# MACHINE LEARNING AND ALGORITHMS APPLIED TO ETHNOGRAPHIC AND BIOMEDICAL CANCER DATA:

STUDIES FROM IRELAND, FINLAND, AND SPAIN

**BY ORNELA BARDHI**

**2021**

UNIVERSIDAD DE DEUSTO

# MACHINE LEARNING AND ALGORITHMS APPLIED TO ETHNOGRAPHIC AND BIOMEDICAL CANCER DATA: STUDIES FROM IRELAND, FINLAND, AND SPAIN

Doctoral thesis by

ORNELA BARDHI

Director

BEGOÑA GARCÍA-ZAPIRAIN SOTO

Bilbao, July 2021

UNIVERSIDAD DE DEUSTO

# MACHINE LEARNING AND ALGORITHMS APPLIED TO ETHNOGRAPHIC AND BIOMEDICAL CANCER DATA: STUDIES FROM IRELAND, FINLAND, AND SPAIN

Doctoral thesis by: ORNELA BARDHI

Industrial Ph.D. between

University of Deusto, Spain

Beacon Hospital, Republic of Ireland and

Success Clinic Oy, Finland

Director:

BEGOÑA GARCÍA-ZAPIRAIN SOTO

Doctoral student Director

Bilbao, July 2021

# DECLARATION

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements.

Ornela Bardhi

July 2021

# ABSTRACT

Technology has seen an increased presence in the healthcare field for many years now. The last decade especially has seen a boom due to the progress of machine learning techniques and algorithms as well as the digitalization of healthcare records. These records are of different formats, such as text data, images, video, etc. and each requires specific ways to preprocess and analyze it. This thesis tackles important health issues faced in our society through ethnographic and biomedical data analysis using statistical analysis, machine learning and deep learning. The thesis is comprised of three case studies conducted in Ireland, Finland, and Spain, and each follows a different methodology and analysis approach.

The first study deals with care pathways, their implementation in the last 20 years around the world, and the Beacon Hospital study. Understanding what factors influence care pathways allow a more person-centered care approach and the redesign of care processes. Four main tasks have been achieved in this study: a literature review of cancer care pathway implementation, an ethnographic study with breast and prostate cancer patients at Beacon Hospital about their perspective on care pathways, creation of two datasets with information coming from electronic health records and one-on-one interviews, and an analysis of the data through statistical analysis to identify the factors influencing care pathways for these two cancer diseases in a hospital setting.

The second study is about the use of electronic health records to predict cancer patient survivability employing various machine learning algorithms. A collaboration with a regional hospital in Finland helped to achieve this task. Two steps were taken to predict survivability. The first one was to select the most relevant variables through various feature selection algorithms, and the second one was to perform survival prediction using nine machine learning algorithms.

The third and the last study is about colorectal polyps detection using deep learning to prevent colorectal cancer from forming or progressing. The tasks performed to complete it follow a comprehensive review of the published scientific research related to colorectal polyp detection, classification, segmentation, localization, and the implementation of combined convolutional neural networks and autoencoders model to detect colorectal polyps without image preprocessing. All three case studies are accepted for publication in high-impact journals; two are already published online, one is currently in press.

**Keywords:** care pathways; EHRs; medical images; breast cancer; prostate cancer; colorectal cancer

# ACKNOWLEDGMENT

There are not enough words to express my gratitude to the many people and institutions who have supported me during my Ph.D. fellowship.

First and foremost, I would like to express my deepest gratitude to my academic supervisor Prof. Begonña Garcia Zapirain, for her support, sharing her expertise, and fruitful discussions. Without your encouragement, this thesis work would not be possible. I would like to thank Roberto Nuño Solinis, Ailish Daly, Prof. Brian Caulfield for being part of my doctoral studies panel and for their valuable guidance during my studies.

I would like to thank the CATCH team, Louise, Gillian, Merce, Francisco, Dominic, Juanmi, Gabriel, Threase, Jennifer, Estefania, Paivi, and Luis.

Thank you to the University of Louisville team, Daniel, Sofia, Pegah, and especially Prof. Adel Elmaghraby for accepting me to conduct part of my research within the Department of Computer Science and Engineering.

An immense appreciation goes to the 83 cancer patients who participated and everyone who considered participating in the Beacon Hospital study. I would like to acknowledge the support from the Beacon Hospital medical and nonmedical staff involved directly and indirectly in the study, consultants Dr. Jennifer Westrup, Dr. Reem Salman, Dr. Pierre Thirion, Mr. Terence Boyle, Dr. Ray McDermott, Dr. Alina Mihai, Prof. John G. Armstrong; the breast care nurses, Denise and Hazel; the liaison nurse, Jennifer, all nurses in the daycare unit, Emma, Audrey, Mary, Aisling, Deidre, Ciara, Annette, Martina, Sandra, Sinead, Noelle, and Ilva: all nurses in the inward unit; the brilliant gatekeepers, Sarah Wright and Annmarie Breen; the UCD Beacon Hospital Academy team, Ruth Whelan and Kathryn Kane; the IT department staff, Ricardo Pereira.

I would like to thank the Success Clinic team, Mika Teikari, for welcoming me, Janne Eho for guiding and supporting me throughout the process of accessing the database, and the rest of the team, Antti, Ilona, Pekka, Annukka, Mikko, Juha, and Esa.

From a personal perspective, I am lucky to also have a long list of people to thank. To my dad, mum, siblings, and niblings, thank you for listening to my struggles, supporting, and believing in me throughout this research towards my dissertation. Jeni të mrekullueshëm, ju dua pafund!

My friends have been another fundamental pillar: You listened when I was worried; you were patient when I was stressed and traveling from one country to the other, and above all, you pushed me to keep going! I feel blessed to have you in my life. Thank you to my beloved friends in Bilbao, Cristian, Toñy, Marta, Liria, Diana, Luana, Elena, Osane, Elena U., Irene, Erik, Carmen, Sofia; my friends in Dublin, Amal, Ray, Patricia, Elnatan, Sarah, Nuria; my PERCCOM friends in Finland Anar, Shola, Victoria; and many more not mentioned here who have a place in my thoughts and my heart.

Lastly, I would like to express my gratitude to the European Commission and the Marie Curie Action for the financial support for the doctoral studies and Mika Teikari at Success Clinic Oy for the generous financial support for purchasing the AURIA database.

Ornela Bardhi

July 2021
Bilbao, Spain

*This thesis is dedicated to my family,*

*for all their love and unconditional support.*

# TABLE OF CONTENTS

# LIST OF FIGURES

# LIST OF TABLES

# 1

# INTRODUCTION

The World Health Organization has estimated that by 2040 the number of incident cases for all cancers will be 28.9 million, an increase of approximately 11.3 million new cases (IARC, 2021). For the period 2018 to 2040, the number of incident cases is estimated to increase by 764,052 new cases for breast cancer (BCa) (female) and by 821,309 for prostate cancer (PCa) (IARC WHO, 2021).

The implementation of guidelines and care pathways has demonstrated improvements in the management of patients and healthcare professionals (Akechi et al., 2015), cost-effective for the hospital and the patients (Akechi et al., 2015; Thorsen et al., 2011), and better satisfied and educated patients (Chen et al., 2000; Munir et al., 2011; Tamburini et al., 2003).

Digital transformation of the healthcare system has brought many changes to the way patient data is stored and used. Patient's medical information is part of that digitalization process, the electronic healthcare records (EHRs). Information stored in such systems varies from simple ones with only demographic information to complex and holistic ones with all types of treatments and examinations patients undergo throughout the course of their care. The latter includes different data types such as text data, images, video, etc., and each of them requires specific ways to preprocess and analyze it. This dissertation deals with EHRs, patient-reported outcomes and medical imaging.

## 1.1 Research hypotheses and objectives

The analysis of the problem introduced in the previous section leads to the following hypotheses.

- *Understanding the past and current state of cancer care pathways implementations using ethnographic analysis can be used to find the main factors influencing care pathways in a hospital setting.*
- *Cancer patient survivability can be predicted using patients' electronic health records and various machine learning algorithms.*
- *Colorectal cancer can be prevented by detecting early colorectal polyps using deep learning algorithms.*

Based on the hypotheses stated above, this dissertation emphasizes the power of qualitative and quantitative analysis to tackle different medical problems. In order to work on each hypothesis, the research was divided into four stages, as presented in *Figure 1*. Each stage has its research question/s (RQ), which is/ are answered by the following specific objectives (SO) divided into:

**Care pathways and ethnographic analysis:**

- SO1: Define the current state of the art of the care pathways implementation. This objective is fulfilled by a comprehensive review of the published scientific research in the last 20 years, taking into consideration all care pathways phases and focusing mainly on breast, prostate, and colorectal cancer diseases. SO1 answers RQ1.

- SO2: Understand the current state of care pathways in a hospital setting. This objective is achieved by running a study at Beacon Hospital with breast and prostate cancer patients using qualitative methods. This objective aims to understand the patients' perspective on their care pathways journey.

- SO3: Construct a database containing information from patients' care pathways. This objective is fulfilled by combining the electronic health records (EHRs) and interview data from the SO2 study. The objective aims to identify the factors influencing care pathways for breast and prostate cancer in a hospital setting. SO2 and SO3 answer RQ2.

**Electronic healthcare records (EHRs) analysis:**

- SO4: Construct a more extensive database containing breast and prostate cancer patients' EHRs information. This objective was achieved by requesting access to a regional hospital EHRs data in Finland. The requested EHRs should at least have the information for the variables retrieved from the SO3 study.

- SO5: Design and implement algorithms for EHRs data analysis. The objective is fulfilled by implementing various algorithms using the Python programming language and its machine learning libraries. The objective aims to determine the variables that most affect patient survivability and explore machine learning algorithms to assist in survivability predictions. SO4 and SO5 answer RQ3 and RQ4.

**Colorectal polyp detection using colonoscopy images:**

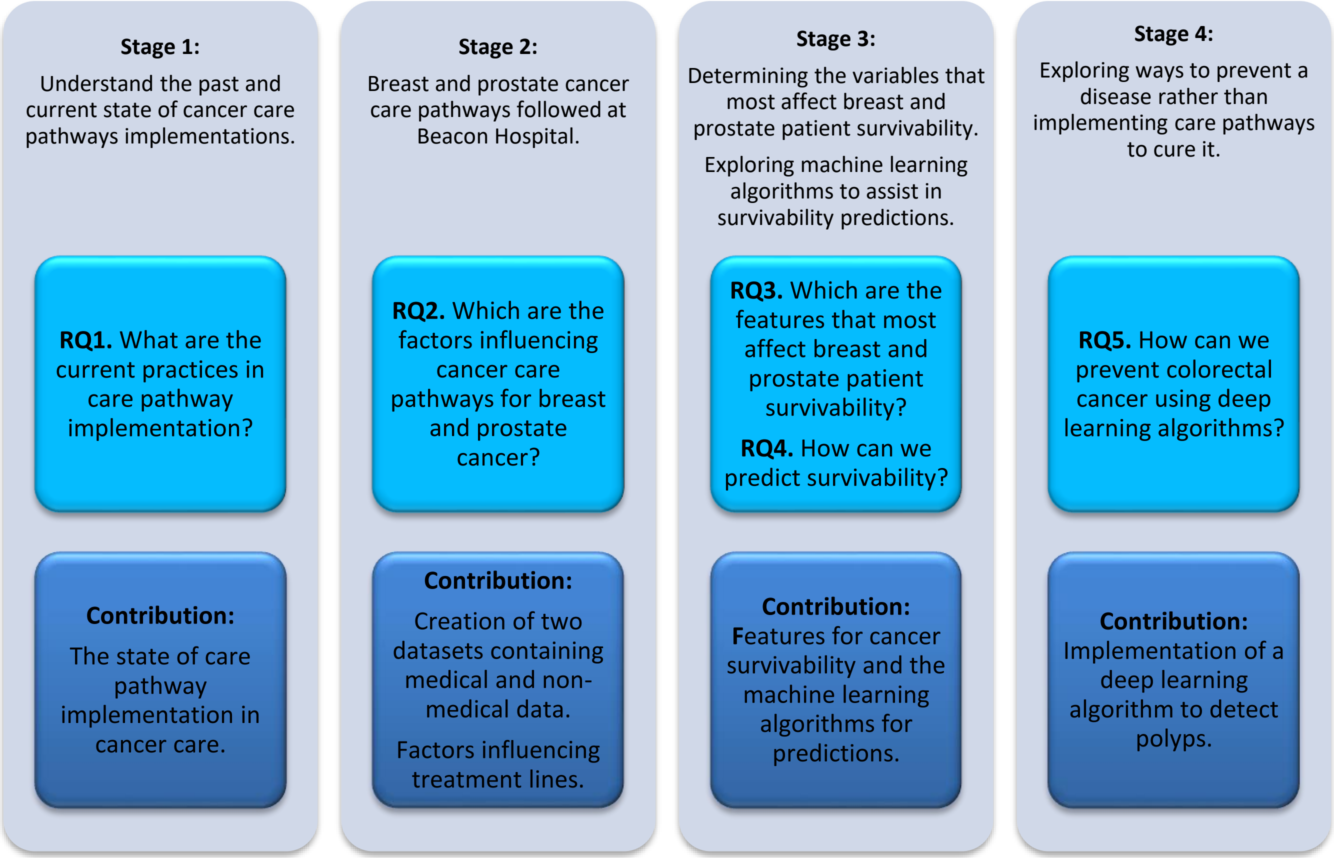
- SO6: Present the current state of the art of implementation of deep learning algorithms for colorectal polyps. This objective is fulfilled by a comprehensive review of the published scientific research regarding colorectal polyp detection, classification, segmentation, localization in both image and video.

- SO7: Design a deep learning architecture based on a combination of convolutional neural networks and autoencoders to detect colorectal polyps without image preprocessing, which outperforms the current state-of-the-art contributions. Both SO6 and SO7 answer RQ5.

In addition to the research objectives mentioned above, the following aims have been included to contribute to the research community and the general public.

- Maximize the scientific contribution of this dissertation with the publication of several articles in scientifically renowned conferences and journals relevant to the topic and science communication outreaches so that scientific contributions could be accessible to the public.

- Maximize the clarity and reproducibility of the different methodologies and data processing algorithms employed. This will allow future developers and researchers to implement, improve and/or replicate all the research questions tackled throughout this document.



**Stage 1:**

Understand the past and current state of cancer care pathways implementations.

**RQ1.** What are the current practices in care pathway implementation?

**Contribution:**

The state of care pathway implementation in cancer care.

**Stage 2:**

Breast and prostate cancer care pathways followed at Beacon Hospital.

**RQ2.** Which are the factors influencing cancer care pathways for breast and prostate cancer?

**Contribution:**

Creation of two datasets containing medical and non-medical data.

Factors influencing treatment lines.

**Stage 3:**

Determining the variables that most affect breast and prostate patient survivability.

Exploring machine learning algorithms to assist in survivability predictions.

**RQ3.** Which are the features that most affect breast and prostate patient survivability?

**RQ4.** How can we predict survivability?

**Contribution:**

Features for cancer survivability and the machine learning algorithms for predictions.

**Stage 4:**

Exploring ways to prevent a disease rather than implementing care pathways to cure it.

**RQ5.** How can we prevent colorectal cancer using deep learning algorithms?

**Contribution:**

Implementation of a deep learning algorithm to detect polyps.

***Figure 1.*** Research Approach Overview.

## 1.2 Social impact

This dissertation is comprised of three major studies, each one tackling different medical problems. The first study, factors influencing care pathways for breast and prostate cancer in a hospital setting – Beacon Hospital case study, deals with the understanding of the implementation of care pathways in hospitals. It is imperative to have the patient's perspective when it comes to implementing projects in any healthcare setting. They bring new insights that healthcare professionals might not be aware of. In this study, the prominent voice is the patients' one through one-on-one interviews. For a 360-degree view, their EHRs are collected as well. The impact of this study is two folds; besides patients, the hospital and the hospital staff profit too. Through analysis of such data, the hospital allocates the resources better, including the staff, and in return, the patients receive a more personalized treatment.

The second study is the large-scale version of the first one, which means the use of machine learning algorithms is better fitted to analyze the data faster. This study shows that not all variables are decisive when predicting breast or prostate cancer patient survivability. By narrowing down the input variables, healthcare professionals are able to focus on the issues that most impact patients and hence devise better, more individualized care plans.

Medical information is stored in various formats. Another form of medical data is medical images which have gained immense importance with the advance of medical imaging devices such as CTs, MRI, ultrasound, etc. It is used for diagnosing diseases, planning treatments, and assessing results. Moreover, medical imaging is used in preventing illness. The third study presented in this dissertation tackles precisely this issue. By automating some part of the diagnosis, in this case detecting colorectal polyps using deep learning algorithms, professional radiologists will be more efficient in their job, less polyp miss rates, and deliver results quicker.

## 1.3 Research methodology

This section explains the methodology followed in conducting this dissertation. This dissertation is presented on the basis of three different studies that vary in nature and in the design and methods used. As such, each study has its own background (related research) section and also its materials and methods sections. However, for each study the steps described below are followed:

- *Literature review*: The main objective of this step is to analyze and understand the current state of the art of cancer care pathway implementation, EHRs analysis using machine learning, and deep learning model for colorectal polyps. To perform this task, the most relevant literature will be selected among the available publications in the scientific medium, such as national and international journals, and conference communications and proceedings will be reviewed. The knowledge obtained during this stage will lead to the formulation of the hypothesis and the generation of articles with the available expertise and future recommendations.

- *Design and Development:* After the literature analysis and the processing of the knowledge acquired in the previous step, this stage will lead to the definition, design, and development of each study. This stage will define the methods each study will use, such as qualitative, quantitative, or mixed methods approaches. The completion of this stage will lead to the final assessment of the established hypothesis.

- *Experiments and Evaluation:* This stage aims to test the different iterations obtained after the design and development step. This section will lead to the completion of the three studies: an ethnographic study with cancer patients about their care pathways experience, an EHR study using similar data as collected in the qualitative study but analyzed through machine learning algorithms, and the third and the last study relate to detection of colorectal polyps using deep learning models. All the knowledge used during these steps will be backed up by the concepts acquired during the first stages of the proposed methodology.

- *Data Analysis and Final Results:* This step aims at comparing the obtained results with the state-of-the-art, which leads to the final assessment of the established hypotheses.

## 1.4 Contributions

The ultimate aim of this work is to analyze various medical data using different methods and techniques or a combination of them if and when needed. The thesis is based on the following published and in press publications; however, the dissertation is not presented in the form of a compendium. The candidate's individual contributions to each publication are also given below.

**Journal Publication I:** O. Bardhi and B. Garcia-Zapirain, "Machine learning techniques applied to electronic healthcare records to predict cancer patient survivability," *Computers, Materials & Continua*, vol. 68, no.2, pp. 1595–1613, 2021. (Impact Factor: 4.89; Q1)

Bardhi applied for permission to get the data, conducted the data cleaning and pre-processing, the data analysis, visualization, and led the writing of the publication.

**Journal Publication II:** O. Bardhi, D. Sierra-Sosa, B. Garcia-Zapirain, and Luis Bujanda, "Deep Learning Models for Colorectal Polyps," *Information*, vol.12, no.6, pp. 245, 2021. (Impact Factor: 3.0; Q2)

Bardhi was part of the conceptualization, methodology, software, and analysis stages. She conducted the review, visualizations and wrote the first draft of the manuscript.

**Journal Publication III:** O. Bardhi, B. Garcia-Zapirain, and R. Nuno-Solinis, "Factors influencing care pathways for breast and prostate cancer in a hospital setting," *International Journal of Environmental Research and Public Health*, vol.18, no.15, pp.7913, 2021. (Impact Factor: 3.39; Q1)

Bardhi designed the study, conducted the interviews, collected the data, and created the datasets. Bardhi was also responsible for the data analysis, data visualization and led the writing of the article.

**Conference Publication I:** O. Bardhi, D. Sierra-Sosa, B. Garcia-Zapirain, and A. Elmaghraby, "Automatic colon polyp detection using Convolutional encoder-decoder model," *2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT)*, pp. 445-448, 2017.

Bardhi was part of the conceptualization, methodology, software, analysis, and visualization stages. She wrote the first draft of the manuscript.

## 1.5 Thesis Outline

This section outlines the structure and content of the different chapters that are part of this dissertation. Chapter 3 – 5 have each their own specific materials, methods, results, and discussions/ conclusions. Each of these chapters presents a study, and each study uses different material and methods, which can be confusing for the reader if described together in a single chapter.

**Chapter 1 – Introduction:** This dissertation begins by explaining the need for this study, positioning the thesis within the appropriate theoretical background, and elaborating on the research gaps, objectives, and research questions.

**Chapter 2 – Background**: This chapter gives an overview of electronic healthcare records, international guidelines, and care pathways and the digitalization of them, medical data analysis, specifically machine learning and deep learning algorithms used throughout this dissertation. A workflow of data training, validation, and evaluation is also presented.

**Chapter 3 – Study 1**: Factors influencing care pathways for breast and prostate cancer in a hospital setting: the Beacon Hospital case study. The chapter starts with a deep dive on care pathway implementation in the last 20 years, followed by the qualitative study conducted at the Beacon Hospital. All the relevant information regarding methods, participants' sample, the duration of the study, the analysis of the data, the results, discussions, and future work will be detailed in this chapter. Part of this chapter is accepted for publication in a Q1 journal and is currently in press.

**Chapter 4 – Study 2**: Machine learning techniques applied to electronic healthcare records to predict cancer patient survivability. The fourth chapter covers the second study conducted as part of this dissertation. The reader can find information regarding the database, the data selection, preparation, and analysis, followed by a description of the machine learning algorithms used for feature selection and later for survivability prediction. Results, discussion, limitations, conclusions, and future work are also described. This chapter is already published in a Q1 journal.

**Chapter 5 – Study 3:** Deep learning models for Colorectal Polyps. Although this study is not strictly related to care pathways, it is however linked with the application of machine learning and deep learning techniques in medicine and with the idea that if we prevent a disease from occurring, we might not need care pathways at all (at least for the treatment, follow-up, or palliative phases). We will look at the current state-of-the-art in colorectal deep learning model implementations. Afterward, a novel convolutional neural network will be presented together with the results, discussions, and future work. Similar to the previous chapter, this one is also already published in a Q2 journal.

**Chapter 6 – Conclusions:** The sixth and final chapter of this dissertation introduces the different thoughts and conclusions extracted from the final evaluation of the research work presented. This chapter will cover the specific objectives introduced in section 1.1 and whether they were successfully met during the process. Future lines will be discussed in this section.

# 2

# BACKGROUND

This chapter gives an overview of the topics discussed in this thesis. A deep dive on the literature review of each topic is presented in their respective study, chapters 3 to 5.

## 2.1 From Paper to Electronic

The documentation of a patient’s medical history and care is usually described in various terms, such as medical records, medical charts, and health records. Traditionally, medical records were written on paper; however, they gained widespread usage during 1900-1920. These records were maintained in folders that were divided into sections based on the type of note, and only one copy was available. The advent of new technologies such as computers in the 1960s and 1970s laid the foundation for the development of the Electronic Health Record (EHR) (Evans, 2016).

The systems behind this complex ecosystem of data did not emerge overnight. It took many years, and this period can be divided into 4 main eras: the 1960s: problem-oriented medical records (POMR), the 1970s: the dawn of the EHR system, the 1990s: the internet's effects on EHR, the 2000s: EHR standardization and adaptation (ICANotes, 2019).

POMRs were developed by Dr. Lawrence Weed in 1968 in response to the stream-of-consciousness-style record-keeping that was commonplace at the time (ICANotes, 2019). It has five main facets: the database, complete problem list, initial planning, daily progress notes, and discharge summary. Although POMR is rigorous and still being used by some medical and behavioral health professionals, it is seen as lengthy and burdensome (ICANotes, 2019).

In the 1970s USA, different types of EHR begin to be developed by academic medical centers, the government, and industry (Virtual Mentor, 2011). The first EHR systems were known as clinical information systems. One of the first EHR systems was Computer Stored Ambulatory Record (COSTAR) at Massachusetts General Hospital, developed in collaboration with Harvard University (Virtual Mentor, 2011). In the early 1980s, EHRs gained widespread recognition. That is when organizations begin to form to tackle the broader issues and create industry-wide standards (Virtual Mentor, 2011).

With the advent of the internet, the EHRs were viewed as a clinician's assistive technology, rather than simply the digitalization of paper records (ICANotes, 2019). They become more powerful and affordable. It was during this period that the systems begin to be deployed in the cloud and the data shared with patients as well (ICANotes, 2019). To tackle the growing tide of EHR systems and patient information protection, the Health Insurance Portability and Accountability Act (HIPAA) was introduced, which laid the structural foundation for the next wave of EHR evolution and EHR standardization (ICANotes, 2019).

The 2000s are the generation of secure EHRs that are still in place today (ICANotes, 2019). The most used safeguards include automatic data backups and logoffs, data encryption, audit trails, and access control. It was during this decade when EHRs became the all-in-one systems as today and partnered with Health Lever 7 International (HL7) (ICANotes, 2019).

Towards the end of 2000s, the adoption of a basic EHRs by hospitals in the US varied from 9% to 73% according to (Everson et al., 2020) that used six different methods to calculate it. Today, the adoption rate in the USA has reached 90% in hospital clinics and 80% in independent clinics (ICANotes, 2019). According to the 2016 OECD survey filled by 15 EU countries, the proportion of primary care practices using EHRs was on average around 80% (OECD/European Union (2018)). A 2021 study (Liang et al., 2021) of adoption rate of EHRs in China and USA over the 2007 and 2018 period showed that the annual average adoption rates in Chinese hospitals was 6.1% an increase from 18.6% to 85.3%, and in USA hospitals 9.6% an increase from 9.4% to 96%.

Some of the main benefits of EHRs system adaptations are excellent continuity of care, improved efficiency between medical professional's communication, but also with diagnostic centers, pharmacies, insurance providers, etc., and better emergency preparedness and response (ICANotes, 2019). Besides providers' benefits, the EHR can improve patient care by improving the accuracy and clarity of medical records by reducing the incidence of medical errors (Centre for Medicare and Medicaid Services, 2021). Other benefits are the availability of health information, reduction of the duplicated tests and delays in treatment, and the patients are well informed to make better decisions (Centre for Medicare and Medicaid Services, 2021).

## 2.2 International Guidelines and Care Pathways

There are several organizations, governmental and non-governmental, that work on implementing guidelines and care pathways and making them publicly accessible worldwide. These guidelines and care pathways aid providers and health IT implementers, implement proper EHR systems. Below are presented some of them.

### 2.2.1 ESMO Guidelines

European Society for Medical Oncology (ESMO) is a European medical oncology organization founded in 1975. It currently has more than 25,000 oncology professionals representing over 160 countries worldwide (ESMO, 2021). The core mission of ESMO is to improve the quality of cancer care, starting from prevention and diagnosis to palliative care and patient follow-up. It fulfills the aim by educating doctors, cancer patients, and the public on the best practices and latest advances in oncology and by promoting access to excellent cancer care for all patients. Optimal care is achieved through the development of integrated cancer care, supporting the professional development of oncologists within the multidisciplinary team, and advocating for sustainable cancer care worldwide.

The main body responsible for the ESMO Clinical Practice Guidelines and Consensus Statements and their update is the ESMO Guidelines Committee (GLC). The GCL follows strict procedures to produce high-quality and well-formulated guidelines with clear instructions. The methodology to create these guidelines is made available for free on the ESMO website.

### 2.2.2 NICE Pathways

NICE stands for National Institute for Health and Care Excellence. The NICE was established in the UK in 1999 to end the unequal treatment dependent on the National Health System (NHS) health authority area (postcode) in which the patient happened to live (NICE, 2021). It has since become a role model internationally for the development of clinical guidelines. One aspect of this is the explicit determination of cost-benefit boundaries for certain technologies that it assesses. NICE also plays a vital role in pioneering technology assessment in other healthcare systems through NICE International, established in May 2008 to help cultivate links with foreign governments.

The role of NICE is to improve outcomes for people using public health and social care services. They fulfill this role by producing guidance and advice for health based on evidence evaluations of efficacy, safety, and cost-effectiveness in various circumstances; developing quality standards and performance metrics for those providing and commissioning health; and providing a range of information services for commissioners, practitioners, and managers across the healthcare sector. The NICE guidance and advice are presented in an integrated view in topic-based interactive flowcharts.

### 2.2.3 NCCN Guidelines

The National Comprehensive Cancer Network (NCCN) is a not-for-profit alliance of 31 cancer centers in the United States of America (NCCN, 2021). The primary purpose is to improve and facilitate quality, effective, efficient, and accessible care for cancer patients. NCCN develops resources for the various stakeholders in the health care delivery system. The primary resources are the NCCN Guidelines appropriate for use by clinicians, patients, and other health care decision-makers in the USA and around the world. These guidelines are grouped according to treatment by cancer type, detection, prevention and risk reduction, supportive care, a specific population, and guidelines for patients. They are updated periodically, and each guideline is posted with the latest update date and version number.

Apart from the USA member organizations, NCCN collaborates with other global organizations. Together they have created the Regional Adaptations of the NCCN Guidelines. These guidelines are translated into multiple languages and adapted according to the local accessibility, consideration of metabolic differences in populations, and regulatory status of health care technologies used in cancer care in the specified country or region.

### 2.2.4 Digitized Guidelines and Care Pathways

The World Health Organization (WHO) is one of the central bodies to develop global guidelines to ensure evidence-based medicine is followed in every country. According to WHO, a guideline is defined as any information product developed by WHO that contains recommendations for clinical practice or public health policy (WHO Guidelines, 2021). These guidelines are created by the Guideline Review Committee by following rigorous methodology and a transparent and evidence-based decision-making process. New and updated guidelines are being approved by the committee continuously.

Technology has a great potential to advance the adoption of guidelines and care pathways. However, digital systems and local policies make it difficult to adapt them. The Standards-based, Machine-readable, Adaptive, Requirements-based, and Testable (SMART) Guidelines are the new approach to systematize and accelerate the consistent application of interventions by WHO in the digital age (Mehl et al., 2021). They are developed with the idea to strengthen the quality of care and accelerate the progress towards national and Sustainable Development Goals.

The SMART Guidelines approach is divided into five ‘knowledge layers’, providing a systematic, transparent, and testable structure for countries to work through, even if they are not fully digital yet, *Figure 2*. The five layers comprise documentation, procedures, and health components to drive guideline implementation through digital systems.

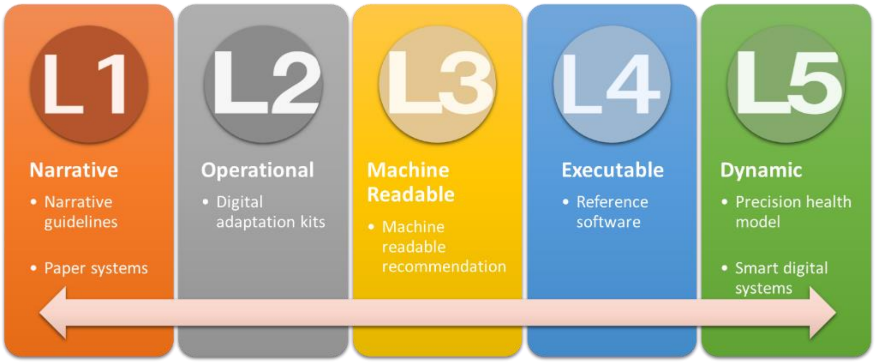


***Figure 2.*** The SMART Guidelines five knowledge layers.

Each layer informs a specific group regarding a particular part of the guideline, e.g., guideline developers on how to translate recommendations into standards; technology professionals on how to integrate recommendations into updatable digital systems that are software-neutral; and countries on how to make these digital systems local, standardized, interoperable, and updatable, consistent with evidence-based recommendations. However, SMART Guidelines are not a standalone solution. Good planning and governance on digital health by all stakeholders is needed when working to integrate digital approaches into health systems.

## 2.3 Analyzing Medical Digital Data

Digitalized medical information has made it easier for researchers and the industry to put to the test new techniques and algorithms in data analysis (Amoon et al., 2020). Many studies have been carried out using patients reported outcomes, EHRs, or medical images. Initially, discriminative features were manually designed and extracted for the classification and detection of abnormalities and segmentation of regions of interest in different medical applications (Xu et al., 2020). This step required the expertise of expert physicians. However, due to data complexity and the limited data interpretation knowledge, machine and deep learning caught the attention of researchers. One of the most potent advantages was the fact that no feature selection was needed to reach the final goal. Deep learning models are composed of multiple processing layers to learn representations of the data with various levels of abstraction. These methods discover complicated structures in large data sets. Thus, they have dramatically improved the state-of-the-art in many fields of machine learning. Some of the most used algorithms in machine learning are logistic regression, support vector machines, random forest, etc. Some of the most used deep learning algorithms are convolutional neural networks, reinforcement learning, natural language processing, etc.

The following sections present the widely used machine and deep learning architectures used in the literature to tackle the aforementioned applications. As for the specific study cases which are addressed in this dissertation, their literature review is presented in their respective chapters, 4 and 5.

### 2.3.1 Machine Learning Algorithms

*Logistic regression*

Logistic regression classifies data by using maximum likelihood functions to predict the probabilities of outcome classes (Lorena et al., 2011) such as alive/dead, healthy/sick, etc. LRs are widely used because they are simple and explicable. In order to model nonlinear relationships between variables with logistic regression, the relationships must be found prior to training or various transformations of variables performed (Tu, 1996).

*Support Vector Machines*

Support vector machines were first introduced by Cortes & Vapnik (Cortes and Vapnik, 1995). Their objective is to find a hyperplane in the N number feature space that maximizes the distance between points corresponding to training dataset subjects in the output classes (Miguel-Hurtado et al., 2016). SVMs are generalizable to different datasets and work well with high-dimensional data (Lorena et al., 2011), and can accurately perform linear and nonlinear classification. Nonlinear classification is performed using the kernel, which maps inputs into high-dimensional feature spaces. However, SVMs require a lot of parameter tuning (Lorena et al., 2011; Pedregosa et al., 2011; Stark et al., 2019).

*Nearest Neighbor*

Nearest neighbor algorithms work by finding a preset number of training samples that are closest in distance to the new point, and later predict the labels (Cover and Hart, 1967). In k-nearest neighbor (KNN) learning, the number of samples is a user-defined constant. By contrast, in radius-based neighbor learning, the constant varies depending on the local density of points (Pedregosa et al., 2011). Despite their simplicity, nearest neighbors have been successful in many classification and regression problems. As a non-parametric method, it often manages to classify situations where the decision boundary is highly irregular.

*Naive Bayes*

Naive Bayes models, unlike the previously described classifiers, are probabilistic classifiers (Lorena et al., 2011) based on the Bayes theorem. NB models generally require less training data and have fewer parameters compared to other models such as SVMs etc. (Al-Aidaroos et al., 2010). NB models are good at disregarding noise or irrelevant inputs (Al-Aidaroos et al., 2010). However, they consider that the input variables are independent, which is not valid for most classification applications (Lorena et al., 2011). Despite this assumption, these models have been successful in many complex problems (Lorena et al., 2011).

*Decision Trees*

Decision trees organize knowledge extracted from data in a recursive hierarchical structure composed of nodes and branches (Quinlan, 1986) [20]. DTs are non-parametric, supervised learning methods used for both classification and regression, whose goal is to create a model that predicts the value of a target feature by learning simple rules inferred from the input features. Besides nodes and branches, DTs are made up of leaves, the last nodes being found at the bottom of the tree (Miguel-Hurtado et al., 2016). Some advantages of DTs are that they are simple to understand and interpret (trees can be visualized), require scarce data preparation (no data normalization is needed), can handle both numerical and categorical data, and the model can be validated by using statistical tests (Pedregosa et al., 2011). Besides all these positive aspects of DTs, particular care should be taken when working with them as over-complex trees can be created that are poorly generalized (Pedregosa et al., 2011). DTs can also be unstable when introducing small variations into data, which can be mitigated by using them within an ensemble (Pedregosa et al., 2011).

*Random Forest*

Random forest is a meta model that fits various decision tree classifiers into a number of sub-samples on the dataset. RF uses averaging to improve predictive accuracy and control overfitting. The sub-sample size is controlled by the max_sample parameter when the bootstrap is set to True (default); otherwise, each tree uses the whole dataset (Pedregosa et al., 2011). Individual DTs generally tend to have high variance and overfit. RFs yield DTs and take an average of the predictions, which leads to some errors being canceled out. RFs achieve reduced variance by combining diverse trees, sometimes to the detriment of a slight increase in bias. In practice, variance reduction is often significant, hence yielding a better overall model.

### 2.3.2 Deep Learning Algorithms

There are several architectures that have been used in medical data analysis:

*Convolutional Neural Networks (CNNs)*

Convolutional neural networks are the most common deep learning architectures used for a diverse set of data analysis problems. However, they are more often utilized for classification and computer vision tasks such as image segmentation and patterns recognition by leveraging principles from linear algebra, specifically matrix multiplication, to identify patterns within an image.

CNNs have three main types of layers. The first and the primary layer is the convolutional layer, where the majority of computation occurs. Its required components are input data, filter, and feature map. This layer is based on convolving an input with kernels to obtain feature maps. As the filter moves along the input, it uses the same parameters for the convolution, *Figure 3.* The featured map that is formed is characterized by a specific pattern. In this sense, CNN is able to recognize distinct patterns and is robust to distortions and geometric transformations. The filter size defines the size of the subregions being convolved. The number of filters represents the number of channels in the convolution layer. The stride determines the step size with which the filter moves along the image.

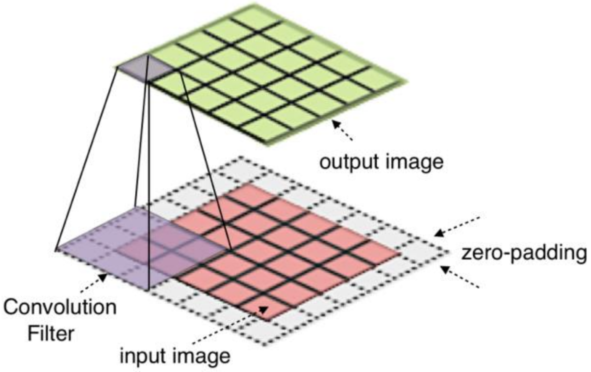


***Figure 3.*** Illustration of the convolution layer with a filter size 3 × 3 and a zero-padding 1 × 1 [Conv].

The pooling layer, also known as downsampling, minimizes the number of parameters used in the network by resizing the previous layer. Similar to the convolutional layer, this layer has a filter that operates across the entire input, but it does not have any weights. It returns the maximum or the mean value of a subregion of the previous layer depending on the type of pooling, max pooling, or average pooling respectively. This layer helps reduce the complexity of the model and also improves efficiency and limits overfitting.

The fully connected layer is the final layer, which connects all its neurons with all the previous neurons. The classification task is performed based on the features extracted through the previous layers and their different filters. While convolutional and pooling layers tend to use ReLu functions, fully-connected layers usually leverage a softmax activation function to classify inputs appropriately, producing a probability from 0 to 1.

Various CNN architectures have emerged since its creation, such as AlexNet (Krizhevsky et al., 2017), VGGNet (Simonyan and Zisserman, 2014), GoogLeNet (Szegedy et al., 2015), ResNet (He et al., 2016), etc.

*Auto-encoders (AEs)*

An auto-encoder is an unsupervised learning algorithm that learns to produce the same output as the one given in the input while using fewer neurons in the hidden layer, as shown in *Figure 4*. It is composed of two components: the encoder and decoder. The encoder part learns the features of the input with fewer parameters, and therefore, it reduces its dimensionality. On the contrary, the decoder part generates the output vector as a representation of the compressed vector in the hidden layer (Le, 2015). When the auto-encoder has multiple hidden units, it is named stacked auto-encoder (SAE). This type of architecture has also been used in biomedical image analysis and has been proven to be efficient in several image processing tasks.

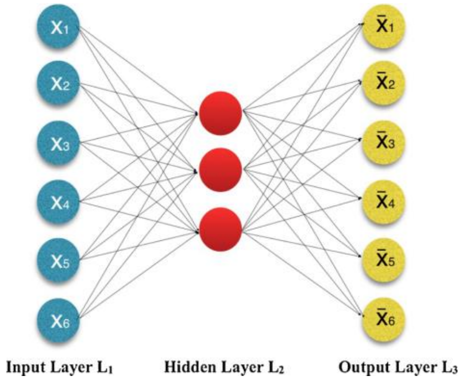


***Figure 4.*** Representation of a simple auto-encoder architecture (Le, 2015).

### *Recurrent Neural Networks (RNNs)*

The essential contributions of RNNs are in the areas of language modeling, such as image labeling, speech processing, and prediction. In such fields, the output is highly correlated with previous data. Hence, the systems should not process the data independently but as a whole sequence. For this very reason, RNNs are the most suitable to handle problems involving sequential data, such as time series or sequences of characters and words, text. RNNs introduce the loop or cycles, where the output of one layer is the input of the same layer or a previous one, as presented in *Figure 5*. However, simple RNN architectures fall short of some processing needs. Actually, input data size is often extensive; therefore, the range of context learned is quite limited, and so the gradients become very small to the point that they vanish to almost zero. This problem is referred to as the vanishing gradient problem (Hochreiter et al., 2001) [26, 2001]. Long short-term memory (LSTM) networks overcome this shortcoming by adding gates to access past information. It enables efficient learning of long sequences by selecting the information to learn and the one to erase.

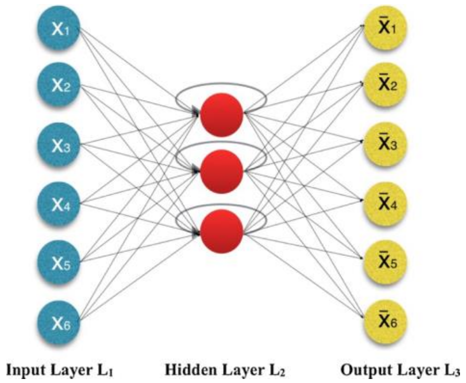


***Figure 5.*** Representation of a recurrent neural network architecture (Le, 2015).

### 2.3.3 Machine and Deep Learning Model Workflow

A typical workflow for applying a machine or deep learning model in some context looks as follows:

*Data preparation*

After obtaining or creating a dataset, one should always prepare the data before any analysis is done. For the purposes of this dissertation, we have assumed that the dataset is made up of a set of pairs (x, y) where x is the input example and y is a label. Then the dataset is split into three folds, commonly a training, validation, and test fold (standard proportions could be 70%, 15%, 15%, respectively). The training set is used for optimizing the parameters, the validation set for hyperparameter optimization, and the test set for evaluation.

*Data preprocessing*

Preprocessing the data is a must for many machine learning models, but also it can help improve model convergence (Lecun et al., 1998). For example, standardizing the data (subtracting the mean and dividing by the standard deviation individually for every input dimension of x) is a common standard preprocessing technique. It is critical to estimate these statistics only on the training data and using these fixed statistics to process the validation and test data, as this deployment simulates more appropriately the real-world application.

*Architecture design*

The next step is deciding on the family of architectures to explore. Although this stage is more of an art than a science, there are a few common heuristics used in practice. It is common to process pixel data, images with convolutions, and sequence data with recurrent networks. Regarding the scale of the architecture, a very rough rule of thumb is that the entire model should have an approximately similar number of parameters as there are examples in the training dataset. An example is the CNN trained on ImageNet, which has over 10M parameters, and regularization techniques (such as L2 regularization, dropout, and data augmentation) are used to further constrain the model to prevent overfitting.

*Hyperparameter optimization*

Hyperparameter optimization is the outer loop that determines good values of hyperparameters that are difficult or impossible to backpropagate into (such as the learning rate, the number of units in the hidden layers, etc.). This process consists of sampling hyperparameters from some search range using grid search technique, optimizing the model, and evaluating the model on the validation set. The final best model is the one that achieves the best performance on the validation set.

*Evaluation*

Once the best-trained model is identified (the lowest validation loss is achieved), the model is validated only once on the test set and reports the performance. Improvements can be obtained by using ensemble models, which average the results of evaluating multiple models trained from different initializations or with different hyperparameters.

The remainder of this dissertation is organized around three projects. Two of them leverage this modeling workflow.

## 2.4 Summary

In this introductory chapter, an overview of the history of EHRs is presented together with some of the most known international agencies that implement guidelines and care pathways, first in paper-based and later digitized ones. Later various machine learning and deep learning architectures are presented, where a few of them are used in studies conducted and included in this dissertation. A typical workflow for applying a machine or deep learning model is also presented. A more detailed state-of-the-art on the specific study topics are presented in the respective studies and can be found in their corresponding chapters.

# 3

# STUDY 1: CARE PATHWAYS – BEACON HOSPITAL

Care pathways, also known as clinical pathways, critical pathways, care paths, integrated care pathways, case management plans, clinical care pathways, or care maps, are used to consistently plan and follow up patients' perioperative and follow-up care (EPA, 2021). They are a way of setting out a process of best practice to be followed in the treatment of a patient with a specific condition or needs and have been implemented internationally since the 1980s (Kinsman et al., 2010). The official definition from European Pathway Association (EPA) is as follows: "*A care pathway is a complex intervention for the mutual decision making and organization of care processes for a well-defined group of patients during a well-defined period.*" (EPA, 2021).

The aim of applying care pathways in practice is to strengthen the quality of care by improving risk-adjusted patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources. Care pathways are created and discussed by a multi-department team, and preferably crossing organizational boundaries. These standardized care programs use standardized documentation, which makes them easy for ongoing audits (Centre for Policy on Ageing, 2021).

In Ireland, according to the National Cancer Registry, 1 in 10 women and 1 in 8 men are at risk of BCa and PCa diagnosis by the age of 74, respectively, and about 30% of all invasive cancers (NCRI, 2021). Although the cumulative lifetime risk of diagnosis is high, the cumulative lifetime risk of death by the age of 74 is 1 in 51 for BCa and 1 in 115 for PCa.

There are different drugs currently used to treat BCa and PCa, and new ones are being developed. Treatments used in breast cancer include chemotherapy, surgery, radiotherapy, endocrine therapy, and recently targeted therapy and immunotherapy. Treatments used in prostate cancer include endocrine therapy, surgery, chemotherapy, radiotherapy, Radium Ra 223 dichloride. Bisphosphonates are used in cases when cancer has metastasized to bones. All these treatments are used in combination with each other in order to cure or control the disease.

This chapter reviews the most relevant literature with regards to one of the main pillars that sustain this dissertation: care pathways (CPs). The analysis of this topic will try to set the scene for the study presented later in this chapter which is conducted at Beacon Hospital. The study's main aim is to understand and analyze the care pathways for breast and prostate cancer patients and to evaluate the association between different treatment lines, the lifestyle and demographic characteristics of these patients.

## 3.1 Background

The implementation of clinical pathways has demonstrated improvements on patients' and healthcare professionals' management (Akechi et al., 2015), cost-effective for the hospital and the patients (Akechi et al., 2015; Thorsen et al., 2011), and better satisfied and educated patients (Chen et al., 2000; Munir et al., 2011; Tamburini et al., 2003). In this context, there is a need to understand and analyze the care pathways cancer patients follow. An overview of the current state of cancer patients' care pathways is presented. *Figure 6* illustrates the distribution of published manuscripts related to care pathways in the last 2 decades (2000 - 2020).

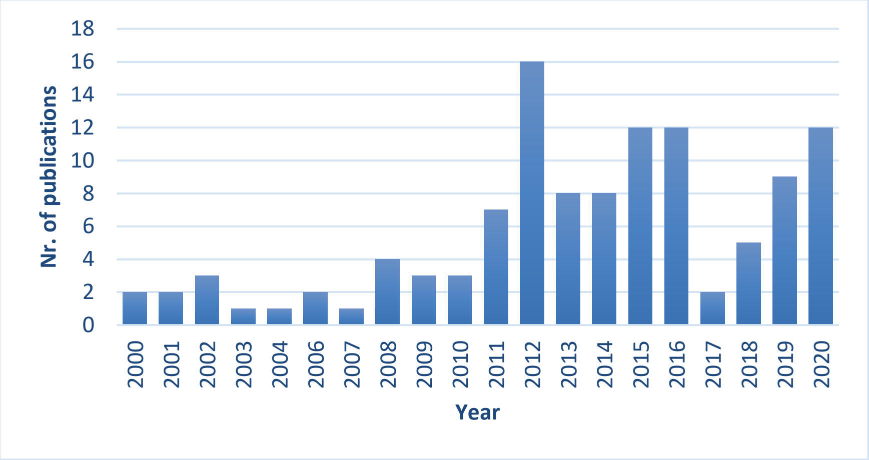


***Figure 6.*** The number of cancer care pathway publications over the years 2000 – 2020.

Most of the publications are between 2011 and 2016, showing that although care pathways had been implemented for more than 30 years, they only gained momentum after 2010. This is also the period when algorithms started to be used to gain insight into care pathways (Caron et al., 2014; Catanuto et al., 2016; Huang et al., 2014, 2013; Meier et al., 2015; Scheuerlein et al., 2012; Yamamoto et al., 2012).

The majority of publications are from the UK and the USA, 22 and 23 respectively, *Figure 7*. Others include countries that had only one published study, and these countries are Czech Republic, Egypt, Ireland, Lebanon, Pakistan, Saudi Arabia, and South Korea. *Figure 8* shows the number of studies according to the continent. The only continent without representation is South America. Eighty-six (65%) studies were conducted in Europe.

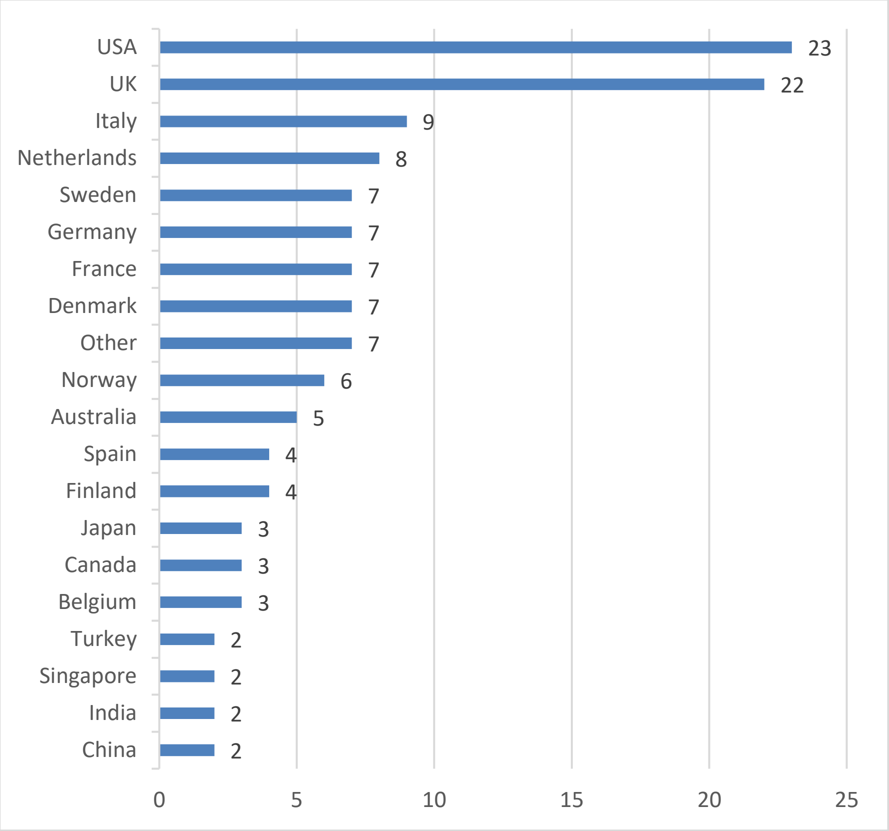


***Figure 7.*** Publications per country.

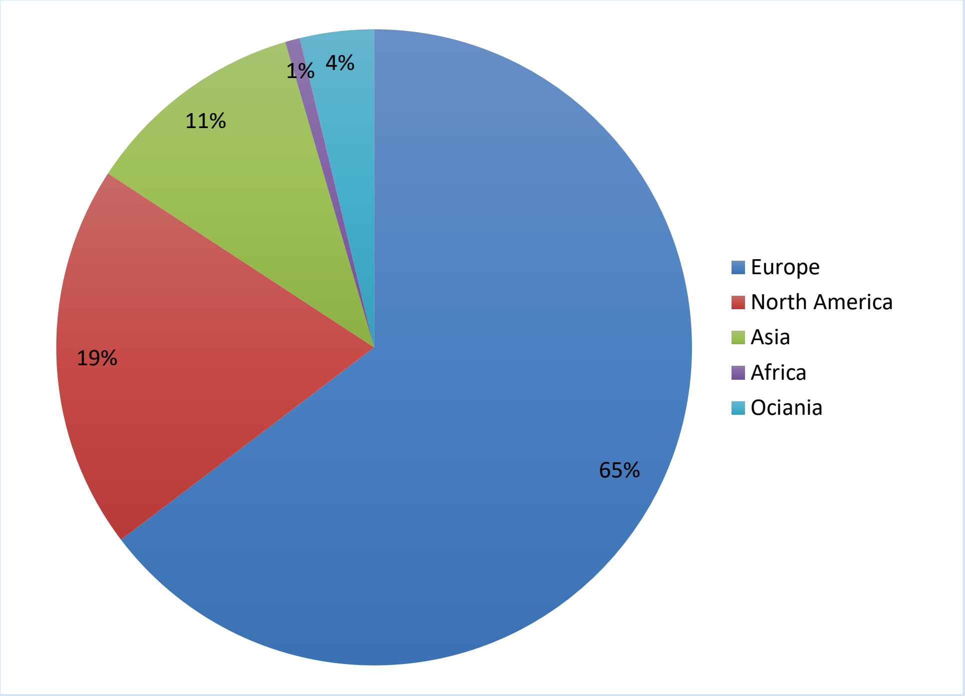


***Figure 8.*** Publications per continent.

Regarding the type of cancer involved, most of the studies were about or included, among other cancers, breast cancer. There was a total of 28 studies. A deep dive on the care pathways involved in breast, prostate, and colorectal cancer will be discussed separately in later sections in this chapter. The other most researched CPs were for lung (21), colorectal (19), and gastric (12) cancer. *Figure 9* shows a treemap of all cancers pertaining to the included CP studies.

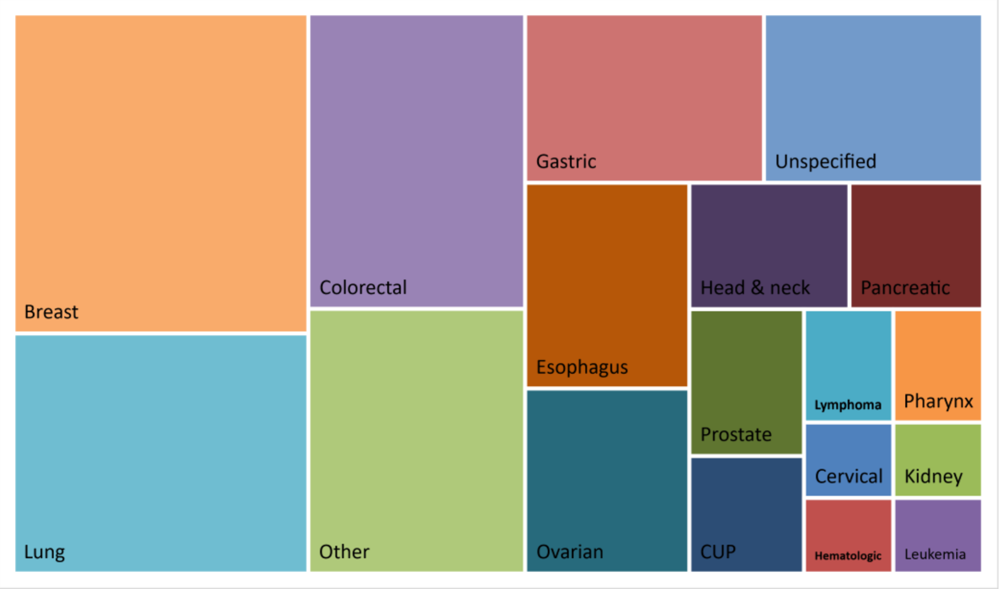


***Figure 9.*** Treemap chart of the most studied cancer care pathway per cancer type.

Across all studies, the number of patients who participated in the study, sample size varied from 8 to 405,635 subjects. Studies with smaller sample sizes were usually qualitative studies, whereas studies with a bigger sample size were the ones involving either electronic healthcare records (EHRs) or national cancer registries. 61% (14/23) of studies with a number of patients 1000 or more are about the diagnosis phase. Two studies are about screening and the 7 about any perioperative phase. In total, 580,037 patients participated in these studies.

### 3.1.1 Care Pathway Phases

More than 60% of the included studies are about the perioperative phase, preoperative phase making the most considerable portion, 61 studies. The diagnosis phase follows with 41, follow-up with 22, and palliative with 13, see *Figure 10*.

A care pathway may have different variations, meaning one or more phases might be implemented and included in one CP. There are studies that their CP include only the diagnosis phase (Falborg et al., 2020; Hameed Khaliq et al., 2019; Muller et al., 2020; Murchie et al., 2020; Næser et al., 2018; Rua et al., 2020; Yap et al., 2018), but there are also CPs that include diagnosis and preoperative phase (Barrett et al., 2010; Colonna et al., 2019; Delaloge et al., 2016; Esteva et al., 2014; Laurent-

Badr et al., 2020; MacPherson et al., 2012; Mousa et al., 2011; Rai et al., 2015; Roennegaard et al., 2018; Väisänen et al., 2014; Vajdic et al., 2015; Jolanda C van Hoeve et al., 2015). Some variations of CPs are as follows:

- CPs that include only one care phase at a time,
- all perioperative phases (Barry et al., 2012; Chiang et al., 2020; Compagna et al., 2014; Fasola et al., 2012; Kay et al., 2020; Kim et al., 2015; Klinkhammer-Schalke et al., 2012; Markar et al., 2014; Numan et al., 2012; Pease et al., 2004; Salamonsen et al., 2016; Sancho et al., 2010; Soria-Aledo et al., 2011) ,
- all care phases expect screening phase (Maher and McConnell, 2011; Viklund and Lagergren, 2007; Yip et al., 2015),
- preoperative and intraoperative phases (Corrao et al., 2020; de Kok et al., 2010; Patel et al., 2014; Scheuerlein et al., 2012; Vijayakumar et al., 2016; Wolf et al., 2015),
- intraoperative and post-operative (Alamoudi et al., 2011; Nussbaum et al., 2014),
- post-operative and follow-up (Klinkhammer-Schalke et al., 2020; Malmström et al., 2016; Nuemi et al., 2013), etc.

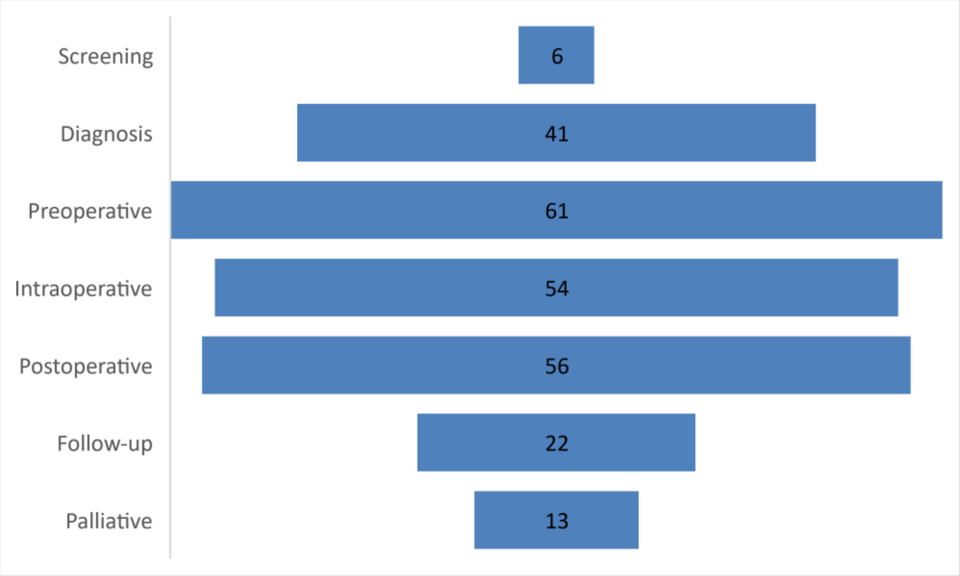


***Figure 10.*** Publications per care pathway phase.

There are cases when the type of cancer is not as important as the group of patients or the purpose the CPs are being implemented. There are 10 studies that fit this profile. Seven out of 10 studies deal with how to maintain cancer, focusing on palliative care and the CPs for this specific care phase. Of the other three, two deal with geriatric patients (Murchie et al., 2020; Schmidt et al., 2017), and one aims to describe the design and development of a model that enables the primary use of EHRs in clinical trials by integrating them with CPs (Yamamoto et al., 2012).

*Screening care programs*

If implemented correctly, screening programs can be critical. Two studies in the USA test the hypothesis that apart from medical benefits, screening programs could result in saving costs. (Santillan et al., 2008) evaluate the feasibility and impact of implementing a CP for Pap test on cost to the USA healthcare system utilization in screening and surveillance of gynecologic cancers. They conclude that indeed such CP is not only feasible, but it also offers opportunities for cost savings in the expenditure of gynecologic oncology healthcare. (Wolf et al., 2015) test the idea if a statewide program could be established to ensure that low-income residents receive colonoscopy for colorectal cancer (CRC) screening and diagnostic evaluation. The study resulted in removing adenomas from 27% of patients, and 1% of the screened patients got diagnosed with cancer. 325 adenomas were removed, thus predicting 325 fewer future CRC incidences and saving future costs.

Two low-dose computed tomography (LDCT) lung screening programs are presented, one by (Baldwin et al., 2011) in the UK and the other one by (Salazar et al., 2020) in the USA. The main result of (Baldwin et al., 2011) is the development of a lung cancer CP and further evaluation through clinical trials. The difference (Salazar et al., 2020) is that they want to make the LDCT eligible for high-risk patients annually by involving primary care clinics and the local knowledge and partnerships for their stepped-wedge trial. The other two studies are both in the UK and deal with cancer of unknown primary (Creak, 2020) and bowel cancer (Blagden et al., 2020). (Creak, 2020) aims at shortening the diagnostic pathway and improve patient support by implementing a referral program. Their three-year study concludes that such a program is feasible and manageable within a tertiary CUP clinic, resulting in high rates of cancer diagnoses, with attendant early support from specialist nursing teams and oncological review.

The last screening program presented aims at exploring processes and beliefs around bowel cancer screening in a UK prison (Blagden et al., 2020). The study resulted in a high willingness amongst prisoners to be screened for bowel cancer; however, severe logistical challenges are encountered in delivering such screening programs. Though challenging, providing good-quality understandable information was vital. Summarization of all screening studies is shown in *Table 1*.

***Table 1.*** Summary table of screening programs.

| *Author* | *N* | *Age* | *Study duration* | *Methodology* | *Analysis* | *Type of cancer* | *Country* |
|---|---|---|---|---|---|---|---|
| (Santillan et al., 2008) | 3280 | NA | Jan 2004 - May 2006 | Before and After CP | Quantitative | Endometrial, Ovarian, Cervical | USA |
| (Baldwin et al., 2011) | 4000 | 50 - 75 | NA | RCT | Quantitative | Lung | UK |
| (Wolf et al., 2015) | 13774 | > =50 | Jan 2006 - June 2012 | surveys | Quantitative | Colon | USA |
| (Creak, 2020) | 258 | 23 - 95 | Sept 2015 - Aug 2018 | NA | Quantitative | CUP | UK |
| (Blagden et al., 2020) | 8 | 60 - 74 | NA | Qualitative | Descriptive | Bowel | UK |
| (Salazar et al., 2020) | NA | NA | NA | NA | Descriptive | Lung | USA |

*N = number of patients; NA = Not Available; RCT = Randomized Control Trial*

*Diagnosis Care Pathways*

When the patient or the general practitioner (GP) has a suspicion of cancer, they refer the patient to the hospital or specialist clinic to perform further examinations. Depending on the type of cancer, the patient goes through different imaging scans and examinations to clarify whether the suspicion can be justified by a finding of physical or radiological changes indicative of malignancy. Milestones defined in this timeframe are the day the referral is received from the GP, the first appointment with the specialist, the decision of treatment, and the start of treatment (Dyrop et al., 2013). This period is defined as the diagnosis phase. Many hospitals and clinics have well-defined diagnosis care plans. These care plans depend on the cancer disease, hospital, national guidelines, stage of cancer, patient, etc.

Understanding if a diagnosis care plan is performing better than the standard of care, hospitals run before and after CP implementation studies (Dyrop et al., 2013; Gerardi et al., 2008; Jakobsen and Jensen, 2016; Tastan et al., 2012), cohort studies (Barrett et al., 2010; Lyhne et al., 2013; Næser et al., 2018; Roennegaard et al., 2018; van Dam et al., 2013), clinical trials (Rua et al., 2020), use EHRs or repositories (Delaloge et al., 2016; Maher and McConnell, 2011; Rai et al., 2015; Vajdic et al., 2015; Yip et al., 2015), or a combination of these.

Different studies have different objectives on why they implement CPs and how to calculate their performance. (Aasebø et al., 2012; Bond et al., 2016; Desandes et al., 2012; Tastan et al., 2012; van Dam et al., 2013) focus on improving the quality of life and quality of care of cancer patients. Three of them are about breast cancer, and apart from (Aasebø et al., 2012), they are about patients aged 18 or older.

In 2007 Denmark introduced a national policy of fast-track CPs. Since then, they have implemented head and neck (Lyhne et al., 2013; Roennegaard et al., 2018), sarcoma (Dyrop et al., 2013), and penile (Jakobsen and Jensen, 2016) fast-track CPs. These CPs reduce waiting times for diagnosis; however, future studies need to address the long patient intervals.

Several papers have studied the financial aspect of the CP implementation. In France, a one-stop breast clinic was created, which proved to be providing timely and cost-efficient diagnosis with high accuracy (Delaloge et al., 2016). (Rua et al., 2020) on the other hand, it evaluates two distinct pathways, CT colonography (CTC) and optical colonoscopy (OC), for initial colonic investigation in low-to-intermediate risk of colorectal cancer patients. CTC was found to be a potential replacer of OC for these patients, leaving OC for high-risk patients. (MacPherson et al., 2012) conclude that early PET-CT scanning for suspected lung cancer patients, which may be suitable for curative therapy, could result in more efficient staging with little additional cost.

### *Perioperative Care Pathways*

Once a diagnosis is established, the next step is to create the treatment plan the patient will undergo. The time interval between the diagnosis and the start of the treatment is crucial. Various studies are published to access factors related to the diagnosis-to-treatment interval (DTI).

(Dang-Tan et al., 2010) assess delay factors related to children and adolescents diagnosed with leukemia and lymphoma in Canada. They conclude that age was found to be positively associated with patient delay for both diseases; the place where the patient was first seen also factored in, with patients first seen by a general practitioner facing a higher risk of delay. The association of DTI and the cancer subtype was noted too. (Pati et al., 2013), explore barriers and enablers in seeking cancer treatment in India, with financial constraints being one of the significant reasons. Other barriers are low awareness of the presenting signs and symptoms of cancer and limited knowledge of the availability of cancer diagnosis and treatment facilities. The main enabling factors are family and friends' support toward seeking treatment. (Sharma et al., 2016), examine practice patterns to determine risk factors for prolonged DTI in patients receiving chemoradiation for oropharyngeal squamous cell carcinoma (OPSCC) using the USA National Cancer Database. Race, intensity-modulated radiation therapy (IMRT), insurance status, and high-volume facilities are significant risk factors for prolonged DTI, which is crucial in survival outcomes. They found that there was a 2.2% increase in the risk of death for every week increase in DTI. A Dutch study, (van de Ven et al., 2019), explores the variation in DTI when the molecular diagnosis is used for patients with stage III and IV non-small cell lung cancer (NSCLC). Results show that indeed, there are variations, especially for patients receiving radiotherapy or targeted therapy. The main variation factors are tumor stage, performance status, and histology.

Some of the most measured outcomes when implementing perioperative CPs are the length of hospital stays (LOS), readmission to hospital, costs, quality of care, improvements in care processes, etc. Shorter LOS were seen in post CP implementations in various countries and different cancer diseases (Chen et al., 2000; Compagna et al., 2014; Kay et al., 2020; Numan et al., 2012; So et al., 2008; Zhu et al., 2020), improved quality of care and life (de Kok et al., 2010; Messager et al., 2016; Numan et al., 2012; Tastan et al., 2012), reduced costs (Chen et al., 2000; Dautremont et al., 2016; So et al., 2008), complications (Compagna et al., 2014; Zhu et al., 2020), and patient anxiety levels (Tastan et al., 2012).

Enhanced recovery after surgery (ERAS) programs are specific CPs for after surgeries said to reduce LOS and postoperative complications when implemented. (Kay et al., 2020) in the USA observed that ovarian cancer patients undergoing open surgery had shorter LOS by 2.5 days when ERAS was implemented, and the use of narcotics in the hospital and after discharge was less compared to the standard perioperative care protocol. In China, (Zhu et al., 2020) concluded that LOS and the incidence of certain surgical complications after ERAS CP for pancreatic cancer were reduced. However, 1-year survival rates kept the same in both ERAS CP and standard protocol. (Joris et al., 2020), tested the feasibility of an ERAS CP for elderly and young patients diagnosed with colorectal cancer in Belgium. Although the elderly patients present with higher risk factors (anemia, COPD, cardiac disease, and cancer), they did not experience more postoperative medical or surgical complications than younger patients. The difference in median LOS between the two groups was 0, demonstrating non-inferiority.

### *Follow-up care pathways*

Support is essential after cancer treatment and is often provided in the form of follow-up programs. Nurse-led follow-up care has been considered more appreciative by patients (Pernilla Viklund et al., 2006), has a positive impact on the patients' experience of received information (Malmström et al., 2016) and supportive therapy (Schmidt et al., 2017), it seems to result in improved care and outcomes for patients undergoing robotic-assisted radical prostatectomy and may also lower impact on hospital resources (Birch et al., 2016). (Yip et al., 2015) highlights the increased number of prostate cancer patients being in the follow-up phase, the challenge to provide them with adequate care, and the importance of collecting and reporting the number of patients following different CPs to help improve future CPs.

### *Palliative care pathways*

Palliative care amounts to optimizing the quality of life for patients with incurable cancers; therefore, the focus is on 1) the alleviation of the symptoms, 2) the up-to-date treatment goals communication, and 3) the patients and their families support throughout the disease (Van Beek et al., 2016). The evidence that palliative care has effectively improved the quality of life of patients with advanced cancer is plentiful (Bakitas et al., 2009; Temel et al., 2010; Zimmermann et al., 2014). However, there are studies that discuss the nonexistence of palliative care, such as (Halawi et al., 2012) and how implementing one could address issues, such as including fatigue as a symptom, considering subgroup differences, managing pain more effectively, and giving special care to vulnerable groups.

### 3.1.2 Care Pathways and Research Methods

Quantitative methods were the most common ones to be used in analyzing the outcomes of the studies, 77% of all studies (*Figure 11*). The most used methods were descriptive statistics (mean, median, percentages, interquartile range), followed by Student t-test, Kruskal-Wallis test, Fisher's exact test, Wilcoxon test, Pearson chi-square test, Manne Whitney U-Test, Friedman's test, ANOVA, logistic and linear regression, etc.

The most used methods to conduct these studies were through various questionnaires, surveys, and data repositories, databases, or electronic healthcare records. Qualitative methods were also used to collect the data and to analyze it. The most used qualitative research methods were interviews, focused groups, and observations. Content, thematic, and interpretative phenomenological analysis were used to analyze the data.

There are studies that use a mixed-methods approach. They use a combination of questionnaires and interviews (Numan et al., 2012; Pati et al., 2013; Phillips et al., 2015; Schmidt et al., 2017; J C van Hoeve et al., 2015; P Viklund et al., 2006).

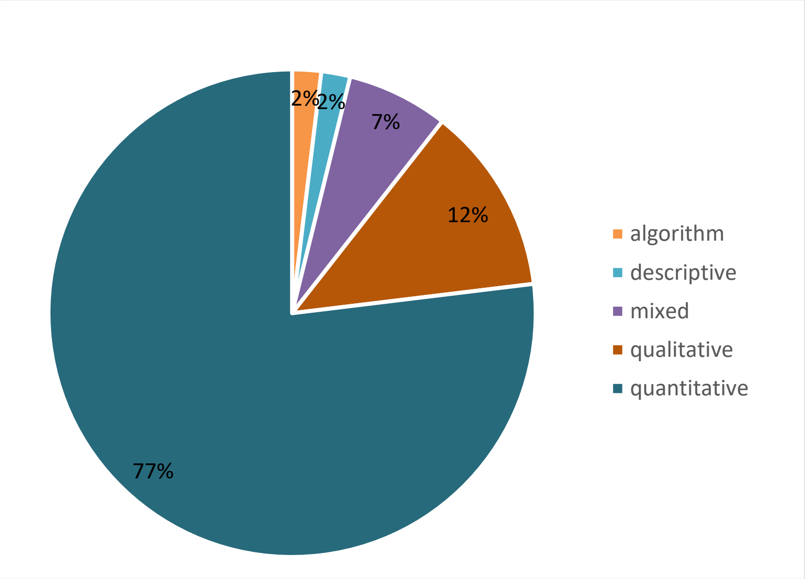


***Figure 11.*** Type of analysis used in selected publications.

### 3.1.3 Breast Cancer Care Pathways

Breast cancer (BCa) is the most common type of cancer in women, and most of the CPs in this review are about breast cancer alone or in combination with other cancers. However, only some of them have included the care pathway itself in the manuscript.

(Bhatnagar et al., 2009) analyze the clinical outcomes for BCa patients treated with intensity-modulated radiation therapy (IMRT) after breast-conserving surgery. Their CP explains the processes pertaining to contours and markers for the Breast IMRT planning on a representative axial CT slice. It includes procedures such as positioning and immobilization, image acquisition which can be either a CT scan or an MRI, structure segmentation, and IMRT treatment planning. They conclude that long-term follow-up is necessary. (Barry et al., 2012) present a case for an enhanced CP for an outpatient axillary lymph node dissection (ALND) procedure. This enhanced CP is implemented with input from a multidisciplinary team and includes that the patients and the staff changing expectations, provide upper limb exercises instructions in preoperative clinic and DVD, moving the postoperative drain management instructions to the preoperative clinic, and home-care assessment. In the intraoperative phase, ET intubation is avoided where appropriate, narcotic/sedative use is reduced, IV sedation +/- LMA is used, postop analgesia is initiated, and longer-acting Bupivacaine wound infiltration is used upon completion of the case. The postoperative phase includes 24-h access emergency phone number, 24-h postoperative nurse phone call, postoperative clinic review within 2 weeks, and record 24-h and 30-day morbidity. (Klinkhammer-Schalke et al., 2012) focuses on the quality of life (QoL) of the patients, and for this reason, their CP includes steps that make sure patients have a good QoL. They have a QoL unit to achieve this, which informs the coordinating practitioner if any deficit is found. Therapeutic options according to QoL deficits include physiotherapy, psychotherapy, social support, pain therapy, nutrition, and fitness. Patients are followed every 3, 6, 9, and 12 months.

(Ryhänen et al., 2012a) and (Ryhänen et al., 2013) present the same CP. The CP is based on (Ryhänen et al., 2012b); however, it is adapted to be internet-based and to empower the patients to understand the content and to use that knowledge in their own treatment and care. The CP includes each milestone, starting from diagnosis (first visit in hospital), before and after surgery, meeting with the oncologist, chemotherapy, radiotherapy, and one year after diagnosis. After each step, they test the expectations on patient knowledge, pathway-related knowledge, received knowledge, and pathway-related received knowledge, knowledge test for the breast cancer patient, source of knowledge, and satisfaction of patient education.

Different from the previous ones, (Tastan et al., 2012) describe the BCa CP in detail. The CP is a checklist of all the examinations that need to be performed on the day of the surgery and during the 4 days of postoperative care and by different healthcare providers. (Baffert et al., 2015) CP spans to other care phases besides perioperative (see *Table 2*). *Table 3* presents the complete list of published studies about BCa CPs.

***Table 2.*** CP implemented by (Baffert et al., 2015).

| **Diagnosis** | **Surgery and Post-Surgery** | **Adjuvant Treatment** | **Follow-up** |
|---|---|---|---|
| • Proposal for participation<br>• Inclusion, signed consent, information.<br>• Delivery of the Logbook<br>• Inpatient and outpatient care use from diagnosis to surgery | • Type of surgery<br>• Anatomopathological results<br>• Decision of the MDT<br>• Mode of hospitalization<br>• Inpatient care use from surgery | • Mode of adjuvant treatment<br>• Inpatient care use from post-surgery visit | Phone call<br>• Reminder of the logbook filling rules.<br>Post follow-up visit<br>• Return of the Logbook.<br>• Follow-up planning |
| Surgery scheduling visit | Post-surgery visit | Post adjuvant treatment visit | |
| | Logbook<br>• Satisfaction questionnaires<br>• Occupational questionnaire<br>• Outpatient care use<br>• Out-of-pocket expenses<br>• Sociodemographic data | | |

***Table 3.*** List of all studies included in this literature review about breast cancer care pathways.

| *Author* | *N* | *Age* | *Study duration* | *Type of research* | *Methods* | *Analysis* | *Country* |
|---|---|---|---|---|---|---|---|
| (Lindop and Cannon, 2001) | 12 | 26 - 58 | Aug 1998 - Oct 1998 | | Interviews | Qualitative | UK |
| (Klinkhammer-Schalke et al., 2008) | 170 | 34 - 86 (median = 58) | Jan 2003 - June 2004 | RCT | Questionnaire | Quantitative | Germany |
| (Bhatnagar et al., 2009) | 495 | 28 - 90 (median = 59) | Dec 2001 - March 2005 | Cohort study | | Quantitative | USA |
| (de Kok et al., 2010) | 282 | ≥ 18 | Dec 2005 - June 2006 & Dec 2006 - June 2007 | Before & after study | Questionnaire | Quantitative | Netherlands |
| (Maher and McConnell, 2011) | EHRs | NA | 2008 | EHRs | | Quantitative | UK |
| (Mousa et al., 2011) | 163 | mean = 51.6; median = 53 | Dec 2009 - Nov 2010 | | Interviews | Qualitative | Egypt |
| (Barry et al., 2012) | 282 | NA | July 2009 - June 2010 | Database | | Descriptive | USA |
| (Halawi et al., 2012) | 100 | 18 - 85 (mean = 51.5) | Jan 2011 - March 2011 | Cross-sectional study | Questionnaire | Qualitative | Lebanon |
| (Klinkhammer-Schalke et al., 2012) | 200 | ≥ 18 | Nov 2004 - Oct 2007 | RCT | Questionnaire, Interviews | Mixed | Germany |
| (Ryhänen et al., 2012b) | 38 | 40 - 66 (mean = 53.5) | 2008 - 2010 | Platform/ educational | Questionnaire | Quantitative | Finland |
| (Ryhänen et al., 2012a) | 90 | 40 - 69 | 2008 - 2010 | RCT | Questionnaire | Quantitative | Finland |
| (Tastan et al., 2012) | 69 | ≥ 18 | March 2004 - April 2005 | | Questionnaire | Quantitative | Turkey |
| (Pati et al., 2013) | 68 | 26 - 85 (mean = 46.5) | April - June 2011 | Cross-sectional study | Questionnaire, Interviews | Mixed | India |
| (Ryhänen et al., 2013) | 90 | 40 - 69 | 2008 - 2010 | RCT | Questionnaire | Quantitative | Finland |
| (van Dam et al., 2013) | 1346 | NA | 2002 - 2010 | Cohort study | | Quantitative | Belgium |
| (Huang et al., 2014) | NA | NA | NA | Event logs | | Algorithms | China |
| (Baffert et al., 2015) | 1000 | NA | NA | Protocol | Questionnaire, Observation | Mixed | France |
| (Bond et al., 2016) | 19440 | ≥ 21 | NA | Protocol | Questionnaire | Quantitative | Italy, Netherlands, Turkey, Germany, Czech Republic, Norway, Poland, UK |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| (Catanuto et al., 2016) | 52 | mean = 64, 51 | Nov 2013 | Software | | Quantitative | Italy |
| (Delaloge et al., 2016) | 10602 | 50 - 74 (mean = 55) | April 2004 - Nov 2012 | Database | | Quantitative | France |
| (Mercadante et al., 2016) | 314 | ≥ 18 (mean = 65.7) | NA | | Survey | Quantitative | Italy |
| (Vijayakumar et al., 2016) | 334 | ≥ 45 | Sept 2011 - Aug 2013 | Clinical audit | | Quantitative | India |
| (Lefeuvre et al., 2017) | 52128 | ≥ 18 | 2012 - 2013 | Database | | Quantitative | France |
| (Colonna et al., 2019) | NA | 27 - 67 (mean: 47.5) | June 2014 - Oct 2016 | Clinical Performance | Vignettes | Quantitative | USA |
| (Hameed Khaliq et al., 2019) | 200 | 22 - 70 | Aug - Dec 2015 | | Interviews | Quantitative | Pakistan |
| (Corrao et al., 2020) | 16753 | ≥ 18 | 2011 - 2016 | Adherence with recommendations | | Quantitative | Italy |
| (Falborg et al., 2020) | 4502 | ≥ 40 (median = 66) | 2013 - 2015 | | | Quantitative | Denmark, Sweden, Norway, Canada, UK, Australia |

*N = number of patients; RCT = Randomized Control Trial; NA = Not Available*

### 3.1.4 Prostate Cancer Care Pathways

Prostate cancer (PCa) is the most common cancer in men. (Yip et al., 2015) segment the population of prostate cancer survivors in the UK into different care phases according to their needs (see *Table 4*). They estimate that approximately a fifth of the patients are either in the treatment phase or have already done so the previous year and are now in the recovery and readjustment phase. As such, patients undergoing post-treatment monitoring will eventually constitute the biggest group of PCa survivors. Adequate follow-up care should be provided, although it is seen as a challenge. There ought to be more data gathered to understand better the CPs followed by PCa patients, which will help create future care programs.

***Table 4.*** Prostate cancer CP phases according to (Yip et al., 2015).

| **Diagnosis & Treatment** | **Recovery & Readjustment** | **Watch & wait** | **Active surveillance** | **Initial monitoring** | **Ongoing monitoring** | **Progressive care** | **End of life** |
|---|---|---|---|---|---|---|---|
| Newly diagnosed <1 year: assumed need of acute sector care | Surviving the first year and < 2 years: assumed need of rehabilitation | Diagnosed but receiving no anticancer treatment (prostate only) | | Up to 5 years from diagnosis: designated as 'initial monitoring' | Beyond 5 & 10 years from diagnosis: designated 'ongoing monitoring' | Incurable disease but not in last year of life: assumed need more treatment and support | End of life care: final year of life and subset of deaths occurring in the first year of diagnosis |

A more detailed CP is described by (Birch et al., 2016), presenting a Robocare pathway for PCa patients in Australia, *Table 5*. The CP coordinates the care between disciplines, length of stay, and readmission rates. It presents as more holistic as it assesses patient satisfaction, sexual and psychological aspects for the care. *Table 6* summarizes the PCa CPs included in this review.

***Table 5.*** Robocare CP for RARP (Birch et al., 2016).

| |
|---|
| **Referral received** |
| **Phone call to the patient**<br>(robotic nurse specialist) |
| **Outpatient clinic**<br>Within 2 weeks of referral received<br>• Consent<br>• Health questionnaire<br>• Prostate cancer pack<br>(urologist/ robotic nurse specialist) |
| **RoCaP Clinic**: Monthly 2-4pm<br>• Preoperative bloods/ ECG<br>• Consent: translational research and tissue banking |

| |
|---|
| Pre-Robotic Prostatectomy Education<br>• Urinary, sexual, psychological function<br>(urologist, anesthetist, robotic nurse specialist, sexual health nurse practitioner, physiotherapist, psychologist) |
| **RARP**<br>Robotic-Assisted Radical Prostatectomy<br>(urologist/ robotic nurse specialist) |
| **Postoperative Day 1**<br>Discharge midday<br>(urologist/ robotic nurse specialist, ward nurse) |
| **Postoperative Day 2**<br>• Phone calls<br>• Bowels, lap sites, IDC<br>(robotic nurse specialist) |
| **Postoperative Day 7 -10**: Outpatient clinic<br>• TOV<br>• Histopathology<br>• Survivorship care plan<br>• PSA follow-up, urinary & sexual function<br>(robotic nurse specialist) |
| **Post TOV Day 4**<br>• Phone calls<br>• Urinary, sexual & psychological function<br>(robotic nurse specialist) |
| **2 months Post RARP:** Outpatient clinic<br>• PSA<br>• Urinary, sexual & psychological function<br>(urologist, robotic nurse specialist, sexual health nurse practitioner, psychologist) |
| **3-monthly follow-up until 1 year, 6-monthly until 5 years, and annually up to 10 years Post RARP:** Nurse-led phone clinics<br>• PSA<br>• Urinary, sexual & psychological function<br>(robotic nurse specialist) |
| *RoCaP = robotic cancer of the prostate perioperative information clinic; RARP = robotic assisted radical prostatectomy; IDC = indwelling catheter; TOV = trial of void.* |

***Table 6.*** The list of included PCa CPs in the review.

| *Author* | *N* | *Age* | *Study duration* | *Type of research* | *Analysis* | *Country* |
|---|---|---|---|---|---|---|
| (Bhayani et al., 2003) | 60 | mean = 60.5 & 57.4 | July 2001 - June 2002 | Database | Quantitative | USA |
| (Davis et al., 2002) | 73 | median = 25 & 28 | 1988 - 1996 | Before & after study | Quantitative | USA |
| (Pease et al., 2004) | 148 | 27 - 88 (mean = 66) | Jan - Dec 1997 & June 1999 – July 2000 | Before & after study | Quantitative | UK |
| (Yip et al., 2015) | EHRs | NA | 2010 | Database | Quantitative | UK |
| (Birch et al., 2016) | 124 | NA | July 2012 - Dec 2013 | Database | Quantitative | Australia |

## 3.1.5 Colorectal Cancer Care Pathways

The third and the last CPs for specific cancers is colorectal cancer (CRC). Unlike breast and prostate cancer, CRC can be prevented. (Sancho et al., 2010) show the evolution of a CRC CP throughout the years and how the self-evaluations of the process have improved several indicators from 2002 to 2007, *Table 7*.

***Table 7.*** Changes to CRC CPs over the years (Sancho et al., 2010).

**2002–2003**

- Initial development of the clinical pathway
- Creation of a multidisciplinary group
- Virtual colonoscopy

**2004–2005**

- High-resolution Magnetic Resonance Imaging
- Fat clearance techniques to optimize lymph node count
- Sedation assisted colonoscopy

**2006–2007**

- Specially dedicated oncologist
- Specially dedicated pathologist
- Extension study protocol with thoracic-abdominopelvic CT
- Virtual colonoscopy (the same day if the colonoscopy was incomplete)
- Entry into the group of 2 colorectal surgeons accredited by the European Union
- Surgery performed only by specially dedicated colorectal surgeons
- Review criteria for admission to the Intensive Care Unit
- Multimodal rehabilitation program (fast track)
- Analysis of the quality of the mesorectum

(Maher and McConnell, 2011) describe the process of how they estimated the number of patients in different care pathways using available data. They categorize the care phases as shown in *Table 8*. The exact estimation process was used by (Yip et al., 2015).

***Table 8.*** Maher & McConnell et al., 2011.

| **Diagnosis & Treatment** | **Rehabilitation** | **Initial monitoring** | **Ongoing monitoring** | **Progressive care** | **End of life** |
|---|---|---|---|---|---|
| Newly diagnosed: assumed need of acute sector care | Surviving the first year: assumed need of rehabilitation | Up to 5 & 10 years from diagnosis: designated as ‘initial monitoring’ | Beyond 10 years from diagnosis: designated ‘ongoing monitoring’ | Incurable disease but not in last year of life: assumed need more treatment and support | End of life care in the final year - a subset of deaths in the first year of diagnosis |

(Soria-Aledo et al., 2011) describe a holistic CP of colorectal carcinoma that is not focused only on the medication aspect but also on diet, information, activity, starting from Day 1 preoperative in the admission ward to Day 7 ward. Check the article for full details on the CP. (Pati et al., 2013) explores the treatment-seeking pathways in India for various cancer patients, among them breast cancer as well. They focus not only on order but also on the time it takes a patient to be referred, diagnosed, and treated. *Figure 12*. (Redaniel et al., 2015), walks through a 2-week referral pathway and identifies points for improvements, *Figure 13*.

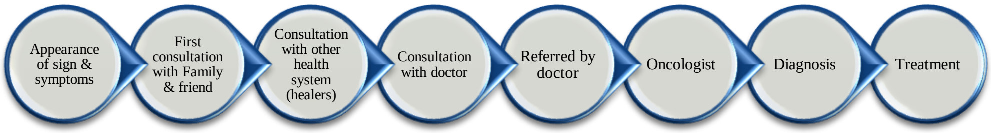

Appearance of sign & symptoms
First consultation with Family & friend
Consultation with other health system (healers)
Consultation with doctor
Referred by doctor
Oncologist
Diagnosis
Treatment

***Figure 12***. The diagnosis CP for cancer patients in India, Pati et al., 2013.

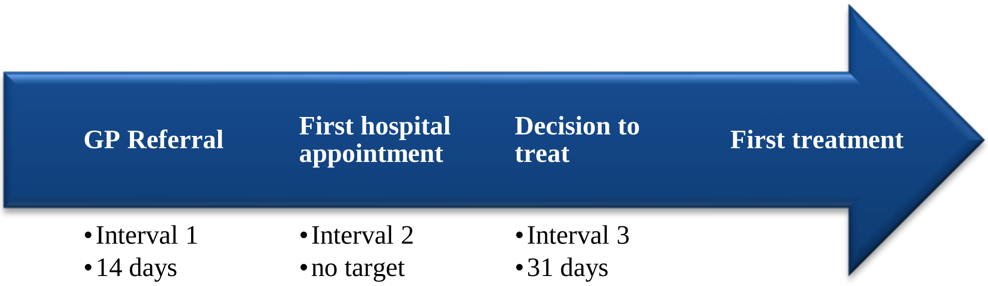


***Figure 13.*** Rapid referral pathway for CRC, Redaniel et al., 2015.

(Wolf et al., 2015) deal with the creation of a CRC program that offering low-income people receive colonoscopy for CRC screening and diagnostic evaluation. Each component of the RCR and associated activities and responsible parties are shown in *Table 9*. The complete list of all selected papers can be found in *Table 10*.

***Table 9.*** Colorado screening program, Wolf et al., 2015.

| **Component** | **Activities** | **Responsible party** |
|---|---|---|
| **Endoscopic screening** | Bowel preparation, colonoscopy or sigmoidoscopy, anesthesia, pathology | Providers |
| **Cancer treatment** | Surgery, chemotherapy, radiotherapy | Specialty providers |
| **Patient navigation** | Clinic in-reach, patient education, screening support, follow up of results, treatment support | Clinic personnel |
| **Quality assurance** | Web-based data collection and analysis for process and quality monitoring | Navigators, Medical Quality, Assurance Committee |
| **Public awareness** | Large and small media campaigns, mailings to promote screening | Coordinating center, Colorado CRC Task Force |

***Table 10.*** A list of included studies about CRC CPs.

| *Author* | *N* | *Age* | *Duration of the study* | *Type of research* | *Methods* | *Analysis* | *Country* |
|---|---|---|---|---|---|---|---|
| (Sancho et al., 2010) | 166 | mean = 71 | Jan 2002 - Dec 2007 | Cohort study | | Quantitative | Spain |
| (Maher and McConnell, 2011) | NA | NA | 2008 | EHRs | | Quantitative | UK |
| (Soria-Aledo et al., 2011) | 270 | mean 68 | Jan 2002 - Jan 2003 & Jan 2004 - Dec 2008 | Before & after study | | Quantitative | Spain |
| (Scheuerlein et al., 2012) | NA | NA | NA | Pilot project | Interviews | Algorithm | Germany |
| (Pati et al., 2013) | 68 | 26 - 85 (mean = 46.5) | April - June 2011 | Cross-sectional study | Questionnaire, Interviews | Mixed | India |
| (Compagna et al., 2014) | 76 | > 70 | April 2010 - Oct 2013 | Comparison | | Quantitative | Italy |
| (Esteva et al., 2014) | 777 | >18 | Sep 2006 - Sep 2008 | Cross sectional study | Interviews, Checklist | Quantitative | Spain |
| (Huang et al., 2014) | NA | NA | NA | Event logs | | Algorithms | China |
| (Redaniel et al., 2015) | NA | NA | March 2013 - Feb 2014 | | Interviews | Qualitative | UK |
| (Wolf et al., 2015) | 13774 | > =50 | Jan 2006 - June 2012 | | Surveys | Quantitative | USA |
| (Næser et al., 2018) | 938 | ≥ 18 (median = 70) | July 2012 - Sep 2014 | Cohort study | | | Denmark |
| (Larentzakis et al., 2019) | 286 | mean = 57.7 | 2002 - June 2015 | Database | | Quantitative | UK |
| (Falborg et al., 2020) | 4502 | ≥ 40 (median = 66) | 2013 - 2015 | | | Quantitative | UK, Canada, Denmark, Sweden, Norway, Australia |
| (Joris et al., 2020) | 302 | 17 - 69 (53.8) 70 - 90 (76.8) | Nov 2015 - Aug 2018 | | | Quantitative | Belgium |
| (Klinkhammer-Schalke et al., 2020) | 220 | ≥ 18 | Jan 2014 - Oct 2015 | RCT | | Quantitative | Germany |
| (Muller et al., 2020) | 405635 | 15 - 99 | Jan 2008 - Dec 2013 | | | Quantitative | UK |
| (Rua et al., 2020) | 173 | ≥ 40 | NA | CT | Observations | Quantitative | UK |

N = number of participants; NA = not available

## 3.2 Factors Influencing Care Pathways for Breast and Prostate Cancer in a Hospital Setting – Beacon Hospital Case Study

### 3.2.1 Materials and Methods

*Study Design*

The study was conducted at Beacon Hospital, a private hospital in Dublin, Republic of Ireland. There were 117 patients selected to be contacted to participate in the study. In total, 83 patients agreed to participate and were interviewed, 41 BCa patients and 42 PCa patients. See *Table 11* for a complete breakdown of all the participants' involvement thought out the study.

***Table 11.*** The patients selected and contacted for the study.

| | **No. Patients** |
|---|---|
| Total number of patients selected to participate in the study | 117 |
| Patients contacted | 109 |
| Patients participated | 83 |
| Breast cancer patients | 41 |
| Prostate cancer patients | 42 |
| Patients who did not want to participate | 18 |
| Patients who withdrew | 1 |
| Patients who forgot about the interview appointment | 1 |
| Patients who rescheduled the interview beyond the study period | 1 |
| Patients who didn't feel well enough to participate | 4 |
| Patients who didn't join because of the length of the interview | 4 |
| Patients who didn’t participate and didn’t give any explanation | 7 |
| Patients who were told about the study but weren’t interviewed because the number of participants was reached | 8 |
| Patients who were selected but not contacted | 8 |

The inclusion criteria were the following: participants ought to be over the age of 18 years; ought to have the capacity to provide informed consent themselves; their current diagnosis ought to be breast cancer or prostate cancer of any cancer stage and care period; they were able to understand and speak English, and they were willing to participate in the study. Exclusion criteria: participants who had been involved in other research projects for the last 8 weeks (2 months) were excluded from the study. This exclusion criterion was included to not burden the patients.

### *Materials*

Demographic, medical, and lifestyle data of the participants were collected through one-on-one interviews with patients performed by the principal investigator (who had no prior knowledge about any of the participants) and electronic health records (EHRs). EHRs were retrieved manually for each patient from both systems used in the hospital. The difference between these systems was one system was used only in the radiotherapy department (EHS2) and stored only information related to radiotherapy treatment, and the other one was used in the rest of the hospital (EHS1). After the categorization of the data retrieved from the interviews and the creation of the EHRs dataset, they were both combined to form two separate datasets, one for breast cancer and another for prostate cancer (*Figure 14*).

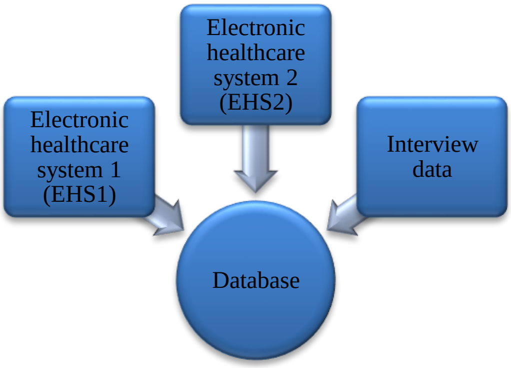

Electronic healthcare system 2 (EHS2)
Electronic healthcare system 1 (EHS1)
Interview data
Database

***Figure 14.*** The dataset contains information gathered from two electronic healthcare systems and the qualitative study conducted between January and November 2018.

### *Demographic Data*

Demographic data included the participant’s current age, age at diagnosis, date of birth, province, marital status, education, employment, and religion. The information for the first 3 variables, age, age at diagnosis, and date of birth, was always found in EHRs. The information about the rest was retrieved through interviews. Marital status and religion could be found in EHRs, although most of the time missing. Participants’ demographic characteristics are presented in *Table 12*.

***Table 12.*** Participant characteristics.

| Characteristics | Breast cancer | Prostate cancer |
|---|---|---|
| Sex | | |
| Male | 0 (0%) | 42 (100%) |
| Female | 41 (100%) | 0 (0%) |
| Age (years) | | |
| Median (range) | 61 (33 - 83) | 74 (46 - 90) |
| Median on diagnosis (range) | 58 (33 - 81) | 66 (38 - 86) |
| Education | | |
| No education | 0 (0%) | 2 (4.8%) |
| Primary school | 0 (0%) | 5 (11.9%) |
| Secondary school | 12 (29.3%) | 4 (9.5%) |
| Professional certificate | 1 (2.4%) | 6 (14.3%) |
| Bachelor’s degree | 20 (48.8%) | 9 (21.4%) |
| Master’s degree | 8 (19.5%) | 1 (2.4%) |
| PhD | 0 (0%) | 4 (9.5%) |
| No information | 0 (0%) | 11 (26.2%) |
| Work | | |
| Employed full time | 18 (43.9%) | 6 (14.3%) |
| Employed part-time | 1 (2.4%) | 2 (4.8%) |
| Unemployed | 2 (4.9%) | 2 (4.8%) |
| Retired | 19 (46.3%) | 25 (59.5%) |
| No information | 1 (2.4%) | 7 (16.7%) |
| Marital status | | |
| Single | 5 (12.2%) | 3 (7.1%) |
| Married | 24 (58.5%) | 34 (81.0%) |
| Partnership | 2 (4.9%) | 0 (0%) |
| Widowed | 6 (14.6%) | 5 (11.9%) |
| Divorced | 1 (2.4%) | 0 (0%) |
| Unmarried | 3 (7.3%) | 0 (0%) |
| Provinces | | |
| Connacht | 0 (0%) | 2 (4.8%) |
| Leinster | 40 (97.6%) | 37 (88.1%) |
| Munster | 1 (2.4%) | 1 (2.4%) |
| Ulster | 0 (0%) | 2 (4.8%) |
| Religion | | |
| Not religious | 7 (17.1%) | 2 (4.8%) |
| Religious | 25 (61.0%) | 30 (71.4%) |
| Not disclosed | 9 (22.0%) | 10 (23.8%) |
| Insurance | | |
| Private insurance | 24 (58.5%) | 37 (88.1%) |
| Self-pay | 17 (41.5%) | 5 (11.9%) |

### *Medical Data*

Medical data were collected mainly from EHRs in both systems (EHS1 and EHS2). That included hearing, vision, allergies, the diagnosis plan, date of biopsy, biopsy results: type of cancer, grade, stage, progesterone receptor and the score, estrogen receptor and the score, HER2 receptor and the score, tumor size, lymph node involvement, Oncotype score; treatment lines including surgery date, type of surgery, side of surgery; chemotherapy drug, the number of cycles, start and end date of chemotherapy, the status of chemotherapy; radiation site, number of sessions of radiotherapy, the gray unit, start and end date of radiation treatment and status; endocrine therapy drug, dose, start and end date of endocrine therapy, and status; targeted therapy drug, number of cycles, start and end date, and status; immunotherapy drug, number of cycles, start and end date, and status; bisphosphonate drug, number of cycles, start and end date, and status; care phase during the interview and comorbidities when the data was collected. All the information about the medical aspect was collected consulting different sections in both systems and sometimes even the participant's patient folder. This was done to get the correct number of chemotherapy cycles and endocrine therapy treatment because it was found that this information was sometimes missing from the EHRs. Participants were asked about their treatment during the interviews to have their views on treatment as well.

*Lifestyle Data*

Lifestyle data was collected from the interviews. This included diet, exercise, smoking, alcohol consumption, support system. Such information was sometimes stored indirectly in EHS1 for a limited number of participants, mainly prostate cancer patients. Financial information was retrieved in the form of insurance they used during their care in the hospital. Due to some participants being in their follow-up phase, the insurance information was changed to Self-pay, and the exact insurance coverage during their treatment is not known. However, as the hospital is a private one, all participants had private insurance. Participant's family history with cancer was collected, and this information was mainly retrieved from interviews. Breast cancer participants were asked about their parity, as nulliparity is one of the risk factors for breast cancer (Fioretti et al., 1999).

During the interviews, participants were asked to elaborate specifically on their diet, alcohol intake, smoking, and exercise habits. The Centre for Disease Control categories were used to group the smoking habit (Centre for Disease Control and Prevention, 2021) as in "never smoker," "former smoker," "current smoker." Alcohol consumption was categorized according to the National Institute on Alcohol Abuse and Alcoholism (National Institute on Alcohol Abuse and Alcoholism, 2021) as "never drinker," "social drinker" (moderate drinker), and "heavy drinker." Diet and exercise habits were found to be more complex to categorize. Diet was categorized into poor, moderate, and healthy, and it was calculated based on the description of a participant’s typical day. Participants were asked about their diet before and after the cancer diagnosis. Participants who expressed having a diet consisting mainly of ready meals and fast food were considered in the poor diet category. Participants with a mixed diet, a diet that consisted of occasional vegetables and fruits, fish or meat, and once per week of ready meals or fast food, would be considered in the moderate diet category. Participants who would have a mixed diet with a variety of food consisting mainly of vegetables, fruits, fish, less or no meat would be considered in the healthy diet category. These participants would sometimes grow their own vegetables and were recorded to follow some dietary programs designed for cancer patients after their diagnosis. Participants would describe these dietary programs as eliminating sugary food, animal-based food such as dairy and meat. The exercise habit was grouped into “sedentary,” “low,” “moderate,” and “high” activity. Participants were asked about their daily activity before and after the cancer diagnosis. The sedentary category was described as no activity during the day apart from walking short distances within the house. The low category was described as little activity in the form of walking and/ or golfing once a week or gardening. The moderate category was described as going for long walks more than twice per week, running or going to the gym at least once per week, or swimming. The high category was described as doing more than one sport during a weeks’ time. Usually, these participants would be everyday runners, would go to the gym or do water aerobics, long-distance cycle, go hiking, besides the other activities mentioned in the abovementioned categories.

### *Methods*

Descriptive statistics were used to summarize and initially inspect the distributions of the study variables. IBM SPSS Statistics version 26 was used to analyze the data. The Kruskal-Wallis test was used for numerical variables and Pearson’s chi-square test for categorical variables. Missing observations were excluded using the SPSS option.

### *Ethics statement*

This research was approved by the Research Ethics Committee at Beacon Hospital, study reference number BEA0084, and approval date January 22, 2018, and the Research Ethics Committee at the University of Deusto, study reference number ETK-08/17-18 and approval date October 18, 2017.

# 3.3 Results

## 3.3.1 Breast Cancer Study

The total number of variables in the breast cancer dataset is p = 163 (see *Table C.1*). The total number of distinctive participants is 41, 4 of them had a cancer recurrence, and one was diagnosed with 2 different types of breast cancer, invasive ductal carcinoma and invasive lobular carcinoma. Some variables, p = 38, were removed from the analysis because they presented with more than 90% of the data missing. The *plan for the staging* variable represents the triple assessment care pathway that includes a mammogram, an ultrasound, and a biopsy examination. Some of the patients had other examinations performed during the diagnosis stage, such as MRI and CT scans. One of the reasons being breast density. Research has shown that for patients with dense breasts, a supplemental MRI is needed (Bakker et al., 2019). The data indicates that 8/46 patients underwent more examinations than stated in the triple assessment care pathway. The care pathway for non-metastatic breast cancer followed at Beacon Hospital can be found in *Table 13*.

***Table 13.*** Care pathway for non-metastatic breast cancer patients.

| |
|---|
| **Pre diagnosis**<br>• Hospital referral |
| **Diagnosis**<br>Breast clinic appointment<br>• Triple assessment (TA): physical examination, mammogram, ultrasound, biopsy (same day)<br>• MRI a few days after TA if dense breast or other reasons<br>Results<br>• Biopsy results and scans<br>• Verbal and written information about cancer<br>• A preliminary care plan, awaiting the receptor status results, described to the patient during clinic visit defining the first line of treatment<br>• Results for receptor status: progesterone, estrogen, and HER2 status defining the course of treatment<br>• Contact information at the hospital<br>• Information about the importance of exercise<br>• Information about fertility for young adult patients |

**Treatment**

- 6 weeks rest between each treatment
- Information about the importance of exercise
- Information about fertility for young adult patients

Chemotherapy

- Information about chemotherapy
- Chemotherapy side effects and how to mitigate them

Surgery

- Information about the surgery
- Length of stay at the hospital after surgery
- Physiotherapy sessions: to recover from surgery and to prepare for radiotherapy

Radiotherapy:

- Information about radiotherapy and its side effects
- Radiotherapy measurements
- Contact information at the radiotherapy department

Endocrine therapy:

- Information about the endocrine treatment and its side effects

**Follow-up**

- Six-month follow-up after the surgery
- Six-month follow-ups with the medical oncologist
- Yearly mammogram
- Yearly appointments with the gynecologist

Before the treatment started, 36/46 went through further examinations, the most common being CT TAP (27) and MRI (18), followed by the Nuclear Medicine bone scan (11).

The average waiting time to start the treatment was 20.65 days; 78.3% (n=36) started their treatment within 31 days (one month). The other 10 participants began their treatment within 50 days after their diagnosis. Of the 10 participants, 6 were diagnosed with metastasized breast cancer. Four out of 6 had cancer metastasized to bones, one in the liver, and the other patient was initially diagnosed with cancer of unknown primary metastasized to the breast. The prolonged start of treatment was due to further examinations to see the extent of the metastasis. However, there were cases where the delay was requested by patients.

The analysis was conducted to see how different lines of treatment are affected by the collected input variables. For this reason, the following tests were run to see which of the input variables played a role if a BCa patient had chemotherapy, radiotherapy, surgery, targeted therapy, immunotherapy, and bisphosphonate.

*Chemotherapy*

The most common chemotherapy combinations were doxorubicin, cyclophosphamide, and docetaxel or paclitaxel (AC + T), docetaxel or paclitaxel, carboplatin and trastuzumab (TCH), and cyclophosphamide, methotrexate, and fluorouracil (CMF). The test showed that there was a statistically significant difference in patient's age at diagnosis (95% Confidence Interval (CI) 51.39 – 58.92, p = 0.047) and hearing (95% CI -0.17 – 0.42, p = 0.029) between the group that was treated with chemotherapy and the other one that did not. The age at diagnosis showed that patients aged 65 (mean) were less predisposed to have chemotherapy than those aged 55 (mean).

*Targeted therapy*

The test showed that there was a statistically significant difference in the patient's HER2 score (95% CI 1.99 – 2.92, p < 0.01), chemotherapy cycles (95% CI 4.22 – 8.38, p = 0.012), and the smoking habits (95% CI 0.20 – 0.50, p = 0.038). Targeted therapy is especially administered when the patient is diagnosed with HER2-positive breast cancer, and the analysis confirmed the same. It was observed that patients who did not have targeted therapy had never smoked, and patients who had targeted therapy tend to have more chemotherapy cycles.

*Surgery*

The test showed that there was a statistically significant difference in patient's diet (95% CI 1.46 – 1.85, p = 0.016), exercise (95% CI 1.20 – 1.96, p = 0.029), chemotherapy cycles (chemotherapy cycles 1: 95% CI 5.06 – 6.33, p = 0.007 and chemotherapy cycles 2: 95% CI 5.06 – 6.33, p = 0.005). Patients who had moderate to a healthy diet and moderate to high active lifestyle showed to have undergone surgeries compared to patients who had self-reported being less active and had a poorer diet.

*Endocrine therapy*

In this cohort study, the majority of patients started their endocrine therapy treatment after they had successfully completed other treatments such as surgery, chemotherapy, and/ or radiotherapy treatments. Endocrine therapy is usually prescribed to be taken for a period of 3 to 5 years. The analysis showed that there was a statistically significant difference in patient's years with cancer (95% CI 1.63 – 4.37, p = 0.016). This treatment line is usually prescribed to patients who have been diagnosed with hormone-positive receptors, the progesterone receptor score (95% CI 3.08 – 6.447, p = 0.016) and estrogen receptor score (95% CI 5.19 – 7.22, p < 0.001). A relation is observed between the chemotherapy cycles and endocrine therapy. The more chemotherapy cycles a patient had, the more chances that the patient had to continue with endocrine therapy (95% CI 4.77 – 13.59, p = 0.044).

*Radiotherapy*

Radiotherapy is another common treatment for breast cancer patients, besides chemotherapy and surgery. Patients with a body mass index above normal (BMI > 25) were observed to have more radiotherapy treatments compared to patients with a smaller BMI (BMI: 95% CI 25.63 – 30.41, p = 0.007). Patients with no radiotherapy treatment were observed to have more chemotherapy cycles (95% CI 3.85 – 16.65, p = 0.049).

*Bisphosphonate*

Bisphosphonate is a medication used for treating bone diseases. In the case of breast cancer, it is used when the disease has metastasized to bones. The cancer stage is one of the indicators if the disease has metastasized or not, starting from 0 to 4, 0 being non-invasive BCa. The test showed that there was a statistically significant difference in the patient's cancer metastasis (stage: 95% CI 4.00 – 4.00, p = 0.022). Patients that underwent bisphosphonate treatments had more chemotherapy cycles (above 8 cycles of chemotherapy) compared to patients that did not have bisphosphonate treatment (these patients usually had less than 8 chemotherapy cycles) (chemo_cycles1: 95% CI 5.19 – 6.60, p = 0.012; chemo_cycles2: 95% CI 4.99 – 7.51, p = 0.028).

The most common first line of treatment was chemotherapy (n=19) followed by surgery (n=18). The other treatments were less common, endocrine therapy (n=4) and radiotherapy (n=3). The most common second line of treatment was chemotherapy (n=18) and targeted therapy (n=14). The most common third line of treatment was radiotherapy (n=11), target therapy (n=9), and surgery (n=8). The most common fourth line of treatment was endocrine therapy (n=11) and radiotherapy (n=9). Metastasized BCa patients had more than 5 lines of treatments. In total 32 participants had surgery, chemotherapy (n=38), targeted therapy (n=25), radiotherapy (n=28), endocrine therapy (n=26), bisphosphonate (n=11), and immunotherapy (n=1).

### 3.3.2 Prostate Cancer Study

The total number of variables in the prostate cancer dataset was p = 77 (see *Table C.2*), and the number of patients was n = 42. Twenty-nine participants were diagnosed with metastatic PCa, 22 of them with bone metastasis. The cancer had spread to other organs such as the bladder, pelvis, liver, mesorectum, and lumbar spine. The information on the metastasized organ for 5 participants was missing.

There was no care pathway for PCa. PCa patients were diagnosed together with other urologic diseases in the Urology department. The plan for staging was very different from one participant to the other. The information for 6 participants was missing; however, 24 participants had undergone at least 3 examinations (n = 36 at least one, and n = 32 at least 2 examinations). The three most common examinations were biopsy (27/42), MRI (21/42), and nuclear bone scan (15/42). Before the treatment started, 19/42 went through further examinations, the most common being CT (10) and MRI (6).

The average waiting time to start the treatment was 19.5 days; 78.6% (n = 33) started their treatment within 31 days (one month). The other 9 participants began their treatment within 4 months after their diagnosis. Of the 9 participants, 6 were diagnosed with metastasized PCa. Four out of 6 had cancer metastasized to bones.

*Endocrine therapy*

Endocrine therapy was the most common treatment for PCa patients in the hospital. Every patient had endocrine therapy. For this reason, no significant difference was observed to play any role in this treatment.

*Chemotherapy*

Docetaxel was the only chemotherapy drug used to treat prostate cancer patients. It was observed that taller (height_cm: 95% CI 176.70 – 183.85, $p = 0.005$), heavier (weight_kg: 95% CI 85.80 – 99.57, $p < 0.001$) patients had chemotherapy compared to the ones who were shorter, lighter, and with BMI less than 25.

*Radiotherapy*

Radiotherapy is a widespread treatment for different types of PCa, and especially for prostate cancer that has metastasized. In such cases, patients who had some discomforts would undergo 1-2 sessions to manage the symptoms. The results showed that patients who were diagnosed with PCa for more than 2 years underwent radiotherapy treatments compared to those who were diagnosed in less than 2 years (95% CI 3.19 – 6.49, p = 0.003). There were some outliers, but these patients were in their follow-up phase, meaning they had completed their treatment. As in other treatments, height and weight variables showed to be significant. Shorter patients underwent radiotherapy treatments, compared to taller ones (95% CI 170.42 – 176.53, p = 0.01). Heavier patients did not undergo radiotherapy compared to the ones who weighed less than 83 kg (95% CI 73.69 – 84.92, p = 0.007).

*Bisphosphonate*

This drug is usually administered in patients who are diagnosed with metastatic cancer. In this study, the analysis showed that patients with an initial prostate-specific antigen level (PSA level) above 25 are treated with this drug (initial_psa: 95% CI 45.51 – 96.14, p = 0.002).

*Radium Ra 223 dichloride*

Radium Ra 223 dichloride is a drug used to treat prostate cancer that has spread to the bone and is causing symptoms but has not spread to other organs. It is used in patients whose cancer is castration-resistant (cancer that keeps growing even when the amount of testosterone in the body is reduced to very low levels). The analysis showed that patients with poor diet tend to be the ones who had Radium Ra 223 dichloride treatments (95% CI -0.51 – 1.71, p = 0.057).

*Surgery*

This treatment was the less preferred treatment. Many opted for less invasive treatments such as endocrine therapy and chemotherapy. Although the number of participants in the study was small (n = 42), the analysis showed that participants who had been diagnosed with prostate cancer within 5 years (95% CI 2.72 – 10.95, p = 0.015) and had a sedentary and low active lifestyle (95% CI -0.21 – 0.88, p = 0.02) were more predisposed to undergo surgery.

The most common first line of treatment for PCa participants was endocrine therapy (n=33), followed by radiotherapy (n=6). The most common second-line of treatments were endocrine therapy (n=22) and radiotherapy (n=7). The most common third line of treatment was endocrine therapy (n=14) and bisphosphonate (n=10). The most common fourth line of treatment was endocrine therapy (n=12) and bisphosphonate (n=10). Patients with 5 lines of treatment and more were using other treatments such as chemotherapy, radiotherapy, and Radium Ra 223 dichloride. In total 7 participants had surgery, chemotherapy (n=17), radiotherapy (n=24), endocrine therapy (n=41), bisphosphonate (n=33), and Radium Ra 223 dichloride (n=8).

### 3.3.3 Breast and Prostate Cancer Analysis

Both datasets were combined to see if there was any significant difference between groups. Treatments used by only one group were excluded, i.e., targeted therapy and Radium Ra 223 dichloride. See Supplement for the graphs for each analysis.

Chemotherapy

As it was observed in the groups independently, the age of diagnosis played a role if a patient had chemotherapy or not. Patients above 70 years of age did not have chemotherapy, whereas patients younger than 70 (median age 59) had chemotherapy (95% CI 55.70 – 62.18, p < 0.001). If the patients were diagnosed within the last 2 years, they had chemotherapy (95% CI 1.25 – 2.31, p < 0.001). Patients with higher BMI (BMI > 25) had chemotherapy, compared to those with a BMI < 25 (95% CI 26.02 – 28.96, p = 0.046).

Radiotherapy

No significant differences were observed apart from a tendency of participants with an average height of 170 who were less likely to receive radiotherapy compared to those with a height of less than 170 cm (mean 168.28) (95% CI 165.88 – 170.67, p = 0.076).

Endocrine therapy

Although no significance was observed in the prostate cancer group, the combined analysis showed that the older the patients when they were diagnosed, the more likely they were to have endocrine therapy treatment (95% CI 61.30 – 67.19, p = 0.05) (this was observed even when the current age of the participants was used instead (95% CI 64.28 – 70.48, p = 0.026)). The analysis showed that taller (height 170 cm (average) and above) and heavier (weight 81 kg (average) and above) patients had endocrine therapy compared to shorter and less heavy patients (height_cm: 95% CI 169.42 – 174.10, p = 0.003; weight_kg: 95% CI 75.14 – 83.71, p = 0.044).

Surgery

The test showed that there was a statistically significant difference in the patient’s height. Shorter patients (height 170 and below) underwent surgery compared to taller patients (95% CI 164.15 – 169.38, p = 0.005).

Bisphosphonate

Older patients (average age 66) had bisphosphonate treatment, compared to younger patients (95% CI 62.88 – 69.63, p = 0.006). Similar to other treatments, the test showed that there was a statistically significant difference in patient’s height and weight, with taller and heavier patients having bisphosphonate treatment compared to shorter and less heavy patients (height_cm: 95% CI 164.33 – 169.73, p = 0.003; weight_kg: 95% CI 68.50 – 78.65, p = 0.027).

## 3.4 Discussion and Conclusions

This study’s main aim was to analyze different treatment lines that BCa and PCa patients underwent while being treated at Beacon Hospital. The following treatment lines for BCa were analyzed: chemotherapy, radiotherapy, endocrine therapy, surgery, targeted therapy, and bisphosphonate. The following treatment lines for PCa were studied: chemotherapy, endocrine therapy, surgery, radiotherapy, bisphosphonate, and Radium Ra 223 dichloride. Chemotherapy, radiotherapy, surgery, endocrine therapy, and bisphosphonate were analyzed for both groups. Immunotherapy treatment was used by only one BCa participant, and

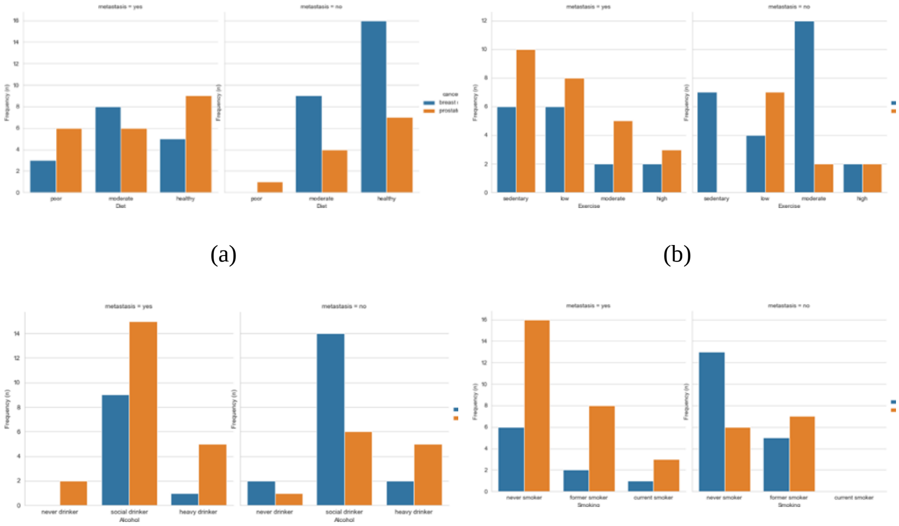
Ipatasertib, an experimental drug, was used by only one PCa participant. For this reason, they were excluded from the analysis.

Our analysis for BCa participants showed the relevance of several lifestyle factors consistent with previous studies (Chajès and Romieu, 2014; Dandamudi et al., 2018; Grosso et al., 2017; Macacu et al., 2015; Mourouti et al., 2015; Zeinomar et al., 2019). Additionally, the number of chemotherapy cycles indicated if a patient underwent other types of treatments such as targeted therapy, surgery, endocrine therapy, radiotherapy, or bisphosphonate. It must be noted that chemotherapy is one of the main treatments for breast cancer. It is used in both non-metastatic and metastatic diseases, compared to surgery which is still under review on metastatic disease (Fitzal et al., 2019; Rashaan et al., 2011).

On the other hand, relations that are proven by the scientific community, such as the relation between the HER2 score and targeted therapy, were noted. Targeted therapy drugs such as trastuzumab and pertuzumab are specifically created to target human epidermal growth factor receptor 2 (HER2 receptor) (Incorvati et al., 2013).

Endocrine therapy drugs, taken in the form of tables, are treatments specifically targeting progesterone and estrogen receptors. These drugs are prescribed to be taken for an extended period of time, 3 to 5 years or even longer periods (Vyas and Kaklamani, 2017); hence there will be a strong relationship between the number of years a patient is diagnosed with cancer and endocrine therapy treatment. According to the data collected from the study for non-metastatic breast cancers, endocrine therapy is the last treatment prescribed to patients. Thus, a newly diagnosed patient will have to go through other treatments first, which usually takes up to 1 year, before starting with endocrine therapy.

Our data showed some relations between lifestyle variables such as diet (*Figure 15* (a)), exercise **(***Figure 15* (b)), alcohol **(***Figure 15* (c)), and smoking **(***Figure 15* (d)). A healthy diet and an active life usually mean a normal BMI. In the data, patients with higher BMI (BMI > 25), especially PCa patients, were observed to be diagnosed with metastatic cancer, *Figure 16*. Studies have shown that, indeed, obese men tend to be diagnosed with advanced PCa, compared to those with normal BMI (Møller et al., 2015). Similar was observed with weight changes. Bigger weight gains were associated with a higher risk of aggressive PCa (Bassett et al., 2012). Recent systematic reviews have shown no clear association between BMI and PCa, but a strong inverse association between BMI and PSA (Harrison et al., 2020).



(a)

(b)

(c)

(d)

***Figure 15.*** Population graphs for BCa and PCa diet (a), exercise (b), alcohol (c), smoking (d) for metastatic and non-metastatic disease.

The same was observed for sedentary behavior; however, sedentary behaviors are considered as modifiable behavior risk factors through the mechanism involving obesity for aggressive PCa (Berger et al., 2019). Seventeen out of 41 BCa participants were diagnosed with metastatic disease, and only 5 of them had undergone surgery. To the best of my knowledge, there is no study published before to see if healthier and fitter patients were more likely to undergo surgeries compared to less healthy and active patients. More studies should be done in this aspect to check if these results are valid in other situations, i.e., more significant study samples, other countries, etc.

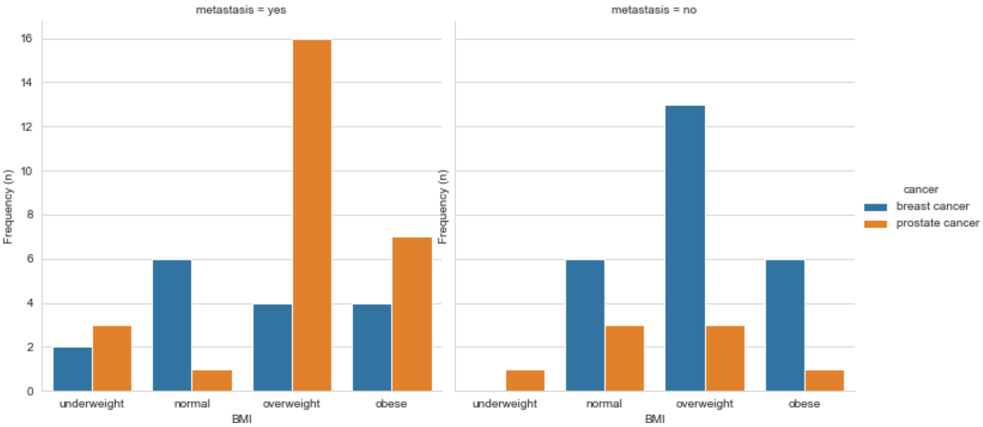


***Figure 16.*** Population graphs for BCa and PCa BMI for metastatic and non-metastatic disease.

In the analysis of the PCa data, height and weight are observed to play different roles when it comes to having chemotherapy and radiotherapy. The relation between height, or tallness, and prostate cancer has been studied for many years. Tall height is associated with an increased risk of high-grade PCa and in PCa mortality (Davies et al., 2015; Khankari et al., 2016; Perez-Cornago et al., 2017; Zuccolo et al., 2008). The analysis showed height and weight play some role if a patient has a specific type of treatment. *Figure 17* shows the height and weight of all participants in the study, indicating that taller and heavier patients are diagnosed with metastatic disease. *Figure 18* shows a comparison of the distribution of participants' age across all BMI groups between BCa and PCa.

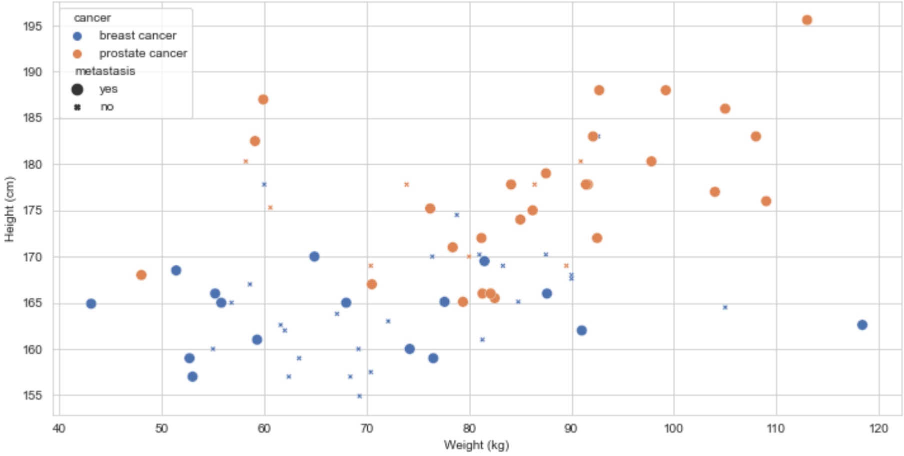


***Figure 17.*** Distribution of height and weight of participants according to the type of cancer and metastatic status.

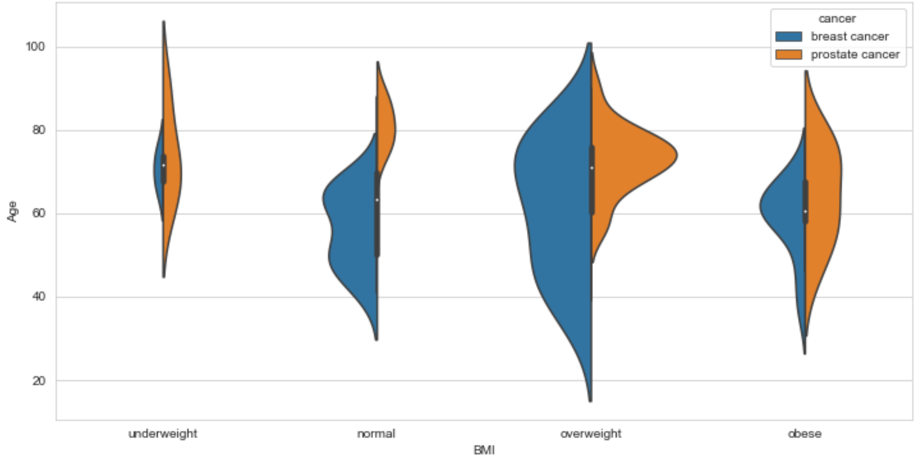


***Figure 18.*** Split violin plots to show the distribution of age across BMI for comparison.

This study presents several limitations. The number of participants was small and represented people with good economic status (all participants had private insurance). Most participants, n = 77, were from the capital region, which does not represent the entire country of the Republic of Ireland. Few of the participants were diagnosed and/or had started their treatment in another hospital before transferring their care to Beacon Hospital. However, these patients were treated by the same doctors, and most of these patients' previous treatments were recorded as well. The start and finish dates of chemotherapy, radiotherapy, endocrine therapy, and bisphosphonate are approximations, not the exact dates. The information about the status of the patient treatment, finished or discontinued, was not available for everyone. The dataset is not balanced. It does not have an equal number of participants between different age groups, cancer metastasis, or between different care phases (treatment, follow-up, palliative).

Studying different types of treatment patient has during their cancer journey is very important. The size of this study is small, and the results may differ if conducted in another country, but it is worth studying what treatments are being used and what role different factors play in these treatments.

## 3.5 Summary

Breast cancer (BCa) and prostate cancer (PCa) are the most prevalent types of cancers. The main aim of this study was to understand and analyze the care pathways for BCa, and PCa patients followed at a hospital setting by analyzing their different treatment lines. The evaluation of the association between different treatment lines and the lifestyle and demographic characteristics of these patients was evaluated as well. Two independent cohorts of patients, one group, diagnosed with breast cancer and another diagnosed with prostate cancer. The information was collected through semi-structured one-on-one interviews with each patient and retrieved data from their electronic health records (EHRs). To conduct the analysis, 125 variables were used for BCa, 77 for PCa, and 39 for the combined group. The variables included demographic, medical, and lifestyle information. Statistical analysis was performed to examine which variable had an impact on the treatment each patient followed. In total, 83 patients participated in the study that ran between January and November 2018 in Beacon Hospital. Chemotherapy was the most common treatment for BCa (38/46). Results showed that chemotherapy cycles indicated if a patient would have other treatments, i.e., patients who had targeted therapy (25/46) had more chemotherapy cycles (95% CI 4.66 – 9.52, p = 0.012), the same was observed with endocrine therapy (95% CI 4.77 – 13.59, p = 0.044). The opposite was observed with radiotherapy, patients who did not have radiotherapy (18/46) had more chemotherapy (95% CI 3.85 – 16.65, p = 0.049). Patients who had bisphosphonate (11/46), an indication of bone metastasis, had more chemotherapy cycles (95% CI 5.19 – 6.60, p = 0.012). Diagnosis age (95% CI 51.39 – 58.92, p = 0.047) showed if a patient would be treated with chemotherapy or not. PCa patients with tall height (95% CI 176.70 – 183.85, p = 0.005), heavier (95% CI 85.80 – 99.57, p < 0.001), and a BMI above 25 (95% CI 1.85 – 2.62, p = 0.017) had chemotherapy compared to patients who were shorter, lighter and with BMI less than 25 who did not have chemotherapy. Patients diagnosed with prostate cancer for more than years underwent radiotherapy treatments compared to those diagnosed in less than 2 years (95% CI 3.19 – 6.49, p = 0.003). Initial prostate-specific antigen level (PSA level) indicated if a patient would be treated with bisphosphonate or not (95% CI 45.51 – 96.14, p = 0.002). Lifestyle variables such as diet (95% CI 1.46 – 1.85, p = 0.016), and exercise (95% CI 1.20 – 1.96, p = 0.029) indicated that healthier and active BCa patients had undergone surgeries. This study shows how different demographic, medical, and lifestyle information affect treatment lines for patients with BCa and PCa. The findings show that chemotherapy cycles and lifestyle for BCa, and tallness and weight for PCa may indicate the rest of these patients' treatment plans. Understanding factors that influence care pathways allow a more person-centered care approach and the redesign of care processes.

# 4

# STUDY 2: MACHINE LEARNING AND ELECTRONIC HEALTH RECORDS – FINLAND

## 4.1 Background

One in three people in Finland will develop cancer at some point during their lifetime (All About Cancer, 2020). Every year, about 30,000 people are diagnosed with cancer. However, only two-thirds will recover from the disease (All About Cancer, 2020). The most common cancer in men in Finland is prostate cancer (PCa) (Finnish Medical Association Duodecim, 2014). In 2018, 5,016 new PCa cases were detected in Finland (Finnish Cancer Registry, 2020); 28% of all new cancers in men. In the same year, 914 men died of PCa, with age-standardized mortality standing at 11.2 per 100,000. PCa patient mortality has remained relatively constant in recent years. By the age of 80, a Finnish man has an 11.6% risk of developing and a 1.6% risk of dying from prostate cancer. The most substantial identified risk factors are age, ethnic background, hereditary susceptibility and environmental factors. Approximately 2–5% of prostate cancers relate to hereditary cancers, and about 15–20% are familial (Bratt, 2002; Grönberg et al., 1996; Pakkanen et al., 2007). A twin Scandinavian study shows that environmental factors play a more significant role in the development of PCa than hereditary factors (Lichtenstein et al., 2000). Excessive consumption of fat, meat and multivitamins may be associated with increased PCa risk (Hori et al., 2011; Patten et al., 2008). Exercise has been found to reduce PCa risk (Liu et al., 2011). Smoking, on the other hand, appears to increase aggressive PCa risk and may also increase its progression (Zu and Giovannucci, 2009).

The relative PCa survival rate one year after diagnosis is 98%: and after five years, 93%. PCa prognosis has remained unchanged over the last ten years (Finnish Cancer Registry, 2020). The 10-year survival forecast for men with local, highly differentiated prostate cancer is the same regardless of treatment (90-94%). Treatments include active monitoring and, if necessary, radical treatments (surgery or radiotherapy), conservative monitoring, and, where needed, endocrine therapy (Finnish Medical Association Duodecim, 2014).

The most common cancer in women in Finland is breast cancer (BCa). In 2018, 4,934 new BCa cases were detected in Finland; 29.8% of all new cancers in women. In the same year, 873 women died of BCa, with age-standardized mortality standing at 12.2 per 100,000 (Finnish Cancer Registry, 2020). BCa patient mortality has remained relatively constant in recent years. By the age of 70, a Finnish woman has an 8.52% risk of developing BCa. The relative BCa survival rate one year after diagnosis is 97.6%: and after five years, 91%. BCa prognosis has slightly improved over the last 15 years (Finnish Cancer Registry, 2020). Among the identified risk factors are gender, age, family history, and hereditary susceptibility, ethnicity, pregnancy and breastfeeding history, weight, alcohol consumption and inactivity. The twin Scandinavian study (Lichtenstein et al., 2000) mentioned above shows that environmental factors play a far more significant role in BCa development than hereditary factors. Only 27 % risk can explain hereditary factors (Lichtenstein et al., 2000). It is worth noting that male breast cancer accounted for just 0.6% of all Finnish BCa in 2018 (Finnish Cancer Registry, 2020), and treatment protocol is mainly based on the principles for female BCa (Mattson and Vehmanen, 2016).

Different drugs are currently in use to treat BCa and PCa, and new ones are frequently clinically trialed. Such treatments include chemotherapy, radiotherapy, endocrine therapy, surgery and, more recently, targeted therapy and immunotherapy. These treatments are administered in combination with each other to cure or keep the disease at bay.

Previous studies have been conducted on predicting the risk of developing BCa and PCa. However, they differ substantially with regard to the different types of information used to make such predictions. In the case of BCa risk prediction (Stark et al., 2019), machine learning (ML) models are developed using Gail model (MDCalc, 2020) inputs only, and models using both Gail model inputs and additional personal health data relevant to BCa risk. Three out of six of the ML models perform better when the additional personal health inputs are added for analysis, improving five-year BCa risk prediction (Stark et al., 2019). Another study assesses ML ensembles of preprocessing methods by improving the biomarker performance for early BCa survival prediction (Gong et al., 2018). The dataset used in this study consisted of genetic data. It concludes that a voting classifier is one way of improving single preprocessing methods. In (Thakur et al., 2018), the authors develop an automated Ki67 scoring method to identify and score the tumor regions using the highest proliferative rates. The authors state that automated Ki67 scores could contribute to models that predict BCa recurrence risk. As in (Gong et al., 2018), genetic inputs, pathologic data and age are used to make predictions.

In the case of PCa risk predictions, (Sapre et al., 2014) show that microRNA profiling of urine and plasma from radical prostatectomy could not predict if PCa is aggressive or slow-growing. Besides RNA data, clinical and pathological data are used to train and test ML. The authors of (Ankerst et al., 2008) add the PCa gene-three biomarker to the Prostate Cancer Prevention Trial risk calculator (PCPTRC), thereby improving PCPTRC accuracy. (Ankerst et al., 2018) is an updated version of the PCPTRC calculator. A recent study in the USA on utilizing neighborhood socioeconomic variables to predict time to PCa diagnosis using ML (Lynch et al., 2020) shows that such data could be useful for men with a high risk of developing PCa.

This chapter presents the results of a study that included Electronic Healthcare Records (EHRs) of breast and prostate cancer patients in a region in Southwest Finland. EHRs are the systematized collection of electronically-stored patient and population health information in digital format. Information stored in such systems varies from demographic information to all types of treatments and examinations that patients undergo throughout the course of their care. This information usually lacks structure or order and requires thorough data cleaning prior to conducting any meaningful analysis. The social impact of analyzing such data is enormous. Understanding the most important variables for a particular disease helps hospitals allocate resources and also helps healthcare professionals individualize care pathways for each patient. Patients thus benefit from a better quality of life. This study aims to determine the most critical variables impacting BCa and PCa patient survivability and how the use of ML models can aid prediction.

## 4.2 Materials and Methods

### 4.2.1 Study Design

A retrospective cohort study (Barrett and Noble, 2019) is conducted using the EHRs of BCa and PCa patients treated at the District of Southwest Finland Hospital via the Turku Centre for Clinical Informatics (TCCI). TCCI provided the Data Analytics Platform (DAP), a remote server where data is accessed and analyzed via a secure shell (SSH) connection.

No ethical approval was required. Nonetheless, it was necessary to apply for authorization to use the data in compliance with privacy and ethical regulations under Finnish law. This study includes anonymized patient data only.

### 4.2.2 Materials

The BCa and PCa data is stored in a PostgreSQL database engine in 24 separate tables according to treatment or the department where the information was collected in the hospital. Structured Query Language (SQL) is utilized to retrieve data for each treatment line (e.g., chemotherapy, radiotherapy, etc.) for both cancers separately and then each file is stored in CSV format. This approach is selected because the data is unstructured and thorough data cleaning and preprocessing are conducted prior to analysis. In total, there are 20,006 individual patients aged 19 – 103, of whom 9,998 are female and 10,008 male. Of 20,006 patients, 9,922 are diagnosed with prostate cancer and 10,113 with breast cancer; 115 are male, 86 of whom are diagnosed with breast cancer only. The database contains information dating from January 2004 (when the regional repository was initially created) until the end of March 2019.

### 4.2.3 Data

The variables collected in this study are primarily based on the Beacon Hospital study, a mixed-method study aiming at understanding breast and prostate cancer patients' care journeys from their perspective. As mentioned in **Chapter 3**, the data is collected using qualitative methods and EHRs. Hospitals, however, do not collect the kind of data retrieved through qualitative methods in their electronic healthcare systems. An explanation of the type of data available and retrieved from the TCCI is given below.

*Demographic Data*

Demographic data includes the patient's current age, age at diagnosis, date of birth, date of death, and years suffering from cancer from the first date of diagnosis. Although patient residence details are collected as part of the study, they do not form part of the analysis.

*Medical Data*

Medical data includes biopsy results: cancer type, grade, Gleason score, progesterone receptor score, estrogen receptor score, HER2 receptor score, tumor size, lymph node involvement, Prostate-Specific Antigen (PSA). Treatment lines include chemotherapy drugs, number of cycles, chemotherapy start, and finish date; the number of radiotherapy sessions, doses delivered, fractions delivered, radiation treatment start and finish date; endocrine therapy drugs; targeted therapy drugs; bisphosphonate drugs; comorbidities at the time of data collection.

The World Health Organization International Classification of Diseases (ICD) version 10 (World Health Organization, 1992) codes are employed for each disease. The main categories for BCa ICD10 codes are used such as c50, c50.1, c50.2, c50.3, c50.4, c50.5, c50.6, c50.7, c50.8 and c50.9. This was done because there were some inconsistencies when associating male breast cancer with male patients. Some were stored as being diagnosed with female breast cancer. This variable is dropped for PCa as there is only one ICD10 code – c61. Grade categories are grade 1, grade 2, and grade 3, and the Gleason score is 6 to 10. There are 18 separate categories for tumor size and 15 for lymph node involvement. Anatomical Therapeutic Chemical (ATC) Classification System codes are used to code chemotherapy, endocrine therapy, targeted therapy and bisphosphonate drugs.

### *Lifestyle Data*

Lifestyle data include smoking and alcohol consumption. Other information such as diet, exercise, family history or female nulliparity (Fioretti et al., 1999) is not initially collected by hospitals and is therefore not included in this study. Participant demographic characteristics are shown in *Table 14*, created using tableone (Pollard et al., 2018), a Python library for creating patient population summary statistics.

***Table 14***. Patient characteristics grouped according to gender.

| | | *Missing* | *Male* | *Female* |
|---|---|---|---|---|
| *n* | | | 9881 | 9941 |
| *Age on diagnosis (mean (range))* | | 0 | 70.2 (19 - 101) | 63.0 (20 - 103) |
| *Current age* (mean (range))* | | 0 | 76.4 (25 - 107) | 69.7 (20 - 115) |
| *Diagnosis (ICD10) (n (%))* | c61 | 0 | 9766 (98.8) | |
| | c50.4 | | 29 (0.3) | 4236 (42.6) |
| | c50.9 | | 17 (0.2) | 1486 (14.9) |
| | c50.2 | | 8 (0.1) | 1288 (13.0) |
| | c50.5 | | 5 (0.1) | 827 (8.3) |
| *Years suffering from cancer (mean (std))* | | 0 | 5.7 (4.4) | 6.1 (4.7) |
| *N. of comorbidities (mean (std))* | | 0 | 13.9 (11.0) | 13.3 (10.9) |
| *Residence (n (%))* | TURKU | 355 | 2787 (28.7) | 3109 (31.8) |
| | KAARINA | | 550 (5.7) | 625 (6.4) |
| | SALO | | 550 (5.7) | 623 (6.4) |
| | RAISIO | | 455 (4.7) | 481 (4.9) |
| | NAANTALI | | 339 (3.5) | 370 (3.8) |

** Age when data was retrieved, March 2019*

### 4.2.4 Methods

Machine learning methods are employed for both feature selection and classification. Python (version 3.5.2) (Rossum and Drake, 2009) programming is used to preprocess and analyze data utilizing Python libraries. Besides Python, SQL is used since data was stored in a PostgreSQL server. The main libraries used during the preprocessing stage were Pandas and NumPy, both of which are open-source libraries providing high-performance, easy-to-use data structures and data analysis tools for scientific computing. Matplotlib and Seaborn open-source data visualization libraries are also used. The study uses the scikit-learn (sklearn) library (Buitinck et al., 2013) for machine learning analysis and is conducted on the server provided by TCCI.

Most of the variables are categorical. Hence one-hot encoding is utilized for encoding and preparing data for ML analysis. This is due to the fact that machine learning models do not work with categorical variables.

Train_test_split(), a pre-defined method in the sklearn library, is employed to train and test the models. 75% of the dataset is used for training the models and 25% for testing. The stratify parameter is included to split the data in a stratified fashion using the desired variable to predict survivability as class labels.

The effectiveness of nine machine learning classifiers is assessed when predicting the probabilities that individuals are likely to survive or die within the first 15 years of diagnosis. The nine classifier types are logistic regression (LR), support vector machine (SVM), nearest neighbor, naïve Bayes (NB), decision tree (DT), and random forest (RF). These machine learning models are selected because each model has significant advantages, which could make it the best model to predict survivability/mortality risk based on the inputs chosen during the feature selection stage. Each of these algorithms is described in *Chapter 2*.

The LR, NB, DT, SVM, and KNN models are implemented using the Python scikit-learn package (version 0.23.1) (Buitinck et al., 2013; Pedregosa et al., 2011). The "linear_model.LogisticRegression" function is used for logistic regression, and "naive_bayes.GaussianNB" and "naive_bayes.BernoulliNB" for naive Bayes. The "tree.DecisionTreeClassifier" function is used to create a decision tree, and "ensemble.RandomForestClassifier" to create a random forest classifier. "svm.SVC" implementation is applied with probability predictions enabled, and "svm.LinearSVC" for the support vector machine. The "neighbors.KNeighborsClassifier" model is used for the nearest neighbor and a grid search technique to extract the best parameters for each function.

Finally, all the features/variables used to train the machine learning models are scaled to be centered around 0 and transformed to unit variance since the datasets have features on different scales, e.g., height in meters and weight in kilograms. Rescaling variables is mandatory because machine learning models assume that data is normally distributed. Also, doing so helps to train the models quickly and generalize more effectively (Saleh, 2018). StandardScaler is chosen to scale the data since it is one of the most popular rescaling methods (Saleh, 2018).

## 4.3 Results

This section is structured in two parts. The first explains feature selection, and the second addresses the classification analysis performed in relation to the features selected from part one.

### 4.3.1 Feature Selection

Feature selection is the process of selecting a set of variables that are significant to the analysis to be conducted. The objective of feature selection is manifold: (i) it provides a better understanding of the underlying process generating data, (ii) faster and more cost-effective predictors, and (iii) improves predictor prediction performance (Guyon and Elisseeff, 2003).

There are different techniques to select the relevant variables. The first technique employed is recursive feature elimination (RFE), whose goal is to remove features step-by-step by using an external estimator that assigns weights to features (Pedregosa et al., 2011). The estimator is trained on the initial dataset, which contains all the features. Each feature’s importance is obtained via two attributes: (i) coef_; or (ii) feature_importances_ (Pedregosa et al., 2011). The least important features are eliminated from the current set of features recursively until the set number of features to be selected is reached. The estimators used to perform RFE are logistic regression, stochastic gradient descent, random forest, linear SVM, and perceptron. *Table 15* shows the estimators used in analysis and accuracy for each number of features selected when predicting whether a patient will survive.

***Table 15.*** Feature selection algorithms and accuracy score.

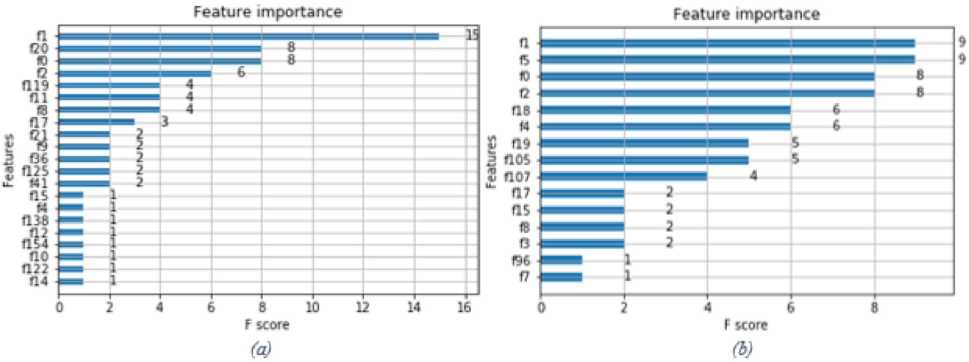
| ***Estimator*** | ***Feature selection*** | ***n_features_to_select/ max_features*** | ***Accuracy Breast*** | ***Accuracy Prostate*** |
|---|---|---|---|---|
| *LogisticRegression(solver='liblinear')* | RFE | 15 | 84.50% | 78.0% |
| *LogisticRegression(solver='liblinear')* | RFE | 25 | 84.70% | 79.3% |
| *LogisticRegression(solver='liblinear')* | RFE | 50 | 85.60% | 79.6% |
| *SGDClassifier()* | RFE | 15 | 73.30% | 67.3% |
| *SGDClassifier()* | RFE | 25 | 73.30% | 67.5% |
| *SGDClassifier()* | RFE | 50 | 82.50% | 77.1% |
| *RandomForestClassifier()* | RFE | 15 | 86.30% | 82.7% |
| *RandomForestClassifier()* | RFE | 25 | 87.50% | 83.4% |
| *RandomForestClassifier()* | RFE | 50 | 87.50% | 83.6% |
| *LinearSVC(C=0.001,max_iter=5000)* | RFE | 15 | 84.20% | 79.4%* |
| *LinearSVC(C=0.001,max_iter=5000)* | RFE | 25 | 84.50% | 79.9%* |
| *LinearSVC(C=0.001,max_iter=5000)* | RFE | 50 | 85.20% | 80.6%* |
| *Perceptron()* | RFE | 15 | 73.30% | 61.2% |
| *Perceptron()* | RFE | 25 | 73.30% | 61.2% |
| *Perceptron()* | RFE | 50 | 74.90% | 64.5% |

** The parameter C was set to 0.01 in the case of prostate cancer data*

Besides RFE, SelectFromModel with a Lasso estimator is used. SelectFromModel is a meta-transformer used alongside an estimator. After fitting, the estimator has an attribute stating feature importance, such as the coef_ or feature_importances_ attributes. In order to control the feature selection algorithms, the same parameters are used to set a limit on the number of features to be selected, the n_features_to_select for RFE and max_features for SelectFromModel.

In order to verify the results obtained from RFE and the SelectFromModel algorithms, the Random Forest Classifier and XGBoost (Chen and Guestrin, 2016) are used. Both these algorithms have a specific attribute to select the best features. The feature_importances_ attribute is used for the Random Forest Classifier and the plot_importance() (Chen and Guestrin, 2016; XGBoost for Python, 2021) method for XGBoost with height set to 0.5 as the parameter. XGBoost is employed on the basis of being an optimized distributed gradient boosting library designed to be flexible, efficient, and portable (Chen and Guestrin, 2016). It uses machine learning algorithms under the Gradient Boosting framework as well as providing parallel tree boosting, which has proven to be highly efficient at solving various problems.

The XGBoost results with the most important features and scores are shown in *Figure 19*. In total, 21 features were selected after running the XGBoost estimator for BCa data and 15 features for PCa data. The results from Random Forest are shown in *Table 16*. All features selected by the algorithms are shown for both BCa and PCa.



*(a)* *(b)*

***Figure 19.*** Feature selection and importance extracted from XGBoost for (a) breast cancer and (b) prostate cancer. Features for both databases are specific to the diseases, and indexes for each feature are different, ex. f0 in the breast cancer dataset represents feature c50_diag_age, whereas, in prostate cancer, it represents c61_diag_age, etc.

Apart from the features shown in *Table 16*, there are six more features (total 21) that were selected but not shown in the table: her2_neg, alcohol_no, alcohol_yes, L02BG04, tumor_size_1, lymph_node_0. All features mentioned above had an F score of at least 1, also shown in *Figure 19 (a)*. All feature indexes refer to the features themselves when shown in *Table 16* and *Table 17*.

***Table 16.*** Features selected using different estimators for breast cancer.

| **RandomForest** | ***XGBoost*** | ***RFE - LR*** | ***RFE - RF*** | ***RFE - LSVC*** | ***SFM - LVC*** |
|---|---|---|---|---|---|
| c50_diag_age | years_cancer_all | c50_diag_age | c50_diag_age | c50_diag_age | years_cancer_all |
| years_cancer_all | c50_diag_age | years_cancer_all | years_cancer_all | years_cancer_all | c50_diag_age |
| nr_comorbidities | doses_delivered | nr_comorbidities | nr_comorbidities | nr_comorbidities | L02BG04 |
| weight | nr_comorbidities | side_left | height | side_left | nr_comorbidities |
| height | height | side_right | weight | side_right | doses_delivered |
| er | er | alcohol_yes | pr | er | L02BA03 |
| pr | L02BA03 | no_smoking | er | alcohol_yes | no_smoking |
| side_right | no_smoking | fractions_delivered | alcohol_yes | no_smoking | alcohol_yes |
| side_left | weight | doses_delivered | no_smoking | doses_delivered | c50_diag_c50.9 |
| her2_neg | cycles | L01CA04 | fractions_delivered | L01CA04 | tumor_size_1b |
| alcohol_no | L01BC06 | L02BA03 | doses_delivered | L02BA03 | side_right |
| her2_pos | L01CA04 | L02BG04 | cycles | L02BG04 | L01CA04 |
| grade_1 | nr_interv_tots | nr_interv_tots | L02BA03 | c50_diag_c50.9 | L01BC06 |
| grade_2 | side_right | c50_diag_c50.9 | L02BG04 | tumor_size_1b | alcohol_no |
| grade_3 | pr | tumor_size_1b | nr_interv_tots | tumor_size_1c | side_left |

***Table 17.*** Features selected using different estimators for prostate cancer.

| **RandomForest** | ***XGBoost*** | ***RFE - LR*** | ***RFE - RF*** | ***RFE - LSVC*** | ***SFM - LVC*** |
|---|---|---|---|---|---|
| c61_diag_age | years_cancer_all | c61_diag_age | c61_diag_age | c61_diag_age | c61_diag_age |
| psa | psa | years_cancer_all | years_cancer_all | years_cancer_all | years_cancer_all |
| years_cancer_all | c61_diag_age | nr_comorbidities | nr_comorbidities | nr_comorbidities | gleason_7 |
| nr_comorbidities | nr_comorbidities | psa | height | psa | nr_comorbidities |
| weight | weight | gleason_6 | weight | gleason_6 | no_smoking |
| height | doses_delivered | gleason_7 | psa | gleason_7 | cycles |
| gleason_7 | cycles | no_smoking | gleason_7 | gleason_9 | gleason_6 |
| gleason_6 | L02BX02 | doses_delivered | alcohol_yes | no_smoking | L02BX02 |
| alcohol_yes | nr_interv_tots | cycles | no_smoking | doses_delivered | L02BX03 |
| has_quit | height | L01XX11 | fractions_delivered | cycles | gleason_9 |
| alcohol_no | gleason_7 | L02AE02 | doses_delivered | L02AE02 | alcohol_yes |
| gleason_9 | no_smoking | L02AE04 | cycles | L02BX02 | doses_delivered |
| gleason_8 | fractions_delivered | L02BX02 | L02BX02 | tumor_size_1c | L02AE02 |
| gleason_5 | gleason_6 | tumor_size_2a | nr_interv_tots | tumor_size_2a | tumor_size_2c |
| gleason_10 | L02AE02 | tumor_size_2c | metastasis_0 | tumor_size_2c | tumor_size_3 |

The final features selected for analysis are shown in *Table 18*. All the features that are included were chosen by at least two estimators, which is shown in the "times" (how many estimators chose the feature) columns for each cancer disease separately.

***Table 18.*** Features selected for breast and prostate cancer data analysis.

| | *Breast* | | *Prostate* | |
|---|---|---|---|---|
| *nr* | **features** | **times** | **features** | **times** |
| *1* | c50_diag_age | 6 | c61_diag_age | 6 |
| *2* | years_cancer_all | 6 | gleason_7 | 6 |
| *3* | doses_delivered | 5 | years_cancer_all | 6 |
| *4* | L02BA03 | 5 | cycles | 5 |
| *5* | L02BG04 | 5 | doses_delivered | 5 |
| *6* | alcohol_yes | 5 | gleason_6 | 5 |
| *7* | side_right | 5 | nr_comorbidities | 5 |
| *8* | L01CA04 | 4 | PSA | 5 |
| *9* | no_smoking | 4 | L02AE02 | 4 |
| *10* | nr_comorbidities | 4 | L02BX02 | 4 |
| *11* | side_left | 4 | no_smoking | 4 |
| *12* | er | 3 | alcohol_yes | 3 |
| *13* | pr | 3 | tumor_size_2c | 3 |
| *14* | alcohol_no | 3 | weight | 3 |
| *15* | height | 3 | fractions_delivered | 2 |
| *16* | tumor_size_1b | 3 | gleason_9 | 2 |
| *17* | weight | 3 | height | 2 |
| *18* | c50_diag_c50.9 | 2 | nr_interv_tots | 2 |
| *19* | cycles | 2 | tumor_size_2a | 2 |
| *20* | her2_neg | 2 | | |
| *21* | L01BC06 | 2 | | |
| *22* | nr_interv_tots | 2 | | |

### 4.3.2 Classification Using Machine Learning

Nine different classification algorithms/estimators are selected for analysis, which is carried out after having chosen the features via the feature selection process. All estimators have several hyperparameters. A GridSearchCV is performed – an exhaustive search over specified parameter values for an estimator – to obtain the best hyperparameters for each algorithm. All parameters and values for each estimator are as follows.

- LogisticRegression parameters:
    - 'penalty': ['l1', 'l2', 'elasticnet'],
    - 'solver': ['lbfgs', 'liblinear', 'sag', 'saga'],
    - 'max_iter': [1000, 3000, 5000]
- LinearSVC and SVC parameters:
    - 'max_iter': [1000, 3000, 5000],
    - 'C': [0.001, 0.01, 0.1]
- SGDClassifier parameters:
    - 'loss': ['hinge', 'log', 'squared_hinge', 'perceptron'],
    - 'alpha': [0.0001, 0.001, 0.01, 0.1],
    - 'penalty': ['l1', 'l2', 'elasticnet']
- KNeighborsClassifier parameters:
    - 'n_neighbors': [3,4,5,6],
    - 'algorithm': ['auto', 'ball_tree', 'kd_tree', 'brute']
- BernoulliNB parameters:
    - 'alpha': [0.1,0.2,0.4,0.6,0.8,1]
- GaussianNB parameters: defaults
- RandomForestClassifier and DecisionTreeClassifier parameters:
    - 'max_depth': [2, 3, 4, 5],
    - 'min_samples_leaf': [0.1, 0.12, 0.14, 0.16, 0.18]

The best value for each hyperparameter is displayed below in *Table 19* for each estimator and disease:

***Table 19.*** Selected best hyperparameters for each type of cancer.

| ***Estimator*** | ***BCa parameters*** | ***PCa parameter*** |
|---|---|---|
| ***LogisticRegression*** | penalty = ‘l2’, solver = 'lbfgs', max_iter = 5000 | penalty = 'l1', solver = 'liblinear', max_iter = 1000 |
| ***LinearSVC*** | C = 0.01, max_iter = 7000 | C = 0.1, max_iter = 5000 |
| ***SVC*** | C = 0.1, max_iter = 3000 | C = 0.1, max_iter = 3000 |
| ***KNeighborsClassifier*** | n_neighbors = 6, algorithm = 'ball_tree' | n_neighbors = 6, algorithm = ‘brute’ |
| ***SGDClassifier*** | alpha = 0.001, loss = 'log' | alpha = 0.01, loss = 'log', penalty = 'l2' |
| ***BernoulliNB*** | alpha = 0.1 | alpha = 0.2 |
| ***GaussianNB*** | default values | default values |
| ***RandomForestClassifier*** | min_samples_leaf = 0.1, max_depth = 4 | min_samples_leaf =0.1, max_depth = 4 |
| ***DecisionTreeClassifier*** | min_samples_leaf = 0.1, max_depth = 4 | min_samples_leaf =0.1, max_depth = 5 |

The Receiver Operating Characteristic (ROC) and AUC metric are used to assess classifier quality. The ROC curve features a true positive rate on the Y-axis and a false positive rate on the X-axis, meaning that the top left corner of the plot is the "ideal" point (a zero false-positive rate and one true positive rate) (Fawcett, 2006). Although the “ideal point” is not realistic, it usually indicates that a larger AUC is preferable. The ROC curve’s "steepness" is also essential since it is ideal for maximizing the true positive rate while minimizing the false positive rate.

Cross-validation is performed for each estimator using scikit-learn StratifiedKFold with the default value of the number of splits set to 5 (5-fold cross-validation). The ROC AUC curve for each estimator with cross-validation for breast cancer is shown in *Figure 20* and in *Figure 21* for prostate cancer.

It can be clearly seen that the support vector machine classifier achieved the best ROC AUC curve for the breast cancer dataset with an area under the curve = 0.83 ± 0.01, followed by KNeighborsClassifier with AUC = 0.82 ± 0.01. Whereas for the prostate cancer dataset, the random forest classifier and KNeighborsClassifier have the best ROC, both yielding AUC = 0.82 ± 0.01.

Conversely, the worst performances for the breast cancer dataset are identified by the following classifiers: Bernoulli Naïve Bayes with ROC AUC = 0.71 ± 0.02, LinearSVC with ROC AUC = 0.72 ± 0.01, and LogisticRegression with ROC AUC = 0.73 ± 0.01. These same classifiers also perform poorly on the prostate cancer dataset, with ROC AUC = 0.64 ± 0.01 for Bernoulli Naïve Bayes, 0.66 ± 0.01 for LinearSVC, and 0.67 ± 0.01 for LogisticRegression. In general, Decision Trees, Random Forest, and Nearest Neighbors perform very well on both datasets with ROC AUC above 0.80.

In addition, ensemble learning is performed using bagging and voting with cross-validation. BaggingClassifier is used for bagging and VotingClassifier for voting. In the case of BaggingClassifier, the number of trees is set to 500, and the KFold cross-validator is used for cross-validation. The ROC-AUC curve for the breast cancer dataset is shown in *Figure 22*, and for the prostate cancer dataset in *Figure 23*.

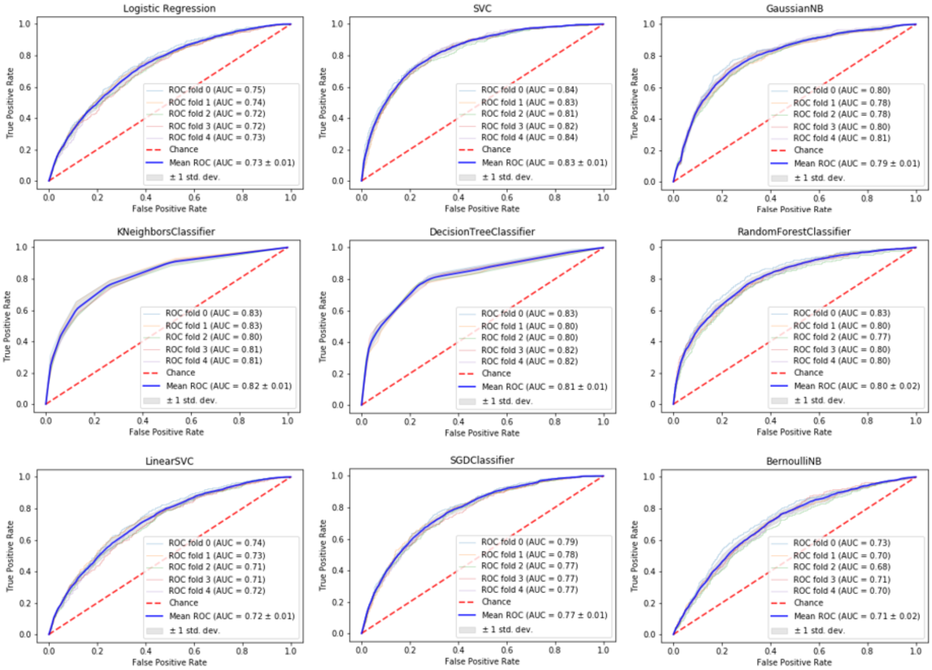


***Figure 20.*** ROC AUC for breast cancer.

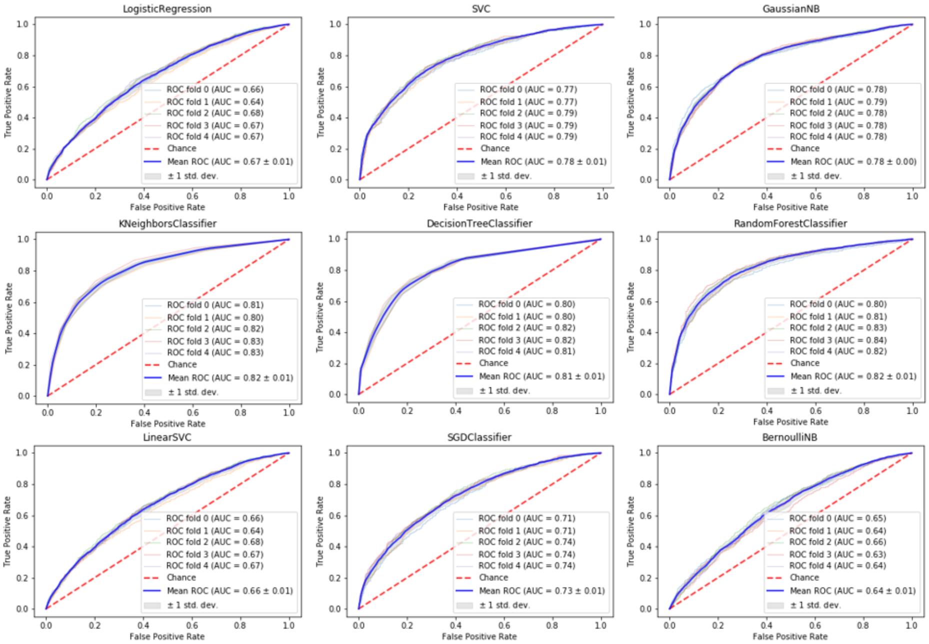


***Figure 21.*** ROC AUC for prostate cancer.

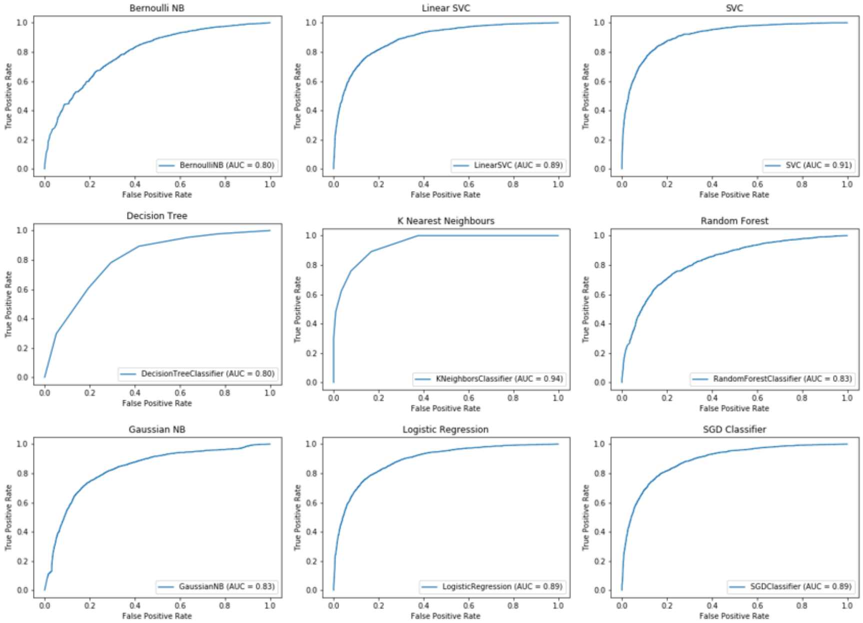


***Figure 22.*** BaggingClassifier for breast cancer dataset.

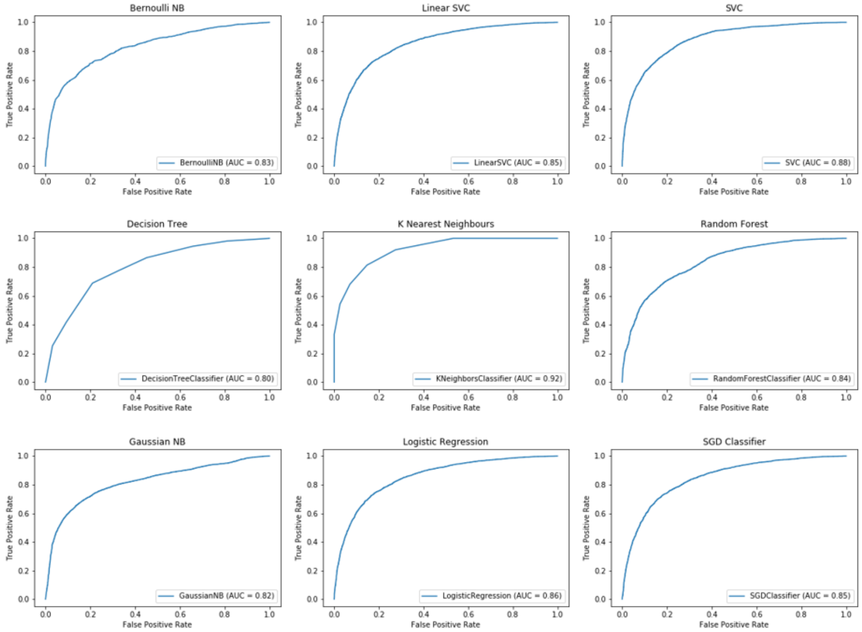


***Figure 23.*** BaggingClassifier for prostate cancer dataset.

As in the previous cross-validation analysis, the best results for BaggingClassifier, in the case of the breast cancer dataset, are yielded by KNeighborsClassifier with a ROC AUC score = 0.94, followed by a ROC

AUC score = 0.91 for SVC. The worst performers are BernoulliNB and DecisionTreeClassifier, both with a ROC AUC score = 0.80. Similarly, in the bagging analysis for the prostate cancer dataset, the best classifiers were KNeighborsClassifier and SVC with ROC AUC scores = 0.92 and 0.88, respectively. Finally, the worst classifiers are DecisionTree and GaussianNB, with ROC AUC scores = 0.80 and 0.82, respectively.

### 4.3.3 Comparing Machine Learning Models

The accuracy score, precision, recall, and F1 score are selected in the training and test sets in order to compare how each model scores when predicting each patient's survivability. Since the problem is a binary classification problem, the results for both classes are presented; the first class, class 0, being patients still alive, and the second, class 1, those who have died. *Table 20* shows the results for the breast cancer dataset and *Table 21* for the prostate cancer dataset. These results are obtained by using the classification_report imported from the sklearn library metrics module.

***Table 20.*** Comparison of machine learning models results for breast cancer dataset.

| ***Estimator*** | ***Accuracy Train*** | ***Accuracy Test*** | ***Precision class 0*** | ***Recall class 0*** | ***F1-score class 0*** | ***Precision class 1*** | ***Recall class 1*** | ***F1-score class 1*** |
|---|---|---|---|---|---|---|---|---|
| *LogisticRegression* | 0.84 | 0.85 | 0.88 | 0.92 | 0.90 | 0.75 | 0.65 | 0.69 |
| *LinearSVC* | 0.84 | 0.85 | 0.88 | 0.92 | 0.90 | 0.75 | 0.64 | 0.69 |
| *SVC* | 0.89 | 0.87 | 0.89 | 0.93 | 0.91 | 0.79 | 0.69 | 0.74 |
| *KNN* | 0.89 | 0.84 | 0.87 | 0.91 | 0.89 | 0.72 | 0.62 | 0.67 |
| *SGDClassifier* | 0.84 | 0.84 | 0.87 | 0.93 | 0.90 | 0.75 | 0.62 | 0.68 |
| *BernoulliNB* | 0.75 | 0.74 | 0.86 | 0.77 | 0.82 | 0.52 | 0.66 | 0.58 |
| *GaussianNB* | 0.80 | 0.79 | 0.82 | 0.92 | 0.87 | 0.67 | 0.43 | 0.52 |
| *RandomForestClassifier* | 0.81 | 0.81 | 0.80 | 0.99 | 0.88 | 0.89 | 0.31 | 0.46 |
| *DecisionTreeClassifier* | 0.81 | 0.79 | 0.80 | 0.95 | 0.87 | 0.73 | 0.37 | 0.49 |
| *VotingClassifier** | 0.83 | 0.83 | 0.84 | 0.95 | 0.89 | 0.79 | 0.49 | 0.60 |
| *VotingClassifier*** | 0.85 | 0.85 | 0.86 | 0.95 | 0.90 | 0.80 | 0.58 | 0.67 |

**non- standardized data*
*** standardized data*

***Table 21***. Comparison of machine learning models results for prostate cancer dataset.

| ***Estimator*** | ***Accuracy Train*** | ***Accuracy Test*** | ***Precision class 0*** | ***Recall class 0*** | ***F1-score class 0*** | ***Precision class 1*** | ***Recall class 1*** | ***F1-score class 1*** |
|---|---|---|---|---|---|---|---|---|
| *LogisticRegression* | 0.79 | 0.78 | 0.80 | 0.86 | 0.83 | 0.75 | 0.66 | 0.70 |
| *LinearSVC* | 0.79 | 0.79 | 0.80 | 0.87 | 0.83 | 0.76 | 0.67 | 0.71 |
| *SVC* | 0.80 | 0.78 | 0.78 | 0.89 | 0.83 | 0.78 | 0.61 | 0.69 |
| *KNN* | 0.83 | 0.78 | 0.78 | 0.88 | 0.83 | 0.77 | 0.61 | 0.68 |
| *SGDClassifier* | 0.79 | 0.78 | 0.79 | 0.86 | 0.83 | 0.75 | 0.65 | 0.70 |
| *BernoulliNB* | 0.77 | 0.77 | 0.81 | 0.83 | 0.82 | 0.72 | 0.69 | 0.70 |
| *GaussianNB* | 0.78 | 0.78 | 0.76 | 0.92 | 0.83 | 0.81 | 0.55 | 0.66 |
| *RandomForestClassifier* | 0.75 | 0.75 | 0.74 | 0.93 | 0.82 | 0.82 | 0.47 | 0.60 |
| *DecisionTreeClassifier* | 0.75 | 0.74 | 0.80 | 0.78 | 0.79 | 0.67 | 0.69 | 0.68 |
| *VotingClassifier** | 0.80 | 0.80 | 0.78 | 0.94 | 0.85 | 0.86 | 0.57 | 0.69 |
| *VotingClassifier*** | 0.80 | 0.80 | 0.80 | 0.90 | 0.85 | 0.80 | 0.64 | 0.71 |

** non- standardized data*
*** standardized data*

In addition, the selected models are trained and tested using the voting technique, with and without data standardization. It is noted that when data standardization techniques are employed, such as StandardScaler(), better results are obtained on all counts for the BCa dataset. However, this is not the case for the PCa dataset. Recall in class 1 and precision in class 2 are slightly worse, but the others either remain unchanged, such as the accuracy scores and F1 score in class 1 or are marginally better.

In general, the algorithms perform better on the breast cancer dataset compared to prostate cancer. One reason could be dataset size; the BCa dataset is slightly larger and more balanced than the PCa dataset. Another reason could be the features. Despite using feature selection algorithms to select the most appropriate variables, other features that were omitted may improve the results.

## 4.4 Discussions and Conclusions

There are multiple variables for each of these two types of cancer. This study sought to analyze which variables are of most importance when predicting patient survivability, or the mortality risk, within the first 15 years of cancer diagnosis. In total, 179 features are included on the breast cancer dataset and 144 on the prostate cancer dataset.

Valid results are obtained by only selecting 15 features after running different feature selection algorithms with different numbers of selected features. In other words, the difference in accuracy achieved by including all 179 features or just 15 features is insignificant.

The selected features are some of the main risk factors of these diseases. In both cancers, it is clear that age at diagnosis and years suffering from cancer are two of the main features that predict whether a patient will survive. Among the selected features, there are few relating to medications and lifestyle (see *Table 22*). Medications for BCa include L02BA03, L02BG04, L01CA04, and L01BC06, and L02AE02 and L02BX02 for PCa.

When attempting to predict the progression of these cancers, it is difficult to make comparisons between studies. This is due to the lack of large, publicly available datasets, the number of records, and the number of variables the datasets contain. Moreover, there is a sheer number of hypotheses that these studies test. This can even be seen in the feature selection algorithms used by various authors. Earlier studies used the F-Score to reduce the number of variables (Akay, 2009; Huang et al., 2008), with more recent studies moving toward more sophisticated algorithms such as random forest (Nguyen et al., 2013) and genetic algorithms (Aličković and Subasi, 2015).

***Table 22.*** Generic names and ATC codes for medication selected during the feature selection process.

| *ATC code* | *Generic name* | *Description* |
|---|---|---|
| *L01BC06* | capecitabine | Chemotherapy drug |
| *L01CA04* | vinorelbine | Chemotherapy drug |
| *L02BA03* | fulvestrant | Endocrine therapy |
| *L02BG04* | letrozole | Endocrine therapy |
| *L02AE02* | leuprorelin | Endocrine therapy |
| *L02BX02* | degarelix | Endocrine therapy |

The database is very comprehensive and covers a wealth of data. This study has endeavored to include as much data as possible in its analytical approach. Nevertheless, laboratory results have not been included. The reason being that blood tests are routinely performed, and results vary depending on the treatment the patient is undergoing. Analyzing the averages of such results would fail to yield any meaningful results. However, other ways of incorporating this information into the analysis are being investigated. Another analysis method could be developing a similar study with different deep learning models and compare these results with the results obtained from the machine learning analysis.

Also, it should be noted that these results are specific to this Finnish population. Each country has its own guidelines and approved medications for certain diseases, so training the same models on a different dataset could deliver different results.

## 4.5 Summary

Breast cancer (BCa) and prostate cancer (PCa) are the two most common types of cancer. Various factors play a role in these cancers, and discovering the most important ones might help patients live longer better lives. This study aims to determine the variables that most affect patient survivability and how the use of different machine learning algorithms can assist in such predictions. The AURIA database was used, which contains electronic healthcare records (EHRs) of 20,006 individual patients diagnosed with either breast or prostate cancer in a particular region in Finland. In total, there were 178 features for BCa and 143 for PCa. Six feature selection algorithms were used to obtain the 21 most important variables for BCa, and 19 for PCa. These features were then used to predict patient survivability by employing nine different machine learning algorithms. Seventy-five percent of the dataset was used to train the models and 25% for testing. Cross-validation was carried out using the StratifiedKfold technique to test the effectiveness of the machine learning models. The support vector machine classifier yielded the best ROC with an area under the curve (AUC) = 0.83, followed by the KNeighborsClassifier with AUC = 0.82 for the BCa dataset. The two algorithms that yielded the best results for PCa are the random forest classifier and KNeighborsClassifier, both with AUC = 0.82. This study shows that not all variables are decisive when predicting breast or prostate cancer patient survivability. By narrowing down the input variables, healthcare professionals are able to focus on the issues that most impact patients and hence devise better, more individualized care plans.

# 5

# STUDY 3: DEEP LEARNING MODELS FOR COLORECTAL POLYPS

Medical imaging has gained immense importance in healthcare throughout history. It has been used in diagnosing diseases, planning treatments, and assessing results. Furthermore, medical imaging is currently used in preventing illness, usually through screening programs. Aggregating it with demographic and other healthcare data can bring novel insights and help scientists discover breakthrough treatments (Esteva et al., 2019).

A lot of research has been done in automating the delivery of medical imaging results. These results still rely on professional radiologists being present when finalizing them. However, automation can help radiologists be more efficient in their job and deliver results quicker.

A review of deep learning (DL) applications in medical imaging (Litjens et al., 2017) shows that AI algorithms will have a significant impact in the healthcare field. The application areas span from digital pathology and microscopy to brain, eye, chest, breast, cardiac, abdomen, etc. These algorithms are for all types of imaging machines used nowadays: computed tomography (CT), ultrasound, MRI, X-ray, microscope, cervigram, photographs, endoscopy/ colonoscopy, tomosynthesis (TS), mammography, etc. Most of these applications deal with classification, segmentation, or detection problems and convolutional neural networks (CNNs), auto-encoders (AE) or stacked auto-encoders (SAE), recurrent neural networks (RNNs), deep belief networks, and restricted Boltzmann machines (RBM) are the most used architectures for these settings. The architecture of some of the most used algorithms is depicted in *Figure 24.*

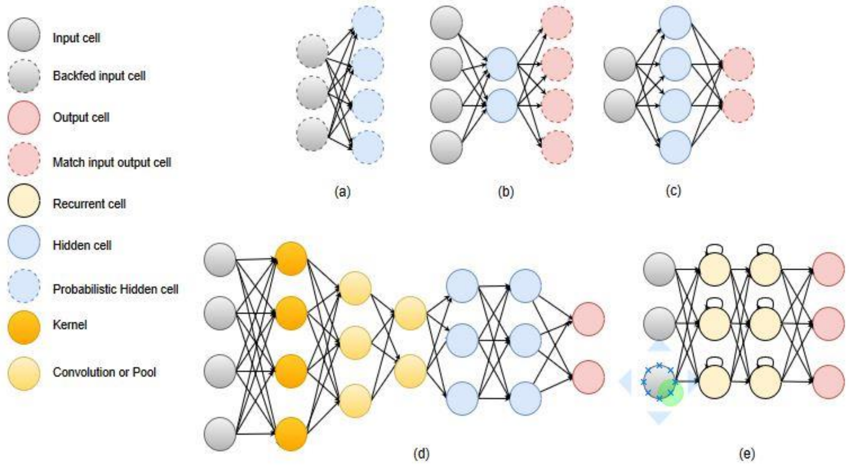


***Figure 24.*** Graph representation of some of the commonly used architectures in medical imaging. (a) AE, (b) RBM, (c) RNN, (d) CNN, (e) MS-CNN.

The focus of this chapter is on colorectal cancer (CRC) and how deep learning algorithms can help detect colon polyps. The World Health Organization, through the International Agency for Research on Cancer, has recognized colorectal cancer as responsible for around 881 thousand deaths, or 9.2% of the total cancer deaths (Cancer Today, 2020). The main concern is that the incidence rates have been rising, with more than 1.85 million cases (Cancer Today, 2020). This increase could be prevented by conducting effective screening tests (Lieberman, 2005). However, a 2020 European study on colorectal cancer shows that total cancer mortality rates are predicted to decline, and these numbers for colorectal cancer are 4.2% in men and 8.3% in women (Carioli et al., 2020). These declines are expected in all age groups (Zauber, 2015). Another study done in the USA shows declining numbers in the USA as well (Siegel et al., 2020). The implementation of screening programs is an essential factor in the declining numbers various countries have seen. Colonoscopy is the preferred technique among the used screening tests to diagnose CRC. It is also used as a prevention procedure for CRC. CRC starts as growth in the lining of the colon or rectum. These growths are called polyps. Polyps are benign neoplasms; some types can transform into CRC over the years. Within the latter are adenomatous polyps and serrated polyps. Not all polyps develop into CRC. The adenomatous colon polyps (adenomas) and polyps larger than 1 cm have a higher risk of malignancy. Sometimes polyps are flat or hide between the folds of the colon, which makes their detection difficult.

One of the procedures to screen for colon polyps is the colonoscopy, which examines the large bowel and the distal part of the small bowel with a camera. The advantages of this procedure include visualization of the polyps and their removal before they grow bigger and, for biopsy purposes, if the medical personnel suspect a cancerous polyp. According to (Siegel et al., 2020), colonoscopy is very well established as a procedure to prevent the development of CRC playing a significant role in rapid declines in incidence cases during the 2000s but not so much during the recent years. Another study on the impact of CRC screening mortality found that using colonoscopy indicates a more than 50% decline in CRC mortality (Zauber, 2015). Although colonoscopy has shown meaningful improvements, the colon polyp miss rate continues the same. A 2017 retrospective study done with 659 patients indicates that among these patients, the colon polyp miss rate was 17% (372 out of 2158 polyps), and 39% of patients (255 out of 659 patients) had at least one missed polyp (Lee et al., 2017). As mentioned before, an undetected polyp, be it benign or malignant, may lead to a late CRC diagnosis, which is associated with a less than 10% survival rate for metastatic CRC. Many elements contribute to missed polyps during a colonoscopy. Two of them are the quality of bowel preparation and the experience of the colonoscopists (Bonnington and Rutter, 2016). While the first problem cannot be fixed by technology, the second one can, and computer-aided tools can assist colonoscopists in detecting polyps and reducing polyp miss rates.

The key contributions of this study are (i) presenting the state-of-the-art deep learning techniques to detect, classify, and localize colon polyps; and (ii) introducing the convolutional neural network with autoencoders (CNN-AE) algorithm for detection of polyps with no previous image pre-processing.

## 5.1 Background

Researchers have been applying deep learning techniques and algorithms in various healthcare applications. Considerable progress is seen in detecting colon polyps (Poon et al., 2020; Tajbakhsh et al., 2016). Having a public database of colon polyp images played a big role. Examples of such contributions include using a pre-trained deep convolutional neural network to detect colon polyps (Tajbakhsh et al., 2016), dividing images into small patches or in sub-images to increase the database′s size, and then classifying different regions of the same image (Ribeiro et al., 2016a). Other works include exploring deep learning to automatically classify polyps using various configurations, such as training the CNN model from scratch or modifying different CNN architectures pre-trained in other databases and testing them in an 8-HD-endoscopic image database (Ribeiro et al., 2017). Authors in (Ribeiro et al., 2016b) take advantage of transfer learning, a technique where a model is trained on a task and later re-purposed and used for another task similar to the previous one. (Ribeiro et al., 2016b) uses CNN as a feature descriptor and generates features for the classification of colon polyps. Another CNN was developed to detect hyperplastic and adenomatous polyps and classify them by modifying different low-level CNN layer features learned from non-medical datasets (Zhang et al., 2017).

The authors in (Shin et al., 2018) use a deep CNN model as a transfer learning scheme. Besides image augmentation strategies for training deep networks, they propose two post-learning methods, automatic false-positive learning, and offline learning. (Shin and Balasingham, 2017) compare a handcraft feature method with a CNN method to classify colorectal images. For the handcraft feature approach, they use the shape and color features together with a support vector machine (SVM) for classification. On the other hand, the CNN approach uses three convolutional layers with pooling to do the same. They compare the strategies by testing them in three public polyp databases. Results show the CNN-based deep learning framework leads better classification performance by achieving an accuracy, sensitivity, specificity, and precision of over 90%. Authors in (Korbar et al., 2017) build an automatic image analysis method that classifies different types of colorectal polyps on whole-slide images with an accuracy of about 93%. (Mahmood and Durr, 2018) use a deep CNN together with a conditional random field (CRF) called (CNN-CRF), a framework for estimating the depth of a monocular endoscopy. Estimated depth is used to reconstruct the topography of the surface of the colon from a single image. They train the framework on over 200,000 synthetic images of an anatomically realistic colon, which they generated by developing an endoscope camera model.

The validation is done using endoscopy images from a porcine colon, transferred to a synthetic-like domain via adversarial training. The relative error of the CNN-CRF approach is 0.152 for synthetic endoscopy images and 0.242 for real endoscopy images. They show that the depth map can be used to reconstruct the mucosa topography.

Three 2020 studies focus more on polyp classification by approaching the problem in different ways. (Carneiro et al., 2020) studies the roles of confidence and classification uncertainty in deep learning models and proposes and tests a new Bayesian deep learning method to improve classification accuracy and model interpretability on a privately owned polyp image dataset. (Gao et al., 2020) use DL methods to establish colorectal lesion detection, positioning, and classification based on white light endoscopic images. The CNN model is used to detect whether the image contains lesions (CRC, colorectal adenoma, and other types of polyps), and the instance segmentation model is used to locate and classify the lesions on the images. They compare some of the most used CNN models to do so, such as ResNet50, AlexNet, VGG19, ResNet18, and GoogleNet. (Song et al., 2020) developed a computer-aided diagnostic system (CAD) for predicting colorectal polyp histology using deep-learning technology with near-focus narrow-band imaging (NBI) pictures of the privately-owned colorectal polyps image dataset. The performance of the CAD is validated with two test datasets. Polyps were classified into three histological groups. The CAD accuracy (81.3–82.4%) shows to be higher than that of trainee colonoscopists (63.8–71.8%) but comparable with that of expert colonoscopists (82.4–87.3%).

There are other works that are focused on colon polyp detection on colonoscopy videos besides images. Such work includes (Urban et al., 2018), where authors explore the idea of applying a deep CNN model to a large set of images taken from 20 videos approximately 5 h long (~500,000 frames). In (Misawa et al., 2018), the authors develop a three-dimensional (3D) CNN model and train it on 155 short videos. In (Mohammed et al., 2018) deep learning method called Y-Net is proposed that consists of two encoder networks with a decoder network that relies on efficient use of pre-trained and un-trained models with novel sum-skip-concatenation operations. The encoders are trained with a learning rate specific to encoders and the same for the decoder. (Yu et al., 2017) proposes an offline and online framework by leveraging the 3D fully convolutional network (3D-FCN). Their 3D-FCN framework is able to learn more representative spatial-temporal features from colonoscopy videos by showing a more powerful discrimination capability. Their proposed online learning scheme deals with limited training data by harnessing the specific information of an input video in the learning process. They integrate offline learning to the online one to reduce the number of false positives, which brings detection performance improvements. Another work (Byrne et al., 2019) includes using a deep CNN model based on inception network architecture trained in colonoscopy videos. They use only unaltered NBI video frames to train and validate the model. A test dataset of 125 videos of consecutively encountered diminutive polyps was used to test the model. However, the confidence mechanism of the model did not generate sufficient confidence to predict the detection of 19 polyps in the test set, which represented 15% of the polyps. In a more recent study (Poon et al., 2020), the authors design an Artificial Intelligent Endoscopist (AI-doscopist) to localize polyps during colonoscopy with the purpose of evaluating the agreement between endoscopists and AI-doscopist for set localization. Another recent study that deals with colorectal videos is (Wang et al., 2020), which is the first double-blind, randomized controlled trial to assess the effectiveness of automatic polyp detection using computer-aided detection (CADe) system during colonoscopy. This is also the only clinical trial that deals with the use of artificial intelligence (AI) in colorectal image/video detection, localization, and/ or classification.

There are studies that train and test models in both images and videos. One of them is (Yamada et al., 2019), where they develop an AI system that detects early signs of colorectal cancer during colonoscopy by decomposing tensor metrics in the trained model. Their AI system consists of a Faster R-CNN and the VGG16 model. *Table 23* summarizes the articles included in this minireview, together with some characteristics of these studies.

The presented model is a combination of CNN and autoencoders. This model was trained on three different colon polyp databases, CVC-ColonDB (Bernal et al., 2012), CVC-ClinicDB (Bernal et al., 2015), and ETIS-LaribPolypDB (Silva et al., 2014). All these datasets are open source and can be used for research purposes to develop techniques to detect colon and rectal polyps making them in a way the standard datasets in the field.

***Table 23.*** Summary of the reviewed work.

| Year | Authors | Nr of images | Format | Objective | Network | Metrics | Datasets | Novelties |
|---|---|---|---|---|---|---|---|---|
| 2017 | (Yu et al., 2017) | Train: 1.1 M non-med<br>Test: 20 | Video | Detection | 3D-FCN | F1 = 78.6%,<br>F2 = 73.9% | Asu-Mayo Clinic Polyp Database | An integrated framework with online and offline 3D representation learning |
| 2017 | (Shin and Balasingham, 2017) | Train: 1525<br>Test: 366 | Image | Classification | HOG + SVM,<br>Combined feature + SVM,<br>CNN (gray),<br>CNN(RGB) | Accu = 91.3%,<br>Sens = 90.8%,<br>Spec = 91.8%,<br>Prec = 92.7% | CVC-Clinic,<br>ETIS-Larib,<br>Asu-Mayo | Compare handcraft feature based SVM method and CNN method for polyp image frame classification |
| 2017 | (Korbar et al., 2017) | Train: 2074 crop images<br>Test: 239 full images | Image | Classification | AlexNet8,<br>VGG19,<br>GoogleNet22,<br>ResNet50,<br>ResNet101,<br>ResNet152,<br>ResNet152 | Accu = 93.0%,<br>Prec = 89.7%,<br>Rec = 88.3%,<br>F1 = 88.8% | Private dataset | Identify polyps and their types on whole-slide images by breaking them into smaller, overlapping patches |
| 2018 | (Mahmood and Durr, 2018) | Synthetic colon: 100,000<br>Phantom colon: 100,000<br>Porcine colon: 1460 | Image | Detection | CNN + CRF | RE = 0.242 | synthetic data, real endoscopy images from a porcine colon | Synthetically generated endoscopy images |
| 2018 | (Urban et al., 2018) | Train: 8641 images<br>Test: 20 videos | Image/Video | Detection | CNN | Accu = 96.4%,<br>AUC ROC = 0.991 | Private dataset | Localization model by optimizing the size and location, optimizing the Dice loss, and a variation of the “you only look once” algorithm (“internal ensemble”) |
| 2019 | (Byrne et al., 2019) | Train: 223<br>Test: 125 | Video | Detection | DCNN based on inception network architecture | Accu = 94%,<br>Sens = 98%,<br>Spec = 83%,<br>NPV = 97%,<br>PPV = 90% | Private dataset | AI differentiating diminutive adenomas from hyperplastic polyps on unaltered videos of colon polyps.<br>The model operates in quasi-real-time |
| 2019 | (Yamada et al., 2019) | Train: 4840 images<br>Test: 77 videos | Image/Video | Detection | Faster R-CNN + VGG16 | Sens = 97.3%,<br>Spec = 99.0%,<br>ROC = 0.975 | Private dataset | Included 5000 images of more than 2000 lesions, and 3000 images of more than 500 non-polypoid superficial lesions |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | | It is nearly real-time processing |
| 2020 | (Carneiro et al., 2020) | 940 | Image | Classification | ResNet-101 & DenseNet-121 | Accu = 51%, Avg Prec = 48% (Z = 0.7) | Private dataset (Australian & Japanese) | Deep learning classifier using classification uncertainty and calibrated confidence to reject the classification of test samples |
| 2020 | (Gao et al., 2020) | 3413 | Image | Detection + Classification | AlexNet, VGG19, ResNet18, GoogLeNet, ResNet50, Mask R-CNN | Accu = 93.0%, Sens = 94.3%, Spec = 90.6% | Private dataset | Detection and classification models based on white light endoscopic images |
| 2020 | (Poon et al., 2020) | Pre-trained: 1.2 M non-med images<br>Fine-tuned: 291,090 polyp & non-med images<br>Test: 144 videos | Video | Localizing | ResNet50 + YOLOv2 + a temporal tracking algorithm | Sens = 96.9%, Spec = 93.3% | CVC-ColonDB, CVC-ClinicDB, ETIS-LaribDB, AsuMayoDB, CU-ColonDB, ACP-ColonDB, Selected Google Images | Real-time AI algorithm for localizing polyps in colonoscopy videos, using different medical and non-medical datasets for training |
| 2020 | (Song et al., 2020) | Train: 12,480 image patches of 624 polyps<br>Test: two DBs of 545 polyp images | Image | Classification | CAD based on NBI near-focus images + ResNet-50, DenseNet-201 | Accu = 82.4% | Private dataset | A CAD system for predicting CR polyp histology using near-focus narrow-band imaging (NBI) pictures and deep-learning technology |
| 2020 | (Wang et al., 2020) | CADe group: 484 patients<br>non-CADe group: 478 patients | Video | Detection | CAD + AI | ADR = 34% | Private dataset | The first double-blind, randomized controlled trial to assess the effectiveness of automatic polyp detection using a CADe system during colonoscopy. |

Accu = accuracy, Prec = precision, Spec = Specificity, Sens = Sensitivity, Rec = recall, NPV = negative predictive value, PPV = positive predictive value, RE = relative error, ADR = adenoma detection rate, non-med = non-medical, CAD = computer-aided device, CADe = computer-aided detection.

## 5.2 Materials and Methods

### 5.2.1 Databases

Three colorectal polyp image datasets, namely CVC-ColonDB, CVC-ClinicDB, and ETIS-LaribPolypDB, are used for this study. The first colorectal polyp image dataset to be made available for researchers is CVC-ColonDB, and it contains 380 images. All the images are part of 15 colonoscopy videos, and each sequence has various numbers of polyp pictures. The same group that published CVC-ColonDB later made available the CVC-ClinicDB dataset, which has 612 images taken from 29 sequences. The third dataset is ETIS-LaribPolypDB which has 196 images, *Table 24.* Each dataset consists of 2 main folders, the raw original images, and the masked images, the ground truths, of the corresponding one in the original image. *Figure 25* shows images of polyps taken during several colonoscopies. As seen from the figure, polyps come in various shapes and sizes, and some of them are not significantly distinguishable from the mucosa of the colon.

***Table 24.*** Databases used to train and test the CNN-AE model.

| Datasets | Nr of Images |
|---|---|
| CVC-ColonDB | 380 |
| CVC-ClinicDB | 612 |
| ETIS-LaribPolypDB | 196 |

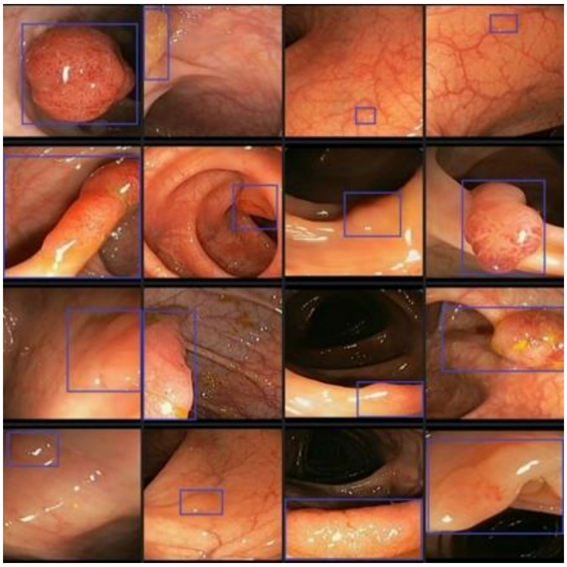


***Figure 25.*** Different shapes and textures of colon polyps taken from colonoscopy videos.

### 5.2.2 The Proposed Model

There are some deep learning libraries that can be used to build a neural network model. One of them is TensorFlow (Abadi et al., 2016), an open-source library created by Google and community contributors, currently on its 2.0 version. This library is used to train and test the proposed convolutional encoder-decoder model. The model uses the same architecture as the SegNet architecture (Badrinarayanan et al., 2017), an algorithm programmed in Caffe, another deep learning library created by Berkeley AI Research and community contributors. The training and testing are performed on a computer with NVIDIA Titan X GPU.

*Figure 26* shows the architecture of the CNN-Autoencoder model. The model has two parts, the encoder, and the decoder. The structure of the encoder is similar to some image classification neural networks, such as the convolutional layer, which includes the batch normalization, the rectified linear unit (the ReLu) activation function, and the pooling layer. The decoder part has the inversed layers used in the encoder, such as deconvolution layers and de-max_pool layers.

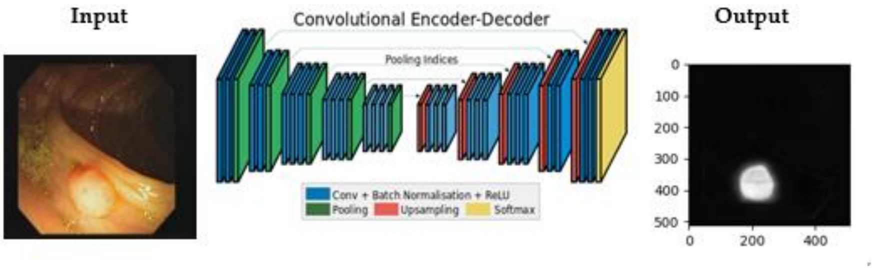


***Figure 26.*** Convolutional encoder-decoder architecture.

The encoder part has 13 convolutional layers and 5 max_pooling layers, where the first 3 layers of the model have these characteristics: the first convolution layer is with stride 2, followed by the second convolution layer with stride 1, and a non-overlapping 2 × 2 window max_pooling layer with stride 2. As mentioned above, each decoder layer contains the corresponding layer of the encoder, which means the decoder network has 13 layers. The output of the last decoder is fed to the Softmax classifier, which produces for each pixel the probabilities if it is a polyp or a normal colon tissue. The network input-output dimensions are equal:

- use the same layer for the non-shrinking convolution layer.

- use transposed deconvolution for the shrinking convolution layer adjusted with the same parameters.
- use the nearest neighbor upsampling for the max_pooling layer.

Open-source medical image datasets lack the number of images in them, often only a couple of hundred images. However, for deep learning algorithms to work, a large amount of data is needed. In the case of image databases, researchers have used image augmentation techniques to increase the number of training images. The image augmentation used in this case is Imgaug Library (imgaug, 2020), a Python image augmentation library. *Figure 27* shows the results after applying some image augmentations that are used in this study, which include:

- Crop—parameter: px = (0, 16) which crops images from each side by 0 to 16 pixels chosen randomly.
- Fliplr—parameter: 0.5, which flips horizontally 50% of all images.
- Flipud—parameter: 0.5, which flips vertically 50% of all images.
- GaussianBlur—parameter: (0, 3.0), blurs each image with varying strength using gaussian blur (sigma between 0 and 3.0).
- Dropout—parameter: (0.02, 0.1), drop randomly 2 to 10% of all pixels (i.e., set them to black).
- AdditiveGaussianNoise—parameter: scale = 0.01*255, adds white noise pixel by pixel to images.
- Affine—parameter: translate_px = {“x”: (-network.IMAGE_HEIGHT // 3, net-work.IMAGE_WIDTH // 3)}, applies translate/move of images (affine transformation).

The use of image augmentation not only increases the number of images to train the model but also increases the robustness and reduces the overfitting of the model. Another technique to deal with the overfitting problem is the Dropout technique, with a rate of 0.2. Each dataset is divided into a train and validation set. The majority of the data in each dataset is used for training 70%, 15% is used to validate and the other 15% to test the model.

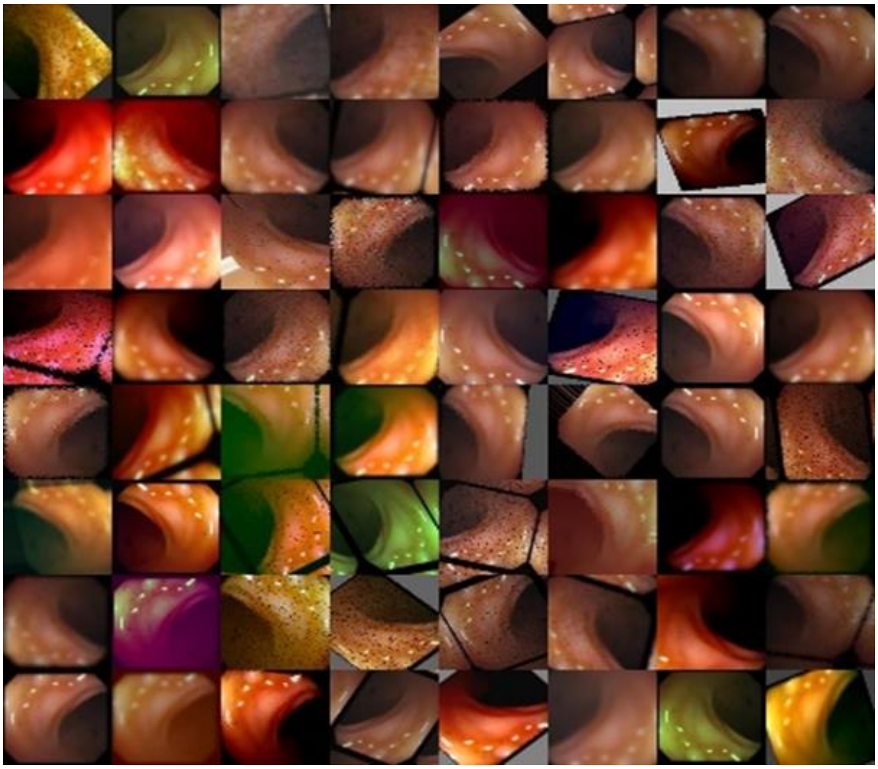


***Figure 27.*** One image of colon polyp after applying different image augmentations.

## 5.3 Results

The model is trained on the selected databases using only the training sets and then validated and tested with the validate and test sets. Each database has a different number of images, the time to train the model varied. The same batch size of 100 was used for all datasets. The accuracy and the total training time for each database are depicted in *Table 25.* The best accuracy is achieved on ETIS-LaribPolypDB's last batch with a score of 0.967.

***Table 25***. The accuracy and the total training time for each dataset.

| Datasets | Best Accuracy | Batch | Total Time |
|---|---|---|---|
| ETIS-LaribPolypDB | 0.967 | 1300 | 1120.48 |
| CVC-ClinicDB | 0.951 | 2200 | 2186.97 |
| CVC-ColonDB | 0.937 | 2000 | 3659.52 |

Apart from the accuracy results from each batch and the final test accuracy, the predicted images are obtained as well. The test input, test targets, and test predictions are set to a gray scale before all the results are drawn. *Figure 28* depicts one example from each dataset. The three columns represent the three datasets (left to right: ETIS-LaribPolypDB, CVC-ClinicDB, and CVC-ColonDB) and the three rows, from top to bottom, the test image, the test ground truth (target), and the result of the segment obtained from the model.

Polyps have various shapes and characteristics, see *Figure 25*, ranging from prominent and recognizable polyps to barely distinguishable circular shapes. In *Figure 28*, the polyp in the first column is not easily detectable by the human eye, while the polyp in the last is recognizable. This wide variation induces errors in polyp recognition.

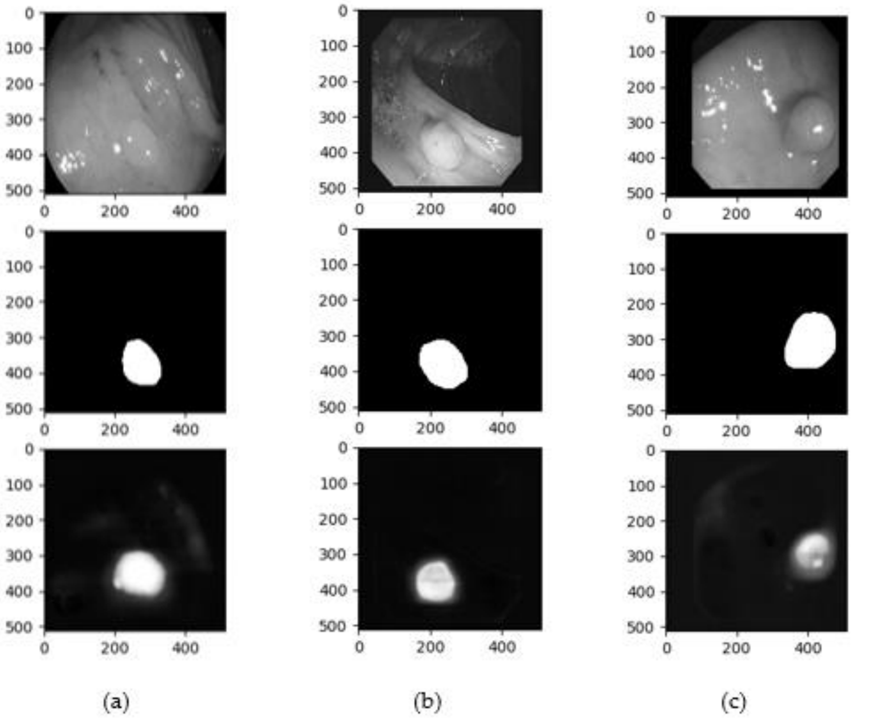


***Figure 28.*** Images showing the results after training the convolutional encoder-decoder model on (a) ETIS-LaribPolypDB, (b) CVC-ClinicDB, and (c) CVC-ColonDB database.

## 5.4 Discussion and Conclusions

Many techniques and algorithms used these recent years are presented in the background section. A quick glance at summary *Table 23* depicts how diverse these techniques are, but also how diverse the metrics to evaluate them are. Accuracy was one of the most used metrics, followed by the other metrics such as precision, recall, etc. Although the main topic is the same, colorectal polyps, comparing results is difficult. The first reason is the one explained above, different metrics. The others are related to the objectives, for what purpose these algorithms are used (classification, segmentation, detection, or classification), and the databases these algorithms are trained.

Among the cited papers, two other similar studies are found, meaning they focus on detection problems and use the same metric and database/s. The accuracy of 96.7 is obtained by using the CNN-Autoencoder model, which is slightly better than the current state-of-the-art models that used the same metric, *Table 26.*

***Table 26.*** Accuracy comparison for the proposed model and previously published studies on colon polyp detection.

| **Model** | **Accuracy (%)** |
|---|---|
| CNN-Autoencoder (proposed model) | 96.7 |
| DLL [23] | 96.4 |
| AI-APD [24] | 76.5 |

The main challenges with colonoscopy images seem to fall on the shape and texture of the polyps (Korbar et al., 2017; Lequan et al., 2017; Song et al., 2020) and the quality of the images (Byrne et al., 2019; Gao et al., 2020; Lequan et al., 2017; Song et al., 2020). The quality of the images depends on the colonoscopy device itself (Gao et al., 2020; Lequan et al., 2017) or on the expertise of the endoscopist (Byrne et al., 2019; Korbar et al., 2017; Song et al., 2020). Furthermore, in the case of polyp classification, class imbalance poses another problem (Carneiro et al., 2020). Considering these challenges, an image results check is performed, and it is verified that indeed some of the segments the model predicted are not as expected. The unexpected bad masks are shown in *Figure 29,* and again this shows the implications that the shape and texture of the polyp has, but also the conditions the colonoscopy image was taken. The lighting used during the examination plays a negative role when it comes to colon polyp detection as the models misrecognize the normal tissue as a polyp. This phenomenon happens because the inner surface of the colon is smooth, and the light attached to the colonoscopy used by the endoscopists to exam the colon reflects, confusing the models to consider healthy colon tissues as polyps. It is also worth mentioning that the patient needs to prepare well and follow the doctor’s instructions as per the regular colonoscopy session.

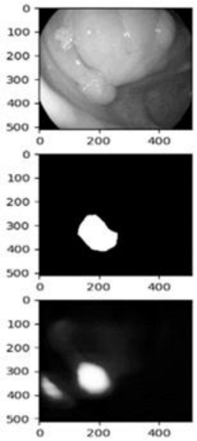


***Figure 29.*** False detection of a polyp due to lighting conditions.

Technology has helped progress the medical field enormously, especially when it comes to medical imaging. Colorectal cancer has been one of the diseases which has gained attention, and many researchers have worked towards detecting and preventing such disease. CAD systems have shown that the polyp miss rate has gone down. However, research shows deep learning has revealed even more progress aiding colonoscopists/ endoscopists to perform better.

In this work, the current state of the art of deep learning techniques in colon polyp detection, classification, segmentation, and localization is presented. The main contribution is the CNN-AE novel algorithm for the detection of polyps, which appears promising considering that no image preprocessing was performed prior to training the model. The model shows better results than the current state of the art, although not very significant. Better results may be achieved if the number of images in the dataset is increased. Moreover, having a diverse range of polyp images may improve the algorithm's performance. The same model is tested on other medical image databases, namely iris and pressure ulcer datasets, and the results obtained are better than with the colon polyp images. To address these issues, future work includes making changes to the model and adding other image augmentations currently not implemented in the Imgaug library. Besides the technical aspect, another problem is the lack of polyp image datasets. A more extensive and diverse dataset of colon polyp images is being created. The model will be tested as soon as the dataset is completed, which will be made available to researchers for academic purposes as well.

## 5.5 Summary

Colorectal cancer is one of the leading causes of cancer incident cases and cancer deaths worldwide. Undetected colon polyps, be they benign or malignant, leading to late diagnosis of colorectal cancer. Computer-aided devices have helped to decrease the polyp miss rate. The application of deep learning algorithms and techniques has escalated during this last decade. Many scientific studies are published to detect, localize, and classify colon polyps. Section 5.1 presents a brief review of the latest published studies. The results obtained from this study from training and testing three independent datasets using a convolutional neural network and autoencoder model are compared with the accuracy of previous studies. A train, validate and test split was performed for each dataset, 70%, 15%, and 15%, respectively. Accuracy of 0.937 was achieved for CVC-ColonDB, 0.951 for CVC-ClinicDB, and 0.967 for ETIS-LaribPolypDB. The results suggest slight improvements compared to the algorithms used to date.

# 6

# CONCLUSIONS

This chapter presents the most relevant conclusions drawn after the development of this work. These findings are introduced following the completion of the objectives outlined in the Introduction chapter.

This dissertation started firstly with understanding the domain of care pathways in cancer disease. It moved on to analyze the care pathways in a hospital with a cohort of 83 patients diagnosed with breast or prostate cancer as the first study. The second study involved the implementation of various machine learning algorithms for feature selection and survival prediction for a database of 20006 unique records and more than 200 variables. The third, and last study, presented colorectal polyps' detection, which represents a serious health issue if undetected, using a combination of two deep learning algorithms, CNNs, and autoencoders.

The hypotheses stated in section 1.1 of the Introduction chapter were:

- *Understanding the past and current state of cancer care pathways implementations using ethnographic analysis can be used to find the main factors influencing care pathways in a hospital setting.*
- *Cancer patient survivability can be predicted using patients' electronic health records and various machine learning algorithms.*
- *Colorectal cancer can be prevented by detecting early colorectal polyps using deep learning algorithms.*

From the obtained results, which were presented in the previous chapters, we can state that the use of the machine and deep learning can definitely take medical decisions to the next level, improving the quality of diagnosis of several health issues by providing reliable results to caregivers from the analysis of patients' data.

## 6.1 Objectives and Research Questions

This dissertation work has provided several contributions in medical applications in the form of end-to-end frameworks, including collecting and preprocessing of the data, the design of the machine and deep learning architectures suitable for each type of data, and reliable results using typical validation metrics of each application. The completion of the different stages and contributions of this thesis has been made possible by the achievement of the objectives presented in section 1.1. All the seven main objectives of this thesis were fulfilled during the research process. Following are the specific objectives and their respective research questions that were answered in this thesis.

For care pathways analysis:

- SO1: *Define the current state of the art of care pathways*. This objective was successfully completed and presented in Chapter 3, section 2. It presents the analysis of published articles related to care pathways implementation and outcomes grouped by each care phase, methodology used to analyze the outcomes, and by the three types of cancer diseases this thesis is focused on. A total of 113 articles were included in this analysis. SO1 answers RQ1.

- SO2: *Understand the current state of care pathways in a hospital setting.* This objective is successfully completed by conducting an ethnographic study at Beacon Hospital. Methods used in this study included one-on-one semi-structured interviews, observations, and collection of patients data through their medical folders and both electronic healthcare records used in the hospital.

- SO3: *Construct a database containing information from patients' care pathways*. This objective was completed by creating two datasets one for breast cancer and another one for prostate cancer. The datasets included demographic, medical, lifestyle and financial information. Statistical analysis were performed to identify the factors influencing care pathways for these two cancer disease in a hospital setting. SO2 and SO3 answer RQ2.

For electronic health records:

- SO4: *Construct a more extensive breast and prostate cancer database*. This objective was successfully completed by collaborating with a regional hospital in Finland. The database was saved as a relational database and was accessed through a remote access server. It contained 24 tables with EHRs of 20006 unique records.

- SO5: *Design and implement algorithms for EHRs data analysis*. This objective was successfully completed by the publication of an article that consisted of implementing various algorithms to firstly select the best features through feature selection algorithms and later perform survival prediction using nine machine learning algorithms. Refer to Chapter 4 for the complete details of the study. SO4 and SO5 answer RQ3 and RQ4.

For colorectal polyp detection:

- SO6: *Present the current state of the art of deep learning algorithms for colorectal polyps*. This objective was successfully completed following a comprehensive review of the published scientific research related to colorectal polyp detection, classification, segmentation, localization in both formats, images, and video.

- SO7: *Implementing a deep learning architecture for colorectal polyp detection.* The architecture is a combination of convolutional neural networks and autoencoders to detect colorectal polyps without image preprocessing, which outperformed the current state-of-the-art contributions. Both SO1 and

SO2 are published in a top-rated journal, and they can be found in Chapter 5 of this thesis. Both SO6 and SO7 answer RQ5.

## 6.2 Scientific contribution

The following presents a complete relation of the different publications that took part throughout this research work. Although this dissertation is structured as a monography, two of the studies are published, and one is accepted and is currently in press, all in international journals with impact factors. Other contributions have also been made to the scientific community in the shape of communications to different international conferences, which are summarized below.

### 6.3.1 Articles in international journals with impact factor

The articles detailed in this section are the ones that compound this Ph.D. dissertation. Two of them have already been accepted and published in international journals, while the last one is currently in press.

***Table 27.*** Publication I - Details of the publication.

| | | | |
|---|---|---|---|
| **Title** | Machine learning techniques applied to electronic healthcare records to predict cancer patient survivability | | |
| **Authors** | Ornela Bardhi and Begonya Garcia-Zapirain | | |
| **Journal** | *Computers, Materials & Continua* | | |
| **Publication Date** | 13 April 2021 | | |
| **Impact Factor** | 4.89 | Quartile | Q1 |
| **DOI** | 10.32604/cmc.2021.015326 | | |

***Table 28.*** Publication II - Details of the publication.

| | | | |
|---|---|---|---|
| **Title** | Deep Learning Models for Colorectal Polyps | | |
| **Authors** | Ornela Bardhi, Daniel Sierra-Sosa, Begonya Garcia-Zapirain and Luis Bujanda | | |
| **Journal** | *Information* | | |
| **Publication Date** | 08 June 2021 | | |
| **Impact Factor** | 3.0 | Quartile | Q2 |
| **DOI** | https://doi.org/10.3390/info12060245 | | |

*Table 29.* Publication III - Details of the publication.

| | | | |
|---|---|---|---|
| **Title** | Factors influencing care pathways for breast and prostate cancer in a hospital setting | | |
| **Authors** | Ornela Bardhi, Begonya Garcia-Zapirain, Roberto Nuno-Solinis | | |
| **Journal** | *International Journal of Environmental Research and Public Health* | | |
| **Publication Date** | Accepted, in Press | | |
| **Impact Factor** | 2.849 | Quartile | Q1 |
| **DOI** | - | | |

### 6.3.2 Communications in international conferences

These articles, posters, and communications have been produced during the Ph.D. research. All the communications in this section are related to the topic and research field of this dissertation.

*Table 30.* Publication IV - Conference publication.

| | | | |
|---|---|---|---|
| **Title** | A convolutional neural network algorithm for colon polyp detection | | |
| **Authors** | Ornela Bardhi, Daniel Sierra-Sosa, Begonya Garcia-Zapirain and Adel Elmaghraby | | |
| **Conference** | The 8th International Conference on Biomedical Engineering and Biotechnology (ICBEB 2019) | | |
| **Year** | 22-25 October 2019 | **Location** | Republic of South Korea |
| **Publisher** | Wiley Online Library | | |
| **DOI** | https://doi.org/10.1111/bcpt.13326 | | |

*Table 31.* Publication V - Conference publication.

| | | | |
|---|---|---|---|
| **Title** | Automatic colon polyp detection using Convolutional encoder-decoder model | | |
| **Authors** | Ornela Bardhi, Daniel Sierra-Sosa, Begonya Garcia-Zapirain and Adel Elmaghraby | | |
| **Conference** | 2017 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT) | | |
| **Publisher** | IEEE | | |
| **Year** | 18-20 December 2017 | **Location** | Spain |
| **DOI** | 10.1109/ISSPIT.2017.8388684 | | |

*Table 32.* Publication VI - Conference publication.

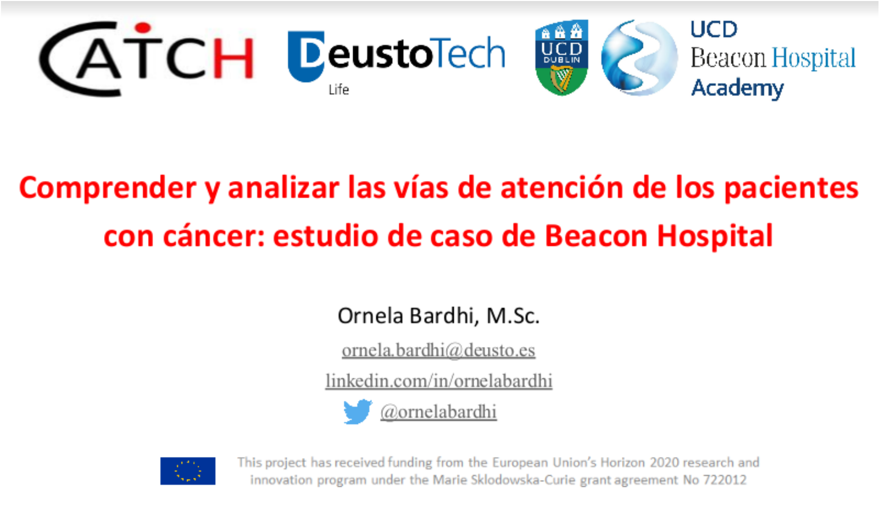
| | | | |
|---|---|---|---|
| **Title** | ICPs as an enabler of transformation towards integrated care | | |
| **Authors** | Elena Urizar, Massimiliano Panella, Carles Blay, Ornela Bardhi | | |
| **Conference** | International Conference on Integrated Care | | |
| **Publisher** | International Journal of Integrated Care | | |
| **Year** | 01-03 April 2019 | **Location** | Spain |
| **DOI** | http://doi.org/10.5334/ijic.s3635 | | |

### 6.3.3 Communications in other scientific venues

Besides international conferences from renowned publishers, the work is presented in other scientific venues, such as the Congress of People with Cancer and their families in Spain, CATCH Conferences, the Marie Sklodowska-Curie Action Falling Walls Lab Ph.D. competition in Belgium, and other mainstream media in Spain and Albania.

*Table 33.* Presentation - Details of the presentation.

| | | | |
|---|---|---|---|
| **Title** | Comprender y analizar las vías de atención de los pacientes con cáncer: estudio de caso de Beacon Hospital | | |
| **Authors** | Ornela Bardhi | | |
| **Conference** | Congreso de Personas con Cáncer y sus Familiares | | |
| **Year** | 15-17 November 2019 | **Location** | Spain |


CATCH
DeustoTech
Life
UCD DUBLIN
UCD
Beacon Hospital
Academy

Comprender y analizar las vías de atención de los pacientes con cáncer: estudio de caso de Beacon Hospital

Ornela Bardhi, M.Sc.

ornela.bardhi@deusto.es

linkedin.com/in/ornelabardhi

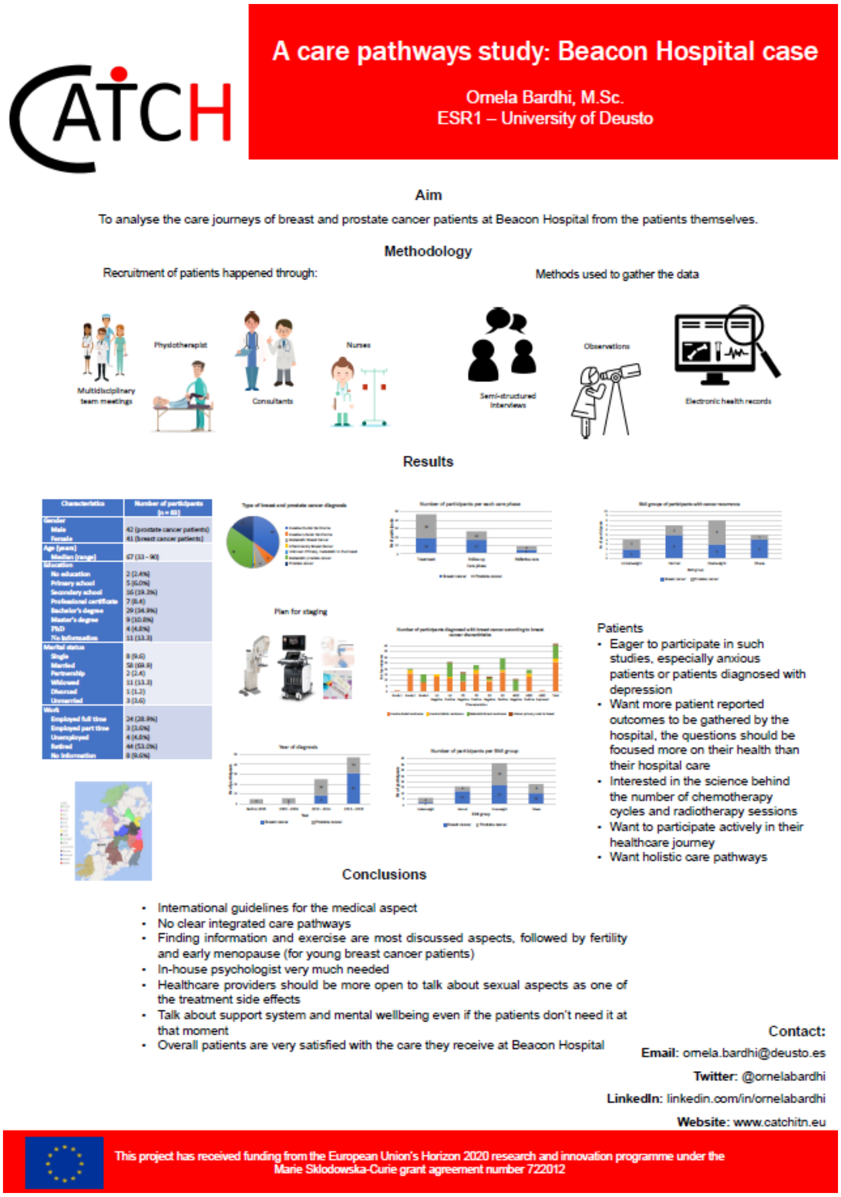
@ornelabardhi

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 722012

***Figure 30.*** Title page of the presentation for the Congreso de Personas con Cáncer y sus Familiares.

***Table 34.*** Poster I - Details of the poster.

| | | | |
|---|---|---|---|
| **Title** | A care pathways study: Beacon Hospital case | | |
| **Authors** | Ornela Bardhi | | |
| **Conference** | CATCH Conference 2019 | | |
| **Year** | 13-16 August 2019 | **Location** | Denmark |



## A care pathways study: Beacon Hospital case

Ornela Bardhi, M.Sc.
ESR1 – University of Deusto

### Aim

To analyse the care journeys of breast and prostate cancer patients at Beacon Hospital from the patients themselves.

### Methodology

Recruitment of patients happened through:

Methods used to gather the data

CATCH
Multidisciplinary team meetings
Physiotherapist
Consultants
Nurses
Semi-structured interviews
Observations
Electronic health records
Type of breast and prostate cancer diagnosis
Number of participants per each care phase
BMI groups of participants with cancer recurrence
Plan for staging
Number of participants diagnosed with breast cancer according to breast cancer characteristics
Year of diagnosis
Number of participants per BMI group

### Results

| Characteristics | Number of participants (n = 83) |
|---|---|
| Gender | |
| Male | 42 (prostate cancer patients) |
| Female | 41 (breast cancer patients) |
| Age (years) | |
| Median (range) | 67 (33 - 90) |
| Education | |
| No education | 2 (2.4%) |
| Primary school | 5 (6.0%) |
| Secondary school | 16 (19.3%) |
| Professional certificate | 7 (8.4) |
| Bachelor's degree | 29 (34.9%) |
| Master's degree | 9 (10.8%) |
| PhD | 4 (4.8%) |
| No information | 11 (13.3) |
| Marital status | |
| Single | 8 (9.6) |
| Married | 58 (69.9) |
| Partnership | 2 (2.4) |
| Widowed | 11 (13.3) |
| Divorced | 1 (1.2) |
| Unmarried | 3 (3.6) |
| Work | |
| Employed full time | 24 (28.9%) |
| Employed part time | 3 (3.6%) |
| Unemployed | 4 (4.8%) |
| Retired | 44 (53.0%) |
| No information | 8 (9.6%) |

Patients

- Eager to participate in such studies, especially anxious patients or patients diagnosed with depression
- Want more patient reported outcomes to be gathered by the hospital, the questions should be focused more on their health than their hospital care
- Interested in the science behind the number of chemotherapy cycles and radiotherapy sessions
- Want to participate actively in their healthcare journey
- Want holistic care pathways

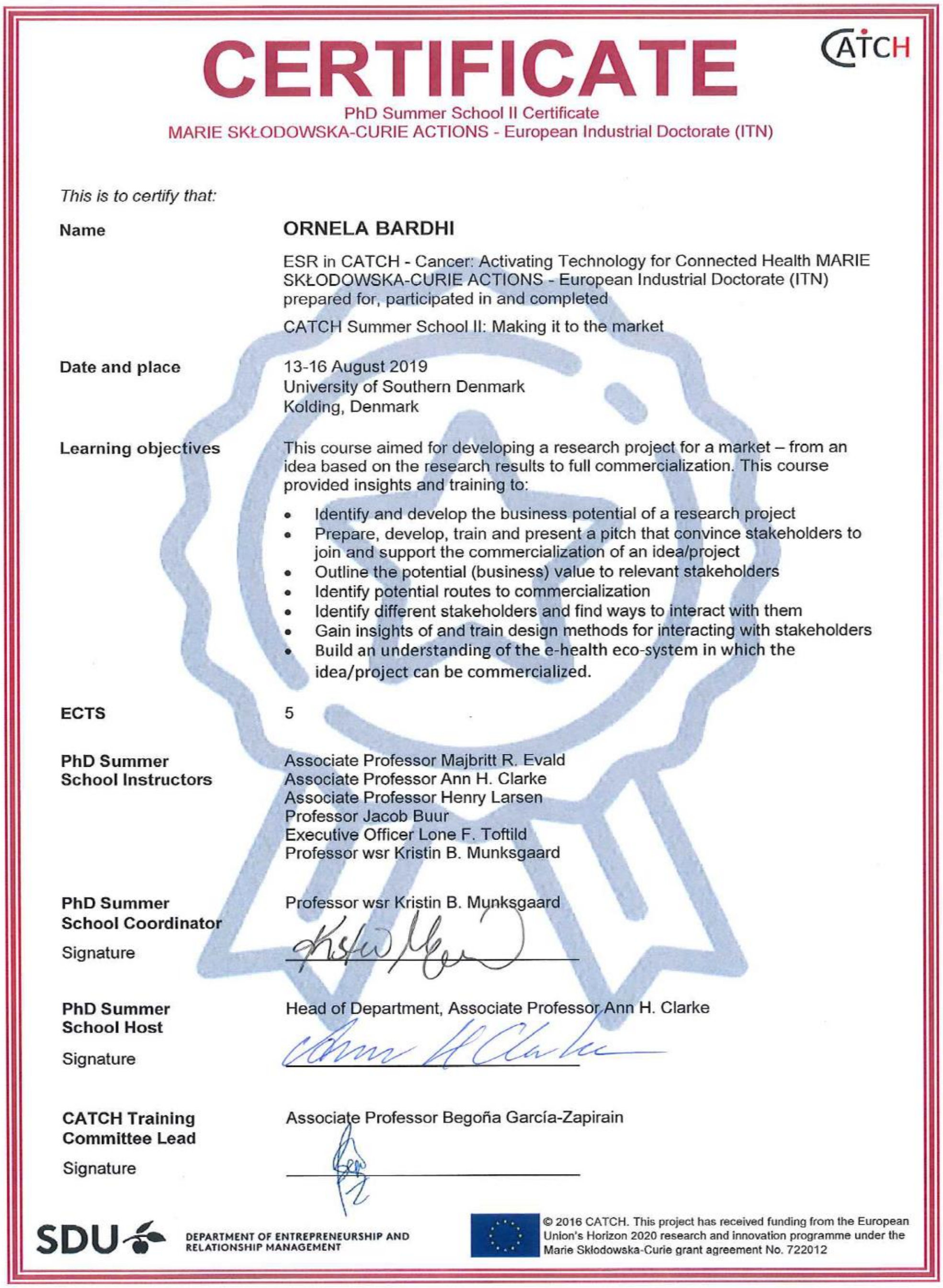
### Conclusions

- International guidelines for the medical aspect
- No clear integrated care pathways
- Finding information and exercise are most discussed aspects, followed by fertility and early menopause (for young breast cancer patients)
- In-house psychologist very much needed
- Healthcare providers should be more open to talk about sexual aspects as one of the treatment side effects
- Talk about support system and mental wellbeing even if the patients don't need it at that moment
- Overall patients are very satisfied with the care they receive at Beacon Hospital

**Contact:**
**Email:** ornela.bardhi@deusto.es
**Twitter:** @ornelabardhi
**LinkedIn:** linkedin.com/in/ornelabardhi
**Website:** www.catchitn.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 722012

***Figure 31.*** Poster for the CATCH Conference 2019.



# CERTIFICATE

CATCH

PhD Summer School II Certificate

MARIE SKŁODOWSKA-CURIE ACTIONS - European Industrial Doctorate (ITN)

*This is to certify that:*

**Name** ORNELA BARDHI

ESR in CATCH - Cancer: Activating Technology for Connected Health MARIE SKŁODOWSKA-CURIE ACTIONS - European Industrial Doctorate (ITN) prepared for, participated in and completed

CATCH Summer School II: Making it to the market

**Date and place**

13-16 August 2019
University of Southern Denmark
Kolding, Denmark

**Learning objectives**

This course aimed for developing a research project for a market – from an idea based on the research results to full commercialization. This course provided insights and training to:

- Identify and develop the business potential of a research project
- Prepare, develop, train and present a pitch that convince stakeholders to join and support the commercialization of an idea/project
- Outline the potential (business) value to relevant stakeholders
- Identify potential routes to commercialization
- Identify different stakeholders and find ways to interact with them
- Gain insights of and train design methods for interacting with stakeholders
- Build an understanding of the e-health eco-system in which the idea/project can be commercialized.

**ECTS** 5

**PhD Summer School Instructors**

Associate Professor Majbritt R. Evald
Associate Professor Ann H. Clarke
Associate Professor Henry Larsen
Professor Jacob Buur
Executive Officer Lone F. Toftild
Professor wsr Kristin B. Munksgaard

**PhD Summer School Coordinator**

Professor wsr Kristin B. Munksgaard

Signature

**PhD Summer School Host**

Head of Department, Associate Professor Ann H. Clarke

Signature Ann H Clarke

**CATCH Training Committee Lead**

Associate Professor Begoña García-Zapirain

Signature

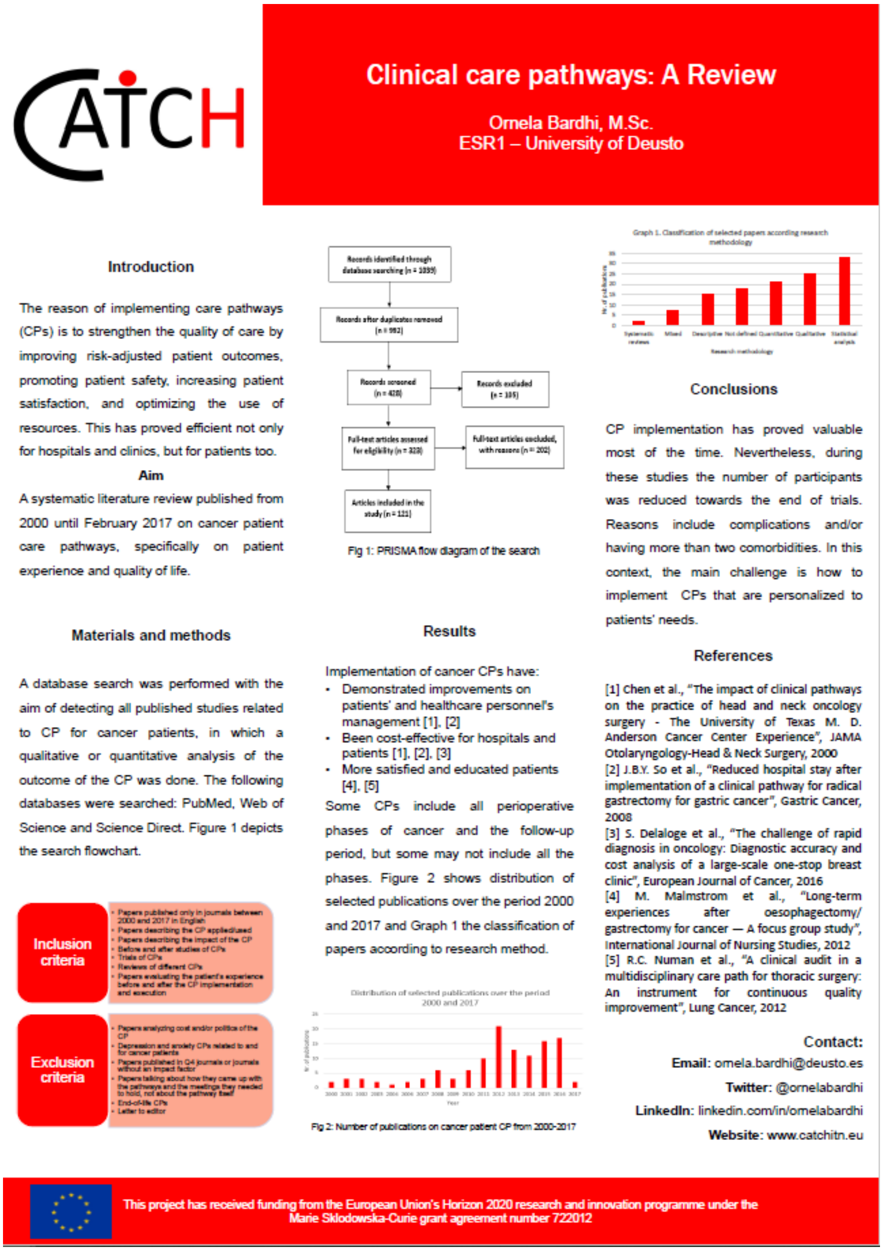
SDU DEPARTMENT OF ENTREPRENEURSHIP AND RELATIONSHIP MANAGEMENT

© 2016 CATCH. This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 722012

***Figure 32.*** Certificate of attendance for the CATCH 2019 Conference.

***Table 35.*** Poster II - Details of the poster.

| | | | |
|---|---|---|---|
| **Title** | Clinical care pathways – A review | | |
| **Authors** | Ornela Bardhi | | |
| **Conference** | CATCH Conference 2017 | | |
| **Year** | 5-9 June 2017 | **Location** | Ireland |



# Clinical care pathways: A Review

Ornela Bardhi, M.Sc.
ESR1 – University of Deusto

## Introduction

The reason of implementing care pathways (CPs) is to strengthen the quality of care by improving risk-adjusted patient outcomes, promoting patient safety, increasing patient satisfaction, and optimizing the use of resources. This has proved efficient not only for hospitals and clinics, but for patients too.

### Aim

A systematic literature review published from 2000 until February 2017 on cancer patient care pathways, specifically on patient experience and quality of life.

## Materials and methods

A database search was performed with the aim of detecting all published studies related to CP for cancer patients, in which a qualitative or quantitative analysis of the outcome of the CP was done. The following databases were searched: PubMed, Web of Science and Science Direct. Figure 1 depicts the search flowchart.

**Inclusion criteria**

- Papers published only in journals between 2000 and 2017 in English
- Papers describing the CP applied/used
- Papers describing the impact of the CP
- Before and after studies of CPs
- Trials of CPs
- Reviews of different CPs
- Papers evaluating the patient's experience before and after the CP implementation and execution

**Exclusion criteria**

- Papers analyzing cost and/or politics of the CP
- Depression and anxiety CPs related to and for cancer patients
- Papers published in Q4 journals or journals without an impact factor
- Papers talking about how they came up with the pathways and the meetings they needed to hold, not about the pathway itself
- End-of-life CPs
- Letter to editor

Records identified through database searching (n = 1039)
Records after duplicates removed (n = 982)
Records screened (n = 428)
Records excluded (n = 105)
Full-text articles assessed for eligibility (n = 323)
Full-text articles excluded, with reasons (n = 202)
Articles included in the study (n = 121)

Fig 1: PRISMA flow diagram of the search

## Results

Implementation of cancer CPs have:

- Demonstrated improvements on patients' and healthcare personnel's management [1], [2]
- Been cost-effective for hospitals and patients [1], [2], [3]
- More satisfied and educated patients [4], [5]

Some CPs include all perioperative phases of cancer and the follow-up period, but some may not include all the phases. Figure 2 shows distribution of selected publications over the period 2000 and 2017 and Graph 1 the classification of papers according to research method.

Fig 2: Number of publications on cancer patient CP from 2000-2017

## Conclusions

CP implementation has proved valuable most of the time. Nevertheless, during these studies the number of participants was reduced towards the end of trials. Reasons include complications and/or having more than two comorbidities. In this context, the main challenge is how to implement CPs that are personalized to patients' needs.

## References

[1] Chen et al., "The impact of clinical pathways on the practice of head and neck oncology surgery - The University of Texas M. D. Anderson Cancer Center Experience", JAMA Otolaryngology-Head & Neck Surgery, 2000

[2] J.B.Y. So et al., "Reduced hospital stay after implementation of a clinical pathway for radical gastrectomy for gastric cancer", Gastric Cancer, 2008

[3] S. Delaloge et al., "The challenge of rapid diagnosis in oncology: Diagnostic accuracy and cost analysis of a large-scale one-stop breast clinic", European Journal of Cancer, 2016

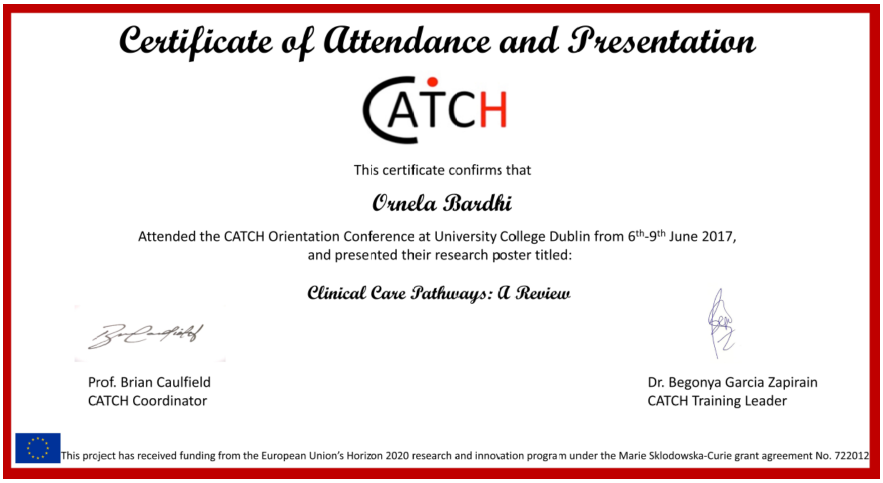
[4] M. Malmstrom et al., "Long-term experiences after oesophagectomy/ gastrectomy for cancer — A focus group study", International Journal of Nursing Studies, 2012

[5] R.C. Numan et al., "A clinical audit in a multidisciplinary care path for thoracic surgery: An instrument for continuous quality improvement", Lung Cancer, 2012

Contact:
**Email:** ornela.bardhi@deusto.es
**Twitter:** @ornelabardhi
**LinkedIn:** linkedin.com/in/ornelabardhi
**Website:** www.catchitn.eu

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 722012

***Figure 33.*** Poster for the CATCH 2017 Conference.



Certificate of Attendance and Presentation

CATCH

This certificate confirms that

Ornela Bardhi

Attended the CATCH Orientation Conference at University College Dublin from 6th-9th June 2017,
and presented their research poster titled:

Clinical Care Pathways: A Review

Prof. Brian Caulfield
CATCH Coordinator

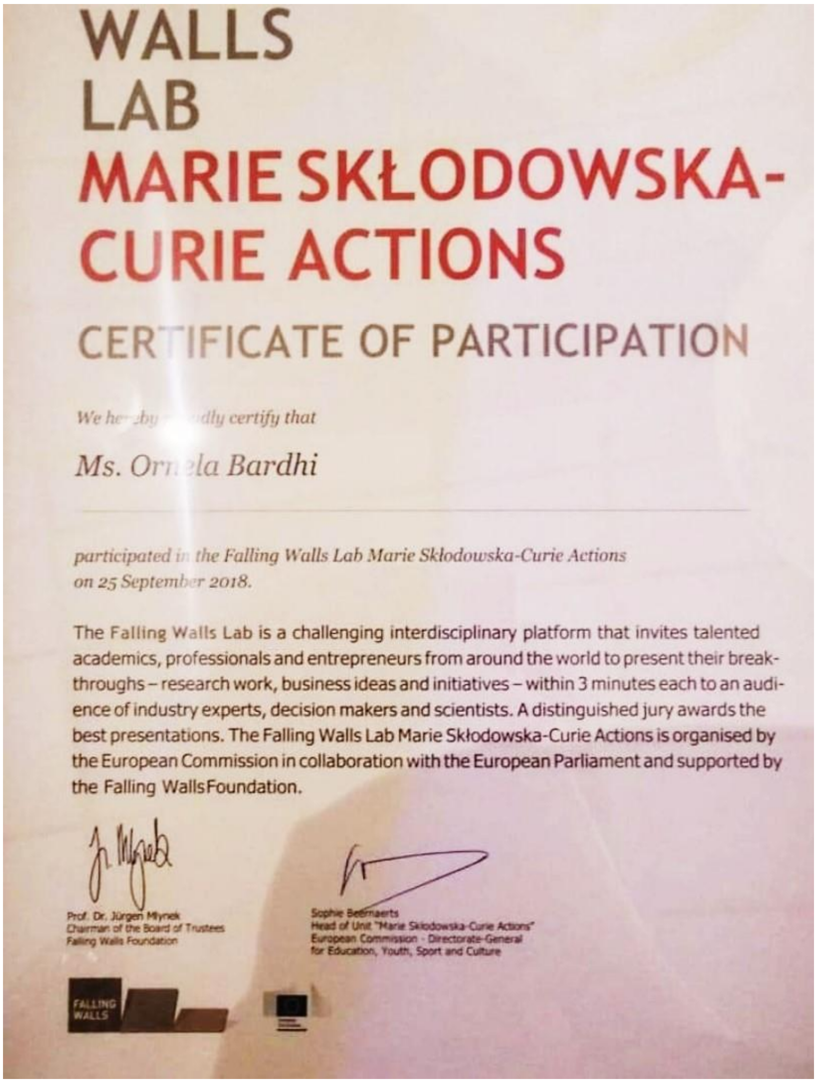
Dr. Begonya Garcia Zapirain
CATCH Training Leader

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No. 722012

***Figure 34.*** Certificate of attendance for CATCH Conference 2017.

***Table 36***. Falling Walls Lab - Details of the competition.

| | | | |
|---|---|---|---|
| **Title** | Breaking the Walls of Cancer Care Pathways | | |
| **Authors** | Ornela Bardhi | | |
| **Conference** | Falling Walls Lab | | |
| **Year** | 25 September 2018 | **Location** | Belgium |



WALLS
LAB
MARIE SKŁODOWSKA-CURIE ACTIONS

CERTIFICATE OF PARTICIPATION

*We hereby proudly certify that*

*Ms. Ornela Bardhi*

*participated in the Falling Walls Lab Marie Skłodowska-Curie Actions on 25 September 2018.*

The Falling Walls Lab is a challenging interdisciplinary platform that invites talented academics, professionals and entrepreneurs from around the world to present their breakthroughs – research work, business ideas and initiatives – within 3 minutes each to an audience of industry experts, decision makers and scientists. A distinguished jury awards the best presentations. The Falling Walls Lab Marie Skłodowska-Curie Actions is organised by the European Commission in collaboration with the European Parliament and supported by the Falling WallsFoundation.

Prof. Dr. Jürgen Mlynek
Chairman of the Board of Trustees
Falling Walls Foundation

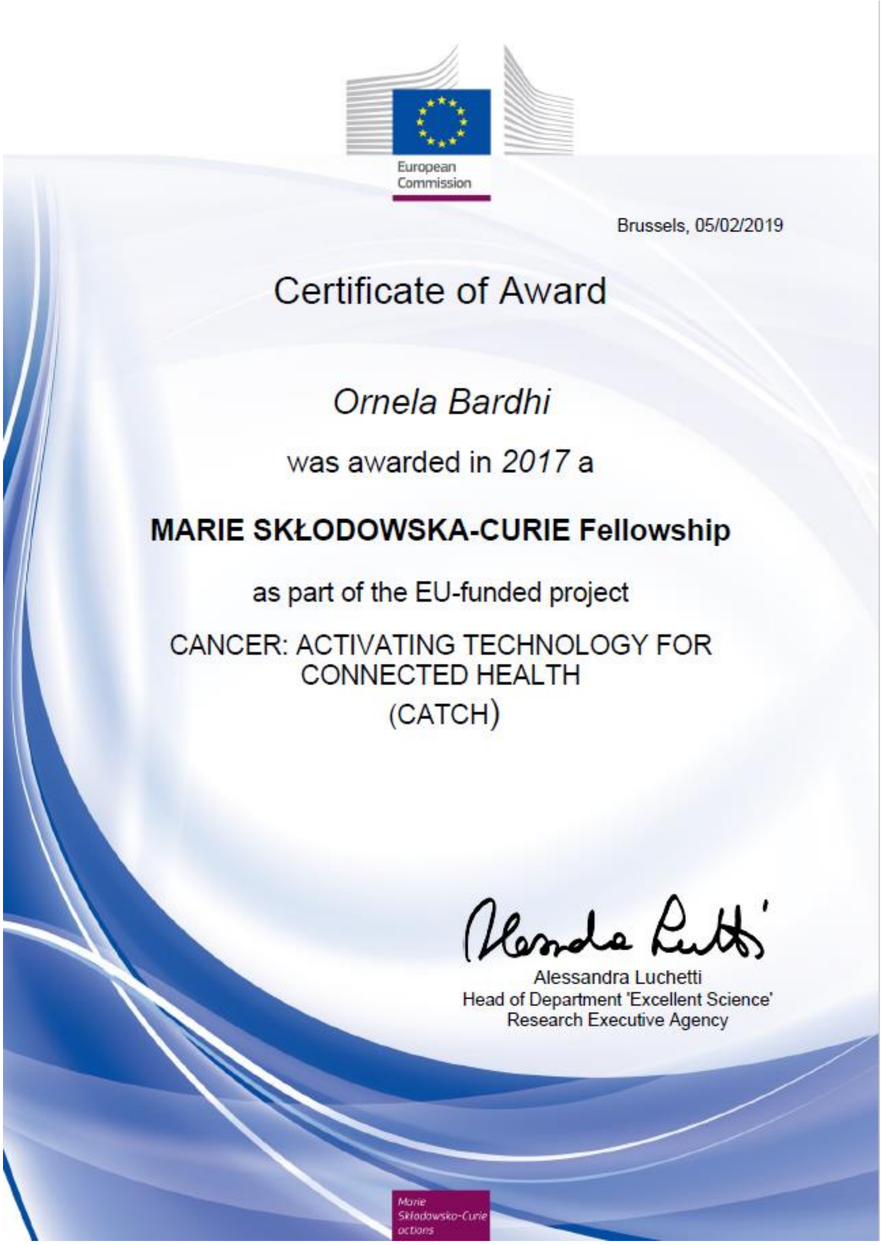
Sophie Beernaerts
Head of Unit "Marie Sklodowska-Curie Actions"
European Commission - Directorate-General
for Education, Youth, Sport and Culture

FALLING WALLS

***Figure 35.*** Certificate of Participation.

### 6.3.4 Research projects

**CATCH: Cancer – Activating Technology for Connected Health** a Horizon 2020 project funded by European Commission under the Marie Sklodowska Curie Action grant number 722012.
Start date: 9 January 2017
End date: 31 May 2020



European Commission

Brussels, 05/02/2019

Certificate of Award

*Ornela Bardhi*

was awarded in *2017* a

**MARIE SKŁODOWSKA-CURIE Fellowship**

as part of the EU-funded project

CANCER: ACTIVATING TECHNOLOGY FOR CONNECTED HEALTH
(CATCH)

Alessandra Luchetti
Head of Department 'Excellent Science'
Research Executive Agency

*Marie Skłodowska-Curie actions*

***Figure 36.*** Certificate of award for the Marie Sklodowska-Curie Fellowship.

## 6.4 Limitations and Recommendations for Future Research

Many aspects of the research have been taken into account when designing and implementing this research work; however, there are some limitations that are worth pointing out. Each of these limitations and future work are presented in their respective chapters; however, a broad overview is presented here as well.

When it comes to literature reviews, the majority of the publications are conducted in Europe, North America, and Australia. In order to not limit the scope to only these countries, more studies from Africa, Asia, and South America need to be published. A typical problem was encountered with the care pathways literature review. Although there are various international guidelines, there is very little information about local guidelines not only in the above-mentioned continents but worldwide. NCCN has an initiative to implement its cancer care programs in various countries (NCCN, 2021). These countries need to step up and test these guidelines and adapt them according to their population's needs. With digitalized care pathways being introduced by WHO and being implemented in academic settings or by private companies, the adoption of care programs in countries and hospitals not yet implemented would become easier.  An additional limitation stems from the confinement to papers published in English. Hence, one may not exclude the possibility that more studies exist in other languages.

Other limitations stem from the data collected for the first and second studies. Both these studies use medical data specific to those two countries, namely the Republic of Ireland and Finland. The care plan and care treatment vary a lot from one country to the other, so it is essential to have in mind that the results are for these specific populations, and one should generalize with caution. Another limitation is the data itself. Depending on the data you collect, you may receive different outcomes. The lifestyle information is limited in the second study compared to the first one, and the problem is that such data is usually not stored in EHRs.

Future work includes testing new hypotheses with the datasets used throughout this research work. Publish the data collected in the first study so that other researchers could reproduce my work or come up with new hypotheses. Test the same hypothesis for the second study but instead of machine learning algorithms, use deep learning algorithms and compare the results. And lastly, collaborate with hospitals in the Basque region to create a sizable dataset with colorectal polyps images and make it public so that the research field can progress.

## 6.5 Concluding Remarks

In summary, this dissertation covered the entire research process from the definition of the research hypotheses, the current state of the art in care pathways implementation to deep learning for colorectal polyp detection applications, implementation of various models for feature selection to making the predictions, and the analysis of the obtained results and how they outperform the current state-of-the-art contributions.

The aging population and the increase in incident cases of cancer diagnosis have been the motivation for conducting this research. International collaborations with people from different backgrounds, including doctors and patients from various countries, have been a crucial part of conducting this research so that a better quality of life throughout the treatment and afterward is provided to patients, or even prevent such diseases from happening.

# REFERENCES

Aasebø, U., Strøm, H.H., Postmyr, M., 2012. The Lean method as a clinical pathway facilitator in patients with lung cancer. Clin. Respir. J. 6, 169–174. https://doi.org/10.1111/j.1752-699X.2011.00271.x

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y., Zheng, X., 2016. TensorFlow: A System for Large-Scale Machine Learning. Proc. 12th USENIX Conf. Oper. Syst. Des. Implement., OSDI'16 265–283.

Akay, M.F., 2009. Support vector machines combined with feature selection for breast cancer diagnosis. Expert Syst. Appl. 36, 3240–3247. https://doi.org/10.1016/j.eswa.2008.01.009

Akechi, T., Uchida, M., Nakaguchi, T., Okuyama, T., Sakamoto, N., Toyama, T., Yamashita, H., 2015. Difference of patient's perceived need in breast cancer patients after diagnosis. Jpn. J. Clin. Oncol. 45, 75–80. https://doi.org/10.1093/jjco/hyu165

Al-Aidaroos, K.M., Bakar, A.A., Othman, Z., 2010. Naïve bayes variants in classification learning. pp. 276–281. https://doi.org/10.1109/INFRKM.2010.5466902

Alamoudi, O., Hamour, O.A., Mudawi, I., Khayyat, E., Batwail, N., Elhadd, T.A., 2011. Consensus-based management of differentiated thyroid cancer in a tertiary care set-up. Int. J. Surg. 9, 96–100. https://doi.org/10.1016/j.ijsu.2010.10.005

Aličković, E., Subasi, A., 2015. Breast cancer diagnosis using GA feature selection and Rotation Forest. Neural Comput. Appl. 28, 753–763. https://doi.org/10.1007/s00521-015-2103-9

All About Cancer, 2020. [Online]. Available on: www.allaboutcancer.fi. Accessed: 29 August 2020.

Amoon, M., Altameem, T., Altameem, A., 2020. Internet of things sensor assisted security and quality analysis for health care data sets using artificial intelligent based heuristic health management system. Measurement 161, 107861. https://doi.org/https://doi.org/10.1016/j.measurement.2020.107861

Ankerst, D.P., Jack, G., John, R.D., Amy, B., Harry, R., Brad, H.P., Cathy, T., Dipen, P., Robin, J.L., Ian, T., 2008. Predicting Prostate Cancer Risk Through Incorporation of Prostate Cancer Gene 3. J. Urol. 180, 1303–1308. https://doi.org/10.1016/j.juro.2008.06.038

Ankerst, D.P., Straubinger, J., Selig, K., Guerrios, L., Hoedt, A. De, Hernandez, J., Liss, M.A., Leach, R.J., Freedland, S.J., Kattan, M.W., Nam, R., Haese, A., Montorsi, F., Boorjian, S.A., Cooperberg, M.R., Poyet, C., Vertosick, E., Vickers, A.J., 2018. A Contemporary Prostate Biopsy Risk Calculator Based on Multiple Heterogeneous Cohorts. Eur. Urol. 74, 197–203. https://doi.org/10.1016/j.eururo.2018.05.003

Badrinarayanan, V., Kendall, A., Cipolla, R., 2017. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. - IEEE Trans. Pattern Anal. Mach. Intell. 39, 2481–2495. https://doi.org/10.1109/TPAMI.2016.2644615

Baffert, S., Hoang, H.L., Brédart, A., Asselain, B., Alran, S., Berseneff, H., Huchon, C., Trichot, C., Combes, A., Alves, K., Koskas, M., Nguyen, T., Roulot, A., Rouzier, R., Héquet, D., 2015. The patient-breast cancer care pathway: how could it be optimized? BMC Cancer 15, 394. https://doi.org/10.1186/s12885-015-1417-4

Bakitas, M., Lyons, K.D., Hegel, M.T., Balan, S., Brokaw, F.C., Seville, J., Hull, J.G., Li, Z., Tosteson, T.D., Byock, I.R., Ahles, T.A., 2009. Effects of a palliative care intervention on clinical outcomes in patients with advanced cancer: the Project ENABLE II randomized controlled trial. JAMA 302, 741–749. https://doi.org/10.1001/jama.2009.1198

Bakker, M.F., de Lange, S. V, Pijnappel, R.M., Mann, R.M., Peeters, P.H.M., Monninkhof, E.M., Emaus, M.J., Loo, C.E., Bisschops, R.H.C., Lobbes, M.B.I., de Jong, M.D.F., Duvivier, K.M., Veltman, J., Karssemeijer, N., de Koning, H.J., van Diest, P.J., Mali, W.P.T.M., van den Bosch, M.A.A.J., Veldhuis, W.B., van Gils, C.H., 2019. Supplemental MRI Screening for Women with Extremely Dense Breast Tissue. N. Engl. J. Med. 381, 2091–2102. https://doi.org/10.1056/NEJMoa1903986

Baldwin, D.R., Duffy, S.W., Wald, N.J., Page, R., Hansell, D.M., Field, J.K., 2011. UK Lung Screen (UKLS) nodule management protocol: modelling of a single screen randomised controlled trial of low-dose CT screening for lung cancer. Thorax 66, 308–313. https://doi.org/10.1136/thx.2010.152066

Barrett, D., Noble, H., 2019. What are cohort studies? Evid. Based. Nurs. 22, 95–96. https://doi.org/10.1136/ebnurs-2019-103183

Barrett, J., Sharp, D.J., Stapley, S., Stabb, C., Hamilton, W., 2010. Pathways to the diagnosis of ovarian cancer in the UK: A cohort study in primary care. BJOG An Int. J. Obstet. Gynaecol. 117, 610–614. https://doi.org/10.1111/j.1471-0528.2010.02499.x

Barry, M., Weber, W.P., Lee, S., Mazzella, A., Sclafani, L.M., 2012. Enhancing the clinical pathway for patients undergoing axillary lymph node dissection. Breast 21, 440–443. https://doi.org/10.1016/j.breast.2011.10.002

Bassett, J.K., Severi, G., Baglietto, L., MacInnis, R.J., Hoang, H.N., Hopper, J.L., English, D.R., Giles, G.G., 2012. Weight change and prostate cancer incidence and mortality. Int. J. Cancer 131, 1711–1719. https://doi.org/10.1002/ijc.27414

Bernal, J., Sánchez, F.J., Fernández-Esparrach, G., Gil, D., Rodríguez, C., Vilariño, F., 2015. WM-DOVA maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians. Comput. Med. Imaging Graph. 43, 99–111. https://doi.org/https://doi.org/10.1016/j.compmedimag.2015.02.007

Bernal, J., Sánchez, J., Vilariño, F., 2012. Towards automatic polyp detection with a polyp appearance model. Pattern Recognit. 45, 3166–3182. https://doi.org/https://doi.org/10.1016/j.patcog.2012.03.002

Bhatnagar, A.K., Beriwal, S., Heron, D.E., Flickinger, J.C., Deutsch, M., Huq, M.S., Sontag, M., Shogan, J., 2009. Initial outcomes analysis for large multicenter integrated cancer network implementation of intensity modulated radiation therapy for breast cancer. Breast J. 15, 468–474. https://doi.org/10.1111/j.1524-4741.2009.00761.x

Bhayani, S.B., Pavlovich, C.P., Hsu, T.S., Sullivan, W., Su, L.M., 2003. Prospective comparison of short-term convalescence: Laparoscopic radical prostatectomy versus open radical retropubic prostatectomy. Urology 61, 612–616. https://doi.org/10.1016/S0090-4295(02)02416-0

Birch, E., van Bruwaene, S., Everaerts, W., Schubach, K., Bush, M., Krishnasamy, M., Moon, D.A., Goad, J., Lawrentschuk, N., Murphy, D.G., 2016. Developing and evaluating Robocare; an innovative, nurse-led robotic prostatectomy care pathway. Eur. J. Oncol. Nurs. 21, 120–125. https://doi.org/10.1016/j.ejon.2016.02.002

Blagden, S., Simpson, C., Limmer, M., 2020. Bowel cancer screening in an English prison: a qualitative service evaluation. Public Health 180, 46–50. https://doi.org/10.1016/j.puhe.2019.10.024

Bond, C., Bruhn, H., de Bont, A., van Exel, J., Busse, R., Sutton, M., Elliott, R., 2016. The iMpact on practice, oUtcomes and costs of New roles for health pROfeSsionals: a study protocol for MUNROS. BMJ Open 6, e010511. https://doi.org/10.1136/bmjopen-2015-010511

Bonnington, S.N., Rutter, M.D., 2016. Surveillance of colonic polyps: Are we getting it right? World J. Gastroenterol. 22, 1925–1934. https://doi.org/10.3748/wjg.v22.i6.1925

Bratt, O., 2002. Hereditary prostate cancer: clinical aspects. J. Urol. 168, 906–913. https://doi.org/S0022-5347(05)64541-7

Buitinck, L., Louppe, G., Blondel, M., 2013. API design for machine learning software: experiences from the scikit-learn project.

Byrne, M.F., Chapados, N., Soudan, F., Oertel, C., Pérez, M.L., Kelly, R., Iqbal, N., Chandelier, F., Rex, D.K., 2019. Real-time differentiation of adenomatous and hyperplastic diminutive colorectal polyps during analysis of unaltered videos of standard colonoscopy using a deep learning model. Gut 68, 94. https://doi.org/10.1136/gutjnl-2017-314547

Carioli, G., Bertuccio, P., Boffetta, P., Levi, F., Vecchia, C. La, Negri, E., Malvezzi, M., 2020. European cancer mortality predictions for the year 2020 with a focus on prostate cancer. Ann. Oncol. 31, 650–658. https://doi.org/10.1016/j.annonc.2020.02.009

Carneiro, G., Zorron Cheng Tao Pu, L., Singh, R., Burt, A., 2020. Deep learning uncertainty and confidence calibration for the five-class polyp classification from colonoscopy. Med. Image Anal. 62, 101653. https://doi.org/https://doi.org/10.1016/j.media.2020.101653

Caron, F., Vanthienen, J., Vanhaecht, K., Limbergen, E. Van, De Weerdt, J., Baesens, B., 2014. Monitoring care processes in the gynecologic oncology department. Comput. Biol. Med. 44, 88–96. https://doi.org/10.1016/j.compbiomed.2013.10.015

Catanuto, G., Pappalardo, F., Rocco, N., Leotta, M., Ursino, V., Chiodini, P., Buggi, F., Folli, S., Catalano, F., Nava, M.B., 2016. Formal analysis of the surgical pathway and development of a new software tool to assist surgeons in the decision making in primary breast surgery. The Breast 29, 74–81. https://doi.org/10.1016/j.breast.2016.06.004

Center for Medicare and Medicaid Services. Electronic Health Records. [Online]. Available on: https://www.cms.gov/Medicare/E-Health/EHealthRecords, Accessed: June 2021.

Centre for Policy on Ageing. Glossary of health and social care. [Online]. Available on: http://www.cpa.org.uk/glossary/glossary.html#C, Accessed: April 2021.

Chajès, V., Romieu, I., 2014. Nutrition and breast cancer. Maturitas 77, 7–11. https://doi.org/10.1016/j.maturitas.2013.10.004

Chen, A.Y., Callender, D., Mansyur, C., Reyna, K.M., Limitone, E., Goepfert, H., 2000. The impact of clinical pathways on the practice of head and neck oncologic surgery: the University of Texas M. D. Anderson Cancer Center Experience. Arch. Otolaryngol. Head neck Surg. 126, 322–6.

Chen, T., Guestrin, C., 2016. XGBoost: A Scalable Tree Boosting System, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD ’16. Association for Computing Machinery, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785

Chiang, H.A., Cheng, P.J., Speed, J.M., Steinberg, J., Najjar, P.A., Steele, G.S., Trinh, Q.-D., Eswara, J.R., Chang, S.L., Kibel, A.S., Stopfkuchen-Evans, M.F., Preston, M.A., 2020. Implementation of a Perioperative Venous Thromboembolism Prophylaxis Program for Patients Undergoing Radical Cystectomy on an Enhanced Recovery After Surgery Protocol.

Eur. Urol. Focus 6, 74–80. https://doi.org/10.1016/j.euf.2018.08.025

Colonna, S., Sweetenham, J., Burgon, T.B., Buys, S.S., Lynch, R., Au, T., Johnson, E., Kubal, T., Paculdo, D., Acelajado, M.C., Peabody, J.W., 2019. A Better Pathway? Building Consensus and Engaging Providers with Feedback to Improve and Standardize Cancer Care. Clin. Breast Cancer 19, e376–e384. https://doi.org/10.1016/j.clbc.2018.12.010

Compagna, R., Aprea, G., De Rosa, D., Gentile, M., Cestaro, G., Vigliotti, G., Bianco, T., Massa, G., Amato, M., Massa, S., Amato, B., 2014. Fast track for elderly patients: Is it feasible for colorectal surgery? Int. J. Surg. 12, S20–S22. https://doi.org/10.1016/j.ijsu.2014.08.389

Corrao, G., Rea, F., Di Felice, E., Di Martino, M., Davoli, M., Merlino, L., Carle, F., De Palma, R., 2020. Influence of adherence with guideline-driven recommendations on survival in women operated for breast cancer: Real-life evidence from Italy. Breast 53, 51–58. https://doi.org/10.1016/j.breast.2020.06.010

Cortes, C., Vapnik, V., 1995. Support-Vector Networks. Mach. Learn. 20, 273–297. https://doi.org/10.1023/A:1022627411411

Cover, T., Hart, P., 1967. Nearest neighbor pattern classification. IEEE Trans. Inf. Theory. https://doi.org/10.1109/TIT.1967.1053964

Creak, A., 2020. Prospective Cohort of Referrals to a Cancer of Unknown Primary Clinic, including Direct Access from Primary Care. Clin. Oncol. (R. Coll. Radiol). 32, e87–e92. https://doi.org/10.1016/j.clon.2019.09.059

Dandamudi, A., Tommie, J., Nommsen-Rivers, L., Couch, S., 2018. Dietary Patterns and Breast Cancer Risk: A Systematic Review. Anticancer Res. 38, 3209–3222. https://doi.org/10.21873/anticanres.12586

Dang-Tan, T., Trottier, H., Mery, L.S., Morrison, H.I., Barr, R.D., Greenberg, M.L., Franco, E.L., 2010. Determinants of delays in treatment initiation in children and adolescents diagnosed with leukemia or lymphoma in Canada. Int. J. Cancer 126, 1936–1943. https://doi.org/10.1002/ijc.24906

Dautremont, J.F., Rudmik, L.R., Nakoneshny, S.C., Chandarana, S.P., Matthews, T.W., Schrag, C., Fick, G.H., Dort, J.C., 2016. Understanding the impact of a clinical care pathway for major head and neck cancer resection on postdischarge healthcare utilization. Head Neck 38 Suppl 1, E1216-20. https://doi.org/10.1002/hed.24196

Davies, N.M., Gaunt, T.R., Lewis, S.J., Holly, J., Donovan, J.L., Hamdy, F.C., Kemp, J.P., Eeles, R., Easton, D., Kote-Jarai, Z., Olama, A.A. Al, Benlloch, S., Muir, K., Giles, G.G., Wiklund, F., Gronberg, H., Haiman, C.A., Schleutker, J., Nordestgaard, B.G., Travis, R.C., Neal, D., Pashayan, N., Khaw, K.-T., Stanford, J.L., Blot, W.J., Thibodeau, S., Maier, C., Kibel, A.S., Cybulski, C., Cannon-Albright, L., Brenner, H., Park, J., Kaneva, R., Batra, J., Teixeira, M.R., Pandha, H., Lathrop, M., Smith, G.D., Martin, R.M., 2015. The effects of height and BMI on prostate cancer incidence and mortality: a Mendelian randomization study in 20,848 cases and 20,214 controls from the PRACTICAL consortium. Cancer Causes Control 26, 1603–1616. https://doi.org/10.1007/s10552-015-0654-9

Davis, J.W., Pisters, L.L., Doviak, M.J., Donat, S.M., 2002. Early nasogastric tube removal combined with metoclopramide after postchemotherapy retroperitoneal lymph node dissection for metastatic testicular nonseminomatous germ cell tumor. Urology 59, 579–583. https://doi.org/S0090429501016545

de Kok, M., van der Weijden, T., Kessels, A.G.H., Dirksen, C.D., Sixma, H.J.M., van de Velde, C.J.H., Roukema, J.A., Finaly-Marais, C., van der Ent, F.W.C., von Meyenfeldt, M.F., 2010. Patients' opinions on quality of care before and after implementation of a short stay programme following breast cancer surgery. Breast 19, 404–409. https://doi.org/10.1016/j.breast.2010.04.002

Delaloge, S., Bonastre, J., Borget, I., Garbay, J.R., Fontenay, R., Boinon, D., Saghatchian, M., Mathieu, M.C., Mazouni, C., Rivera, S., Uzan, C., André, F., Dromain, C., Boyer, B., Pistilli, B., Azoulay, S., Rimareix, F., Bayou, E.H., Sarfati, B., Caron, H., Ghouadni, A., Leymarie, N., Canale, S., Mons, M., Arfi-Rouche, J., Arnedos, M., Suciu, V., Vielh, P., Balleyguier, C., 2016. The challenge of rapid diagnosis in oncology: Diagnostic accuracy and cost analysis of a large-scale one-stop breast clinic. Eur. J. Cancer 66, 131–137. https://doi.org/10.1016/j.ejca.2016.06.021

Desandes, E., Bonnay, S., Berger, C., Brugieres, L., Demeocq, F., Laurence, V., Sommelet, D., Tron, I., Clavel, J., Lacour, B., 2012. Pathways of Care for Adolescent Patients With Cancer in France from 2006 to 2007 Emmanuel. Pediatr Blood Cancer 924–929. https://doi.org/10.1002/pbc

Dyrop, H.B., Safwat, A., Vedsted, P., Maretty-Nielsen, K., Hansen, B.H., Jørgensen, P.H., Baad-Hansen, T., Bünger, C., Keller, J., 2013. Cancer Patient Pathways shortens waiting times and accelerates the diagnostic process of suspected sarcoma patients in Denmark. Health Policy 113, 110–117. https://doi.org/10.1016/j.healthpol.2013.09.012

ESMO. About ESMO. [Online]. Available on: https://www.esmo.org/about-esmo. Accessed: June 2021.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G., Thrun, S., Dean, J., 2019. A guide to deep learning in healthcare. Nat. Med. 25, 24–29. https://doi.org/10.1038/s41591-018-0316-z

Esteva, M., Ruiz, A., Ramos, M., Casamitjana, M., Sánchez-Calavera, M.A., González-Luján, L., Pita-Fernández, S., Leiva, A., Pértega-Díaz, S., Costa-Alcaraz, A.M., Macià, F., Espí, A., Segura, J.M., Lafita, S., Novella, M.T., Yus, C., Oliván, B., Cabeza, E., Seoane-Pillado, T., López-Calviño, B., Llobera, J., Martín-Rabadán, M.M., Teresa Novella, M., Ripoll, J., Manzano, H., Amengual, I., Forteza, A., Company, M., De Lluch Bennassar, M., Sánchez, M.A., Magallón, R., Olivan, B., Maciá, F., Pita, S., Louro, A., Serrano, J., Arnal, F., González-Santamaría, P., Seoane, T., González-Lujan, L., Costa-Alcaraz, A., Espí, A., Bosca, M.M., Balza, N., Villagrasa, R.A., Vázquez, J.F., González-Timoneda, y. A., 2014. Age differences in presentation, diagnosis pathway and management of colorectal cancer. Cancer Epidemiol. 38, 346–353. https://doi.org/10.1016/j.canep.2014.05.002

European Pathway Association. What is care pathways?. [Online]. Available on: http://e-p-a.org/care-pathways/, Accessed: April 2021.

Evans, R.S., 2016. Electronic Health Records: Then, Now, and in the Future. Yearb. Med. Inform. Suppl 1, 48. https://doi.org/10.15265/IYS-2016-s006

Everson, J., Rubin, J.C., Friedman, C.P., 2020. Reconsidering hospital EHR adoption at the dawn of HITECH: implications of the reported 9% adoption of a “basic” EHR. J. Am. Med. Informatics Assoc. 27, 1198–1205. https://doi.org/10.1093/jamia/ocaa090

Falborg, A.Z., Vedsted, P., Menon, U., Weller, D., Neal, R.D., Reguilon, I., Harrison, S., Jensen, H., 2020. Agreement between questionnaires and registry data on routes to diagnosis and milestone dates of the cancer diagnostic pathway. Cancer Epidemiol. 65, 101690. https://doi.org/10.1016/j.canep.2020.101690

Fasola, G., Rizzato, S., Merlo, V., Aita, M., Ceschia, T., Giacomuzzi, F., Lugatti, E., Meduri, S., Morelli, A., Rocco, M., Tozzi, V., 2012. Adopting Integrated Care Pathways in Non–Small-Cell Lung Cancer: From Theory to Practice. J. Thorac. Oncol. 7, 1283–1290. https://doi.org/10.1097/JTO.0b013e318257fbfe

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27, 861–874. https://doi.org/https://doi.org/10.1016/j.patrec.2005.10.010

Finnish Cancer Registry, 2020. [Online]. Available on: https://cancerregistry.fi/. Accessed: 20 August 2020.

Fioretti, F., Tavani, A., Bosetti, C., Vecchia, C. La, Negri, E., Barbone, F., Talamini, R., Franceschi, S., 1999. Risk factors for breast cancer in nulliparous women. Br. J. Cancer 79, 1923–1928. https://doi.org/10.1038/sj.bjc.6690306

Fitzal, F., Bjelic-Radisic, V., Knauer, M., Steger, G., Hubalek, M., Balic, M., Singer, C., Bartsch, R., Schrenk, P., Soelkner, L., Greil, R., Gnant, M., 2019. Impact of Breast Surgery in Primary Metastasized Breast Cancer: Outcomes of the Prospective Randomized Phase III ABCSG-28 POSYTIVE Trial. Ann. Surg. 269, 1163–1169. https://doi.org/10.1097/SLA.0000000000002771

Gao, J., Guo, Y., Sun, Y., Qu, G., 2020. Application of Deep Learning for Early Screening of Colorectal Precancerous Lesions under White Light Endoscopy. Comput. Math. Methods Med. 2020, 8374317. https://doi.org/10.1155/2020/8374317

Gerardi, M.A., Santillan, A., Meisner, B., Zahurak, M.L., Diaz Montes, T.P., Giuntoli, R.L., Bristow, R.E., 2008. A clinical pathway for patients undergoing primary cytoreductive surgery with rectosigmoid colectomy for advanced ovarian and primary peritoneal cancers. Gynecol. Oncol. 108, 282–286. https://doi.org/10.1016/j.ygyno.2007.10.014

Gong, I.Y., Fox, N.S., Huang, V., Boutros, P.C., 2018. Prediction of early breast cancer patient survival using ensembles of hypoxia signatures. PLoS One 13, e0204123.

Grönberg, H., Damber, L., Damber, J.E., 1996. Familial prostate cancer in Sweden. A nationwide register cohort study. Cancer 77, 138–143. https://doi.org/10.1002/(SICI)1097-0142(19960101)77:1<138::AID-CNCR23>3.0.CO;2-5

Grosso, G., Bella, F., Godos, J., Sciacca, S., Del Rio, D., Ray, S., Galvano, F., Giovannucci, E.L., 2017. Possible role of diet in cancer: systematic review and multiple meta-analyses of dietary patterns, lifestyle factors, and cancer risk. Nutr. Rev. 75, 405–419. https://doi.org/10.1093/nutrit/nux012

Guyon, I., Elisseeff, A. ’e, 2003. An Introduction to Variable and Feature Selection. J.Mach.Learn.Res. 3, 1157–1182.

Halawi, R., Saad Aldin, E., Baydoun, A., Dbouk, H., Nahleh, Z., Nasser, Z., Tfayli, A., 2012. Physical symptom profile for adult cancer inpatients at a Lebanese cancer unit. Eur. J. Intern. Med. 23, e185–e189. https://doi.org/10.1016/j.ejim.2012.08.018

Hameed Khaliq, I., Mahmood, H.Z., Sarfraz, M.D., Masood Gondal, K., Zaman, S., 2019. Pathways to care for patients in Pakistan experiencing signs or symptoms of breast cancer. Breast 46, 40–47. https://doi.org/10.1016/j.breast.2019.04.005

Harrison, S., Tilling, K., Turner, E.L., Martin, R.M., Lennon, R., Lane, J.A., Donovan, J.L., Hamdy, F.C., Neal, D.E., Bosch, J.L.H.R., Jones, H.E., 2020. Systematic review and meta-analysis of the associations between body mass index, prostate cancer, advanced prostate cancer, and prostate-specific antigen. Cancer Causes Control 31, 431–449. https://doi.org/10.1007/s10552-020-01291-3

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep Residual Learning for Image Recognition, in: - 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 770–778. https://doi.org/10.1109/CVPR.2016.90

Helsinki: Finnish Medical Association Duodecim, 2014. [Online]. Available on: www.kaypahoito.fi. Accessed: 02 August 2020.

Hochreiter, S., Bengio, Y., Frasconi, P., Schmidhuber, J., n.d. Gradient Flow in Recurrent Nets: the Difficulty of Learning Long-Term Dependencies.

Hori, S., Butler, E., McLoughlin, J., 2011. Prostate cancer and diet: food for thought? BJU Int. 107, 1348–1359. https://doi.org/10.1111/j.1464-410X.2010.09897.x

Huang, C.-L., Liao, H.-C., Chen, M.-C., 2008. Prediction model building and feature selection with support vector machines in breast cancer diagnosis. Expert Syst. Appl. 34, 578–587. https://doi.org/https://doi.org/10.1016/j.eswa.2006.09.041

Huang, Z., Dong, W., Ji, L., Gan, C., Lu, X., Duan, H., 2014. Discovery of clinical pathway patterns from event logs using probabilistic topic models. J. Biomed. Inform. 47, 39–57. https://doi.org/10.1016/j.jbi.2013.09.003

Huang, Z., Lu, X., Duan, H., Fan, W., 2013. Summarizing clinical pathways from event logs. J. Biomed. Inform. 46, 111–127. https://doi.org/10.1016/j.jbi.2012.10.001

ICANotes, 2019. A History of EHR Through the Years. [Online]. Available on: https://www.icanotes.com/2019/04/16/a-history-of-ehr-through-the-years/, Accessed: June 2021.

Incorvati, J.A., Shah, S., Mu, Y., Lu, J., 2013. Targeted therapy for HER2 positive breast cancer. J. Hematol. Oncol. 6, 38. https://doi.org/10.1186/1756-8722-6-38

International Agency for Research on Cancer, WHO. Cancer Tomorrow. [Online]. Available on: https://gco.iarc.fr/tomorrow/en/dataviz/bars?mode=population, Accessed: April 2021.

Jakobsen, J.K., Jensen, J.B., 2016. DaPeCa-2: Implementation of fast-track clinical pathways for penile cancer shortens waiting time and accelerates the diagnostic process--A comparative before-and-after study in a tertiary referral centre in Denmark. Scand. J. Urol. 50, 80–87. https://doi.org/10.3109/21681805.2015.1077472

Joris, J., Hans, G., Coimbra, C., Decker, E., Kaba, A., 2020. Elderly patients over 70 years benefit from enhanced recovery programme after colorectal surgery as much as younger patients. J. Visc. Surg. 157, 23–31. https://doi.org/10.1016/j.jviscsurg.2019.07.011

Jung, A., 2017. imgaug 0.2.5. [Online]. Available on: http://imgaug.readthedocs.io/en/latest, Accessed: 10 June 2020.

Kay, A.H., Venn, M., Urban, R., Gray, H.J., Goff, B., 2020. Postoperative narcotic use in patients with ovarian cancer on an Enhanced Recovery After Surgery (ERAS) pathway. Gynecol. Oncol. 156, 624–628.

https://doi.org/10.1016/j.ygyno.2019.12.018

Khankari, N., Shu, X.-O., Wen, W., Kraft, P., Lindström, S., Peters, U., Schildkraut, J., Schumacher, F., Bofetta, P., Risch, A., Bickeböller, H., Amos, C., Easton, D., Eeles, R., Eeles, R., Gruber, S., Haiman, C., Hunter, D., Chanock, S., Pierce, B., Zheng, W., Discovery, B., 2016. Association between Adult Height and Risk of Colorectal, Lung, and Prostate Cancer: Results from Meta-analyses of Prospective Studies and Mendelian Randomization Analyses. PLoS Med. https://doi.org/10.1371/journal.pmed.1002118

Kim, H.S., Kim, S.O., Kim, B.S., 2015. Use of a clinical pathway in laparoscopic gastrectomy for gastric cancer. World J. Gastroenterol. 21, 13507–13517. https://doi.org/10.3748/wjg.v21.i48.13507

Kinsman, L., Rotter, T., James, E., Snow, P., Willis, J., 2010. What is a clinical pathway? Development of a definition to inform the debate. BMC Med. 8, 31. https://doi.org/10.1186/1741-7015-8-31

Klinkhammer-Schalke, M., Koller, M., Ehret, C., Steinger, B., Ernst, B., Wyatt, J.C., Hofstädter, F., Lorenz, W., QoL, S.G.R., 2008. Implementing a system of quality-of-life diagnosis and therapy for breast cancer patients: results of an exploratory trial as a prerequisite for a subsequent RCT. Br. J. Cancer 99, 415–422. https://doi.org/10.1038/sj.bjc.6604505

Klinkhammer-Schalke, M., Koller, M., Steinger, B., Ehret, C., Ernst, B., Wyatt, J.C., Hofstädter, F., Lorenz, W., 2012. Direct improvement of quality of life using a tailored quality of life diagnosis and therapy pathway: Randomised trial in 200 women with breast cancer. Br. J. Cancer 106, 826–838. https://doi.org/10.1038/bjc.2012.4

Klinkhammer-Schalke, M., Steinger, B., Koller, M., Zeman, F., Fürst, A., Gumpp, J., Obermaier, R., Piso, P., Lindberg-Scharf, P., 2020. Diagnosing deficits in quality of life and providing tailored therapeutic options: Results of a randomised trial in 220 patients with colorectal cancer. Eur. J. Cancer 130, 102–113. https://doi.org/10.1016/j.ejca.2020.01.025

Korbar, B., Olofson, A.M., Miraflor, A.P., Nicka, C.M., Suriawinata, M.A., Torresani, L., Suriawinata, A.A., Hassanpour, S., 2017. Deep Learning for Classification of Colorectal Polyps on Whole-slide Images. J. Pathol. Inform. 8, 30. https://doi.org/10.4103/jpi.jpi_34_17

Krizhevsky, A., Sutskever, I., Hinton, G.E., 2017. ImageNet Classification with Deep Convolutional Neural Networks. Commun.ACM 60, 84–90. https://doi.org/10.1145/3065386

Larentzakis, A., O'Dwyer, S.T., Becker, J., Shuweihdi, F., Aziz, O., Selvasekar, C.R., Fulford, P., Renehan, A.G., Wilson, M., 2019. Referral pathways and outcome of patients with colorectal peritoneal metastasis (CRPM). Eur. J. Surg. Oncol. J. Eur. Soc. Surg. Oncol. Br. Assoc. Surg.

Oncol. 45, 2310–2315. https://doi.org/10.1016/j.ejso.2019.07.008

Laurent-Badr, Q., Barbe, C., Brugel, M., Hautefeuille, V., Volet, J., Grelet, S., Desot, E., Botsen, D., Deguelte, S., Pitta, A., Abdelli, N., Brasseur, M., De Mestier, L., Neuzillet, C., Bouché, O., 2020. Time intervals to diagnosis and chemotherapy do not influence survival outcome in patients with advanced pancreatic adenocarcinoma. Dig. liver Dis. Off. J. Ital. Soc. Gastroenterol. Ital. Assoc. Study Liver 52, 658–667. https://doi.org/10.1016/j.dld.2020.03.014

Le, Q. V, 2015. A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks.

Lecun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. - Proc. IEEE. https://doi.org/10.1109/5.726791

Lee, J., Park, S.W., Kim, Y.S., Lee, K.J., Sung, H., Song, P.H., Yoon, W.J., Moon, J.S., 2017. Risk factors of missed colorectal lesions after colonoscopy. Medicine (Baltimore). 96. https://doi.org/10.1097/MD.0000000000007468

Lefeuvre, D., Le Bihan-Benjamin, C., Pauporté, I., Medioni, J., Bousquet, P.-J., 2017. French Medico-Administrative Data to Identify the Care Pathways of Women With Breast Cancer. Clin. Breast Cancer 17, e191–e197. https://doi.org/10.1016/j.clbc.2017.01.008

Lequan, Y., Hao, C., Qi, D., Jing, Q., Ann, H.P., 2017. Integrating Online and Offline Three-Dimensional Deep Learning for Automated Polyp Detection in Colonoscopy Videos. IEEE J. Biomed. Heal. informatics 21, 65–75. https://doi.org/10.1109/JBHI.2016.2637004

Liang, J., Li, Y., Zhang, Z., Shen, D., Xu, J., Zheng, X., Wang, T., Tang, B., Lei, J., Zhang, J., 2021. Adoption of Electronic Health Records (EHRs) in China During the Past 10 Years: Consecutive Survey Data Analysis and Comparison of Sino-American Challenges and Experiences. J Med Internet Res 23, e24813. https://doi.org/10.2196/24813

Lichtenstein, P., Holm, N. V, Verkasalo, P.K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., Hemminki, K., 2000. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N. Engl. J. Med. 343, 78–85. https://doi.org/10.1056/NEJM200007133430201

Lieberman, D., 2005. Quality and colonoscopy: a new imperative. Gastrointest. Endosc. 61, 392–394. https://doi.org/10.1016/S0016-5107(05)00133-1

Lindop, E., Cannon, S., 2001. Experiences of women with a diagnosis of breast cancer: a clinical pathway approach. Eur. J. Oncol. Nurs. 5, 91–99. https://doi.org/https://doi.org/10.1054/ejon.2000.0116

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A.W.M., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88. https://doi.org/https://doi.org/10.1016/j.media.2017.07.005

Liu, Y., Hu, F., Li, D., Wang, F., Zhu, L., Chen, W., Ge, J., An, R., Zhao, Y., 2011. Does physical activity reduce the risk of prostate cancer? A systematic review and meta-analysis. Eur. Urol. 60, 1029–1044. https://doi.org/10.1016/j.eururo.2011.07.007

Lorena, A.C., Jacintho, L.F.O., Siqueira, M.F., Giovanni, R. De, Lohmann, L.G., de Carvalho, A.C.P.L.F., Yamamoto, M., 2011. Comparing machine learning classifiers in potential distribution modelling. Expert Syst. Appl. 38, 5268–5275. https://doi.org/10.1016/j.eswa.2010.10.031

Lyhne, N.M., Christensen, A., Alanin, M.C., Bruun, M.T., Jung, T.H., Bruhn, M.A., Jespersen, J.B.B., Kristensen, C.A., Andersen, E., Godballe, C., Buchwald, C., Bundgaard, T., Johansen, J., Lambertsen, K., Primdahl, H., Toustrup, K., Sørensen, J.A., Overgaard, J., Grau, C., 2013. Waiting times for diagnosis and treatment of head and neck cancer in Denmark in 2010 compared to 1992 and 2002. Eur. J. Cancer 49, 1627–1633. https://doi.org/10.1016/j.ejca.2012.11.034

Lynch, S.M., Handorf, E., Sorice, K.A., Blackman, E., Bealin, L., Giri, V.N., Obeid, E., Ragin, C., Daly, M., 2020. The effect of neighborhood social environment on prostate cancer development in black and white men at high risk for prostate cancer. PLoS One 15, e0237332.

Macacu, A., Autier, P., Boniol, M., Boyle, P., 2015. Active and passive smoking and risk of breast cancer: a meta-analysis. Breast Cancer Res. Treat. 154, 213–224. https://doi.org/10.1007/s10549-015-3628-4

MacPherson, R., Benamore, R., Panakis, N., Sayeed, R., Breen, D., Bradley, K., Carter, R., Baldwin, D., Craig, J., Gleeson, F., 2012. A proposed new imaging pathway for patients with suspected lung cancer. Clin. Radiol. 67, 564–573. https://doi.org/10.1016/j.crad.2011.10.032

Maher, J., McConnell, H., 2011. New pathways of care for cancer survivors: Adding the numbers. Br. J. Cancer 105, S5–S10. https://doi.org/10.1038/bjc.2011.417

Mahmood, F., Durr, N.J., 2018. Deep learning and conditional random fields-based depth estimation and topographical reconstruction from conventional endoscopy. Med. Image Anal. 48, 230–243. https://doi.org/https://doi.org/10.1016/j.media.2018.06.005

Malmström, M., Ivarsson, B., Klefsgård, R., Persson, K., Jakobsson, U., Johansson, J., 2016. The effect of a nurse led telephone supportive care programme on patients' quality of life, received information and health care contacts after oesophageal cancer surgery—A six month RCT-follow-up study. Int. J. Nurs. Stud. 64, 86–95. https://doi.org/10.1016/j.ijnurstu.2016.09.009

Markar, S.R., Schmidt, H., Kunz, S., Bodnar, A., Hubka, M., Low, D.E., 2014. Evolution of Standardized Clinical Pathways: Refining Multidisciplinary Care and Process to Improve Outcomes of the Surgical Treatment of Esophageal Cancer. J. Gastrointest. Surg. 18, 1238–1246. https://doi.org/10.1007/s11605-014-2520-6

Mattson, J., Vehmanen, L., 2016. [Male breast cancer]. Duodecim. 132, 627–631.

MDCalc. Gail Model for Breast Cancer Risk. [Online]. Available on: https://www.mdcalc.com/gail-model-breast-cancer-risk. Accessed: 20 August 2020.

Mehl, G., Tunçalp, Ö., Ratanaprayul, N., Tamrat, T., Barreix, M., Lowrance, D., Bartolomeos, K., Say, L., Kostanjsek, N., Jakob, R., Grove, J., Jr, B.M., Swaminathan, S., 2021. WHO SMART guidelines: optimising country-level use of guideline recommendations in the digital age. Lancet Digit. Heal. 3, e213–e216. https://doi.org/10.1016/S2589-7500(21)00038-8

Meier, J., Dietz, A., Boehm, A., Neumuth, T., 2015. Predicting treatment process steps from events. J. Biomed. Inform. 53, 308–319. https://doi.org/10.1016/j.jbi.2014.12.003

Mercadante, S., Adile, C., Caruselli, A., Ferrera, P., Costanzi, A., Marchetti, P., Casuccio, A., 2016. The Palliative-Supportive Care Unit in a Comprehensive Cancer Center as Crossroad for Patients' Oncological Pathway. PLoS One 11, e0157300. https://doi.org/10.1371/journal.pone.0157300

Messager, M., de Steur, W., Boelens, P.G., Jensen, L.S., Mariette, C., Reynolds, J. V., Osorio, J., Pera, M., Johansson, J., Ko??odziejczyk, P., Roviello, F., De Manzoni, G., M??nig, S.P., Allum, W.H., 2016. Description and analysis of clinical pathways for oesophago-gastric adenocarcinoma, in 10 European countries (the EURECCA upper gastro intestinal group ??? European Registration of Cancer Care). Eur. J. Surg. Oncol. 42, 1432–1447. https://doi.org/10.1016/j.ejso.2016.01.001

Miguel-Hurtado, O., Guest, R., Stevenage, S. V, Neil, G.J., Black, S., 2016. Comparing Machine Learning Classifiers and Linear/Logistic Regression to Explore the Relationship between Hand Dimensions and Demographic Characteristics. PLoS One 11, e0165521. https://doi.org/10.1371/journal.pone.0165521

Misawa, M., Kudo, S.E., Mori, Y., Cho, T., Kataoka, S., Yamauchi, A., Ogawa, Y., Maeda, Y., Takeda, K., Ichimasa, K., Nakamura, H., Yagawa, Y., Toyoshima, N., Ogata, N., Kudo, T., Hisayuki, T., Hayashi, T., Wakamura, K., Baba, T., Ishida, F., Itoh, H., Roth, H., Oda, M., Mori, K., 2018. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. Gastroenterology 154, 2027-2029.e3. https://doi.org/S0016-5085(18)30415-3

Mohammed, A.K., Yildirim, S., Farup, I., Pedersen, M., Hovde, O., 2018. Y-Net: A deep Convolutional Neural Network for Polyp Detection. CoRR abs/1806.0.

Møller, H., Roswall, N., Hemelrijck, M. Van, Larsen, S.B., Cuzick, J., Holmberg, L., Overvad, K., Tjønneland, A., 2015. Prostate cancer incidence, clinical stage and survival in relation to obesity: A prospective cohort study in Denmark. Int. J. Cancer 136, 1940–1947. https://doi.org/10.1002/ijc.29238

Mourouti, N., Kontogianni, M.D., Papavagelis, C., Panagiotakos, D.B., 2015. Diet and breast cancer: a systematic review. Int. J. Food Sci. Nutr. 66, 1–42. https://doi.org/10.3109/09637486.2014.950207

Mousa, S.M., Seifeldin, I.A., Hablas, A., Elbana, E.S., Soliman, A.S., 2011. Patterns of seeking medical care among Egyptian breast cancer patients: Relationship to late-stage presentation. Breast 20, 555–561. https://doi.org/10.1016/j.breast.2011.07.001

Muller, P., Woods, L., Walters, S., 2020. Temporal and geographic changes in stage at diagnosis in England during 2008-2013: A population-based study of colorectal, lung and ovarian cancers. Cancer Epidemiol. 67, 101743. https://doi.org/10.1016/j.canep.2020.101743

Munir, F., Kalawsky, K., Lawrence, C., Yarker, J., Haslam, C., Ahmed, S., 2011. Cognitive intervention for breast cancer patients undergoing adjuvant chemotherapy: a needs analysis. Cancer Nurs. 34, 385–392. https://doi.org/10.1097/NCC.0b013e31820254f3

Murchie, P., Adam, R., Khor, W.L., Smith, S., McNair, E., Swann, R., Witt, J., Weller, D., 2020. Impact of geography on Scottish cancer diagnoses in primary care: Results from a national cancer diagnosis audit. Cancer Epidemiol. 66, 101720. https://doi.org/10.1016/j.canep.2020.101720

Næser, E., Møller, H., Fredberg, U., Vedsted, P., 2018. Mortality of patients examined at a diagnostic centre: A matched cohort study. Cancer Epidemiol. 55, 130–135. https://doi.org/10.1016/j.canep.2018.06.008

National Cancer Registry Ireland. Cancer Factsheet. [Online]. Available on: https://www.ncri.ie/factsheets, Accessed: April 2021.

NCCN. About. [Online]. Available on: https://www.nccn.org/home/about. Accessed: June 2021.

NICE. About. [Online]. Available on: https://www.nice.org.uk/about. Accessed: June 2021.

Nguyen, C., Wang, Y., Nguyen, H.N., 2013. Random forest classifier combined with feature selection for breast cancer diagnosis and prognostic. J. Biomed. Sci. Eng. 6, 551–560. https://doi.org/10.4236/jbise.2013.65070

Nuemi, G., Afonso, F., Roussot, A., Billard, L., Cottenet, J., Combier, E., Diday, E., Quantin, C., 2013. Classification of hospital pathways in the management of cancer: Application to lung cancer in the region of burgundy. Cancer Epidemiol. 37, 688–696. https://doi.org/10.1016/j.canep.2013.06.007

Numan, R.C., Klomp, H.M., Li, W., Buitelaar, D.R., Burgers, J.A., Van Sandick, J.W., Wouters, M.W., 2012. A clinical audit in a multidisciplinary care path for thoracic surgery: An instrument for continuous quality improvement. Lung Cancer 78, 270–275. https://doi.org/10.1016/j.lungcan.2012.08.006

Nussbaum, D.P., Penne, K., Speicher, P.J., Stinnett, S.S., Perez, A., White, R.R., Clary, B.M., Tyler, D.S., Blazer, D.G., 2014. The role of clinical care pathways: An experience with distal pancreatectomy. J. Surg. Res. 190, 64–71. https://doi.org/10.1016/j.jss.2014.02.026

OECD/European Union, 2018.Adoption and use of Electronic Medical Records and ePrescribing. in Health at a Glance: Europe 2018: State of Health in the EU Cycle, OECD Publishing, Paris/European Union, Brussels. DOI: https://doi.org/10.1787/health_glance_eur-2018-56-en

Pakkanen, S., Baffoe-Bonnie, A.B., Matikainen, M.P., Koivisto, P.A., Tammela, T.L., Deshmukh, S., Ou, L., Bailey-Wilson, J.E., Schleutker, J., 2007. Segregation analysis of 1,546 prostate cancer families in Finland shows recessive inheritance. Hum. Genet. 121, 257–267. https://doi.org/10.1007/s00439-006-0310-2

Patel, A., Golan, S., Razmaria, A., Prasad, S., Eggener, S., Shalhav, A., 2014. Early discharge after laparoscopic or robotic partial nephrectomy: Care pathway evaluation. BJU Int. 113, 592–597. https://doi.org/10.1111/bju.12278

Pati, S., Hussain, M.A., Chauhan, A.S., Mallick, D., Nayak, S., 2013. Patient navigation pathway and barriers to treatment seeking in cancer in India: A qualitative inquiry. Cancer Epidemiol. 37, 973–978. https://doi.org/10.1016/j.canep.2013.09.018

Patten, C.L. Van, de Boer, J.G., Guns, E.S.T., 2008. Diet and dietary supplement intervention trials for the prevention of prostate cancer recurrence: a review of the randomized controlled trial evidence. J. Urol. 180, 2312–2314. https://doi.org/10.1016/j.juro.2008.08.078

Pease, N.J., Harris, R.J., Finlay, I.G., 2004. Development and audit of a care pathway for the management of patients with suspected malignant spinal cord compression. Physiotherapy 90, 27–34. https://doi.org/10.1016/S0031-9406(03)00006-3

Pedregosa, F., el Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, 'Edouard, 2011. Scikit-Learn: Machine Learning in Python.

J.Mach.Learn.Res. 12, 2825–2830.

Perez-Cornago, A., Appleby, P.N., Pischon, T., Tsilidis, K.K., Tjønneland, A., Olsen, A., Overvad, K., Kaaks, R., Kühn, T., Boeing, H., Steffen, A., Trichopoulou, A., Lagiou, P., Kritikou, M., Krogh, V., Palli, D., Sacerdote, C., Tumino, R., Bueno-de-Mesquita, H.B., Agudo, A., Larrañaga, N., Molina-Portillo, E., Barricarte, A., Chirlaque, M.-D., Quirós, J.R., är Stattin, P., Häggström, C., Wareham, N., Khaw, K.-T., Schmidt, J.A., Gunter, M., Freisling, H., Aune, D., Ward, H., Riboli, E., Key, T.J., Travis, R.C., 2017. Tall height and obesity are associated with an increased risk of aggressive prostate cancer: results from the EPIC cohort study. BMC Med. 15, 115. https://doi.org/10.1186/s12916-017-0876-7

Phillips, J.L., Lovell, M., Luckett, T., Agar, M., Green, A., Davidson, P., 2015. Australian survey of current practice and guideline use in adult cancer pain assessment and management: The community nurse perspective. Collegian 22, 33–41. https://doi.org/10.1016/j.colegn.2013.11.002

Pollard, T.J., Johnson, A.E.W., Raffa, J.D., Mark, R.G., 2018. tableone: An open source Python package for producing summary statistics for research papers. JAMIA Open 1, 26–31. https://doi.org/10.1093/jamiaopen/ooy012

Poon, C.C.Y., Jiang, Y., Zhang, R., Lo, W.W.Y., Cheung, M.S.H., Yu, R., Zheng, Y., Wong, J.C.T., Liu, Q., Wong, S.H., Mak, T.W.C., Lau, J.Y.W., 2020. AI-doscopist: a real-time deep-learning-based algorithm for localising polyps in colonoscopy videos with edge computing devices. NPJ Digit. Med. 3, 73-z. eCollection 2020. https://doi.org/10.1038/s41746-020-0281-z

Prostate Cancer Prevention Trial risk calculator, [Online]. Available on: http://riskcalc.org:3838/PCPTRC/. Accessed: 20 August 2020.

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106. https://doi.org/10.1007/BF00116251

Rai, N., Nevin, J., Downey, G., Abedin, P., Balogun, M., Kehoe, S., Sundar, S., 2015. Outcomes following implementation of symptom triggered diagnostic testing for ovarian cancer. Eur. J. Obstet. Gynecol. Reprod. Biol. 187, 64–69. https://doi.org/10.1016/j.ejogrb.2015.02.011

Rashaan, Z.M., Bastiaannet, E., Portielje, J.E.A., van de Water, W., van der Velde, S., Ernst, M.F., van de Velde, C.J.H., Liefers, G.J., 2011. Surgery in metastatic breast cancer: Patients with a favorable profile seem to have the most benefit from surgery. Eur. J. Surg. Oncol. 38, 52–56. https://doi.org/10.1016/j.ejso.2011.10.004

Redaniel, M.T., Ridd, M., Martin, R.M., Coxon, F., Jeffreys, M., Wade, J., 2015. Rapid diagnostic pathways for suspected colorectal cancer: views of primary and secondary care clinicians on challenges and their potential solutions. BMJ Open 5, e008577. https://doi.org/10.1136/bmjopen-2015-008577

Ribeiro, E., Häfner, M., Wimmer, G., Tamaki, T., Tischendorf, J.J.W., Yoshida, S., Tanaka, S., Uhl, A., 2017. Exploring texture Transfer Learning for Colonic Polyp Classification via Convolutional Neural Networks. - 2017 IEEE 14th Int. Symp. Biomed. Imaging (ISBI 2017) 1044–1048. https://doi.org/10.1109/ISBI.2017.7950695

Ribeiro, E., Uhl, A., Hafner, M., 2016a. Colonic polyp classification with convolutional neural networks, in: Proceedings - IEEE Symposium on Computer-Based Medical Systems. https://doi.org/10.1109/CBMS.2016.39

Ribeiro, E., Uhl, A., Wimmer, G., Häfner, M., 2016b. Exploring Deep Learning and Transfer Learning for Colonic Polyp Classification. Comput. Math. Methods Med. https://doi.org/10.1155/2016/6584725

Roennegaard, A.B., Rosenberg, T., Bjørndal, K., Sørensen, J.A., Johansen, J., Godballe, C., 2018. The Danish Head and Neck Cancer fast-track program: a tertiary cancer centre  experience. Eur. J. Cancer 90, 133–139. https://doi.org/10.1016/j.ejca.2017.10.038

Rossum, G. Van, Drake, F.L., 2009. Python 3 Reference Manual. CreateSpace, Scotts Valley, CA.

Rua, T., Watson, H., Malhotra, B., Turville, J., Razavi, R., Peacock, J.L., McCrone, P., Goh, V., Shearer, J., Griffin, N., 2020. An observational study to compare the utilisation of computed tomography  colonography with optical colonoscopy as the first diagnostic imaging tool in patients with suspected colorectal cancer. Clin. Radiol. 75, 712.e23-712.e31. https://doi.org/10.1016/j.crad.2020.04.014

Ryhänen, A.M., Rankinen, S., Siekkinen, M., Saarinen, M., Korvenranta, H., Leino-Kilpi, H., 2013. The impact of an empowering Internet-based Breast Cancer Patient Pathway program on breast cancer patients' clinical outcomes: A randomised controlled trial. J. Clin. Nurs. 22, 1016–1025. https://doi.org/10.1111/jocn.12007

Ryhänen, A.M., Rankinen, S., Siekkinen, M., Saarinen, M., Korvenranta, H., Leino-Kilpi, H., 2012a. The impact of an empowering Internet-based Breast Cancer Patient Pathway programme on breast cancer patients' knowledge: A randomised control trial. Patient Educ. Couns. 88, 224–231. https://doi.org/10.1016/j.pec.2012.02.013

Ryhänen, A.M., Rankinen, S., Tulus, K., Korvenranta, H., Leino-Kilpi, H., 2012b. Internet based patient pathway as an educational tool for breast cancer patients. Int. J. Med. Inform. 81, 270–278. https://doi.org/10.1016/j.ijmedinf.2012.01.010

Salamonsen, A., Kiil, M.A., Kristoffersen, A.E., Stub, T., Berntsen, G.R., 2016. My cancer is not my deepest concern": Life course disruption influencing patient pathways and health care needs among persons living with colorectal cancer. Patient Prefer. Adherence 10, 1591–1600. https://doi.org/10.2147/PPA.S108422

Salazar, A.S., Sekhon, S., Rohatgi, K.W., Nuako, A., Liu, J., Harriss, C., Brennan, E., LaBeau, D., Abdalla, I., Schulze, C., Muenks, J., Overlot, D., Higgins, J.A., Jones, L.S., Swick, C., Goings, S., Badiu, J., Walker, J., Colditz, G.A., James, A.S., 2020. A stepped-wedge randomized trial protocol of a community intervention for increasing lung screening through engaging primary care providers (I-STEP). Contemp. Clin. Trials 91, 105991. https://doi.org/https://doi.org/10.1016/j.cct.2020.105991

Saleh, H., 2018. Machine learning fundamentals: use Python and scikit-learn to get up and running with the hottest developments in machine learning," Birmingham, United Kingdom, Packt Publishing, Chapter 1: Introduction to scikit-learn; p. 1-37.

Sancho, C., Villalba, F.L., García-Coret, M.J., Vázquez, A., Safont, M.J., Hernández, A., Martínez, E., Martínez-Sanjuán, V., García-Armengol, J., Roig, J. V, 2010. Self-evaluation of a clinical pathway to improve the results of rectal cancer. Cirugía Española (English Ed. 87, 231–238. https://doi.org/https://doi.org/10.1016/S2173-5077(10)70053-3

Santillan, A., Govan, L., Zahurak, M.L., Diaz-Montes, T.P., Giuntoli, R.L., Bristow, R.E., 2008. Feasibility and economic impact of a clinical pathway for pap test utilization in Gynecologic Oncology practice. Gynecol. Oncol. 109, 388–393. https://doi.org/10.1016/j.ygyno.2008.01.006

Sapre, N., Hong, M.K.H., Macintyre, G., Lewis, H., Kowalczyk, A., Costello, A.J., Corcoran, N.M., Hovens, C.M., 2014. Curated MicroRNAs in Urine and Blood Fail to Validate as Predictive Biomarkers for High-Risk Prostate Cancer. PLoS One 9, e91729.

Scheuerlein, H., Rauchfuss, F., Dittmar, Y., Molle, R., Lehmann, T., Pienkos, N., Settmacher, U., 2012. New methods for clinical pathways - Business Process Modeling Notation (BPMN) and Tangible Business Process Modeling (t.BPM). Langenbeck's Arch. Surg. 397, 755–761. https://doi.org/10.1007/s00423-012-0914-z

Schmidt, H., Boese, S., Lampe, K., Jordan, K., Fiedler, E., Müller-Werdan, U., Wienke, A., Vordermark, D., 2017. Trans sectoral care of geriatric cancer patients based on comprehensive geriatric assessment and patient-reported quality of life - Results of a multicenter study to develop and pilot test a patient-centered interdisciplinary care concept for geriatric o. J. Geriatr. Oncol. 8, 262–270. https://doi.org/10.1016/j.jgo.2017.04.002

Sharma, S., Bekelman, J., Lin, A., Lukens, J.N., Roman, B.R., Mitra, N., Swisher-McClure, S., 2016. Clinical impact of prolonged diagnosis to treatment interval (DTI) among patients with oropharyngeal squamous cell carcinoma. Oral Oncol. 56, 17–24.

https://doi.org/10.1016/j.oraloncology.2016.02.010

Shin, Y., Balasingham, I., 2017. Comparison of hand-craft feature based SVM and CNN based deep learning framework for automatic polyp classification. - 2017 39th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. 3277–3280. https://doi.org/10.1109/EMBC.2017.8037556

Shin, Y., Qadir, H.A., Aabakken, L., Bergsland, J., Balasingham, I., 2018. Automatic Colon Polyp Detection Using Region Based Deep CNN and Post Learning Approaches. - IEEE Access 6, 40950–40962. https://doi.org/10.1109/ACCESS.2018.2856402

Siegel, R.L., Miller, K.D., Jemal, A., 2020. Cancer statistics, 2020. CA. Cancer J. Clin. 70, 7–30. https://doi.org/https://doi.org/10.3322/caac.21590

Silva, J., Histace, A., Romain, O., Dray, X., Granado, B., 2014. Toward embedded detection of polyps in WCE images for early diagnosis of colorectal cancer. Int. J. Comput. Assist. Radiol. Surg. 9, 283–293. https://doi.org/10.1007/s11548-013-0926-3

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv Prepr. 1–10. https://doi.org/10.1016/j.infsof.2008.09.005

So, J.B.Y., Lim, Z.L., Lin, H.A., Ti, T.K., 2008. Reduction of hospital stay and cost after the implementation of a clinical pathway for radical gastrectomy for gastric cancer. Gastric Cancer 11, 81–85. https://doi.org/10.1007/s10120-008-0458-7

Song, E.M., Park, B., Ha, C.-A., Hwang, S.W., Park, S.H., Yang, D.-H., Ye, B.D., Myung, S.-J., Yang, S.-K., Kim, N., Byeon, J.-S., 2020. Endoscopic diagnosis and treatment planning for colorectal polyps using a deep-learning model. Sci. Rep. 10, 30. https://doi.org/10.1038/s41598-019-56697-0

Soria-Aledo, V., Mengual-Ballester, M., Pellicer-Franco, E., Carrillo-Alcaraz, A., Cases-Baldó, M.J., Carrasco-Prats, M., Campillo-Soto, A., Flores-Pastor, B., Aguayo-Albasini, J.L., 2011. Evaluation of a Clinical Pathway to Improve Colorectal Cancer Outcomes. Am. J. Med. Qual. 26, 396–404. https://doi.org/10.1177/1062860611404049

Stark, G.F., Hart, G.R., Nartowt, B.J., Deng, J., 2019. Predicting breast cancer risk using personal health data and machine learning models. PLoS One 14, e0226765. https://doi.org/10.1371/journal.pone.0226765

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions, in: - 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional Neural Networks for Medical Image

Analysis: Full Training or Fine Tuning? - IEEE Trans. Med. Imaging 35, 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Tamburini, M., Gangeri, L., Brunelli, C., Boeri, P., Borreani, C., Bosisio, M., Karmann, C.F., Greco, M., Miccinesi, G., Murru, L., Trimigno, P., 2003. Cancer patients' needs during hospitalisation: a quantitative and qualitative study. BMC Cancer 3, 12. https://doi.org/10.1186/1471-2407-3-12

Tastan, S., Hatipoglu, S., Iyigun, E., Kilic, S., 2012. Implementation of a clinical pathway in breast cancer patients undergoing breast surgery. Eur. J. Oncol. Nurs. 16, 368–374. https://doi.org/10.1016/j.ejon.2011.07.003

Temel, J.S., Greer, J.A., Muzikansky, A., Gallagher, E.R., Admane, S., Jackson, V.A., Dahlin, C.M., Blinderman, C.D., Jacobsen, J., Pirl, W.F., Billings, J.A., Lynch, T.J., 2010. Early palliative care for patients with metastatic non-small-cell lung cancer. N. Engl. J. Med. 363, 733–742. https://doi.org/10.1056/NEJMoa1000678

Thakur, S.S., Li, H., Chan, A.M.Y., Tudor, R., Bigras, G., Morris, D., Enwere, E.K., Yang, H., 2018. The use of automated Ki67 analysis to predict Oncotype DX risk-of-recurrence categories in early-stage breast cancer. PLoS One 13, e0188983.

Thorsen, L., Gjerset, G.M., Loge, J.H., Kiserud, C.E., Skovlund, E., Fløtten, T., Fosså, S.D., 2011. Cancer patients' needs for rehabilitation services. Acta Oncol. 50, 212–222. https://doi.org/10.3109/0284186X.2010.531050

Tu, J. V, 1996. Advantages and disadvantages of using artificial neural networks versus logistic regression for predicting medical outcomes. J. Clin. Epidemiol. 49, 1225–1231. https://doi.org/10.1016/s0895-4356(96)00002-9

Urban, G., Tripathi, P., Alkayali, T., Mittal, M., Jalali, F., Karnes, W., Baldi, P., 2018. Deep Learning Localizes and Identifies Polyps in Real Time With 96% Accuracy in Screening Colonoscopy. Gastroenterology 155, 1069-1078.e8. https://doi.org/S0016-5085(18)34659-6

Väisänen, J.A., Syrjälä, A.-M.H., Pesonen, P.R.O., Pukkila, M.J., Koivunen, P.T., Alho, O.-P., 2014. Characteristics and medical-care-seeking of head and neck cancer patients: A population-based cross-sectional survey. Oral Oncol. 50, 740–745. https://doi.org/10.1016/j.oraloncology.2014.04.009

Vajdic, C.M., Schaffer, A.L., Dobbins, T.A., Ward, R.L., Er, C.C., Pearson, S.-A., 2015. Health service utilisation and investigations before diagnosis of cancer of unknown primary (CUP): A population-based nested case–control study in Australian Government Department of Veterans' Affairs clients. Cancer Epidemiol. 39, 585–592. https://doi.org/https://doi.org/10.1016/j.canep.2015.02.006

Van Beek, K., Siouta, N., Preston, N., Hasselaar, J., Hughes, S., Payne, S., Radbruch, L., Centeno, C., Csikos, A., Garralda, E., van der Eerden, M., Hodiamont, F., Radvanyi, I., Menten, J., 2016. To what degree is palliative care integrated in guidelines and pathways for adult cancer patients in Europe: a systematic literature review. BMC Palliat. Care 15, 26. https://doi.org/10.1186/s12904-016-0100-0

van Dam, P.A., Verheyden, G., Sugihara, A., Trinh, X.B., Van Der Mussele, H., Wuyts, H., Verkinderen, L., Hauspy, J., Vermeulen, P., Dirix, L., 2013. A dynamic clinical pathway for the treatment of patients with early breast cancer is a tool for better cancer care: implementation and prospective analysis between 2002-2010. World J. Surg. Oncol. 11, 70. https://doi.org/10.1186/1477-7819-11-70

van de Ven, M., Retèl, V.P., Koffijberg, H., van Harten, W.H., IJzerman, M.J., 2019. Variation in the time to treatment for stage III and IV non-small cell lung cancer patients for hospitals in the Netherlands. Lung Cancer 134, 34–41. https://doi.org/10.1016/j.lungcan.2019.05.023

van Hoeve, J C, Elferink, M.A., Klaase, J.M., Kouwenhoven, E.A., Schiphorst, P.P., Siesling, S., 2015. Long-term effects of a regional care pathway for patients with rectal cancer. Int J Color. Dis 30, 787–795. https://doi.org/10.1007/s00384-015-2209-7

van Hoeve, Jolanda C, Elferink, M.A.G., Klaase, J.M., Kouwenhoven, E.A., Schiphorst, P.P.J.B.M., Siesling, S., 2015. Long-term effects of a regional care pathway for patients with rectal cancer. Int. J. Colorectal Dis. 30, 787–795. https://doi.org/10.1007/s00384-015-2209-7

Vijayakumar, C., Maroju, N.K., Srinivasan, K., Reddy, K.S., 2016. Clinical audit system as a quality improvement tool in the management of breast cancer. Int. J. Surg. 35, 44–50. https://doi.org/10.1016/j.ijsu.2016.09.011

Viklund, P., Lagergren, J., 2007. A care pathway for patients with oesophageal cancer. Eur. J. Cancer Care (Engl). 16, 533–538. https://doi.org/ECC790

Viklund, Pernilla, Wengström, Y., Lagergren, J., 2006. Supportive care for patients with oesophageal and other upper gastrointestinal cancers: The role of a specialist nurse in the team. Eur. J. Oncol. Nurs. 10, 353–363. https://doi.org/10.1016/j.ejon.2006.01.009

Viklund, P, Wengström, Y., Lagergren, J., 2006. Supportive care for patients with oesophageal and other upper gastrointestinal cancers: The role of a specialist nurse in the team. Eur. J. Oncol. Nurs. 10, 353–363. https://doi.org/S1462-3889(06)00036-6

Virtual Mentor, 2011. Development of the Electronic Health Record. American Medical Association Journal of Ethics;13(3):186-189. DOI: 10.1001/virtualmentor.2011.13.3.mhst1-1103.

Vyas, O., Kaklamani, V., 2017. Evaluating the Role of Extended Aromatase Inhibitor Therapy in Early Hormone-Positive Breast Cancer. Curr. Breast Cancer Rep. 9, 183–187. https://doi.org/10.1007/s12609-017-0250-y

Wang, P., Liu, X., Berzin, T.M., R., J.G.B., Liu, P., Zhou, C., Lei, L., Li, L., Guo, Z., Lei, S., Xiong, F., Wang, H., Song, Y., Pan, Y., Zhou, G., 2020. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study. Lancet Gastroenterol. Hepatol. 5, 343–351. https://doi.org/10.1016/S2468-1253(19)30411-X

Wolf, H.J., Dwyer, A., Ahnen, D.J., Pray, S.L., Rein, S.M., Morwood, K.D., Lowery, J.T., Masias, A., Collins, N.J., Brown, C.E., Demaio Goheen, C.A., McAbee, K.E., Sauaia, A., Byers, T.E., 2015. Colon cancer screening for colorado’s underserved: A community clinic/academic partnership. Am. J. Prev. Med. 48, 264–270. https://doi.org/10.1016/j.amepre.2014.09.016

World Health Organization. WHO guidelines. [Online]. Available on: https://www.who.int/publications/who-guidelines, Accessed: June 2021.

World Health Organization, 1992. International Statistical Classification of Diseases and Related Health Problems. 10th Revision (ICD-10). Geneva: WHO.

XGBoost for Python. [Online]. Available on: https://xgboost.readthedocs.io/en/latest/python/index.html. Accessed: 02 July 2020.

Xu, Y., Liu, Z., Li, Y., Hou, H., Cao, Y., Zhao, Y., Guo, W., Cui, L., 2020. Feature data processing: Making medical data fit deep neural networks. Futur. Gener. Comput. Syst. 109, 149–157. https://doi.org/https://doi.org/10.1016/j.future.2020.02.034

Yamada, M., Saito, Y., Imaoka, H., Saiko, M., Yamada, S., Kondo, H., Takamaru, H., Sakamoto, T., Sese, J., Kuchiba, A., Shibata, T., Hamamoto, R., 2019. Development of a real-time endoscopic image diagnosis support system using deep learning technology in colonoscopy. Sci. Rep. 9, 14465. https://doi.org/10.1038/s41598-019-50567-5

Yamamoto, K., Yamanaka, K., Hatano, E., Sumi, E., Ishii, T., Taura, K., Iguchi, K., Teramukai, S., Yokode, M., Uemoto, S., Fukushima, M., 2012. An eClinical trial system for cancer that integrates with clinical pathways and electronic medical records. Clin. Trials 9, 408–417. https://doi.org/10.1177/1740774512445912

Yap, S., Goldsbury, D., Yap, M.L., Yuill, S., Rankin, N., Weber, M., Canfell, K., O’Connell, D.L., 2018. Patterns of care and emergency presentations for people with non-small cell lung cancer in New South Wales, Australia: A population-based study. Lung Cancer 122, 171–179. https://doi.org/10.1016/j.lungcan.2018.06.006

Yip, K., Mcconnell, H., Alonzi, R., Maher, J., 2015. Using routinely collected data to stratify prostate cancer patients into phases of care in the United Kingdom: Implications for resource allocation and the cancer survivorship programme. Br. J. Cancer 112, 1594–1602. https://doi.org/10.1038/bjc.2014.650

Yu, L., Chen, H., Dou, Q., Qin, J., Heng, P.A., 2017. Integrating Online and Offline Three-Dimensional Deep Learning for Automated Polyp Detection in Colonoscopy Videos. IEEE J. Biomed. Heal. Informatics. https://doi.org/10.1109/JBHI.2016.2637004

Zauber, A.G., 2015. The Impact of Screening on Colorectal Cancer Mortality and Incidence: Has It Really Made a Difference? Dig. Dis. Sci. 60, 681–691. https://doi.org/10.1007/s10620-015-3600-5

Zeinomar, N., Knight, J.A., Genkinger, J.M., Phillips, K.-A., Daly, M.B., Milne, R.L., Dite, G.S., Kehm, R.D., Liao, Y., Southey, M.C., Chung, W.K., Giles, G.G., McLachlan, S.-A., Friedlander, M.L., Weideman, P.C., Glendon, G., Nesci, S., Andrulis, I.L., Buys, S.S., John, E.M., MacInnis, R.J., Hopper, J.L., Terry, M.B., 2019. Alcohol consumption, cigarette smoking, and familial breast cancer risk: findings from the Prospective Family Study Cohort (ProF-SC). Breast Cancer Res. 21, 128. https://doi.org/10.1186/s13058-019-1213-1

Zhang, R., Zheng, Y., Mak, T.W.C., Yu, R., Wong, S.H., Lau, J.Y.W., Poon, C.C.Y., 2017. Automatic Detection and Classification of Colorectal Polyps by Transferring Low-Level CNN Features from Nonmedical Domain. IEEE J. Biomed. Heal. Informatics 21, 41–47. https://doi.org/10.1109/JBHI.2016.2635662

Zhu, J., Li, X., Li, H., Liu, Z., Ma, J., Kou, J., He, Q., 2020. Enhanced recovery after surgery pathways benefit patients with soft pancreatic texture following pancreaticoduodenectomy. Am. J. Surg. 219, 1019–1023. https://doi.org/10.1016/j.amjsurg.2019.08.002

Zimmermann, C., Swami, N., Krzyzanowska, M., Hannon, B., Leighl, N., Oza, A., Moore, M., Rydall, A., Rodin, G., Tannock, I., Donner, A., Lo, C., 2014. Early palliative care for patients with advanced cancer: a cluster-randomised controlled trial. Lancet (London, England) 383, 1721–1730. https://doi.org/10.1016/S0140-6736(13)62416-2

Zu, K., Giovannucci, E., 2009. Smoking and aggressive prostate cancer: a review of the epidemiologic evidence. Cancer Causes Control 20, 1799–1810. https://doi.org/10.1007/s10552-009-9387-y

Zuccolo, L., Harris, R., Gunnell, D., Oliver, S., Lane, J.A., Davis, M., Donovan, J., Neal, D., Hamdy, F., Beynon, R., Savovic, J., Martin, R.M., 2008. Height and Prostate Cancer Risk: A Large Nested Case-Control Study (ProtecT) and Meta-analysis. Cancer Epidemiol. Biomarkers Prev. 17, 2325–2336. https://doi.org/10.1158/1055-9965.EPI-08-0342

# Appendix A

The interviews were conducted in the form of conversations which were based on the following questions. However, the order, the depth, and the speed depended on the participant.

1. What is your age?
2. What is the highest degree you have obtained?
3. Do you work? (What’s your profession?)

**Diagnosis phase**

4. When were you diagnosed?
5. Did you suffer from other diseases before cancer diagnosis?
6. How was your cancer discovered?
7. Talk to me about the journey you came to the hospital until you got your diagnosis.
8. Are you satisfied with the care you received during the first stages of your diagnosis?
9. What steps/protocol did you need to follow?
10. How did you find those steps/ the protocol you needed to follow?
11. What do you think about those steps?
12. Did you talk to anyone about what you were going through? (internet – which websites, doctor, nurse, friends, support groups, family) Why did you go to that person?
13. Where did you find information about the cancer diagnosis? (internet – which websites, doctor, nurse, friends, support groups, family)

**Treatment phase**

14. What were the steps you followed the moment you were diagnosed until you started your treatment?
15. How many days did it take to start with the treatment?
16. How does it feel? Was it enough time?
17. Did you have any other illness before the treatment? What about any complications during the treatment?
18. Were you given information about treatment?

19. Who was your support during this period?
20. Did you change your diet during treatment? What was your diet before, during, and after treatment?
21. Do you smoke? If the person is a smoker: How many cigarettes/cigars do you smoke in a day/ week?
22. What is your alcohol intake?
23. What treatment did you need? What treatment are you following/ followed? (chemotherapy/ hormone therapy/ radiotherapy/ targeted therapy/ immunotherapy etc)
24. How long did the treatment last? (in days, or weeks, or months)
25. What do you think about clinical studies? Did you participate in any of them?
26. Besides the medical aspect, what other discussions did you have with your GP/ nurses/ consultants at the hospital?

**Follow up period**

27. When did you finish your treatment?
28. How was the preparation for the follow-up phase?
29. Did you get any information about it (the follow-up phase)?
30. How do you consider the follow-up meetings/ appointments so far?
31. What lifestyle changes did you make during this period?
32. Passing through this journey:
    a. How did the diagnosis affect your mental health? (Did you go to a support group?; Did you see a psychologist?; Did you take any medication for this?; Did you join a meditation class, online or in-person?)
    b. How did the diagnosis affect your sexuality?
    c. How did the diagnosis affect your fertility?
    d. How did the diagnosis affect your body image?
    e. How did the diagnosis affect your relationships? (family, work, friends, partner, etc.)
    f. How did the diagnosis affect your fear of recurrence?
    g. For young breast cancer patients: motherhood, breastfeeding.

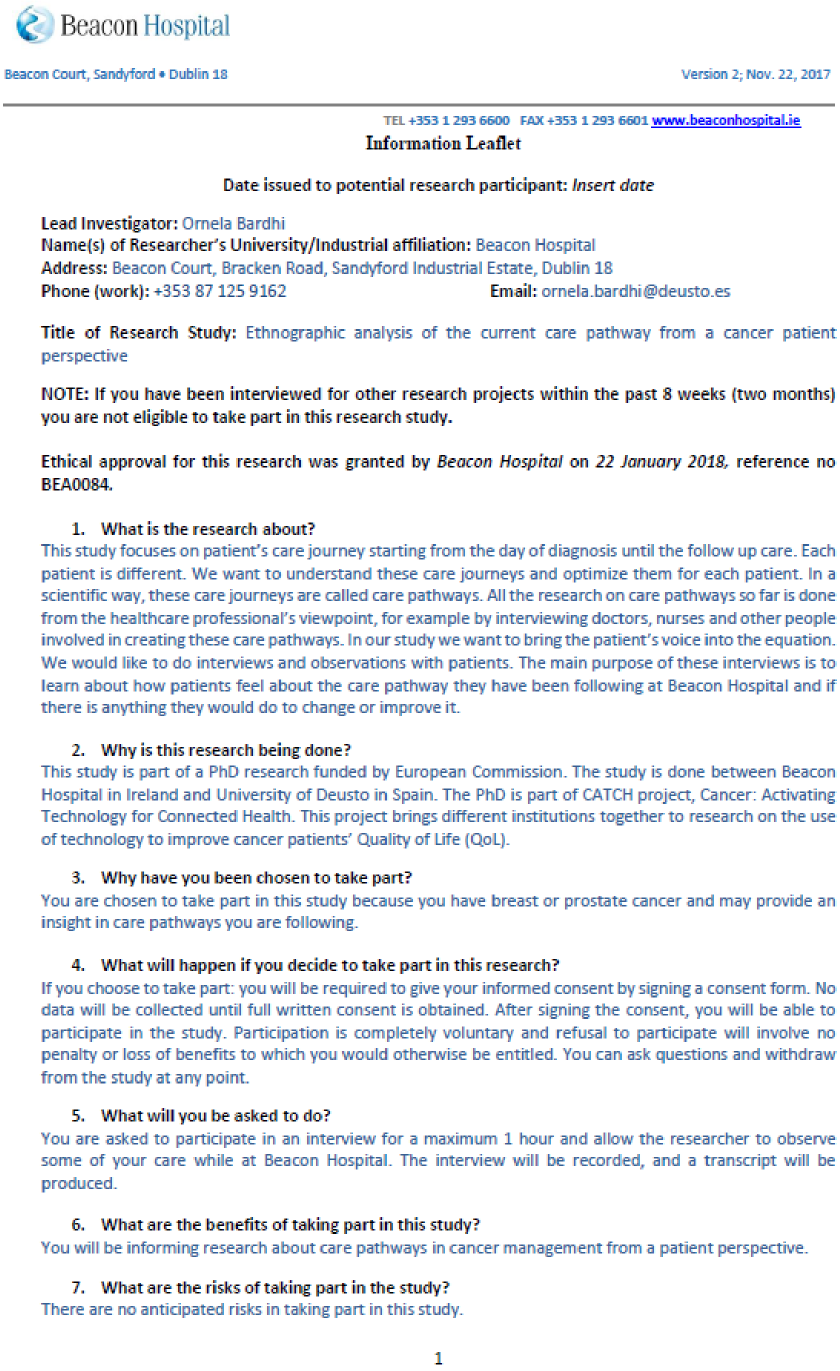
33. Any further comments about your care journey here at Beacon Hospital?

34. If you think of any improvements at any phase of the care journey you believe the hospital should consider, what would those be?

35. Any feedback about the interview? Any suggestions for improvements for the interviewer?

# Appendix B

Information leaflet and informed consent for the Beacon Hospital study.



Beacon Hospital

Beacon Court, Sandyford • Dublin 18

Version 2; Nov. 22, 2017

TEL +353 1 293 6600 FAX +353 1 293 6601 www.beaconhospital.ie

**Information Leaflet**

**Date issued to potential research participant:** ***Insert date***

**Lead Investigator:** Ornela Bardhi
**Name(s) of Researcher's University/Industrial affiliation:** Beacon Hospital
**Address:** Beacon Court, Bracken Road, Sandyford Industrial Estate, Dublin 18
**Phone (work):** +353 87 125 9162 **Email:** ornela.bardhi@deusto.es

**Title of Research Study:** Ethnographic analysis of the current care pathway from a cancer patient perspective

**NOTE: If you have been interviewed for other research projects within the past 8 weeks (two months) you are not eligible to take part in this research study.**

**Ethical approval for this research was granted by *Beacon Hospital* on *22 January 2018*, reference no BEA0084.**

1. **What is the research about?**

This study focuses on patient's care journey starting from the day of diagnosis until the follow up care. Each patient is different. We want to understand these care journeys and optimize them for each patient. In a scientific way, these care journeys are called care pathways. All the research on care pathways so far is done from the healthcare professional's viewpoint, for example by interviewing doctors, nurses and other people involved in creating these care pathways. In our study we want to bring the patient's voice into the equation. We would like to do interviews and observations with patients. The main purpose of these interviews is to learn about how patients feel about the care pathway they have been following at Beacon Hospital and if there is anything they would do to change or improve it.

2. **Why is this research being done?**

This study is part of a PhD research funded by European Commission. The study is done between Beacon Hospital in Ireland and University of Deusto in Spain. The PhD is part of CATCH project, Cancer: Activating Technology for Connected Health. This project brings different institutions together to research on the use of technology to improve cancer patients' Quality of Life (QoL).

3. **Why have you been chosen to take part?**

You are chosen to take part in this study because you have breast or prostate cancer and may provide an insight in care pathways you are following.

4. **What will happen if you decide to take part in this research?**

If you choose to take part: you will be required to give your informed consent by signing a consent form. No data will be collected until full written consent is obtained. After signing the consent, you will be able to participate in the study. Participation is completely voluntary and refusal to participate will involve no penalty or loss of benefits to which you would otherwise be entitled. You can ask questions and withdraw from the study at any point.

5. **What will you be asked to do?**

You are asked to participate in an interview for a maximum 1 hour and allow the researcher to observe some of your care while at Beacon Hospital. The interview will be recorded, and a transcript will be produced.

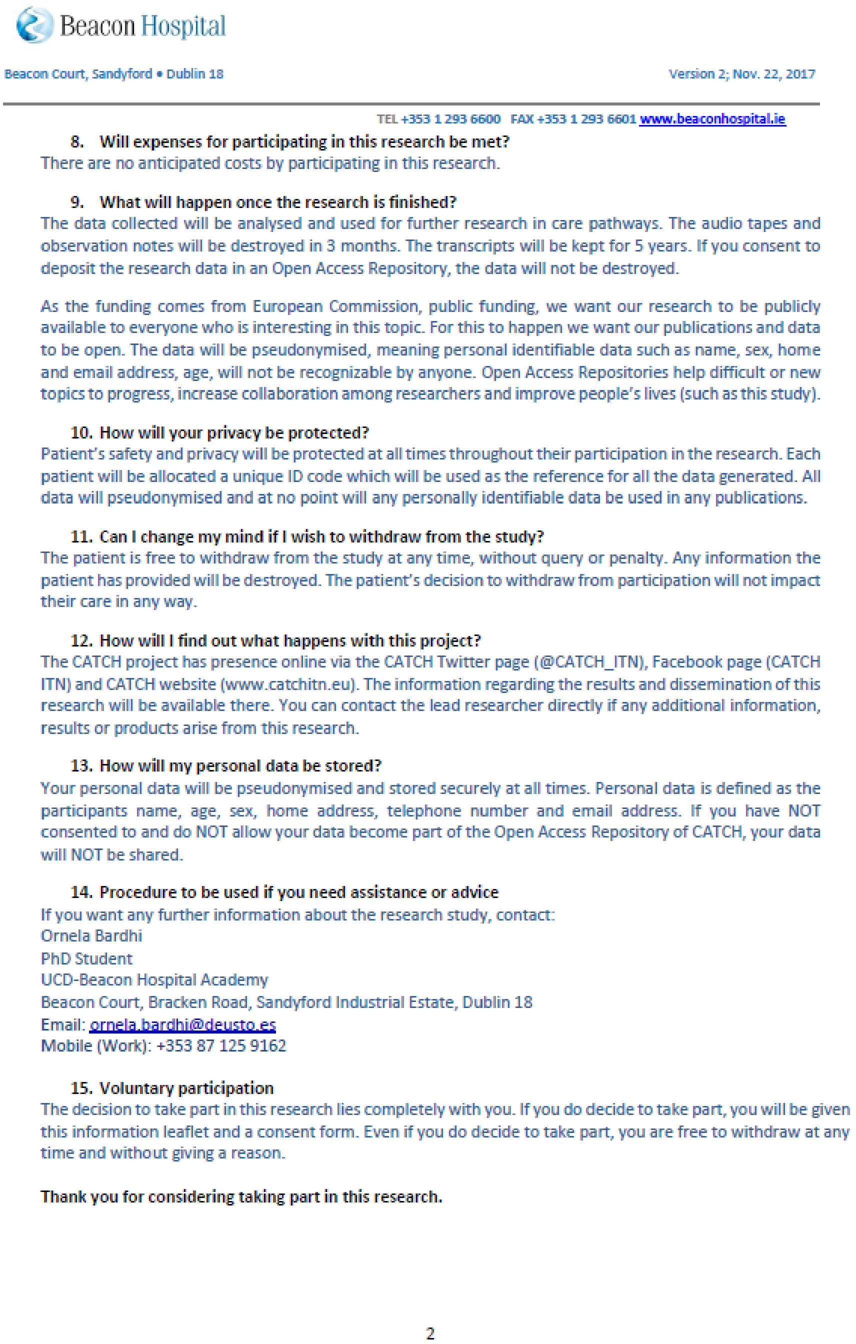
6. **What are the benefits of taking part in this study?**

You will be informing research about care pathways in cancer management from a patient perspective.

7. **What are the risks of taking part in the study?**

There are no anticipated risks in taking part in this study.

***Figure B.1.*** The first page of the information leaflet for the Beacon Hospital study.



Beacon Hospital

Beacon Court, Sandyford • Dublin 18

Version 2; Nov. 22, 2017

TEL +353 1 293 6600 FAX +353 1 293 6601 www.beaconhospital.ie

**8. Will expenses for participating in this research be met?**

There are no anticipated costs by participating in this research.

**9. What will happen once the research is finished?**

The data collected will be analysed and used for further research in care pathways. The audio tapes and observation notes will be destroyed in 3 months. The transcripts will be kept for 5 years. If you consent to deposit the research data in an Open Access Repository, the data will not be destroyed.

As the funding comes from European Commission, public funding, we want our research to be publicly available to everyone who is interesting in this topic. For this to happen we want our publications and data to be open. The data will be pseudonymised, meaning personal identifiable data such as name, sex, home and email address, age, will not be recognizable by anyone. Open Access Repositories help difficult or new topics to progress, increase collaboration among researchers and improve people's lives (such as this study).

**10. How will your privacy be protected?**

Patient's safety and privacy will be protected at all times throughout their participation in the research. Each patient will be allocated a unique ID code which will be used as the reference for all the data generated. All data will pseudonymised and at no point will any personally identifiable data be used in any publications.

**11. Can I change my mind if I wish to withdraw from the study?**

The patient is free to withdraw from the study at any time, without query or penalty. Any information the patient has provided will be destroyed. The patient's decision to withdraw from participation will not impact their care in any way.

**12. How will I find out what happens with this project?**

The CATCH project has presence online via the CATCH Twitter page (@CATCH_ITN), Facebook page (CATCH ITN) and CATCH website (www.catchitn.eu). The information regarding the results and dissemination of this research will be available there. You can contact the lead researcher directly if any additional information, results or products arise from this research.

**13. How will my personal data be stored?**

Your personal data will be pseudonymised and stored securely at all times. Personal data is defined as the participants name, age, sex, home address, telephone number and email address. If you have NOT consented to and do NOT allow your data become part of the Open Access Repository of CATCH, your data will NOT be shared.

**14. Procedure to be used if you need assistance or advice**

If you want any further information about the research study, contact:
Ornela Bardhi
PhD Student
UCD-Beacon Hospital Academy
Beacon Court, Bracken Road, Sandyford Industrial Estate, Dublin 18
Email: ornela.bardhi@deusto.es
Mobile (Work): +353 87 125 9162

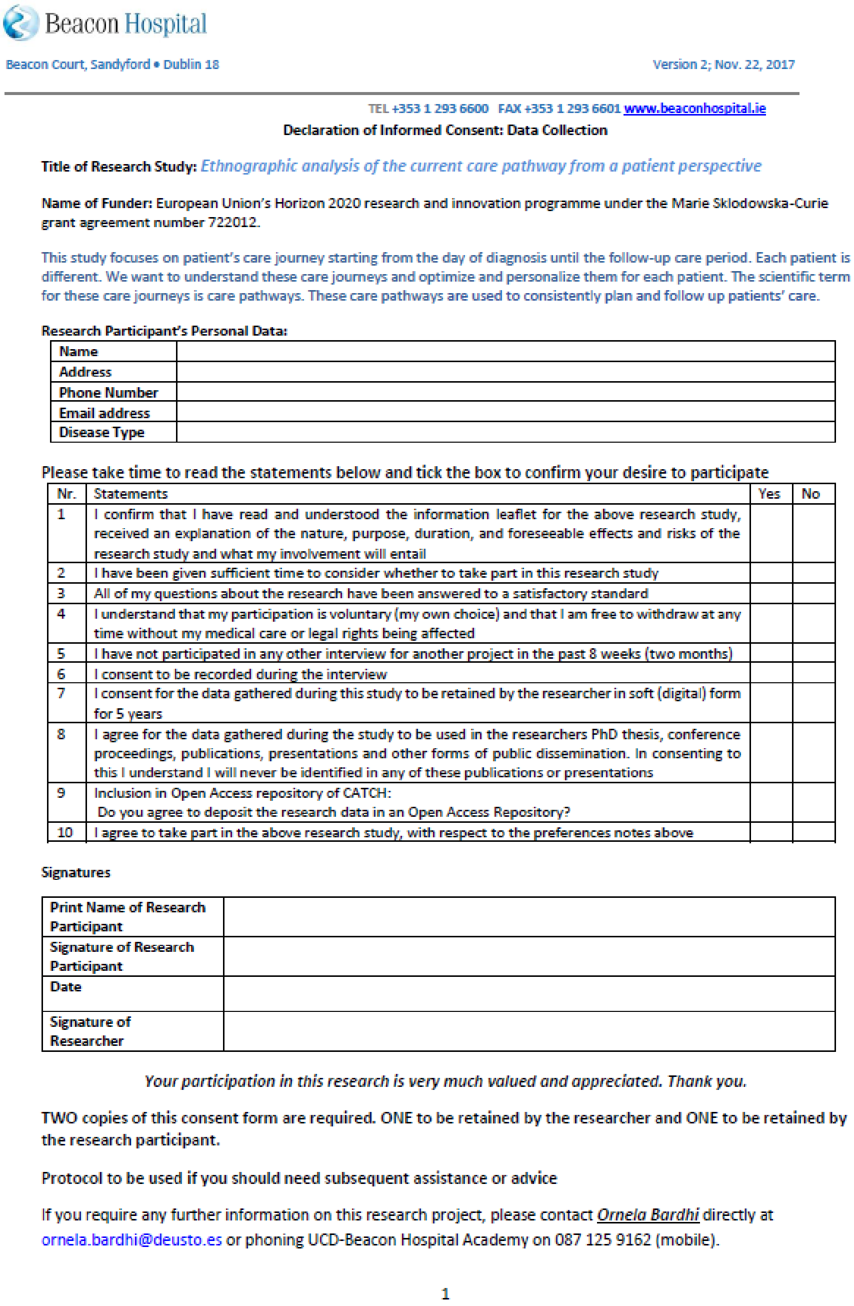
**15. Voluntary participation**

The decision to take part in this research lies completely with you. If you do decide to take part, you will be given this information leaflet and a consent form. Even if you do decide to take part, you are free to withdraw at any time and without giving a reason.

**Thank you for considering taking part in this research.**

2

***Figure B.2.*** The second page of the information leaflet for the Beacon Hospital study.



Beacon Hospital

Beacon Court, Sandyford • Dublin 18 Version 2; Nov. 22, 2017

TEL +353 1 293 6600 FAX +353 1 293 6601 www.beaconhospital.ie

**Declaration of Informed Consent: Data Collection**

**Title of Research Study:** *Ethnographic analysis of the current care pathway from a patient perspective*

**Name of Funder:** European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement number 722012.

This study focuses on patient's care journey starting from the day of diagnosis until the follow-up care period. Each patient is different. We want to understand these care journeys and optimize and personalize them for each patient. The scientific term for these care journeys is care pathways. These care pathways are used to consistently plan and follow up patients' care.

**Research Participant's Personal Data:**

| | |
|---|---|
| **Name** | |
| **Address** | |
| **Phone Number** | |
| **Email address** | |
| **Disease Type** | |

Please take time to read the statements below and tick the box to confirm your desire to participate

| Nr. | Statements | Yes | No |
|---|---|---|---|
| 1 | I confirm that I have read and understood the information leaflet for the above research study, received an explanation of the nature, purpose, duration, and foreseeable effects and risks of the research study and what my involvement will entail | | |
| 2 | I have been given sufficient time to consider whether to take part in this research study | | |
| 3 | All of my questions about the research have been answered to a satisfactory standard | | |
| 4 | I understand that my participation is voluntary (my own choice) and that I am free to withdraw at any time without my medical care or legal rights being affected | | |
| 5 | I have not participated in any other interview for another project in the past 8 weeks (two months) | | |
| 6 | I consent to be recorded during the interview | | |
| 7 | I consent for the data gathered during this study to be retained by the researcher in soft (digital) form for 5 years | | |
| 8 | I agree for the data gathered during the study to be used in the researchers PhD thesis, conference proceedings, publications, presentations and other forms of public dissemination. In consenting to this I understand I will never be identified in any of these publications or presentations | | |
| 9 | Inclusion in Open Access repository of CATCH:<br>Do you agree to deposit the research data in an Open Access Repository? | | |
| 10 | I agree to take part in the above research study, with respect to the preferences notes above | | |

**Signatures**

| | |
|---|---|
| **Print Name of Research Participant** | |
| **Signature of Research Participant** | |
| **Date** | |
| **Signature of Researcher** | |

*Your participation in this research is very much valued and appreciated. Thank you.*

TWO copies of this consent form are required. ONE to be retained by the researcher and ONE to be retained by the research participant.

Protocol to be used if you should need subsequent assistance or advice

If you require any further information on this research project, please contact *Ornela Bardhi* directly at ornela.bardhi@deusto.es or phoning UCD-Beacon Hospital Academy on 087 125 9162 (mobile).

1

***Figure B.3.*** Informed consent for the Beacon Hospital study.

# Appendix C

***Table C.1.*** Variables collected for the breast cancer study and their explanation.

| ABBREVIATION | EXPLANATION |
|---|---|
| **DEMOGRAPHIC DATA** | |
| P_id | The ID of the participant (not related to their medical ID used at the hospital) |
| Age | age of the participant the time of the interview |
| Age_diagnosis | age of the participant when diagnosed |
| Dod | date of death of the participant |
| Province | one of the 4 provinces of the Republic of Ireland: Connacht, Leinster, Munster, Ulster |
| M_status | marital status: single, married, partnership, widowed, divorced, unmarried |
| Edu_irish | Irish education system |
| Employment | employment status of the participant during the interview and the care phase: not working, no info, part-time, full time, retired |
| Religion | the religion of the participant: religious, not religious |
| Insurance | all participants had private insurance (private hospital) |
| **MEDICAL DATA** | |
| Hearing | hearing status of the participant: normal, impaired |
| Vision | vision status of the participant: normal, impaired |
| Allergies | allergy status if a participant had either drug or food allergies: no, yes |
| Height_cm | the last measured height of the participant in centimeters |
| Weight_kg | the last measured weight of the participant in kilograms |
| Bmi | body mass index |
| Bmi_group | body mass index groups: underweight, normal, overweight, obese |
| Plan_staging1 | the first examination in the diagnosis of cancer |
| Plan_staging2 | the second examination in the diagnosis of cancer |
| Plan_staging3 | the third examination in the diagnosis of cancer |
| Plan_staging4 | the fourth examination in the diagnosis of cancer |
| Date_biopsy | the date when the breast biopsy was done |
| Cancer_type | the type of breast cancer: invasive ductal carcinoma, invasive lobular carcinoma, metastatic breast carcinoma, Cancer of unknown primary |
| Metastasis | has cancer metastasized: no, yes |
| Mets_org1 | the first organ cancer metastasized |
| Mets_org2 | the second organ cancer metastasized |
| Mets_org3 | the third organ cancer metastasized |
| Tumor_size_ mm | the size of the tumor in millimeters |
| Grade | grade of tumor: 1, 2, 3 |
| Stage | state of cancer: 1, 2, 3, 4 |

| | |
|---|---|
| L_nodes | lymph node involvement: no, yes |
| Pr | progesterone receptor: negative, positive |
| Pr_score | the progesterone receptor score: 0 to 8 |
| Er | estrogen receptor: negative, positive |
| Er_score | the estrogen receptor score: 0 to 8 |
| HER2 | human epidermal growth factor receptor 2 (a gene that can play a role in the development of breast cancer): negative, positive, equivocal |
| HER2_score | the her2 score: 0 to 3 |
| Pretreat_investigations1 | first examination before treatment |
| Pretreat_investigations2 | second examination before treatment |
| Pretreat_investigations3 | third examination before treatment |
| Pretreat_investigations4 | fourth examination before treatment |
| Diag_treat_days | days between the diagnosis day and the start of the treatment |
| Treatment_started | the day treatment started |
| Treatment_line1 | the first line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, immunotherapy |
| Treatment_line2 | the second line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, immunotherapy |
| Treatment_line3 | the third line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, immunotherapy |
| Treatment_line4 | the fourth line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, immunotherapy |
| Treatment_line5 | the fifth line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, immunotherapy |
| Surgery | if the participant had surgery as part of the treatment: yes/ no |
| Surgery_date | the date when the breast surgery was done |
| Surgery_side | the side of the breast the surgery was done: left, right, both |
| Type_of_surgery | the type of surgery that was undergone: wire-guided wide local excision (WLE) = lumpectomy = breast-conserving surgery or mastectomy, including sentinel lymph node biopsy and axillary lymph node dissection. |
| Care_phase | the care phase the participant was in at the time of the interview: diagnosis, treatment, follow-up, palliative (participants were selected to have at least started their treatment, so there were no participants on their diagnosis phase. These patients were asked to participate in the study at a later stage when they had gone through some part of their treatment phase.) |
| Tumor_size_after_chemo | the size of the tumor after the chemotherapy, when the first line of treatment was chemotherapy |
| Post_neochemo_examination | examinations were done after the chemotherapy and before surgery |
| Comorbidities | other illnesses the participant had besides the cancer |
| Oncotype_score | Oncotype DX is a test that predicts how likely breast cancer is to come back. The test gives a score between 0 and 100. |
| **LIFESTYLE DATA** | |
| Diet | Diet categories: poor diet, moderate diet, healthy diet, very healthy diet |
| Exercise | Exercise categories: little exercise, moderate, active, very active |
| Smoking | Categories for smoking according to the USA Centre for Disease Control (CDC): never smoker, former smoker, everyday smoker |
| Drinking | Drinking categories: never drinker, moderate drinker, social drinker, heavy drinker |

***Table C.2.*** Variables collected for the prostate cancer study and their explanation.

| **ABBREVIATION** | **MEANING** |
|---|---|
| **DEMOGRAPHIC DATA** | |
| P_id | The ID of the participant (not related to their medical ID used at the hospital) |
| Age | age of the participant the time of the interview |
| Age_diag | age of the participant when diagnosed |
| Age_group | Age groups: < 55; 55 - 64; 65 - 74; >75 |
| Years_cancer | Number of years a patient has been living with cancer |
| Dod | date of death of the participant |
| Province | one of the 4 provinces of the Republic of Ireland: Connacht, Leinster, Munster, Ulster |
| M_status | marital status: single, married, partnership, widowed, divorced, unmarried |
| Education_irish | Irish education system |
| Employment | employment status of the participant during the interview and the care phase: not working, no info, part-time, full time, retired |
| Religion | the religion of the participant: religious, not religious |
| Insurance | all participants had private insurance (private hospital) |
| **MEDICAL DATA** | |
| Hearing | hearing status of the participant: normal, impaired |
| Vision | vision status of the participant: normal, impaired |
| Allergies | allergy status if a participant had either drug or food allergies: no, yes |
| Height_cm | the last measured height of the participant in centimeters |
| Weight_kg | the last measured weight of the participant in kilograms |
| Bmi | body mass index |
| Bmi_groups | body mass index groups: underweight, normal, overweight, obese |
| Plan_staging1 | the first examination in the diagnosis of cancer |
| Plan_staging2 | second examination in the diagnosis of cancer |
| Plan_staging3 | third examination in the diagnosis of cancer |
| Plan_staging4 | fourth examination in the diagnosis of cancer |
| Plan_staging5 | fifth examination in the diagnosis of cancer |
| Date_biopsy | the date when the breast biopsy was done |
| Type_biopsy | clinical trial participation |
| Cancer_type | the type of breast cancer: invasive ductal carcinoma, invasive lobular carcinoma, metastatic breast carcinoma, Cancer of unknown primary |
| Metastasis | has cancer metastasized: no, yes |
| Mets_organ1 | the first organ cancer metastasized |
| Mets_organ2 | the second organ cancer metastasized |
| Mets_organ3 | the third organ cancer metastasized |
| Seminal_vesicle_invasion | If cancer had invaded the seminal vesicle |

| | |
|---|---|
| Perineural_invasion | If cancer had invaded the perineural |
| Path_stage | Pathological stage of the cancer |
| Gleason_s | Gleason score of the cancer |
| L_nodes | Lymph node invasion |
| Initial_psa | The initial prostate-specific antigen level |
| Pretreat_exam1 | the first examination before treatment |
| Pretreat_exam2 | the second examination before treatment |
| Pretreat_exam3 | the third examination before treatment |
| Diag_treat_days | days between the diagnosis day and the start of the treatment |
| Diag_ehr | the day the diagnosis was made |
| Treat_ehr | the day treatment started |
| Clinical_trials | participation in clinical trials |
| Care_phase | the care phase the participant was in at the moment of the interview: diagnosis, treatment, follow-up, palliative (participant was selected to have at least started their treatment, so there were no participants on their diagnosis phase. These patients were asked to participate in the study at a later stage when they had gone through some part of their treatment phase.) |
| Treatment_line1 | the first line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Treatment_line2 | the second line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Treatment_line3 | the third line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Treatment_line4 | the fourth line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Treatment_line5 | the fifth line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Treatment_line6 | the sixth line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Treatment_line7 | the seventh line of treatment: surgery, chemotherapy, radiotherapy, endocrine therapy, bisphosphonate, Xofigo |
| Postt_exam1 | first examination after treatment |
| Postt_exam2 | second examination after treatment |
| Postt_exam3 | third examination after treatment |
| Postt_exam4 | fourth examination after treatment |
| Postt_exam5 | fifth examination after treatment |
| Postt_exam6 | sixth examination after treatment |
| Surgery | if the participant had surgery as part of the treatment: yes/ no |
| Surgery_date | the date when the breast surgery was done |
| Type_of_surgery | the type of surgery that was undergone: prostatectomy, TURP |
| Icd10_code1 | comorbidity 1 coded in ICD10 |
| Icd10_code2 | comorbidity 2 coded in ICD10 |
| Icd10_code3 | comorbidity 3 coded in ICD10 |
| Icd10_code4 | comorbidity 4 coded in ICD10 |
| Icd10_code5 | comorbidity 5 coded in ICD10 |
| Chemotherapy | Chemotherapy treatment: yes/no |

| | |
|---|---|
| Radiotherapy | Radiotherapy treatment: yes/no |
| Endocrine_therapy | Endocrine therapy treatment: yes/no |
| Bisphosphonate | Bisphosphonate treatment: yes/no |
| Xofigo | Xofigo treatment: yes/no |
| Recurrence | Cancer recurrence: yes/no |
| Family_history | Family history of any cancer: yes/no |
| **LIFESTYLE DATA** | |
| Diet | Diet categories: poor diet, moderate diet, healthy diet, very healthy diet |
| Drinking | Exercise categories: little exercise, moderate, active, very active |
| Smoking | Categories for smoking according to the USA Centre for Disease Control (CDC): never smoker, former smoker, everyday smoker |
| Exercising | Drinking categories: never drinker, moderate drinker, social drinker, heavy drinker |