the filtered cancer location does not affect it. On the right, an age pyramid plot analyzes the mortality in each age group by gender. This plot can also be recalculated by the period in years and by cancer location. At the bottom, a table has the tumor locations and the number of patients who passed away, sorted in descending order. The information is displayed by the chosen period of years and gender; the cancer location filter will not affect it. Finally, an evolution plot is calculated to analyze the increase or decrease in deaths for all locations or specific tumors. This plot is recalculated depending on the chosen year, gender, or tumor location.

Figure 4. Mortality view.



Customization

The research team designed the system for easy deployment. Therefore, the users only need to consider these items:

- Deploy the Mongo database by executing the docker-compose file. The system will download the Mongo image (if it is the first time it runs), build the Docker Container, and deploy the database. Finally, add the information to show in the dashboard web application.
- Download the web application project and specify the user and password in the config.js file. Next, execute the docker-compose file to build the containers for the API system and R Shiny application. The system will download the image to make these containers if it is the first time and then deploy the containers.

Discussion

Principal Findings

The research team designed and implemented a web application to rapidly analyze the cancer situation in the Lleida region. It contains information about the incidence of each cancer by subregion, related risk factors, and the cancer mortality registered in this region. The application can be used in computer and mobile browsers because it has been designed responsively. It has been implemented using open-source technologies such as Docker, MongoDB, NodeJS, and R Shiny, which permit easy deployment of cancer registries in other hospitals. The code is also free to download and can be deployed within 1 day.

Recently, new applications have been designed to facilitate the analysis of data sets. Some studies have suggested that the latest technologies can help to extract information and value of the data rapidly and obtain the results instantly in different contexts. Luz et al [27] designed an application called RadarR to analyze infection management. They described an accessible web application to analyze infection and antimicrobial stewardship information. Another study implemented a Shiny application for automatically coding text responses [28]. They offer an application in which users can add text to train a model to analyze this added information. For completely different information but with the same technologies, Möller et al [29] presented an R Shiny application for the visualization and extraction of phenological windows in Germany. As the literature shows, these kinds of applications are increasing for all themes as well as cancer. Miller and Shalhout [30] designed

and implemented an application to generate anatomical visualizations of cancer lesions. They concluded that data visualizations of the characteristics of clinical tumors could help to understand the natural history of malignancies. Therefore, this interactive data visualization application could permit analysis of the tumor characteristics. Another R Shiny application related to cancer data was published by Zhang et al [31]. The researchers designed a platform to analyze cell line responses to an anticancer drug. They concluded that it helped researchers understand the response of tumor cell lines to 15 therapeutic agents. Finally, a similar platform was implemented by Xia et al [32]. This platform visualizes cancer risk factors and mortality [32]. They shared a data warehouse and R Shiny application to improve their understanding of spatial and temporal trends across the population served by the University of Kansas Cancer Center.

This system helped the research team rapidly analyze the cancer information and reach some conclusions about the data and the use of these technologies. Therefore, regarding cancer incidence, the analysis detected that the number of cases is higher in men than in women in all periods and years [33]. Regarding age, the average age was 67 years, considering both genders. Men aged 65 years to 79 years registered a significant number of cases. However, cases for women occurred more often between 65 years and 69 years of age and between 75 years and 84 years of age [34]. Additional observable information was that the most common were cancers of the colon, lung, breast, prostate, and bladder [33,34]. Finally, an evolution of the incidence in Lleida showed an increase in the cases until 2015. The specific cancer incidence view also gave important information about some regions in Lleida. We observed that some areas, considered more urban than rural, had a higher incidence of some kinds of cancer, such as colon or lung [35,36].

As the incidence showed, the risk factors view also provided the previous situation of patients with cancer. Regarding risky drinking, 2.2% of the patients diagnosed consumed high amounts of alcohol daily [37]. The same percentage, 2.2%, of patients had diabetes. However, smokers represented 9.3% of the patients, one of the highest risk factors related to cancer [38]. Finally, the percentage with excess weight was high (57.3%), and some studies have pointed out that excess weight is significantly associated with the risk of cancer [39]. These results, including the number of cases for each risk factor, were obtained by the implementation of this application, which also helps to understand the cancer situation better, as other research teams have done before [32,40].

The cancer mortality registry permitted us to analyze the severity and impact of this disease, considered the second cause of death globally [41]. As we showed previously, analysts need tools like our web application offers. The application indicated that more men than women died between 2012 and 2019 [42], which might be related to the number of observed cases of cancer diagnosed among men and women [33]. The application also permitted us to know that lung cancer was the most lethal cancer among men [43] and breast cancer was the most lethal cancer in women [44]. Regarding age, the age group of 85 years to 89 years registered the highest number of deaths in both genders. Finally, we observed a general decrease in cancer deaths until 2018, when the number of patients passing away increased significantly. In case a user wanted to analyze a specific cancer location, the web platform recalculates the plots and tables for this variable.

The application presents some strengths and limitations that should be noted. This kind of implementation increases the data's potential and adds value to the cancer registries. It permits an analysis and comparison of cancer information trends in specific areas in real time and helps make decisions about public health and the impact of cancer. The risk factor situation among cancer patients suggests some associations between risk factors and cancer. The scalability of the technologies used helps to deploy them to other cancer registries. Regarding limitations, the map plot has to be adapted to the region where it is deployed. The inconsistency between the cancer registry and cancer mortality did not permit them to be merged and analyzed in depth. The codification of some risk factors suggested underdiagnosis. A future systematic link between the cancer registry and the primary care medical records could improve the registry of risk factors. Related to the software, R Shiny presented some restrictions and incompatibility with some new libraries even though they were supplied with others that are accepted and adapted perfectly. MongoDB, in the beginning, requires extra effort to understand how it works, which delayed other parts of the application.

Conclusions

The web application discussed in this study offers an analytical model of population cancer information. In addition, the technologies used to build this system permit its deployment into other cancer registries. Although there are web applications based on similar technologies, none use population cancer registry data to show the cancer situation in a specific region.

The views presented in the platform show the incidence of cancer detected in a specific time and particular areas, allowing it to be filtered by such inputs as year, gender, and tumor location. It also shows the evolution of cancer in the years analyzed. In addition, it studies the impact of some risk factors among the patients in the registry. Finally, it permits users to explore cancer mortality and its evolution in the Lleida region, filtering by year, gender, and tumor location.

Regarding future work, the research team is designing new views to analyze cancer incidence and the impact of the second primary tumor in depth. They are also creating a new risk factor view to offer a filter to give the risk factors for specific gender and tumor locations and integrating treatment data, such as for radiotherapy and chemotherapy. Finally, new web views are being created to build machine learning algorithms, train models, and analyze the results.

Acknowledgments

This work was supported by contract 2019-DI-43 from the Industrial Doctorate Program of the Government of Catalonia and by the Spanish Ministry of Economy and Competitiveness under contract PID2020-113614RB-C22. Some of the authors are members of the research group 2014-SGR163, funded by the Generalitat de Catalunya.

The authors wish to thank to the Arnau de Vilanova University Hospital, Santa Maria University Hospital, and the Catalan Health Service in Lleida for the support and resources to conduct this study.

Data Availability

The data set is available from the corresponding author upon reasonable request.

Conflicts of Interest

None declared.

References

1. Soerjomataram I, Bray F. Planning for tomorrow: global cancer incidence and the role of prevention 2020-2070. Nat Rev Clin Oncol 2021 Oct 02;18(10):663-672. [doi: 10.1038/s41571-021-00514-z] [Medline: 34079102]
2. Piñeros M, Saraiya M, Baussano I, Bonjour M, Chao A, Bray F. The role and utility of population-based cancer registries in cervical cancer surveillance and control. Prev Med 2021 Mar;144:106237 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106237] [Medline: 33678223]
3. Redondo-Sánchez D, Rodríguez-Barranco M, Ameijide A, Alonso FJ, Fernández-Navarro P, Jiménez-Moleón JJ, et al. Cancer incidence estimation from mortality data: a validation study within a population-based cancer registry. Popul Health Metr 2021 Mar 23;19(1):18 [FREE Full text] [doi: 10.1186/s12963-021-00248-1] [Medline: 33757540]
4. Bray F, Parkin DM. Evaluation of data quality in the cancer registry: principles and methods. Part I: comparability, validity and timeliness. Eur J Cancer 2009 Mar;45(5):747-755. [doi: 10.1016/j.ejca.2008.11.032] [Medline: 19117750]
5. Piñeros M, Znaor A, Mery L, Bray F. A global cancer surveillance framework within noncommunicable disease surveillance: making the case for population-based cancer registries. Epidemiol Rev 2017 Jan 01;39(1):161-169. [doi: 10.1093/epirev/mxx003] [Medline: 28472440]
6. Sung H, Siegel R, Rosenberg P, Jemal A. Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. The Lancet Public Health 2019 Mar;4(3):e137-e147 [FREE Full text] [doi: 10.1016/s2468-2667(18)30267-6]
7. Tucker TC, Durbin EB, McDowell JK, Huang B. Unlocking the potential of population-based cancer registries. Cancer 2019 Nov 01;125(21):3729-3737 [FREE Full text] [doi: 10.1002/cncr.32355] [Medline: 31381143]
8. Petrov I, Alexeyenko A. EviCor: interactive web platform for exploration of molecular features and response to anti-cancer drugs. J Mol Biol 2022 Jun 15;434(11):167528 [FREE Full text] [doi: 10.1016/j.jmb.2022.167528] [Medline: 35662462]
9. Deng M, Brägelmann J, Schultze JL, Perner S. Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. BMC Bioinformatics 2016 Feb 06;17(1):72 [FREE Full text] [doi: 10.1186/s12859-016-0917-9] [Medline: 26852330]
10. Yang IS, Son H, Kim S, Kim S. ISOexpresso: a web-based platform for isoform-level expression analysis in human cancer. BMC Genomics 2016 Aug 12;17(1):631 [FREE Full text] [doi: 10.1186/s12864-016-2852-6] [Medline: 27519173]
11. Dwivedi B, Mumme H, Satpathy S, Bhasin SS, Bhasin M. Survival Genie, a web platform for survival analysis across pediatric and adult cancers. Sci Rep 2022 Feb 23;12(1):3069 [FREE Full text] [doi: 10.1038/s41598-022-06841-0] [Medline: 35197510]
12. van de Water LF, van den Boorn HG, Hoxha F, Henselmans I, Calff MM, Sprangers MAG, et al. Informing patients with esophagogastric cancer about treatment outcomes by using a web-based tool and training: development and evaluation study. J Med Internet Res 2021 Aug 27;23(8):e27824 [FREE Full text] [doi: 10.2196/27824] [Medline: 34448703]
13. Xu X, Yu Z, Ge Z, Chow EPF, Bao Y, Ong JJ, et al. Web-based risk prediction tool for an individual's risk of HIV and sexually transmitted infections using machine learning algorithms: development and external validation study. J Med Internet Res 2022 Aug 25;24(8):e37850 [FREE Full text] [doi: 10.2196/37850] [Medline: 36006685]
14. The R Project for Statistical Computing. The R Foundation. URL: https://www.r-project.org/ [accessed 2023-03-25]
15. Posit Software. URL: https://posit.co/ [accessed 2023-03-25]
16. MongoDB. URL: https://www.mongodb.com/ [accessed 2023-03-25]
17. Docker. URL: https://www.docker.com/ [accessed 2023-03-25]
18. didacflorensa / CancerRegistryPlatform. GitHub. URL: https://github.com/didacflorensa/CancerRegistryPlatform [accessed 2023-03-25]
19. Shiny. URL: https://shiny.rstudio.com/ [accessed 2023-03-25]
20. Plotly. URL: https://plotly.com/ [accessed 2023-03-25]
21. DataTables. URL: https://datatables.net/ [accessed 2023-03-25]

22. GeoJSON. URL: https://geojson.org/ [accessed 2023-03-25]
23. Node.js. URL: https://nodejs.org/en [accessed 2023-03-25]
24. Gyorödi C, Gyorödi R, Sotoc R. A comparative study of relational and non-relational database models in a web-based application. International Journal of Advanced Computer Science and Applications 2015;6(11):1. [doi: 10.14569/IJACSA.2015.061111]
25. Rad B, Bhatti HJ, Ahmadi M. An introduction to Docker and analysis of its performance. International Journal of Computer Science and Network Security 2017:228-235 [FREE Full text]
26. Ibrahim MH, Sayagh M, Hassan AE. A study of how Docker Compose is used to compose multi-component systems. Empir Software Eng 2021 Sep 23;26(6):1. [doi: 10.1007/s10664-021-10025-1]
27. Luz CF, Berends MS, Dik JH, Lokate M, Pulcini C, Glasner C, et al. Rapid analysis of diagnostic and antimicrobial patterns in R (RadaR): interactive open-source software app for infection management and antimicrobial stewardship. J Med Internet Res 2019 May 24;21(6):e12843 [FREE Full text] [doi: 10.2196/12843] [Medline: 31199325]
28. Andersen N, Zehner F. shinyReCoR: a Shiny application for automatically coding text responses using R. Psych 2021 Aug 16;3(3):422-446. [doi: 10.3390/psych3030030]
29. Möller M, Boutarfa L, Strassemeyer J. PhenoWin – an R Shiny application for visualization and extraction of phenological windows in Germany. Computers and Electronics in Agriculture 2020 Aug;175:105534. [doi: 10.1016/j.compag.2020.105534]
30. Miller DM, Shalhout SZ. BodyMapR: an R package and Shiny application designed to generate anatomical visualizations of cancer lesions. JAMIA Open 2022 Apr;5(1):ooac013 [FREE Full text] [doi: 10.1093/jamiaopen/ooac013] [Medline: 35274087]
31. Zhang P, Palmisano A, Kumar R, Li MC, Doroshow JH, Zhao Y. TPWshiny: an interactive R/Shiny app to explore cell line transcriptional responses to anti-cancer drugs. Bioinformatics 2022 Jan 03;38(2):570-572. [doi: 10.1093/bioinformatics/btab619] [Medline: 34450618]
32. Xia Q, Mudaranthakam DP, Chollet-Hinton L, Chen R, Krebill H, Kuo H, et al. shinyOPTIK, a user-friendly R Shiny application for visualizing cancer risk factors and mortality across the University of Kansas Cancer Center catchment area. JCO Clinical Cancer Informatics 2022 May(6):1. [doi: 10.1200/cci.21.00118]
33. Ferlay J, Colombet M, Soerjomataram I, Mathers C, Parkin D, Piñeros M, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int J Cancer 2019 Apr 15;144(8):1941-1953 [FREE Full text] [doi: 10.1002/ijc.31937] [Medline: 30350310]
34. Sánchez MJ, Payer T, De Angelis R, Larrañaga N, Capocaccia R, Martinez C, CIBERESP Working Group. Cancer incidence and mortality in Spain: estimates and projections for the period 1981-2012. Ann Oncol 2010 May;21 Suppl 3:iii30-iii36 [FREE Full text] [doi: 10.1093/annonc/mdq090] [Medline: 20427358]
35. Florensa D, Godoy P, Mateo J, Solsona F, Pedrol T, Mesas M, et al. The use of multiple correspondence analysis to explore associations between categories of qualitative variables and cancer incidence. IEEE J. Biomed. Health Inform 2021 Sep;25(9):3659-3667. [doi: 10.1109/jbhi.2021.3073605]
36. Munker R, Midis G, Owen-Schaub L, Andreff M. Soluble FAS (CD95) is not elevated in the serum of patients with myeloid leukemias, myeloproliferative and myelodysplastic syndromes. Leukemia 1996 Sep;10(9):1531-1533. [Medline: 8751476]
37. Larsson SC, Carter P, Kar S, Vithayathil M, Mason AM, Michaëlsson K, et al. Smoking, alcohol consumption, and cancer: A mendelian randomisation study in UK Biobank and international genetic consortia participants. PLoS Med 2020 Jul 23;17(7):e1003178 [FREE Full text] [doi: 10.1371/journal.pmed.1003178] [Medline: 32701947]
38. Gandini S, Botteri E, Iodice S, Boniol M, Lowenfels AB, Maisonneuve P, et al. Tobacco smoking and cancer: a meta-analysis. Int J Cancer 2008 Jan 01;122(1):155-164 [FREE Full text] [doi: 10.1002/ijc.23033] [Medline: 17893872]
39. Sung H, Siegel RL, Torre LA, Pearson-Stuttard J, Islami F, Fedewa SA, et al. Global patterns in excess body weight and the associated cancer burden. CA Cancer J Clin 2019 Mar 12;69(2):88-112 [FREE Full text] [doi: 10.3322/caac.21499] [Medline: 30548482]
40. Moraga P. SpatialEpiApp: A Shiny web application for the analysis of spatial and spatio-temporal disease data. Spat Spatiotemporal Epidemiol 2017 Nov;23:47-57. [doi: 10.1016/j.sste.2017.08.001] [Medline: 29108690]
41. Rahib L, Wehner MR, Matrisian LM, Nead KT. Estimated projection of US cancer incidence and death to 2040. JAMA Netw Open 2021 Apr 01;4(4):e214708 [FREE Full text] [doi: 10.1001/jamanetworkopen.2021.4708] [Medline: 33825840]
42. Torre LA, Siegel RL, Ward EM, Jemal A. Global cancer incidence and mortality rates and trends--an update. Cancer Epidemiol Biomarkers Prev 2016 Jan 14;25(1):16-27. [doi: 10.1158/1055-9965.EPI-15-0578] [Medline: 26667886]
43. Yang X, Man J, Chen H, Zhang T, Yin X, He Q, et al. Temporal trends of the lung cancer mortality attributable to smoking from 1990 to 2017: A global, regional and national analysis. Lung Cancer 2021 Feb;152:49-57. [doi: 10.1016/j.lungcan.2020.12.007] [Medline: 33348250]
44. Wojtyla C, Bertuccio P, Wojtyla A, La Vecchia C. European trends in breast cancer mortality, 1980-2017 and predictions to 2025. Eur J Cancer 2021 Jul;152:4-17 [FREE Full text] [doi: 10.1016/j.ejca.2021.04.026] [Medline: 34062485]

Abbreviations

API: application programming interface

CEIC: Committee of Ethics and Clinical Research of Lleida
CI5: Cancer Incidence in Five Continents
PBCR: population-based cancer registry

Edited by A Mavragani; submitted 30.11.22; peer-reviewed by CM Moore, N Jiwani; comments to author 29.01.23; revised version received 13.02.23; accepted 07.03.23; published 09.04.23

Please cite as:
Florensa D, Mateo-Fornes J, Lopez Sorribes S, Torres Tuca A, Solsona F, Godoy P
Exploring Cancer Incidence, Risk Factors, and Mortality in the Lleida Region: Interactive, Open-source R Shiny Application for Cancer Data Analysis
JMIR Cancer 2023;9:e44695
URL: https://cancer.jmir.org/2023/1/e44695
doi: 10.2196/44695
PMID:

4.2 PAPER II: MCA TO CANCER INCIDENCE PATTERNS

Authors: *Didac Florensa, Jordi Mateo, Francesc Solsona, Tere Pedrol, Miquel Mesas and Ramon Piñol and Pere Godoy*
Journal: IEEE Journal of Biomedical and Health Informatics.
Publisher: IEEE
Year: 2021
DOI: https://doi.org/10.1109/JBHI.2021.3073605
ISSN: 2168-2194
Keywords: *Cancer, Cancer registry, Multiple Correspondence Analysis, Rural, Urban*

The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables and Cancer Incidence

Abstract: Previous works have shown that risk factors for some kinds of cancer depend on people's lifestyle (e.g. rural or urban residence). This article looks into this, seeking relationships between cancer, age group, gender and population in the region of Lleida (Catalonia, Spain) using Multiple Correspondence Analysis (MCA). The dataset analysed was made up of 3,408 cancer episodes between 2012 and 2014, extracted from the Population-based Cancer Registry (PCR) for Lleida province. The cancers studied were colon and rectal (1,059 cases), lung (551 cases), urinary bladder (446 cases), prostate (609 cases) and breast (743 cases). The MCA technique was applied and used to search relationships among the main qualitative features. The basic statistics were the percentage explaining (variance), the inertia and the contribution of each qualitative variable. General outcomes showed a low and moderate contribution of living in rural areas to colorectal and male prostate cancer. Males in urban areas were slightly and heavily affected by lung and urinary bladder cancer respectively. The analysis of each cancer provided additional information. Colorectal cancer greatly affected males aged <60, urban residents aged 70-79, and rural females aged ≥ 80. The impact of lung cancer was high among urban females <60, moderate among males aged 70-79 and high among rural females aged ≥ 80. The results for urinary bladder cancer results were similar to those for lung cancer. Prostate cancer affected both the <60 and ≥ 80 age groups significantly in rural areas. Breast cancer hit the 70-79 group significantly and, somewhat less so, rural females aged ≥ 80. MCA was a significant help for detecting the contributions of qualitative variables and the associations between them. MCA has proven to be an effective technique for analyzing the incidence of cancer. The outcomes obtained help to corroborate suspected trends, as well as detecting and stimulating new hypotheses about the risk factors associated with a specific area and cancer. These findings will be helpful for encouraging new studies and prevention campaigns to highlight observed singularities.

The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables and Cancer Incidence

Dídac Florensa, Pere Godoy, Jordi Mateo, Francesc Solsona, Tere Pedrol, Miquel Mesas, and Ramon Piñol

***Abstract*—Background: Previous works have shown that risk factors for some kinds of cancer depend on people's lifestyle (e.g. rural or urban residence). This article looks into this, seeking relationships between cancer, age group, gender and population in the region of Lleida (Catalonia, Spain) using Multiple Correspondence Analysis (MCA). Methods: The dataset analysed was made up of 3408 cancer episodes between 2012 and 2014, extracted from the Population-based Cancer Registry (PCR) for Lleida province. The cancers studied were colon and rectal (1059 cases), lung (551 cases), urinary bladder (446 cases), prostate (609 cases) and breast (743 cases). The MCA technique was applied and used to search relationships among the main qualitative features. The basic statistics were the percentage explaining (variance), the inertia and the contribution of each qualitative variable. Results: General outcomes showed a low and moderate contribution of living in rural areas to colorectal and male prostate cancer. Males in urban areas were slightly and heavily affected by lung and urinary bladder cancer respectively. The analysis of each cancer provided additional information. Colorectal cancer greatly affected males aged <60, urban residents aged 70–79, and rural females aged ≥ 80. The impact of lung cancer was high among urban females <60, moderate among males aged 70–79 and high among rural females aged ≥ 80. The results for urinary bladder cancer results were similar to those for lung cancer. Prostate cancer affected both the <60 and ≥ 80 age groups significantly in rural areas. Breast cancer hit the 70–79 group significantly and, somewhat less so, rural females aged ≥ 80. Conclusions: MCA was a significant help for detecting the contributions of qualitative variables and the associations between them. MCA has proven to be an effective technique for analyzing the incidence of cancer. The outcomes obtained help to corroborate suspected trends, as well as detecting and stimulating new hypotheses about the risk factors associated with a specific area and cancer. These findings will be helpful for encouraging new studies and prevention campaigns to highlight observed singularities.**

***Index Terms*—Cancer, cancer registry, multiple correspondence analysis, rural, urban.**

I. Background

CANCER is the second leading cause of death globally. Between 30–50% of cancers can currently be prevented by avoiding risk factors and implementing existing evidence-based prevention strategies. The continuous rise of this disease over recent decades is attributed to the impact of aging among an increasingly elderly population [1].

Cancer recording is considered a key factor in controlling the disease [2]. The purpose of the registers is to detect and fully record all cases of cancer diagnosed among the residents of the reference area [2], [3]. There are three population-based cancer registers (PCR) in Catalonia (Spain), these being the PCRs of the provinces of Lleida, Girona and Tarragona [4]. Barcelona is the fourth province. However, it has no PCR. These records indicate the existence of territorial differences that would need to be studied. Recent studies suggest differences in the incidence of cancer, temporal trends, and mortality among urban and rural areas which are attributable to exposure to different risk factors, access to screening programs, and regular diagnosis and treatment [5]. Specifically, the population of the Lleida region presents life styles, risk factors and work activity which can be traduced to a specific incidence for certain types of cancer. Nearly half of the population of Lleida province live in rural areas. As a consequence, their lifestyle is different from that of the more urban populations in other Catalan provinces. A peculiarity of this region is the work environment. In rural areas,

Manuscript received June 29, 2020; revised November 4, 2020, December 17, 2020, and March 17, 2021; accepted April 11, 2021. Date of publication April 15, 2021; date of current version September 3, 2021. This work was supported in part by the Industrial Doctorate Program of the Government of Catalonia under Contract 2019-DI-43 in part by the Ministerio de Economía y Competitividad under Contract TIN2017-84553-C2-2-R, and in part by the Generalitat de Catalunya (some of the authors are members of the research group 2014-SGR163). *(Corresponding author: Didac Florensa.)*

Dídac Florensa, Jordi Mateo, and Francesc Solsona are with the Department of Computer Science, University of Lleida, 25001 Lleida, Spain (e-mail: didac.florensa@gencat.cat; jordi.mateo@udl.cat; francesc.solsona@udl.cat).

Pere Godoy is with Epidemiology Service, Department of Health, 25006 Lleida, Spain, and also with CIBER Epidemiology and Public Health (CIBERESP), 25006 Lleida, Spain (e-mail: pere.godoy@gencat.cat).

Tere Pedrol is with the Health Department, Population-Based Cancer Registry in Lleida, 25006 Lleida, Spain (e-mail: mtpedrol.ics.lleida@gencat.cat).

Miquel Mesas is with the Department Computer, Santa Maria University Hospital, 25198 Lleida, Spain (e-mail: mmesas@gss.cat).

Ramon Piñol is with the Department of Health, Catalan Health Service, 08023 Barcelona, Spain (e-mail: rpinol@catsalut.cat).

Digital Object Identifier 10.1109/JBHI.2021.3073605

the main activity is the agri-food industry and, in urban areas, it is service sector activities such as education, health and catering.

In the literature, there are several reports that present the incidence of cancer in rural and urban areas. In 1992, the University of North Carolina presented a rural-urban pattern study of cancer mortality [6] and explained differences in its incidence among rural versus urban populations. It concluded that cancer is diagnosed at more advanced and more disseminated stages of the disease in rural populations because they are typically older, less educated, poorer and have less access to such health care services as early-cancer detection. Potential explanations were given for lower overall incidence rates in rural areas compared with urban zones. These include smoking (more prevalent in urban areas) and exposure to environmental pollutants. Whitney E Zahnd *et al* ([5]) presented a report about rural-urban differences in cancer incidence and trends in the United States. The study analyzed age-adjusted incidence rates, ratios and annual percentage change (APC) for all cancers detected between 2009 and 2013. Concretely, this report concludes that cancer rates associated with modifiable risks-tobacco, human papillomavirus, and some preventive screening modalities (e.g., colorectal and cervical cancers)- were higher in rural settings compared with urban populations. Next, the work in [7] concluded that, although cigarette smoking is the primary cause of lung cancer, there are other risk factors which may differ by geographic region. These include passive smoking, exposure to indoor radon and asbestos. Finally, [8] present a work investigating urban-rural variations in the incidence of several cancers after adjusting them for socioeconomic status. This interesting article concluded that the risk of some cancers varied with area and gender. For example, the risk of prostate cancer was higher in rural areas and as was that of breast cancer in females in urban areas.

Recently, the PCR team in Lleida presented a descriptive-analytic study highlighting the preliminary results of the impact and incidence of cancer in urban and rural areas [9]. The article compared the number of cancer cases between rural and urban areas according to the crude data rates from the Catalan Population-based Cancer Registry. Tumour ranking and rates obtained in the different areas of the province of Lleida suggested that some cancers have particular features that should be investigated. Jointly with this incipient work, the related literature [5]–[8] has led us to study the incidence of major cancers by rural and urban areas. Many efforts have been made to measure the incidence of cancer by using such traditional methods as density rates, annual rates or the Spearman rank correlation coefficient. However, none of them has provided enough evidence of a relation between cancer and population.

To address these limitations, this paper proposes studying the differences between urban and rural populations. The method proposed is the application of Multiple Correspondence Analysis (MCA) to stimulate new hypotheses and relations between the characteristics of the patients and the incidence of cancer in the province of Lleida.

As the main contribution of this study, we propose the use of MCA as a statistical technique to search for associations between the registered data for cancer in the province of Lleida (Catalonia). This province has a good balance between rural and urban populations and the dataset is mainly made up of categorical variables. In [10], the authors asserted that MCA helped them to classify the degree of tumor regression with a categorical dataset. This led us to assess our challenge with the same statistics using MCA. The outcomes obtained demonstrate the usefulness of this technique in this kind of data analysis, made up exclusively of categorical variables.

II. Methods

The Population-based Cancer Registry (PCR) of the health region of the province of Lleida (HRPLL) was the basis for this descriptive epidemiological study into cancer. The main information sources were hospital records (ICD-9 codes-140.0 to 208.9) and reports from pathological anatomy. Before extracting the information for this study, the cases were reviewed and validated using ASEDAT[1]. Then, an accurate description of the data and basic concepts of the MCA statistical technique used in this work are explained in this section.

A. Data

Lleida is the largest province in Catalonia with a population density of 36 people per square kilometre. More specifically, the population was 438,001 in 2014 [12], 221 891 men and 216,110 women. Approximately half of the population lives in rural areas. In accordance with [13], people living in cities with a population of more than 10 000 are classified as "urban" in this study and the rest as "rural".[2] In 2014, the respective urban and rural populations were 199,300 and 238,701. Thus, this is a well-balanced dataset for studying differences between urban and rural populations in the risk-factors for cancer.

The data are made up of the information registered between 2012-2014 in the Lleida PCR [14], [15] for cancer patients in the main hospitals in the health region of the province of Lleida. These are the Arnau de Vilanova University Hospital (HUAV) and the Santa Maria University Hospital (HUSM). The study is GDPR[3]-compliant, maintaining the anonymity of the patients. Cancer episodes were recorded according to international criteria. These go from the case definition to the operation system and the final results obtained in order to ensure reliability, the validation of the data and comparison with other hospital registers.

The initial dataset consisted of 3,423 new cancer diagnoses in the HRPLL during 2012-2014. After applying data cleaning by using the Box Plot technique to discard statistical outliers, the data collection became 3408 cancer diagnosis (See box plot graphs in the Github repositoy [17]). These box plots graphs are based on each cancer and by age and population and gender. As Figure 1 in the annex shows, this allows outliers for colorectal cancer by age and gender to be detected. And so

[1]ASEDAT: Software of the Catalan Institute of Oncology to select, extract and validate cancer data [11]

[2]The Spanish National Statistics Institute (Spanish initials: INE) has defined rural areas as those with fewer than 2,000 inhabitants; semi-urban areas as those with between 2001 and 10,000 inhabitants; and urban areas as those with more than 10,000 inhabitants.

[3]GDPR: General Data Protection Regulation (EU)

on in the rest of the graphs. This technique uses the median, approximate quartiles, and the lowest and highest data points to convey the level, spread, and symmetry of a distribution of data values [16]. In addition, all the scripts implemented for data cleaning (done with Python) and data analysis (done with R) can be freely downloaded from this Github repository [17]. All the data provided in this link were generated randomly.

Each register contains the following fields: **age group** (<60, 60–69, 70–79, ≥ 80); **gender** (male, female); **population** (rural, urban) and **cancer** type. Only the five most frequent types of cancer were analysed (see incidence tables in the Github repository [17]). These being colon and rectal (1059), urinary bladder (446), breast (743), prostate (609) and lung (551). Gender was divided between males (2088) and females (1319). The population was divided into rural (1821) and urban (1587). Finally, age was divided into 4 balanced intervals: <60 years old with 834 cases, 60 to 69 years old with 927 cases, 70 to 79 with 968 cases and aged ≥ 80 with 679 cases.

B. Statistics

All the information presented was analyzed using Multiple Correspondence Analysis (MCA), an extension of Correspondence Analysis (CA). MCA is an unsupervised learning algorithm for visualizing the patterns in large tables and for multi-dimensional categorical data [18]. This method can be used to describe, explore, summarize and visualize information contained on individuals described by categorical variables within a data table [19]. Unlike CA, MCA can deal with more than one categorical variable. This is the main advantage of the MCA technique. In our case, MCA was firstly used to evaluate the relationships between the four features. It was then used to evaluate the relationships between population, age and gender for each cancer. Associations between features were represented graphically [10]. The graphs aim to visualize the similarities and/or differences in the profiles simultaneously, identifying those dimensions that contain most of the data variability. Features or their categories close to each other are significantly related statistically.

The factors produced were interpreted with the help of various statistical coefficients which complemented each other to provide a better interpretation. The most common and important are inertia, the eigenvalue and the contribution and factorial coordinates. Inertia is a measurement of the dispersion of the set of computed distances between points. Analogously, in Principal Correspondence Analysis (PCA), inertia corresponds to the explained variance of dimensions. The eigenvalue allows the inertia that a specific category produces to be quantified. The contribution enables us to consider how much influence a category has in determining a certain percentage relative to the entire set of the active category. The percentage coordinates (x- and y-axis) of the graph enable the category points in a graph to be represented and established. In MCA, the distance between two or more categories of different variables can be interpreted in terms of the associations and correlations between these. If two categories present high coordinates and are close in space, this means that they tend to be directly associated. If two categories present high coordinates but are distant from each other (e.g. they have opposite signs), this means that they tend to be inversely associated. If two categories present the same coordinate sign, they can be related to each other [20], [21].

Thus, the graphic depiction aims to visualize the similarities and/or differences in the profiles simultaneously, identifying those dimensions that contain most of the data variability. These MCA representations can be read like those from a PCA: the coordinates of a product are its values for the common factors; the coordinates of a variable are its correlation with these factors [22]. Categories depicted in the same direction on the dimension will be significantly related statistically and have patterns of relative frequencies. This association is also valuable statistically when the points are located far from the origin of the graph, representing a mean, uninformative profile [23], [24].

In this study, the MCA method was applied in scripts performed with R [25], an open-source programming language and environment for statistical computing and graphics. It provides a wide variety of statistical and graphical techniques, and is highly extensible. Specifically, the main library used to implement the methods and obtain the results was FactoMineR [26].

III. Results

In this section, we first present a general analysis of the results obtained from applying the MCA technique to the dataset presented in section II. Then, a similar analysis was applied to each cancer to evaluate these in isolation. It is important to clarify the differences between contribution and correlation to understand and interpret the results and figures presented in this section. The contribution is used to denote which variables explain better the variations in the data set and are most important in the construction of the axes. In contrast, correlation represents the relation between two variables or, in other words, the degree of influence of one variable compared with the other.

A. Multiple Correspondence Analysis for All Cancers

The variance obtained was 20.5% (eigenvalue: 0.46) for dimension 1 (x-axis) and 12.7% (eigenvalue: 0.285) for the second one (y-axis). The inertia (sum of the variances) for these two dimensions was 33.2%. Age variance scored 0.259 and 0.582 in dimensions 1 (x-axis) and 2 (y-axis) respectively. Cancer was 0.809 and 0.443, gender 0.765 and 0.007, and population 0.008 and 0.107. The variable that gave the worst results for percentage explanation was population.

Similar results were obtained when discarding the gender variable. The variances in this case were 16.3% (eigenvalue: 0.433) for dimension 1 and 14.1% (eigenvalue: 0.376) for dimension 2, and an inertia for these two dimensions of 30.4%. The percentages of variances explained for population were 0.062 and 0.035.

Removing the age variable, the variances were 28.4% (eigenvalue: 0.568) and 17.3% (eigenvalue: 0.346) for dimensions 1 and 2 respectively. Thus, the inertia for these dimensions was 45.7%. This was the two-dimension combination (the dimensions are ranked with the variance) that gave the highest inertia.

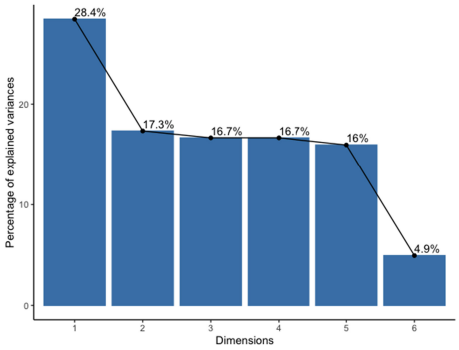


Fig. 1. Percentage of explained variances of the overall dimensions.

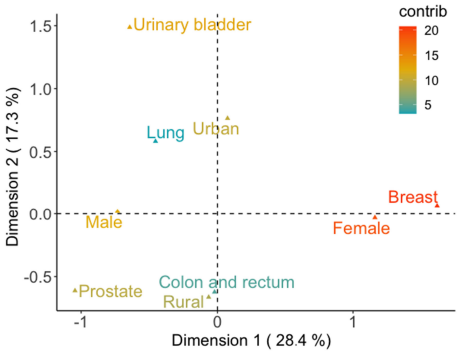


Fig. 2. Two-dimensional MCA plot. Correlations between the variables.

Each variable variance usually increases with the inertia. Fortunately, in this case, the population variances (0.004 and 0.506) improved significantly. This was also the best combination for population, the main goal of the present work.

Fig. 1 shows the variances of the overall dimensions (6) for the combinations of variables obtained. Note that the dimensions are ranked in descending order. It can be seen that dimensions 1 and 2 have variances of 28.4% and 17.3% respectively. The sum of the variances of the overall dimensions is 100%. In this figure, the main idea was to show the percentage of explained variance in every dimension and not the influence of all the variables.

Fig. 2 presents the results of the MCA algorithm in a two-dimensional plot (x- and y-axis representing dimensions 1 and 2 respectively) that shows the correlations between the variables. A two-dimensional plot gives more information about correlations between variables than higher dimensional ones. Thus, no higher dimensional-results were presented. It can be seen that colorectal cancer and rural are very close and appear in the negative y-axis (dimension 2). This means that they are

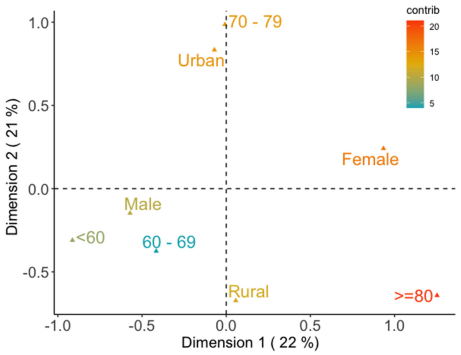


Fig. 3. **Colorectal** cancer. The positive and negative x-axis (representing gender variable) depicts females and males. The positive and negative y-axis (representing population variable) depicts urban and rural.

correlated. Lung and urinary bladder cancers appear on the positive y-axis (dimension 2) where urban contributes and on the negative x-axis (dimension 1) where males appear. This suggests that these cancers are correlated with urban males. Moreover, prostate appears on the negative y-axis (dimension 2) meaning that it is significant in rural areas. Finally, the only breast cancer is correlated with females, with the same contribution in both areas.

The general outcomes showed a low contribution for colorectal cancer in rural areas. The lowest contributions are depicted in blue in the ranking. No differences between gender are observed, due to the location of coordinate 0 on the x-axis. Prostate cancer (its ranked color is located in the middle of the key) affected males in rural areas moderately. A low affection of lung cancer was observed in urban males. Urinary bladder cancer affected urban dwellers severely, mostly males. Finally, as expected, breast cancer heavily affected females independently of the area.

B. Multiple Correspondence Analysis by Cancer

This section presents the MCA results for colorectal, lung, urinary bladder, prostate and breast cancers.

The first cancer studied was **colorectal** (Fig. 3). The variance obtained for the first dimension was 22% (eigenvalue: 0.366) and 21% for the second dimensions (eigenvalue: 0.349). The total inertia was 43%. The correlation between the population variable and dimensions was 0.003 on the first and 0.56 on the second. The gender correlation was 0.533 (dimension 1) and 0.035 (dimension 2), and the age group correlation was 0.561 (dimension 1) and 0.453 (dimension. 2). The urban population was represented on the positive y-axis (29.55% of the total category contributions in dimension 2) and the rural on the negative y-axis (23.85% of the total category contributions in dimension 2). Gender is represented on the x-axis (dimension 1). The female contribution was 30.13% on the positive x-axis and

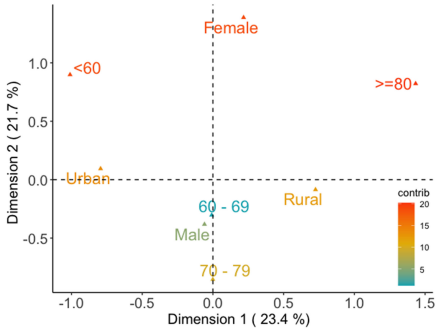


Fig. 4. **Lung cancer**. The positive and negative x-axis (representing population variable) depicts rural and urban respectively. The positive and negative y-axis (representing gender variable) depicts females and males respectively.

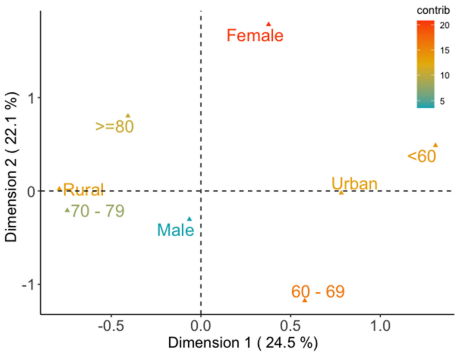


Fig. 5. **Urinary bladder** cancer. The positive and negative x-axis (representing population variable) depicts urban and rural respectively. The positive and negative y-axis (representing gender variable) depicts females and males respectively.

the male contribution was 18.43% on the negative (dimension 1). The group aged ≥ 80 contributed 32.62% in dimension 1 and 8.96% in dimension 2. The contribution of the 70-79 group was 0.001% on the former and 28.85% on the later. The contributions of those aged 60-69 were 4.31% and 3.67%, and the group aged <60 contributed 14.12% and 1.70%.

The 70-79 age group was closely related to the urban population, regardless of gender. This is a very significant result because it has an important y-axis component (close to 1). In the <60 age band, it affected males slightly more. The most important result was the ≥ 80 age group, where there is a higher incidence among women. The incidence among the 60-69 age group was not significant but mainly affected men.

Fig. 4 shows the results obtained for **lung** cancer. The variance for dimension 1 was 23.4% (eigenvalue: 0.390) and 21.7% for dimension 2 (eigenvalue: 0.362), so the total inertia was 45.1%. In this study, the population correlation was 0.576 on dimension 1 and 0.008 on dimension 2, the gender correlation was 0.012 and 0.530, and finally, the age group was 0.581 and 0.548. Regarding the categories variables, urban areas contributed on the negative x-axis (dimension 1) with 25.71% and the rural with 23.48% on the positive x-axis. In the case of gender, the male contribution was 10.53% on the negative y-axis (dimension 2) and the female contribution was 38.24% on the positive y-axis. The ≥ 80 age group contributed 29.28% in the first dimension and 10.37% in the second. The contributions of the 70-79 age group were 0.01% and 20.21%, then 0.019 and 2.51% for the 60-69 age group and the <60 age group contributed 20.40% and 17.38%.

The high position of urban females <60 shows that the contribution of this relation was very high. A moderate significance can be observed for males aged 70-79. Finally, the contribution for rural females aged ≥ 80 reached the same significance as urban females <60.

In **urinary bladder** cancer (Fig. 5), the variance for the first dimension was 24.5% (eigenvalue: 0.407) and 22.1% (eigenvalue: 0.367) for the second, and in consequence, the total inertia was 46.6%. The population correlation was 0.617 on dimension 1 and 0.0004 on dimension 2, the gender correlations were 0.024 and 0.542 and those for the age group were 0.581 and 0.559. The urban category contributed 25.13% on the negative x-axis and the rural on the positive x-axis with 25.36% (dimension 1). The male category contribution was 7.16% on the negative y-axis and that of female was 42.01%. The ≥ 80 age group contributed with 4.12% in dimension 1 and 17.86% in dimension 2, the 70-79 age group contributed 13.74% and 1.2%, the 60-69 age group with 6.13% and 28.08%, and the <60 age group with 23.50% and 3.61%.

The results for urinary bladder cancer were similar to those for lung cancer. Specifically, females aged <60 contributed moderately in urban areas (23.50% in dimension 1 and 3.61% in dimension 2). In the 60-69 age cohort, it affected men in urban areas moderately but this incidence decreased among those aged 70-79 living in rural areas (13.74 in dimension 1 and 1.20% in dimension 2). The contribution of men between 60-69 was 6.13% in dimension 1 and 28.08% in dimension 2. In the ≥ 80 age group, rural women were slightly affected (4.12 in dimension 1 and 17.86% in dimension 2).

Moderate importance was moved to urban males aged 60-69, dropping when reaching the 70-79 age group, in the rural zone. And, attenuated importance was to rural females aged ≥ 80.

Females are ruled out of **prostate** cancer (Fig. 6). The variance for the first dimension was 25% (eigenvalue: 0.5) and 25% (eigenvalue: 0.5) in the second, resulting in total inertia of 50%. The population correlation for dimension 1 was 0.527 and this was 0 for dimension 2, and the correlations for age group were 0.527 and 1. In this case, the gender variable was not included because this type of cancer only affects men. The contribution of the urban category was 28.32% on the positive x-axis and the rural contribution was 21.67% on the negative x-axis (dimension 1). The ≥ 80 age group's contribution was 27.12% in the first dimension and 30.24% in the second. The contributions of the

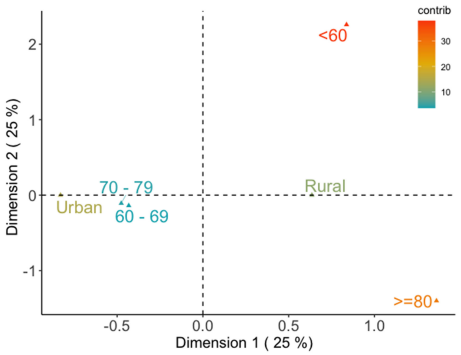


Fig. 6. **Prostate** cancer. The positive and negative x-axis (representing population variable) depicts rural and urban respectively. The y-axis has no meaning on this occasion.

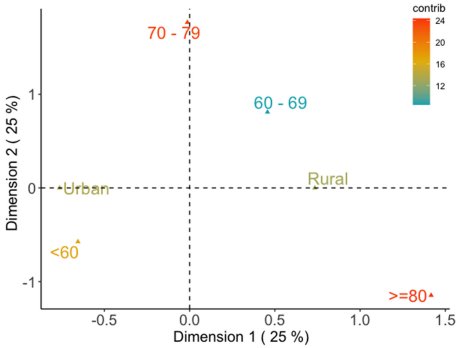


Fig. 7. **Breast** cancer. The positive and negative x-axis (representing population variable) depicts for rural and urban respectively. The y-axis has no meaning on this occasion.

70-79 age group were 7.58% and 0.413%, for 60-69 age group, 6.34% and 0.691%. Finally, the <60 age group contributed 8.94% and 68.64%.

In contrast to the previous cancers analyzed, rural males <60 suffered heavily. Hardly any contributions appeared in the 60-69 and 70-79 urban age ranges. In the absence of females, prostate cancer affected rural males aged ≥ 80 significantly.

Fig. 7 shows the results obtained when applied the MCA to **breast** cancer for females only. The variance obtained for the first dimension was 25% (eigenvalue: 0.559) and 25% (eigenvalue: 0.5) for the second (total inertia was 50%). The population correlation was 0.559 for dimension 1 and 0.0 for dimension 2, and the age group correlations were 0.559 and 1.0. As explained in subsection II-A, gender was not considered. The contribution of the urban category was 25.47% on the negative x-axis and rural contribution was 24.52% on the positive x-axis (dimension 1). The ≥ 80 age group's contribution was 27.84% in the first dimension and 20.42% in the second. The 70-79 age group contributed 0.002% and 50.25%. For the 60-69 age group, the percentages were 3.85% and 13.58%, and finally, for the <60 age group, 18.29% and 15.73%.

In the data resulting after applying the screening technique, breast cancer only presented females cases even though males can also suffer from it [27]. This cancer affects rural females aged <60 moderately. Among the 60-69 age group, urban women were hardly affected. The group which contributed the most and with a great significance was those aged 70-79, although the type of population did not influence the results. In the ≥ 80 group there was, as usual, a high incidence among rural females.

Figs 8 and 9 show the contributions obtained for all categories of cancer in the first and second dimensions, respectively. The figures enable the categories that contribute significantly to be detected and interpreted. They also allow the associations to be detected by the contribution in the same dimension. As they show, the x-axis represents each cancer and the y-axis, the contribution. The categories are represented in each stacked bar. For example, in Fig. 8, colorectal cancer, the ≥ 80 age group suggests an association between gender because it presents a higher contribution than the others groups. However, in Fig. 9 and for the same cancer, the 70-79 age group suggests an association with the population.

IV. Discussion

The MCA technique enables the analysis and detection of new relations between categories not observed in the literature. It helped to detect that prostate and colorectal cancer have a high incidence in rural areas. Another important finding was the correlation between lung and urinary bladder cancer and urban areas.

During the period presented (2012-2014), the PCR of Lleida registered approximately 6,000 cases of all possible types of cancer. In this study, only the types most frequently diagnosed cancers (see incidence tables in the Github repository [17]) in the region were selected (colorectal, lung, urinary bladder, prostate, breast cancer). These covered a total of 3408 cases. New relationships were found through applying MCA to detect relations with the features used in a health region with a good balance between urban and rural populations. We based this on a preliminary study [28] which concludes that the incidence of some cancers depends more on the area. However, it does not search for relationships geographic areas and cancers. Another starting point is the work presented in [8]. This studied the most important cancers and their relationship with rural and urban areas. Their most important findings were that the incidence of prostate cancer was most significant in rural areas and that of breast cancer, in urban settings. These articles studied the urban-rural incidence but did not use the MCA technique to explore associations between categories of qualitative variables as we do. To understand the application of MCA, we based ourselves on a study [29] about healthy ageing, and one [30] which concluded that bad driving and crashes could be affected

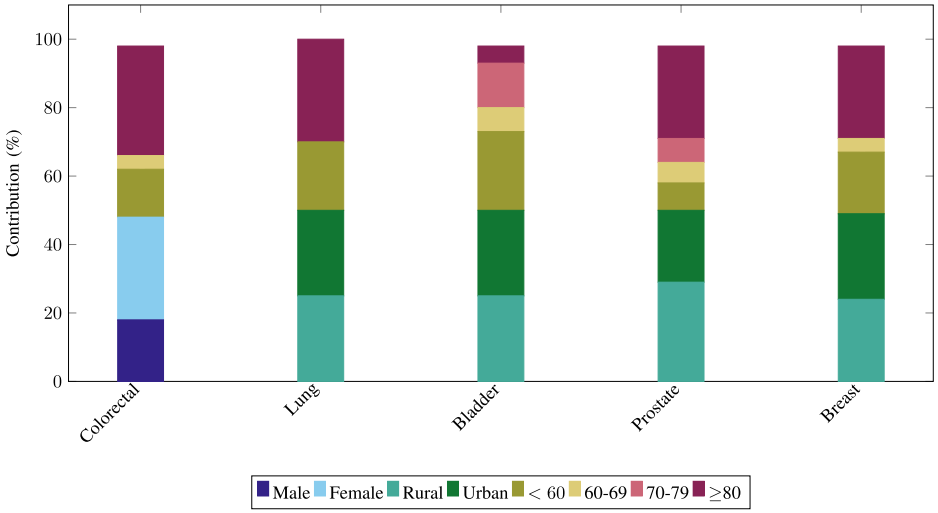


Fig. 8. Contributions by cancer and categories in dimension 1.

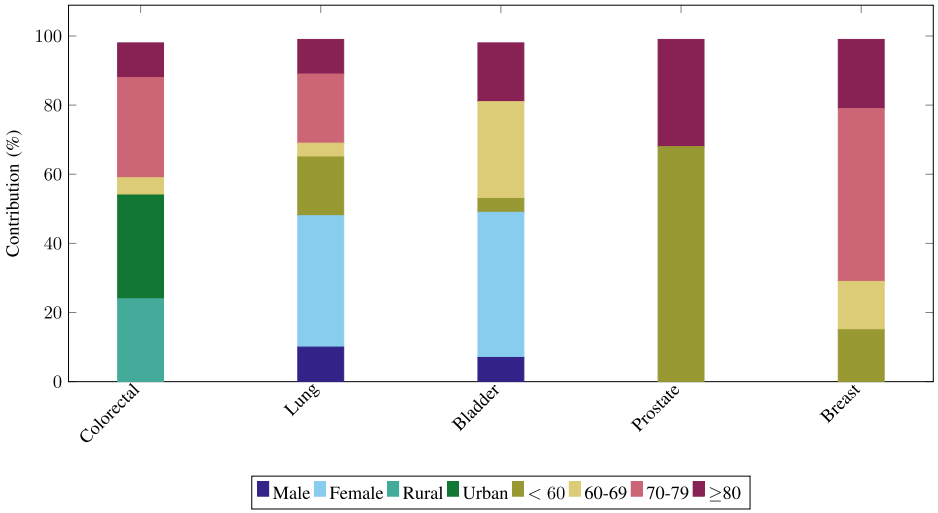


Fig. 9. Contributions by cancer and categories in dimension 2.

by differences between urban and rural areas, traffic volume, driver age and more. In addition, a previous study used MCA to analyse the prognosis in surgery for low rectal cancer [10]. However, to the best of our knowledge no prior studies have used MCA to link types of cancers to rural or urban areas.

Firstly, all the cancers were analyzed together with MCA. The total inertia was 32.9% and the population variance obtained was close to 0 in both dimensions (0.008 in dimension 1 and 0.107 in dimension 2), meaning that this variable combination performed poorly in associating the dataset centered on the population. Next, discarding the gender variable, the total inertia worsened (30.4%), as did the explained population variances (0.062 and 0.035). On also removing the age variable, the population variances improved significantly, to 0.004 and 0.506. This was the best result obtained with the population variable. However, these good results were at the expense of discarding such an important feature as age group.

Some important outcomes were found. These include the lower incidence of colorectal cancer (for either gender) and the moderate rate of prostate cancers among men in rural areas. Males were also more significantly affected by lung and urinary bladder cancer in urban areas. As expected, breast cancer had a high incidence among females. This suggest new hypotheses to deepen and study these specific cancers.

The analysis then studied each cancer separately. Among the population aged <60, **colorectal** cancer affects males severely. On reaching the age of 70-79, this shifted to the urban population. Significant outcomes were obtained in rural females aged ≥ 80. This can be related to the greater age of females in the rural population [31]. These associations with rural populations suggest a high incidence in rural areas. A similar rural incidence was obtained in the study into metropolitan and non-metropolitan areas in the United States [28].

In **lung** cancer, the goodness of the results obtained can be contrasted with human behaviour and genetics. First, the migration of young people influences this in both urban and rural areas. This is seen in females aged <60 in urban areas, contrasting with rural zones, where it tends to affect those aged ≥ 80. In both cases, the red in the picture shows that this associating contribution is very high. These results corroborate the findings of the work presented in [32]. Furthermore, this cancer affects males in the 70-79 age group slightly more. The 60-69 contribution is insignificant in any sense.

The results for **urinary bladder** cancer were similar to those for lung cancer. Urban females aged <60 and then urban males (60-69) contributed moderately, but this decreased even more for the population aged 70-79. Rural females aged ≥ 80 are hardly affected.

As expected, **prostate** cancer only affected males, with a high incidence among the rural population aged <60 and ≥ 80. In contrast the results were insignificant among the other groups in urban environments. The major incidence among those aged <60 in the rural environment is very significant and much attention should be paid to it. In this case, the significant association between rural areas and prostate cancer differs from the incidence of this cancer in other regions analyzed [28]. However, a study into the incidence of cancer in Ireland obtained outcomes that concluded that the risk was higher in rural areas [8].

Again, of course, **breast** cancer only affected females. There was a higher incidence among urban women aged <60 (as with colorectal, lung and urinary bladder cancer). Surprisingly, rural females in the 60-69 age group were hardly affected. The group which contributed the most was those aged 70-79 whatever the population. This contribution was very significant and is a clear example of a case to be studied. Again it affected rural females aged ≥ 80 heavily.

This study has some limitations that should be noted. The postal address registered for each case was where the patient lived at the moment of cancer diagnose. However, this may have changed during the study. Despite this, the number of cases with changed addresses would be very low and this factor is not expected to produce bias in the results. Some lifestyle aspects, such as tobacco and alcohol consumption, profession or other risk factors that could explain some of the differences observed, were not taken into account.

V. Conclusion

There are incipient research efforts to search for correlations between cancers and lifestyle, such as the effect of incidence from living in rural or urban environments. Research using MCA has been applied in various fields, but no one has focused on analysing relationships between cancers and urban and rural lifestyles. This was our main research aim.

Some important outcomes were found, such as the contribution of colorectal cancer (whatever the gender) and prostate cancers among men in rural areas. Also, there was a low incidence of lung cancer but high rate of bladder cancer, especially in urban areas, and the incidence of breast cancer has high in both areas. These outcomes suggest new hypotheses to deepen the study of these specific cancers.

The analysis of each cancer provided additional information. Colorectal cancer severely affected males aged <60, and those in urban areas aged 70-79, as well as women aged ≥ 80 in rural areas. Lung cancer had a high impact on urban females <60, a moderate one on males between 70 and 80 and high again among females aged ≥ 80. Similar but lower results were obtained for urinary bladder cancer. This was moderate in urban females <60 and urban males aged 60-69, decreasing for rural residents aged 70-79 and even more for rural females aged ≥ 80. Prostate cancer, as expected, only affected males. There was a high rate among the rural population aged <60, but this was lower in urban dwellers aged 60-69 and 70-79 before becoming significant again among rural men aged ≥ 80. In contrast, cases of breast cancer were only registered in females in the selected period. Whatever the area, those aged 70-79 were affected the most while the incidence among rural females aged ≥ 80, was somewhat less.

The outcomes obtained help to corroborate suspected trends in several of the relationships detected and stimulate new hypotheses about the risk factors and new techniques to analyse the incidence of cancer. They also help the public health system to focus advice on specific areas and cancers. In future work, it is important to delve deeper into each cancer in order to study its risk factors. This means using new variables, such as tumor characteristics (size, cancer stage or degree of aggressiveness),

treatments, socioeconomic ratio, environmental conditions and mortality. Also, new artificial intelligence algorithms can be explored to search for behavior patterns of cancer, unsupervised clusters or to analyze risk factors and prior patient comorbidities.

References

[1] "World Health Organization. Cancer," Accessed: Jun. 1, 2020. [Online]. Availabel: https://www.who.int/news-room/fact-sheets/detail/cancer

[2] S. Siesling *et al.*, "Uses of cancer registries for public health and clinical research in Europe: Results of the European network of cancer registries survey among 161 population-based cancer registries during 2010–2012," *Eur. J. Cancer*, vol. 51, no. 9, pp. 1039–1049, Jun. 2015.

[3] M. C. White *et al.*, "The history and use of cancer registry data by public health cancer control programs in the United States," *Cancer*, vol. 123, no. Suppl 24, pp. 4969–4976, Dec. 2017.

[4] D. de Salut, *El Càncer a Catalunya*. Monografia 2016 (In catalan), Barcelona: Dept de Salut, 2017, ch. 1, pp. 1–109.

[5] W. E. Zahnd *et al.*, "Rural-urban differences in cancer incidence and trends in the united states," *Cancer Epidemiol., Biomarkers Prevention: A Pub. Amer. Assoc. Cancer Res., cosponsored by the Amer. Soc. Prev. Oncol.*, vol. 27, no. 11, pp. 1265–1274, Nov. 2018.

[6] A. C. Monroe, T. C. Ricketts, and L. A. Savitz, "Cancer in rural versus urban populations: A review," *J. Rural Health*, vol. 8, no. 3, pp. 212–220, 1992.

[7] M. E. O'Neil, S. J. Henley, E. A. Rohan, T. D. Ellington, and M. S. Gallaway, "Lung cancer incidence in nonmetropolitan and metropolitan counties - United States, 2007–2016," *MMWR. Morbidity Mortality Weekly Rep.*, vol. 68, no. 44, pp. 993–998, 2019.

[8] L. Sharp *et al.*, "Risk of several cancers is higher in urban areas after adjusting for socioeconomic status. results from a two-country population-based study of 18 common cancers," *J. Urban Health*, vol. 91, no. 3, pp. 510–525, 2014.

[9] D. Florensa *et al.*, "El registre poblacional de càncer a lleida en zones urbanes i rurals. resultats de l'any 2014." *Butlletí Epidemiol. Catalunya*, vol. 40, no. 12, pp. 252–264, 2020.

[10] R. Mancini *et al.*, "Tumor regression grade after neoadjuvant chemoradiation and surgery for low rectal cancer evaluated by multiple correspondence analysis: Ten years as minimum follow-up," *Clin. Colorectal Cancer*, vol. 17, no. 1, pp. e13–e19, 2018.

[11] I. C. d'Oncologia, "Registre Hospitalari De Tumors ICO/CSUB," [Online]. Available: http://pdo.iconcologia.net/rht/registres.htm

[12] "Idescat, Anuari Estadístic De Catalunya. Densitat De Població. Comarques I Aran, àmbits I Províncies," 2014. [Online]. Available: https://www.idescat.cat/pub/?id=aec&n=249&t=2014

[13] J. García González, "La población rural de España, De los desequilibrios a la sostenibilidad social," *Encrucijadas - Revista Crítica de Ciencias Sociales*, vol. 6, no. 14, pp. 146–149, 2013.

[14] P. Godoy, T. Pedrol, I. Mòdol, and A. Salud, "El registre poblacional de càncer a lleida: Resultats I perspectives," *Butlletí Epidemiològic Catalunya*, vol. 37, no. 7, pp. 161–172, 2016.

[15] P. Godoy-Garcia, T. Pedrol, I. Mòdol-Pena, and A. Salud, "El registre poblacional de càncer a lleida: Resultats de l'any 2013," *Butlletí Epidemiològic Catalunya*, vol. 39, no. 1, pp. 1–11, 2018.

[16] S. K. Kwak and J. H. Kim, "Statistical data preparation: Management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, pp. 407–411, 2017.

[17] D. Florensa, P. Godoy, J. Mateo, and F. Solsona, "Github repository - A new proposal for analysing cancer in urban versus rural populations," [Online]. Available: https://github.com/didacflorensa/MCA-Cancer

[18] F. Murtagh, "Multiple correspondence analysis and related methods," *Psychometrika*, vol. 72, no. 2, pp. 275–277, 2007, doi: 10.1007/s11336-006-1579-x.

[19] F. Husson and J. Josse, *Multiple Correspondence Analysis*, Boca Raton: Chapman and Hall, Jan. 2014, ch. 11, pp. 165–184.

[20] G. D. Franco, "Multiple correspondence analysis: one only or several techniques?," *Qual. Quantity*, vol. 50, no. 3, pp. 1299–1315, 2016.

[21] C. E. Heckler, "Applied multivariate statistical analysis," *Technometrics*, vol. 47, no. 4, pp. 517–518, 2005.

[22] J. Pagès, "Multiple factor analysis: Main features and application to sensory data," *Revista Colombiana de Estadistica*, vol. 27, no. 1, pp. 1–22, 2004.

[23] M. Greenacre, *Correspondence Analysis in Practice*. Boca Raton: Chapman and Hall, ch. 6, pp. 56–62, 2017, doi: 10.1201/9781315369983.

[24] B. L. Roux and H. Rouanet, *Geometric data analysis: From correspondence analysis to structured data analysis*, Springer, 2005, ch. 1, pp. 18-24, doi: 10.1007/1-4020-2236-0.

[25] R Core Team, *R: A. Language and Environment for Statistical Computing, R. Foundation for Statistical Computing*, Vienna, Austria, 2019. [Online]. Available: https://www.R-project.org/

[26] S. Lê, J. Josse, and F. Husson, "Factominer: An R package for multivariate analysis," *J. Statist. Softw., Articles*, vol. 25, no. 1, pp. 1–18, 2008.

[27] A. J. Abdelwahab Yousef, "Male breast cancer: Epidemiology and risk factors," *Seminars Oncol.*, vol. 44, no. 4, pp. 267–272, 2017.

[28] S. Henley, R. Anderson, C.C. Thomas, G.M. Massetti and B. Peaker, "Invasive cancer incidence, 2004–2013, and deaths, 2006–2015, in non-metropolitan metropolitan counties - United States," *MMWR Surveill Summ*, vol. 66, no. SS-14, pp. 1–13, 2017.

[29] P. S. Costa, N. C. Santos, P. Cunha, J. Cotter, and N. Sousa, "The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing," *J. Aging Res.*, vol. 72, no. 2, pp. 257–284, 2013.

[30] S. Das, R. Avelar, K. Dixon, and X. Sun, "Investigation on the wrong way driving crash patterns using multiple correspondence analysis," *Accident Anal. Prevention*, vol. 111, no. 2018, pp. 43–55, 2018.

[31] L. C. M. Fernández and J. M. D. Urrecho, "Envejecimiento Y desequilibrios poblacionales en las regiones españolas con desafíos demográficos," *Ería*, vol. 37, no. 1, pp. 21–43, 2017.

[32] C. D. Viñas, "Depopulation processes in european rural areas: A case study of cantabria (Spain)," *Eur. Countryside*, vo. 11, no. 3, pp. 341–369, 2019, doi: 10.2478/euco-2019-0021.

4.3 PAPER III: MCA AND K-MEANS TO CANCER PATTERNS

Authors: *Didac Florensa, Jordi Mateo, Francesc Solsona, Tere Pedrol, Miquel Mesas and Ramon Piñol and Pere Godoy*
Journal: Journal of Medical Internet Research
Publisher: JMIR Publications
Year: 2022
DOI: https://doi.org/10.2196/29056
ISSN: 1438-8871
Keywords: *Colorectal cancer, Cancer registry, Multiple Correspondence Analysis, K-means, Risk factors*

Use of Multiple Correspondence Analysis and K-means to Explore Associations Between Risk Factors and Likelihood of Colorectal Cancer: Cross-sectional Study.

Abstract: Previous works have shown that risk factors are associated with an increased likelihood of colorectal cancer. The purpose of this study was to detect these associations in the region of Lleida (Catalonia) by using multiple correspondence analysis (MCA) and k-means. This cross-sectional study was made up of 1,083 colorectal cancer episodes between 2012 and 2015, extracted from the population-based cancer registry for the province of Lleida (Spain), the Primary Care Centers database, and the Catalan Health Service Register. The data set included risk factors such as smoking and BMI as well as sociodemographic information and tumor details. The relations between the risk factors and patient characteristics were identified using MCA and k-means. The combination of these techniques helps to detect clusters of patients with similar risk factors. Risk of death is associated with being elderly and obesity or being overweight. Stage III cancer is associated with people aged ≥ 65 years and rural/semiurban populations, while younger people were associated with stage 0. MCA and k-means were significantly useful for detecting associations between risk factors and patient characteristics. These techniques have proven to be effective tools for analyzing the incidence of some factors in colorectal cancer. The outcomes obtained help corroborate suspected trends and stimulate the use of these techniques for finding the association of risk factors with the incidence of other cancers.

Original Paper

Use of Multiple Correspondence Analysis and K-means to Explore Associations Between Risk Factors and Likelihood of Colorectal Cancer: Cross-sectional Study

Dídac Florensa[1,2], MPH; Jordi Mateo-Fornés[1], PhD; Francesc Solsona[1], PhD; Teresa Pedrol Aige[3], MPH; Miquel Mesas Julió[2], MPH; Ramon Piñol[4], BSc; Pere Godoy[5,6,7], PhD

[1]Department of Computer Science, University of Lleida, Lleida, Spain
[2]Department of Computer Systems, Santa Maria University Hospital, Lleida, Spain
[3]Hospital-based Cancer Registry, Arnau de Vilanova University Hospital, Lleida, Spain
[4]Catalan Health Service, Department of Health, Lleida, Spain
[5]Biomedical Institute Research of Lleida, Lleida, Spain
[6]Centro de Investigación Biomédica en Red, Madrid, Spain
[7]Santa Maria University Hospital, Population Cancer Registry, Lleida, Spain

Corresponding Author:
Dídac Florensa, MPH
Department of Computer Science
University of Lleida
Jaume II, 69
Lleida, 25001
Spain
Phone: 34 973 70 27 00
Email: didac.florensa@gencat.cat

Abstract

Background: Previous works have shown that risk factors are associated with an increased likelihood of colorectal cancer.

Objective: The purpose of this study was to detect these associations in the region of Lleida (Catalonia) by using multiple correspondence analysis (MCA) and k-means.

Methods: This cross-sectional study was made up of 1083 colorectal cancer episodes between 2012 and 2015, extracted from the population-based cancer registry for the province of Lleida (Spain), the Primary Care Centers database, and the Catalan Health Service Register. The data set included risk factors such as smoking and BMI as well as sociodemographic information and tumor details. The relations between the risk factors and patient characteristics were identified using MCA and k-means.

Results: The combination of these techniques helps to detect clusters of patients with similar risk factors. Risk of death is associated with being elderly and obesity or being overweight. Stage III cancer is associated with people aged ≥65 years and rural/semiurban populations, while younger people were associated with stage 0.

Conclusions: MCA and k-means were significantly useful for detecting associations between risk factors and patient characteristics. These techniques have proven to be effective tools for analyzing the incidence of some factors in colorectal cancer. The outcomes obtained help corroborate suspected trends and stimulate the use of these techniques for finding the association of risk factors with the incidence of other cancers.

(J Med Internet Res 2022;24(7):e29056) doi: 10.2196/29056

KEYWORDS

colorectal cancer; cancer registry; multiple correspondence analysis; k-means; risk factors

Introduction

Colorectal cancer is the third most common type of cancer worldwide [1,2]. In Europe, around 250,000 new colorectal cancer cases are diagnosed each year, accounting for around 9% of all malignancies. The rates of this cancer increase with industrialization and urbanization. In general, the evidence shows that the incidence increases in countries where the overall risk of large bowel cancer is low, while in countries with high incidence, the rate has either stabilized or decreased, particularly among younger age groups [3].

In the province of Lleida (Spain), the population-based cancer registries allow the identification and counting of all incident cases (new cases) diagnosed among the residents of this geographical area [4]. The residents of the Lleida region present lifestyles, risk factors, and work activity, which can be used to determine the specific incidence of certain types of cancer. Nearly half the population of the Lleida province live in rural and semiurban areas. As a consequence, their lifestyle is different from that of the more urban populations in other Catalan provinces [5,6]. Thus, they can present different risk factors and socioeconomic status (SES).

Some studies have demonstrated a higher incidence of colorectal cancer among those with low SES and risk factors such as BMI and smoking. A pooled European cohort study [7] demonstrated that adult weight gain was associated with increased risk of several major cancers. They also concluded that the degree, timing, and duration of being overweight and obesity also seemed to be important. More specifically for colon cancer, Guo et al [8] presented a prospective cohort study in northern China. They concluded that obesity increased the risk of colon cancer in males. Regarding smoking, Mizoue et al [9] presented a report evaluating the association in the Japanese population based on a systematic review of epidemiological evidence. This report concluded that tobacco smoking may increase the risk of colorectal cancer in the Japanese population. However, there is still insufficient epidemiological evidence to demonstrate any clear association with colon cancer. Kim et al [10] studied a possible association between SES and the risk of colorectal cancer in women. Their findings suggested that high SES may protect against colorectal cancer in women. The methodology used in these studies was similar, namely, the multivariate regression analysis.

Recent research has applied the techniques used in this study, but none of these studies were for cancer and risk factors. Ugurlu and Cicek [11] used the multiple correspondence analysis (MCA) method to search for relations in ship collisions [11]. However, the k-means algorithm was more widely used in some cancer aspects. Rustam et al [12] applied this technique to obtain the centroid of each cluster and predict the class of every data point in the validation set. Recently, Ronen et al [13] used k-means as an initial step in a deep learning method to evaluate the colorectal cancer subtypes. K-means allowed the detection of relevant clinical patterns that improved the prediction model. Therefore, the use of MCA and k-means to search for the relationship between risk factors and cancer incidence is a novel method.

Several studies [7-10] have found new associations among risk factors, demographic information, and SES in patients with colorectal cancer. These studies have taken a great effort to analyze and compare risk factors such as obesity, cigarette smoking, and SES in patients with colorectal cancer. They used statistical methods, including Cox regression, Spearman rank correlation coefficient, and multilevel logistic regression to estimate the association between variables. However, none of them used a combination of a statistical method like MCA and an artificial intelligence algorithm such as k-means to search for associations between a group of categorical variables.

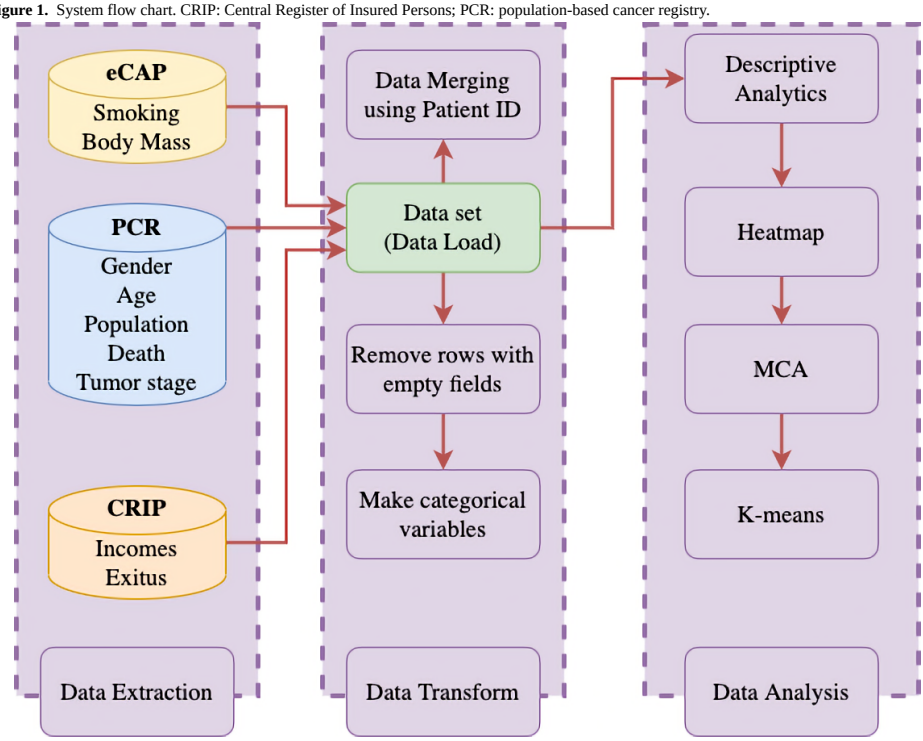
As the main contribution of this study, we propose the use of MCA as a statistical technique to detect relations between risk factors and patients' characteristics and k-means as an unsupervised learning algorithm to search for clusters of patients with similar risk factor profiles for colorectal cancer.

Methods

Preprocessing

The main information sources were the population-based cancer registry of the health region of the province of Lleida, the eCAP (a computerized medical history program used by doctors, pediatricians, and nurses in primary care centers when they see their patients [14]) software, and the Central Register of Insured Persons (a register that allows the unique identification of those covered by the Catalan Health Service through the personal identification code, the management and consultation of their data, and their updates [15]). Before applying the statistical technique, the information was validated by experienced professionals (doctors, nurses, and documentalists) in the Lleida population-based cancer registry who reviewed the clinical history of each patient. After that, the International Agency for Research on Cancer tool was applied to detect unlikely or impossible codes or combinations of codes [16]. Then, an accurate description of the data and basic concepts of the MCA and k-means used in this work are explained in this section. See the system flow chart of the whole process in Figure 1; it shows the different registers used to extract the data, its process and transformation, and its applied analysis. The patients with empty fields were removed.

Figure 1. System flow chart. CRIP: Central Register of Insured Persons; PCR: population-based cancer registry.



Study Population

The colorectal cancer data were extracted from the new cases registered between 2012 and 2015 in the Lleida population-based cancer registry [5,17,18] for patients with cancer in the main hospitals in the health care region of the Lleida province. Specifically, the data set consisted of 1083 new colorectal cancer cases. These hospitals were the Arnau de Vilanova University Hospital and the Santa Maria University Hospital, and the primary information sources were hospital records (International Classification of Diseases, ninth revision codes-140.0 to 208.9) and reports from pathological anatomy. Additionally, these reports confirmed >92% of cases included in the sample. Risk factors such as BMI and smoking were extracted from eCAP software and the SES was extracted from the Central Register of Insured Persons. The study is compliant with the General Data Protection Regulation (European Union), thereby maintaining the anonymity of the patients. Cancer episodes were recorded according to international criteria. In addition, the data analysis (done with R) can be freely downloaded from this GitHub repository [19]. It also included a mock data set randomly generated to test the models. The original data set could not be uploaded due to General Data Protection Regulation, which does not permit sharing patients' information.

The BMI was used to calculate the obesity of each patient by standard weight status categories [20]. We categorized the BMI as the established table: <24.9 as normal weight, 25-29.9 as overweight, and >30 as obesity. Regarding SES, we categorized the variable according to the annual income available from the Central Register of Insured Persons. According to the legislation [21], we created 2 groups: annual income <€18,000 (low income) and >€18,000 (high income) (€1=US $1.04). The population was categorized as rural, semiurban, and urban. In accordance with [22], people living in cities with a population of more than 10,000 were classified as urban, population between 10,000 and 2000 in towns as semiurban, and the rest as rural. The Spanish National Statistics Institute has defined rural areas as those with a population of less than 2000, semiurban areas as those with a population between 2001 and 10,000, and urban areas as those with a population with more than 10,000 people. All the cancer cases that did not conform to one of these fields were discarded automatically. To sum up, each register contains the following fields: age group (50-64 years, 65-74 years, ≥75 years); gender (male, female); population (rural, semiurban, urban); exitus (death, alive); BMI (normal, overweight, obesity); smoking (ex-smoker/smoker, nonsmoker); income (high income, low income); and stage (0, I, II, III, undefined). Table 1 shows the number of cases for each category.

Table 1. Principal comorbidities groups included in this study: patients with colorectal cancer between 2012 and 2015, where all the comorbidities were properly registered (N=1083).

Characteristics	Values, n (%)
Gender	
Male	689 (63.6)
Female	394 (36.4)
Age group (years)	
50-64	319 (29.5)
65-74	328 (30.3)
≥75	436 (40.2)
Exitus	
Death	221 (20.4)
Alive	862 (79.6)
Income[a]	
<€18,000/year	863 (79.7)
>€18,000/year	220 (20.3)
Population	
Rural	228 (21.1)
Semiurban	333 (30.7)
Urban	522 (48.2)
BMI	
Normal	234 (21.6)
Overweight	506 (46.7)
Obesity	343 (31.7)
Smoker	
Smoker/Ex-smoker	232 (21.4)
Nonsmoker	851 (78.6)
Stage	
0	64 (5.9)
I	115 (10.6)
II	168 (15.5)
III	91 (8.4)
Undefined	645 (59.6)

[a]€1=US $1.04.

MCA Algorithm

MCA is an unsupervised learning algorithm for visualizing the patterns in large and multidimensional categorical data [23]. This method can be used to analyze, explore, summarize, and visualize information contained of individuals described by categorical variables [24]. Unlike correspondence analysis (CA), MCA can deal with more than one categorical variable. This is the main advantage of the MCA technique. In our case, MCA was first used to evaluate the relationships between all the features. MCA was then used to evaluate the relationships among population, age, gender, exitus, BMI, smoking, and tumor stage. Associations between features are represented graphically [25]. The graphs aim to visualize the similarities or differences in the profiles simultaneously, identifying those dimensions that contain most of the data variability. Features or their categories close to each other are significantly related statistically.

The factors were interpreted with the help of various statistical coefficients, which complemented each other to provide a better interpretation. The most common and important are inertia, eigenvalue, contribution, and factorial coordinates. Inertia is a measurement of the dispersion of the set of computed distances between points. Analogously, in principal CA, inertia corresponds to the explained variance of dimensions. The

eigenvalue allows the inertia that a specific category produces to be quantified determining a certain percentage relative to the entire set of the active category. The percentage coordinates (x- and y-axis) of the graph enable the category points in a graph to be represented and established. In MCA, the distance between 2 or more categories of different variables can be interpreted in terms of the associations and correlations between these. If 2 categories present high coordinates and are close in space, this means that they tend to be directly associated [26,27]. If 2 categories present high coordinates but are distant from each other (eg, they have opposite signs), this means that they tend to be inversely associated [28,29]. A heatmap was created to help the interpretation of the MCA. This plot used the intensity of the colors to show the level of association between the variables. Our graphs showed the association by the distance between the categories in the MCA plot.

K-means

K-means [30] is a nonsupervised learning algorithm used in data mining and pattern recognition. The algorithm partitions the data set in *k* predefined distinct nonoverlapping subgroups (clusters) where each data point belongs to only one group. It tries to make the intracluster data points as similar as possible while also keeping the clusters as different (far) as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneity (similarity) there is between the data points within the same cluster. The k-means algorithm is composed of the following steps: (1) it places *k* points in the space represented by the patients who are being clustered, (2) it assigns each patient to the group that has the closest centroid, and (3) when all patients have been assigned, it recalculates the positions of the *k* centroids. Steps 2 and 3 are repeated until the centroids no longer move. This produces a separation of the patients into homogenous groups while maximizing heterogeneity across groups. The optimal number of clusters was obtained by the elbow method [31]. This consists of plotting the explained variation as a function of the number of clusters and picking the elbow of the curve as the number of groups to use. To assess internal cluster quality, cluster stability of the optimal solution was computed using Jaccard bootstrap values with 10,000 runs [32].

Statistical Analysis

All the information presented was analyzed using MCA, an extension of CA, and the k-means algorithm. The combination of MCA and k-means benefits the effectiveness of the calculation process and, in consequence, the k-means results. MCA helps to reduce the noise, which allows the k-means algorithm to obtain more accurate distances. The MCA dimension reduction automatically performs data clustering according to the k-means objective function [33]. In addition, the potential confounding factors in this study were assessed by calculating the distances between the variables (inertia) that take into account their relative weight in the database as a whole. However, these variables were related to each other depending on the similarity of each register. Previously, the patients with empty fields were removed.

The MCA method was implemented in scripts performed with R [34], an open-source programming language and environment for statistical computing and graphics. Specifically, the main library used to implement the methods and obtain the results was FactoMineR [35]. K-means was written in Python [36], and the main library used scikit-learn [37]. These methods were launched by their default configuration and using a personal computer.

Results

MCA and K-means Without the Tumor Staging

The analysis of the MCA and k-means without the stage variable included 1083 registers. Figure 2 shows the different categories and their possible associations. The variance for dimension 1 was 15% (eigenvalue 0.21) and that for dimension 2 was 12% (eigenvalue 0.17). Figure 2 also shows the position of each category in the plot and its contribution on the dimensions. Note the contribution of mortality (15% on the negative x-axis and 10.2% on the positive y-axis), the ≥75 years age group (18.8% on the negative x-axis and 4.5% on the positive y-axis), and the ex-smoker/smoker (16.5% and 12.3% on positive x-y axis). Figure 3 shows the relation between the categories. The associations between the points were significant when they were closer and the distance was minimum. For example, females and obesity were represented in the same dimension in the MCA plot. Therefore, the heatmap also demonstrated this association with a distance of 0.4 between the points in the MCA plot.

Figure 2. 2D multiple correspondence analysis plot showing the correlations between the categories and their contributions for all data sets.

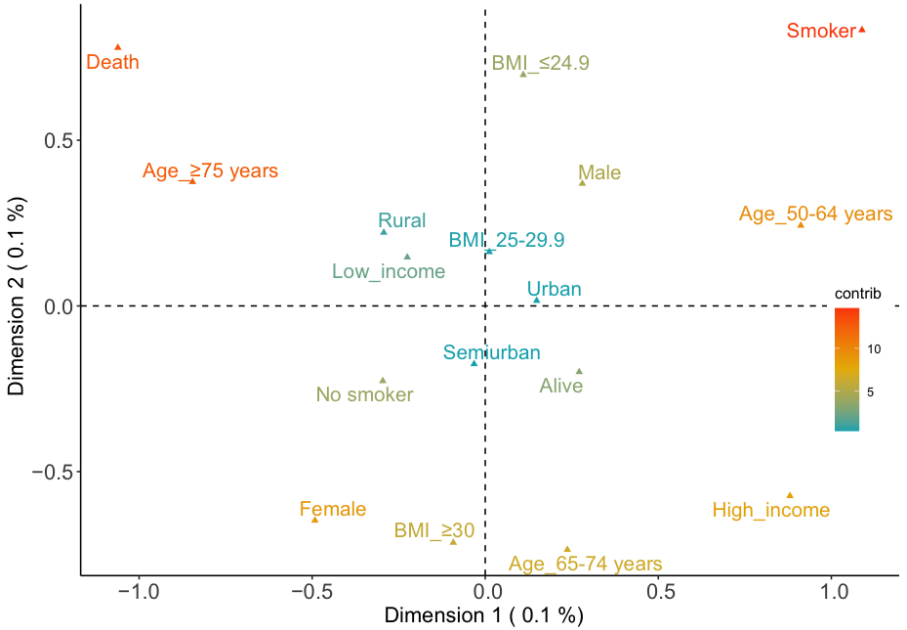
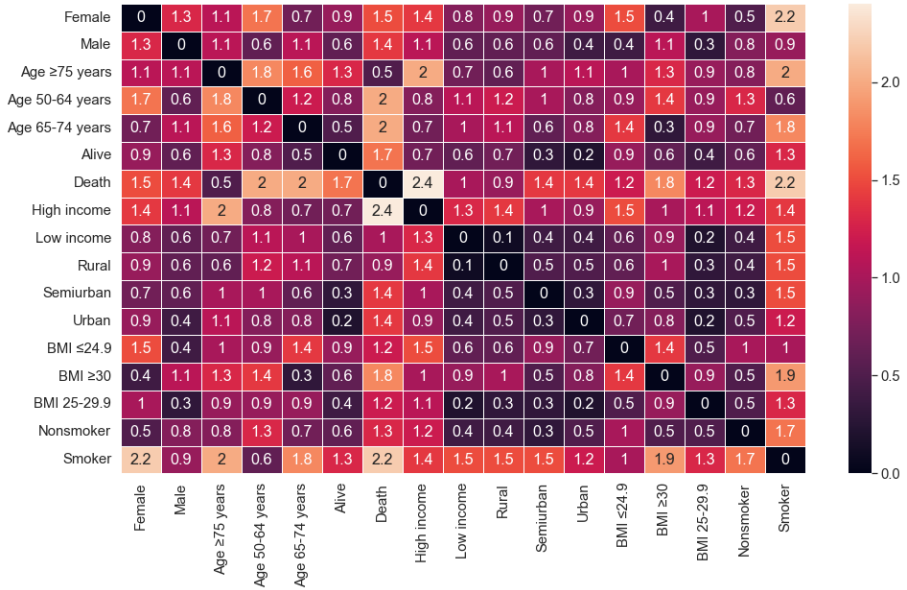


Figure 3. Correlations between the categories by the distance between them.



Graphically, the points closer to each other or the points represented in the same direction of the axis suggest associations. As can be seen, mortality and older age are very close in the plot. This suggests a possible association. Another possible relation observed could be between females and obesity. Then, a cloud on the positive x-axis and the negative y-axis was made up of the 65-74 years age group, high income, and survival. Finally, additional associations could be made up of the 50-64 years age group, males, smokers or ex-smokers, and normal weight.

Table 2 shows the centroids of the main clusters obtained after applying the k-means algorithm. The recommended number of optimal clusters was 5 [31] (see the GitHub [19] repository to evaluate the plot). The first cluster grouped 242 registers among which the main register was males aged ≥75 years from urban populations, with low income, nonsmokers who were overweight, and with a low risk of dying. The next cluster (259 registers) represented females aged between 50 and 64 years with high income. It grouped the cases from rural populations with normal weight and survival. Cluster number 3 was made up of 180 registers. These were mostly males aged ≥75 years with low income and from semiurban populations. They were nonsmokers but were obese and unfortunately included exitus. It was the only cluster that included mortality. The fourth cluster represented urban males aged between 65 and 74 years and with low income. In this case, they were smokers or ex-smokers with normal weight and no mortality. It contained 194 registers. Finally, the last cluster was made up of 208 cases, which included semiurban females aged between 65 and 74 years with low income. They were not smokers but they were overweight. Fortunately, surviving patients predominated in this cluster and the risk of dying was low. See these clusters represented graphically in the GitHub [19].

Table 2. Centroids of the main clusters obtained from the k-means algorithm for all data sets.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Urban	Rural	Semiurban	Urban	Semiurban
Age ≥75 years	Age 50-64 years	Age ≥75 years	Age 65-74 years	Age 65-74 years
Low income	High income	Low income	Low income	Low income
Male	Female	Male	Male	Female
Nonsmoker	Nonsmoker	Nonsmoker	Smoker/Ex-smoker	Nonsmoker
Overweight	Normal weight	Obesity	Normal weight	Overweight
Alive	Alive	Death	Alive	Alive

MCA and K-means Including the Tumor Staging

This subsection presents the outcomes, including the stage of the tumor. The data set used for this analysis discarded the registers, which did not contain the stage (647 registers). Therefore, the number of cases analyzed was 438 (Table 1). Figure 4 shows the outcomes obtained after applying MCA. The variance of dimension 1 was 11.4% (eigenvalue 0.18) and that of dimension 2 was 10.2% (eigenvalue 0.16). Mortality was also one of those with the highest contribution (26.4% on the positive x-axis and 10.5% on the positive y-axis). Near this was stage III with a high contribution (16.3% on the positive x-axis and 13.7% on the positive y-axis). Ex-smoker/smoker contributed significantly compared with the rest of categories (9.1% on the negative x-axis and 1.3% on the positive y-axis). The relations between these and other categories are shown in Figure 5. See the death and its correlation between stage III. The heatmap differentiated this association clearly, as the MCA plot also showed. The location of the categories in the plot and their contributions suggested possible associations. The main association was between stage III and mortality and with females with stage II, the ≥75 years age group, and nonsmokers. Another relation could be males with high income, aged between 50 and 64 years, stage 0, and ex-smokers or smokers. However, these results could be affected by the decrease in cases.

Figure 4. 2D multiple correspondence analysis plot showing the correlations between the categories and their contributions.

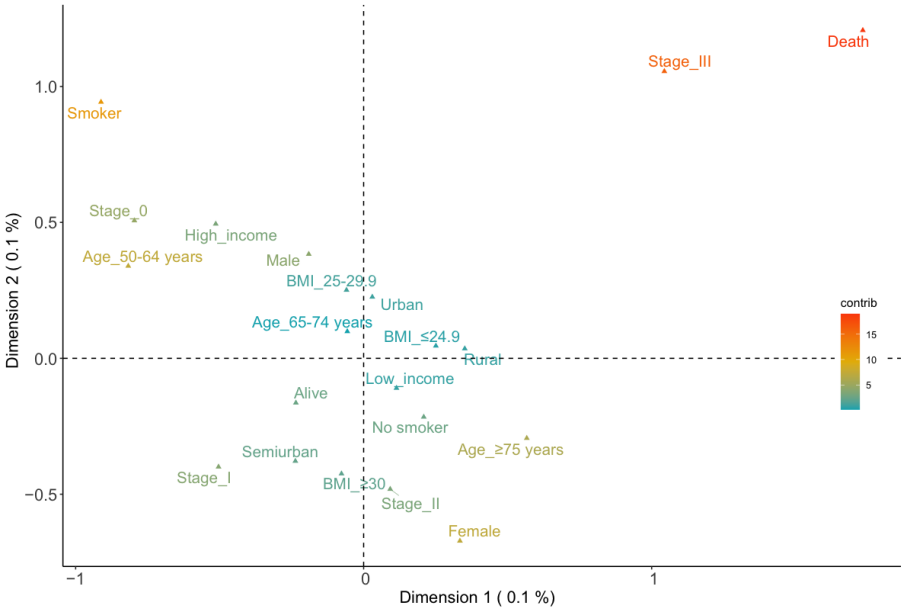


Figure 5. Correlations between the categories by the distance between them including the tumor staging.

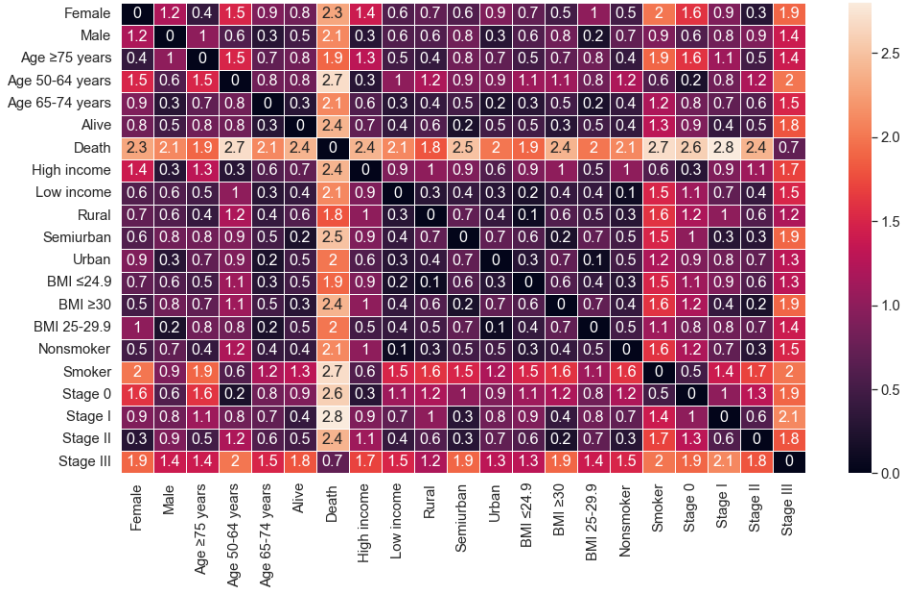


Table 3 shows the clusters obtained from the data set with the tumor stage. All the clusters obtained were male nonsmokers owing to the decrease in the number of registers in the data set. The first cluster with 135 cases represented obese urban patients aged between 65 and 74 years with stage II reached and a low risk of death. The second cluster had 120 registers of patients with stage II and age ≥75 years from semiurban populations. Their risk of death was also low. The next cluster included 76 registers and they were from the urban population but overweight. They included the younger patients (50-64 years age group), with a low risk of death and the lowest stage (stage 0). The fourth cluster (n=72) represented rural inhabitants, aged between 65 and 74 years. They were obese with stage III cancer but low risk of death. However, the fifth cluster was patients from the semiurban population, aged ≥75 years, overweight, in an advanced stage (III), and with a high risk of death. See these clusters represented graphically in the k-means folder of GitHub [19].

Table 3. Centroids of the main clusters obtained from the k-means algorithm: the final data set after including the stage of the tumor.

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Urban	Semiurban	Urban	Rural	Semiurban
Age 65-74 years	Age ≥75 years	Age 50-64 years	Age 65-74 years	Age ≥75 years
High income	Low income	Low income	Low income	Low income
Male	Male	Male	Male	Male
Nonsmoker	Nonsmoker	Nonsmoker	Nonsmoker	Nonsmoker
Obesity	Obesity	Overweight	Obesity	Overweight
Alive	Alive	Alive	Alive	Death
Stage II	Stage II	Stage 0	Stage III	Stage III

Discussion

The MCA technique and the k-means algorithm permit the analysis and detection of clusters of patients with similar risk factors and outcomes not observed in the literature. The population-based cancer registry for the province of Lleida registered 1083 colorectal cancers between 2012 and 2015. This cancer is the most incident in our region [5,17,18] and by applying MCA and k-means, some relationships were found between some aspects that corroborate the usefulness of these techniques. They helped to detect that in colorectal cancer, the age group and BMI risk factors are related. Another important corroboration was the risk of death in older people (≥75 years age group) either obese or overweight and in an advanced stage. Related to this latter factor, the advanced stage was observed in older people with obesity. Stages II and III were 65% (119/181) of the total in the ≥75 years age group.

Previous studies have used clustering techniques to detect associations, but none of them were used for associating patient profiles with risk factors. We based our study on a preliminary paper [38], which evaluated the relationship between air pollution, particulate matter components, and risk of breast cancer in a United States–wide prospective cohort by using a clustering technique. That study concluded that air pollution measures were related to both invasive breast cancer and ductal carcinoma in situ within certain geographic regions. Another starting point was the study presented in [39], which used the combination of MCA and k-means to ascertain multimorbidity patterns. That study concluded that these techniques could help to identify these patterns. Another study our work was based on is the one presented in [40], which studied the trends in the incidence of cancers associated with being overweight and obese. Another study [41] analyzed the possible relation between obesity and colorectal cancer. These papers studied the impact of the risk factors on colorectal cancer but did not use the MCA technique or k-means algorithm to explore associations between these and their impact. In addition, a previous study used MCA to analyze the prognosis in surgery for low rectal cancer [42]. Another study used k-means to search patterns in patients with colorectal cancer, but its main aim was to detect emotion regulation patterns and personal resilience [43]. However, to the best of our knowledge, no prior studies have used MCA or k-means to link types of risk factors, SES, tumor stage, and patients' characteristics in cases of colorectal cancer.

One MCA outcome was the inertia (27%). Further, various variables had high contributions. A strong relation was obtained between older patients (≥75 years age group) and mortality. This may suggest an increase in the risk of mortality for colorectal cancer in older adults, as previous studies showed [44]. On the opposite side of previous associations, it showed another association between survival, high SES, and the 65-75 years age group. Even though the contributions of these are lower than those of mortality and the older population, it is suggested that the risk of death is lower in people with high SES [45] and among younger people. An association was detected between females and obesity although this was not reflected in the k-means. This relation may be because 37% (146/394) of all the women were obese. However, obese men represented 29% (205/689) of the male population, and the percentage of obesity in the data set was 31% (343/1083). This relation suggests that obese women could more likely develop colorectal cancer than men. In general, the probability of colorectal cancer in obese patients can increase by 30%-70% [46]. However, although the contribution is too low to establish a strong relation, the position of males and normal weight in the plot might suggest that there may be some other factors that increase the risk of this cancer and that these techniques

highlighted other associations. Some additional patient clinical history would be necessary.

Regarding the k-means analysis, the third cluster confirmed the mortality in the older population with obesity [44]. The first cluster also represented the ≥75 years age group but who were overweight and had no exitus. These differences between clusters suggested that obesity may be a determining factor in older persons that increases the risk of death. In addition, these 2 clusters were males. Similar outcomes were obtained in the fifth cluster when the tumor stage was added. Stage III was directly related with the ≥75 years age group, the semiurban population, and mortality, thereby suggesting that for older persons, being overweight or obese and in an advanced stage could increase the risk of death. The fourth cluster was made up of smokers or ex-smokers. Although tobacco is not usually directly related with colorectal cancer, some studies also support this result [47,48].

The analysis then studied the data set filtered by tumor stage. The final data set was made up of 438 registers. The MCA technique obtained a significant relation between stage III and mortality. However, screening programs and technology decrease this risk, as recent studies concluded [49]. We can also see that stage 0 was related with younger people (50-64 years age group). The k-means results gave similar conclusions as in the MCA. The younger people, stage 0, and survival appeared in the same cluster as demonstrated in the previous k-means analysis with the second cluster. This suggests the importance of screening programs to detect tumors at an early stage [50]. The fourth cluster in the second analysis related rural and stage III. This association may insinuate a possible delay in diagnosis or difficulties in accessing the health care system and mass screening testing in rural areas [51]. Finally, note that all clusters that had stage II or III also included obesity or excess weight. This may suggest that the BMI may be a determinant for having an aggressive colorectal tumor. However, no significant outcomes related to income were obtained, although 80% (863/1083) of the cases were low-income patients. This high percentage of low-income cases could be explained by the fact that the average annual net income per person in Catalonia in 2015 was €12,283 [52].

The strengths of using the MCA and k-means cluster analysis are that the results are less susceptible to outliers in the data, the influence of chosen distance measures, or the inclusion of inappropriate or irrelevant variables [53]. This study had some limitations that should be noted. Regarding the techniques, it tends to take into account the relative weight of each variable concerning the set of study variables and allows control for potential confounding factors such as sex, age, and survival. However, some residual confounding effects cannot be ruled out. Further, these include the low number of cases with tumor stage (438/1083, 40% of total). In consequence, the final data set also made it difficult to analyze the strength of the causal relationship between different prediction parameters and outcomes because it contained few registers. The postal address registered for each case was the patient's home address at the time of cancer diagnosis. However, this address may have changed during the study. Despite this, the number of cases with changed addresses would be very low and this factor is not expected to produce bias in the results. Some lifestyle aspects such as alcohol consumption, diabetes, or profession were not considered. The lack of cause of death is another limitation. The results showed that there is room for other kinds of risk factors. Additional patient clinical history would be required in order to find these. Further, related to the comorbidities, the Charlson index could not be added because approximately only 15% of the sample received it. A future study may be the study of the causality, adding synthetic data to enlarge the data set. Finally, some associations could hide others due to these techniques even though they showed the most significant relationships. In addition, the genetic and hereditary conditions were not considered.

In conclusion, many studies demonstrate that some risk factors such as BMI, tobacco smoking, or SES could influence the incidence of colorectal cancer by using traditional techniques. This study used new techniques such as MCA and k-means to analyze the relationships between colorectal cancer and risk factors. The outcomes obtained demonstrated that the combination of these techniques could help to detect relations between risk factors and patient characteristics. Obesity and being overweight in the older population (≥75 years age group) increases the risk of developing aggressive tumors and death. Stage 0 was related with younger people and survival. This highlights the importance of screening programs for colorectal cancer. The presence of tobacco in a cluster indicated that it must be considered as a risk factor in colorectal cancer. The results of our study help to corroborate suspected trends in several of the relationships detected and confirm the usefulness of these techniques. Further, they encourage applying these methods to other cancers and detecting how the risk factors could be associated. In future work, it is important to delve deeper into the patients' characteristics and risk factors. This means including new variables such as diabetes, alcoholism, or the cause of death. The findings obtained in this study motivate us to search for relations between risk factors in other cancers. Moreover, new techniques and artificial intelligence algorithms can be implemented to explore patterns of pretumor and posttumor detection from the clinical history.

Acknowledgments

This work was supported by contract 2019-DI-43 from the Industrial Doctorate Program of the Government of Catalonia and by the Spanish Ministry of Science and Innovation under contract PID2020-113614RB-C22. Some of the authors are members of the research group 2014-SGR163, funded by the Generalitat de Catalunya.

Conflicts of Interest

None declared.

References

1. Ferlay J, Soerjomataram I, Dikshit R, Eser S, Mathers C, Rebelo M, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. Int J Cancer 2015 Mar 01;136(5):E359-E386 [FREE Full text] [doi: 10.1002/ijc.29210] [Medline: 25220842]
2. Ferlay J, Colombet M, Soerjomataram I, Dyba T, Randi G, Bettio M, et al. Cancer incidence and mortality patterns in Europe: Estimates for 40 countries and 25 major cancers in 2018. Eur J Cancer 2018 Nov;103:356-387. [doi: 10.1016/j.ejca.2018.07.005] [Medline: 30100160]
3. Labianca R, Beretta GD, Kildani B, Milesi L, Merlin F, Mosconi S, et al. Colon cancer. Crit Rev Oncol Hematol 2010 May;74(2):106-133. [doi: 10.1016/j.critrevonc.2010.01.010] [Medline: 20138539]
4. Parkin DM. The evolution of the population-based cancer registry. Nat Rev Cancer 2006 Aug;6(8):603-612. [doi: 10.1038/nrc1948] [Medline: 16862191]
5. Florensa D, Pedrol T, Mòdol I, Farré X, Salud A, Mateo J, et al. Resultats de l'any 2014. Butlletí Epidemiològic Catalunya 2020 2020;40(12):252-264.
6. Florensa D, Godoy P, Mateo J, Solsona F, Pedrol T, Mesas M, et al. The Use of Multiple Correspondence Analysis to Explore Associations Between Categories of Qualitative Variables and Cancer Incidence. IEEE J Biomed Health Inform 2021 Sep;25(9):3659-3667. [doi: 10.1109/JBHI.2021.3073605] [Medline: 33857006]
7. Bjørge T, Häggström C, Ghaderi S, Nagel G, Manjer J, Tretli S, et al. BMI and weight changes and risk of obesity-related cancers: a pooled European cohort study. Int J Epidemiol 2019 Dec 01;48(6):1872-1885. [doi: 10.1093/ije/dyz188] [Medline: 31566221]
8. Guo L, Li N, Wang G, Su K, Li F, Yang L, et al. [Body mass index and cancer incidence:a prospective cohort study in northern China]. Zhonghua Liu Xing Bing Xue Za Zhi 2014 Mar;35(3):231-236. [Medline: 24831616]
9. Mizoue T, Inoue M, Tanaka K, Tsuji I, Wakai K, Nagata C, Research Group for the Development, Evaluation of Cancer Prevention Strategies in Japan. Tobacco smoking and colorectal cancer risk: an evaluation based on a systematic review of epidemiologic evidence among the Japanese population. Jpn J Clin Oncol 2006 Jan;36(1):25-39. [doi: 10.1093/jjco/hyi207] [Medline: 16423841]
10. Kim D, Masyn KE, Kawachi I, Laden F, Colditz GA. Neighborhood socioeconomic status and behavioral pathways to risks of colon and rectal cancer in women. Cancer 2010 Sep 01;116(17):4187-4196 [FREE Full text] [doi: 10.1002/cncr.25195] [Medline: 20544839]
11. Ugurlu H, Cicek I. Analysis and assessment of ship collision accidents using Fault Tree and Multiple Correspondence Analysis. Ocean Engineering 2022 Feb;245:110514. [doi: 10.1016/j.oceaneng.2021.110514]
12. Rustam Z, Hartini S, Yunus R, Pratama R, Yunus R, Hidayat R. Analysis of Architecture Combining Convolutional Neural Network (CNN) and Kernel K-Means Clustering for Lung Cancer Diagnosis. Artic Int J Adv Sci Eng Inf Technol Internet 2020 2020;10(3):1200-1206. [doi: 10.18517/ijaseit.10.3.12113]
13. Ronen J, Hayat S, Akalin A. Evaluation of colorectal cancer subtypes and cell lines using deep learning. Life Sci Alliance 2019 Dec;2(6):e201900517 [FREE Full text] [doi: 10.26508/lsa.201900517] [Medline: 31792061]
14. eCAP. Departament de Salut. URL: https://salutweb.gencat.cat/ca/ambits_actuacio/linies_dactuacio/tic/sistemes-informacio/gestio-assistencial/ecap/ [accessed 2020-12-16]
15. Registre central de població del CatSalut. Català de la Salut. URL: https://catsalut.gencat.cat/ca/proveidors-professionals/registres-catalegs/registres/central-poblacio/index.html [accessed 2020-12-16]
16. International Agency for Research on Cancer. URL: http://www.iacr.com.fr/index.php?option=com_content&view=article&id=72:iarccrgtools&catid=68&Itemid=445 [accessed 2020-12-14]
17. Godoy P, Pedrol T, Mòdol I, Salud A. El Registre poblacional de càncer a Lleida: resultats i perspectives. Butlletí Epidemiològic Catalunya. 2016. URL: https://scientiasalut.gencat.cat/handle/11351/3052?show=full&locale-attribute=en [accessed 2022-07-06]
18. Godoy-Garcia P, Pedrol T, Mòdol-Pena I, Salud A. El registre poblacional de càncer a Lleida: resultats de l'any 2013. Butlletí Epidemiològic Catalunya. 2018. URL: https://scientiasalut.gencat.cat/handle/11351/3665?show=full [accessed 2022-07-06]
19. Florensa D, Godoy P, Solsona P, et al. The use of multiple correspondence analysis to explore associations between categories of qualitative variables and cancer incidence. Github Repository. URL: https://github.com/didacflorensa/MCA-Cancer [accessed 2021-01-01]
20. Weisell R. Body mass index as an indicator of obesity. Asia Pac J Clin Nutr 2002;11(8):681-684. [doi: 10.1046/j.1440-6047.11.s8.5.x]
21. España, Real Decreto-ley 16/2012, de 20 de abril, de medidas urgentes para garantizar la sostenibilidad del Sistema Nacional de Salud y mejorar la calidad y seguridad de sus prestaciones. Boletín Oficial del Estado. URL: https://www.boe.es/diario_boe/txt.php?id=BOE-A-2012-5403 [accessed 2022-07-06]
22. García GJ. La Población Rural de España: De los Desequilibrios a la Sostenibilidad Social. Spain: Barcelona: Fundación La Caixa; 2013:146-149.

23. Murtagh F. Multiple correspondence analysis and related methods. Psychometrika 2007 Mar 24;72(2):275-277. [doi: 10.1007/s11336-006-1579-x]
24. Husson F, Josse J. Multiple correspondence analysis. In: Visualization and Verbalization of Data. Boca Raton, Florida: CRC/PRESS; 2014.
25. Sourial N, Wolfson C, Zhu B, Quail J, Fletcher J, Karunananthan S, et al. Correspondence analysis is a useful tool to uncover the relationships among categorical variables. J Clin Epidemiol 2010 Jun;63(6):638-646 [FREE Full text] [doi: 10.1016/j.jclinepi.2009.08.008] [Medline: 19896800]
26. Greenacre M. Correspondence Analysis in Practice, Third Edition. Boca Raton, Florida: Chapman and Hall/CRC; 2017.
27. Roux BL, Rouanet H. Geometric Data Analysis: From Correspondence Analysis to Structured Data Analysis. New York: Springer; 2005.
28. Di Franco G. Multiple correspondence analysis: one only or several techniques? Qual Quant 2015 Apr 21;50(3):1299-1315. [doi: 10.1007/s11135-015-0206-0]
29. Heckler CE. Applied Multivariate Statistical Analysis. Technometrics 2005 Nov;47(4):517-517. [doi: 10.1198/tech.2005.s319]
30. Likas A, Vlassis N, J. Verbeek J. The global k-means clustering algorithm. Pattern Recognition 2003 Feb;36(2):451-461. [doi: 10.1016/s0031-3203(02)00060-2]
31. Bholowalia P, Kumar A. EBK-Means: A Clustering Technique based on Elbow MethodK-Means in WSN. Int J Comput Appl 2014;105:17-24. [doi: 10.5120/18405-9674]
32. Hennig C. Cluster-wise assessment of cluster stability. Computational Statistics & Data Analysis 2007 Sep;52(1):258-271. [doi: 10.1016/j.csda.2006.11.025]
33. Ding C, He X. K-means clustering via principal component analysis. 2004 Presented at: Proceedings of the 21st Int Conf Mach Learn (ICML); 2004; Banff, Canada. [doi: 10.1145/1015330.1015408]
34. The R Project for Statistical Computing. 2019. URL: https://www.r-project.org/ [accessed 2020-12-16]
35. Lê S, Josse J, Husson F. FactoMineR: An R Package for Multivariate Analysis. J Stat Soft 2008;25(1):1-18. [doi: 10.18637/jss.v025.i01]
36. Welcome to Python.org. URL: https://www.python.org/ [accessed 2020-12-16]
37. scikit-learn: MLIP. URL: https://scikit-learn.org/stable/ [accessed 2020-12-16]
38. White A, Keller J, Zhao S, Carroll R, Kaufman J, Sandler D. Air Pollution, Clustering of Particulate Matter Components, and Breast Cancer in the Sister Study: A U.S.-Wide Cohort. Environ Health Perspective 2019;127(10):107002. [doi: 10.1289/ehp5131]
39. Violán C, Roso-Llorach A, Foguet-Boreu Q, Guisado-Clavero M, Pons-Vigués M, Pujol-Ribera E, et al. Multimorbidity patterns with K-means nonhierarchical cluster analysis. BMC Fam Pract 2018 Jul 03;19(1):108 [FREE Full text] [doi: 10.1186/s12875-018-0790-x] [Medline: 29969997]
40. Steele CB, Thomas CC, Henley SJ, Massetti GM, Galuska DA, Agurs-Collins T, et al. Vital Signs: Trends in Incidence of Cancers Associated with Overweight and Obesity - United States, 2005-2014. MMWR Morb Mortal Wkly Rep 2017 Oct 03;66(39):1052-1058 [FREE Full text] [doi: 10.15585/mmwr.mm6639e1] [Medline: 28981482]
41. Liu P, Wu K, Ng K, Zauber AG, Nguyen LH, Song M, et al. Association of Obesity With Risk of Early-Onset Colorectal Cancer Among Women. JAMA Oncol 2019 Jan 01;5(1):37-44 [FREE Full text] [doi: 10.1001/jamaoncol.2018.4280] [Medline: 30326010]
42. Mancini R, Pattaro G, Diodoro MG, Sperduti I, Garufi C, Stigliano V, et al. Tumor Regression Grade After Neoadjuvant Chemoradiation and Surgery for Low Rectal Cancer Evaluated by Multiple Correspondence Analysis: Ten Years as Minimum Follow-up. Clin Colorectal Cancer 2018 Mar;17(1):e13-e19. [doi: 10.1016/j.clcc.2017.06.004] [Medline: 28865674]
43. Baziliansky S, Cohen M. Emotion Regulation Patterns among Colorectal Cancer Survivors: Clustering and Associations with Personal Coping Resources. Behav Med 2021;47(3):214-224. [doi: 10.1080/08964289.2020.1731674] [Medline: 32275195]
44. Liu X, Bi Y, Wang H, Meng R, Zhou W, Zhang G, et al. Different trends in colorectal cancer mortality between age groups in China: an age-period-cohort and joinpoint analysis. Public Health 2019 Jan;166:45-52. [doi: 10.1016/j.puhe.2018.08.007] [Medline: 30447645]
45. Hastert TA, Beresford SAA, Sheppard L, White E. Disparities in cancer incidence and mortality by area-level socioeconomic status: a multilevel analysis. J Epidemiol Community Health 2015 Mar;69(2):168-176. [doi: 10.1136/jech-2014-204417] [Medline: 25288143]
46. Bardou M, Barkun AN, Martel M. Obesity and colorectal cancer. Gut 2013 Jun;62(6):933-947. [doi: 10.1136/gutjnl-2013-304701] [Medline: 23481261]
47. Rawla P, Sunkara T, Barsouk A. Incidence, mortality, survival, and risk factors. Gastroenterology Review 2019;14(2):89-103. [doi: 10.5114/pg.2018.81072]
48. Botteri E, Iodice S, Bagnardi V, Raimondi S, Lowenfels AB, Maisonneuve P. Smoking and colorectal cancer: a meta-analysis. JAMA 2008 Dec 17;300(23):2765-2778. [doi: 10.1001/jama.2008.839] [Medline: 19088354]

49. Brouwer NPM, Bos ACRK, Lemmens VEPP, Tanis PJ, Hugen N, Nagtegaal ID, et al. An overview of 25 years of incidence, treatment and outcome of colorectal cancer patients. Int J Cancer 2018 Dec 01;143(11):2758-2766 [FREE Full text] [doi: 10.1002/ijc.31785] [Medline: 30095162]
50. Brenner H, Chang-Claude J, Jansen L, Knebel P, Stock C, Hoffmeister M. Reduced risk of colorectal cancer up to 10 years after screening, surveillance, or diagnostic colonoscopy. Gastroenterology 2014 Mar;146(3):709-717. [doi: 10.1053/j.gastro.2013.09.001] [Medline: 24012982]
51. Henley SJ, Anderson RN, Thomas CC, Massetti GM, Peaker B, Richardson LC. Invasive Cancer Incidence, 2004-2013, and Deaths, 2006-2015, in Nonmetropolitan and Metropolitan Counties - United States. MMWR Surveill Summ 2017 Jul 07;66(14):1-13 [FREE Full text] [doi: 10.15585/mmwr.ss6614a1] [Medline: 28683054]
52. Average annual net income of households. Statistical Institute of Catalonia. 2015. URL: https://www.idescat.cat/pub/?id=aec&n=414&t=2015&lang=en [accessed 2020-12-16]
53. Liao M, Li Y, Kianifard F, Obi E, Arcona S. Cluster analysis and its application to healthcare claims data: a study of end-stage renal disease patients who initiated hemodialysis. BMC Nephrol 2016 Mar 02;17:25 [FREE Full text] [doi: 10.1186/s12882-016-0238-2] [Medline: 26936756]

Abbreviations

CA: correspondence analysis
MCA: multiple correspondence analysis
SES: socioeconomic status

Edited by A Mavragani; submitted 25.03.21; peer-reviewed by CM Moore, E Mohammadi, Y Chu, L Espinosa-Leal; comments to author 28.12.21; revised version received 22.02.22; accepted 23.05.22; published 19.07.22

Please cite as:
Florensa D, Mateo-Fornés J, Solsona F, Pedrol Aige T, Mesas Julió M, Piñol R, Godoy P
Use of Multiple Correspondence Analysis and K-means to Explore Associations Between Risk Factors and Likelihood of Colorectal Cancer: Cross-sectional Study
J Med Internet Res 2022;24(7):e29056
URL: https://www.jmir.org/2022/7/e29056
doi: 10.2196/29056
PMID:

4.4 PAPER IV: SMOKING AND ALCOHOL ASSOCIATED WITH SPC

Authors: *Dídac Florensa, Jordi Mateo, Francesc Solsona, Carme Miret, Miquel Mesas, Ramon Piñol, Pere Godoy*
Journal: Cancers
Status: Under Review
Keywords: *Subsequent Primary Cancer, Risk factors, Smoking, Heavy drinking, Retrospective cohort study*

Association of smoking and heavy drinking with risk of Subsequent Primary Cancer

Abstract: Smoking and heavy drinking are important preventable risk factors for cancer. This article investigated the relationship between smoking and heavy drinking and the risk of a subsequent primary cancer (SPC). A retrospective cohort study on subsequent primary cancer (SPC) risk factors was conducted in patients with primary cancer. Participants were patients with primary cancers from the population registry of cancer in Lleida during 2012-2016. The dependent variable was an SPC. Risk factors were studied with the adjusted hazard ratio (aHR) with 95% confidence intervals (CI) using a Cox proportional hazard model. We studied 5,658 primary cancers: 234 cases developed an SPC (4.1%) of which 2.7% (HR=1.7; 95%CI: 1.3-2.3) were males and 3% (HR=3.0; 95%CI: 1.5-3.3) were from the 70-79 age group. There were also higher hazard ratios for SPC in smokers (HR=1.5; 95%CI: 1.1-1.9) and heavy drinkers (HR=2.7; 95%CI: 1.4-5.4). The Cox proportional hazard model showed a relationship with the risk of SPC in the 60-69 and 70-79 years age groups, males (aHR=1.4; 95%CI: 1.1-1.9), smokers (aHR=1.3; 95%CI: 1.0-1.7) and heavy drinkers (aHR=2.4; 95%CI: 1.3-4.8). The results confirm the association between smoking and heavy drinking and the risk of a SPC during the early follow-up years. Patients with a primary cancer should be informed about the risk of an SPC.

INTRODUCTION

Smoking and heavy alcohol use increase the risk of cancer [25, 26]. Smoking is mainly associated with lung cancer and many other types of primary cancer [27–29]. There is also strong evidence that heavy drinking causes cancer [31]. In addition to liver cancer [30], other cancers, such as oesophageal, gastric and colorectal cancer are associated with heavy drinking [32]. Cancer survival trends are generally increasing, even for some of those with the highest mortality, such as liver, pancreas and lung cancer [16]. Long-term survivors face physical, psychosocial, medical, behavioural, and socioeconomic consequences due to cancer and its treatment [17]. One medical consequence is an increased likelihood of subsequent diagnosis with another cancer [18]. Cancer survivors might be especially prone to developing new cancers for various reasons. These include common etiologic risk factors with the primary cancer (i.e., environmental exposure, genetics, lifestyle choices) and the after-effects of cancer treatment [19]. Risk factors such as obesity, smoking and heavy drinking could be determinants of a subsequent primary cancer (SPC) [20, 21] which is defined as first subsequent primary cancer occurring at least 6 months after the first cancer. Specifically, some risk factors are more closely related to specific cancers, such as smoking with the larynx or obesity with the stomach [22]. Several studies have reported an association between a first primary cancer and the risk of an SPC [93, 94]. They suggested that cancer survivors have a higher risk of an SPC than the general population and that some SPCs share lifestyle-associated risk factors with some primary cancers. Heavy drinking and smoking may be independent additive SPC risk factors [95]. Other studies considered obesity, insulin resistance and diabetes as risk factors for several SPCs [96–98]. Supramaniam et al. [18] reported that 35% of all excess risk for SPCs may be attributable to behavioural factors, such as drinking and smoking. The study by Jassem [99] also suggested that smoking increased the relative risk of SPCs among patients free of cancer for two or more years. Sung et al. [22] highlighted that larynx, Hodgkin's lymphoma, pancreas and oesophageal cancer meant a higher risk of an SPC.

The literature focused on analysing the association between primary cancer risk and risk factors. However, relatively few studies have explored the role of these factors on SPC risk. Previ-ous studies investigated the relationship between smoking, heavy drinking or body mass index with the risk of developing the first primary cancer [23, 100] and the role this may play in the future [2] . Few studies have focused on the relationship between SPC and risk factors in the years following the first cancer diagnosis [22, 93, 101].

The objective of this study was to analyse the relationship between smoking and heavy drinking and the risk of an SPC in the Lleida region (Catalonia).

METHODS

Study population and design

A retrospective cohort study on SPCs and their association with some risk factors and sociodemographic information was carried out. Data on cancer diagnoses were obtained from the Population-based Cancer Registry (PCR) of the health region of the province of Lleida (HRPLL) with 5 consecutive years of incidence data from 2012 to 2016. Information on risk factors was extracted from the eCAP [102] software, a computerized medical record program used by doctors, paediatricians and nurses in Primary Care Centres. Before applying statistical techniques, the information was validated by professionals in the Lleida PCR. Potential cancer cases were validated by checking their medical records to confirm them. We used hospital records and pathological anatomy records as the main information sources. Study participants included all persons diagnosed with their first ever cancer from January 1, 2012 to December 31, 2016. All patients with a primary cancer in the 5 years before January 1, 2012, were excluded. The cancer cases detected as FPC during 2016 were also excluded because there was no follow-up time to register an SPC. These cases were initially included to detect potential SPCs. Following the first cancer diagnosis every patient was followed up from the first cancer diagnosis to the detection of an SPC, date of death or December 31, 2016. An SPC was defined as a first subsequent primary cancer occurring at least 6 months after the first cancer in patients aged 50 years or more at time of first primary cancer diagnosis. Primary cancers were identified according to the rules defined by the International Association of Cancer Registries (IACR), the International Association for Research on Cancer (IARC) and the European Network of Cancer Registries (ENCR).

Information on heavy drinking and smoking consumption, body mass index and diabetes were extracted from the eCAP software. The values of these variables at the time of the first cancer diagnoses were obtained. Body mass index (BMI) was calculated by the weight and height of the patient using the formula

$$BMI = (weight(kg))/height(m)^2$$

and categorized as follows: 18.5–24.9 normal weight, 25-29.9 overweight and >30 obese. Heavy drinking, smoking and diabetes were identified by the ICD-10 international criteria. The ICD-10 code for alcohol use was F10.9 (alcohol use)

previously registered by a primary care doctor. We defined this exposure as heavy drinking that was diagnosed as the consumption of >40 grams/day in men and >24 grams/day in women for 1 or more years [103]. To determine diabetes, we used the code E10-E14 (diabetes mellitus). And, for smoking, we used Z72.0 (tobacco use). Once the patients were detected to be using tobacco, we analyzed how many cigarettes they smoked per day. We defined this exposure as smoking when somebody smoked 6 or more cigarettes/day (moderate or severe smoker) [104]. The eCAP software permits determination of the smoking exposure per year. Therefore, patients defined as smokers were cases that were exposed for more than five years before cancer detection. Former smokers were considered as smokers because the observed cases in the dataset were minimum.

Person-years at risk were calculated as the time from the first cancer diagnosis until December 31, 2016 or the date of the SPC diagnosis or date of death [94]. Cancer was grouped according to the type and its association with risk factors. Type 1 covered cancers associated with smoking (larynx, oral cavity and pharynx, stomach, oesophagus, lung and bronchus, urinary bladder and leukaemia). Type 2 included cancers associated with obesity (colon and rectum, stomach-cardia, gallbladder (and other biliary), thyroid, corpus and uterus, breast, pancreas, ovary, kidney and renal pelvis, myeloma). Type 3 was associated with infections (liver, anus, cervix uteri, vulva and other genital organs, lymphoma, penis and other genital organs). Type 0 represents other cancers [22].

Data collection

The data consisted of the information registered between 2012-2016 in the Lleida PCR for cancer patients in the main hospitals in the health region of the province of Lleida. These are the Arnau de Vilanova University Hospital and the Santa Maria University Hospital (HUSM). This information included sociodemographic variables and exposure to smoking and heavy drinking at the moment the cancer was detected. The variables of this study were: gender (male, female), age group (50-59, 60-69, 70-79, 80-); person-years at risk, cancer type (type 0, type 1, type 2, type 3), smoking habits (Yes, No), heavy drinking (Yes, No), classification of BMI (normal weight, overweight, obesity), diabetes (Yes, No).

Statistical Analysis

Patients diagnosed with and without SPCs were compared by socio-demographic information and risk factors. A bivariate analysis was performed to investigate the relationships between the dependent variable (primary cancer with SPC and

non-SPC) and the independent variables, with the crude hazard ratios (HR) and their 95% confidence intervals (CIs) for SPC in cancer patients. A Cox proportional hazard model was used to estimate the adjusted hazard ratios (aHR) and their corresponding 95% CIs. In addition, Cox proportional hazard models were constructed separately for men and women. The probability values for statistical tests were two-tailed, and a CI that does not contain 1.0 was regarded as statistically significant. Results with wide CIs should be interpreted cautiously. All statistical analyses were performed using R 3.6.2 (R Core Team 2019), an open-source programming language and environment for statistical analysis and graphic representation.

RESULTS

We studied 5,658 cancer patients (3,477:61.4% males), of whom 234 (4.1%) had an SPC. Most patients were from the 60-69 years (31.5%) and 70-79 years (32.7%) age groups. With respect to the first cancer type, 152 cases (2.7%) were infection cancer types (type 3), 1,190 cases (21.0%) were obese cancer types (type 2) and 1,573 cases (27.8%) were smoking cancer types. The median patient follow-up was 2 years and the mean was 1.9 years. Smokers represented 30.5% (1,727 cases), there were 216 cases (3.8%) diagnosed with diabetes, and 97 cases of heavy drinking (1.7%) (table 4.1).

The risk factors for patients with an SPC (table 4.2) were analysed by bivariate analysis. There were 69 (1.6%) cases of women with a diagnosed SPC and 165 (2.7%) of men (HR=1.7; 95% CI: 1.3-2.2). The age groups associated with SPCs were those aged 60-69 (HR=1.6; 95% CI: 1.1-2.6) and 70-79 years (HR=2.1; 95% CI: 1.5-3.3). There was also a relationship between SPC and smoking (HR=1.5; 95% CI: 1.1-1.9) and heavy drinking (HR=2.6; 95% CI: 1.4-5.4). No associations were found for the primary cancer type, body mass index and diabetes.

With respect to the outcomes of the Cox proportional hazard model (table 4.3), gender, age, smoking habits and heavy drinking were significantly associated with the risk of SPCs. The aHR for males was 1.4 (95% CI: 1.1-1.9). With respect to the age group, the aHR of the 60-69 years age group was 1.6 (95% CI: 1.1-2.5) and 2.2 (95% CI: 1.5-3.4) for the 70-79 years age group. For the 80 years and over age group, the aHR was 1.2 (95% CI: 0.7-2.0). The aHR for smoking was 1.3 (95% CI: 1.0-1.7) and for heavy drinking, 2.4 (95% CI: 1.3-4.8).

Table 4.4 shows the results of the Cox proportional hazard analysis stratified by gender. The outcomes for males were similar to those of table 4.2. For females, the risk of SPC increased in the 70-79 years age group (HR=2.6; 95%

	Total	
	N	%
Gender		
Female	2,181	38.5
Male	3,477	61.4
Age		
50-59	1,104	19.5
60-69	1,784	31.5
70-79	1,851	32.7
80-	919	16.2
Cancer type		
Type 3	152	2.7
Type 2	1,314	23.3
Type 1	1,446	25.5
Type 0	2,743	48.5
Body mass index		
Normal weight	1,537	27.1
Overweight	2,437	43.1
Obese	1,684	29.7
Smoking		
No	3,931	69,4
Yes	1,727	30,5
Diabetes		
No	5,442	96.2
Yes	216	3.8
Heavy drinking		
No	5,561	98.3
Yes	97	1.7

TABLE 4.1: General characteristics of patients with cancer, Lleida cancer register, 2012-2016.

	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
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

TABLE 4.2: Bivariate analysis by the density incidence and subsequent primary cancer.

a Subsequent primary cancer

CI: 1.2-5.7) and the 80 years and over age group (HR=3.0; 95% CI: 1.3-7.1). Smoking and heavy drinking presented similar aHRs but were not statistically significant.

Finally, potential interactions were calculated between smoking, heavy drinking, and diabetes with groups of cancer locations (types 0, 1, 2, and 3). However, there were no statistically significant interactions found. Please see the tables in the annex for further details (tables 4.7, 4.8, 4.9).

	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

TABLE 4.3: Cox proportional hazard analysis of risk of subsequent primary cancer.

a Confidence interval

DISCUSSION

This study highlighted that 4.1% of patients had an SPC after a mean follow-up of 1.9 years. Other studies registered up to 10% of patients with SPCs, but over a longer period of time [22]. This risk is increased by heavy drinking and smoking during the first follow-up years and suggests that these factors are risk factors for an SPC. Age also played an important role in the appearance of SPCs.

Previous studies analysed the association of SPC risk with risk factors. Barclay et al. studied the incidence of second primary cancer related to smoking in lung cancer cases [105]. They concluded that the incidence of subsequent cancer increased among lung cancer survivors. Another study also suggested the association of subsequent risk cancer with smoking and heavy alcohol use, especially among prostate cancer cases [106]. Another risk factor recently related to the SPC risk was the human papillomavirus (HPV). Wang et al. [107] highlighted the importance of finding strategies that prevent or detect SPC early

	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

TABLE 4.4: Cox proportional hazard analysis of risk of subsequent primary cancer stratified by men and women.

a Hazard ratio
b Confidence interval

in cancer survivors. Gilbert et al. [108] quantified the increased risk of a second HPV-associated cancer following diagnosis and also remarked on the implications for follow-up, screening and future therapeutic trials.

There were large differences in the percentage of cases among men during the first years following detection of the first primary cancer. Men were at higher risk for SPC, and both men and women with primary cancer would be at higher risk for a new cancer than men and women in the general population [22, 109]. This can be related to having a higher exposure to heavy drinking and smoking [110]. Ragusa et al., in a population study based on the Sicilian Registry (Italy) also observed a higher risk of SPC in men [111].

There were also significant differences between age groups. The 60-69 and 70-79 years age groups represented the vast majority of cases with SPCs (31.5% and 32.7% respectively). During the first primary follow-up years, these groups had a higher risk. The increase in risk with age has been observed in other studies and is related to the cumulative effect of exposure to possible risk factors, such as smoking, heavy drinking and obesity [112]. After the age of 80, the risk decreased notably because the risk of dying, in general, is higher and there were fewer follow-up years [113]. Similarly to the incidence of primary cancer in older people, genetics, underreporting and the age-related physiological effect could also explain the reduced risk [114].

No significant results were obtained for the primary cancer type and the body mass index. However, other studies have established an increased risk of SPC

from being overweight and obese. These factors are a risk for a significant group of primary cancers and could also behave as risk factors for SPC [96, 115]. In the case of diabetes, the study also estimated a higher risk of SPCs (HR=1.2), but this was not statistically significant. Other studies also suggest that diabetes could be a risk factor for primary cancer and SPCs [116].

We did not find significant associations for cancer location and the risk of SPCs. This may be related to the relatively short follow-up or the explanatory role of other variables included in the model such as smoking. However, other authors concluded that several types of primary cancer are significantly associated with a greater risk of developing and dying from an SPC compared with the general population. Chen et al. analysed the risk of an SPC among patients with oesophageal cancer regardless of risk factors [117]. The authors confirmed that the risk of an SPC remained high in the follow-up > 10 years. Liu et. al focused on the risk of SPCs in patients with head and neck carcinoma [118]. They concluded that there are significant differences in SPC incidence but did not take into account the exposure to risk factors. Another study demonstrated that some specific cancers increased the risk of SPC [22]. Sung et al. demonstrated that larynx and oesophageal cancers were closely associated with smoking and carried a risk of SPC.

Smoking and heavy drinking were significantly associated with the risk of SPCs and cor-roborated previous studies that reached similar conclusions [119, 120]. Both factors explain part of the risk of SPC associated with age and being male.

The Cox proportional hazard analysis included all the variables that remained in the model as risk factors for SPCs. The gender and age groups confirmed the correlation with the risk of SPCs during the first follow-up years after diagnosis. Smoking and heavy drinking were also significant in the independent model. These results corroborated the outcomes of previous studies [95]. Smokers were 1.3 times more at risk than non-smokers. Heavy drinking patients were 2.4 times more at risk than the others [121, 122]. A recent study found similar smoking outcomes where the authors concluded that smoking increases the SPC risk. In terms of heavy drinking, our study presented a higher risk even though both were significant [123]. The Cox proportional hazard analysis stratified by gender obtained similar results. Males maintained the smoking and heavy drinking associated with the risk of SPCs. Females also maintained similar risks as the previous model, even though smoking and heavy drinking were not significant due to fewer cases and a lower statistical power of the model.

Supplementary tables are included to analyse the association between SPC risk, risk factors and sociodemographic information stratified by BMI (table 4.5) and smoking (table 4.6). Table 4.5 shows the adjusted hazard ratios stratified by

normal weight, overweight and obesity. Similar risks were observed even though they lost statistical power due to fewer cases. In the case of table 4.6, similar risks were also obtained. The results for the BMI were similar to those obtained in the bivariate analysis and the same occurred with diabetes. However, for heavy drinking there was suggested association among non-smokers and smokers.

The study has some limitations. The short follow-up period after the first diagnosis may have reduced the probability of detecting risk factors related to body mass index and diabetes. This shorter follow-up did not enable more SPCs to be observed; therefore, a greater number of years of observation would improve the quality of the sample. Additionally, this shorter period caused the first patients included in the cohort to present a higher risk of an SPC due to the longer follow-up period. Moreover, the codification of the risk factors varied depending on the coding by the GP. Some cases of primary cancer without an SPC could develop SPCs later and become cases in future years. In this case, they also share risk factors for SPCs. Therefore, the effects of the risk factors might be underestimated, which could explain the observed lack of relationship between BMI and primary cancer location. In the case of smoking, consumers of electronic cigarettes and smokeless tobacco were not included because the software did not register them at the beginning of the cohort, although their prevalence among patients aged 50 or more years is low. The study's strengths included the fact that data are presented on risk factors, such as excess weight, smoking and heavy drinking. The study was performed with information from clinical practice, with physicians unaware of the study objectives, thus avoiding investigator bias.

CONCLUSIONS

This study carried out between 2012 and 2016 found an association between smoking and drinking habits and the risk of SPC during the first follow-up years. It also showed that men were more at risk than women. By age group, the risk of SPC increased until the age of 80 or more years during the first years after the primary cancer diagnosis. Therefore, despite some limitiations, such as the lack information on electronic cigarettes, smokeless tobacco and dietery factors, the results are in line with recently published reports. In general, these results reinforce the need for public health messaging about the harmful effects of smoking and heavy drinking. They also encourage continued research into SPCs to find new factors associated with these cancers and help the health system focus on preventing them. Patients with a primary cancer should be informed about the risks of SPCs.

SUPPLEMENTARY

	Normal-weight		Overweight		Obese	
	Adjusted HR[a]	95% CI[b]	Adjusted HR	95% CI	Adjusted HR	95% CI
Female	1.0	Ref. group	1.0	Ref. group	1.0	Ref. group
Man	1.6	0.9 – 2.8	1.3	0.8 – 2.1	1.4	0.8 – 2.4
Age 50-59	1.0	Ref. group	1.0	Ref. group	1.0	Ref. group
Age 60-69	1.7	0.8 – 3.6	1.5	0.8 - 2.9	2.0	0.9 – 4.7
Age 70-79	2.0	0.7 - 3.2	2.7	1.5 – 5.0	2.2	1.0 – 5.3
Age 80-	0.5	0.6 – 3.4	0.6	0.2 – 1.7	2.0	0.7 – 5.9
Diabetes	1.0	0.2 – 4.3	1.0	0.3 – 2.8	1.2	0.4 – 3.3
Smoking	1.2	0.7 – 2.0	1.6	1.0 – 2.4	1.2	0.6 – 2.0
Heavy drinking	2.3	0.7 – 7.7	1.2	0.3 – 4.7	4.6	1.6 – 13.1

TABLE 4.5: Cox proportional hazard analysis of risk of subsequent primary cancer stratified by normal-weight, overweight and obese.

a Hazard ratio
b Confidence interval

	Non-smoking		Smoking	
	Adjusted HR[a]	95% CI[b]	Adjusted HR	95% CI
Female	1.0	Ref. group	1.0	Ref. group
Man	1.4	1.0 – 2.1	1.2	0.6 – 2.5
Age 50-59	1.0	Ref. group	1.0	Ref. group2
Age 60-69	1.8	1.0 – 3.4	1.5	0.8 - 2.8
Age 70-79	2.4	1.3 – 4.3	2.0	1.1 – 3.8
Age 80-	1.5	0.7 – 3.0	0.5	0.1 – 1.8
Diabetes	0.6	0.2 – 1.7	1.9	0.8 – 4.5
Normal weight	1.0	Ref. group	1.0	Ref. group
Overweight	0.9	0.6 – 1.3	1.1	0.7 – 1.9
Obese	0.9	0.6 – 1.4	0.8	0.5 – 1.5
Heavy drinking	3.7	1.5 – 9.3	1.6	0.6 – 4.4

TABLE 4.6: Cox proportional hazard analysis of risk of subsequent primary cancer stratified by normal-weight, overweight and obese.

a Hazard ratio
b Confidence interval

	Hazard ratio	95% CI [a]
Type 3	1.0	Ref. group
Type 2	0.9	0.3 - 2.6
Type 1	2.6	1.0 - 7.1
Type 0	1.9	0.8 - 5.3
Diabetes	1.0	0.2 - 7.5
Diabetes * Type 2	N/A	N/A - N/A
Diabetes * Type 1	0.8	0.1 - 9.4
Diabetes * Type 0	1.2	0.1 - 10.3

TABLE 4.7: Cox proportional hazard analysis of risk of subsequent primary cancer. Interactions between diabetes and cancer location groups.

a Confidence interval

	Hazard ratio	95% CI[a]
Type 3	1.0	Ref. group
Type 2	0.9	0.3 - 2.6
Type 1	2.6	1.0 - 7.1
Type 0	1.9	0.8 - 5.3
Smoking	1.3	1.0 - 1.7
Smoking * Type 3	2.7	0.8 - 9.1
Smoking * Type 2	N/A	N/A - N/A
Smoking * Type 1	2.0	0.9 - 4.5
Smoking * Type 0	1.6	0.7 - 3.4

TABLE 4.8: Cox proportional hazard analysis of risk of subsequent primary cancer. Interactions between smoking and cancer location groups.

a Confidence interval

	Hazard ratio	95% CI [a]
Type 3	1.0	Ref. group
Type 2	0.9	0.3 - 2.6
Type 1	2.6	1.0 - 7.1
Type 0	1.9	0.8 - 5.3
Heavy drinking	2.4	1.3 - 4.8
Heavy drinking * Type 3	0.0	0.0 - Inf.
Heavy drinking * Type 2	N/A	N/A - N/A
Heavy drinking * Type 1	1.1	0.1 - 9.8
Heavy drinking * Type 0	1.2	0.1 - 10.8

TABLE 4.9: Cox proportional hazard analysis of risk of subsequent primary cancer. Interactions between heavy drinking and cancer location groups.

a Confidence interval

4.5 PAPER V: EFFECT OF ASPIRIN ON COLORECTAL CANCER

Authors: *Dídac Florensa, Jordi Mateo, Francesc Solsona, Leonardo Galvan, Ramon Piñol, Miquel Mesas, Leonardo Espinosa-Leal, Pere Godoy*
Journal: International Journal of Environmental Research and Public Health
Publisher: MDPI
Year: 2023
DOI: https://doi.org/10.3390/ijerph20054104
Keywords: *Colorectal cancer, aspirin use, excess weight, smoking, risky drinking*

Acetylsalicylic Acid Effect in Colorectal Cancer Taking into Account the Role of Tobacco, Alcohol and Excess Weight

Abstract: Excess weight, smoking and risky drinking are preventable risk factors for colorectal cancer (CRC). However, several studies have reported a protective association between aspirin and the risk of CRC. This article looks deeper into the relationships between risk factors and aspirin use with the risk of developing CRC. We made a retrospective cohort study of CRC risk factors and aspirin use in persons aged >50 years in Lleida province. The participants were inhabitants pre-scribed with medication between 2007 and 2016 and patients with CRC diagnosed between 2012 and 2016. Risk factors and aspirin use were studied using adjusted HR (aHR) with 95% confi-dence intervals (CI) using a Cox proportional hazard model. We included 154,715 inhabitants of Lleida aged > 50 years. Of patients with CRC, 62% were male (HR=1.8; 95% CI: 1.6-2.2), 39.5% were overweight (HR=2.8; 95% CI: 2.3-3.4) and 47.3% were obese (HR=3.0; 95% CI: 2.6-3.6). Cox regression showed an association between aspirin and CRC (aHR=0.7; 95% CI: 0.6-0.8), confirm-ing a protective effect against CRC and an association between the risk of CRC and excess weight (aHR=1.4; 95% CI: 1.2-1.7), smoking (aHR=1.4; 95% CI: 1.3-1.7) and risky drinking (aHR=1.6; 95% CI: 1.2-2.0). Our results show that aspirin use decreased the risk of CRC and corroborated the relationship between overweight, smoking and risky drinking and the risk of CRC.

International Journal of
Environmental Research and Public Health

MDPI

Article

Acetylsalicylic Acid Effect in Colorectal Cancer Taking into Account the Role of Tobacco, Alcohol and Excess Weight

Didac Florensa [1,2,*], Jordi Mateo [1], Francesc Solsona [1], Leonardo Galván [3], Miquel Mesas [4], Ramon Piñol [5], Leonardo Espinosa-Leal [6] and Pere Godoy [2,7,8]

1 Department of Computer Engineering and Digital Design, University of Lleida, Jaume II 69, 25001 Lleida, Spain
2 Population Cancer Registry in Lleida, Santa Maria University Hospital, Av. Alcalde Rovira Roure 44, 25198 Lleida, Spain
3 Pharmacy Unit, Catalan Health Service, Av. Alcalde Rovira Roure 2, 25006 Lleida, Spain
4 SAP-Argos Department, Santa Maria University Hospital, Av. Alcalde Rovira Roure 44, 25198 Lleida, Spain
5 Catalan Health Service, Department of Health, Av. Alcalde Rovira Roure 2, 25006 Lleida, Spain
6 Graduate School and Research, Arcada University of Applied Science, Jan-Magnus Janssonin Aukio 1, 00550 Helsinki, Finland
7 Lleida Biomedical Research Institute, Av. Alcalde Rovira Roure 80, 25198 Lleida, Spain
8 CIBER Epidemiology and Public Health (CIBERESP), Health Institute Carlos III, 28029 Madrid, Spain
* Correspondence: didac.florensa@gencat.cat

Abstract: Excess weight, smoking and risky drinking are preventable risk factors for colorectal cancer (CRC). However, several studies have reported a protective association between aspirin and the risk of CRC. This article looks deeper into the relationships between risk factors and aspirin use with the risk of developing CRC. We performed a retrospective cohort study of CRC risk factors and aspirin use in persons aged >50 years in Lleida province. The participants were inhabitants with some medication prescribed between 2007 and 2016 that were linked to the Population-Based Cancer Registry to detect CRC diagnosed between 2012 and 2016. Risk factors and aspirin use were studied using the adjusted HR (aHR) with 95% confidence intervals (CI) using a Cox proportional hazard model. We included 154,715 inhabitants of Lleida (Spain) aged >50 years. Of patients with CRC, 62% were male (HR = 1.8; 95% CI: 1.6–2.2), 39.5% were overweight (HR = 2.8; 95% CI: 2.3–3.4) and 47.3% were obese (HR = 3.0; 95% CI: 2.6–3.6). Cox regression showed an association between aspirin and CRC (aHR = 0.7; 95% CI: 0.6–0.8), confirming a protective effect against CRC and an association between the risk of CRC and excess weight (aHR = 1.4; 95% CI: 1.2–1.7), smoking (aHR = 1.4; 95% CI: 1.3–1.7) and risky drinking (aHR = 1.6; 95% CI: 1.2–2.0). Our results show that aspirin use decreased the risk of CRC and corroborate the relationship between overweight, smoking and risky drinking and the risk of CRC.

Keywords: colorectal cancer; aspirin use; excess weight; smoking; risky drinking

Citation: Florensa, D.; Mateo, J.; Solsona, F.; Galván, L.; Mesas, M.; Piñol, R.; Espinosa-Leal, L.; Godoy, P. Acetylsalicylic Acid Effect in Colorectal Cancer Taking into Account the Role of Tobacco, Alcohol and Excess Weight. *Int. J. Environ. Res. Public Health* **2023**, *20*, 4104. https://doi.org/10.3390/ijerph20054104

Academic Editor: Paul B. Tchounwou

Received: 4 January 2023
Revised: 17 February 2023
Accepted: 22 February 2023
Published: 24 February 2023

1. Introduction

Colorectal cancer (CRC) is the third leading cause of cancer death globally and the second in Europe, and its incidence is steadily rising in developing nations [1], with nearly 520,000 new cases in Europe in 2020 [2], even though a large proportion of these case are highly preventable [3]. A study in nine European countries found that approximately 20% of CRC cases may be related to overweight, smoking and risky drinking [4]. In contrast, studies have shown that long-term aspirin use may prevent CRC [5,6].

Shaukat et al. found a direct relationship between the body mass index (BMI) and long-term CRC mortality and suggested that BMI modulation may reduce the risk of CRC mortality [7]. A recent study has shown the role of obesity and overweight in early-onset CRC, and concluded that obesity is a strong risk factor [8]. Ghazaleh Dashti et al. found an

association between risky drinking and an increased risk of CRC [9]. Likewise, a study has suggested an association between passive smoking and the risk of CRC [10].

Some studies have found a protective effect of aspirin against CRC [11] and various studies have concluded that aspirin reduces the overall risk of CRC recurrence and mortality and colorectal adenomas. Ma et al. recently found that aspirin, including low-dose aspirin, reduced the risk of CRC [12]. A recent study by Zhang et al. on the effect of aspirin use for 5 and 10 years found that the continuous use of aspirin increases the protective effect on CRC [13]. A Danish study also found that the continuous use of low-dose aspirin was associated with a reduced CRC risk [14]. Some studies have shown differing results on the protective effect of aspirin due to the different designs used, the type of follow-up, the recorded aspirin consumption and the size and type of population. Although the data seem compelling, a limitation of these analyses is that they do not take into account risk factors for CRC [15]. These previous studies investigated the association between the use of aspirin and CRC, but they did not study the role played by risk factors such as tobacco smoking, alcohol or excess weight. In this study, we explore how these factors, combined with aspirin use, affect the risk of CRC in a particular society.

The objective of this study was to determine the protective effect of aspirin against CRC, taking into account the effect of other risk factors (overweight/obesity, risky drinking and smoking), in Lleida, a province in Catalonia, Spain, with a large rural population and an agri-food industry that may present specific risk factors [16,17].

2. Materials and Methods

2.1. Study Population

We conducted a retrospective cohort study of aspirin use and risk factors to analyze the impact of these factors on the risk of CRC. We carried out the study on 154,715 inhabitants of Lleida aged >50 years at the start of the study period, with data available on aspirin use from 1 January 2007 to 31 December 2016 in the Catalan Health Service (CatSalut) system. The reason for selecting this period was to ensure that those CRC cases detected in 2012 had the opportunity to be exposed to aspirin for at least five years. This population was linked to the Lleida Population-based Cancer Registry to detect CRC diagnosed between 2012 and 2016.

Data on aspirin use were obtained from the number of packages dispensed by pharmacies. Catalonia has a public health system in which medicines are dispensed in pharmacies after presenting a doctor's prescription. Drugs administered to hospitalized patients and those prescribed by private providers are not registered in the CatSalut system, and therefore were not included in this study. The CRC cases in the sample were obtained from the Lleida Population-Based Cancer Registry, and the demographic characteristics of participants, including age and sex, were obtained from the CatSalut system. Figure 1 shows a flowchart of the study population. Initially, the pharmacy database registered 724,070 inhabitants with any prescription, although 346,365 were excluded because they did not reside in the Lleida region. Another exclusion criterion was age. We only included inhabitants aged >50 years at the start of the observed period (2007), resulting in 154,717 inhabitants. We also excluded those inhabitants who did not register the risk factors correctly, although the cases excluded were minimal.

As has been presented before, this study included different databases. To enable this linkage, it was necessary to use a personal identification code called CIP. This code is unique to each inhabitant who resides in Catalonia and permits us to identify them in the Catalan Health Service and its registers (hospitals, pharmacies or primary care centers).

2.2. Data Collection

Data on CRC diagnoses were obtained from the Lleida Population-Based Cancer Registry using five consecutive years of incidence data, from 2012 to 2016. This period was chosen as the available years validated by the professionals of the register. Potential CRC cases were validated by checking medical records. We used hospital and pathological

anatomy records as the main information sources. Cancers were identified following the rules defined by the International Association of Cancer Registries, the International Association for Research on Cancer and the European Network of Cancer Registries.

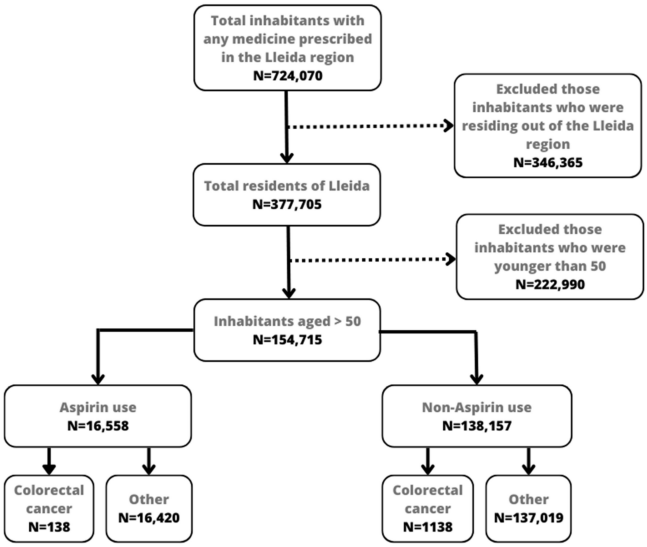


Figure 1. Flowchart of subjects included for the analysis.

The risk factors included were risky drinking, smoking and body mass index. This information was extracted using the eCAP software (V 20.4.3) used by primary care physicians to record all patient information, which registers information from 2001. The values of these variables at the time that this study started were obtained. Body mass index (BMI) was calculated by the weight and height of the patient using the formula $\text{BMI} = \text{weight(kg)}/\text{height(m)}^2$ and categorized as follows: 18.5–24.9 normal weight, 25–29.9 overweight and >30 obesity [18]. The ICD-10 international criteria identified risky drinking and smoking. The ICD-10 code for risky drinking is F10.2, and those for smoking are F17 (mental and behavioral disorders due to tobacco use) and Z72 (tobacco use). Risky drinking was defined as consumption of >40 g/day of alcohol in men and >24 g/day in women [19]. The Spanish Health Ministry defined these grams per day with the supervision of the WHO [20]. The software also states the date of smoking onset. Smokers were defined as exposure for >5 years before the start of the study. The reason for using a period of five years was due to a previous study that suggested that this period might increase the risk of cancer [21]. Former smokers were considered smokers because the observed points in the dataset were minimal, and adding this new category could have imbalanced the dataset. General characteristics are represented in Table 1.

2.3. Exposure

Aspirin was categorized according to the Anatomical Therapeutic Chemical (ATC) classification system as A01AD05 (acetylsalicylic acid) medication. The use of aspirin was evaluated based on the defined daily dose (DDD) and the milligrams (mg) accumulated dose consumed by each patient throughout the study period. The DDD is a technical unit of measurement that corresponds to the daily maintenance dose of a drug for its main indication in adults and a given route of administration. The DDDs of active ingredients

are established by the WHO and published on the WHO Collaborating Center for Drug Statistics Methodology website [22,23].

Table 1. General characteristics of the inhabitants included in this study.

	Total		Men		Women	
	N	%	N	%	N	%
Gender						
Female	80,865	52.3	-	-	-	-
Male	73,850	47.7	-	-	-	-
Age						
[50–59)	46,454	30.0	24,080	32.6	22,374	27.7
[60–69)	35,819	23.2	17,655	23.9	18,164	22.5
[70–79)	28,138	18.2	12,842	17.4	15,296	18.9
[80–89)	23,651	15.3	10,176	13.8	13,475	16.7
[90–)	20,653	13.3	9097	12.3	11,556	14.3
Aspirin						
Non-use	138,157	88.1	66,695	90.3	71,462	88.4
Use	16,558	11.9	7155	9.7	9403	11.6
Body Mass Index						
Normal weight	47,761	30.9	21,088	28.6	26,673	33.0
Overweight	51,022	33.0	27,005	36.6	24,017	29.7
Obesity	55,932	36.1	25,757	34.9	30,175	37.3
Risky drinking						
No	151,323	97.8	71,997	97.5	79,326	98.1
Yes	3392	2.2	1853	2.5	1539	1.9
Smoking						
No	140,749	90.9	62,995	85.3	77,754	96.2
Yes	13,966	9.1	10,855	14.7	3111	3.8

Exposure was determined from computerized pharmacy data and consisted of the total DDD dispensed to an individual during the study period. For instance, if a person consumed aspirin for a while, then stopped using it and later started again, the total DDD consumed during the following period was considered. To be considered as exposed to aspirin, the total number of years of consumption had to be $\geq$5 years. The number of years was based on previous studies, which suggested this period as the minimum for aspirin to have a protective effect [13,24]. To consider exposure to aspirin, the minimum consumed daily was >75 mg [25,26]. The number of DDD calculated this value in mg.

2.4. Statistical Analysis

Descriptive analyses were performed to evaluate the association between characteristics at baseline, exposure and outcomes. Patients' characteristics, risk factors and aspirin exposure were analyzed to determine the association with the risk of CRC. The incidence rate of CRC was calculated to each factor over a specified period. A bivariate analysis was initially used to estimate the crude hazard ratios for the association between aspirin consumption and the risk of incident CRC.

A Cox proportional hazard model was used to determine the HR and the corresponding 95% CI. The models were adjusted by sex, age, aspirin exposure, BMI, risky drinking and smoking. Subsequently, stratified models were calculated by sex.

The probability values for the statistical tests were two-tailed, and a CI that did not contain 1.0 was regarded as statistically significant. Results with wide CIs should be interpreted cautiously. All statistical analyses were performed using R (R Core Team 2019), an open-source programming language and environment for statistical analysis and graphic representation.

3. Results

We analyzed 154,715 inhabitants of Lleida aged >50 years, of whom 1276 (0.8%) had CRC between 2012 and 2016. The mean CRC incidence rate and the total cases by sex and age group for the five study years are shown in Figure 2a,b.

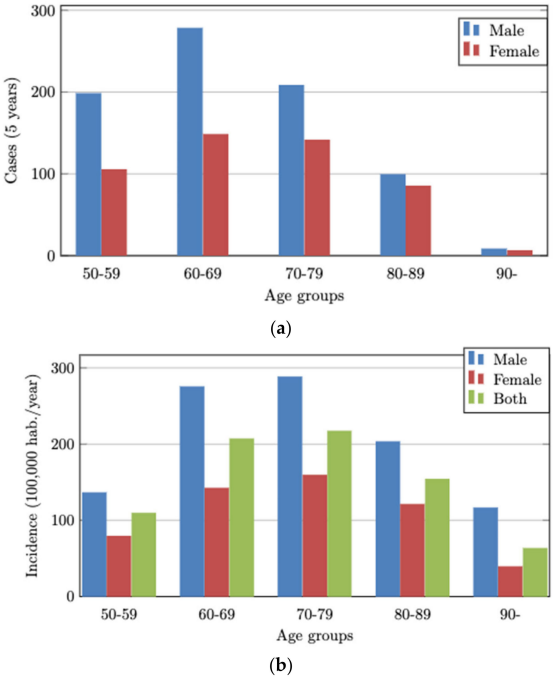


Figure 2. (**a**) Total CRC cases by age groups and sex for 5 years. (**b**) Mean observed CRC incidence by age groups and sex during one year.

The sociodemographic information and aspirin exposure in patients with CRC (Table 2) were analyzed in the bivariate analysis.

We recorded 485 (0.8 × 1000) females and 791 (1.4 × 1000) males (HR = 1.9; 95% CI; 1.6–2.0) with CRC. Most patients were from the 60–69 years (HR = 1.8; 95% CI; 1.6–2.1) and 70–79 years age groups (HR = 2.0; 95% CI; 1.9–2.6). There were 1138 (1.2 × 1000) CRC cases without aspirin consumption and 138 with aspirin consumption (1.0 × 1000) (HR = 0.9; 95% CI; 0.8–1.1). There were 504 (1.2 × 1000) cases with overweight (HR = 2.5; 95% CI; 2.2–3.1) and 603 (1.3 × 1000) with obesity (HR = 2.7; 95% CI; 2.3–3.3), and there were 56 (2.2 × 1000) cases with risky drinking (HR = 2.1; 95% CI; 1.6–2.7), while 220 (2.0 × 1000) were smokers (HR = 2.0; 95% CI; 1.8–2.4).

Cox regression showed variations in the outcomes (Table 3). Sex, age and aspirin exposure were significantly associated with CRC. The adjusted HR (aHR) for males was 1.8 (95% CI: 1.6–2.1) and 1.8 (95% CI: 1.6–2.1) in the 60–69 years age group, 2.3 (95% CI: 1.9–2.7) in the 70–79 years age, 2.2 (95% CI: 1.8–2.6) in the 80–89 years age group and 0.2 (95% CI: 0.1–0.3) in the 90 years age group. Aspirin consumption had an aHR of 0.7 (95% CI: 0.6–0.8). The BMI also was significant. Overweight had an aHR of 1.4 (95% CI: 1.2–1.7) and obesity of 1.5 (95% CI: 1.3–1.8). Risky drinking had a significant aHR of 1.6 (95%

CI: 1.2–2.0) and smoking an aHR of 1.4 (95% CI: 1.3–1.7). Figure 3 represents the adjusted hazard ratios graphically.

Table 2. Bivariate analysis with the observed years and CRC patients.

	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.

Table 3. Multivariate analysis—Cox regression.

	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.

HRs were adjusted by gender, age, aspirin use, BMI, risky drinking and smoking.

Table 4 shows the results of the Cox regression stratified by sex. In the case of males, the results were similar to the general table. In this model, aspirin exposure remained significant (aHR: 0.7; 95% CI: 0.6–0.8), as did the BMI, risky drinking and smoking. In females, aspirin use remained significant (aHR: 0.6; 95% CI: 0.4–0.8), but, of the risk factors, only obesity remained significant (aHR: 1.4; 95% CI: 1.2–1.9). Figure 4 represents the adjusted hazard ratios graphically.

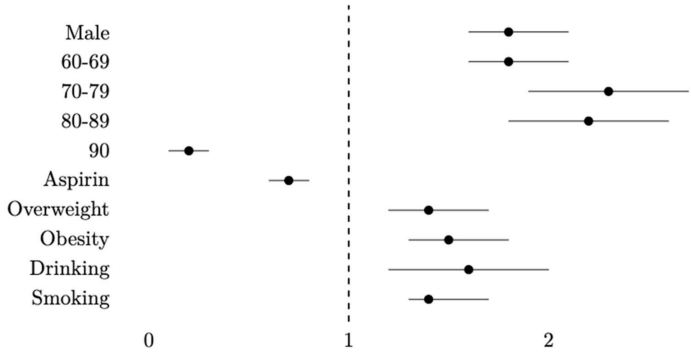


Figure 3. Hazard ratios by Cox regression adjusted for gender, age, aspirin use, BMI, risky drinking and tobacco smoking.

Table 4. Adjusted Cox regression model stratified by men and women.

	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.

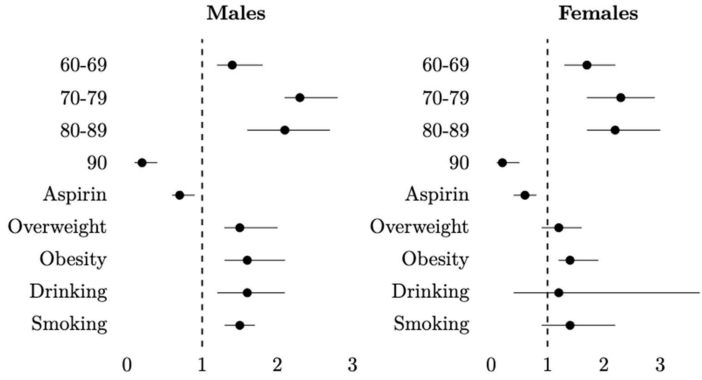


Figure 4. Hazard ratios by Cox regression stratified by males and females. Adjusted for age, aspirin use, BMI, risky drinking and tobacco smoking.

HRs were adjusted by age, aspirin use, BMI, risky drinking and smoking.

4. Discussion

Our results confirm the negative association between aspirin consumption and CRC independently of the other risk factors measured. Males may be at a higher risk of CRC than females but aspirin may be slightly more protective in females.

Reports support a delayed effect of aspirin on CRC [27]. A meta-analysis by Rothwell et al. examined the long-term effects of aspirin on CRC outcomes using trials of aspirin [28]. Studies on the impact of aspirin in CRC prevention have been published [6,29], although the effects of risk factors and aspirin use together have not yet been analyzed. Therefore, our findings corroborate the research in the field highlighting the protective effect of aspirin and go beyond comparing this positive effect with the negative effects caused by several risk factors.

Several recent studies have suggested an association between aspirin use and some specific cancers. Ciu et al. concluded that high-dose aspirin reduced the risk of pancreatic cancer [30]. Jacobo et al. analyzed studies on the relationship between aspirin and breast cancer [31] and concluded that aspirin consumption reduced the relative risk of breast cancer. Sieros et al. suggested that aspirin reduced the risk of esophageal cancer [32].

We found significant differences according to sex, suggesting that men have a higher risk of developing CRC. It has been reported that men have higher cumulative levels of smoking than women and a higher alcohol intake, which may explain the higher risk [33].

People aged between 60 and 80 years had a higher risk of CRC and the 80–89 years and 90–99 years age groups had a lower risk [34,35]. Older adults may have a differential mechanism compared with younger people. For example, aging is associated with alterations in DNA methylation, which may affect the susceptibility to cancer. The gut microbiota of older people differs from that of younger adults, which may influence drug metabolism and inflammatory processes. Genetics, underreporting and age-related physiological effects could explain the reduced risk [36].

We found some differences with respect to risk factors, such as overweight/obesity, risky drinking and smoking. Overweight represented 39.5% of total CRC cases and obesity 47.3%. Therefore, approximately 85% of patients with CRC presented excess weight, suggesting exposure to a poor diet. These results corroborated previous studies [37–40]. Excess weight is one of the most important risk factors for CRC. Individuals with a higher BMI have higher levels of chronic inflammation, and obesity may act through the gut microbiome on colorectal tumorigenesis and also promotes colorectal cancer in mice. There were notable differences in risky drinking. Patients with risky drinking had a higher risk of CRC (HR = 2.2). Meta-analyses of case–control and cohort studies suggest that high alcohol consumption might be associated with an increased risk of colorectal cancer. The epidemiological evidence has been complemented by molecular evidence on the mechanisms that could explain this association [17,41]. Similar results were obtained for smoking (HR = 2.0). The crude HR obtained also indicated this association between smoking and CRC [42]. Smoking was more closely associated with colorectal tumors that arose from non-conventional pathways, such as the serrated polyps pathway, and smoking was significantly associated with the risk of advanced serrated polyps in a screening population.

The Cox regression included all the remaining model variables, such as risk factors and aspirin exposure for CRC. Sociodemographic variables such as gender and age confirmed the correlation with CRC. Males were 1.8 times more at risk than females. This may be related to men having excess body weight and higher exposure to alcohol and smoking than women [43]. Regarding the age groups, the results confirmed that the 70–79 years age group had the highest risk, which was 2.3 times greater than the 50–59 years (ref. group) and the 69–69 years and 80–89 years age groups. Other studies found similar outcomes on the incidence and association related to CRC [44,45].

The use of aspirin for $\geq$5 years was significant in the Cox regression. The analysis suggested that aspirin decreased the risk of CRC. The HR was 0.7 (95% CI: 0.6–0.8), meaning

that it reduced the risk of CRC by 30% [46,47]. Studies have found reductions of 20–30% [46] and 27% [47] in the risk for CRC. The risk factors were correlated with an increased risk of CRC. Overweight and obesity were significantly associated with a CRC risk 1.4 and 1.5 times higher, respectively. Obesity had a higher risk, although the HR was similar [48]. Risky drinking and smoking also had a significant HR. Risky drinking had a 1.6 times higher risk and smoking a 1.4 times higher risk. Other studies also found these associations [49,50], with a 1.3 and 1.2 times higher risk for risky drinking, respectively, and a 1.2 higher risk for smoking [50].

The Cox regression stratified by sex also obtained significant results. Men and women had similar outcomes according to age. The trends were the same as the non-stratified regression. The risk of CRC was higher in people aged 60–89 years in both sexes. The use of aspirin also maintained the association with a reduced CRC risk. Specifically, in females, aspirin could prevent CRC, in the best case, by up to 40%. A similar percentage was obtained by Cook et al. [51] in a randomized controlled trial, which showed 42% aspirin protection against CRC risk among women. These results corroborated the fact that aspirin reduces the risk of CRC in both sexes [52]. However, the results related to risk factors were significant in males. Overweight and obesity were associated with a 1.5 and 1.6 times higher risk of CRC, respectively [53]. Risky drinking and smoking were also correlated with the CRC risk. The differences between males and females may be that males more often have a poor diet and drink and smoke more than females [54]. In females, only obesity was significantly associated with an increased risk of CRC. Moreover, as a previous study concluded [55], only excess weight among men was significantly associated with increased CRC risk. In addition, the authors suggested that this risk might be reversed in obese men taking aspirin. Similarly, in our study, in the analysis stratified by normal weight and overweight/obesity, aspirin was protective against CRC in both groups, but it was only statistically significant in overweight/obese patients (Supplementary Table S1) [55]. The remaining risk factors were not related to CRC, although the HR was >1 in all of them. Individual susceptibility and the type of exposure may explain these results. Men probably have a different pattern of consumption than women and are more intensely exposed to alcohol and smoking. In addition, it seems that without the effect of aspirin, these factors are related to CRC (Supplementary Table S2).

The preventive effect of aspirin has been attributed to the inhibition of cyclooxygenase (COX), the enzyme responsible for the synthesis of prostaglandins [56,57]. COX-2 is abnormally expressed in many cancer cell lines and is involved in the processes of carcinogenesis, angiogenesis and tumor growth. Additional mechanisms of aspirin include the induction of apoptosis through COX-independent pathways. Future research should also study the role of aspirin metabolites and the role of the intestinal microbiota in cancer prevention against CRC.

Long-term aspirin is prescribed for patients with a high cardiovascular risk of non-focal continuous pain due to arthritis, and the results of this study may support this indication [58,59].

The study has some limitations. Firstly, some patients could buy aspirin directly in pharmacies without a doctor's prescription, and this consumption is underreported. Second, some patients may not take the medication, even if they have purchased it at the pharmacy, and, in this case, aspirin use will be overreported. Third, although the Population-Based Cancer Registry is exhaustive, it cannot be ruled out that some cases were diagnosed in hospitals in other territories and some cases have not been correctly registered. We were unable to study the dose–response relationship between low-dose aspirin and CRC because more than 90% of aspirin use in this study was at a dose of 100 mg/day, which did not allow us to assess the highest related dose effect. Another limitation is the lack of specification of the types of CRC, such as familial polyposis or familiar cancer genetics, as a possible bias. This information was not taken into account in the register. A limitation that must be considered concerns the CRC cases diagnosed before 2012. These cases were not included because the Cancer Registry started registering cases in

2012. However, CRC cases prior to 2012 would not have had the opportunity to be exposed to risk factors or aspirin for a period of 5 or more years and would not have been recorded as incident cases in this study. Despite this, CRC is a type of cancer that can present another primary cancer a few years later; therefore, some CRC cases may be included. moreover, related to the risk factors, some bias was present due to under-reporting, although the percentage of our cases was similar to the prevalence observed in Catalonia. Finally, the impact of these excluded cases was minimal because they were younger than 50, where cancer may be unrelated to risk factors, or they were cases from other regions, and few patients had to be excluded due to a lack of information on risk factors.

The study's strengths included the fact that data are presented on risk factors, such as excess weight, smoking and risky drinking. The study was performed with information from clinical practice, with physicians unaware of the study objectives, which avoided investigator bias.

5. Conclusions

This retrospective study found an association between aspirin use for $\geq$5 years and a reduced risk of CRC. The protective effect due to aspirin was higher in women. The results also showed an association between the risk of CRC and risk factors such as overweight, obesity, smoking and risky drinking, specifically in men. Moreover, the risk of CRC in women was significantly associated with obesity. The 70–79 and 80–89 age groups had a higher risk of CRC in men and women. Therefore, despite some limitations, such as the lack of information on food or dietary factors or some bias in the aspirin prescriptions, the results are according to the recently published literature.

In general, these results reinforce the need for public health messaging about the harmful effects of smoking, alcohol use and excess weight, and the use of aspirin to prevent CRC under prescription. They also encourage continued research into CRC to find new factors or interactions among them associated with this cancer. They also may help the health system to focus on preventing them and recommend the continuous use of aspirin under medical supervision.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/ijerph20054104/s1, Table S1: Cox regression stratified by excess weight (obesity-overweight) and normal-weight; Table S2: Cox regression stratified by aspirin and non-aspirin use.

Author Contributions: D.F. designed the study, analyzed the data and wrote the manuscript. J.M. analyzed the data and supervised the study. F.S. and P.G. supervised the study. R.P. and L.G. collected the medicine dataset. M.M. extracted the cancer information. L.E.-L. supervised the analysis. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by contract 2019-DI-43 of the Industrial Doctorate Program of the Government of Catalonia. Project PID2020-113614RB-C22 was funded by the Spanish Ministry of Science and Innovation. Some authors are members of the 2014-SGR163 research group, funded by the Generalitat de Catalunya.

Institutional Review Board Statement: The study was approved by the Clinical Investigation Ethical Committee (CEIC 21/190-P, approval 8 September 2021) of IDIAP Jordi Gol. As it was a retrospective cohort study and the patients were blinded to the investigators, no written informed consent was necessary according to the CEIC. All statistical calculations were carried out in accordance with the relevant guidelines and regulations.

Informed Consent Statement: As this was a retrospective cohort study and the patients were blinded to the investigators, no written informed consent was necessary according to the CEIC.

Data Availability Statement: The dataset is available from the corresponding author upon reasonable request.

Acknowledgments: The authors acknowledge the CSC-IT Center for Science, Finland, for the computing resources, and the Arnau de Vilanova University Hospital, Santa Maria University Hospital and Catalan Health Service in Lleida for the support and resources provided to perform this study.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rawla, P.; Sunkara, T.; Barsouk, A. Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. *Gastroenterol. Rev. Przegląd Gastroenterol.* **2019**, *14*, 89–103. [CrossRef]
2. Ferlay, J.; Ervik, M.; Lam, F.; Colombet, M.; Mery, L.; Piñeros, M.; Znaor, A.; Soerjomataram, I. *Global Cancer Observatory: Cancer Today*; International Agency for Research on Cancer: Lyon, France, 2020. Available online: https://gco.iarc.fr/today (accessed on 15 February 2023).
3. Cardoso, R.; Guo, F.; Heisser, T.; Hackl, M.; Ihle, P.; De Schutter, H.; Van Damme, N.; Valerianova, Z.; Atanasov, T.; Májek, O.; et al. Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: An international population-based study. *Lancet Oncol.* **2021**, *22*, 1002–1013. [CrossRef]
4. Aleksandrova, K.; Pischon, T.; Jenab, M.; Bueno-de-Mesquita, H.B.; Fedirko, V.; Norat, T.; Romaguera, D.; Knüppel, S.; Boutron-Ruault, M.C.; Dossus, L.; et al. Combined impact of healthy lifestyle factors on colorectal cancer: A large European cohort study. *BMC Med.* **2014**, *12*, 168. [CrossRef] [PubMed]
5. Burn, J.; Sheth, H.; Elliott, F.; Reed, L.; Macrae, F.; Mecklin, J.-P.; Möslein, G.; McRonald, F.E.; Bertario, L.; Evans, D.G.; et al. Cancer prevention with aspirin in hereditary colorectal cancer (Lynch syndrome), 10-year follow-up and registry-based 20-year data in the CAPP2 study: A double-blind, randomised, placebo-controlled trial. *Lancet* **2020**, *395*, 1855–1863. [CrossRef] [PubMed]
6. Serrano, D.; Patrignani, P.; Stigliano, V.; Turchetti, D.; Sciallero, S.; Roviello, F.; D'Arpino, A.; Grattagliano, I.; Testa, S.; Oliani, C.; et al. Aspirin Colorectal Cancer Prevention in Lynch Syndrome: Recommendations in the Era of Precision Medicine. *Genes* **2022**, *13*, 460. [CrossRef] [PubMed]
7. Shaukat, A.; Dostal, A.; Menk, J.; Church, T.R. BMI Is a Risk Factor for Colorectal Cancer Mortality. *Dig. Dis. Sci.* **2017**, *62*, 2511–2517. [CrossRef] [PubMed]
8. Li, H.; Boakye, D.; Chen, X.; Hoffmeister, M.; Brenner, H. Association of Body Mass Index With Risk of Early-Onset Colorectal Cancer: Systematic Review and Meta-Analysis. *Am. J. Gastroenterol.* **2021**, *116*, 2173–2183. [CrossRef] [PubMed]
9. Dashti, S.G.; Buchanan, D.D.; Jayasekara, H.; Ouakrim, D.A.; Clendenning, M.; Rosty, C.; Winship, I.M.; MacRae, F.A.; Giles, G.G.; Parry, S.; et al. Alcohol consumption and the risk of colorectal cancer for mismatch repair gene mutation carriers. *Cancer Epidemiol. Biomarkers Prev.* **2017**, *26*, 366–375. [CrossRef] [PubMed]
10. Yang, C.; Wang, X.; Huang, C.H.; Yuan, W.J.; Chen, Z.H. Passive Smoking and Risk of Colorectal Cancer: A Meta-analysis of Observational Studies. *Asia Pac. J. Public Health* **2016**, *28*, 394–403. [CrossRef] [PubMed]
11. Coyle, C.; Cafferty, F.H.; Langley, R.E. Aspirin and Colorectal Cancer Prevention and Treatment: Is It for Everyone? *Curr. Colorectal Cancer Rep.* **2016**, *12*, 27–34. [CrossRef]
12. Ma, S.; Han, T.; Sun, C.; Cheng, C.; Zhang, H.; Qu, G.; Bhan, C.; Yang, H.; Guo, Z.; Yan, Y.; et al. Does aspirin reduce the incidence, recurrence, and mortality of colorectal cancer? A meta-analysis of randomized clinical trials. *Int. J. Colorectal Dis.* **2021**, *36*, 1653–1666. [CrossRef]
13. Guo, C.G.; Ma, W.; Drew, D.A.; Cao, Y.; Nguyen, L.H.; Joshi, A.D.; Ng, K.; Ogino, S.; Meyerhardt, J.A.; Song, M.; et al. Aspirin Use and Risk of Colorectal Cancer Among Older Adults. *JAMA Oncol.* **2021**, *7*, 428–435. [CrossRef]
14. Friis, S.; Riis, A.H.; Erichsen, R.; Baron, J.A.; Sørensen, H.T. Low-Dose Aspirin or Nonsteroidal Anti-inflammatory Drug Use and Colorectal Cancer Risk. *Ann. Intern. Med.* **2015**, *163*, 347–355. [CrossRef] [PubMed]
15. Cho, M.H.; Yoo, T.G.; Jeong, S.M.; Shin, D.W. Association of Aspirin, Metformin, and Statin use with gastric cancer incidence and mortality: A nationwide cohort study. *Cancer Prev. Res.* **2021**, *14*, 95–104. [CrossRef] [PubMed]
16. Henley, S.J.; Anderson, R.N.; Thomas, C.C.; Massetti, G.M.; Peaker, B.; Richardson, L.C. Invasive cancer incidence, 2004–2013, and deaths, 2006–2015, in nonmetropolitan and metropolitan counties—United States. *MMWR Surveill. Summ.* **2017**, *661*, 1–13. [CrossRef] [PubMed]
17. Florensa, D.; Godoy, P.; Mateo, J.; Solsona, F.; Pedrol, T.; Mesas, M.; Pinol, R. The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables and Cancer Incidence. *IEEE J. Biomed. Health Inform.* **2021**, *25*, 3659–3667. [CrossRef] [PubMed]
18. WHO | World Health Organization. Available online: https://www.who.int/en (accessed on 18 January 2023).
19. Sierosławski, J.; Foster, J.; Moskalewicz, J. Survey of European drinking surveys. Alcohol survey experiences of 22 European countries. *Drugs Educ. Prev. Policy* **2013**, *20*, 383–398. [CrossRef]
20. Low Risk Alcohol Consumption Thresholds. Update on the Risks Related to Alcohol Consumption Levels, Consumption Patterns and Type of Alcoholic Beverages. Available online: https://www.sanidad.gob.es/profesionales/saludPublica/prevPromocion/Prevencion/alcohol/docs/Low_Risk_Alcohol_Consumption_Thresholds_Part1.pdf (accessed on 20 January 2023).
21. Kenfield, S.A.; Stampfer, M.J.; Rosner, B.A.; Colditz, G.A. Smoking and smoking cessation in relation to mortality in women. *JAMA* **2008**, *299*, 2037–2047. [CrossRef]
22. WHOCC—ATC/DDD Index. Available online: https://www.whocc.no/atc_ddd_index/ (accessed on 22 December 2022).

23. Torres-Bondia, F.; Dakterzada, F.; Galván, L.; Buti, M.; Besanson, G.; Gill, E.; Buil, R.; de Batlle, J.; Piñol-Ripoll, G. Proton pump inhibitors and the risk of Alzheimer's disease and non-Alzheimer's dementias. *Sci. Rep.* **2020**, *10*, 21046. [CrossRef]
24. Chan, A.T.; Giovannucci, E.L.; Meyerhardt, J.A.; Schernhammer, E.S.; Wu, K.; Fuchs, C.S. Aspirin Dose and Duration of Use and Risk of Colorectal Cancer in Men. *Gastroenterology* **2008**, *134*, 21–28. [CrossRef]
25. Hwang, I.C.; Chang, J.; Kim, K.; Park, S.M. Aspirin Use and Risk of Hepatocellular Carcinoma in a National Cohort Study of Korean Adults. *Sci. Rep.* **2018**, *8*, 4968. [CrossRef] [PubMed]
26. Thun, M.J.; Jacobs, E.J.; Patrono, C. The role of aspirin in cancer prevention. *Nat. Rev. Clin. Oncol.* **2012**, *9*, 259–267. [CrossRef] [PubMed]
27. Rothwell, P.M.; Wilson, M.; Elwin, C.E.; Norrving, B.; Algra, A.; Warlow, C.P.; Meade, T.W. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *Lancet* **2010**, *376*, 1741–1750. [CrossRef] [PubMed]
28. Rothwell, P.M.; Fowkes, F.G.R.; Belch, J.F.; Ogawa, H.; Warlow, C.P.; Meade, T.W. Effect of daily aspirin on long-term risk of death due to cancer: Analysis of individual patient data from randomised trials. *Lancet* **2011**, *377*, 31–41. [CrossRef]
29. Maniewska, J.; Jeżewska, D. Non-Steroidal Anti-Inflammatory Drugs in Colorectal Cancer Chemoprevention. *Cancers* **2021**, *13*, 594. [CrossRef] [PubMed]
30. Cui, X.J.; He, Q.; Zhang, J.M.; Fan, H.J.; Wen, Z.F.; Qin, Y.R. High-dose aspirin consumption contributes to decreased risk for pancreatic cancer in a systematic review and meta-analysis. *Pancreas* **2014**, *43*, 135–140. [CrossRef]
31. Jacobo-Herrera, N.J.; Pérez-Plasencia, C.; Camacho-Zavala, E.; Figueroa González, G.; López Urrutia, E.; Garćia-Castillo, V.; Zentella-Dehesa, A. Clinical evidence of the relationship between aspirin and breast cancer risk (review). *Oncol. Rep.* **2014**, *32*, 451–461. [CrossRef]
32. Liao, L.M.; Vaughan, T.L.; Corley, D.A.; Cook, M.B.; Casson, A.G.; Kamangar, F.; Abnet, C.C.; Risch, H.A.; Giffen, C.; Freedman, N.D.; et al. Nonsteroidal Anti-inflammatory Drug Use Reduces Risk of Adenocarcinomas of the Esophagus and Esophagogastric Junction in a Pooled Analysis. *Gastroenterology* **2012**, *142*, 442–452.e5. [CrossRef]
33. Yang, Y.; Wang, G.; He, J.; Ren, S.; Wu, F.; Zhang, J.; Wang, F. Gender differences in colorectal cancer survival: A meta-analysis. *Int. J. Cancer* **2017**, *141*, 1942–1949. [CrossRef]
34. Favoriti, P.; Carbone, G.; Greco, M.; Pirozzi, F.; Pirozzi, R.E.M.; Corcione, F. Worldwide burden of colorectal cancer: A review. *Updat. Surg.* **2016**, *68*, 7–11. [CrossRef]
35. Permanyer, I.; Scholl, N. Global trends in lifespan inequality: 1950–2015. *PLoS ONE* **2019**, *14*, e0215742. [CrossRef]
36. Nolen, S.C.; Evans, M.A.; Fischer, A.; Corrada, M.M.; Kawas, C.H.; Bota, D.A. Cancer—Incidence, prevalence and mortality in the oldest-old. A comprehensive review. *Mech. Ageing Dev.* **2017**, *164*, 113–126. [CrossRef]
37. Shahjehan, F.; Merchea, A.; Cochuyt, J.J.; Li, Z.; Colibaseanu, D.T.; Kasi, P.M. Body mass index and long-term outcomes in patients with colorectal cancer. *Front. Oncol.* **2018**, *8*, 620. [CrossRef]
38. Liu, P.H.; Wu, K.; Ng, K.; Zauber, A.G.; Nguyen, L.H.; Song, M.; He, X.; Fuchs, C.S.; Ogino, S.; Willett, W.C.; et al. Association of Obesity With Risk of Early-Onset Colorectal Cancer Among Women. *JAMA Oncol.* **2019**, *5*, 37–44. [CrossRef]
39. Sanford, N.N.; Giovannucci, E.L.; Ahn, C.; Dee, E.C.; Mahal, B.A. Obesity and younger versus older onset colorectal cancer in the United States, 1998–2017. *J. Gastrointest. Oncol.* **2020**, *11*, 121–126. [CrossRef] [PubMed]
40. Jaspan, V.; Lin, K.; Popov, V. The impact of anthropometric parameters on colorectal cancer prognosis: A systematic review and meta-analysis. *Crit. Rev. Oncol. Hematol.* **2021**, *159*, 103232. [CrossRef]
41. Bradbury, K.E.; Murphy, N.; Key, T.J. Diet and colorectal cancer in UK Biobank: A prospective study. *Int. J. Epidemiol.* **2020**, *49*, 246–258. [CrossRef] [PubMed]
42. Akter, S.; Islam, Z.; Mizoue, T.; Sawada, N.; Ihira, H.; Tsugane, S.; Koyanagi, Y.N.; Ito, H.; Wang, C.; Tamakoshi, A.; et al. Smoking and colorectal cancer: A pooled analysis of 10 population-based cohort studies in Japan. *Int. J. Cancer* **2021**, *148*, 654–664. [CrossRef] [PubMed]
43. Wong, M.C.S.; Huang, J.; Lok, V.; Wang, J.; Fung, F.; Ding, H.; Zheng, Z.J. Differences in Incidence and Mortality Trends of Colorectal Cancer Worldwide Based on Sex, Age, and Anatomic Location. *Clin. Gastroenterol. Hepatol.* **2021**, *19*, 955–966.e61. [CrossRef] [PubMed]
44. Rasool, S.; Kadla, S.A.; Rasool, V.; Ganai, B.A. A comparative overview of general risk factors associated with the incidence of colorectal cancer. *Tumor Biol.* **2013**, *34*, 2469–2476. [CrossRef]
45. Siegel, R.L.; Fedewa, S.A.; Anderson, W.F.; Miller, K.D.; Ma, J.; Rosenberg, P.S.; Jemal, A. Colorectal Cancer Incidence Patterns in the United States, 1974–2013. *JNCI J. Natl. Cancer Inst.* **2017**, *109*, djw322. [CrossRef]
46. Rodríguez-Miguel, A.; García-Rodríguez, L.A.; Gil, M.; Montoya, H.; Rodríguez-Martín, S.; de Abajo, F.J. Clopidogrel and Low-Dose Aspirin, Alone or Together, Reduce Risk of Colorectal Cancer. *Clin. Gastroenterol. Hepatol.* **2019**, *17*, 2024–2033.e2. [CrossRef] [PubMed]
47. Bosetti, C.; Santucci, C.; Gallus, S.; Martinetti, M.; La Vecchia, C. Aspirin and the risk of colorectal and other digestive tract cancers: An updated meta-analysis through 2019. *Ann. Oncol.* **2020**, *31*, 558–568. [CrossRef]
48. Steele, C.B.; Thomas, C.C.; Henley, S.J.; Massetti, G.M.; Galuska, D.A.; Agurs-Collins, T.; Puckett, M.; Richardson, L.C. Vital Signs: Trends in Incidence of Cancers Associated with Overweight and Obesity—United States, 2005–2014. *Morb. Mortal. Wkly. Rep.* **2017**, *66*, 1052. [CrossRef]

49. Park, S.Y.; Wilkens, L.R.; Setiawan, V.W.; Monroe, K.R.; Haiman, C.A.; Le Marchand, L. Alcohol Intake and Colorectal Cancer Risk in the Multiethnic Cohort Study. *Am. J. Epidemiol.* **2019**, *188*, 67–76. [CrossRef] [PubMed]
50. Choi, Y.J.; Lee, D.H.; Han, K.D.; Kim, H.S.; Yoon, H.; Shin, C.M.; Park, Y.S.; Kim, N. The relationship between drinking alcohol and esophageal, gastric or colorectal cancer: A nationwide population-based cohort study of South Korea. *PLoS ONE* **2017**, *12*, e0185778. [CrossRef]
51. Cook, N.R.; Lee, I.M.; Zhang, S.M.; Moorthy, M.V.; Buring, J.E. Alternate-day, low-dose aspirin and cancer risk: Long-term observational follow-up of a randomized trial. *Ann. Intern. Med.* **2013**, *159*, 77–85. [CrossRef] [PubMed]
52. Brasky, T.M.; Potter, J.D.; Kristal, A.R.; Patterson, R.E.; Peters, U.; Asgari, M.M.; Thornquist, M.D.; White, E. Non-steroidal anti-inflammatory drugs and cancer incidence by sex in the VITamins and Lifestyle (VITAL) cohort. *Cancer Causes Control* **2012**, *23*, 431–444. [CrossRef]
53. Kim, H.; Giovannucci, E.L. Sex differences in the association of obesity and colorectal cancer risk. *Cancer Causes Control* **2016**, *28*, 1–4. [CrossRef]
54. Bentham, J.; Di Cesare, M.; Bilano, V.; Bixby, H.; Zhou, B.; Stevens, G.A.; Riley, L.M.; Taddei, C.; Hajifathalian, K.; Lu, Y.; et al. Worldwide trends in body-mass index, underweight, overweight, and obesity from 1975 to 2016: A pooled analysis of 2416 population-based measurement studies in 128·9 million children, adolescents, and adults. *Lancet* **2017**, *390*, 2627–2642. [CrossRef]
55. Movahedi, M.; Bishop, D.T.; Macrae, F.; Mecklin, J.P.; Moeslein, G.; Olschwang, S.; Eccles, D.; Evans, D.G.; Maher, E.R.; Bertario, L.; et al. Obesity, aspirin, and risk of colorectal cancer in carriers of hereditary colorectal cancer: A prospective investigation in the CAPP2 study. *J. Clin. Oncol.* **2015**, *33*, 3591–3597. [CrossRef] [PubMed]
56. Kuo, C.N.; Pan, J.J.; Huang, Y.W.; Tsai, H.J.; Chang, W.C. Association between nonsteroidal anti-inflammatory drugs and colorectal cancer: A population-based case-control study. *Cancer Epidemiol. Biomark. Prev.* **2018**, *27*, 737–745. [CrossRef]
57. Sankaranarayanan, R.; Kumar, D.R.; Altinoz, M.A.; Bhat, G.J. Mechanisms of Colorectal Cancer Prevention by Aspirin—A Literature Review and Perspective on the Role of COX-Dependent and -Independent Pathways. *Int. J. Mol. Sci.* **2020**, *21*, 9018. [CrossRef] [PubMed]
58. Gu, Q.; Dillon, C.F.; Eberhardt, M.S.; Wright, J.D.; Burt, V.L. Preventive aspirin and other antiplatelet medication use among U.S. adults aged $\geq$ 40 years: Data from the national health and nutrition examination survey, 2011–2012. *Public Health Rep.* **2015**, *130*, 643–654. [CrossRef]
59. Duffy, D.; Kelly, E.; Trang, A.; Whellan, D.; Mills, G. Aspirin for Cardioprotection and Strategies to Improve Patient Adherence. *Postgrad. Med.* **2015**, *126*, 18–28. [CrossRef] [PubMed]

4.6 PAPER VI: EFFECT OF ASPIRIN ON CANCERS

Authors: *Dídac Florensa, Jordi Mateo, Francesc Solsona, Leonardo Galvan, Ramon Piñol, Miquel Mesas, Leonardo Espinosa-Leal, Pere Godoy*
Journal: Annals of Epidemiology
Status: Under Review
Keywords: *Aspirin use, risk factors, comorbidities, cancer, retrospective cohort study*

Low-dose of acetylsalicylic acid for cancer prevention taking into account risk factors: A retrospective cohort study

Abstract: Aspirin has been recently reported as a protector against different cancers. Excess weight, smoking and risky drinking are risk factors significantly associated with cancer. This study looks deeper into the relationships between aspirin consumption, risk factors and diabetes with the risk of developing cancer. A retrospective cohort study of aspirin use and cancer risk factors was carried out in persons aged >50 years in Lleida province. The participants were prescribed medication between 2007 and 2016, and patients with cancer were diagnosed between 2012 and 2016. Using a Cox hazard model, risk factors, comorbidities and aspirin use were studied using adjusted hazard ratios (aHR) with 95% confidence intervals (CI). : There were 154,715 patients in the study, 16,598 of whom consumed aspirin, and 4,714 had cancer. Aspirin use had a significant protective effect against colorectal cancer (aHR: 0.7; 95% CI: 0.6-0.8), pancreatic cancer (aHR: 0.4; 95% CI: 0.2-0.9), prostate cancer (aHR: 0.6; 95% CI: 0.5-0.7) and lymphoma (aHR: 0.5; 95% CI: 0.2-0.9). A protective effect was also suggested for the oesophageal cancer (aHR: 0.5; 95% CI: 0.2-1.8), stomach cancer (aHR: 0.7; 95% CI: 0.4-1.3), liver cancer (aHR: 0.7; 95% CI: 0.3-1.6), breast cancer (aHR: 0.8; 95% CI: 0.6-1.1) and lung cancer (aHR: 0.9; 95% CI: 0.7-1.2) although this relation was not significant. For leukaemia (aHR: 1.0; 95% CI: 0.7-1.4) and bladder cancer (aHR: 1.0; 95% CI: 0.8-1.3), aspirin intake was not significant. Our results show that aspirin use decreased the risk of colorectal, pancreatic, prostate cancer and lymphoma.

INTRODUCTION

Aspirin has long been known to prevent cardiovascular and cerebrovascular [40, 41]. Daily administration of a low dose of aspirin has been proven beneficial for preventing recurrent cardiovascular events [124] . Recently, many studies have shown that long-term use of aspirin can significantly reduce the risk of cancer [42, 43]. Specifically, aspirin consumption has been strongly related with a protective effect against colorectal cancer (CRC) [125]. Some studies have found that aspirin has a protective effect against some cancers. Patrignani et al. demonstrated a potent chemopreventive effect against CRC in adults age 50 to 59 [126]. For gastrointestinal cancers, Cho et al. concluded that a long-use of aspirin was related with reduced incidence and mortality from gastric cancer [127]. Specifically, a similar reduction was also associated for stomach cancer [128]. Other cancer that has been associated with the long-use of aspirin is the oesophageal. A recent study carried out in the United Kingdom confirmed this protective effect [67]. Other studies investigated the effect of aspirin on further cancers. Tracey et al. concluded that a low-dose of aspirin was associated with a significantly lower risk of hepatocellular carcinoma [129]. Another cancer where a lower risk could be related to the use of aspirin was pancreatic cancer [69]. In the case of lung and bronchial cancer, the combination of aspirin and metformin also had independent protective associations [70]. Liebow et al. concluded that aspirin and other nonsteroidal anti-inflammatory drugs were associated with a lower risk of non-Hodgkin lymphoma. Similar conclusions were obtained for breast cancer, in which the use of aspirin was related to a lower risk [71]. Although some studies have concluded that aspirin protects against some types of cancer, its effect may vary depending on the patient's lifestyle, comorbidities and risk factors. Therefore, further research is recommended to confirm these associations [130]. Excess weight was associated with an increased risk of cancer [33]. Smoking and risky drinking were also associated with cancer risk [59]. Finally, recent some studies presented relation between diabetes and cancer risk, specifically in pancreas cancer [131]. This study aimed to analyse the association between aspirin use and different cancers, considering the effect of risk factors (overweight/obesity, risky drinking and smoking) and comorbidities such as diabetes in Lleida, Catalonia, Spain.

METHODS

Study population

A retrospective cohort study of aspirin use and risk factors for specific cancer cases was carried out based on data available from 1st January 2007 to 31st December 2016 in the Catalan Health Service (CatSalut) system, which provided care for 154,715 inhabitants in Lleida aged > 50 years at the start of the study period. The cancers included in this study were oesophageal, stomach, colorectal, liver, pancreatic, lung and bronchial cancers, leukaemia, breast (only women), prostate and bladder cancers and lymphomas. Person-years at risk were calculated as the time from the January 1, 2007 until December 31, 2016 or the date of the cancer diagnosis or date of death [94]. Data on aspirin use was obtained from the number of packages dispensed by pharmacies. Catalonia has a public health system in which medicines are dispensed in pharmacies on presentation of a doctor's prescription. Drugs administered to hospitalised patients and those prescribed by private providers did not appear the CatSalut system, and were therefore not included in this study. The cancer cases in the sample were obtained from the Lleida Population-based Cancer Registry. The demographic characteristics of the participants, including age and sex, were obtained from the CatSalut system.

Data collection

Data on cancer diagnoses were obtained from the Lleida Population-based Cancer Registry using five consecutive years of incidence data from 2012 to 2016. Potential cancer cases were validated by checking medical records. Hospital and pathological anatomy records were used as the main sources of information. Cancers were identified following the rules defined by the International Association of Cancer Registries, the International Association for Research on Cancer and the European Network of Cancer Registries. The risk factors included were risky drinking, smoking and body mass index, and the comorbidity included was diabetes. This information was extracted using the eCAP software used by primary care physicians to record all patient information. The values of these variables at the time this study started were obtained. Body mass index (BMI) was calculated from the weight and height of the patient using the formula

$$BMI = weight(kg)/height(m)^2$$

and categorised as: 18.5–24.9 normal weight, 25-29.9 overweight and >30 obese. Heavy drinking, smoking and diabetes were identified by the ICD-10 international criteria. The ICD-10 code for heavy alcohol use was F10.9 (alcohol use) previously registered by a primary care doctor. We defined this exposure as heavy drinking that was diagnosed as the consumption of >40 grams/day in men and >24 grams/day in women for 1 or more years [103]. To determine diabetes, we used the code E10-E14 (diabetes mellitus). And, for smoking, we used Z72.0 (tobacco use). Once the patients were detected to be using tobacco, we analyzed how many cigarettes they smoked per day. We defined this exposure as smoking when somebody smoked 6 or more cigarettes/day (moderate or severe smoker) [104]. The eCAP software permits determination of the smoking exposure per year. Therefore, patients defined as smokers were cases that were exposed for more than five years before cancer detection.

Exposure

Aspirin was categorised according to the Anatomical Therapeutic Chemical classification system (ATC) as A01AD05 (Acetylsalicylic acid). Medication. The use of aspirin was evaluated based on the defined daily dose (DDD) and the milligrams (mg) of accumulated dose consumed by each patient throughout the study period. The DDD is a technical unit of measurement that corresponds to the daily maintenance dose of a drug for its main indication in adults and a given route of administration. The DDDs of active ingredients are established by the WHO and published on the WHO Collaborating Center for Drug Statistics Methodology website [37, 132]. Exposure was determined from computerised pharmacy data and consisted of the total DDDs dispensed to an individual during the study period. For instance, if a person consumed aspirin for a while, then stopped using it, and later started again, the total DDDs consumed during the following period were considered. To be considered exposed to aspirin, the total time of consumption had to be $\geq$5 years [133] and the minimum consumed daily was >75 mg [68, 134]. The number of DDD calculated this value in mg.

Statistical analysis

Descriptive analyses were performed to evaluate the association between characteristics at baseline, exposure and outcomes. Crude hazard ratios were initially used to estimate the association between aspirin consumption and the risk of incident cancer. A Cox proportional hazard model was used to determine the HR and the corresponding 95% CI. The models for estimating the effect of

aspirin consumption were adjusted by sex, age, aspirin exposure, risky drinking, smoking and diabetes for each cancer. The probability values for the statistical tests were two-tailed, and a CI that did not contain 1.0 was regarded as statistically significant. Results with wide CIs should be interpreted cautiously. All statistical analyses were performed using R 3.6.2 (R Core Team 2019), an open-source programming language and environment for statistical analysis and graphic representation.

RESULTS

Patients' characteristics

From a total of 154,715 inhabitants of Lleida aged > 50, 16,598 (10.7%) had been prescribed aspirin for five or more years from 2007 to 2016, as observed in this study (table 4.10). For the aspirin group, there were a total of 5,150 (31.0%) people registered in the 70-79 age group. There were similar numbers registered in the 60-60 and 80-89 age groups, respectively 3,954 (23.8%) and 4,537 people (27.3%). However, for the non-aspirin group, the largest age group was the 50-59 (32.2%). Males were dominant in the aspirin group (54.7%) against 46.9% in the non-aspirin group. Regarding the risk factors, 83.5% of the total patients receiving aspirin had excess weight represented. 66.9 %. of those in the non-aspirin group, had excess weight. The total number of smokers registered in the aspirin group was 2,589 (15.6%), and there were 466 alcohol consumers (2.8%). In the non-aspirin group, smokers represented 8.2% and heavy alcohol drinkers, 2.1%. Diabetes was diagnosed in 8.5% and 2.4%, respectively. Finally, aspirin consumers diagnosed with cancer totalled 568 (3.4%), and for non-aspirin consumers this number was 4,146 (3.0%).

Cancer incidence

The cumulative incidence of each of the cancers was calculated for all patients and for users and non-users of aspirin (table 4.10).

Some of the results shown in Table 4.11 suggested the protective effect of aspirin against cancer. These include oesophageal cancer (RR: 0.86; 95% CI: 0.26-2.82), pancreatic cancer (RR: 0.63; 95% CI: 0.31-1.31) and lymphoma (RR: 0.54; 95% CI: 0.26-1.11). For stomach cancer (RR: 1.28; 95% CI: 0.78-2.1), CRC (RR: 1.01; 95% CI: 0.8-1.2), liver cancer (RR: 1.09; 95% CI: 0.5-2.42) and breast cancer (RR: 0.77; 95% CI: 0.58-1.02) the relative risk was not significant. However, for lung and bronchial cancer (RR: 1.39; 95% CI: 1.06-1.82), leukaemia

	Total n= 154,715 (%)	Aspirin group n = 16,598 (%)	Non-aspirin n = 138,117 (%)
Age, years			
50-59	46,454 (30.03)	1,936 (11.76)	44,518 (32.29)
60-69	35,819 (23.15)	3,945 (23.86)	31,874 (23.14)
70-79	28,138 (18.19)	5,150 (31.03)	22,988 (16.45)
80-89	23,651 (15.29)	4,537 (27.30)	19,114 (13.87)
90-	20,653 (13.35)	1,030 (6.25)	19,623 (14.22)
Sex, male	73,850 (47.73)	9,081 (54.77)	64,769(46.93)
Medication use, dosage		>75 mg	NA
Body Mass Index			
Normal weight	47,761 (30.87)	2,116 (12.71)	45,645 (33.17)
Overweight	51,022 (32.98)	5,938 (35.82)	45,084 (32.62)
Obesity	55,932 (36.15)	8,544 (51.55)	47,388 (34.30)
Smoking, yes	13,966 (9.03)	2,589 (15.60)	11,377 (8.27)
Risky drinking, yes	3,392 (2.19)	466 (2.83)	2,926 (2.19)
Diabetes, yes	4,724 (3.05)	1,422 (8.57)	3,302 (2.43)
Cancers, all	4,714 (3.05)	568 (3.43)	4,146 (3.07)

TABLE 4.10: Characteristics of the patients in the aspirin and non-aspirin groups. NA: Not applicable

(RR: 1.71; 95% CI: 1.28-2.28) and bladder cancer (RR: 1.78; 95% CI: 1.43-2.21) the results were significant.

	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)

TABLE 4.11: Relative risk of cancer incidence between aspirin and non-aspirin groups.

a Confidence interval

Cox regression showed variations in the outcomes (Table 4.12). Males have more risk of suffering cancer than women in all the cancers studied. And in

general, the age with the highest risk of cancer was between 70 and 89. Such risk factors as obesity, risky drinking and smoking increase the risk of cancer in most of the cases studied, as does smoking in lung cancer (aHR: 1.7; 95% CI: 1.4-2.1) or obesity for CRC (aHR: 1.5; 95% CI: 1.3-1.8). Diabetes was a significant factor for pancreatic cancer (aHR: 3.0; 95% CI: 1.4-6.3). Aspirin use had a significant protective effect for CRC (aHR: 0.7; 95% CI: 0.6-0.8), pancreatic cancer (aHR: 0.4; 95% CI: 0.2-0.9), prostate cancer (aHR: 0.6; 95% CI: 0.5-0.7) and lymphoma (aHR: 0.5; 95% CI: 0.2-0.9). For oesophageal cancer (aHR: 0.5; 95% CI: 0.2-1.8), stomach cancer (aHR: 0.7; 95% CI: 0.4-1.3), liver cancer (aHR: 0.7; 95% CI: 0.3-1.6), breast cancer (aHR: 0.8; 95% CI: 0.6-1.1) and lung cancer (aHR: 0.9; 95% CI: 0.7-1.2), aspirin consumption was not significant although a decreased risk of cancer was suggested. For leukaemia (aHR: 1.0; 95% CI: 0.7-1.4) and bladder cancer (aHR: 1.0; 95% CI: 0.8-1.3), aspirin intake was not significant.

Adjusted Hazard Ratio (95% CI[a])											
	Oesophagus	Stomach	Colorectal	Liver	Pancreas	Lung	Leukaemia	Breast	Prostate	Bladder	Lymphoma
Female	-	-	-	-	-	-	-	-	NA	-	-
Male	5.5 (2.1-14.8)	1.9 (1.3-2.8)	1.8 (1.6-2.1)	2.1 (1.2-3.8)	1.6 (1.1-2.5)	4.3 (3.4-5.6)	1.7 (1.4-2.2)	NA	-	6.4 (5.1-8.2)	1.2 (0.8-1.7)
Age [50-59)	-	-	-	-	-	-	-	-			
Age [60-69)	6.5 (0.2-1.7)	1.7 (1.0-3.0)	1.8 (1.6-2.1)	2.8 (1.3-6.0)	1.4 (0.8-2.5)	2.1 (1.6-2.7)	3.5 (2.4-5.3)	1.1 (0.9-1.4)	3.6 (3.0-4.5)	1.8 (1.4-2.2)	1.6 (1.0-2.5)
Age [70-79)	1.2 (0.5-3.2)	3.4 (2.0-5.7)	2.3 (1.9-2.7)	3.7 (1.7-8.1)	2.4 (1.4-4.3)	2.9 (2.2-3.8)	6.7 (4.5-9.9)	1.0 (0.8-1.2)	5.2 (4.2-6.4)	3.2 (2.5-4.0)	2.7 (1.7-4.2)
Age [80-89)	1.9 (0.7-5.4)	4.7 (2.8-8.1)	2.2 (1.8-2.6)	3.7 (1.5-9.1)	2.9 (1.6-5.5)	2.1 (1.5-2.9)	9.1 (6.1-13.6)	1.0 (0.8-1.3)	3.5 (2.7-4.9)	3.2 (2.4-4.2)	1.3 (0.6-2.5)
Age [90-)	3.9 (0-NA)	0.6 (0.2-1.6)	0.2 (0.1-0.3)	0.3 (0.0-2.6)	0.3 (0.1-1.3)	0.1 (0.1-0.3)	0.5 (0.2-1.3)	0.1 (0.0-0.2)	0.1 (0.0-0.3)	0.2 (0.1-0.4)	NA
Aspirin use	0.5 (0.2-1.8)	0.7 (0.4-1.3)	0.7 (0.6-0.8)	0.7 (0.3-1.6)	0.5 (0.2-0.9)	0.9 (0.7-1.2)	1.0 (0.7-1.4)	0.8 (0.6-1.1)	0.6 (0.5-0.7)	1.0 (0.8-1.3)	0.5 (0.2-0.9)
Normal weight	-	-	-	-	-	-	-	-			
Overweight	0.7 (0.3-1.4)	1.0 (0.6-1.6)	1.4 (1.2-1.7)	1.0 (0.5-2.2)	1.2 (0.6-2.2)	0.7 (0.6-0.8)	0.9 (0.7-1.3)	1.3 (1.1-1.6)	1.8 (1.4-2.3)	1.0 (0.8-1.3)	0.9 (0.6-1.5)
Obese	0.5 (0.2-1.4)	0.9 (0.5-1.5)	1.5 (1.3-1.8)	1.1 (0.5-2.4)	1.2 (0.7-2.3)	0.6 (0.5-1.0)	1.1 (0.8-1.5)	1.6 (1.3-2.0)	2.0 (1.6-2.6)	1.1 (0.8-1.4)	1.0 (0.6-1.7)
Smoking	1.9 (0.8-4.5)	1.4 (0.8-2.3)	1.4 (1.3-1.7)	1.6 (0.8-3.2)	1.5 (0.8-2.7)	1.7 (1.4-2.1)	0.9 (0.6-1.3)	1.0 (0.7-1.5)	1.0 (0.8-1.1)	1.8 (1.5-2.2)	0.8 (0.4-1.5)
Risky drinking	3.7 (1.3-10.9)	0.9 (0.3-2.9)	1.6 (1.2-2.0)	1.2 (0.3-4.8)	1.6 (0.6-4.5)	1.9 (1.4-2.9)	1.8 (1.0-3.3)	1.0 (0.4-2.6)	0.9 (0.6-1.2)	1.1 (0.7-1.7)	1.4 (0.5-3.8)
Diabetes	2.3 (0.5-10.3)	1.6 (0.7-3.5)	1.1 (0.8-1.5)	1.0 (0.2-4.1)	3.0 (1.4-6.3)	0.9 (0.6-1.7)	0.8 (0.4-1.5)	0.7 (0.4-1.3)	1.2 (0.9-1.7)	1.2 (0.8-1.8)	0.3 (0.1-2.0)

TABLE 4.12: Adjusted Hazard Ratios for cancer type

a Confidence interval

DISCUSSION

This retrospective study confirmed that aspirin reduced the risk of CRC, pancreatic cancer, prostate cancer and lymphomas. The benefit of aspirin also suggested the reduction in the risk of oesophageal, stomach, liver and breast cancer, although the effect is not statistically significant. Risky drinkers and smokers were significantly associated with the risk of most cancers and diabetes, specifically with the risk of pancreatic cancer. Several recent studies have suggested an association between aspirin use and some specific cancers, although not all concluded similar outcomes. Tsoi et al. analysed the effect of aspirin on different cancers [42]. They demonstrated aspirin's protective effect against

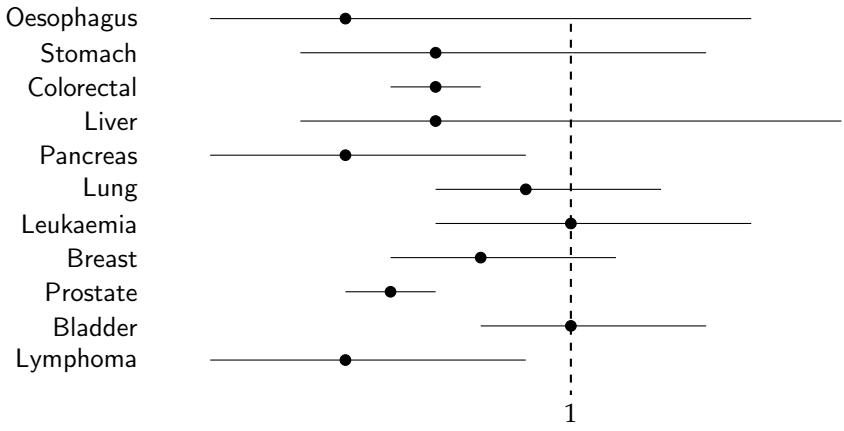


FIGURE 4.1: Aspirin adjusted hazard ratio for cancer.

liver, colorectal or pancreatic cancers, but it was considered a risk factor for breast cancer. However, other studies have demonstrated that low-dose aspirin intake is associated with a significantly lower risk [135]. Lower risk of CRC was also significantly associated with the use of aspirin [133, 136, 137]. Other cancers, such as oesophageal [138], stomach [67], liver [129] or pancreatic [69] were associated with lower risk. Studies on the impact of aspirin use have been published, even though most of these do not include the other risk factors and comorbidities that we present in this analysis. Also, some studies focus on a specific cancer. We included different cancers to have an overview of the role of aspirin exposure in most cancers. Therefore, our findings corroborate the research in the field highlighting the protective effect of aspirin in some cancers and go beyond comparing this positive effect with the negative effects caused by several risk factors and comorbidities.

We found significant differences according to sex, suggesting that men have a higher risk of developing CRC. It has been reported that men have higher cumulative levels of smoking than women and a higher alcohol use, which may explain the higher risk [139]. Another factor that could explain this discrepancy is the high intake of animal fats and proteins in men's diets [140]. Regarding age, people aged between 70 and 89 had a higher risk of cancer in most cases, although there were exceptions. The risk of lung and prostate cancer was similar in the 60-69 age group. This fact could be explained because older adults may have a different mechanism to younger people. For example, ageing is

associated with alterations in DNA methylation, which may affect susceptibility to cancer [114].

Some significant differences were observed in risk factors and comorbidities. These differences showed that some cancers may be associated with being overweight/obese. CRC was significantly associated with excess weight [141]. Obese people had 1.5 times the risk [87] and those who were overweight, 1.4 [142]. Excess weight was also significantly related to breast and prostate cancer. Females (breast) who were overweight or obese had 1.3 and 1.6 times the risk [143] and males (prostate) 1.8 and 2.0 respectively [144]. The other cancers suggested an association between excess weight and risk of cancer even though this was not significant. Smoking and risky drinking were also strongly related to some cancers.

Smoking increased the risk of lung and bronchial cancer by a factor of 1.7 [145], as several studies have demonstrated over a long time. Another cancer associated with smoking was bladder. It gave a risk 1.8 times higher, similar to that in previous studies [146]. A similar outcome was obtained for CRC [147]. Other cancers, such as the oesophageal, stomach, liver and pancreatic, may be associated even though these possible relations were not significant.

Regarding risky drinking, our results corroborated previous conclusions. CRC was significantly associated, with an increase in the risk of 1.6 times, as Park et al. concluded [148]. Lung cancer has been related to smoking, and an association was also obtained between risky drinking and lung cancer with a risk 1.9 times higher [149]. Another cancer significantly related to risky drinking was leukaemia, with 1.8 times the risk. Other previous studies suggested this relation [150, 151], even though their results were not significant. Our results suggest that an exhaustive analysis should be carried out to confirm this association. Finally, other such cancers as liver, pancreas, bladder and lymphomas suggested an association with risk drinking, but these were non-significant, although other studies into liver and pancreas cancer have confirmed the link [131, 152]. However, another studies into bladder cancer and lymphomas have not shown a relationship between alcohol use and these [153, 154]. Therefore, more research is required. Diabetes was significantly associated with the risk of pancreas cancer, with a 3.0 times greater risk. However, with the rest of the cancers, there were no associations.

The results of the possible association between aspirin use and cancer were promising. There was a significant association shown between CRC and the use of aspirin. This confirmed a protective effective of around 30% against CRC [155, 156]. Garcia et al. [67] obtained similar results when they studied CRC. Another cancer strongly associated with lower risk and the use of aspirin was pancreatic [42]. The results suggested a risk reduction of 50%. The reduction

of the risk of pancreatic cancer could be because aspirin inhibits the proliferation and stimulates the apoptosis of pancreatic cancer cells by inactivating the P13K/Akt/mTOR signalling pathway [157]. Aspirin consumption (only males) was also strongly related to a reduction in prostate cancer [158]. Our results suggested a protection of 40%. The aspirin suppressed prostate cancer cell invasion by reducing MMP-9 activity and uPA expression by decreasing IKK-mediated NF-kB activation [159]. Finally, the lower risk of lymphoma (Hodgkin and non-Hodgkin) was also significantly associated with aspirin consumption. The results for these suggested a protective effect of 50% [160]. Although there is not so much literature related to lymphomas, the studies available suggest that the use of aspirin decreased the risk of Hodgkin lymphoma [161]. However, aspirin was associated with an increase in the risk of non-Hodgkin lymphoma [162]. The controversy in the published literature needs a deeper into the association between this kind of cancer and the use of certain medicaments, although a recent study confirmed the association between the use of dipyridamole and a lower risk of lymphoid neoplasms [163].

Other outcomes for cancers included in the study suggested the protective effect of aspirin against them. The use of aspirin against oesophageal cancer suggested a protective effect of 50% [42, 138], although the outcome was not statistically significant in our study. This may be due by the small number of oesophageal cancer cases included in the study (see table 4.11). Similar outcomes were obtained for stomach cancer. The use of aspirin suggested a 30% lower risk of stomach cancer, although the result was not significant. Other studies have recently obtained significant results with a similar protective effect of 32% [164]. The same level of protection was obtained for liver cancer by using aspirin [129]. This is due to the inhibition of the adenosine-monophosphate-activated protein kinase (AMPK)-TOR pathway, thus resulting in the suppression of the mTORC1 activity [165]. A lower risk of breast cancer was also associated with aspirin consumption [166], although our results were not significant [167]. Finally, other cancers, such as lung, bladder and leukaemia, and their association with aspirin use were not significant.

The study has some limitations. Firstly, some patients could buy aspirin directly in pharmacies without a doctor's prescription, and this consumption is underreported. Second, some patients may not take the medication even though they have purchased it at the pharmacy, and, in this case, aspirin use will be over-reported. Third, although the population-based cancer registry is exhaustive, it cannot be ruled out that some cases were diagnosed in hospitals in other territories, and some cases were not correctly registered. In addition, there were few registered cases of some cancers such as oesophagus or liver. We could not

study a dose-response relationship between low-dose aspirin and cancers because more than 90% of aspirin use in this study was at a dose of 100 mg/day, which did not allow us to assess the effect of higher doses.

The strengths of this study included the fact that data is presented by risk factors, such as excess weight, smoking and risky drinking. The study was made with information from clinical practice, with physicians unaware of the aims of the study, thereby preventing researcher bias.

In conclusion, this retrospective cohort study found an association between aspirin use for $\geq$ 5 years and a reduced risk of CRC, pancreas cancer, prostate cancer and lymphomas. In addition, aspirin use among the patients diagnosed with oesophagus, stomach, liver and breast cancer suggested a possible protective effect. The results also showed an association between cancers and such risk factors as excess weight (overweight/obesity), smoking and risky drinking. They also showed an association between diabetes and pancreatic cancer, which stimulated more research. In general, these results reinforce the use of aspirin, under prescription, as a protective factor for some cancers, and they also reinforce the need for public health messaging about the harmful effects of smoking, alcohol use and excess weight. They also encourage deep associations between specific cancers and the use of aspirin to help the health system to focus on preventing these or the effect of some comorbidities such as diabetes. Also, the use of other techniques to consolidate the suggestions about the protective effect of aspirin use in some cancers.

5

GLOBAL DISCUSSION OF RESULTS

I dream my painting and I paint my dream.
— Vincent van Gogh

The main contribution of this thesis was to explore and develop machine learning models and an analytical cloud platform to access and use the PBCR data. The work proposed answers questions about the use of artificial intelligence and statistical techniques to determine the cancer situation in Lleida region, integrating and analyzing externals databases with information on factors related to cancer. Therefore, this thesis presents artificial intelligence solutions to search cancer patterns, and cancer risk and protective factors analysis.

At the beginning of the research, efforts were focused on understanding and analyzing the epidemiological cancer situation in Lleida region. For this, we designed and implemented an interactive cloud platform based on DSS. The cancer incidence rate is essential in public health surveillance. The incidence rate approximates the average risk of developing cancer, allowing geographic comparisons of the disease risk in different populations. In this field, only a few proposals aimed to represent the incidence and mortality in a DSS web app. A similar platform was implemented by Xia *et al.*, which visualizes cancer risk factors and mortality [74]. They shared a data warehouse and the Rshiny app to improve their understanding of spatial and temporal trends across the population served by the Kansas University Cancer Center. Our proposal offers a wholly adaptable and scalable solution which permits analysis of cancer incidence and mortality in depth and for specific cancers, regions, gender or age periods. This system helped the research team analyze the cancer information rapidly and draw some conclusions about the data and the use of these technologies. It also helps to offer data accessibility to determine the cancer situation. In incidence, the user can observe the number of cases by gender, age group, region or years. Evolution plots also help to compare how cancer affects these factors. Specific maps for different areas are printed to analyze the incidence in real-time. A summary of risk factors offers a general image of them. Similarly, the mortality front-end implementation is based on the incidence web-views. General trends were described to determine the analyzed region's cancer mortality situation. In addition, the technologies used to build this system permit deployment in other cancer registries. Therefore, this analysis motivated the next studies about

the application of ML algorithms which permits the information shown in the web-platform to be analyzed and predicted.

Then, efforts were focused on training machine learning algorithms based on sociodemographic patient information, risk factors and tumor information to detect cancer patterns among specific cancers. The results of this research permitted us to discover some regions where specific cancers present a higher incidence. They also provided affirmative evidence that these algorithms can be used to detect cancer patterns and cancer associations. A rapidly growing trend in the use of machine learning applied to cancer opens the door to new ways of detecting and analyzing cancer. Machine learning, cancer, and risk factors lie at the heart of the work realized in this thesis aimed at addressing these challenges. Cancer demands research to understand, analyze and promote public policies to raise awareness that factors that increase the cancer risk should be avoided. Understanding cancer in a specific region and sub-region permits the public government to understand cancer incidence, and this was the first step of this thesis. Authors in [60] present evidence about the association between some specific cancers and specific regions. They observe that female breast and prostate cancer register a greater incidence in metropolitan areas, and colorectal or cervix cancer in nonmetropolitan areas. On the other hand, in [82], the authors propose an unsupervised algorithm (MCA) to detect relations between car accidents and different aspects. They concluded that bad driving and crashes could be affected by differences between urban and rural areas, traffic volume, driver age and more. In this thesis, we proposed using MCA to detect associations between cancer incidence and specific areas such as rural or urban. Before starting the training of the algorithm, a statistical technique is employed to eliminate potential outliers. This approach results in only a negligible loss of information, and consequently, any bias in the results resulting from the removal of these cases is residual.

First of all, the algorithm was trained for all the cancers together and then, it was stratified by each one. Once the model was trained, we evaluated the performance to ensure the statistical potential. Then, the results were corroborated by the previous literature to confirm the obtained relations. The conclusions give us strong evidence the MCA algorithm proposed can be used to search for cancer associations among possible related factors. In addition, they allowed us to find hidden associations that had not previously been raised. Note that when using this type of technique, certain shared risk factors among some cancers may be affected when displaying correlations. However, these factors were not included in the analysis, and therefore, the bias cannot be diluted with this analysis.

At this point, the MCA technique could be used to find patterns of cancer. Thus, we focused on combining it with another unsupervised algorithm that helped corroborate the same results. This thesis proposed the combination of MCA and K-means to detect patterns of cancer among colorectal cancer patients by risk factors, tumor information and sociodemographic information. To conduct our study and identify risk factors such as BMI, smoking, and alcohol use, we initially examined hospital records and primary care databases to extract information. However, the use of this information is often disputed due to its potential bias in epidemiological studies. Despite this, a previous study concluded that the prevalence observed in primary care databases is similar to those from nationally representative surveys [168]. In our region, previous studies have also demonstrated the potential of primary care databases in cardiovascular disease [169]. The authors concluded that the prevalence of risk factors associated with cardiovascular disease and their association with the emergence of vascular diseases is consistent with the findings of a population-based epidemiological investigation that used a comprehensive, standardized methodology. Additionally, a study by Bolibar *et al.* [170] demonstrated the potential of primary care databases as a source of information for epidemiological studies. The results obtained in the thesis using this information may be subject to bias, even though the associations have been justified by the previous literature [171].

MCA allowed associations to be discovered, and subsequently K-means found the main cancer profiles. This analysis strategy has been extensively tested in multiple areas of knowledge, including bio-medicine [172, 173]. The authors in [85] presented a study about using the combination of MCA and K-means to ascertain multimorbidity patterns. It concluded that these techniques could help to identify these patterns. Also, in terms of cancer analysis, the clusters allowed us to discover that smoking may be associated with colorectal cancer and that obesity is more incident in older people. In addition, the younger population presents early tumor staging, which results in higher survival statistics. Therefore, these techniques have proven to be effective tools for analyzing the incidence of some factors in colorectal cancer. The outcomes obtained help corroborate suspected trends and stimulate the use of these techniques to find the association of risk factors and the incidence of other cancers.

The next milestone of this thesis was to analyze the association between risk factors and second primary cancer to build models to predict the cancer risk. Inspired by a previous study [22], we analyzed the situation of the SPC in the Lleida region and the role of risk factors in the SPC risk. We found an increased risk of SPC and compared this risk between genders, age groups or the first primary cancer. We also found that smoking and risky drinking increase the risk

of SPC. In general, these results reinforce the need for public health messaging about the harmful effects of tobacco and alcohol. They also encourage continued research into SPCs to find new factors associated with these cancers and help the health system to focus on preventing them. These results also added the statistical potential to train the algorithms in the next study.

These results motivated the next study which focused on building predictive models to predict the risk of SPC by previous risk factors. Although the experiment is currently in its preliminary stages, it is designed to build on the previous work by [174] and utilize many supervised machine learning algorithms to develop the most effective model.

Finally, to highlight a possible use of PBCRs, this thesis focused on the impact of aspirin, taking into account risk factors, on some cancers. We based this on previous literature which showed the protective effect of aspirin against cancer, especially colorectal cancer [65, 66]. In the first study, we corroborated this protective effect against colorectal cancer, as had previous reports. We found an association between aspirin use for $\geq$5 years and a reduced risk of CRC. The protective effect due to aspirin was higher in women. We also confirmed that risk factors such as risky drinking, excess weight and smoking increase the risk of colorectal cancer, especially in men. It has been reported that men have higher cumulative levels of smoking than women and a higher alcohol intake, which may explain the higher risk [139]. In general, these results confirm the potential offered by the combination of these databases, which motivates us to continue to explore other cancers. Therefore, the second study focused also on the impact of aspirin on some cancers, such as breast, prostate or lung, among others. For this study, we used a previous study that analyzed the specific relationship between aspirin and cancer. The authors in [42] showed the protective effect of aspirin against liver, colorectal or pancreatic cancers, among others. However, aspirin was considered a risk factor for breast cancer. In this thesis, the results confirm that aspirin use decreased the risk of colorectal, pancreatic, prostate cancer and lymphoma. On the other hand, it was not associated as either a risk or protective factor in the other cancers. However, excess weight, smoking and risky drinking were significantly associated with the risk of most cancers, which demonstrate that, in this region, these factors increase the risk of cancer, as the literature concludes.

The thesis at hand acknowledges certain limitations that must be taken into account. Specifically, it should be noted that the Population-based Cancer Registry in Lleida is a relatively new initiative that began its operations in 2017. As a result, the registry's coverage of cancer cases in the region is not yet comprehensive. This limited coverage has also resulted in a relatively low number

of registered years, and consequently, a restricted total number of cases. These factors may impact the generalizability and scope of the findings presented, and should be taken into consideration when interpreting the results. However, the outcomes are demonstrated and confirmed by recent published literature.

The use of machine learning techniques in this thesis also has certain limitations that should be noted. The Multiple Correspondence Analysis (MCA) and K-means algorithms employed in this study are designed for exploratory data analysis, where the goal is to identify patterns and significant associations in large databases without relying on pre-existing hypotheses. However, in cancer epidemiology, the majority of studies are hypothesis-driven, which has resulted in a relatively low number of studies using similar techniques as those presented here. Nevertheless, the potential of these methods to explore complex databases, such as cancer registries, should not be underestimated. Despite the limitations of MCA and K-means in the context of hypothesis-driven research, they offer valuable insights into the structure and distribution of data, which can inform future studies in cancer epidemiology. Another limitation that should be mentioned is related to the use of primary care databases as a source of information for risk factors. It is important to recognize that the analysis is not completely free from bias, and risk factors can be subject to bias. One of the strengths of the study was that physicians were unaware of the study objectives, which helped to avoid investigator bias.

There are some limitations that need to be taken into account regarding the studies on subsequent primary cancer. The short follow-up period after the first diagnosis might have reduced the probability of detecting risk factors related to body mass index and diabetes. This shorter follow-up did not allow for more SPCs to be observed; therefore, a longer observation period would improve the quality of the sample. Additionally, this shorter period caused the first patients included in the cohort to have a higher risk of SPC due to the longer follow-up period. However, the results are promising and adding more years to the cohort can improve them further.

Regarding medicine consumption, there are some limitations that need to be noted. Firstly, some patients may purchase aspirin directly from pharmacies without a doctor's prescription, and this consumption may be underreported. Secondly, some patients may have purchased the medication but not taken it, leading to an overreported use of aspirin. Despite these limitations, the results obtained in this study are consistent with previous studies conducted in other territories.

6

GENERAL CONCLUSIONS AND FUTURE DIRECTIONS

This chapter summarises the research work on the use of the Population-based Cancer Registry to analyze the cancer situation and associated factors in Lleida region by the implementation of a Cloud platform and Machine Learning algorithms.

It also discusses open research problems in the area and outlines many future research directions.

6.1 CONCLUSIONS

The objective of this thesis was to contribute new solutions to integrate, analyze, and search for associations and to detect cancer patterns. It mainly researched DSS-based cloud platforms and applied machine learning to transfer qualified knowledge from academia to the medical sector and society. The research in this thesis highlights two essential fields in three sciences: Cancer Epidemiology, Machine Learning and Cloud platforms.

The first paper permitted us to integrate risk factor databases and to determine the epidemiological cancer situation in Lleida through an interactive cloud application and offers data accessibility to society. This mainly allows cancer incidence and mortality to be analyzed. The next paper demonstrates the use of an unsupervised machine-learning algorithm to detect associations between cancer incidence and demographic information. The following article focuses on combining two unsupervised machine learning algorithms to detect relationships and cancer patterns between colorectal cancer, risk factors, demographic information and tumor information. Hence, the following paper focus on secondary primary cancer (SPC) risk which analyzed the association between risk factors and the SPC risk. The encouraging findings suggest a promising new direction for future research into cancer risk prediction using machine learning techniques. Finally, the last two studies demonstrated the protective effect of aspirin against some cancers, especially colorectal cancer.

The main conclusion that can be drawn from this thesis are:

- The significance of Population-based Cancer Registries for monitoring and analyzing cancer and their research potential when integrated with other databases, such as risk factors or medication exposures.
- Smoking and heavy alcohol use increase the risk of secondary primary cancer during the first follow-up years, especially among men.
- In terms of risk factors and colorectal cancer, excess weight in the older population increases the risk of developing aggressive tumors and death.
- The contribution of colorectal cancer (whatever the gender) and prostate cancers among men in rural areas. There is a high incidence of lung cancer in urban areas and breast cancer in both areas.
- Aspirin use decreases the risk of colorectal cancer and excess weight, smoking and heavy drinking increase this risk. Aspirin also decreases the risk of pancreatic, prostate cancer and lymphoma.
- The cloud applications offer accessibility, elasticity, scalability 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.
- The combination of MCA and K-means is a perfect alliance for detecting patterns and associations of cancer from information about patients and risk factors.

To sum up, this thesis worked on integrating hospital and primary care databases to achieve cancer analysis solutions based on machine learning algorithms to detect cancer associations and patterns through risk factors, sociodemographic information and tumor information. Consequently, introducing this methodology into the cancer epidemiology sector allows population-based cancer registries to benefit from academic expertise and create new ways to understand, detect and predict cancer incidence.

6.2 FUTURE DIRECTIONS

Despite the contributions of the current thesis in Artificial Intelligence and Population-based Cancer Registries, there are many open research challenges to address to advance further in these areas.

This thesis has demonstrated the importance and the potential of Population-based Cancer Registries to search possible association that increase the cancer

risk. We focused on specific cancer or risk factors, even though the proposed solutions can be extended to analyze other cancers and find other combinations. Furthermore, this study aims to utilize supervised machine learning algorithms to predict cancer risk based on exposure to various risk factors, and investigate the extent to which these factors influence cancer risk.

Advanced technologies can automate the validation process related to population-based cancer registries. The use of new technologies, such as the Natural Language Process, permits the automation of a process that currently requires more time. Another point is the comparability with other registries, which can offer the possibility to compare different regions. The public administration is working on this integration.

Regarding the risk factors associated with cancer, several can be analyzed to detect new associations that increase these risks. Lleida presents a particular lifestyle. Consequently, the associations may vary depending on the area or region. The impact of radiation exposure could be determinant in some cancers. The same may be true of certain infections, such as human papillomavirus (HPV), helicobacter pylori, VIH and hepatitis B and C, which can increase the risk of certain types of cancer. Therefore, the analysis by these associations can enable us to understand part of the incidence in Lleida. Finally, the exposure to some drugs opens a new door for ongoing research into their effects on some cancers.

Artificial Intelligence and Cloud Computing technology perform a vital role in the future of health and cancer. Data science applied to the medical sector opens a new way of understanding the diagnosis and prognosis of these diseases. The prediction of some cancers depending on risk factors or genes can be crucial to a better prognosis in the future. Also, personalized medicine can help to adapt treatments and, in consequence, also improve the prognosis. Moreover, analyzing all the historical information collected in the medical record can help identify patterns and enable better predictions and prescriptions. Thus, researching this technology's use in specific cancers may highlight particularities. Therefore, the future is the complex models which can help medical decision-making when fed with intelligent data delivered by cloud-based decision support tools. This thesis plants the seed of a much bigger and more ambitious idea.

A

DOCTORAL STAY: TECHNIQUE PAPER

Authors: *Didac Florensa, Jordi Mateo, Francesc Solsona, Leonardo Galván, Pere Godoy and Leonardo Espinosa-Leal*

Predicting the risk of cancer by the consumption of medicines

INTRODUCTION

Cancer is the leading cause of death worldwide, accounting for nearly 10 million deaths in 2020 [132], and many of these cancers are highly preventable [175]. Approximately 20% of these cases could be related to excess weight, high alcohol consumption and smoking [176]. On the other side, several studies have been recently published about the consumption of some drugs and cancer risk [125, 177]. Some concluded that the extensive exposition of the specific drug could increase some particular risks. These studies searched for associations between drugs exposition and cancer risk, but they were not used this to predict it. This study aims to predict the risk of cancer, in general, by the consumption of the drugs for five years in the Lleida region (Catalonia, Spain). Also, to detect if medicines could be used to predict cancer risk.

METHODS

This is a community-based retrospective cohort study. The observed period was between January 1, 2007, and December 31, 2016. The dataset included 724,120 patients exposed to some medicine. The information about the drugs exposition was obtained from the Catalan Health Service (CatSalut). It represents the public health system of Catalonia, and it registers the medicines dispensed by the pharmacies after presenting a doctor's prescription. And the cancer information was obtained by the Population Cancer Registry of Lleida, which documented all the cancer cases diagnosed in the leading hospitals in this region. The cancer cases included in the dataset were diagnosed between 2012 and 2016. Therefore, the dataset contained 425 medicines groups (features), age groups and gender. To be considered a drug risk consumer, the person had to be exposed to each specific drug for five years. The exposition was evaluated based on the defined daily dose (DDD). The DDD is a technical unit of measurement that indicates the route of administration. The DDD minimum to be considered had to be >30 DDDs. To build the prediction models, the researchers implemented machine learning algorithms. Specifically, they create a model based on random forest (RF), neural network (NN) and logistic regression (LR). Accuracy, precision and recall metrics were used to evaluate these models.

RESULTS

Random forest obtained an accuracy of 86%, which represented that 8,6 of 10 patients will be predicted correctly depending on their medicine's expositions.

The neural network also kept similar results. Its accuracy was 85%. Finally, the logistic regression obtained worse accuracy than previous ones. It was 75%. These accuracies were obtained with the dataset balanced (100%). See figure A.1.

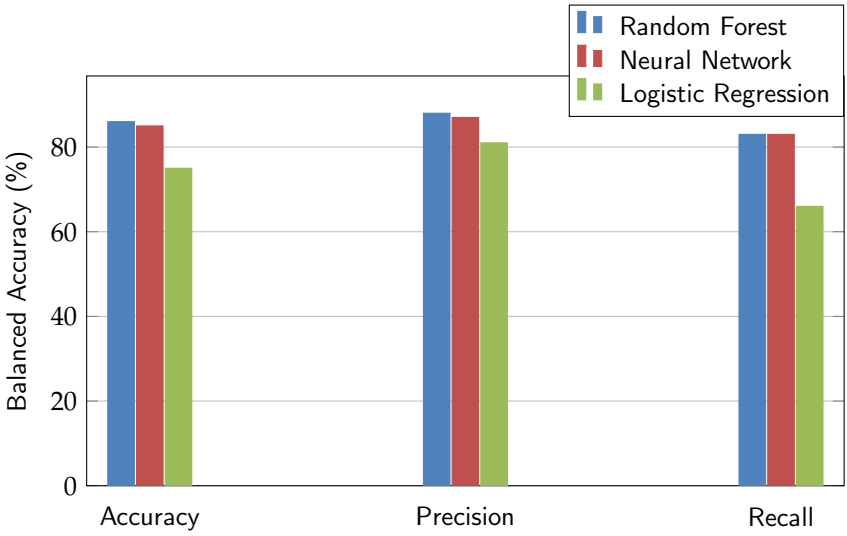


FIGURE A.1: Obtained metrics with the balanced dataset (100%) for the different algorithms.

CONCLUSIONS

Different machine learning algorithms could predict the risk of cancer. In this case, the best-obtained metrics were from the Random Forest and Neural Network. However, the metrics of the logistic regression decreased. These results suggested that the medicine exposition could be used for building predictive models about the risk of cancer. They are also encouraged to work with specific cancers and medicines, including other risk factors such as smoking, alcohol consumption and excess weight.

ACKNOWLEDGEMENT

The work has been performed under the Project HPC-EUROPA3 (INFRAIA-2016-1-730897), with the support of the EC Research Innovation Action under

the H2020 Programme; in particular, the author gratefully acknowledges the support of Leonardo Espinosa from the Department of Business Management and Analytics and the computer resources and technical support provided by CSC-Finland.

B

SUPPLEMENTARY TABLES

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Prostate	122.5	276	180	65.2
Colorectal	91.4	206	190	92.2
Lung (and Bronchus)	89.4	201	144	71.6
Urinary bladder	66.5	150	75	50.0
Oral Cavity and Pharynx	24.4	55	37	67.3
Leukaemia	19.1	43	45	104.7
Stomach	17.7	40	38	95.0
Non-Hodgkin lymphoma	17.2	39	28	71.8
Liver	15.4	35	32	91.4
Larynx	15.1	34	23	67.6
Kidney	13.1	30	45	153.3
Pancreas	11.3	25	26	104.0
Melanoma	9.5	21	12	57.1
Brain and Other Nervous System	8.6	19	9	47.4
Total without non-melanoma	581.4	1,310	941	71.8

TABLE B.1: Comparation of expected cases in men against the residents of Lleida of Population Cancer Registry. 2012

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Breast	108.8	237	218	92.0
Colorectal	59.9	130	118	90.8
Uterus	19.3	42	38	90.5
Lung (and Bronchus)	14.8	32	43	134.4
Non-Hodgkin lymphoma	13.3	29	18	62.1
Ovary	13.1	29	14	48.3
Cervix Uteri	12.4	27	25	92.6
Pancreas	11.6	25	23	92.0
Urinary bladder	11.1	16	32	66.7
Melanoma	10.9	24	18	75.0
Stomach	9.7	21	13	61.9
Kidney	6.7	15	18	120.0
Brain and Other Nervous System	6.5	14	17	121.4
Liver	6.3	14	4	28.6
Total without non-melanoma	374.8	816	668	81.9

TABLE B.2: Comparation of expected cases in women against the residents of Lleida of Population Cancer Registry. 2012

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Prostate	120.6	270	185	68.5
Colorectal	87.2	195	179	91.8
Lung (and Bronchus)	82.8	185	179	73.5
Urinary bladder	50.3	113	108	95.6
Oral Cavity and Pharynx	21	47	27	57.4
Leukaemia	13.6	30	49	163.3
Stomach	17.2	38	38	100
Non-Hodgkin lymphoma	17.5	39	34	87.1
Liver	17	38	13	34.1
Larynx	12.1	27	22	81.5
Kidney	18.2	41	40	97.6
Pancreas	13.9	31	34	109.7
Melanoma	10.7	24	28	116.6
Brain and Other Nervous System	9.3	21	16	76.2
Total without non-melanoma	557.6	1,248	1,108	88.7

TABLE B.3: Comparation of expected cases in men against the residents of Lleida of Population Cancer Registry. 2013

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Breast	112.4	244	227	93.0
Colorectal	63.1	137	122	89.1
Uterus	21.5	47	44	93.6
Lung (and Bronchus)	17.6	38	32	84.2
Non-Hodgkin lymphoma	14.1	31	31	100
Ovary	11.9	26	22	84.6
Cervix Uteri	8.1	18	13	72.2
Pancreas	12.8	28	19	67.9
Urinary bladder	9.2	20	17	85.0
Melanoma	11	24	28	116.7
Stomach	11.3	25	22	88.0
Kidney	8.9	19	18	94.7
Brain and Other Nervous System	7.3	16	13	81.3
Liver	6.6	14	8	57.1
Total without non-melanoma	393.4	854	795	93.1

TABLE B.4: Comparation of expected cases in women against the residents of Lleida of Population Cancer Registry. 2013

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Prostate	120,7	268	243	90,8
Colorectal	87,2	194	218	112,6
Lung (and Bronchus)	82,8	184	149	81,1
Urinary Bladder	50,3	112	157	140,6
Oral Cavity and Pharynx	21,0	47	30	64,3
Kidney	18,2	40	36	89,4
Non-Hodgkin lymphoma	17,5	39	16	41,3
Stomach	17,2	38	40	104,8
Liver	17,0	38	24	63,7
Pancreas	13,9	31	27	87,6
Leukaemia	13,6	30	46	152,8
Larynx	12,1	27	30	111,8
Melanoma	10,7	24	30	126,0
Brain and Other Nervous System	9,3	21	18	87,5
Oesophagus	8,4	19	12	64,6
Myleoma	6,5	14	16	111,1
Testis	5,6	12	8	64,6
Gallbladder and Other Biliary	4,2	9	12	128,5
Hodgkin lymphoma	2,6	6	4	68,8
Thyroid	2,6	6	9	155,4
Total without non-melanoma	557,6	1237	1342	108,5

TABLE B.5: Comparation of expected cases in men against the residents of Lleida of Population Cancer Registry. 2014 (221,891 men)

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Breast	112,4	243	253	104,2
Colorectal	63,1	136	148	108,5
Cos Uterusí (endometri)	21,5	46	52	112,0
Lung (and Bronchus)	17,6	38	43	113,2
Non-Hodgkin lymphoma	14,1	30	10	32,9
Pancreas	12,8	28	20	72,4
Ovary	11,9	26	23	89,1
Stomach	11,3	24	26	106,9
Melanoma	11,0	24	23	96,8
Thyroid	10,3	22	19	85,6
Leukaemia	10,2	22	39	176,2
Urinary bladder	9,2	20	26	131,3
Kidney	8,9	19	15	78,3
Cervix Uteri	8,1	18	14	79,7
Brain and Other Nervous System	7,3	16	22	139,3
Liver	6,6	14	10	70,6
Myeloma	6,3	14	17	124,9
Oral Cavity and Pharynx	6,2	13	16	119,0
Gallbladder and Other Biliary	4,6	10	14	139,9
Hodgkin lymphoma	2,6	6	4	72,6
Total without non-melanoma	393,4	850	829	97,5

TABLE B.6: Comparation of expected cases in women against the residents of Lleida of Population Cancer Registry. 2014 (216,110 women)

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Prostate	146,4	323	251	77,7
Colorectal	108,7	240	289	120,5
Lung (and Bronchus)	98,4	217	160	73,7
Urinary bladder	76,5	169	141	83,5
Stomach	22,6	50	36	72,2
Oral Cavity and Pharynx	21,9	48	37	76,5
Liver	18,7	41	25	60,6
Non-Hodgkin lymphoma	18,4	41	24	59,1
Leukaemia	16,6	37	67	182,9
Kidney	15,8	35	42	120,4
Pàncreas	15,4	34	33	97,1
Larynx	15	33	38	114,8
Melanoma	11,3	25	35	140,3
Brain and Other Nervous System	10,2	23	32	142,1
Oesophagus	8,7	19	15	78,1
Myeloma	6,5	14	15	104,6
Gallbladder and Other Biliary	4,8	11	15	141,6
Testis	4,3	9	7	73,8
Hodgkin lymphoma	4,1	9	10	110,5
Thyroid	3,8	8	5	59,6
Total without non-melanoma	653,1	1.442	1.329	92,2

TABLE B.7: Comparation of expected cases in men against the residents of Lleida of Population Cancer Registry. 2015 (220,719 men)

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Breast	117,5	253	246	97,2
Colorectal	70,6	152	157	103,3
Cos Uterusí	26,1	56	36	64,1
Lung (and Bronchus)	25,1	54	49	90,7
Urinary bladder	15,5	33	21	62,9
Non-Hodgkin lymphoma	14,7	32	20	63,2
Pancreas	14,4	31	19	61,3
Stomach	14,0	30	30	99,5
Ovary	13,7	29	26	88,1
Leukaemia	11,6	25	38	152,1
Thyroid	10,3	22	27	121,7
Cervix Uteri	10,2	22	25	113,8
Melanoma	9,8	21	19	90,0
Kidney	8,4	18	16	88,5
Brain and Other Nervous System	7,2	16	33	212,9
Oral Cavity and Pharynx	7,2	16	16	103,2
Liver	6,8	15	8	54,6
Myeloma	5,3	11	9	78,9
Gallbladder and Other Biliary	4,3	9	6	64,8
Hodgkin lymphoma	2,7	6	5	86,0
Oesophagus	1,6	3	2	58,1
Total without non-melanoma	419,0	902	845	93,7

TABLE B.8: Comparation of expected cases in women against the residents of Lleida of Population Cancer Registry. 2015 (215,310 women)

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Prostate	146,4	322	240	74,5
Colorectal	108,7	239	228	95,4
Lung (and Bronchus)	98,4	216	155	71,6
Urinary bladder	76,5	168	138	82,0
Stomach	22,6	50	41	82,5
Oral Cavity and Pharynx	21,9	48	24	49,8
Liver	18,7	41	24	58,4
Non-Hodgkin lymphoma	18,4	40	19	47,0
Leukaemia	16,6	37	60	164,4
Kidney	15,8	35	48	138,1
Pancreas	15,4	34	26	76,8
Larynx	15	33	27	81,8
Melanoma	11,3	25	29	116,7
Brain and Other Nervous System	10,2	22	44	196,2
Oesophagus	8,7	19	12	62,7
Myeloma	6,5	14	20	139,9
Gallbladder and Other Biliary	4,8	11	11	104,2
Testis	4,3	9	9	95,2
Hodgkin lymphoma	4,1	9	8	88,7
Thyroid	3,8	8	15	179,5
Total without non-melanoma	653,1	1.436	1.211	84,3

TABLE B.9: Comparation of expected cases in men against the residents of Lleida of Population Cancer Registry. 2016 (219,917 men)

Cancer locations	Crude rate	Expected cases	Observed cases	% Coverage
Breast	117,5	252	255	101,4
Colorectal	70,6	151	136	90,0
Cos Uterusí	26,1	56	43	76,9
Lung (and Bronchus)	25,1	54	45	83,7
Urinary bladder	15,5	33	22	66,3
Non-hodgkin lymphoma	14,7	31	29	92,1
Pancreas	14,4	31	21	68,1
Stomach	14	30	27	90,1
Ovary	13,7	29	19	64,8
Leukaemia	11,6	25	30	120,8
Thyroid	10,3	22	18	81,6
Cervix Uteri	10,2	22	16	73,3
Melnoma	9,8	21	23	109,6
Kidney	8,4	18	19	105,6
Brain and Other Nervous System	7,2	15	42	272,4
Oral Cavity and Pharynx	7,2	15	15	97,3
Liver	6,8	15	7	48,1
Myeloma	5,3	11	16	141,0
Gallbladder and Other Biliary	4,3	9	9	97,7
Hodgkin lymphoma	2,7	6	4	69,2
Oesophagus	1,6	3	4	116,8
Total without non-melanoma	419	897	824	91,8

TABLE B.10: Comparation of expected cases in women against the residents of Lleida of Population Cancer Registry. 2016 (214,124 women)

Cancer locations	**AC[a] (2012)**	**AC (2020)**	**APC[b] (2012-2020)**	**CI[c] 95%**
Oral Cavity and Pharynx	1,8	1,66	-3,79	(-11,94 - 5,14)
Esophagus	1,06	1,76	3,29	(-4,80 - 12,06)
Stomach	4,47	3,29	-3,32	(-6,77 - 0,26)
Colorectal	11,02	10,14	-1,52	(-3,40 - 0,39)
Liver	3,36	2,90	-0,25	(-4,39 - 4,08)
Gallbladder and Other Biliary	0,07	0	-5,24	(-21,53 - 14,45)
Pancreas	5,39	4,89	-0,02	(-2,02 - 2,02)
Larynx	1,74	1,27	-5,24	(-10,75 - 0,60)
Lung (and bronchus)	17,33	16,36	-0,43	(-2,17 - 1,34)
Bones and Joints	0,23	0,07	-1,71	(-23,30 - 25,97)
Melanoma	0,68	0,81	0,56	(-6,36 - 7,99)
Non-melanoma	0,4	0,56	19,84	(-1,25 - 45,43)
Breast	6,53	5,34	-1,66	(-5,29 - 2,10)
Cervix Uteri	0,77	0,9	6,91	(-7,00 - 22,90)
Cos uterí	1,28	1,13	-4,88	(-11,11 - 1,78)
Ovary	1,42	1,65	3,52	(-1,24 - 8,50)
Prostate	3,85	2,81	-4,38	(-6,69 - -2,01)
Kidney	2,86	2,18	0,16	(-6,57 - 7,36)
Urinary bladder	2,84	1,62	-2,96	(-7,18 - 1,45)
Brain and Other Nervous System	3,39	2,68	-1,21	(-8,18 - 6,29)
Hodgkin lymphoma	0	0,13	1,82	(-20,62 - 30,61)
Non-Hodgkin lymphoma	0	0	21,44	(-34,99 - 126,86)
Myeloma	1,24	0,7	-5,60	(-10,16 - -0,80)
Leukaemia	2,42	2,5	2,68	(-4,60 - 10,51)
Total	77,5	68,78	-0,98	(-1,73 - -0,22)
Total without non-melanoma	77,1	68,18	0,09	(-6,15 - 6,74)

TABLE B.11: Annual percentage change in mortality in both sexes (2012-2020).

a Age-adjusted rate in the world population per 100,000 inhabitants per year.
b Estimation of the Annual Percentage Change
c Confidence Interval

Cancer locations	**AC[a] (2012)**	**AC (2020)**	**APC[b] (2012-2020)**	**CI[c] 95%**
Oral Cavity and Pharynx	3,64	3,22	-5,24	(-13,30 - 3,57)
Esophagus	2,32	3,5	2,62	(-5,78 - 11,78)
Stomach	7,49	5,19	-3,36	(-7,52 - 0,99)
Colorectal	14,91	12,48	-1,21	(-3,48 - 1,11)
Liver	5,61	4,29	-0,96	(-5,74 - 4,05)
Gallbladder and Other Biliary	0	0	-11,72	(-35,62 - 21,06)
Pancreas	6,31	7,19	1,77	(-1,62 - 5,28)
Larynx	3,44	2,33	-5,44	(-10,13 - -0,52)
Lung (and bronchus)	31,48	26,07	-1,06	(-3,31 - 1,25)
Bones and Joints	0,39	0,08	-7,54	(-29,42 - 21,13)
Melanoma	1,14	1,03	-2,55	(-12,31 - 8,29)
Non-melanoma	0,77	0,8	7,59	(-7,24 - 24,79)
Prostate	9,34	6,77	-4,44	(-6,85 - -1,97)
Kidney	4,3	3,48	1,08	(-5,77 - 8,44)
Urinary bladder	5,39	3,08	-3,95	(-6,85 - -0,96)
Brain and Other Nervous System	3,44	2,88	0,19	(-8,18 - 9,32)
Hodgkin lymphoma	0	0,22	7,70	(-13,73 - 34,46)
Non-Hodgkin lymphoma	0	0	27,60	(-23,90 - 113,94)
Myeloma	2,43	0,93	-7,00	(-17,41 - 4,73)
Leukaemia	3,18	4,23	5,50	(-3,22 - 15,00)
Total	110,03	92,81	-1,13	(-2,56 - 0,32)
Total without non-melanoma	108,89	91,78	-1,12	(-2,61 - 0,41)

TABLE B.12: Annual percentage change in mortality in men (2012-2020).

a Age-adjusted rate in the world population per 100,000 males per year.
b Estimation of the Annual Percentage Change
c Confidence Interval

Cancer locations	**AC[a] (2012)**	**AC (2020)**	**APC[b] (2012-2020)**	**CI[c] 95%**
Oral Cavity and Pharynx	0,06	0,22	3,50	(-17,44 - 29,76)
Esophagus	0,06	0,05	1,55	(-26,51 - 40,32)
Stomach	2,09	1,63	-4,36	(-12,48 - 4,52)
Colorectal	8,15	8,57	-2,24	(-6,71 - 2,46)
Liver	1,43	1,71	1,52	(-3,92 - 7,27)
Gallbladder and Other Biliary	0,12	0	2,06	(-20,31 - 30,72)
Pancreas	4,62	2,72	-3,63	(-8,86 - 1,89)
Larinx	0,34	0,27	-17,95	(-47,81 - 29,01)
Lung (and bronchus)	4,78	8,11	2,04	(-2,85 - 7,18)
Bones and Joints	0,06	0,07	26,32	(-11,41 - 80,11)
Melanoma	0,22	0,57	4,36	(-4,63 - 14,20)
Non-melanoma	0,11	0,4	37,81	(-7,06 - 104,35)
Breast	12,66	10,37	-1,77	(-5,80 - 2,44)
Cervix Uteri	1,56	1,82	7,42	(-6,97 - 24,04)
Cos uterí	2,46	2,16	-4,96	(-11,41 - 1,96)
Ovary	2,72	3,21	3,74	(-1,24 - 8,96)
Kidney	1,75	1,06	-4,93	(-13,46 - 4,44)
Urinary bladder	0,91	0,64	3,43	(-10,13 - 19,03)
Brain and Other Nervous System	3,55	2,48	-3,59	(-13,02 - 6,86)
Hodgkin lymphoma	0	0,05	-4,82	(-35,13 - 39,65)
Myeloma	0,17	0,54	-1,72	(-14,46 - 12,91)
Leukeamia	1,77	0,98	-2,45	(-11,44 - 7,46)
Total	52,48	50,27	-1,05	(-2,77 – 0,70)
Total without non-melanoma	52,26	49,69	-1,14	(-2,96 – 0,72)

TABLE B.13: Annual percentage change in mortality in women (2012-2020).

a Age-adjusted rate in the world population per 100,000 females per year.
b Estimation of the Annual Percentage Change
c Confidence Interval

ICD-O-3.2	Description	Diagnoses code
C01.X	Base of tongue	C02.9
C02.X	Other and unspecified parts of tongue	C02.9
C00.X	Lip	C06.9
C03.X	Gum	C06.9
C04.X	Floor of mouth	C06.9
C05.X	Palate	C06.9
C06.X	Other and unspecified parts of mouth	C06.9
C09.X	Tonsil	C14.0
C10.X	Oropharynx	C14.0
C12.X	Pyriform sinus	C14.0
C13.X	Hypopharyinx	C14.0
C14.X	Other and ill-defined sites in lip, oral cavity and pharynx	C14.0
C19.X	Rectosigmoid junction	C20.9
C20.X	Rectum	C20.9
C23.X	Gallbladder	C24.9
C24.X	Other and unspecified parts of biliary tract	C24.9
C33.X	Trachea	C34.9
C34.X	Bronchus and lung	C34.9
C40.X	Bones joints and articular cartilage of limbs	C41.9
C41.X	Bones joints and articular cartilage of other and unspecified sites	C41.9
C65.X	Renal pelvis	C68.9
C66.X	Ureter	C68.9
C67.X	Bladder	C68.9
C68.X	Other and unspecified urinary organs	C68.9

TABLE B.14: Groups of topographic codes that are considered a single topography.

BIBLIOGRAPHY

1. *Cancer*
2. Soerjomataram, I. & Bray, F. Planning for tomorrow: global cancer incidence and the role of prevention 2020–2070. *Nature Reviews Clinical Oncology* **18**, 663 (2021).
3. Piñeros, M., Saraiya, M., Baussano, I., Bonjour, M., Chao, A. & Bray, F. The role and utility of population-based cancer registries in cervical cancer surveillance and control. *Preventive Medicine* **144**, 106237 (2021).
4. Redondo-Sánchez, D., Rodríguez-Barranco, M., Ameijide, A., Alonso, F. J., Fernández-Navarro, P., Jiménez-Moleón, J. J. & Sánchez, M. J. Cancer incidence estimation from mortality data: a validation study within a population-based cancer registry. *Population Health Metrics* **19**, 1 (1 2021).
5. Bray, F. & Parkin, D. M. Evaluation of data quality in the cancer registry: Principles and methods. Part I: Comparability, validity and timeliness. *European Journal of Cancer* **45**, 747 (5 2009).
6. Piñeros, M., Znaor, A., Mery, L. & Bray, F. A Global Cancer Surveillance Framework Within Noncommunicable Disease Surveillance: Making the Case for Population-Based Cancer Registries. *Epidemiologic Reviews* **39**, 161 (2017).
7. Sung, H., Siegel, R. L., Rosenberg, P. S. & Jemal, A. Emerging cancer trends among young adults in the USA: analysis of a population-based cancer registry. *The Lancet Public Health* **4**, e137 (2019).
8. Tucker, T. C., Durbin, E. B., McDowell, J. K. & Huang, B. Unlocking the potential of population-based cancer registries. *Cancer* **125**, 3729 (2019).
9. *The population-based cancer registries | ECIS*
10. *Cancer Registry*
11. Navarro, C., Martos, C., Ardanaz, E., Galceran, J., Izarzugaza, I., Peris-Bonet, R., Martínez, C., Argüelles, M., Garau, I., Almar, E., Izquierdo, Á., Díaz, J. M., Rojas, D., Perucha, J., Torrella, A. & Vicente-Raneda, M. L. Population-based cancer registries in Spain and their role in cancer control. *Annals of Oncology* **21**, iii3 (2010).

12. Coebergh, J. W., Van Den Hurk, C., Louwman, M., Comber, H., Rosso, S., Zanetti, R., Sacchetto, L., Storm, H., Van Veen, E. B., Siesling, S. & Van Den Eijnden-Van Raaij, J. EUROCOURSE recipe for cancer surveillance by visible population-based cancer RegisTrees® in Europe: From roots to fruits. *European Journal of Cancer* **51**, 1050 (2015).

13. *Home | Redecan*

14. Ribes, J., Gálvez, J., Melià, À., Clèries, R., Messeguer, X. & Bosch, F. X. Automatización de un registro hospitalario de tumores. *Gaceta Sanitaria* **19**, 221 (2005).

15. International Agency for Research on Cancer, World Health Organization, I. A. o. C. & Registries, E. N. o. C. R. *International Rules for Multiple Primary Cancers (ICD-O Third Edition)* tech. rep. (2004).

16. Allemani, C., Matsuda, T., Di Carlo, V., Harewood, R., Matz, M., Nikšić, M. & et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet* **391**, 1023 (2018).

17. El-Shami, K., Oeffinger, K. C., Erb, N. L., Willis, A., Bretsch, J. K., Pratt-Chapman, M. L., Cannady, R. S., Wong, S. L., Rose, J., Barbour, A. L., Stein, K. D., Sharpe, K. B., Brooks, D. D. & Cowens-Alvarado, R. L. American Cancer Society Colorectal Cancer Survivorship Care Guidelines. *CA: A Cancer Journal for Clinicians* **65**, 427 (2015).

18. Supramaniam, R. New malignancies among cancer survivors: SEER cancer registries, 1973-2000. *Journal of Epidemiology & Community Health* **62**, 375 (2008).

19. Vogt, A., Schmid, S., Heinimann, K., Frick, H., Herrmann, C., Cerny, T. & Omlin, A. Multiple primary tumours: challenges and approaches, a review. eng. *ESMO open* **2**, e000172 (2017).

20. Druesne-Pecollo, N., Keita, Y., Touvier, M., Chan, D. S., Norat, T., Hercberg, S. & Latino-Martel, P. Alcohol drinking and second primary cancer risk in patients with upper aerodigestive tract cancers: A systematic review and meta-Analysis of observational studies. *Cancer Epidemiology Biomarkers and Prevention* **23**, 324 (2014).

21. Tabuchi, T., Ito, Y., Ioka, A., Nakayama, T., Miyashiro, I. & Tsukuma, H. Tobacco smoking and the risk of subsequent primary cancer among cancer survivors: A retrospective cohort study. *Annals of Oncology* **24**, 2699 (2013).

22. Sung, H., Hyun, N., Leach, C. R., Yabroff, K. R. & Jemal, A. Association of First Primary Cancer With Risk of Subsequent Primary Cancer Among Survivors of Adult-Onset Cancers in the United States. *JAMA* **324**, 2521 (2020).

23. Tran, K. B., Lang, J. J., Compton, K., Xu, R., Acheson, A. R., Henrikson, H. J., Kocarnik, J. M. & et al. The global burden of cancer attributable to risk factors, 2010–19: a systematic analysis for the Global Burden of Disease Study 2019. *The Lancet* **400**, 563 (2022).

24. Stein, C. J. & Colditz, G. A. *Modifiable risk factors for cancer* 2004.

25. Jacob, L., Freyn, M., Kalder, M., Dinas, K. & Kostev, K. Impact of tobacco smoking on the risk of developing 25 different cancers in the UK: A retrospective study of 422,010 patients followed for up to 30 years. *Oncotarget* **9**, 17420 (2018).

26. Klein, W. M., Jacobsen, P. B. & Helzlsouer, K. J. *Alcohol and Cancer Risk: Clinical and Research Implications* 2020.

27. Cho, A., Chang, Y., Ahn, J., Shin, H. & Ryu, S. Cigarette smoking and thyroid cancer risk: a cohort study. *British Journal of Cancer 2018 119:5* **119**, 638 (2018).

28. Su, B., Qin, W., Xue, F., Wei, X., Guan, Q., Jiang, W., Wang, S., Xu, M. & Yu, S. The relation of passive smoking with cervical cancer: A systematic review and meta-analysis. *Medicine* **97**, e13061 (2018).

29. (US), S. A., Administration, M. H. S. & (US), O. o. t. S. G. *Smoking Cessation: A Report of the Surgeon General* (US Department of Health and Human Services, 2020).

30. Chuang, S. C., Lee, Y. C. A., Wu, G. J., Straif, K. & Hashibe, M. Alcohol consumption and liver cancer risk: a meta-analysis. *Cancer Causes and Control* **26**, 1205 (2015).

31. *Limit alcohol consumption | WCRF International*

32. Choi, Y. J., Lee, D. H., Han, K. D., Kim, H. S., Yoon, H., Shin, C. M., Park, Y. S. & Kim, N. The relationship between drinking alcohol and esophageal, gastric or colorectal cancer: A nationwide population-based cohort study of South Korea. *PLOS ONE* **12**, e0185778 (2017).

33. Sung, H., Siegel, R. L., Torre, L. A., Pearson-Stuttard, J., Islami, F., Fedewa, S. A., Sauer, A. G., Shuval, K., Gapstur, S. M., Jacobs, E. J., Giovannucci, E. L. & Jemal, A. Global patterns in excess body weight and the associated cancer burden. *CA: A Cancer Journal for Clinicians* **69**, 88 (2019).

34. Florensa, D., Pedrol, T., Mòdol, I., Farré, X., Salud, A., Mateo, J. & Godoy, P. El registre poblacional de càncer a Lleida en zones urbanes i rurals. Resultats de l'any 2014. *Butlletí Epidemiològic Catalunya* **40**, 252 (2020).

35. Florensa, D., Godoy, P., Mateo, J., Solsona, F., Pedrol, T., Mesas, M. & Pinol, R. The Use of Multiple Correspondence Analysis to Explore Associations between Categories of Qualitative Variables and Cancer Incidence. *IEEE Journal of Biomedical and Health Informatics* **25**, 3659 (2021).

36. *Targeta sanitària individual (TSI). CatSalut. Servei Català de la Salut*

37. Torres-Bondia, F., Dakterzada, F., Galván, L., Buti, M., Besanson, G., Gill, E., Buil, R., de Batlle, J. & Piñol-Ripoll, G. Proton pump inhibitors and the risk of Alzheimer's disease and non-Alzheimer's dementias. *Scientific Reports* **10**, 1 (2020).

38. Eslami, L. & Nasseri-Moghaddam, S. Meta-analyses: Does long-term PPI use increase the risk of gastric premalignant lesions? *Archives of Iranian Medicine* **16**, 449 (2013).

39. Rothwell, P. M., Wilson, M., Elwin, C. E., Norrving, B., Algra, A., Warlow, C. P. & Meade, T. W. Long-term effect of aspirin on colorectal cancer incidence and mortality: 20-year follow-up of five randomised trials. *The Lancet* (2010).

40. Patrono, C. & Baigent, C. Role of aspirin in primary prevention of cardiovascular disease. *Nature Reviews Cardiology* **16**, 675 (2019).

41. Liu, C., Du, L., Wang, S., Kong, L., Zhang, S., Li, S., Zhang, W. & Du, G. Differences in the prevention and control of cardiovascular and cerebrovascular diseases. *Pharmacological Research* **170**, 105737 (2021).

42. Tsoi, K. K., Ho, J. M., Chan, F. C. & Sung, J. J. Long-term use of low-dose aspirin for cancer prevention: A 10-year population cohort study in Hong Kong. *International Journal of Cancer* **145**, 267 (2019).

43. Bosetti, C., Santucci, C., Gallus, S., Martinetti, M. & La Vecchia, C. Aspirin and the risk of colorectal and other digestive tract cancers: an updated meta-analysis through 2019. *Annals of Oncology* **31**, 558 (2020).

44. Sankaranarayanan, R., Kumar, D. R., Altinoz, M. A. & Bhat, G. J. Mechanisms of Colorectal Cancer Prevention by Aspirin—A Literature Review and Perspective on the Role of COX-Dependent and -Independent Pathways. *International Journal of Molecular Sciences 2020, Vol. 21, Page 9018* **21**, 9018 (2020).

45. Murtagh, F. Multiple correspondence analysis and related methods. *Psychometrika* **72**, 275 (2007).

46. Husson, F. & Josse, J. in, 165 (Chapman and Hall, Boca Raton, 2014).

47. Di Franco, G. Multiple correspondence analysis: one only or several techniques? *Quality and Quantity* **50**, 1299 (2016).

48. Heckler, C. E. Applied Multivariate Statistical Analysis. *Technometrics* **47**, 517 (2005).

49. Likas, A., Vlassis, N. & J. Verbeek, J. The global k-means clustering algorithm. *Pattern Recognition* **36**, 451 (2003).

50. Bholowalia, P. & Kumar, A. *EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN* tech. rep. 9 (2014), 975.

51. Cutler, A., Cutler, D. R. & Stevens, J. R. *Random Forests* 157 (Springer, Boston, MA, Boston, MA, 2012).

52. Nowozin, S., Rother, C., Bagon, S., Sharp, T., Yao, B. & Kohli, P. Decision tree fields. *Proceedings of the IEEE International Conference on Computer Vision*, 1668 (2011).

53. Wang, S.-C. *Artificial Neural Network* 81 (Springer, Boston, MA, Boston, MA, 2003).

54. Xu, S. Bayesian Naïve Bayes classifiers to text classification: *Journal of Information Science* **44**, 48 (2016).

55. Izenman, A. J. *Linear Discriminant Analysis* 237 (Springer, New York, NY, 2013).

56. Lemon, S. C., Roy, J., Clark, M. A., Friedmann, P. D. & Rakowski, W. Classification and regression tree analysis in public health: Methodological review and comparison with logistic regression. *Annals of Behavioral Medicine 2003 26:3* **26**, 172 (2003).

57. Suthaharan, S. *Support Vector Machine* 207 (Springer, Boston, MA, 2016).

58. Tharwat, A. Classification assessment methods. *Applied Computing and Informatics* (2018).

59. Larsson, S. C., Carter, P., Kar, S., Vithayathil, M., Mason, A. M., Michaëlsson, K. & Burgess, S. Smoking, alcohol consumption, and cancer: A mendelian randomisation study in UK Biobank and international genetic consortia participants. *PLOS Medicine* **17**, e1003178 (2020).

60. Henley, S., Anderson, R., CC, T., GM, M., B, P. & LC, R. Invasive Cancer Incidence, 2004–2013, and Deaths, 2006–2015, in Nonmetropolitan and Metropolitan Counties — United States. *MMWR Surveill Summ* **66**, 1 (2017).

61. Pilleron, S., Sarfati, D., Janssen-Heijnen, M., Vignat, J., Ferlay, J., Bray, F. & Soerjomataram, I. Global cancer incidence in older adults, 2012 and 2035: A population-based study. *International Journal of Cancer* **144**, 49 (2018).

62. Darlington, W. S. & Green, A. L. The role of geographic distance from a cancer center in survival and stage of AYA cancer diagnoses. *Cancer* **127**, 3508 (2021).

63. Sharp, L., Donnelly, D., Hegarty, A., Carsin, A. E., Deady, S., McCluskey, N., Gavin, A. & Comber, H. Risk of several cancers is higher in urban areas after adjusting for socioeconomic status. Results from a two-country population-based study of 18 common cancers. *Journal of Urban Health* **91**, 510 (2014).

64. Zahnd, W. E., James, A. S., Jenkins, W. D., Izadi, S. R., Fogleman, A. J., Steward, D. E., Colditz, G. A. & Brard, L. Rural-Urban Differences in Cancer Incidence and Trends in the United States. eng. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **27**, 1265 (2018).

65. Rothwell, P. M., Fowkes, F. G. R., Belch, J. F., Ogawa, H., Warlow, C. P. & Meade, T. W. Effect of daily aspirin on long-term risk of death due to cancer: analysis of individual patient data from randomised trials. *The Lancet* **377**, 31 (2011).

66. Maniewska, J. & Jeżewska, D. Non-Steroidal Anti-Inflammatory Drugs in Colorectal Cancer Chemoprevention. *Cancers 2021, Vol. 13, Page 594* **13**, 594 (2021).

67. García Rodríguez, L. A., Soriano-Gabarró, M., Vora, P. & Cea Soriano, L. Low-dose aspirin and risk of gastric and oesophageal cancer: A population-based study in the United Kingdom using The Health Improvement Network. *International Journal of Cancer* **147**, 2394 (2020).

68. Hwang, I. C., Chang, J., Kim, K. & Park, S. M. Aspirin Use and Risk of Hepatocellular Carcinoma in a National Cohort Study of Korean Adults. *Scientific Reports* **8**, 4968 (2018).

69. Risch, H. A., Lu, L., Streicher, S. A., Wang, J., Zhang, W., Ni, Q., Kidd, M. S., Yu, H. & Gao, Y. T. Aspirin use and reduced risk of pancreatic cancer. *Cancer Epidemiology Biomarkers and Prevention* **26**, 68 (2017).

70. Kang, J., Jeong, S. M., Shin, D. W., Cho, M., Cho, J. H. & Kim, J. The Associations of Aspirin, Statins, and Metformin With Lung Cancer Risk and Related Mortality: A Time-Dependent Analysis of Population-Based Nationally Representative Data. *Journal of Thoracic Oncology* **16**, 76 (2021).

71. Ma, S., Guo, C., Sun, C., Han, T., Zhang, H., Qu, G., Jiang, Y., Zhou, Q. & Sun, Y. Aspirin Use and Risk of Breast Cancer: A Meta-analysis of Observational Studies from 1989 to 2019. *Clinical Breast Cancer* **21**, 552 (2021).

72. Miller, D. M. & Shalhout, S. Z. BodyMapR: an R package and Shiny application designed to generate anatomical visualizations of cancer lesions. *JAMIA Open* **5** (2022).

73. Zhang, P., Palmisano, A., Kumar, R., Li, M.-C., Doroshow, J. H. & Zhao, Y. TPWshiny: an interactive R/Shiny app to explore cell line transcriptional responses to anti-cancer drugs. *Bioinformatics* **38**, 570 (2022).

74. Xia, Q., Mudaranthakam, D. P., Chollet-Hinton, L., Chen, R., Krebill, H., Kuo, H. & Koestler, D. C. shinyOPTIK, a User-Friendly R Shiny Application for Visualizing Cancer Risk Factors and Mortality Across the University of Kansas Cancer Center Catchment Area. *JCO Clinical Cancer Informatics* (2022).

75. Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V. & Fotiadis, D. I. Machine learning applications in cancer prognosis and prediction. **13**, 8 (2015).

76. Delen, D., Walker, G. & Kadam, A. Predicting breast cancer survivability: A comparison of three data mining methods. *Artificial Intelligence in Medicine* **34**, 113 (2005).

77. Chen, Y. C., Ke, W. C. & Chiu, H. W. Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Computers in Biology and Medicine* **48**, 1 (2014).

78. Goldenberg, S. L., Nir, G. & Salcudean, S. E. A new era: artificial intelligence and machine learning in prostate cancer. **16**, 391 (2019).

79. Behravan, H., Hartikainen, J. M., Tengström, M., Kosma, V.-M. .-. & Mannermaa, A. Predicting breast cancer risk using interacting genetic and demographic factors and machine learning. *Scientific Reports* **10**, 1 (2020).

80. Alhazmi, A., Alhazmi, Y., Makrami, A., Masmali, A., Salawi, N., Masmali, K. & Patil, S. Application of artificial intelligence and machine learning for prediction of oral cancer risk. *Journal of Oral Pathology & Medicine* **50**, 444 (2021).

81. Costa, P. S., Santos, N. C., Cunha, P., Cotter, J. & Sousa, N. The use of multiple correspondence analysis to explore associations between categories of qualitative variables in healthy ageing. *Journal of Aging Research* **72**, 257 (2013).

82. Das, S., Avelar, R., Dixon, K. & Sun, X. Investigation on the wrong way driving crash patterns using multiple correspondence analysis. *Accident Analysis and Prevention* **111**, 43 (2018).

83. White, A. J., Keller, J. P., Zhao, S., Carroll, R., Kaufman, J. D. & Sandler, D. P. Air Pollution, Clustering of Particulate Matter Components, and Breast Cancer in the Sister Study: A U.S.-Wide Cohort. *Environmental Health Perspectives* **127**, 107002 (2019).

84. Mancini, R., Pattaro, G., Diodoro, M. G., Sperduti, I., Garufi, C., Stigliano, V., Perri, P., Grazi, G. L. & Cosimelli, M. Tumor Regression Grade After Neoadjuvant Chemoradiation and Surgery for Low Rectal Cancer Evaluated by Multiple Correspondence Analysis: Ten Years as Minimum Follow-up. *Clinical Colorectal Cancer* **17**, e13 (2018).

85. Violán, C., Roso-Llorach, A., Foguet-Boreu, Q., Guisado-Clavero, M., Pons-Vigués, M., Pujol-Ribera, E. & Valderas, J. M. Multimorbidity patterns with K-means nonhierarchical cluster analysis. *BMC Family Practice* **19**, 108 (2018).

86. Steele, C. B., Thomas, C. C., Henley, S. J., Massetti, G. M., Galuska, D. A., Agurs-Collins, T., Puckett, M. & Richardson, L. C. Vital Signs : Trends in Incidence of Cancers Associated with Overweight and Obesity — United States, 2005–2014. *MMWR. Morbidity and Mortality Weekly Report* **66**, 1052 (2017).

87. Liu, P. H., Wu, K., Ng, K., Zauber, A. G., Nguyen, L. H., Song, M., He, X., Fuchs, C. S., Ogino, S., Willett, W. C., Chan, A. T., Giovannucci, E. L. & Cao, Y. Association of Obesity with Risk of Early-Onset Colorectal Cancer among Women. *JAMA Oncology* **5**, 37 (2019).

88. Baziliansky, S. & Cohen, M. Emotion Regulation Patterns among Colorectal Cancer Survivors: Clustering and Associations with Personal Coping Resources. *Behavioral Medicine* **47**, 214 (2021).

89. Brockmoeller, S., Echle, A., Ghaffari Laleh, N., Eiholm, S., Malmstrøm, M. L., Plato Kuhlmann, T., Levic, K., Grabsch, H. I., West, N. P., Saldanha, O. L., Kouvidi, K., Bono, A., Heij, L. R., Brinker, T. J., Gögenür, I., Quirke, P. & Kather, J. N. Deep learning identifies inflamed fat as a risk factor for lymph node metastasis in early colorectal cancer. *The Journal of Pathology* **256**, 269 (2022).

90. Xiao, Y., Wu, J., Lin, Z. & Zhao, X. A deep learning-based multi-model ensemble method for cancer prediction. *Computer Methods and Programs in Biomedicine* **153**, 1 (2018).

91. Zhang, Y., Zhu, S., Yuan, Z., Li, Q., Ding, R., Bao, X., Zhen, T., Fu, Z., Fu, H., Xing, K., Yuan, H. & Chen, T. Risk factors and socio-economic burden in pancreatic ductal adenocarcinoma operation: a machine learning based analysis. *BMC Cancer 2020 20:1* **20**, 1 (2020).

92. X, Y., H, L., T, S. & PW, S. Ensemble Feature Learning to Identify Risk Factors for Predicting Secondary Cancer. *International journal of medical sciences* **16**, 949 (2019).

93. Youlden, D. R. & Baade, P. D. The relative risk of second primary cancers in Queensland, Australia: A retrospective cohort study. *BMC Cancer* **11** (2011).

94. Feller, A., Matthes, K. L., Bordoni, A., Bouchardy, C., Bulliard, J. L., Herrmann, C., Konzelmann, I., Maspoli, M., Mousavi, M., Rohrmann, S., Staehelin, K., Arndt, V., Staehelin, K., Bouchardy, C., Mousavi, M., Bulliard, J. L., Maspoli, M., Mousavi, M., Bordoni, A., Konzelmann, I., Blanc-Moya, R. & Rohrmann, S. The relative risk of second primary cancers in Switzerland: A population-based retrospective cohort study. *BMC Cancer* **20** (2020).

95. Tabuchi, T., Ozaki, K., Ioka, A. & Miyashiro, I. Joint and independent effect of alcohol and tobacco use on the risk of subsequent cancer incidence among cancer survivors: A cohort study using cancer registries. *International Journal of Cancer* **137**, 2114 (2015).

96. Sang, M. P., Min, K. L., Kyu, W. J., Soon, A. S., Yoo, K. Y., Young, H. Y. & Bong, Y. H. Prediagnosis smoking, obesity, insulin resistance, and second primary cancer risk in male cancer survivors: National Health Insurance Corporation Study. *Journal of Clinical Oncology* **25**, 4835 (2007).

97. Park, S. M., Li, T., Wu, S., Li, W. Q., Qureshi, A. A., Stampfer, M. & Cho, E. Risk of second primary cancer associated with pre-diagnostic smoking, alcohol, and obesity in women with keratinocyte carcinoma. *Cancer Epidemiology* **47**, 106 (2017).

98. Li, C. I., Daling, J. R., Tang, M. T. C. & Malone, K. E. Relationship between diabetes and risk of second primary contralateral breast cancer. *Breast Cancer Research and Treatment* **125**, 545 (2011).

99. Jassem, J. *Tobacco smoking after diagnosis of cancer: Clinical aspects* 2019.

100. Sawicki, T., Ruszkowska, M., Danielewicz, A., Niedźwiedzka, E., Arłukowicz, T. & Przybyłowicz, K. E. *A review of colorectal cancer in terms of epidemiology, risk factors, development, symptoms and diagnosis* 2021.

101. Youlden, D. R., Youl, P. H., Soyer, H. P., Aitken, J. F. & Baade, P. D. Distribution of subsequent primary invasive melanomas following a first primary invasive or in situ melanoma Queensland, Australia, 1982-2010. *JAMA dermatology* **150**, 526 (2014).

102. *eCAP. Departament de Salut*

103. Sierosławski, J., Foster, J. & Moskalewicz, J. Survey of European drinking surveys. Alcohol survey experiences of 22 European countries. *Drugs: Education, Prevention and Policy* **20**, 383 (2013).

104. *Tobacco*

105. Barclay, M. E., Lyratzopoulos, G., Walter, F. M., Jefferies, S., Peake, M. D. & Rintoul, R. C. Incidence of second and higher order smoking-related primary cancers following lung cancer: A population-based cohort study. *Thorax* **74**, 466 (2019).

106. Sawada, N., Inoue, M., Iwasaki, M., Sasazuki, S., Yamaji, T., Shimazu, T. & Tsugane, S. Alcohol and smoking and subsequent risk of prostate cancer in Japanese men: The Japan Public Health Center-based prospective study. *International Journal of Cancer* **134**, 971 (2014).

107. Wang, M., Sharma, A., Osazuwa-Peters, N., Simpson, M. C., Schootman, M., Piccirillo, J. F., Huh, W. K. & Adjei Boakye, E. Risk of subsequent malignant neoplasms after an index potentially-human papillomavirus (HPV)-associated cancers. *Cancer Epidemiology* **64**, 101649 (2020).

108. Gilbert, D. C., Wakeham, K., Langley, R. E. & Vale, C. L. Increased risk of second cancers at sites associated with HPV after a prior HPV-associated malignancy, a systematic review and meta-analysis. *British Journal of Cancer* **120**, 256 (2019).

109. Radkiewicz, C., Johansson, A. L., Dickman, P. W., Lambe, M. & Edgren, G. Sex differences in cancer risk and survival: A Swedish cohort study. *European Journal of Cancer* **84**, 130 (2017).

110. Wilsnack, R. W., Wilsnack, S. C., Kristjanson, A. F., Vogeltanz-Holm, N. D. & Gmel, G. Gender and alcohol consumption: Patterns from the multinational GENACIS project. *Addiction* **104**, 1487 (2009).

111. Ragusa, R., Torrisi, A., Di Prima, A. A., Torrisi, A. A., Ippolito, A., Ferrante, M., Madeddu, A. & Guardabasso, V. Cancer Prevention for Survivors: Incidence of Second Primary Cancers and Sex Differences—A Population-Based Study from an Italian Cancer Registry. *International Journal of Environmental Research and Public Health* **19**, 12201 (2022).

112. Harding, C., Pompei, F. & Wilson, R. Peak and decline in cancer incidence, mortality, and prevalence at old ages. *Cancer* **118**, 1371 (2012).

113. Permanyer, I. & Scholl, N. Global trends in lifespan inequality: 1950-2015. *PLOS ONE* **14**, 1 (2019).

114. Nolen, S. C., Evans, M. A., Fischer, A., Corrada, M. M., Kawas, C. H. & Bota, D. A. Cancer—Incidence, prevalence and mortality in the oldest-old. A comprehensive review. *Mechanisms of Ageing and Development* **164**, 113 (2017).

115. Moke, D. J., Hamilton, A. S., Chehab, L., Deapen, D. & Freyer, D. R. Obesity and risk for second malignant neoplasms in childhood cancer survivors: A case–control study utilizing the California cancer registry. *Cancer Epidemiology Biomarkers and Prevention* **28**, 1612 (2019).

116. Chen, S. T., Hsueh, C., Chiou, W. K. & Lin, J. D. Disease-Specific Mortality and Secondary Primary Cancer in Well-Differentiated Thyroid Cancer with Type 2 Diabetes Mellitus. *PLOS ONE* **8**, e55179 (2013).

117. Chen, S. C., Teng, C. J., Hu, Y. W., Yeh, C. M., Hung, M. H., Hu, L. Y., Ku, F. C., Tzeng, C. H., Chiou, T. J., Chen, T. J. & Liu, C. J. Secondary primary malignancy risk among patients with esophageal cancer in Taiwan: A nationwide population-based study. *PLoS ONE* **10**, e0116384 (2015).

118. Liu, W. S., Chang, Y. J., Lin, C. L., Liang, J. A., Sung, F. C., Hwang, I. M. & Kao, C. H. Secondary primary cancer in patients with head and neck carcinoma: The differences among hypopharyngeal, laryngeal, and other sites of head and neck cancer. *European Journal of Cancer Care* **23**, 36 (2014).

119. Aredo, J. V., Luo, S. J., Gardner, R. M., Sanyal, N., Choi, E., Hickey, T. P., Riley, T. L., Huang, W. Y., Kurian, A. W., Leung, A. N., Wilkens, L. R., Robbins, H. A., Riboli, E., Kaaks, R., Tjønneland, A., Vermeulen, R. C., Panico, S., Le Marchand, L., Amos, C. I., Hung, R. J., Freedman, N. D., Johansson, M., Cheng, I., Wakelee, H. A. & Han, S. S. Tobacco Smoking and Risk of Second Primary Lung Cancer. *Journal of Thoracic Oncology* **16**, 968 (2021).

120. Gallaway, M. S., Henley, S. J., Steele, C. B., Momin, B., Thomas, C. C., Jamal, A., Trivers, K. F., Singh, S. D. & Stewart, S. L. Surveillance for cancers associated with Tobacco Use - United States, 2010-2014. *MMWR Surveillance Summaries* **67**, 1 (2018).

121. Mahabir, S., Leitzmann, M. F., Virtanen, M. J., Virtamo, J., Pietinen, P., Albanes, D. & Taylor, P. R. Prospective study of alcohol drinking and renal cell cancer risk in a cohort of finnish male smokers. *Cancer Epidemiology Biomarkers and Prevention* **14**, 170 (2005).

122. Ishiguro, S., Sasazuki, S., Inoue, M., Kurahashi, N., Iwasaki, M. & Tsugane, S. Effect of alcohol consumption, cigarette smoking and flushing response on esophageal cancer risk: A population-based cohort study (JPHC study). *Cancer Letters* **275**, 240 (2009).

123. Tian, F., Fang, F., Shen, Q., Ye, W., Valdimarsdóttir, U. A. & Song, H. Stress-related disorders and subsequent cancer risk and mortality: a population-based and sibling-controlled cohort study in Sweden. *European Journal of Epidemiology* **37**, 947 (2022).

124. Capodanno, D. & Angiolillo, D. J. Aspirin for Primary Cardiovascular Risk Prevention and Beyond in Diabetes Mellitus. *Circulation* **134**, 1579 (2016).

125. Burn, J., Sheth, H., Elliott, F., Reed, L., Macrae, F., Mecklin, J. P. & etal. Cancer prevention with aspirin in hereditary colorectal cancer (Lynch syndrome), 10-year follow-up and registry-based 20-year data in the CAPP2 study: a double-blind, randomised, placebo-controlled trial. *The Lancet* **395**, 1855 (2020).

126. Patrignani, P. & Patrono, C. Aspirin and Cancer. *Journal of the American College of Cardiology* **68**, 967 (2016).

127. Cho, M. H., Yoo, T. G., Jeong, S. M. & Shin, D. W. Association of Aspirin, Metformin, and Statin use with gastric cancer incidence and mortality: A nationwide cohort study. *Cancer Prevention Research* **14**, 95 (2021).

128. Wang, Y., Shen, C., Ge, J. & Duan, H. Regular aspirin use and stomach cancer risk in China. *European Journal of Surgical Oncology* **41**, 801 (2015).

129. Simon, T. G., Duberg, A.-S., Aleman, S., Chung, R. T., Chan, A. T. & Ludvigsson, J. F. Association of Aspirin with Hepatocellular Carcinoma and Liver-Related Mortality. *New England Journal of Medicine* **382**, 1018 (2020).

130. Hurwitz, L. M., Michels, K. A., Cook, M. B., Pfeiffer, R. M. & Trabert, B. Associations between daily aspirin use and cancer risk across strata of major cancer risk factors in two large U.S. cohorts. *Cancer Causes and Control* **32**, 57 (2021).

131. McGlynn, K. A., Petrick, J. L. & El-Serag, H. B. Epidemiology of Hepatocellular Carcinoma. *Hepatology (Baltimore, Md.)* **73 Suppl 1**, 4 (2021).

132. *WHO | World Health Organization*

133. Guo, C. G., Ma, W., Drew, D. A., Cao, Y., Nguyen, L. H., Joshi, A. D., Ng, K., Ogino, S., Meyerhardt, J. A., Song, M., Leung, W. K., Giovannucci, E. L. & Chan, A. T. Aspirin Use and Risk of Colorectal Cancer Among Older Adults. *JAMA oncology* **7**, 428 (2021).

134. Thun, M. J., Jacobs, E. J. & Patrono, C. The role of aspirin in cancer prevention. *Nature Reviews Clinical Oncology 2012 9:5* **9**, 259 (2012).

135. Jacobo-Herrera, N. J., Pérez-Plasencia, C., Camacho-Zavala, E., Figueroa González, G., López Urrutia, E., Garćia-Castillo, V. & Zentella-Dehesa, A. Clinical evidence of the relationship between aspirin and breast cancer risk (review). *Oncology Reports* **32**, 451 (2014).

136. Drew, D. A., Cao, Y. & Chan, A. T. Aspirin and colorectal cancer: the promise of precision chemoprevention. *Nature Reviews Cancer 2016 16:3* **16**, 173 (2016).

137. Keum, N. N. & Giovannucci, E. Global burden of colorectal cancer: emerging trends, risk factors and prevention strategies. *Nature Reviews Gastroenterology & Hepatology 2019 16:12* **16**, 713 (2019).

138. Liao, L. M., Vaughan, T. L., Corley, D. A., Cook, M. B., Casson, A. G., Kamangar, F., Abnet, C. C., Risch, H. A., Giffen, C., Freedman, N. D., Chow, W., Sadeghi, S., Pandeya, N., Whiteman, D. C., Murray, L. J., Bernstein, L., Gammon, M. D. & Wu, A. H. Nonsteroidal Anti-inflammatory Drug Use Reduces Risk of Adenocarcinomas of the Esophagus and Esophagogastric Junction in a Pooled Analysis. *Gastroenterology* **142**, 442 (2012).

139. Yang, Y., Wang, G., He, J., Ren, S., Wu, F., Zhang, J. & Wang, F. Gender differences in colorectal cancer survival: A meta-analysis. *International Journal of Cancer* **141**, 1942 (2017).

140. Wu, K., Feskanich, D., Fuchs, C. S., Willett, W. C., Hollis, B. W. & Giovannucci, E. L. A Nested Case–Control Study of Plasma 25-Hydroxyvitamin D Concentrations and Risk of Colorectal Cancer. *JNCI: Journal of the National Cancer Institute* **99**, 1120 (2007).

141. Shahjehan, F., Merchea, A., Cochuyt, J. J., Li, Z., Colibaseanu, D. T. & Kasi, P. M. Body mass index and long-term outcomes in patients with colorectal cancer. *Frontiers in Oncology* **8**, 620 (2018).

142. Jaspan, V., Lin, K. & Popov, V. The impact of anthropometric parameters on colorectal cancer prognosis: A systematic review and meta-analysis. *Critical Reviews in Oncology/Hematology* **159**, 103232 (2021).

143. Cleary, M. P. & Grossmann, M. E. Obesity and Breast Cancer: The Estrogen Connection. *Endocrinology* **150**, 2537 (2009).

144. Vidal, A. C., Oyekunle, T., Howard, L. E., De Hoedt, A. M., Kane, C. J., Terris, M. K., Cooperberg, M. R., Amling, C. L., Klaassen, Z., Freedland, S. J. & Aronson, W. J. Obesity, race, and long-term prostate cancer outcomes. *Cancer* **126**, 3733 (2020).

145. Jeon, J., Holford, T. R., Levy, D. T., Feuer, E. J., Cao, P., Tam, J., Clarke, L., Clarke, J., Kong, C. Y. & Meza, R. Smoking and Lung Cancer Mortality in the United States From 2015 to 2065: A Comparative Modeling Approach. *Annals of internal medicine* **169**, 684 (2018).

146. Antoni, S., Ferlay, J., Soerjomataram, I., Znaor, A., Jemal, A. & Bray, F. Bladder Cancer Incidence and Mortality: A Global Overview and Recent Trends. *European Urology* **71**, 96 (2017).

147. Botteri, E., Borroni, E., Sloan, E. K., Bagnardi, V., Bosetti, C., Peveri, G., Santucci, C., Specchia, C., Van Den Brandt, P., Gallus, S. & Lugo, A. Smoking and Colorectal Cancer Risk, Overall and by Molecular Subtypes: A Meta-Analysis. *American Journal of Gastroenterology* **115**, 1940 (2020).

148. Park, S. Y., Wilkens, L. R., Setiawan, V. W., Monroe, K. R., Haiman, C. A. & Le Marchand, L. Alcohol Intake and Colorectal Cancer Risk in the Multiethnic Cohort Study. *American Journal of Epidemiology* **188**, 67 (2019).

149. Álvarez-Avellón, S. M., Fernández-Somoano, A., Navarrete-Muñoz, E. M., Vioque, J. & Tardón, A. Efecto del alcohol y sus metabolitos en el cáncer de pulmón: estudio CAPUA. *Medicina Clínica* **148**, 531 (2017).

150. Gorini, G., Stagnaro, E., Fontana, V., Miligi, L., Ramazzotti, V., Nanni, O., Rodella, S., Tumino, R., Crosignani, P., Vindigni, C., Fontana, A., Vineis, P. & Costantini, A. S. Alcohol consumption and risk of leukemia: A multicenter case–control study. *Leukemia Research* **31**, 379 (2007).

151. Rota, M., Porta, L., Pelucchi, C., Negri, E., Bagnardi, V., Bellocco, R., Corrao, G., Boffetta, P. & La Vecchia, C. Alcohol drinking and risk of leukemia—A systematic review and meta-analysis of the dose–risk relation. *Cancer Epidemiology* **38**, 339 (2014).

152. Tramacere, I., Scotti, L., Jenab, M., Bagnardi, V., Bellocco, R., Rota, M., Corrao, G., Bravi, F., Boffetta, P. & La Vecchia, C. Alcohol drinking and pancreatic cancer risk: a meta-analysis of the dose-risk relation. *International Journal of Cancer* **126**, 1474 (2010).

153. Tramacere, I., Pelucchi, C., Bonifazi, M., Bagnardi, V., Rota, M., Bellocco, R., Scotti, L., Islami, F., Corrao, G., Boffetta, P., La Vecchia, C. & Negri, E. A meta-analysis on alcohol drinking and the risk of Hodgkin lymphoma. *European Journal of Cancer Prevention* **21**, 268 (2012).

154. Pelucchi, C., Galeone, C., Tramacere, I., Bagnardi, V., Negri, E., Islami, F., Scotti, L., Bellocco, R., Corrao, G., Boffetta, P. & La vecchia, C. Alcohol drinking and bladder cancer risk: a meta-analysis. *Annals of Oncology* **23**, 1586 (2012).

155. Aspirin use for the primary prevention of cardiovascular disease and colorectal cancer: U.S. Preventive Services Task Force Recommendation Statement. *Ann Intern Med* **164**, 836 (2016).

156. Ma, S., Han, T., Sun, C., Cheng, C., Zhang, H., Qu, G., Bhan, C., Yang, H., Guo, Z., Yan, Y., Cao, C., Ji, Z. & Zhou, Q. Does aspirin reduce the incidence, recurrence, and mortality of colorectal cancer? A meta-analysis of randomized clinical trials. *International Journal of Colorectal Disease* **36**, 1653 (2021).

157. Lin, S., Pan, Y. & Xu, C. Effects of aspirin on pancreatic cancer cells PANC-1 and its potential molecular mechanism. *Journal of B.U.ON. : official journal of the Balkan Union of Oncology* **25**, 2449 (2020).

158. Huang, T. B., Yan, Y., Guo, Z. F., Zhang, X. L., Liu, H., Geng, J., Yao, X. D. & Zheng, J. H. Aspirin use and the risk of prostate cancer: a meta-analysis of 24 epidemiologic studies. *International urology and nephrology* **46**, 1715 (2014).

159. Shi, C., Zhang, N., Feng, Y., Cao, J., Chen, X. & Liu, B. Aspirin Inhibits IKK-β-mediated Prostate Cancer Cell Invasion by Targeting Matrix Metalloproteinase-9 and Urokinase-Type Plasminogen Activator. *Cellular Physiology and Biochemistry* **41**, 1313 (2017).

160. Chang, E. T., Zheng, T., Weir, E. G., Borowitz, M., Mann, R. B., Spiegelman, D. & Meuller, N. E. Aspirin and the Risk of Hodgkin's Lymphoma in a Population-Based Case–Control Study. *JNCI: Journal of the National Cancer Institute* **96**, 305 (2004).

161. Chang, E. T., Frøslev, T., Sørensen, H. T. & Pedersen, L. A nationwide study of aspirin, other non-steroidal anti-inflammatory drugs, and Hodgkin lymphoma risk in Denmark. *British Journal of Cancer 2011 105:11* **105**, 1776 (2011).

162. Cerhan, J. R., Anderson, K. E., Janney, C. A., Vachon, C. M., Witzig, T. E. & Habermann, T. M. Association of aspirin and other non-steroidal anti-inflammatory drug use with incidence of non-hodgkin lymphoma. *International Journal of Cancer* **106**, 784 (2003).

163. Huang, W., Sundquist, K., Sundquist, J. & Ji, J. Use of dipyridamole is associated with lower risk of lymphoid neoplasms: a propensity score-matched cohort study. *British Journal of Haematology* **196**, 690 (2022).

164. Win, T. T., Aye, S. N., Fern, J. L. C. & Fei, C. O. Aspirin and Reducing Risk of Gastric Cancer: Systematic Review and Meta-Analysis of the Observational Studies. *Journal of gastrointestinal and liver diseases : JGLD* **29**, 191 (2020).

165. Sun, D., Liu, H., Dai, X., Zheng, X., Yan, J., Wei, R., Fu, X., Huang, M., Shen, A., Huang, X., Ding, J. & Geng, M. Aspirin disrupts the mTOR-Raptor complex and potentiates the anti-cancer activities of sorafenib via mTORC1 inhibition. *Cancer Letters* **406**, 105 (2017).

166. Zhong, S., Chen, L., Zhang, X., Yu, D., Tang, J. & Zhao, J. Aspirin use and risk of breast cancer: Systematic review and meta-analysis of observational studies. *Cancer Epidemiology Biomarkers and Prevention* **24**, 1645 (2015).

167. Henry, W. S., Laszewski, T., Tsang, T., Beca, F., Beck, A. H., McAllister, S. S. & Toker, A. Aspirin suppresses growth in PI3K-mutant breast cancer by activating AMPK and inhibiting mTORC1 signaling. *Cancer Research* **77**, 790 (2017).

168. Booth, H. P., Prevost, A. T. & Gulliford, M. C. Validity of smoking prevalence estimates from primary care electronic health records compared with national population survey data for England, 2007 to 2011 †. **20**, 1357 (2013).

169. Ramos, R., Balló, E., Marrugat, J., Elosua, R., Sala, J., Grau, M., Vila, J., Bolíbar, B., García-Gil, M., Martí, R., Fina, F., Hermosilla, E., Rosell, M., Muñ, M. A., Prieto-Alhambra, D. & Quesada, M. Validity for Use in Research on Vascular Diseases of the SIDIAP (Information System for the Development of Research in Primary Care): the EMMA Study. *Rev Esp Cardiol* **65**, 29 (2012).

170. Bolíbar, B., Fina Avilés, F., Morros, R., Del Mar Garcia-Gil, M., Hermosilla, E., Ramos, R., Rosell, M., Rodríguez, J., Medina, M., Calero, S. & Prieto-Alhambra, D. Base de datos SIDIAP: la historia clínica informatizada de Atención Primaria como fuente de información para la investigación epidemiológica. *Medicina Clínica* **138**, 617 (2012).

171. Coma, E., Ferran, M., Méndez, L., Iglesias, B., Fina, F. & Medina, M. Creation of a synthetic indicator of quality of care as a clinical management standard in primary care. *SpringerPlus* **2**, 1 (2013).

172. Prieto-Bonete, G., Pérez-Cárceles, M. D., Maurandi-López, A., Pérez-Martínez, C. & Luna, A. Association between protein profile and post-mortem interval in human bone remains. *Journal of Proteomics* **192**, 54 (2019).

173. Samadoulougou, S., Idzerda, L., Letarte, L., McKay, R., Quesnel-Vallée, A. & Lebel, A. Self-perceived health status among adults with obesity in Quebec: a cluster analysis. *Annals of Epidemiology* **67**, 43 (2022).

174. CC, C. & YC, C. Advanced Machine Learning in Prediction of Second Primary Cancer in Colorectal Cancer. *Studies in health technology and informatics* **270**, 1191 (2020).

175. Cardoso, R., Guo, F., Heisser, T., Hackl, M., Ihle, P., De Schutter, H. & et al. Colorectal cancer incidence, mortality, and stage distribution in European countries in the colorectal cancer screening era: an international population-based study. *The Lancet Oncology* **22**, 1002 (2021).

176. Aleksandrova, K., Pischon, T., Jenab, M., Bueno-de-Mesquita, H. B., Fedirko, V., Norat, T., Romaguera, D. & et al. Combined impact of healthy lifestyle factors on colorectal cancer: A large European cohort study. *BMC Medicine* **12**, 1 (2014).

177. Serrano, D., Patrignani, P., Stigliano, V., Turchetti, D., Sciallero, S., Roviello, F., D'Arpino, A. & et al. Aspirin Colorectal Cancer Prevention in Lynch Syndrome: Recommendations in the Era of Precision Medicine. *Genes 2022, Vol. 13, Page 460* **13**, 460 (2022).