SAPIENZA
UNIVERSITÀ DI ROMA

A Bayesian probabilistic record linkage method to perform survival analysis for joint prostheses when the operated side is not available

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica

Dottorato di Ricerca in Scuola di dottorato in Scienze Statistiche – XXXIV Ciclo

Candidate

Enrico Ciminello
ID number 1391205

Thesis Advisors

Prof. Marco Alfò
Ing. Marina Torre

January 2023

A Bayesian probabilistic record linkage method to perform survival analysis for joint prostheses when the operated side is not available

Ph.D. thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: enrico.ciminello@uniroma1.it

Contents

Introduction **7**

1 Arthroplasty Registries: an overview **9**

1.1 Registries today: definition, features and aim 11

1.1.1 Features of a registry 11

1.1.2 Statistical analysis of registry data 13

1.1.3 The present and the future: patient-oriented registries 15

1.2 The Italian Arthroplasty Registry 16

1.2.1 Data collection and participating regions 17

1.2.2 The Italian Registry now 19

2 Survival Analysis based on Hospital Discharge Records **23**

2.1 The model 25

2.1.1 The probability of a clinical path: the case with two observed primaries 29

2.1.2 The probability of a clinical path: the case with only one observed primary 33

2.1.3 Deriving the complete Model 37

2.2 The use of prior information in the Probability Matrix 39

2.2.1 Causes of revision and failure times 39

2.2.2 Elicitation of the prior information 40

2.3 Formalizing the model 42

3 An application to simulated data **45**

3.1 Data generation process 46

3.2 Model performance assessment on simulated data 47

3.2.1 Methodological differences between Mode and Median approaches 54

3.2.2 Comparison with literature 55

4 Application to Registry data 59

4.1 Probability for failure time by cause of revision 59

4.2 Data exploration and preprocessing 60

4.3 Results 62

4.3.1 Comparison with standard record linkage methods on real data 65

Concluding Remarks 67

A Range of variation for Kaplan-Meier estimates 69

B Theorems 73

C ICD9-CM codes mapping table 81

D R code 87

Bibliography 96

Abstract

Surveillance on implantable joint prostheses is crucial to assess performance of devices and ensure safety of patients. In data collection, unique identifiers for patients and information about the operated side are necessary to correctly follow implanted devices over time and perform accurate survival analysis. Hospital Discharge Data is the largest source of information on joint replacement surgery in Italy; however, it does not include the operated side variable, making follow-up impossible and survival estimates unreliable. This work provides a Bayesian Probabilistic Record Linkage model that allows to link, with good accuracy, primary interventions and corresponding revisions (due to implant failure); this makes follow-up possible and precise enough to perform survival analysis with results close to those that can be obtained when the operated side is known.

Introduction

Joint arthroplasty constitutes a major innovation in clinical practice to treat degenerative diseases, like osteoarthritis, or trauma events, like fractures, or other specific joint diseases. The principal benefits of this kind of interventions are relief of pain, improved functionality and a better quality of life [96]. For these reasons and their high effectiveness, hip and knee arthroplasty were defined the intervention of the 20^{th} century [59] and of the first decade of the millennium [70], respectively. The important observed increase in the number of joint replacement performed worldwide through the years, starting from the 80's, confirms these results [40, 53, 85, 76, 8, 91]. Indeed, since 2009, the number of hip and knee replacements has increased rapidly in most OECD countries. On average, hip and knee replacement rates increased by 22% and by 35% [75], respectevely, between 2009 and 2019. In Italy, the number of joint replacements have more than doubled in 19 years, from 105,491 in 2001 to 220,445 in 2019 [14]

In this context, an accurate assessment of safety and effectiveness of joint prostheses is necessary. In many countries, national arthroplasty registries perform this task at national level, mostly applying survival analysis techniques to estimate the lifespan of the implants. In Italy, the Italian Arthroplasty Registry (Registro Italiano ArtroProtesi, RIAP), was established at the Italian National Institute of Health (Istituto Superiore di Sanità, ISS) and started collecting data in 2013, after an initial pilot project [100]. However, due to its limited time window, geographical coverage and low completeness of data collection in most of the participating regions, RIAP can carry out efficient analyses for devices assessment and clinical practice evaluation only in few limited contexts and not at national level.

The National Hospital Discharge Records (HDRs) Database contains information on all hospital admissions performed within the national territory [17]; it is transmitted, every year, by the Ministry of Health to ISS, to support research on public health. For arthroplasties, this research is carried on by RIAP. Even if such data does not contain any information about implanted devices, it might fruitfully be used to perform statistical analysis to assess the clinical practice in arthroplasty surgery at national and regional level. To this aim, the major drawback to overcome is the absence of the "operated side" variable in HDRs, that currently prevents to associate every revision procedure (and associated failure) with its related previous implant, making survival analysis impossible.

The aim of this work is to propose a probabilistic record linkage model to match the record of the first implantation with subsequent records about revisions performed on the same joint, despite the missing information about the side; by exploiting such a procedure, reliable survival analyses can be produced when using the national database of Hospital Discharge Records.

Chapter 1 contains an overview on arthroplasty registries and statistical methods that are routinely used to assess safety and effectiveness of implanted devices.

In chapter 2 the probabilistic record linkage method is presented and the underlying mathematical structure is described.

The performance of the model is evaluated in Chapter 3, by the use of simulated data.

Finally, the model validation on real data, collected in the Autonomous Provinces of Trento and Bolzano from 2010 to 2018, is presented in Chapter 4. For such data, the operated side information is available and, therefore, results obtained via the proposed linkage procedure can be compared with ground truth.

Chapter 1

Arthroplasty Registries: an overview

Joint replacements are a well-recognized valuable solution to relieve pain, restore joint function and improve patients' quality of life [74]. For this reason, hip and knee replacements have been called the prosthesis of the century [59] and the prosthesis of the decade [70], respectevely. The development of hip arthroplasty to restore joint function has its roots in the late XIX century [51]. After decades of various attempts, the revolutionary concept of "low friction arthroplasty" has been introduced in the early 1960s by Sir John Charnley, paving the way to the current hip [11] and knee arthroplasties [97]. In the following 50 years, remarkable advances were observed. However, despite the development of new clinical and surgical techniques and many successes, patients have been harmed by many failures [32]. Besides the most common side effects, such as infections, thromboembolic complications, prosthetic component breakage or wear, loss of fixation, osteolysis and many others requiring revision surgery [63], in the new millennium, a high incidence of important problems related to surface arthroplasties, metal-on-metal, modular neck, taper corrosion were observed [32], with serious long-term effects on patient health [65].

To monitor devices and patients' safety, the need for collecting information about interventions and measuring outcomes became evident, leading surgeons to understand the importance of a detailed follow-up of their patients and the need of the development of systems for tracking their patients and documenting outcome over time. The first example of this approach comes from the USA: it was the Mayo Registry, established at the Mayo Clinic in Rochester, Minnesota in 1969. This registry includes detailed data on surgical procedure, joint features and information on short and long-term complications for hip, knee, shoulder, and elbow replacements performed since 1969. Purpose of a registry is "to collect institutional, regional, or national data in order to analyze and draw statistically significant conclusions regarding patient, surgical technique, and implant associated risk factors that lead to good or poor outcomes" [63]. Nevertheless, the number of patients and interventions collected in the Mayo registry was too small to perform proper statistical analysis, leading to an inherent risk for performance bias. Moreover, only the few kind of devices and surgical techniques used in that specific institution could be assessed. Even considering all the limitations of a registry collecting data from a single institution, the Mayo registry was an important first step to create a tool to perform epidemiological studies and analysis of risk factors to gain knowledge about the treatment of several complications and side effects of arthroplasty.

The first nationwide orthopaedic registry was created in Sweden in 1975 and it collected data only about knee arthroplasty [84]. Four years later, in 1979, always in Sweden, the first national registry tracking hip replacement interventions was established [62]. Starting from such experience, several countries started developing their own national registry and, nowadays, arthroplasty registries are fundamental tools for surveillance and vigilance of medical devices worldwide. Their role is to collect information on joint replacement surgery and to monitor the performance of implanted devices [63].

Registries can assess performance, effectiveness and safety of the different kinds of implants used in joint replacements. This task is performed by collecting data about the surgical practice, with a key focus on patient quality of life. Thanks to registry data, clinical practice might be improved, providing useful feedback to the whole orthopaedic sector, manufacturers, industry and procuring stakeholders [29].

The effectiveness of registries in improving clinical practice and safety of devices is evident when considering that since the 80's, by getting feedback from annual reports, the decision making of Swedish surgeons about surgical techniques and devices performances has driven them to a 10-year total knee arthroplasty revision rate lower than 4% [110]. The importance of an efficient and high quality registry as a surveillance and vigilance tool can be understood by using a real life example from 2007. Outcomes from the National Joint Registry for England and Wales highlighted poor performances in terms of survival of metal-on-metal joint replacements [19]. Moreover, this kind of devices led to unwanted effects, that are unusual for arthroplasty, in a large amount of patients, resulting in an increased risk for safety

[36]. The concern exposed by the registry allowed for the Articular Surface Replacement (ASR) metal-on-metal hip arthroplasty system by DePuy Orthopaedics to be removed from the market, even though with a certain delay [5]. This shows how assessing and monitoring devices is important in order to detect poorly performing implants to guarantee the safety of patients [107].

1.1 Registries today: definition, features and aim

The International Medical Device Regulators Forum (IMDRF), a worldwide group of medical device regulators that includes several institutions - like the European Commission Directorate-General for Internal Market, Industry, Entrepreneurship and SMEs and the U.S. Food and Drug Administration - and that is directly observed by the World Health Organization, defines a registry as follows:

> ”An organized system with a primary aim to increase the knowledge on medical devices contributing to improve the quality of patient care that continuously collects relevant data, evaluates meaningful outcomes and comprehensively covers the population defined by exposure to particular device(s) at a reasonably generalizable scale (e.g. international, national, regional, and health system)” [67].

When considering arthroplasty registries, the main goal is then to collect information about any joint replacement of interest performed in a certain hospital, region or country, covering primary interventions and revisions of previously implanted devices. Individual patient data such as age, sex, diagnosis are collected, along with surgical technique and type of implant used.

By using this data, it is possible to get statistical summaries regarding the reoperated patients (that is, to estimate the survival time of an implanted device until its failure) and provide surgeons and manufacturers with important information, useful to improve both clinical practice and device performance, in terms of effectiveness and safety. This information may concern surgical methods, types of prostheses or their materials, design and other features.

1.1.1 Features of a registry

One of the key features of a high quality registry is the linkability, namely the ability to link a primary implantation of a device in a joint to possible future revisions, for the same patient on the same side. This is necessary to study effectiveness and safety of implantable devices, as it allows to estimate the corresponding survival time and detect possible side effects. The main tool registries use to guarantee linkability is a single identifier for each patient, that allows for unique identification and remains the same over time and space. That is, the identifier of a patient has to be the same through the years and even if the patient undergoes surgery in different hospitals or regions. Such identifier might be the health insurance number, the national identity number or any kind of randomized unique code that ensures the linkability of the procedures performed on a patient. However, since a patient might need to have both joints replaced, at the same time or in different occasions, recording information on the operated side is crucial, as it allows to uniquely identify a device, avoiding bias in the estimate of the survival time of prostheses. Thus, collecting the identifier and the operated side ensures that a record about a prosthesis implantation during a primary procedure at a given institution and date, can be linked to subsequent revisions of the device, even if performed elsewhere.

The other fundamental dimensions for a registry that ensure reliable outcomes are high coverage and completeness. Coverage is defined as the ratio between the number of institutions participating in the registry and the total number of institutions performing interventions of interest for the registry in that country [109]. A high level of coverage implies a large territorial extension of the data collection and the ability to detect the performance of a wide range of devices and surgical techniques used in interventions. This allows also to track the use and consequent performance of a specific kind of prosthesis, implanted in patients with different characteristics and using different clinical approaches, avoiding bias due to ability of surgeons or to the not comparable effectiveness of the same device in different conditions.

Completeness measures the percentage of interventions of interest performed in a single institution caught by the registry out of the total number of interventions of interest performed in that institution [109]. Obviously, despite a good coverage, while completeness is low, analyses and reporting could be misleading.

Along with quantity, measured by high levels of completeness and coverage, data quality is a key factor for the ability of a registry to improve in terms of reliability of feedback and to be able to drive all the stakeholders to progress in clinical research. Quality and harmonization of data submitted by hospitals and agreement on a common set of statistical tools are crucial prerequisites for comparison between different institutions or countries conducting joint replacement surgery [1].

1.1.2 Statistical analysis of registry data

The vital status of patients is not of crucial interest for arthroplasty registries, as the main focus is about implants. This means that if a patient dies while the prosthetic implant is still correctly in place, this is not considered as a failure whether the death of the patient is not strictly related to the device itself (as it can happen because of infections), but rather as a censoring of the observation time. In this context, implant revision is the primary and most important endpoint. Therefore, it is useful to properly define what a revision is, as this is a notion of major importance in the following: a revision is defined as a "new operation in a previously replaced joint during which one or more of the components are exchanged, removed, manipulated or added" [60]

Time to revision is the most important criterion to assess performance of implants in total joint surgery. Estimates of failure probabilities are used as a measure of the risk of suffering a reoperation and the Kaplan-Meier (KM) estimator is the main tool used by arthroplasty registries to estimate the probability of revision after a primary procedure [34]. The revision-free survivorship, estimated via KM estimator and usually expressed as the complement of the estimated survival function, is widely used and is often referred to as "revision rate" in literature about arthroplasty [15]. The KM estimator is adequate because the estimate it provides for the probability of reoperation results to be an easily interpretable, understandable, and clinically relevant quantity. Then, at the time being, the KM estimator sets the gold standard and it hardly will be superseded [106].

Nevertheless, some authors argue the effectiveness of this approach because of the potential presence of competiting events, with subsequent violation of the assumption of identical revision risk in censored and uncensored patients. Those authors claim that the patients' death, implying the presence of censored time to event when using a KM approach, may rather be considered as a competing event, to avoid biased estimates of the revision risk [7, 6]. In particular, the KM estimator is based on the assumption the patients with a censored time have the same probability of revision at any subsequent time of patients with a observed event. Therefore, when KM estimator is used in the analysis of registry data, deaths are handled as right-censored observations. Essentially, the implicit assumption when using the KM method, is that dead patients would have the same probability to undergo a revision intervention than alive patients.

Models for the analysis of competing events can be used to account for concurrent mortality in order to estimate failure probabilities [95]. This alternative approach leads to the use of cumulative incidence functions (CIF) [47, 27], where patients' deaths are considered to be competing events, while patients who reach the end of the follow-up alive and unrevised are censored. The assumption behind the KM approach generally leads to an overestimate of the revision rate with a statistically significant difference between the estimates obtained by KM approach and the one obtained by competing risks approach [55, 56].

However, this bias is not always observed, as it depends on the distribution of deaths in the groups being compared and on statistical tests used to investigate the differences in estimated survival curves. Indeed, the difference between the performance of KM and CIF methods, with death considered as a competing event, may be dependent on the death probability in the group of patients under analysis, as CIF estimates of the revision rate are generally lower when the failure rate for the competing event (death) is high [24]. Indeed, this difference in estimates is more evident in groups of patients with higher risk of death and in elder people in general, leading to greater difference between KM and CIF estimates of the survival curve [64, 7].

The choice of the estimating method depends on the aim of the study: while the aim of the analysis is to describe the failure of an implant, compare longevity performances or give regulatory headlines, the KM method is the most appropriate, as death with a device still in place, without adverse event occuring, is not considered as a failure, but rather as a censoring of the observation time. On the other hand, the CIF estimator is recommended when the topic of the study is about resource planning, health economics, or communication with patients on their risk of undergoing a revision intervention [82, 93].

To have a complete overview on standard models to estimate the failure probability in arthroplasty studies, it is worth mentioning the Cox model, that is often used to estimate the effects of covariates on the hazard of the failure occurrence. While it is widely used and considered as a benchmark in literature for survival analysis [83, 58], the hypothesis of proportional hazards on which it relies on does not generally hold when data from arthroplasty registries are considered [81]. Some adjustments with respect to the standard model have to be done to circumvent this issue. Therefore, a weighted Cox regression model has been proposed to deal with the problem of non-proportional hazards providing estimates of average hazard ratio [94].

1.1.3 The present and the future: patient-oriented registries

As described so far, estimated survival time of prostheses is used as the main endpoint to determine whether surgery has been a success or a failure and to assess safety and efficacy of devices. It is of primary interest to learn how long implants can correctly be in place without occurrence of infections, re-admissions, and other adverse events such as thromboembolic and cardiovascular complications. It is also of main importance to detect any devices leading to early failures. But, mainly in last decades , these quantities have not been considered as sufficient anymore to measure the ultimate outcome for patients [113]. Indeed, arthroplasty surgery is performed to decrease pain, restore functionality and movement and improve quality of life. Therefore, it is logical to consider these factors as further endpoints when assessing efficacy and safety of implanted devices. Revision is an event that is straightforward to understand and with immediate clinical relevance, but it is insufficient to measure the success of performed surgeries. In fact, survival after one year is close to 100% while self-reported satisfaction is 80% after total knee arthroplasty and 90% after total hip arthroplasty [112]. If an implant is considered to have failed only when a subsequent revision happens, then interventions on patients with a poor outcome in terms of pain, range of movement, walking and radiological appearance, which do not imply a revision surgery, are considered as a success [12, 71].

To satisfy the need to collect information on patients' quality of life after surgery, the Swedish registry started in 2002 with a pilot project to introduce the Patient-Reported Outcomes Measures (PROMs) follow-up programme [86]. A Patient Reported Outcome consists in patients reporting about their own health status, in terms of relief of pain, effectiveness of the surgery and possible improvement of quality of life, without interpretation or any kind of filtering by clinical experts [88]. By this approach, a generic health status and any health-related quality of life improvement are reported, along with data about specific features concerning the performed treatment, such as, in case of arthroplasty surgery, range of movement of the joint, ability to walk and any pain and functionality-related information. Usually, these outcomes are collected by making patients fill a self-completed questionnaire [111].

The most developed, high quality and functioning national orthopaedic arthroplasty registries, such as those established in Northern Europe, collect Patient-Reported Outcome Measures on all or on samples of patients undergoing hip and knee arthroplasty and produce feedback and reports on a yearly basis [87]. The information collected can be of primary importance to focus on patients' needs and improve their care, to establish a data-driven decision making process about surgical timing and clinical approach. They can also improve the effectiveness of provided care, leading to better choices in terms of procuring of devices, surgical techniques and postoperative follow-up and rehabilitation [9].

1.2 The Italian Arthroplasty Registry

In Italy, joint replacements showed a huge increase in terms of volume of activity since early 2000's, with an average yearly rate of 4.2% between 2001 and 2019 [14], reaching over 200,000 arthroplasties performed in 2019, with an impact of almost two billions euros on National Health System [105].

That is why, in 2000, regional registries were developed, collecting data for patients resident in the region and operated within the region limits [98, 22]. Nevertheless, this approach resulted to be not effective because of the high observed interregional mobility due to the freedom for patients to move from a region to another to be medically assisted. Patients operated in a region and revised in another region refer to two distinct regional healthcare systems and the tracking of their device is lost. This means that a regional registry might consider still in place a number of implanted devices that are instead revised in another region, with a huge bias in performance and safety estimates [89, 90].

In 2006, the General Directorate of Medical Devices and Pharmaceutical Service of the Ministry of Health funded a pilot project of the Italian National Institute of Health aimed at implementing a national arthroplasty registry, organized as a federation of regional registries. All the stakeholders are involved: public health institutions, scientific societies, biomedical industries and patients associations, by including their representatives in the Registry Steering Committee, namely the body responsible for the work and progress of the Registry [103]. Regions are enrolled on a voluntary basis and, at the time being, eleven regions and two autonomous provinces are participating in the project (Figure 1.1) [101].

The importance of the registries as a tool to support the medical device governance is recognized by the EU Regulation on medical devices 2017/745 of 5 April 2017, stating that:

> ″The Commission and the Member States shall take all appropriate measures to encourage the establishment of registers and databanks

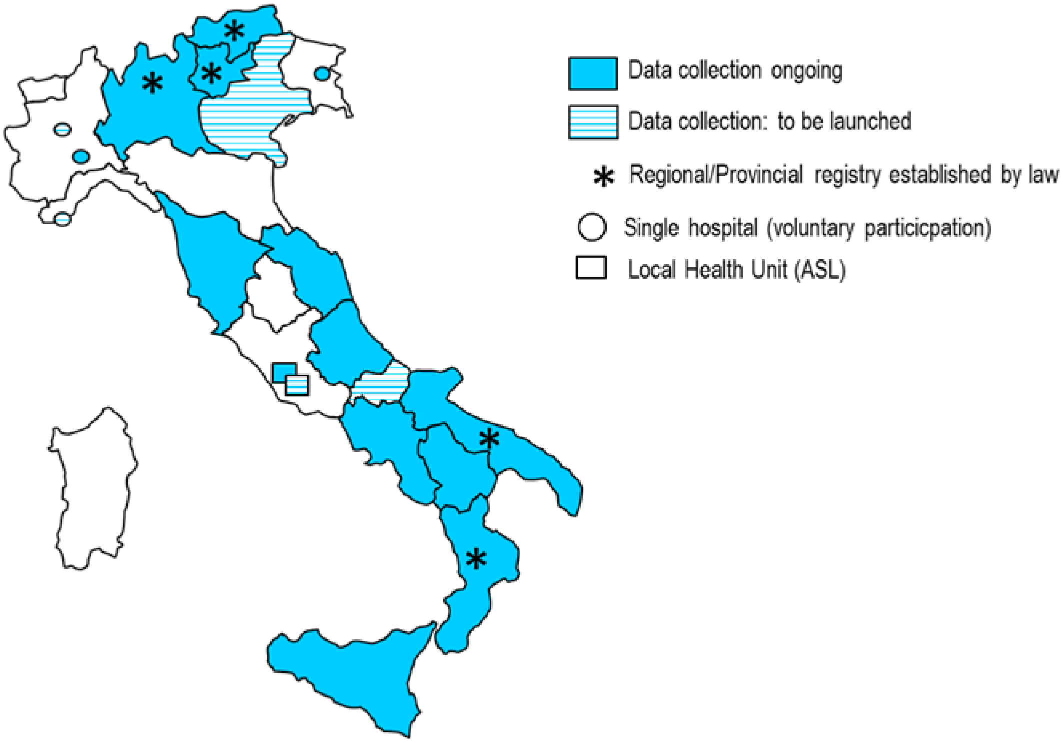

Data collection ongoing
Data collection: to be launched
✱ Regional/Provincial registry established by law
○ Single hospital (voluntary particicpation)
□ Local Health Unit (ASL)

Figure 1.1. Regions and institutions participating in RIAP at 31 december 2020 [102].

> for specific types of devices setting common principles to collect comparable information. Such registers and databanks shall contribute to the independent evaluation of the long-term safety and performance of devices, or the trackability of implantable devices, or all of such characteristics.” [20]

At that time, the Italian government was already working in this direction: by the Prime Minister decree of May, 12th, 2017 [44], according to the Decree Law n. 179 of October, 18th, 2012 [43], the government had just established the Italian Registry of Implantable Prostheses (Registro Italiano Protesi Impiantabili, RIPI) at the Italian National Institute of Health. The RIPI includes the Italian Arthroplasty Registry (Registro Italiano ArtroProtesi, RIAP), aimed at collecting data about the use of prosthetic devices, in order to assess their effectiveness and safety once they are on the market, and tracking patients in case of recall of a specific device.

1.2.1 Data collection and participating regions

A record from the Italian Registry is composed by two parts: 1) a set of variables directly picked from the original Hospital Discharge Record; 2) the so-called Minimum

Data Set, derived by a form compiled for each performed intervention, including information of specific interest to Registry purposes.

The Hospital Discharge Record section collects data on:

- hospital;
- patient unique identifier;
- patient demographics;
- clinical procedures performed during the hospitalization (coded sccording to the International Classification of Diseases, 9^{th} Revision, Clinical Modification, the ICD-9-CM);
- patient general health and possible comorbidities (ICD-9-CM);
- administrative information.

The Minimum Data Set section collects information on:

- operated joint;
- operated side;
- surgical procedures (approach, fixation, others);
- diagnosis for primary intervention and revision;
- class of the implanted device according to the National Classification of Medical Devices (CND) [41];
- manufacturer;
- device catalog code;
- lot

Part of the Minimum Data Set is about the implanted device. Collected data allows for unique tracking of the prosthesis by using the patient identifier along with the side specification. This information is of crucial interest in case of patients with bilateral prostheses, that is when a patient has both joints of the same kind (e.g. both hips) totally or partially replaced by an implantable device. This usually happens because of degenerative diseases like osteoarthritis, since patients with severe osteoarthritis in one operated joint may also have symptoms of disease development on the other side [16] and will undergo a second arthroplasty to replace the other joint in the following few years.

Particular attention is paid to the unique identification of every single component of the implant. The main tool used to ensure the correct tracking of these components is a dedicated library (RIAP Medical Devices Dictionary) implemented and continuously updated by manufacturers with their new products and relative features. The information collected in the library includes catalog code, name of manufacturer, description, CND code, and the General Repository of the Medical Devices registration code [103]. This last one is an identifying code that all the medical devices receive when they are registered in the National Medical Device database of the Ministry of Health. Registration is mandatory to allow devices to be marketed in Italy and used within the National Health System [42].

To ease collection of records, in particular for the Minimum Data Set section, and to make data submission more immediate, the Italian National Institute of Health developed two web apps: RaDaR and SoNaR. RaDaR is aimed to facilitate the task for surgeons to compile information required for the Minimum Data Set. SoNaR is the tool used by regional coordinating centers to send data to the Registry, after linking the Hospital Discharge Record and the Minimum Data Set.

The Italian National Institute of Health has access to all data submitted by the participant regions, while regions have access only to the regional data they sent. This data flow is of crucial importance as RIAP needs high quality data with uniform structure from all the participating regions for proper analysis to be carried out. Once data is available, the Registry can produce accurate statistical and epidemiological analyses and provide all the stakeholders with useful feedback. Surgeons and clinicians, in general, can better understand the effectiveness and safety of devices and techniques they use, as well as manufacturers may receive fundamental information for their postmarket surveillance and possible feedback on malfunctioning devices to recall. The data collection flow is resumed in Figure 1.2 [108].

1.2.2 The Italian Registry now

At the time being, the Registry is focused on several activities. The RIAP library dedicated to Medical Devices is continuously updated and, according to the last published report [102], it includes over 80,545 codes for products by 110 manu-

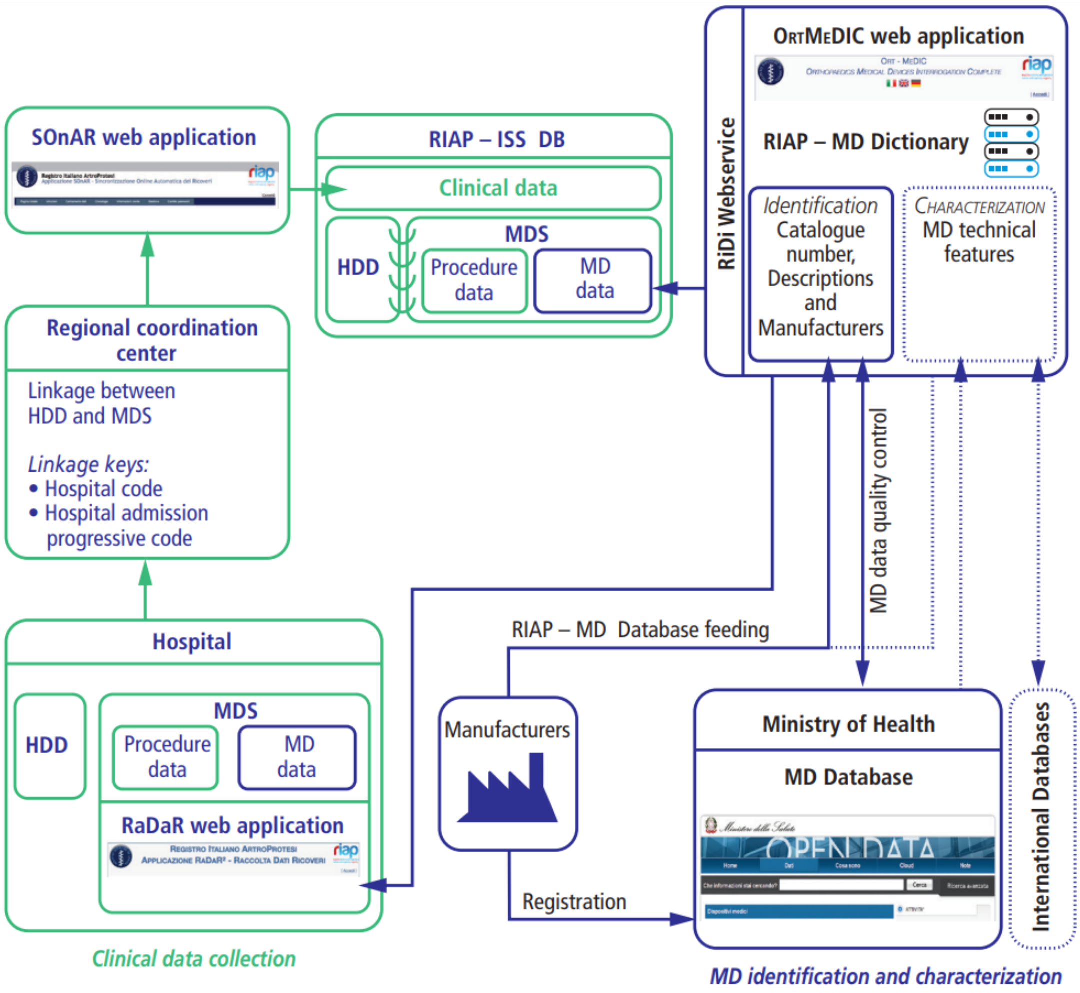


Figure 1.2. Flow diagram of the Italian Arthroplasty Registry data collection model [108].

facturers. This activity and the collaboration with the National Joint Registry of England and Wales (NJR), aimed to connect the RIAP Medical Devices Library and the NJR Component Database, will lead to implement a useful tool that will collect information about a wide range of medical devices used across Europe [68].

The development and maintenance of the web applications SoNaR and RaDaR are of primary interest as well, since participation in the Registry is on a voluntary basis and facilitating the compiling of the forms and submission of data is crucial. Data collection still appears to be a hard task for clinicians, as it requires time and competence, and, in case the method for filling records and submitting data would not be supported and eased by tools like SoNaR and RaDaR, they could be unwilling to participate in the RIAP [99].

In the last years, the RIAP is also trying to match the vision of patient-oriented registries by the development of a Patients Reported Outcome approach. The first pilot project was aimed to measure the quality of life for patients undergoing hip arthroplasty following the guidelines for the Hip disability and Osteoarthritis Outcome Score (HOOS), that is designed to assess outcomes in patients with osteoarthritis after total hip replacement [104].

Participation in the Registry is not mandatory for most regions and this implies low levels of coverage and completeness. By the use of the Hospital Discharge Records collected in the country every year, obtained from the Ministry of Health, the Registry has estimated the completeness for replacement interventions on hip (34.9%), knee (36.5%) and shoulder (11.3%), performed in 2019. The overall number of collected interventions is equal to 75,682, representing in terms of completeness only 34.2% of all interventions performed with respect to the whole national activity volume. For the regions and hospitals participating in RIAP, the average completeness was equal to 65.8% for hip, 63.7% for knee and 52% for shoulder replacements [102].

Unfortunately, these values for completeness are not enough to guarantee reliable analyses and results. Registries outcomes are considered to be acceptable only if based on data with a completeness not lower than 90%, with high quality standards [52]. To implement survival models, another key feature to be considered is the amount of years covered by the data collection. Since the medical devices of interest have usually a long lifespan with high probability (e.g. the National Joint Registry of England and Wales reports that, by using KM estimator, the estimated revision rate for hip implants is 8% after 16 years from primary intervention [4]), a long time window of observation is needed to get reliable estimates. Even if RIAP activities started in 2006, data have been collected with high quality standards only in the last few years. Moreover, the high variability of quality among the regions represents a further limit that does not allow for accurate analysis in Registry reports.

That is why RIAP uses Hospital Discharge Data, that is transmitted every year from the Ministry of Health, to carry out epidemiological analyses and have a frame of the arthroplasty activity in the country. Results on activity by region, focused on demographics and clinical features are produced and published in reports [13, 14] and allow stakeholders to have information and improve their decision-making.

Chapter 2

Survival Analysis based on Hospital Discharge Records

Hospital Discharge Data (HDD) is a huge source of information. Every year the Ministry of Health provides the National Institute of Health with Hospital Discharge Records (HDRs) covering hospital stays occurred in the whole country and, at the time being, such data is available for years since 2001. By using this data, it could be possible to carry out some useful analyses for topics of interest for the RIAP, but limitations exist and have to be considered.

As it comes from an administrative source, HDD does not give any information about the implanted device, such as material, surgical technique or manufacturer, that are usually reported as part of the Minimum Data Set in a Registry record. It only registers that the patient underwent arthroplasty surgery, specifying whether it was a primary intervention or a revision (implying a failure) of a previously implanted device. While assessing effectiveness and safety of implanted devices is impossible, information about hospital, region of hospitalization, demographics of the patient and diagnoses are registered. This information makes the data suitable to analyse different kind of topics: investigation of risk factors for primary arthroplasty or failure, evaluation of performance by region, studies about inter-regional mobility, and, most important, overall performance of arthroplasty surgery in Italy.

To carry out such analyses, a major issue has to be overcome. HDD reports a unique identifier for each patient, that allows to follow patients in time (the same identifier is used since 2001) and space (the identifier is the same whatever the region of the hospital), but the operated side is not available. This means that, for patients undergoing bilateral surgery, the link between a primary and the corresponding revisions is ambiguous.

Let consider, for instance, a patient with osteoarthritis that has the left hip replaced. Two years later, because of the development of the disease, the patient undergoes a further surgery to have a prosthesis implanted also in the right hip. After five years, one of the devices is revised because of a failure; the record corresponding to this last hospitalization reports that the patient underwent a hip revision arthroplasty, but the side is not specified. In such a case, matching the primary and the revision (and consequent failure) is not possible. So, there is no way to know whether the first implanted device survived seven years or the second implant survived five years. This may introduce a (potentially huge) bias in survival estimates, in particular while considering that thousands of patients have been operated on both sides for the same kind of joint.

To explore the magnitude of the resulting bias, due to the missing information on the operated side, Registry data (with side information available) from the Autonomous Provinces of Trento and Bolzano have been used. RIAP referents for Autonomous Province of Trento provided an accurate back reconstruction of registry data [78], while data from the Autonomous Province of Bolzano is continuously collected since 2010 [77]. Thus, in both cases, high quality registry data is available from 2010 to 2018 .

Data at disposal is composed of 10,638 records for primary intervention aimed to implant hip prostheses in 9,588 patients, with 1,050 patients who had both hips replaced. Figure 2.1 shows the result of the analysis: the black line is the complement of the Kaplan-Meier (KM) estimate of the survival curve with known operated side, while in gray there is a surface where the possible KM estimates of the curve lie when the operated side is unknown and the linkage between primary and revision interventions is arbitrary or performed rondomly or in a naive way. Such surface is carried out by taking the extreme values of the KM estimates of the complement of the survival curve in four linkage scenarios, depicted in table A.1 in Appendix A. The analysis of real data from Autonomous Provinces of Trento and Bolzano leads to the KM estimate of the revision-free survivorship at nine years equal to 3%, when the operated side is known; however, while information about the side is missing, the value can be overestimated up to over 4%, as well as underestimated to a value down to 1% if the information about the operated side is unavailable. This would mean an overestimate up to 33% and an underestimate down to 66% with respect to the estimate obtained with complete information. As already stressed, the KM estimate of the survivorship, often referred to only as "revision rate" in

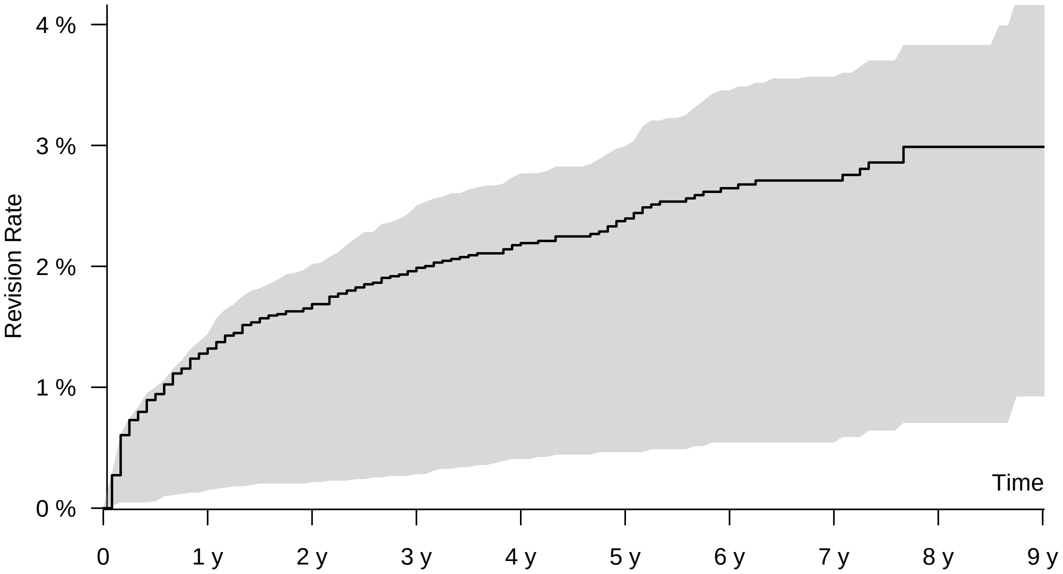


Figure 2.1. Bias observable in the KM estimated survivoship in case of random or naive linkage between primary interventions and revisions. Black line: KM estimate while side is known. Gray surface: region where KM estimates lie with random or naive linkage between primary interventions and revisions while the operated side is unknown.

registries worldwide [15], is the most important and widely used quantity to assess the performance of arthroplasty and such a huge error in the corresponding estimate can be misleading for clinicians and stakeholders in general.

This example shows how producing reliable survival analysis in not straightforward when using HDD. The missing knowledge of the operated side is an issue of crucial importance.

The focus of this work is then about finding a method to overcome this issue and produce an accurate record linkage, so that primary arthroplasty interventions performed on a patient can be matched with corresponding revisions, performed on the same side, to make the estimation of relevant quantities of interest feasible.

2.1 The model

Record linkage has been widely explored in literature as it gives a useful tool to perform population-based studies in epidemiolocical, demographic, clinical and public health domains, among others.

Record linkage has two main applications [23]:

- merging different data about the same statistical unit (e.g. a patient) from different databases: for example, it can be used to link hospital discharge records to laboratory data and other records from financial or demographic databases;

- matching records in the same dataset that are related to a unique statistical unit and may have been compiled in different places and at different time: e.g. multiple records related to different hospital admissions for the same patient.

Record linkage is generally needed because of the absence of a unique identifier for the same patient between or within databases [18]. It relies on the use of a combined set of variables with a huge discriminating power to identify patients, like first and last name, date of birth, postal code and others [80, 92]. These are referred to as partially identifying variables.

Two approaches are typically used in Record Linkage: the Deterministic and the Probabilistic one.

In **Deterministic Record Linkage** all of the chosen items in the set of partially identifying variables have to agree, namely, they have to show the same value in all of the records that are considered to match (belong to the same unit). If there is disagreement in just one variable, then two records are not considered to be linked [28].

In **Probabilistic Record Linkage**, weights for agreement or disagreement are estimated for each item in the set of partially identifying variables. If the total sum of these weights is above a pre-determined threshold, the records are considered to belong to the same unit [46].

Both methods are commonly used to merge information from different datasets or different records; the major drawback consists in the bias introduced by imperfect linkage, corrupted or wrongly filled records, lack of discriminatory variables or missing data. This bias and the subsequent effects in statistical analysis can be huge and have been widely explored in literature [72, 57, 3, 50].

Starting from these two basic approaches, many extensions have been proposed, suggesting the introduction of prior information exploited within a Bayesian framework and the use of multiple imputation techniques to increase the matching accuracy between records [66, 30]. In 2012, an improvement of classical Probabilistic Record Linkage method, based on multiple imputation that exploits prior information, has been proposed [26]. The model allows to deal with one-on-one and

#	Date	Pseudo-code	Type	Side
1	2005/01	AAA	Primary	L
2	2007/01	AAA	Primary	R
3	2007/09	AAA	Revision	R
4	2008/06	AAA	Revision	L

Table 2.1. Example of four simplified records observed for one patient

one-on-many matches in presence of missing data in the matching variables. Moreover, it takes into account information from possible matches with weights lower than the fixed threshold, that are usually not considered in classical Probabilistic Record Linkage.

The problem exposed in this work can be considered to belong to that same setting, as all candidate matching can be easly found by considering the unique identifier available in the HDRs, but the missing information about the operated side makes the deterministic linkage between primary interventions and revisions unfeasible. Unfortunately, up to our knowledge, the approaches proposed by the cited authors do not fit in this case. Indeed, an actual imputation of the operated side is impossible as it has no correlation to the outcome of the intervention or any other variable.

Thus, the proposed model, described in this chapter, is not designed to impute the missing information on the operated side. It is rather aimed to reconstruct, for every patient, the corresponding hospitalization history. The proposed method can be considered a Probabilistic Record Linkage method, used to match all the records of a patient about interventions performed on the same side and subsequently, to link primaries with corresponding revisions.

The model is based on the definition of the "most probable" clinical path for a patient who underwent multiple hospitalizations due to arthroplasty interventions: for every hospitalization where a revision intervention occurred, the goal is to understand which is the corresponding previous interventions performed on the same side.

An example can be useful to better understand the concept of clinical path. Let consider a patient undergoing arthroplasty surgery four times, with two primary interventions performed during the first two hospitalizations and two revisions observed in the last two (Table 2.1). If the operated side is unknown, then revisions can not be linked to the corresponding previous interventions, but four possible scenarios (clinical paths) have to be considered:

1. first revision (intervention #3) is linked to first primary (intervention #1) and second revision (intervention #4) is linked to second primary (intervention #2), so the sequence of operated sides is LRLR;
2. first revision (intervention #3) is linked to second primary (intervention #2) and second revision (intervention #4) is linked to first primary (intervention #1), so the sequence of operated sides is LRRL;
3. first revision (intervention #3) is linked to first primary (intervention #1) and second revision (intervention #4) is a further revision of the first revision and is then linked to it (intervention #3), so the sequence of operated sides is LRLL;
4. first revision (intervention #3) is linked to second primary (intervention #2) and second revision (intervention #4) is a further revision of the first revision and is then linked to it (intervention #3), so the sequence of operated sides is LRRR.

Obviously, if the operated side sequence is the one depicted in Table 2.1, the second scenario can be easily picked up as the true one, but when this information in not available, as it happens in HDD, there is not way to know which clinical path is the true one. Choosing a clinical path at random may lead to erroneous assessment of the lifespan for those devices implanted in the first two interventions.

A formal definition of a clinical path follows:

Definition 2.1 (Clinical Path). *Given the following notation:*

n is number of revisions observed for the patient;

$\boldsymbol{a}_n$ denotes a vector with generic j^{th} element $(\boldsymbol{a}_n)_j$, $j = 1, \dots n$;

$|\cdot|$ denotes the dimension of a set;

then, a clinical path is an element of the set

$$A_n = \left\{ \boldsymbol{a}_n = \left\{ (\boldsymbol{a}_n)_j \right\}_{j=1}^{n} \right\},\ (\boldsymbol{a}_n)_j \in \{1, \dots, j+1\}$$

where the following properties are satisfied $\forall j, j' = 1, \dots n$

i) $|\boldsymbol{a}_n| = n$

ii) $(\boldsymbol{a}_n)_j \leq j+1$

iii) $(\boldsymbol{a}_n)_j \neq (\boldsymbol{a}_n)_{j'}$

The cardinality of A_n is $|A_n| = 2^n$.

From an operational point of view, a clinical path is a vector where the j^{th} position is associated to the j^{th} observed revision and its value is the equal to the position (ordered by time) of the intervention such revision refers to.

According to Definition 2.1, the set of possible clinical paths in the example reported in Table 2.1 is $A_2 = \{(1,2),(1,3),(2,1),(2,3)\}$ and the actual clinical path is $(2,1)$. The element in position one, referring to the first revision, has value equal to 2, while the element in position two, referring to the second revision, has value equal to 1; this means that the first revision is linked to the intervention #2, while the second revision is linked to the intervention #1.

The aim is then to define the probability of each clinical path to occur and to select the "best" one according to such probabilities. Linkage between revisions and primaries will be automatically assigned by choosing the clinical path for the patient.

2.1.1 The probability of a clinical path: the case with two observed primaries

The case where two primary interventions are observed is the simplest to handle. The first tool needed is a probability matrix with the number of columns being equal to the number of observed interventions and the number of rows equal to the total number of interventions minus one, as the last observed intervention is corresponding to a device still in place and that has not been revised yet. Every element of the matrix $B = \{b\}_{i,j}$ is the probability for the i^{th} intervention (row index) being revised in the j^{th} occasion (column index). An intervention can not be revision for a future intervention and primariy interventions can not be revisions for any interventions. Definition 2.2 gives a formal description of such a probability matrix.

Definition 2.2 (Probability Matrix). *Given the following notation:*

n the number of observed revisions;

k the number of observed primaries;

$i, j = 1, \dots (n+k)$ the ordered occasions at which interventions are performed;

t_i the time at which the i^{th} intervention is performed;

I_i the i^{th} observed intervention;

$f(\cdot;\cdot)$ the generic probability measure;

then, the Probability Matrix is a matrix $B \in \mathcal{M}_{(n+k-1)\times(n+k)}$ with the following structure:

$$B = \{b\}_{i,j} \text{ s.t. } b_{i,j} = P(I_j \textit{ is revision for } I_i) = \begin{cases} 0 \textit{ if } i \geq j \\ 0 \textit{ if } I_j \textit{ is primary} \\ f(t_i; t_j) \textit{ otherwise} \end{cases}$$

Examples of the possible structure of the probability matrix are given in Table 2.2.

The Probability Matrix described in Definition 2.2 is easy to interpret and helps visualize the sequence of interventions for a patient; on the other hand, it leads to some issues when it comes to operationality. In order to ease notation and computation, an auxiliary tool is derived from the Probability Matrix.

Definition 2.3 (Operational Probability Matrix). *Given the following notation:*

n the number of observed revisions;

k the number of observed primaries;

$i, j = 1, \ldots (n+k)$ the ordered occasions at which interventions are performed;

$k(j)$ the number of primaries performed until the j^{th} occasion

$j' = j - k(j); j' = 1, \ldots, n;$

t_i the time at which the i^{th} intervention is performed;

I_i the i^{th} observed intervention;

then, the Operational Probability Matrix $C \in \mathcal{M}_{(n+k-1)\times(n)}$ is derived from Probability Matrix B by dropping all the columns of primaries, that is, all the columns with all elements equal to 0 and standardizing to unit sum.

$$C = \{c\}_{i,j} \ \ s.t. \ c_{\cdot,j'} = b_{\cdot,j} \ / \ \sum_{i=1}^{n} b_{i,j} \neq 0 \ \forall j$$

$$\sum_{i=1}^{n+k-1} c_{i,j} = 1 \quad \forall j = 1, \ldots, n$$

As every revision intervention must be revision for one and only one of the previous interventions, the Operational Probability Matrix is a left stochastic matrix, namely column sum is 1 for each column.

(a)

$$\begin{array}{lccccccc} & \text{Prim} & \text{Prim} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim} & 0 & 0 & b_{1,3} & b_{1,4} & b_{1,5} & \cdots \\ \text{Prim} & 0 & 0 & b_{2,3} & b_{2,4} & b_{2,5} & \cdots \\ \text{Rev} & 0 & 0 & 0 & b_{3,4} & b_{3,5} & \cdots \\ \text{Rev} & 0 & 0 & 0 & 0 & b_{4,5} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

(b)

$$\begin{array}{lccccccc} & \text{Prim} & \text{Rev} & \text{Prim} & \text{Rev} & \text{Rev} & \\ \text{Prim} & 0 & b_{1,2} & 0 & b_{1,4} & b_{1,5} & \cdots \\ \text{Rev} & 0 & 0 & 0 & b_{2,4} & b_{2,5} & \cdots \\ \text{Prim} & 0 & 0 & 0 & b_{3,4} & b_{3,5} & \cdots \\ \text{Rev} & 0 & 0 & 0 & 0 & b_{4,5} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

(c)

$$\begin{array}{lccccccc} & \text{Prim} & \text{Rev} & \text{Rev} & \text{Prim} & \text{Rev} & \\ \text{Prim} & 0 & b_{1,2} & b_{1,3} & 0 & b_{1,5} & \cdots \\ \text{Rev} & 0 & 0 & b_{2,3} & 0 & b_{2,5} & \cdots \\ \text{Rev} & 0 & 0 & 0 & 0 & b_{3,5} & \cdots \\ \text{Prim} & 0 & 0 & 0 & 0 & b_{4,5} & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

(d)

$$\begin{array}{lccccccc} & \text{Prim} & \text{Rev} & \text{Rev} & \text{Rev} & \text{Prim} & \\ \text{Prim} & 0 & b_{1,2} & b_{1,3} & b_{1,4} & 0 & \cdots \\ \text{Rev} & 0 & 0 & b_{2,3} & b_{2,4} & 0 & \cdots \\ \text{Rev} & 0 & 0 & 0 & b_{3,4} & 0 & \cdots \\ \text{Rev} & 0 & 0 & 0 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \end{array}$$

Table 2.2. Structure of the Probability Matrix with second primary moving in time. (a) Primaries are performed in a row, all revisions come later. (b);(c);(d) Second primary is performed after one or more revisions linked to the first primary

(a)

$$\begin{array}{lcccc} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim} & c_{1,1} & c_{1,2} & c_{1,3} & \cdots \\ \text{Prim} & c_{2,1} & c_{2,2} & c_{2,3} & \cdots \\ \text{Rev} & 0 & c_{3,2} & c_{3,3} & \cdots \\ \text{Rev} & 0 & 0 & c_{4,3} & \cdots \\ & \vdots & \vdots & \vdots & \end{array}$$

(b)

$$\begin{array}{lcccc} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim} & c_{1,1} & c_{1,2} & c_{1,3} & \cdots \\ \text{Rev} & 0 & c_{2,2} & c_{2,3} & \cdots \\ \text{Prim} & 0 & c_{3,2} & c_{3,3} & \cdots \\ \text{Rev} & 0 & 0 & c_{4,3} & \cdots \\ & \vdots & \vdots & \vdots & \end{array}$$

(c)

$$\begin{array}{lcccc} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim} & c_{1,1} & c_{1,2} & c_{1,3} & \cdots \\ \text{Rev} & 0 & c_{2,2} & c_{2,3} & \cdots \\ \text{Rev} & 0 & 0 & c_{3,3} & \cdots \\ \text{Prim} & 0 & 0 & c_{4,3} & \cdots \\ & \vdots & \vdots & \vdots & \end{array}$$

(d)

$$\begin{array}{lcccc} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim} & c_{1,1} & c_{1,2} & c_{1,3} & \cdots \\ \text{Rev} & 0 & c_{2,2} & c_{2,3} & \cdots \\ \text{Rev} & 0 & 0 & c_{3,3} & \cdots \\ \text{Rev} & 0 & 0 & 0 & \cdots \\ & \vdots & \vdots & \vdots & \end{array}$$

Table 2.3. Structure of the Operational Probability Matrix with second primary moving in time. (a) Primaries are performed in a row, all revisions come later. (b);(c);(d) Second primary is performed after one or more revisions linked to the first primary

Table 2.3 shows the Operational Probability Matrices corresponding to the Probability Matrices in Table 2.2.

Given $\mathcal{P}_n$, the set of all possible tuples of n elements of size n, let define $\boldsymbol{s}_n \in \mathcal{P}_n$ as a generic vector of length n denoting one of the tuples and $(\boldsymbol{s}_n)_j \in \{1, \dots, n\}$ the j^{th} element of $\boldsymbol{s}_n$.

Given the Operational Probability Matrix C, then $\left\{c_{(\boldsymbol{s}_n)_j, j}\right\}_{j=1}^{n}$ defines a generic possible set available when picking one element from each column of C, namely a path from occasion 1 (column 1) to occasion n (column n) within the Operational Probability Matrix.

The probability for a specific path $\boldsymbol{s}_n$ to occur is given by the following:

$$P(\boldsymbol{s}_n) = \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} \tag{2.1}$$

Theorem B.1 in Appendix B guarantees that the sum of the probabilities carried out by (2.1) is equal to 1, namely that:

$$\sum_{\boldsymbol{s}_n \in \mathcal{P}_n} P(\boldsymbol{s}_n) = P\left(\bigcup_{\boldsymbol{s}_n \in \mathcal{P}_n} \boldsymbol{s}_n\right) = 1.$$

Clinical paths described in Definition 2.1 define a subset of the tuples in $\mathcal{P}_n$, namely $A_n \subset \mathcal{P}_n$. Indeed, clinical paths $\boldsymbol{a}_n$ are particular sets that ensure that the object $\left\{c_{(\boldsymbol{a}_n)_j, j}\right\}_{j=1}^{n}$ is built by picking up one element from each column of C, where every element belongs to a different row of the matrix and with row index always smaller than column index plus one. This is guaranteed by properties *i*), *ii*) and *iii*) in Definition 2.1. A_n induces a partition of the set $\mathcal{P}_n$, separating paths fulfilling Definition 2.1 and paths which do not.

Then, given a path $\boldsymbol{s}_n$, the following holds:

$$\begin{aligned}
P(\boldsymbol{s}_n \in \mathcal{P}_n) &= P(\boldsymbol{s}_n \in A_n \cup \boldsymbol{s}_n \in A_n^c) = \\
&= P(\boldsymbol{s}_n \in A_n) + P(\boldsymbol{s}_n \in A_n^c) = \\
&= \sum_{\boldsymbol{s}_n \in A_n} P(\boldsymbol{s}_n) + \sum_{\boldsymbol{s}_n \in A_n^c} P(\boldsymbol{s}_n) = \\
&= \sum_{\boldsymbol{s}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} + \sum_{\boldsymbol{s}_n \in A_n^c} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} = \\
&= \sum_{\boldsymbol{s}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} + \left(1 - \sum_{\boldsymbol{s}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j}\right) = 1.
\end{aligned} \tag{2.2}$$

According to (2.2), the following is straightforward to be proven:

$$\mathbb{1}\left(\boldsymbol{s}_n \in A_n\right) \sim \mathcal{B}er\left(\sum_{\boldsymbol{s}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j,j}\right). \tag{2.3}$$

From a clinical point of view, the set $\left\{c_{(\boldsymbol{a}_n)_j,j}\right\}_{j=1}^{n}$ is composed by the probabilities for the j^{th} revision to be corresponding to the $(\boldsymbol{a}_n)_j^{th}$ intervention and the Definition 2.1 of $\boldsymbol{a}_n$ ensures that j^{th} revision was performed after $(\boldsymbol{a}_n)_j^{th}$ intervention and that no revision can be considered as corresponding to multiple previous interventions.

The objects in (2.1) and (2.3) are needed to find out the probability of a specific clinical path to occur.

Given the following notation:

$\boldsymbol{a}_n = \boldsymbol{s}_n \ s.t. \ \boldsymbol{s}_n \in A_n$;

$(\boldsymbol{a}_n)_j$ j^{th} element of $\boldsymbol{a}_n$;

$A_n \subset \mathcal{P}_n$ the set defined in 2.1;

then, the probability of a clinical path $\boldsymbol{a}_n$ to occur can be written as follows:

$$P(\boldsymbol{a}_n) = P(\boldsymbol{s}_n | \boldsymbol{s}_n \in A_n, \boldsymbol{s}_n = \boldsymbol{a}_n) = \frac{P(\boldsymbol{s}_n \cap \boldsymbol{s}_n \in A_n)}{P(\boldsymbol{s}_n \in A_n)} = \frac{\displaystyle\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\displaystyle\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}. \tag{2.4}$$

The denominator in (2.4) is the probability of selecting a succession of elements $\boldsymbol{s}_n$ that fulfills $i)$, $ii)$, $iii)$ from Definition 2.1, according to probabilities in matrix C. This quantity plays the role of normalizing constant, as the possible realizations of a clinical path are by definition incompatible events. Then it ensures that:

$$\sum_{\boldsymbol{a}_n \in A_n} P(\boldsymbol{a}_n) = \sum_{\boldsymbol{a}_n \in A_n} \frac{\displaystyle\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\displaystyle\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} = 1.$$

2.1.2 The probability of a clinical path: the case with only one observed primary

When only one primary intervention is observed, the derivation of the actual clinical path for a generic patient seems to be straightforward. In that case, by using the

Operational Probability Matrix in Definition 2.3, the only clinical path $\boldsymbol{a}_n \in A_n$ for which $P(\boldsymbol{a}_n) \neq 0$ is $\boldsymbol{a}_n = (1, \dots, n)$ and it is such that

$$P(\boldsymbol{a}_n = (1, \dots, n)) = \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j, j} = \prod_{\substack{j=1 \\ i=j}}^{n} c_{i,j} = 1$$

In fact, every patient can undergo at most two primaries for the same kind of joint. While this is an advantage when dealing with cases where two primaries are actually observed, as the actual number of primaries is known, it can be a drawback when there is only one. In such a case, the existence of another primary, not observed because performed before the beginninig of the data collection time window, must be taken into account.

To handle this, the following object has to be introduced:

Definition 2.4 (Augmented Operational Probability Matrix). *Given the following notation:*

n the number of observed revisions;

k the number of observed primaries;

$i, j = 1, \dots (n + k)$ the ordered occasions at which interventions are performed;

t_i the time at which the i^{th} intervention is performed;

t_0 the time of the first data observation

I_i the i^{th} observed intervention;

$f(t_i | t_0)$ a generic probability measure that returns the probability that the time elapsed between the i^{th} revision and the corresponding (not observed) primary is bigger than $t_i - t_0$;

The Augmented Operational Probability Matrix $C^* \in \mathcal{M}_{(n+k) \times (n)}$ is derived from the Operational Probability Matrix by adding a row $c_{0,\cdot}$ where the j^{th} term is equal to the probability of the j^{th} revision to be revision of a primary occurred before the first time of observation, that is $c_{0,j} = f(t_j | t_0)$. Normalizing the columns, the propriety of left stochasticity holds.

Table 2.4 shows the generic structure of an Augmented Operational Probability Matrix when compared to the corresponding Operational Probability Matrix.

$$\underset{C}{\begin{array}{ccccc} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim} & c_{1,1} & c_{1,2} & c_{1,3} & \cdots \\ \text{Rev} & 0 & c_{2,2} & c_{2,3} & \cdots \\ \text{Rev} & 0 & 0 & c_{3,3} & \cdots \\ & \vdots & \vdots & \vdots & \end{array}} \longrightarrow \underset{C^*}{\begin{array}{ccccc} & \text{Rev} & \text{Rev} & \text{Rev} & \\ \text{Prim}^* & c^*_{0,1} & c^*_{0,2} & c^*_{0,3} & \cdots \\ \text{Prim} & c^*_{1,1} & c^*_{1,2} & c^*_{1,3} & \cdots \\ \text{Prim} & 0 & c^*_{2,2} & c^*_{2,3} & \cdots \\ \text{Prim} & 0 & 0 & c^*_{3,3} & \cdots \\ & \vdots & \vdots & \vdots & \end{array}}$$

Table 2.4. Deriving the generic structure of an Augmented Operational Probability Matrix (C^*) starting from the Operational Probability Matrix (C) in cases where only one primary has been observed.

Taking into account the existence of an unobserved primary intervention is crucial to provide a proper estimate for the implant survivorship; further investigation about those cases when the Augmented Operational Probability Matrix would be suitable follows.

Bilateral arthroplasty often occurs because of degenerative diseases hitting multiple joints. Usually, the spreading of the symptoms requires few years or even months [37, 21, 69], depending on the kind of disease and when it affects the same joint on both sides, this can lead to bilateral arthroplasty.

Since the time necessary for the disease to hit multiple joints may be quite short, an interval of few years can be expected between two primary interventions in case of bilateral arthroplasty. Figure 2.2 shows the distribution of the time elapsed between two primary interventions for patients with bilateral hip prostheses according to registry data from Autonomous Provinces of Trento and Bolzano; data refers to the period 2010 to 2018 (108 months). The median value is equal to 19 months, while the 75% and 95% percentiles are 40 and 72 months respectively. According to these results, the possibility that a primary intervention could have occurred before the starting time of the data collection should be considered. In this scenario, it is assumed that the probability of such an event depends on the time elapsed between the only observed primary and the study start. When, for instance, dealing with data observed over twenty years, if a patient undergoes a first and unique observed primary at the 10^{th} year, observing an arthroplasty intervention for the same kind of joint, but on the other side, performed more than ten years earlier, would be infrequent. To consider this feature in the model, an auxiliary variable has to be introduced.

Given the following notation:

K the number of primaries to take into account;

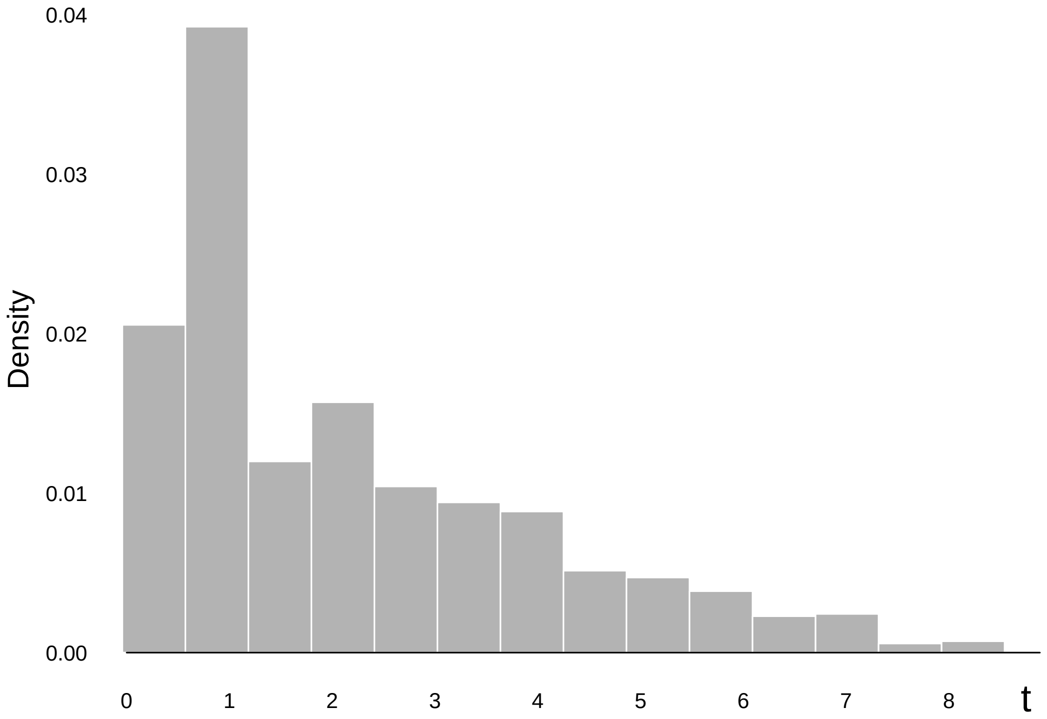


Figure 2.2. Distribution of the time elapsed between two primary interventions for patients with bilateral hip prostheses according to registry data from Autonomous Provinces of Trento and Bolzano from 2010 to 2018, considering only the records of patients for which two primaries were observed in the time window.

k_{oss} is the number of observed primaries;

t the time at which the only primary is observed

t_0 the starting observation time;

$X_{t,t_0} \sim \mathcal{B}in(1, q_{t,t_0})$;

$f(\cdot)$ the probability density describing the time elapsing between two primaries;

then:

$$K = 2 \times \mathbb{1}(k_{oss} = 2) + (1 + X_{t,t_0}) \times \mathbb{1}(k_{oss} = 1), \tag{2.5}$$

where

$$q_{t,t_0} = \int_{t-t_0}^{\infty} f(dx). \tag{2.6}$$

The variable in (2.5) takes into account the distance between the time at which the primary is observed and the starting time of the data observation window. It allows to consider either only the observed primary intervention, or the existence of an unobserved primary performed on the other side before the study started, depending on the time elapsing between the two interventions. The probability of considering an unobserved primary decreases as the distance between the observation time t of the actual intervention and the starting time of data collection t_0 increases and this is consistent with the knowledge obtained from literature and data exploration.

While the number of observed primaries is $k_{oss} = 1$, exploiting the Definitions 2.3, 2.4 and the random variable in (2.5), the probability for a clinical path $\boldsymbol{a}_n \in A_n$ to occur is given by the following:

$$\begin{aligned} P(\boldsymbol{a}_n | k_{oss} = 1) = & \sum_{k=1}^{2} P(\boldsymbol{a}_n \cap K = k | k_{oss} = 1) = \\ = & \sum_{k=1}^{2} P(\boldsymbol{a}_n | K = k, k_{oss} = 1) P(K = k) = \\ = & \ P(\boldsymbol{a}_n | K = 1, k_{oss} = 1) P(K = 1) + \\ & + P(\boldsymbol{a}_n | K = 2, k_{oss} = 1) P(K = 2) = \\ = & \ (1 - q_{t,t_0}) \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j, j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j, j}} + q_{t,t_0} \frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j, j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j, j}} . \end{aligned} \tag{2.7}$$

2.1.3 Deriving the complete Model

The tools introduced so far are the basic ingredients that are necessary to formalize a comprehensive model to link any primary with its most likely corresponding revision, in order to estimate the survival of the implanted device for each observed primary intervention in the observed data.

Considering the objects introduced in (2.4) and (2.7), the generic probability of a clinical path to occur for a patient may be provided, whatever the number of observed primaries:

$$\begin{aligned} P(\boldsymbol{a}_n) &= P(\boldsymbol{a}_n|k_{oss}=2)\mathbb{1}_{k_{oss}=2} + P(\boldsymbol{a}_n|k_{oss}=1)\mathbb{1}_{k_{oss}=1} = \\ &= (k_{oss}-1)\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} + \\ &+ (1-(k_{oss}-1))\left[(1-q_{t,t_0})\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} + q_{t,t_0}\frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}\right]. \end{aligned} \tag{2.8}$$

By exploiting (2.8), deriving the probability of the j^{th} revision to represent the failure of a given primary is possible.

Denoting by $A_n\left((\boldsymbol{a}_n)_j = i\right)$ as the set of clinical paths $\boldsymbol{a}_n$ such that the j^{th} observed revision for the patient is revision for the i^{th} primary intervention, even though it is unobserved as it has been performed before the first data observation time; and by $Y_{i,j} = \left(j^{th} \text{ revision is the failure of } i^{th} \text{ primary}\right)$ as the corresponding event, then

$$Y_{i,j} \sim \mathcal{B}er(p_{i,j}) \tag{2.9}$$

where , by using the probability $P(\boldsymbol{a}_n)$ in (2.8)

$$p_{i,j} = \sum_{\boldsymbol{a}_n\in A_n((\boldsymbol{a}_n)_j=i)} P(\boldsymbol{a}_n) \tag{2.10}$$

The probability in (2.8) has a useful property of convergence. Theorem B.2 in Appendix B shows that

$$\lim_{t_0\to-\infty} P(\boldsymbol{a}_n) = \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}. \tag{2.11}$$

This convergence propriety has a very clear and useful clinical meaning. If data could be observed backward to an infinite time, that is, if data could be collected back in time up to the very first intervention performed, the information about the number of primaries actually performed for a certain patient would be ensured and the possibility to have a previous unobserved primary should not be considered, as data would be collected from the very first intervention.

2.2 The use of prior information in the Probability Matrix

So far, the elements in the Probability Matrix and, consequently, the ones in the Operational Probability Matrix were assumed to be known. In fact, the quantity $f(t_i; t_j)$ in Definition 2.2 is not given and has to be carried out. The proposed approach takes into account the prior information coming from literature, mainly foreign registries, about failure time associated to causes of revision.

2.2.1 Causes of revision and failure times

The relation between causes of revision of joint prostheses and associated failure times has been widely explored in literature. The most common causes of failure are aseptic loosening, infection, dislocation or instability, implant or bone wear and prosthetic or peri-prosthetic fracture [49, 45]. Usually, all of them have a well known pattern with respect to failure times. For instance, infection and dislocation are often causes of early revision, that is, they usually occur within two years from the corresponding primary; on the other hand, implant wear, as one could expect, generally occurs after several years, as a device, mostly made on metal, ceramic or polyethilene, needs a lot of time to wear out [45, 79, 25].

This kind of information can be used to provide linkage between primaries and revisions. If, for example, two primaries and a subsequent revision due to infection are observed, it would be more probable that the failure is related to the device implanted during the primary closer in time to the revision intervention. On the other hand, if a failure due to implant wear is observed, the related primary is likely to be found the farthest in time from the revision intervention.

The goal is then to formalize this knowledge into a probabilistic approach to elicit the density $f(t_i; t_j)$ in Definition 2.2. Since causes of revision are known in HDRs and coded according to ICD9-CM, an estimate of $f(t_i; t_j)$ based on causes of revision is feasible and it could be used to provide the main ingredients for the model exposed in this chapter to work properly.

When a subset of national data provides the side variable (for instance, some regional registries may properly collect data), this information can be used to estimate failure times by cause of revision for the whole country and obtain an estimate for $f(t_i; t_j)$ to fill in matrix B.

However, the subset on which failure times by cause of revision are known may be poorly informative. Information provided may be misleading or not coherent with the knowledge coming from the literature. In such cases, prior information about failure times by causes of revision, elicited from foreign registries, can be used to update the estimates coming from subsets of national data in the spirit of the Bayesian approach.

Indeed, every year, several registries publish data about the revision rate, expressed via KM estimates, by cause of revision. Recovering raw data from these curves allows to estimate the failure time probability of a device by cause of revision.

2.2.2 Elicitation of the prior information

When the survival function estimates are available at a grid of event times, solving the following system is necessary to recover raw data:

$$\begin{cases} S_1 = & \dfrac{l_1 - c_1 - d_1}{l_1 - c_1} \\ S_2 = & S_1 \dfrac{l_2 - c_2 - d_2}{l_2 - c_2} \\ \quad \vdots & \\ S_i = & S_{i-1} \dfrac{l_i - c_i - d_i}{l_i - c_i} \\ \quad \vdots & \end{cases} \tag{2.12}$$

where S_i is the survival function estimate at time t_i, l_i is the number of surviving devices, c_i is the number of censored items at the begin of each interval and d_i is the number of failures. System (2.12) has the same structure of the system leading to the KM estimate of the survival curve [48]; the crucial difference in this case is that S_i's are known and d_i's have to be estimated. The solution is not unique without constraints. Imposing constraints based on the maximum number of devices implanted at every time t, leads to fixing c_i's and l_i's; in this case, the solution is unique and raw data can be recovered.

Figure 2.3 shows the recovering of raw data for failure time T from the KM estimate of survival curves carried out from three different simulated scenarios:

- $T \sim Exp(200)$;
- $T = wX_1 + (1 - w)X_2$ where $X_1 \sim Exp(10)$, $X_2 \sim Weib(20, 100)$ and $w \sim Ber(0.3)$;

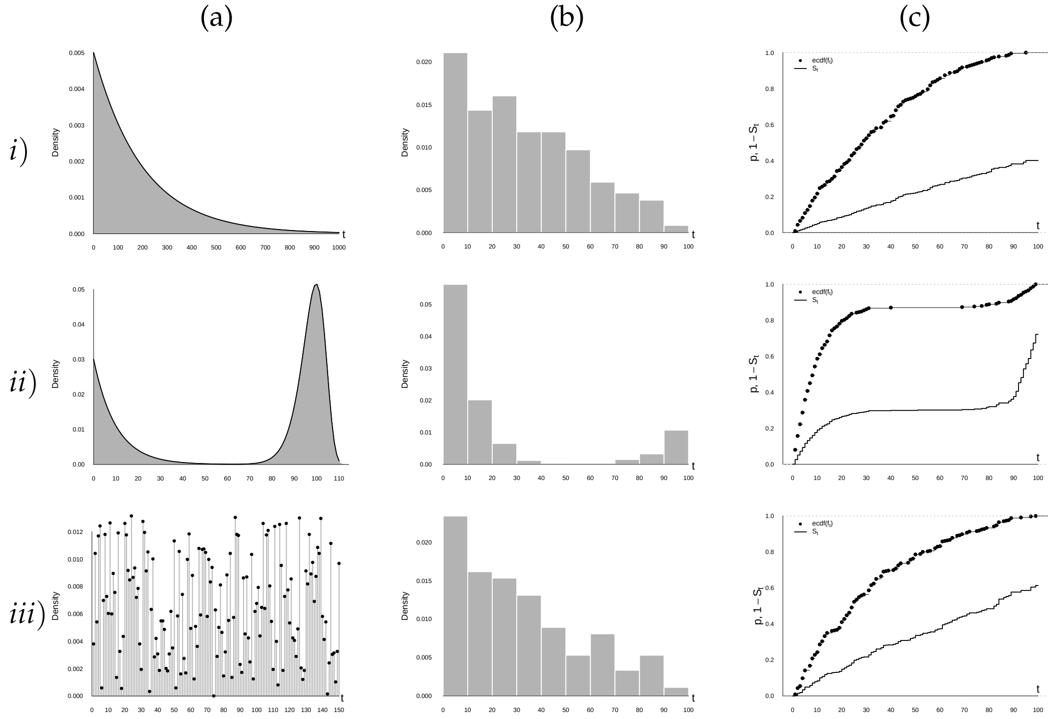


Figure 2.3. Simulation of recovering raw data from KM survival curves. (a):Theoretical density for failure time; (b): Observed relative frequencies given censored data; (c): Observed KM curve and relative ecdf of recovered raw data. *i)* $T \sim Exp(200)$ (Exponential); *ii)* $T = wX_1 + (1-w)X_2$ where $X_1 \sim Exp(10)$, $X_2 \sim Weib(20, 100)$ and $w \sim Ber(0.3)$ (Mixture); *iii)* $T \sim F$ with F non parametric and mass point randomly generated for each t_i, $i = 1, \ldots 150$ and normalized (Non parametric).

- $T \sim F$ with F non parametric and mass point randomly generated for each t_i, $i = 1, \ldots 150$ and normalized.

Such scenarios were chosen to test the recovering method in different contexts, that is for data generated from parametric, mixture and non parametric models and in all of the cases, raw data about the number of failures for each time was recovered.

Unfortunately, the estimate of the survival curve at each time is not always available and all the information comes from the figures published in reports or articles. In such a case, several approaches can be used, based on reading in the coordinates from a raster image, or recovering information from the various electronic formats in which such curves are published (.jpg, .png, .pdf etc.) [31, 61]. Codes in different programming languages can be found in literature and a specific software named DigitizeIt has been implemented for this purpose [10].

Once raw data is recovered, the failure time density can be estimated. Either parametric or non parametric approaches can be used; indeed the goal is to provide an accurate and suitable density estimation for the failure times, conditional on the specific cause of revision, for all causes of revision.

The estimated conditional failure time density by cause of revision can be used as prior information to derive estimates for $f(t_i, t_j)$ in Definition 2.2, given the recorded cause of revision for the j^{th} revision intervention.

So, the estimated densities are necessary to properly fill the Probability Matrix B and make the proposed model work. Eliciting this prior information from foreign registries is necessary when estimates of failure time by cause of revision are totally or partially unavailable for data under analysis because of missing knowledge on the operated side.

2.3 Formalizing the model

By exploiting (2.5), (2.6), (2.8) and (2.9), the model can be formalized as a Bayesian hierarchical model with the following structure:

$$\begin{aligned}
&Y_{i,j} \sim \mathcal{B}er(p_{i,j}) \\
&p_{i,j} = \sum_{a_n \in A_n\left((\boldsymbol{a}_n)_j = i\right)} P(\boldsymbol{a}_n) \\
&P(\boldsymbol{a}_n) = (k_{oss} - 1) \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} + \\
&\quad + (1 - (k_{oss} - 1)) \left[(K-1) \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} + (1 - (K-1)) \frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}} \right]. \\
&c_{\cdot,j} \sim F(\cdot | \text{ cause of revision for the } j^{th} \text{ revision intervention}) \\
&K = 2 \times \mathbb{1}(k_{oss} = 2) + (1 + X_{t,t_0}) \times \mathbb{1}(k_{oss} = 1) \\
&X_{t,t_0} \sim \mathcal{B}in(1, q_{t,t_0}) \\
&q_{t,t_0} = \int_{t-t_0}^{\infty} f(dx)
\end{aligned} \tag{2.13}$$

where $F(\cdot|$ cause of revision for the j^{th} revision intervention$)$ is derived from the estimates of the density for failure times conditional on cause of revision and $f(\cdot)$ is the estimate of the probability density for the time elapsing between two primaries.

Distribution $F(\cdot|$ cause of revision for the j^{th} revision intervention$)$ and $f(\cdot)$ are estimated externally and plugged into the model since, in the context under analysis, there is not a way to estimate them on the available data, as the operated side is missing. Such distributions play the role of power priors used to introduce historical (otherwise not available) information, in the spirit of the proposal of incorporating historical information in parameter estimate [39, 73, 38].

The main difference with the literature standard relies in the absence of a discount parameter, generally used to give a weight to the prior information to incorporate in the model, with respect to the information available via likelihood in classical frameworks. Indeed, in the currently proposed method, information on failure time conditional on causes of revision and interoperative elapsing time is unavailable and the related likelihood is impossible to estimate because of the missing information about the operated side. Then, the posterior distributions are proportional to the corresponding prior distributions, as the likelihood does not exist and the historical information is the only available one.

However, the efficacy of the proposed method should not be affected by the choice of a particular set of historical priors, as failure times by cause of revision are assumed to be consistent in literature among all registries worldwide. This is because of the clinical reasons behind the observed pattern of failure times by cause of revision and since results come from population studies, that are not likely affected by issues due to sample size determination or sampling strategies.

Explicit expressions of $F(\cdot|$ cause of revision for the j^{th} revision intervention$)$ and $f(\cdot)$ are not provided in this framework, as they depend on the chosen estimation mechanism. That is, they can be estimated in several ways, like in a fully non parametric way, as point mass for each considered time on a grid or by a parametric approach or any suitable distribution and a more generalized form may be more suitable in this formalization of the model.

Chapter 3

An application to simulated data

In this chapter, the performance of the proposed model will be assessed in different scenarios in a controlled simulation study. The principal measures to evaluate the model performance are the accuracy in the linkage between primary and revision interventions and accuracy in approximating the KM estimate of the lifetime distribution function derived by the adopted linkage.

For sake of simplicity, from now on, the expressions "survival curve" and "revision rate" will be used referring to the complement to one of the survival function, $1 - S(t)$.

The final aim is to evaluate whether the use of the proposed approach to link primaries and revisions allows for an accurate recovery of the revision rate estimated via the KM estimator when the operated side is available.

Labek et al. [54], in their systematic review, highlight how revision rate is one of the most important outcome measures of joint replacement surgery and suggest to evaluate it on a yearly basis to compare results from registries and clinical studies worldwide. According to this, the average distance between the KM estimate of the revision rate carried out after using the proposed record linkage method and the KM estimate of the revision rate obtained when the operated side is known is used to asses the performance of the proposed record linkage method. This allows to understand how effective the approach is and how much the results obtained after the record linkage are reliable.

The model performance was tested in four different simulated scenarios. Starting from a single dataset, generated according to predetermined rules, over a time window of twenty years, the model was tested on data from the last five years, the last ten years, the last fifteen years and the whole dataset. This approach allows to check whether the performance in correctly linking primaries and revision and in producing KM estimates of the survival curve depends on the length of the observation window.

3.1 Data generation process

Data was generated over 240 observation times t, that is, simulating an observation of twenty years on a monthly basis, according to the following steps, after fixing the generation seed.

1. For all times t, 100 primary interventions were generated. 24000 records were generated. n= 24000.

2. A unique pseudo-code was associated to each of these primaries, resulting in 24000 different "patients".

3. For 1 out of 4 of the patients, at each time t, a second primary intervention for the same patient was generated, with time elapsing between the two primaries following the distribution $\mathcal{E}xp(1/20)$. The choice of this distribution is due to its shape, that can fit with the observed relative frequencies for time elapsing between two primaries in real data (Figure 2.2). 6000 records were generated. n=30000.

4. A first revision for a primary is generated with probability 0.15 with cause of revision A, B, C, each with probability equal to $\frac{1}{3}$. The revision intervention is generated at T times from the corresponding primary. The conditional distribution is assumed to be one of the following (Figure 3.1):

   - $T|A \sim \mathcal{W}eib(2, 15)$;
   - $T|B \sim \mathcal{W}eib(10, 120)$;
   - $T|C \sim \mathcal{W}eib(20, 210)$.

   Those conditional distributions were chosen to simulate a dataset similar to a real one: A plays the role of a cause of early revision; B plays the role of a cause of mid-term revision; C plays the role of a cause of long-term revision. 4402 records were generated. n=34402.

5. Out of those generated revisions, 15% leads to re-revision with causes of revision A, B, C, each with probability equal to $\frac{1}{3}$, according to distributions in point 4. 682 records were generated. n=35084.

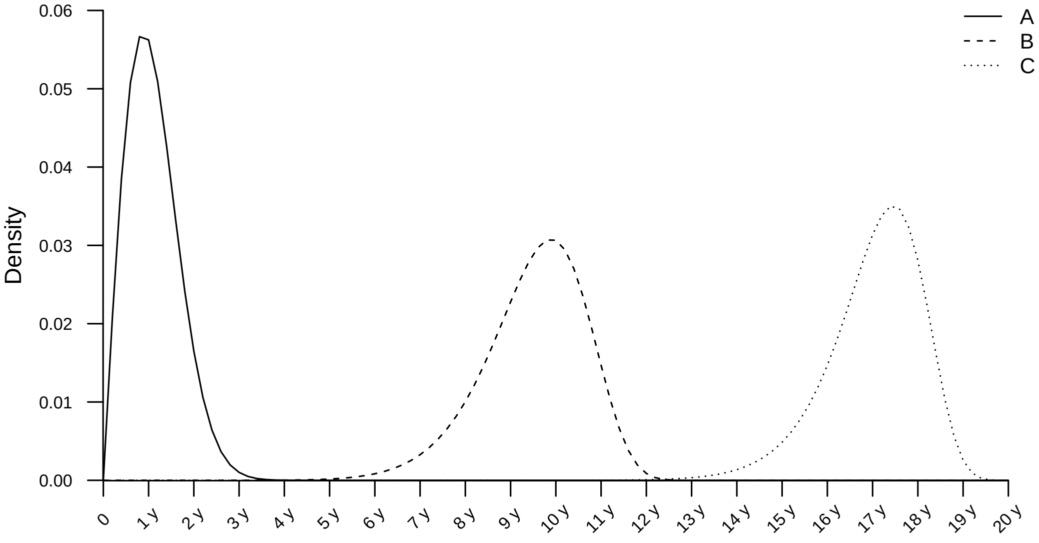


Figure 3.1. Conditional densities of time to revision by cause of revision. A: $\mathcal{W}eib(2, 15)$; B: $\mathcal{W}eib(10, 120)$; C: $\mathcal{W}eib(20, 210)$.

6. Step 5 is repeated for re-revisions in order to have up to 2 levels of re-revisions. 87 records were generated. n=35171.

7. Step 6 is repeated for re-revisions of second level in order to have up to 3 levels of re-revisions. 14 records were generated. n=35185

8. interventions generated with occurring time after $t = 240$ are dropped. 3216 records were dropped. n=31969

Simulated data are composed by 29490 primary interventions and 2479 revisions performed on 24000 fake "patients". At most five interventions were associated to the same pseudo-code, with at most two primaries and three revisions.

In this simulation study, the prior distributions for failure time conditional on causes of revision A, B and C are not elicited according to the technique described in section 2.2.2, but are assumed to be known according to step 4 in data generation process.

3.2 Model performance assessment on simulated data

The performance of the proposed linkage method was evaluated by checking the accuracy in correctly matching primary interventions and revisions and by assessing the ability of the model to lead to an estimated survival curve as close as possible to the one estimated when the operated side is known.

The following approaches were used to carry out the final estimate.

Median approach: the model was run 100 times over the generated dataset. Every possible clinical path was taken into account for each patient. The probability of every possible clinical path for the patient was computed according to (2.8) and exploiting distributions for failure times conditional on causes of revision A, B and C described at the step 4 of the data generation process (Figure 3.1) within the Operational Probability Matrix. A clinical path was picked according to these probabilities. A different KM survival curve estimate was carried out in each run of the model, depending on the imputed linkage between primary and revision interventions, resulting in 100 different KM estimated curves. The median value of the survival curve estimate at each time $t = 1, \dots, 240$ over 100 runs was used as final estimate for the survival curve.

Mode approach: the most probable clinical path was selected for each patient according to (2.8) and exploiting distributions for failure times conditional on causes of revision A, B and C described at the step 4 of the data generation process (Figure 3.1) within the Operational Probability Matrix. Once linkage between primaries and revision was given, the corresponding KM survival curve estimate was straightforward to carry out.

Accuracy in the median approach was computed as the average accuracy in linking primary and revision interventions over the 100 runs on the sample generated according to the procedure exposed in Subsection 3.1. Accuracy in the mode approach is computed as the percentage of primary and revision interventions correctly linked. Figure 3.2 shows the accuracy of the model in linking primary interventions to the corresponding revisions in 100 different runs. In all of the considered scenarios, the model seems to perform almost at the same level, producing a small range for accuracy values, never under 86.5% and with average accuracy equal to 91.5%, 91.5%, 89.8% and 88.1% in scenarios considering five, ten, fifteen and twenty years respectively. Accuracy resulting when using the mode approach is higher in all scenarios, equal to 92%, 92.7%, 91.7% and 90%, respectively. The slightly decreasing accuracy as the time window increases is due to the higher number of patients with two observed primary in a longer observation period.

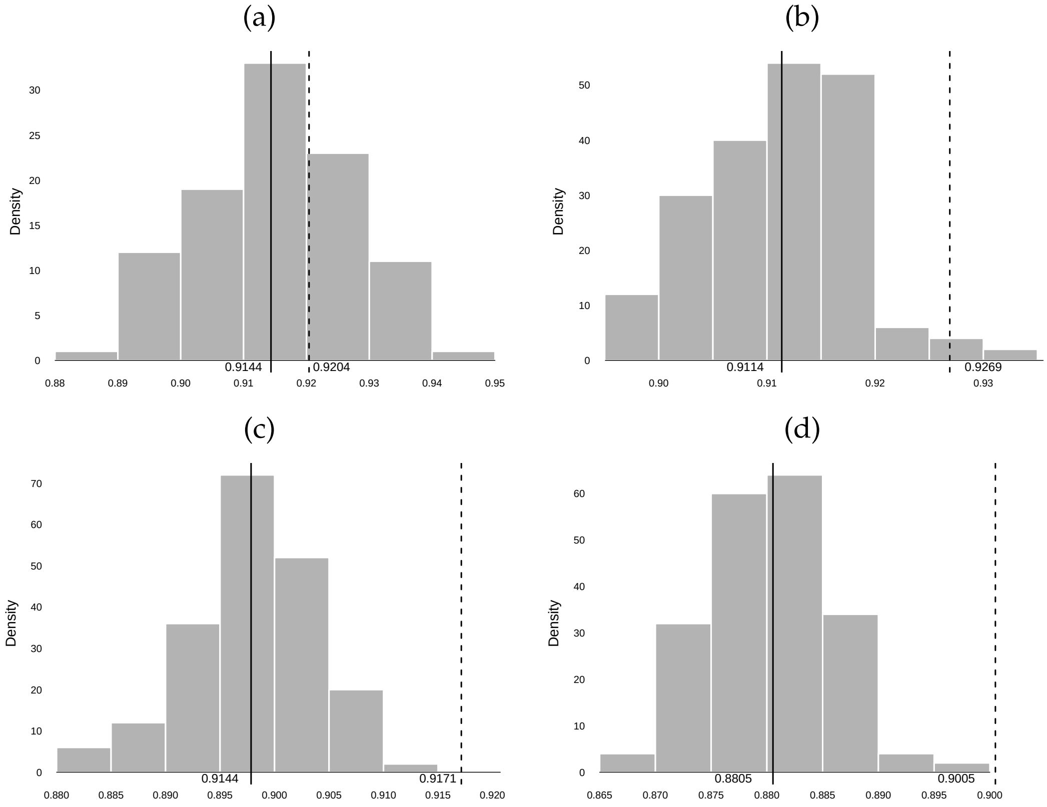


Figure 3.2. Simulated data: accuracy in linking primaries and revisions. Median approach over 100 runs. Black solid line: average accuracy via median approach. Black dashed line: accuracy via mode approach. (a): 5 years; (b): 10 years; (c): 15 years; (d): 20 years

The other considered measure is the ability of the model to recover the KM estimate of the revision rate when compared to the one estimated with known operated side. This assessment is performed both by comparing the survival curves, and checking the difference in revision rate on 5-years basis.

Figures 3.3 and 3.4 show the KM estimate of the of the revision rate obtained after linking primaries and revisions via median approach and mode approach, respectively. The curves are close to the ones estimated while the information on the operated side is available for both the proposed approach. In all the simulated scenarios, the proposed methods produce a slight over-approximation, even though the error is smaller as the number of years of observation increases. Tables 3.1 and 3.2 show the difference between the estimate carried out when the information about the operated side is available and the estimated revision rate obtained after using median and mode approaches, respectively, to link primaries and revision rate at five, ten, fifteen and twenty years in the four considered scenarios. By using both the proposed approaches, the estimates of the survival curve get closer to

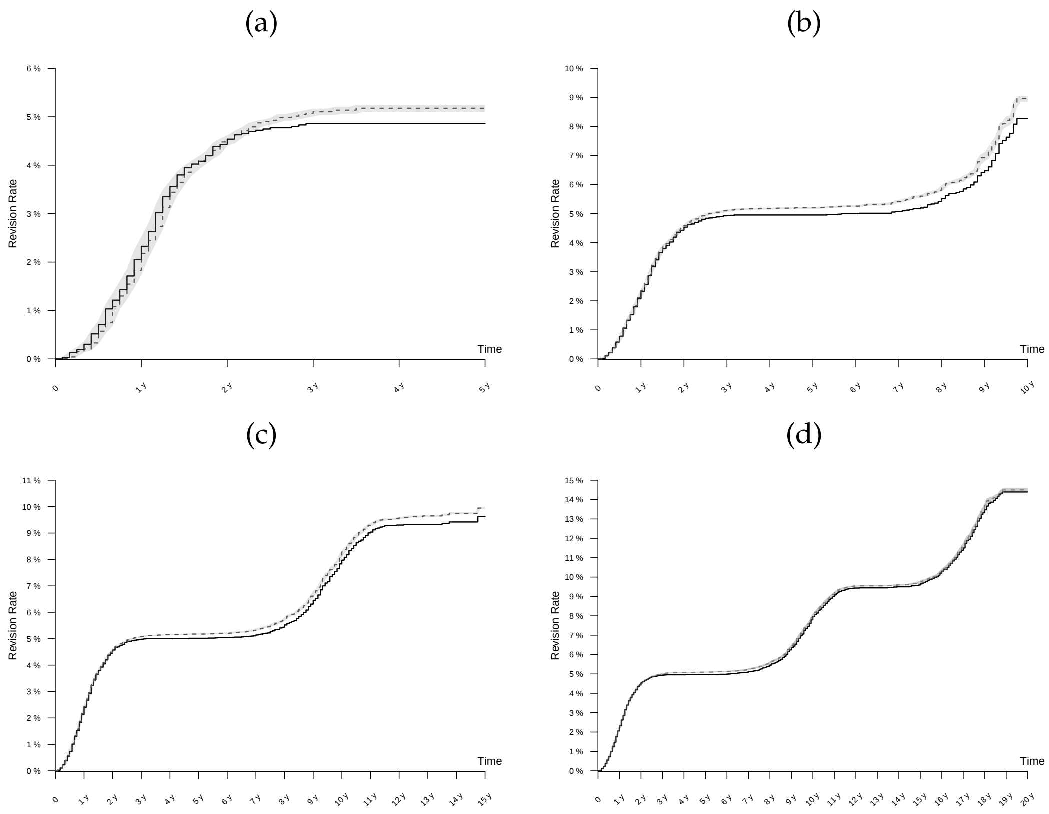


Figure 3.3. Simulated data: performance in the revision rate recovering Median approach. 100 runs of the model. Black solid line: KM estimate of the revision rate under known linkage between primaries and revisions. Grey dotted line: KM estimate of the revision rate when linkage between primaries and revisions is imputed by the proposed median approach. Grey surface: range of the KM estimates obtained in 100 runs. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario

	5-years ΔRR	10-years ΔRR	15-years ΔRR	20-years ΔRR
(a)	0.003145	-	-	-
(b)	0.002483	0.006869	-	-
(c)	0.001589	0.003091	0.003248	-
(d)	0.001263	0.001249	0.001095	0.001044

Table 3.1. Difference (Δ) of the estimated Revision Rate (RR) between the case when linkage between primaries and revisions is imputed by the model and the case when it is known. Median estimate over 100 runs. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

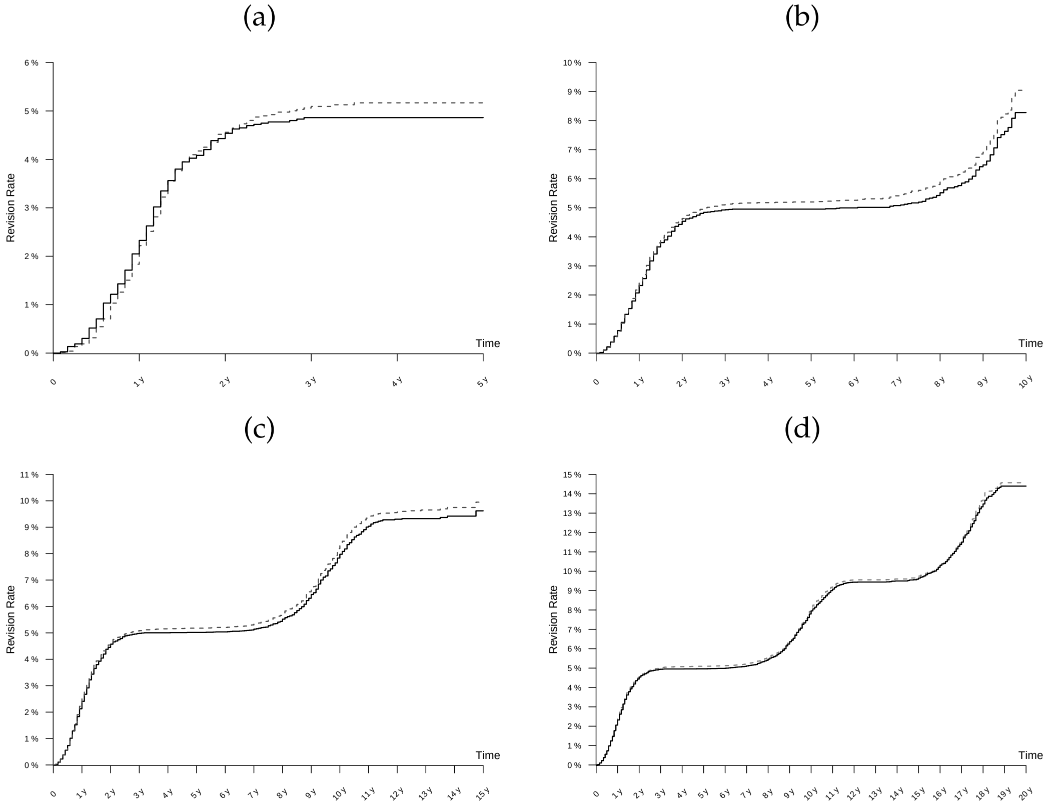


Figure 3.4. Simulated data: performance in the revision rate recovering. Mode approach. Black solid line: KM estimate of the revision rate under known linkage between primaries and revisions. Grey dotted line: KM estimate of the revision rate when linkage between primaries and revisions is imputed by the proposed mode approach. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario

	5-years ΔRR	10-years ΔRR	15-years ΔRR	20-years ΔRR
(a)	0.003053	-	-	-
(b)	0.002474	0.007657	-	-
(c)	0.001612	0.003865	0.003283	-
(d)	0.001279	0.001145	0.001034	0.001761

Table 3.2. Difference (Δ) of the estimated Revision Rate (RR) between the case when linkage between primaries and revisions is imputed by the model and the case when it is known. Median estimate over 100 runs. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

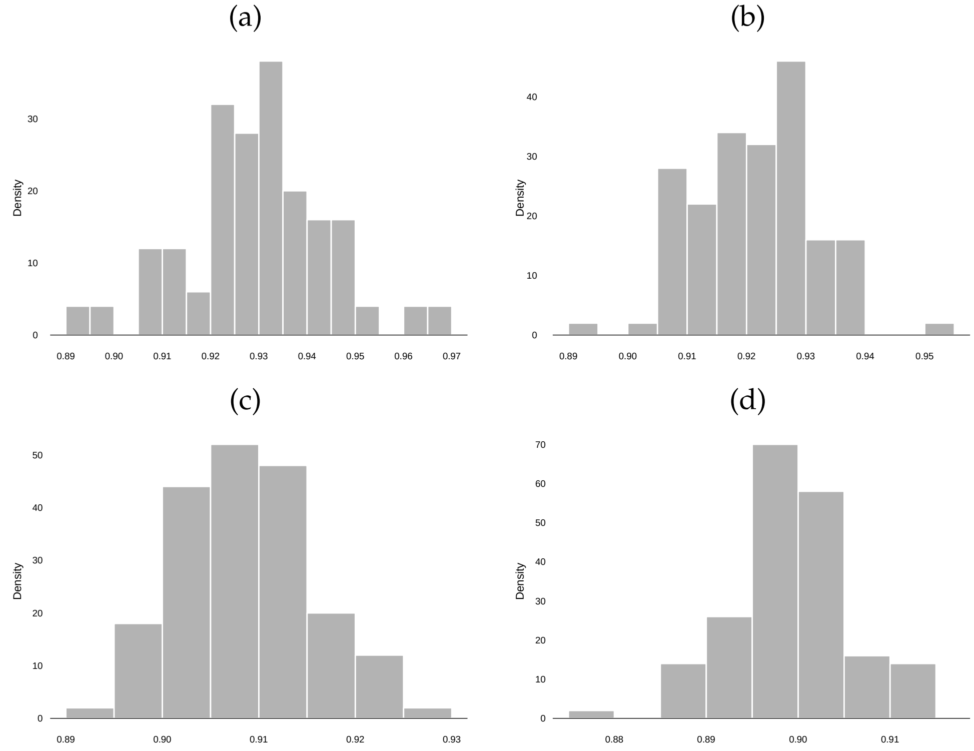


Figure 3.5. Simulated data: accuracy in linking primaries and revisions. Application of the proposed linkage method to 100 different simulated dataset. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

the one available when the linkage between primaries and revisions is known as the years of observation increase. This can be considered as a consequence of the property exposed in (2.11), since the choice of the clinical path for patients with only one observed primary becomes more clear and leads to a smaller error when the observed time period increases. The tight boundaries of the surface shown in Figure 3.3 and the small range of variation for the accuracy estimates reported in Figure 3.2 suggest that the proposed linkage approach works well in general empirical situations. Indeed, all the KM estimates of the revision rate lie in a narrow area.

The estimated curves in Figures 3.3 and 3.4 and the values in Tables 3.1 and 3.2 show that there is not a method that performs systematically better than the other between mode approach and median approach. For this reason, both methods should be taken into account when dealing with real data.

To assess the performance of the proposed linkage method against different data, the mode approach was ran to produce the linkage in 100 simulated datasets, generated according to the steps described in section 3.1, switching the generation seed. The same observation time scenarios were used.

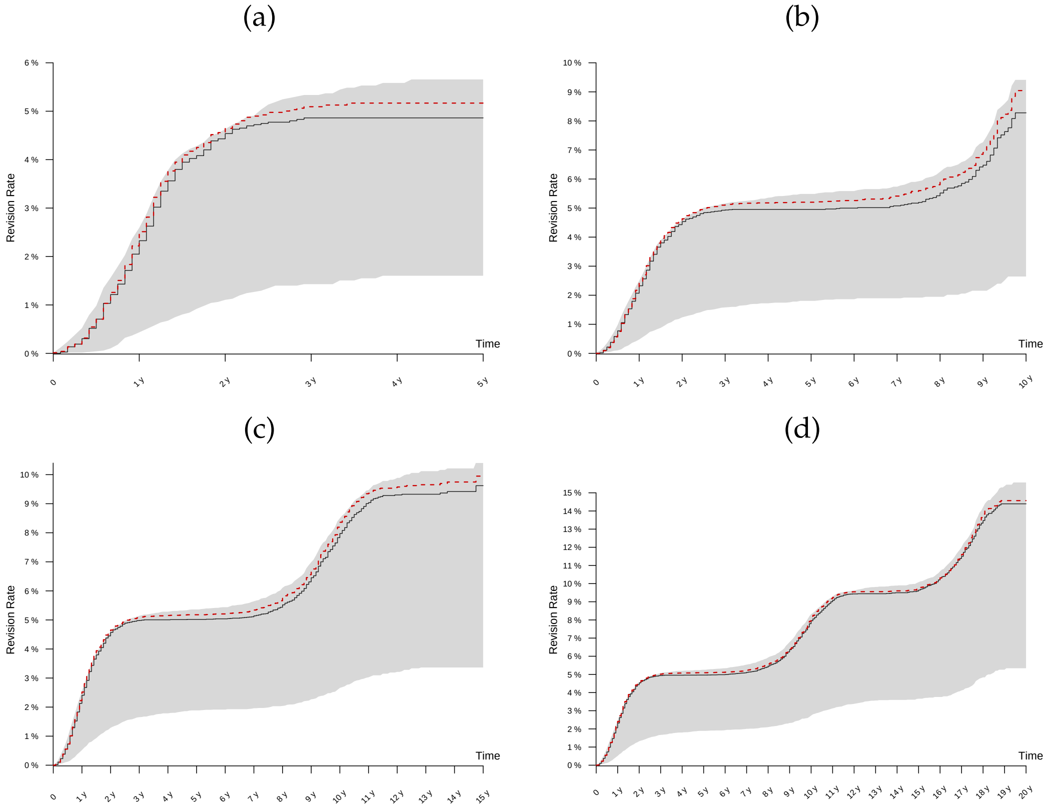


Figure 3.6. Simulated data: bias observed in the KM estimate of the revision rate in case of random linkage between primary interventions and revisions. Black solid line: KM estimate while side is known. Red dashed line: KM estimate performed after linking primaries and revision with the mode approach. Gray surface: region where KM estimates lie while naive or random linkage between primary interventions and revisions is performed. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

The accuracy of the linkage is reported in Figure 3.5. The average accuracy is higher than 89% against all the generated datasets. The KM estimate of the revision rate obtained after linking primaries and revisions with the mode approach is compared to the surface built up by considering all KM estimates obtained by naive linkage, carried out by the method exposed in Appendix A (Figure 3.6). This allows to better understand the improvement gained by using the proposed approach with respect to the use of a naive linkage. The resulting KM estimate of the survival curve is close to the one obtained when the operation side is known, even though it results in a slight over-estimate. Figure 3.7 shows the range of the absolute error in the KM estimate of the revision rate at each time t, when the linkage is performed via mode approach with respect to the KM estimate obtained with known operated side. The range is given by the maximum and the minim error observed in the

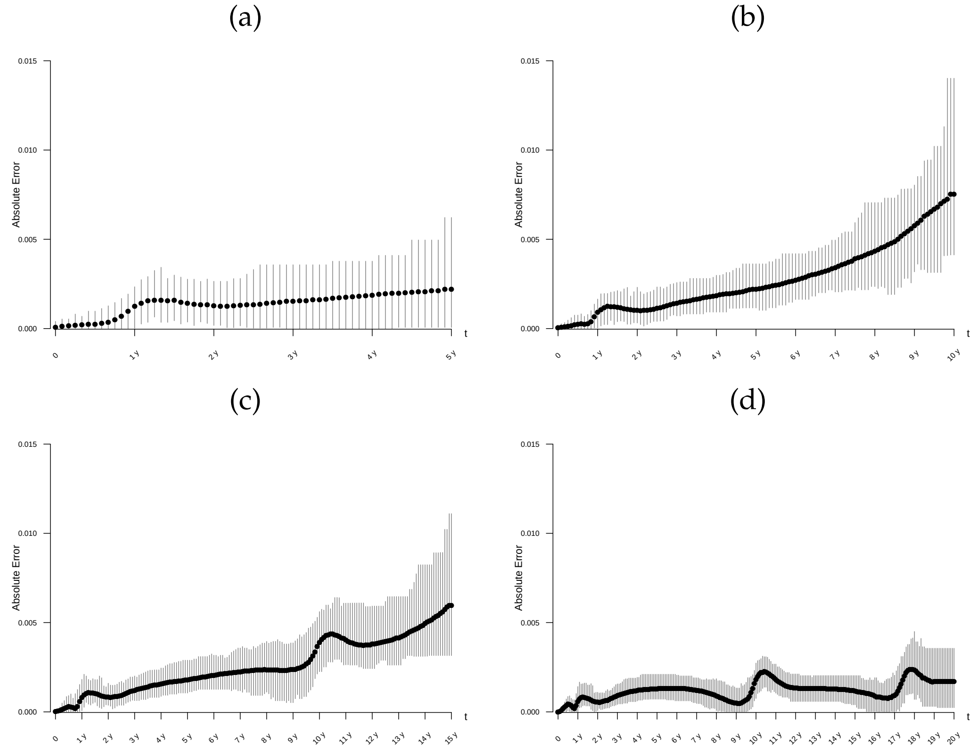


Figure 3.7. Simulated data: absolute error in the KM estimate of the revision rate at each observation time. Record linkage by proposed approach vs known linkage. Black dot: average error. Gray line: minimum-maximum error range (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

estimates against the 100 generated dataset. Such error is always smaller than 0.015 and the worst performance is obtained in the 10-years scenario. Also this metric shows that a wider observation time window allows for better performance in the KM estimate of the revision rate. Indeed, in the 20-years scenario, the absolute error is under 0.005 at each time t.

3.2.1 Methodological differences between Mode and Median approaches

The proposed linkage mechanism assigns probabilities to all the possible linkage scenarios for the records on a patient, finding which records are more probable to be about interventions performed on the same side and which ones are more probable to be about interventions performed on the other side. This is repeated for all patients.

In the Mode approach, only the most probable linkage scenario is considered for all patients (as it is usual in probabilistic record linkage methods); after the linkage between interventions is produced, only one survival curve is estimated, according to the time-to-event values derived from those linkages.

In the Median approach, a linkage scenario is drawn for each patient according to the probabilities assigned by the model; all linkage scenarios are taken into account and can be drawn with different probabilities; after the linkage is produced for each patient according to this rule, one survival curve is estimated, according to the time-to-event values derived from those linkages. This procedure is repeated many times (100 times in the current applications), and at every repetition, different linkage scenarios can be drawn for each patient according to the probabilities defined by the model. Once 100 different estimates of the survival curve are available, the median value of those curves at each observed time (depending on the chosen time grid: days, months, years etc.) is picked as estimate for the survival curve at that time.

By this method, accuracy can range by definition between 0 (in case all drawn linkages are erroneous for all patients) and 100% (in case all drawn linkages are correct for all patients). The aim of the Median approach is to check the behaviour of the model in case some actual realizations of linkages would go against theoretical probabilities. By this approach, it is possible to take into account such cases in survival analysis, with the drawback of a lower accuracy (on average) and a slightly larger (on average) distance from the survival curve estimate available when the linkage between interventions is known.

3.2.2 Comparison with standard probabilistic record linkage methods

As already exposed in section 2.1, probabilistic record linkage methods explored in literature [23, 18, 80, 92, 26, 66, 57, 35] are not suitable in this context. In fact, the proposed linkage methods are based on partially identifying variables or clinical features that allow to estimate the probability of a given set of observetions to be successfully linked.

In the problem under analysis, all candidate records for the linkage belong to the same patient; however, reported interventions can be performed on the two different sides, left and right. This implies that partially identifying variables and clinical features, on which those methods rely, can not be used, as they are identical in all candidate records. For this reason, such methods, when used in this context, produce a totally random linkage between records.

However, to provide a complete report of the performance of the proposed method, it is compared with the probabilistic record linkage based on partially identifying variables, in terms of accuracy (i.e. correctly matching primary interventions and revisions) and by assessing the differences between the resulting survival curves.

Figure 3.8 shows the improvement gained in terms of KM estimate when linking data by the proposed model with mode approach with respect to the performance of the probabilistic record linkage methods that rely on partially identifying variables. Table 3.3 shows the median absolute distance over the years in the four simulated scenarios and, with Table 3.2, this confirms a higher performance of the proposed linkage method. Moreover, the linkage accuracy scores in the four scenarios are 0.6827, 0.6653, 0.6627 and 0.6538 in the 5-years, 10-years, 15-years and 20-years scenarios, respectively, with a worse performance equal to 0.2 at least with respect to the case when linkage is performed by the proposed approach.

These results confirm that classical probabilistic record linkage approach is not suitable in the context of arthroplasty data coming from HDD. Those methods, relying only on partially identifying variables that are identical in all candidate records at each linkage step, result in a random linkage mechanism and need modifications to take into account the possible existence of unobserved records related to primary interventions, that have been performed before the beginning time of data collaction, that may be linked to the observed revisions. Therefore, the subsequent statistical analyses may be misleading and survival estimates unreliable, producing an increased risk for the safety of patients and to incorrect information on the efficacy of implantable devices.

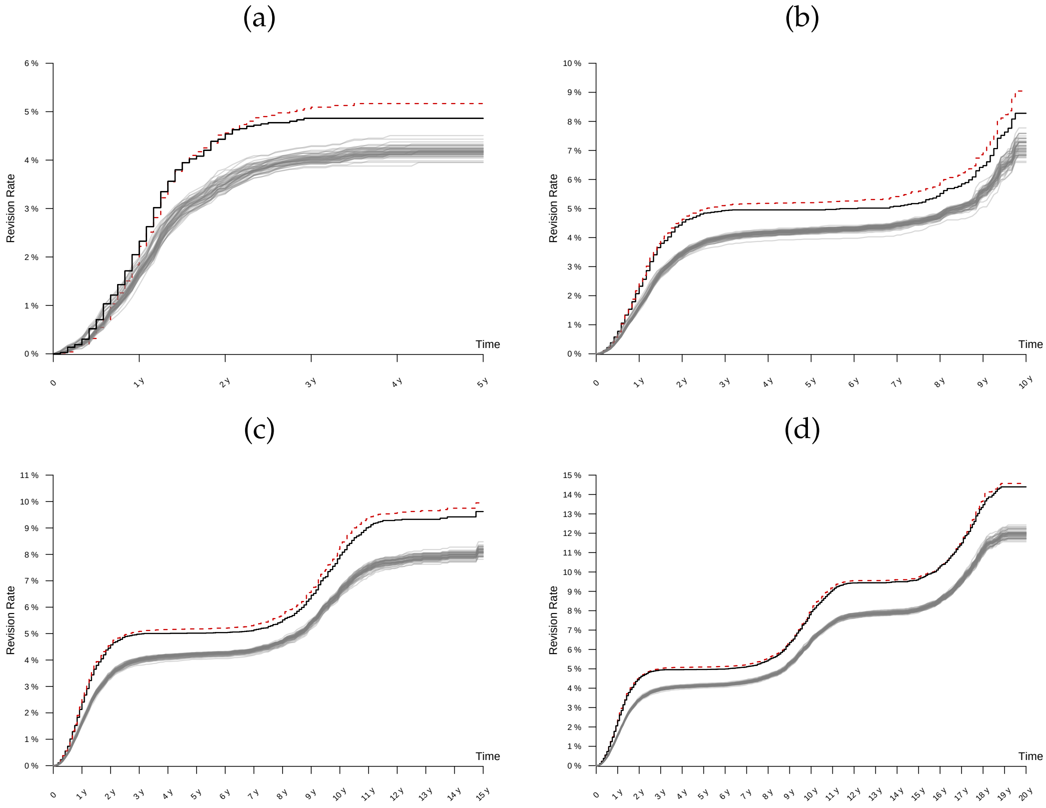


Figure 3.8. Simulated data: bias observed in the KM estimate of the revision rate obtained after 50 different realizations of random linkage between primaries and revisions obtained via probabilistic record linkage method based on partially identifying variables. Black solid line: KM estimate while side is known. Red dashed line: KM estimate performed after linking primaries and revisions with the mode approach. Gray solid lines: KM estimates performed after linking primaries and revisions with probabilistic record linkage based on partially identifying variables over 50 runs (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

	5-years ΔRR	10-years ΔRR	15-years ΔRR	20-years ΔRR
(a)	0.006833	-	-	-
(b)	0.007156	0.011925	-	-
(c)	0.008023	0.013805	0.015165	-
(d)	0.008304	0.015761	0.015954	0.024397

Table 3.3. Difference (Δ) of the estimated Revision Rate (RR) between the case when linkage between primaries and revisions is imputed by probabilistic record linkage model based on partially identifying variables and the case when it is known. Median estimate over 50 runs. (a): 5 years scenario; (b): 10 years scenario; (c): 15 years scenario; (d): 20 years scenario.

Chapter 4

Application to real registry data from the Autonomous Provinces of Trento and Bolzano

The registry data from the Autonomous Provinces of Trento and Bolzano, already introduced in Chapter 2 to build Figures 2.1 and 2.2, was used to assess the proposed linkage method performance in a real world context. Every record is composed by two parts: a set of variables coming from the HDRs; and a set of variables belonging to the Minimum Data Set. In this application only the variables from HDRs were considered, while all the others were dropped.

Type of intervention and causes of revision were assigned to each record according to the mapping from ICD9-CM codes to registry meaningful modalities. These are reported in Tables C.1 and C.2 (Appendix C).

In the simulation study, probability distributions for failure times by cause of revision were assumed to be known, while, in this real world application, densities were estimated by using prior information coming from the 2019 report of the Australian Orthopaedics Association [2].

4.1 Deriving probability densities for failure time by cause of revision

The probability densities for failure time by cause of revision were estimated by using raw data recovered from survival curves in the 2019 report of the Australian Orthopaedic Association National Joint Replacement Registry [2]. The Australian

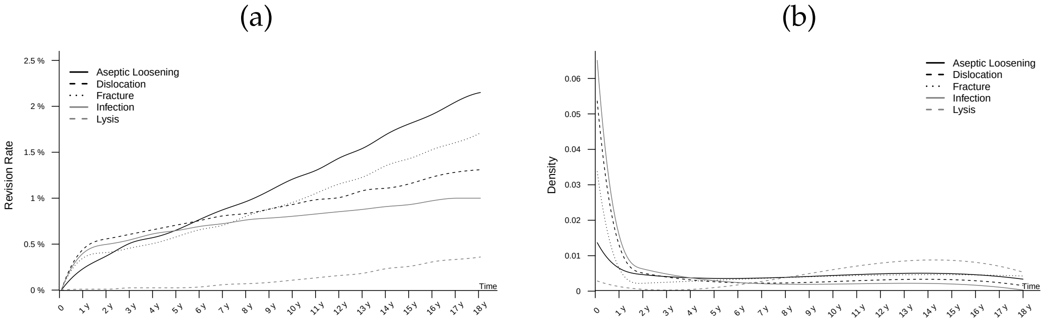


Figure 4.1. (a): KM revision rate by cause of revision for hip prostheses after first implant. Australian Orthopaedic Association National Joint Replacement Registry. Report 2019 [2]. (b): density estimation for failure time associated to the KM estimates by cause of revision for hip prostheses after first implant.

registry considers the following causes of revision: aseptic loosening, dislocation, fracture, infection and lysis. Survival curves are reported in Figure 4.1 and the raw data was recovered from those curves by solving (2.12). The probability densities for failure time were estimated via a non parametric approach, exploiting the smoothing provided by polynomial splines with time as a covariate [33].

As expected, according to literature already discussed in chapter 2, infection and dislocation have high probability to occur within the first two years, while aseptic loosening, fracture and lysis have high probability to occur long-term (Figure 4.1).

The probability densities in Figure 4.1 were used to fill in the Probability Matrix B and the Operational Probability Matrix C as prior information without any updating of observed information expressed via likelihood. Indeed, the aim is to test the model against data that does not provide information about the operated side and, then, information on the distribution of the failure times by cause of revision.

Probability densities for time of first revision after a primary and time of revision of an already revised device, conditional on cause of revision, were assumed to be the same. This hypothesis affects the results in a negligible way, as the aim is to estimate the revision rate of first implantation via KM estimator.

4.2 Data exploration and preprocessing

Data was composed by 12083 records reported by the register of the Autonomous Provinces of Trento and Bolzano between 2010 and 2018. Data was prepared for the analysis according to the rules in the following flow:

1. KEEP: records reporting one of the ICD9-CM codes in Table C.1 in at least one

of the variables describing the type of intervention performed. This decision was made following the same criterion used in the published 2020 RIAP report [14];

DROP: all of the other records.

Number of records kept: 12083 → 11976.

2. KEEP: records reporting either a code for primary intervention or a code for revision;

   DROP: records reporting both a code for primary intervention and a code for revision.

   (Assumption: primary and revision interventions can not be performed during the same hospitalization.)

   Number of records kept: 11976 → 11869.

3. KEEP: records about patients for which at most two hospitalization due to primary interventions are observed;

   DROP: records about patients for which more than two hospitalization due to primary interventions are observed.

   (Assumption: at most two primary interventions are admissible for a single patient, as people have only two hips.)

   Number of records kept: 11869 → 11639.

4. KEEP: records about patients for which at least one primary intervention is observed;

   DROP: records about patients for which only revision interventions are observed.

   (Assumption: if there is not record linkage between primary and revision interventions to impute, then records can not be used.)

   Number of records kept: 11639 → 10989.

5. KEEP: for the records about each patient, all the records for the patient on interventions performed strating from the first observed primary in time.

   DROP: for the records about each patient, all the records about revision intervention performed before the first observed primary.

Cause of revision	Dislocation	Fracture	Infection	Aseptic loosening	Lysis	Total
Frequency	54	44	63	143	9	313
Percentage	17.25%	14.06%	20.13%	45.69%	2.87%	100%

Table 4.1. Absolute and relative frequencies of causes of revision in Data from Autonomous Provences of Trento e Bolzano

(Assumption: if a revision intervention is performed before the first observed primary intervention, it is surely corresponding to an unobserved, left censored primary and there are neither record linkage between primary and revision interventions to impute, nor survival time to measure, then records can not be used.)

Number of records kept: 10989 → 10951.

The final dataset is composed by 10951 records where 10638 are primary interventions and 313 are revisions, performed on 9588 patients.

Causes of revision were assigned to each record associated to a revision intervention by looking at ICD9-CM codes reported in the HDR, according to Table C.2 and rules in Appendix C. Results about absolute and relative frequencies are reported in Table 4.1

4.3 Results

The performance of the propose linkage method was evaluated by using the median and the mode approaches introduced in Chapter 3. The considered metrics are the accuracy in correctly matching primary interventions and revisions and the ability of the model to lead to revision rate estimate as close as possible to the one obtained when the operated side is known.

By using the mode approach, model accuracy was 87.93%, while it was 87.65% on average with the median approach, with a minimum of 82.33% and a maximum of 92.24% over 100 runs (Figure 4.2).

Table 4.2 shows that the absolute difference in revision rate, when using the mode approach to link primaries and revisions, with respect to the KM estimate when the operated side is known, is equal to 0.0028 on average (yearly basis). The highest absolute difference is 0.0043 (year 4), while the smallest is equal to 0.0018 (year 1). When using the median approach the absolute difference is 0.0029 on

Estimate	1 year	2 year	3 year	4 year	5 year	6 year	7 year	8 year	9 year	**Average**
Mode	0.0018	0.0027	0.0040	0.0043	0.0028	0.0031	0.0027	0.0021	0.0021	**0.0028**
Median	0.0011	0.0019	0.0030	0.0034	0.0032	0.0034	0.0036	0.0025	0.0043	**0.0029**

Table 4.2. Revision Rate difference on yearly basis between the KM estimate when the operated side is known and the estimate after the use of the proposed linkage method. Comparison between the performances of mode and median approaches.

average; the minimum distance is 0.0011 (year 1) and the maximum is 0.0043 (year 9).

The significance of the difference between the curves was tested by log-rank test with equivalence as null hypothesis: for the mode approach, the null hypothesis is not rejected (with a p-value equal to 0.1873) and the difference between the curves is considered to be not statistically significant. Tha same occurs with the median approach, as the difference between the curves is not statistically significant (p-value equal to 0.1481).

The two approaches produce quite similar results in terms of accuracy in correctly matching primaries and revisions, with the accuracy in the mode approach that is slightly higher (Figure 4.2). On the other hand, the analysis of recovering of the KM estimate of the survival curve shows a better performance of the mode approach, that leads to an estimate of the revision rate that is very close to the one obtained when the operated side is known. With data at disposal, the mode approach seems to be more efficient and it leads to a reliable approximation of the survival curve in later times (years 5-9), while the median approach leads to a better recovering in early times (years 1-4). However, both approaches lead to a reliable estimate with respect to the surface where all possible KM estimates of the revision rate may lie after a random or naive linkage (Figure 4.3). The reason of the slight over-estimate in the revision rate observed in Figure 4.3 may be found in the short observation time window (only nine years). As already highlighted by looking at the model assessment on simulated data, the approximation gets better as the time window increases, since the linkage model has to deal with a lower number of unobserved primaries the revisions can be related to.

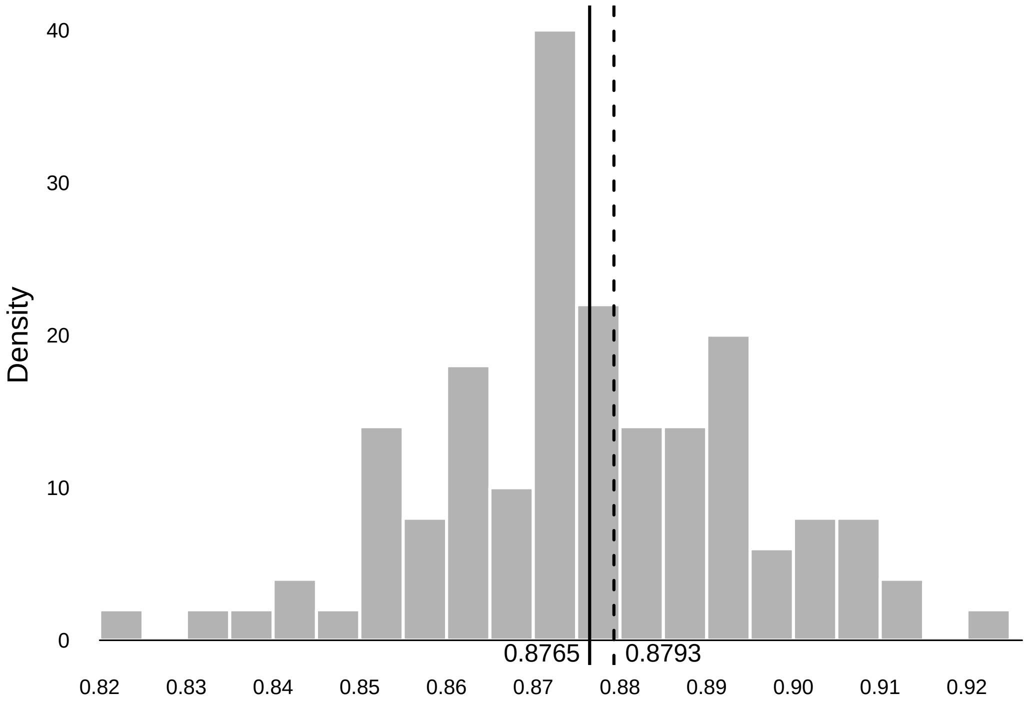


Figure 4.2. Accuracy of the model in matching primaries and revisions. Relative frequencies: accuracy by median approach over 100 runs. Black solid line: average accuracy by median approach. Black dashed line: accuracy in mode approach.

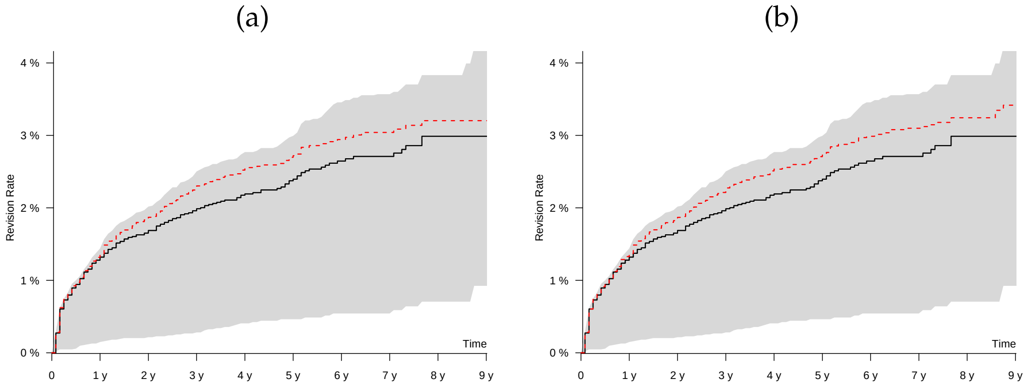


Figure 4.3. Real data: revision rate. Black, solid line: KM estimate when operated side is known. Red, dashed line: KM estimate whien the proposed linkage method is used to link primaries and revisions. Gray surface: region where KM estimates lie while naive or random linkage between primary interventions and revisions is performed. (a): mode approach; (b): median approach.

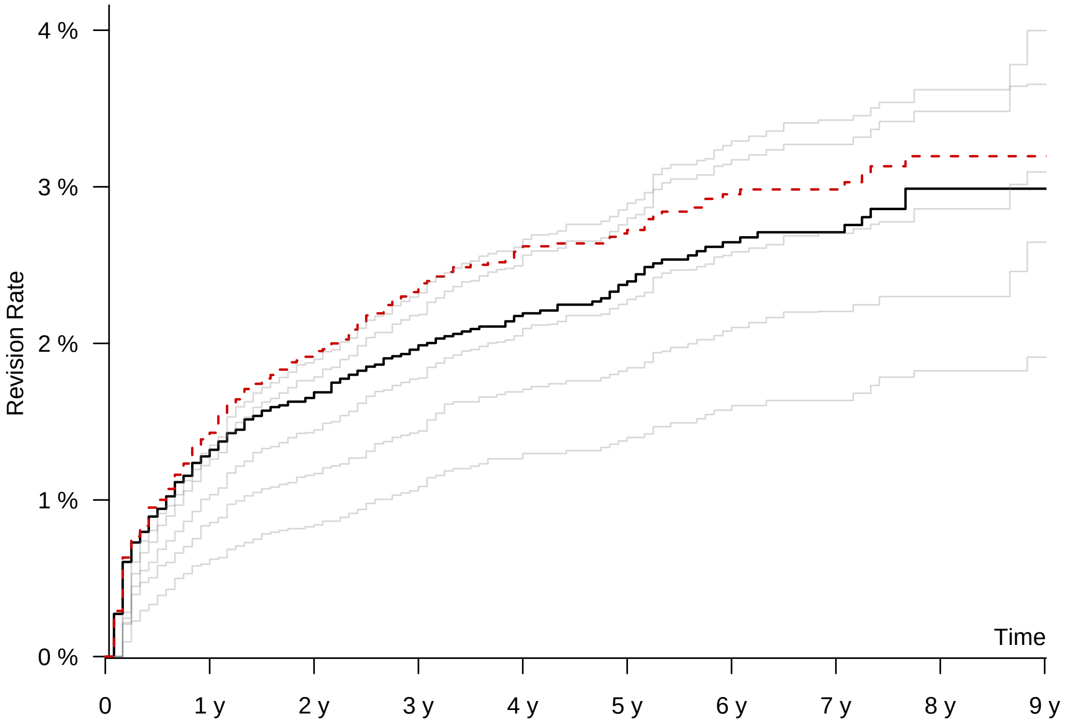


Figure 4.4. Real data: revision rate. Black, solid line: KM estimate when operated side is known. Red, dashed line: KM estimate whien the proposed linkage method is used to link primaries and revisions. Gray lines: realizations of the KM estimates while probabilistic record linkage models relying on partially identifying variables are used to link primary interventions and revisions.

4.3.1 Comparison with standard probabilistic record linkage methods applied to real data

To give a complete insight on the performance of the proposed linkage approach, it is compared to standard probabilistic record linkage model. As exposed in Chapters 2 and 3, probabilistic record linkage methods rely on partially identifiying variables that assign a certain probability to a set of candidate records to belong to the same patient and, then, be linked to each other. Unfortunately, in the context under analysis, as linkage uncertainty is due to the operated side, all candidate records already belong to the same patient and have identical realizations of partially identifying variables. This means that all possible linkages have the same probability to be selected, reducing the probabilistic record linkage method in a totally random linkage procedure.

Such method, when applied to real data from Autonomous Provinces of Trento and Bolzano, results in an average accuracy equal to 69% over 100 runs in linking interventions. Five realizations of the KM estimate derived after linking records by classic probabilistic record linkage method (resulting in a totally random procedure) are shown in Figure 4.4 and are compared to the resulting KM estimate obtained after using the mode approach of the proposed linkage method. Such curves show that a huge over-estimate or under-estimate of the revision rate can be the outcome of using standard probabilistic record linkage models in this context. Even if KM estimates can be very close to the one obtained when operated the side is known, this is because of randomness, and a huge bias is probable to be observed.

Concluding Remarks

Estimating the revision rate of primary implantation of joint prostheses is crucial to assess the clinical performance of the implant. Even if information about the devices is not available in HDRs, such data can be used to carry out epidemiological studies. An accurate analysis may be useful to assess surgical practice and to understand the reasons behind the interregional mobility. Moreover, an exploration at hospital level can help to evaluate the efficacy of the procuring strategies of a structure.

Nevertheless, given the missing information about the operated side, an uncritical procedure to match primaries and revisions can be misleading. An important over-estimate or under-estimate of the revision rate can lead all of the stakeholders to erroneous decisions, implying an increased risk for patients' safety.

The Bayesian probabilistic record linkage method proposed in this work is based on the prior information that causes of revision give about the expected failure time. It provides an effective tool to produce a matching between primary and revision interventions with high accuracy. The results of the application of the proposed method to real and simulated data show that the resulting estimate of the survival curve is close to the one obtained when the operated side is known and the observed error is substantially lower when compared to the one that can be obtained via a naive linkage procedure.

The performance of the model is sensitive with respect to the quality in filling Hospital Discharge Records, in particular in reporting the ICD9-CM codes used to identify causes of revision. Low quality data can lead to a low accuracy in linking primaries and revisions.

Another factor that can affect the performance of the proposed approach is the quality of the estimates of the prior probability densities conditional on the cause of revision derived from literature. Improving the accuracy of such estimates by the use, for instance, of advanced parametric or non parametric models and functional analysis or by the use of mixture models and Markov Chain Monte Carlo techniques, may lead to a better performance of the linkage method and can be the topic for a future work.

A lack of accuracy in filling Hospital Discharge Records and the method used to estimate the prior probability densities conditional to the cause of revision can be the reason of the slight difference in the assessment metrics when evaluating the performance of the proposed linkage method applied to simulated datasets in chapter 3 with respect to the results obtained on real data in chapter 4.

The proposed linkage method, with available national data, allows to produce the first reliable estimate of the revision rate for arthroplasty surgery at a national level. Moreover, after linking primaries and revisions with high accuracy, any further survival study, based on demographic and clinical information available within the Hospital Discharge Records, is possible.

Appendix A

Range of variation for Kaplan-Meier estimates

Table A.1 depicts the four linkage scenarios used to define the surface where KM estimates lie after producing any kind of linkage between primaries and revision when the operated side is unknown.

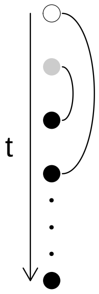
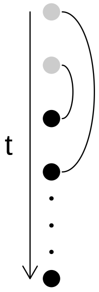
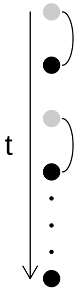
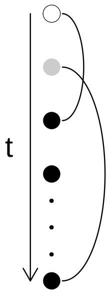
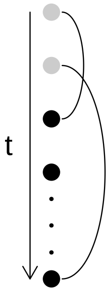
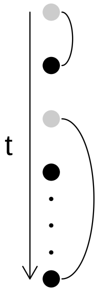
A) The KM estimate of the survival curve resulting from this linkage pattern leads to an estimated curve with the highest possible values in the first times (1-2 years).

B) The KM estimate of the survival curve resulting from this linkage pattern leads to an estimated curve with the highest possible values in the last time (9 years).

C) The KM estimate of the survival curve resulting from this linkage pattern leads to an estimated curve with the lowest possible values in the last time (9 years).

D) The KM estimate of the survival curve resulting from this linkage pattern leads to an estimated curve with the lowest possible values in the first times (1-2 years).

The maximum and the minimum values of those curves at each time *t* are used to define the surface boundaries.

Number of primaries	One primary	Two primaries in a row	Two primaries with revisions in between
Scenario A	The primary is linked to the first revision. All other revision are corresponding to an unobserved primary  	The second primary is linked to the closest revision; the first primary is linked to the second closest revision.  	Both primary are linked to their closest revision in time, respectively. 
Scenario B	The primary is linked to the farthest revision; all revisions in between are corresponding to an unobserved primary.  	The first primary is linked to the first revision; the second primary is linked to the farthest revision.  	The first primary is linked to the first revision; the second primary is linked to the farthest revision. 

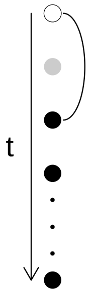
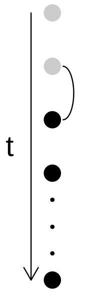
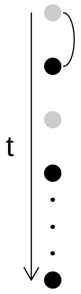
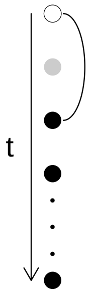
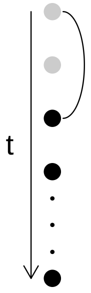
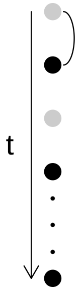
Scenario C	The primary is not linked to any revision; all revisions are corresponding to an unobserved primary.  	The second primary is linked to the closest revision; the first primary is not linked to any revision  	The first primary is linked to the first revision; the second primary is not linked to any revision. 
Scenario D	The primary is not linked to any revision. All revisions are corresponding to an unobserved primary.  	The first primary is linked to the closest revision; the second primary is not linked to any revision.  	The first primary is linked to the first revision; the second primary is not linked to any revision. 

Table A.1. Linkage scenarios to carry out the surface where KM estimates of the revision rate lie after performing any arbitrary linkage. Every revision for which the linkage is not specified is considered as a re-revision of a previously performed revision. Black dot: revision. Gray dot: primary. White dot: unobserved primary. Black line: linkage.

Appendix B

Theorems

Lemma B.1. *Given m vectors $V_1, \dots, V_m$ of size n such that $V_j \in \mathbb{R}^n \quad \forall j = 1, \dots, m$ and v_i^j i^{th} element of vector V_j,; given $\mathcal{P}(n,m)$, set of all the possible permutation with repetition of n integers $1, \dots, n \in \mathbb{N}$ of size m, then*

$$\sum_{(i_1,\dots,i_m)\in\mathcal{P}(n,m)} \prod_{j=1}^{m} v_{i_j}^j = \prod_{j=1}^{m} \sum_{i=1}^{n} v_i^j$$

Proof

The lemma will be proven by induction.

n=2

Using vectorial form and row by column product it is straightforward to show that

$$
\begin{aligned}
\sum_{i_1,i_2\in\mathcal{P}(n,2)} \prod_{j=1}^{2} v_{i_j}^{j} &= \mathbb{1}_n^T V_1 V_2^T \mathbb{1}_n = \\
&= \mathbb{1}_n^T \begin{bmatrix} v_1^1 v_1^2 & \cdots & v_1^1 v_n^2 \\ \vdots & \ddots & \vdots \\ v_n^1 v_1^2 & \cdots & v_n^1 v_n^2 \end{bmatrix}_{n\times n} \mathbb{1}_n = \\
&= \mathbb{1}_n^T \begin{bmatrix} v_1^1 v_1^2 + \cdots + v_1^1 v_n^2 \\ \vdots \\ v_n^1 v_1^2 + \cdots + v_n^1 v_n^2 \end{bmatrix}_{n\times 1} = \\
&= \mathbb{1}_n^T \begin{bmatrix} v_1^1 \sum_{i=1}^{n} v_i^2 \\ \vdots \\ v_n^1 \sum_{i=1}^{n} v_i^2 \end{bmatrix}_{n\times 1} = \\
&= v_1^1 \sum_{i=1}^{n} v_i^2 + v_2^1 \sum_{i=1}^{n} v_i^2 + \cdots + v_n^1 \sum_{i=1}^{n} v_i^2 = \\
&= \left(v_1^1 + \cdots + v_n^1\right) \sum_{i=1}^{n} v_i^2 = \left(\sum_{i=1}^{n} v_i^1\right) \left(\sum_{i=1}^{n} v_i^2\right) = \prod_{j=1}^{2} \sum_{i=1}^{n} v_i^j
\end{aligned}
$$

n=3

$$\begin{aligned}
\sum_{i_1,i_2,i_3\in\mathcal{P}(n,3)} \prod_{j=1}^{3} v_{i_j}^{j} &= \mathbb{1}_n^T V_1 V_2^T \mathbb{1}_n V_3^T \mathbb{1}_n = \\
&= \mathbb{1}_n^T \begin{bmatrix} v_1^1 \sum_{i=1}^n v_i^2 \\ \vdots \\ v_n^1 \sum_{i=1}^n v_i^2 \end{bmatrix}_{n\times 1} \quad V_3^T \mathbb{1}_n = \\
&= \mathbb{1}_n^T \begin{bmatrix} v_1^1 v_1^3 \sum_{i=1}^n v_i^2 & \cdots & v_1^1 v_n^3 \sum_{i=1}^n v_i^2 \\ \vdots & \ddots & \vdots \\ v_n^1 v_1^3 \sum_{i=1}^n v_i^2 & \cdots & v_n^1 v_n^3 \sum_{i=1}^n v_i^2 \end{bmatrix}_{n\times n} \quad \mathbb{1}_n \\
&= \mathbb{1}_n^T \begin{bmatrix} v_1^1 \left(\sum_{i=1}^n v_i^2\right) \left(\sum_{i=1}^n v_i^3\right) \\ \vdots \\ v_n^1 \left(\sum_{i=1}^n v_i^2\right) \left(\sum_{i=1}^n v_i^3\right) \end{bmatrix}_{n\times 1} \quad = \\
(*) &= \mathbb{1}_n^T \begin{bmatrix} v_1^1 \prod_{j=2}^3 \sum_{i=1}^n v_i^j \\ \vdots \\ v_n^1 \prod_{j=2}^3 \sum_{i=1}^n v_i^j \end{bmatrix}_{n\times 1} \quad = \\
&= \left(v_1^1 + \cdots + v_n^1\right) \prod_{j=2}^{3} \sum_{i=1}^{n} v_i^j = \prod_{j=1}^{3} \sum_{i=1}^{n} v_i^j
\end{aligned}$$

Assuming the statement is true for m, let proof it holds for $m+1$.

$$\begin{aligned}
\sum_{i_1,\ldots,i_{m+1}\in\mathcal{P}(n,m+1)} \prod_{j=1}^{m+1} v^j_{i_j} &= \mathbb{1}_n^T V_1 V_2^T \mathbb{1}_n V_3^T \mathbb{1}_n \cdots V_m^T \mathbb{1}_n V_{m+1}^T \mathbb{1}_n = \\
\textit{using } (*) \textit{ for } m \textit{ vectors } &= \mathbb{1}_n^T \begin{bmatrix} v^1_1 \prod_{j=2}^m \sum_{i=1}^n v^j_i \\ \vdots \\ v^1_n \prod_{j=2}^m \sum_{i=1}^n v^j_i \end{bmatrix}_{n\times 1} \quad V_{m+1}^T \mathbb{1}_n = \\
&= \mathbb{1}_n^T \begin{bmatrix} v^1_1 v^{m+1}_1 \prod_{j=2}^m \sum_{i=1}^n v^j_i & \cdots & v^1_1 v^{m+1}_n \prod_{j=2}^m \sum_{i=1}^n v^j_i \\ \vdots & \ddots & \vdots \\ v^1_n v^{m+1}_1 \prod_{j=2}^m \sum_{i=1}^n v^j_i & \cdots & v^1_n v^{m+1}_n \prod_{j=2}^m \sum_{i=1}^n v^j_i \end{bmatrix}_{n\times n} \quad \mathbb{1}_n \\
= &= \mathbb{1}_n^T \begin{bmatrix} v^1_1 \prod_{j=2}^{m+1} \sum_{i=1}^n v^j_i \\ \vdots \\ v^1_n \prod_{j=2}^{m+1} \sum_{i=1}^n v^j_i \end{bmatrix}_{n\times 1} \quad = \\
&= \left(v^1_1 + \cdots + v^1_n\right) \prod_{j=2}^{m+1} \sum_{i=1}^{n} v^j_i = \prod_{j=1}^{m+1} \sum_{i=1}^{n} v^j_i \qquad \square
\end{aligned}$$

Theorem B.1. *Given $\boldsymbol{s}_n \in \mathcal{P}_n$, set of all the possible permutation with repetition of n natural numbers $1, \ldots, n$ of size n; given the matrix C as it is defined in Definition 2.3, then*

$$\sum_{\boldsymbol{s}_n \in \mathcal{P}_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} = 1$$

Proof *Let consider $c_{\cdot,j}$ as j^{th} column vector of matrix C defined in Definition 2.3; $c_{\cdot,j} \in [0,1]^n \subset \mathbb{R}^n$. According to Lemma B.1*

$$\sum_{\boldsymbol{s}_n \in \mathcal{P}_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} = \prod_{j=1}^{n} \sum_{i=1}^{n} c_{i,j}.$$

By definition of Operational Probability Matrix, it is a left stochastic matrix, namely

$$\sum_{i=1}^{n} c_{i,j} = 1 \quad \forall j = 1, \ldots, n.$$

Then,

$$\sum_{\boldsymbol{s}_n \in \mathcal{P}_n} \prod_{j=1}^{n} c_{(\boldsymbol{s}_n)_j, j} = \prod_{j=1}^{n} \sum_{i=1}^{n} c_{i,j} = \prod_{j=1}^{n} 1 = 1. \ \square$$

Theorem B.2. *Given the Operational Probability Matrix, the Augmented Operational Probability Matrix, a Clinical Path $a_n \in A_n$ defined in Definitions 2.3, 2.4 and 2.1 respectively, and the probability of a clinical path to occur in (2.8), then*

$$\lim_{t_0 \to -\infty} P(\boldsymbol{a}_n) = \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}$$

Proof

$$\lim_{t_0 \to -\infty} P(\boldsymbol{a}_n) = \lim_{t_0 \to -\infty} \left((k_{oss} - 1) \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} + \right.$$

$$\left. +(1 - (k_{oss} - 1)) \left[(1 - q_{t,t_0}) \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} + q_{t,t_0} \frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}} \right] \right) =$$

$$= (k_{oss} - 1) \frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} +$$

$$+(1 - (k_{oss} - 1)) \left[\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}} \lim_{t_0 \to -\infty} (1 - q_{t,t_0}) + \frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n \in A_n} \prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}} \lim_{t_0 \to -\infty} q_{t,t_0} \right] =$$

By (2.6) for any probability density function $f(\cdot)$, the following holds:

$$\begin{aligned} \lim_{t_0 \to -\infty} q_{t,t_0} &= \lim_{t_0 \to -\infty} \int_{t-t_0}^{\infty} f(dx) \\ &= \lim_{t_0 \to -\infty} \int_{\infty}^{\infty} f(dx) = 0. \end{aligned}$$

Then,

$$
\begin{aligned}
\lim_{t_0\to-\infty} P(\boldsymbol{a}_n) =& (k_{oss}-1)\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}+\\
&+(1-(k_{oss}-1))\left[\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}\lim_{t_0\to-\infty}(1-q_{t,t_0})+\frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}\lim_{t_0\to-\infty} q_{t,t_0}\right]=\\
=&(k_{oss}-1)\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}+\\
&+(1-(k_{oss}-1))\left[\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}(1-0)+\frac{\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c^*_{(\boldsymbol{a}_n)_j,j}}0\right]=\\
=&(k_{oss}-1)\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}+(1-(k_{oss}-1))\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}=\\
=&\frac{\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}{\sum_{\boldsymbol{a}_n\in A_n}\prod_{j=1}^{n} c_{(\boldsymbol{a}_n)_j,j}}. \quad \square
\end{aligned}
$$

Appendix C

ICD9-CM codes mapping table

Tables C.1 and C.2 show the mapping from ICD9-CM codes to registry modalities for type of intervention and cause of revision, respectively.

In case of multiple possible assignment to a record for type of cause of revision (codes corresponding to both primary and revision in the same record; two or more codes related to two or more different causes of revision in the same record), the following preference hierarchy is considered:

$$\text{Infection} \succ \text{Lysis} \succ \text{Fracture} \succ \text{Dislocation} \succ \text{Aseptic loosening}$$

Type of intervention	ICD9-CM code	Description
Primary	81.51	Total hip arthroplasty
Primary	81.52	Partial hip arthroplasty
Revision	81.53	Hip arthroplasty revision
Revision	00.70	Hip arthroplasty total revision (both acetabular and femural components)
Revision	00.71	Hip arthroplasty partial revision (acetabular component)
Revision	00.72	Hip arthroplasty partial revision (femural component)
Revision	00.73	Hip arthroplasty revision (acetabular insert and/or femural head)
Revision	80.05	Arthrotomy for removal of prosthesis without replacement, hip

Table C.1. Mapping from ICD9-CM codes to regisrty-kind modalities for type of intervention.

Cause of revision	ICD9-CM code	Description
Aseptic loosening	996.4	Mechanical complication of internal orthopedic device implant and graft
	996.40	Unspecified mechanical complication of internal orthopedic device, implant, and graft
	996.41	Mechanical loosening of prosthetic joint
	996.47	Other mechanical complication of prosthetic joint implant
	996.49	Other mechanical complication of other internal orthopedic device, implant, and graft
	996.52	Mechanical complication due to graft of other tissue, not elsewhere classified
	996.59	Mechanical complication due to other implant and internal device, not elsewhere classified
Dislocation	345.80	Other forms of epilepsy and recurrent seizures, without mention of intractable epilepsy
	718.3	Recurrent dislocation of joint
	718.30	Recurrent dislocation of joint, site unspecified
	718.35	Recurrent dislocation of joint, pelvic region and thigh
	718.7	Developmental dislocation of joint
	718.70	Developmental dislocation of joint, site unspecified
	718.75	Developmental dislocation of joint, pelvic region and thigh
	835	Dislocation of hip
	835.0	Closed dislocation of hip
	835.00	Closed dislocation of hip, unspecified site
	835.01	Posterior dislocation of unspecified hip, initial encounter

Dislocation	835.02	Closed obturator dislocation of hip
	835.03	Other closed anterior dislocation of hip
	839.69	Closed dislocation, other location.
	905.6	Late effect of dislocation
	996.42	Dislocation of prosthetic joint
Infection	041.4	Escherichia coli [e. coli] infection in conditions classified elsewhere and of unspecified site
	682.6	Cellulitis and abscess of leg, except foot
	711.95	Infection and inflammatory reaction due to other internal orthopedic device, implant, and graft
	711.98	Unspecified infection of bone, other specified sites
	711.99	Unspecified infection of bone, multiple sites
	730.95	Unspecified infection of bone, pelvic region and thigh
	996.60	Infection and inflammatory reaction due to unspecified device, implant, and graft
	996.66	Infection and inflammatory reaction due to internal joint prosthesis
	996.67	Infection and inflammatory reaction due to other internal orthopedic device, implant, and graft
	998.51	Infected postoperative seroma
	998.59	Other postoperative infection
	998.6	Persistent postoperative fistula
Fracture	733.82	Nonunion of fracture
	808.0	Closed fracture of acetabulum
	820	Fracture of neck of femur
	820.0	Transcervical fracture closed
	820.00	Closed fracture of intracapsular section of neck of femur, unspecified
	820.01	Closed fracture of epiphysis (separation) (upper) of neck of femur

Fracture	820.02	Closed fracture of midcervical section of neck of femur
	820.03	Closed fracture of base of neck of femur
	820.09	Other closed transcervical fracture of neck of femur
	820.1	Transcervical fracture open
	820.10	Open fracture of intracapsular section of neck of femur, unspecified
	820.11	Open fracture of epiphysis (separation) (upper) of neck of femur
	820.12	Open fracture of midcervical section of neck of femur
	820.13	Open fracture of base of neck of femur
	820.19	Other open transcervical fracture of neck of femur
	820.2	Pertrochanteric fracture of femur closed
	820.20	Closed fracture of trochanteric section of neck of femur
	820.21	Closed fracture of intertrochanteric section of neck of femur
	820.22	Closed fracture of subtrochanteric section of neck of femur
	820.3	Pertrochanteric fracture of femur open
	820.30	Open fracture of trochanteric section of neck of femur, unspecified
	820.31	Open fracture of intertrochanteric section of neck of femur
	820.32	Open fracture of subtrochanteric section of neck of femur
	820.8	Closed fracture of unspecified part of neck of femur
	820.9	Open fracture of unspecified part of neck of femur
	821	Fracture of other and unspecified parts of femur

Fracture	821.0	Fracture of shaft or unspecified part of femur closed
	821.00	Closed fracture of unspecified part of femur
	821.01	Closed fracture of shaft of femur
	821.1	Fracture of shaft or unspecified part of femur open
	821.10	Open fracture of unspecified part of femur
	821.11	Open fracture of shaft of femur
	821.2	Fracture of lower end of femur closed
	821.20	Closed fracture of lower end of femur, unspecified part
	821.21	Closed fracture of condyle, femoral
	821.22	Closed fracture of epiphysis, lower (separation) of femur
	821.23	Closed supracondylar fracture of femur
	821.29	Other closed fracture of lower end of femur
	821.3	Fracture of lower end of femur open
	821.30	Open fracture of lower end of femur, unspecified part
	821.31	Open fracture of condyle, femoral
	821.32	Open fracture of epiphysis. Lower (separation) of femur
	821.33	Open supracondylar fracture of femur
	821.39	Other open fracture of lower end of femur
	905.3	Late effect of fracture of neck of femur
	905.4	Late effect of fracture of lower extremities
	996.43	Broken prosthetic joint implant
	996.44	Peri-prosthetic fracture around prosthetic joint
Lysis	718.25	Pathological dislocation of joint, pelvic region and thigh
	996.45	Peri-prosthetic osteolysis

Lysis	996.46	Articular bearing surface wear of prosthetic joint

Table C.2. Mapping from ICD9-CM codes to regisrty-kind modalities for cause of revision.

Appendix D

R code

All applications in Chapters 3 and 4 are implemented by software R, version 3.6.3 (2020-02-29) – "Holding the Windsock". A code implementing Mode and Median approaches of the proposed probabilistic record linkage method follows:

```
# Input: candidate.records

# A data.frame with following variables:


## $myid: charachter; the id of each candidate record

## $primary: boolean (T/F); T = the record corresponds to an hospitalization
## for primary intervention

## $revision boolean (T/F); T = the record corresponds to an hospitalization
## for revision intervention


### $primary and $revision are mutually excluseve


## $t: the time elapsed between the intervention date and the starting time
## of data collection

## $caur: the cause of revision of every revision record (NA if primary)

### $caur must be filled with elements of the set of
### row names of prior.measures


# Input: T.max

# a numeric valure with the total number of grid times
# in the observed time window
# Input: f.param.estimate

# The parameter estimate of the f (estimated density of time elapsing
# between two primaries);
# here it is assumed to be the estimated parameter of an exponential


# Input: prior.measures

# A matrix with T.max columns where each row is the estimated probability
# of revision at a time t conditional on a certan cause of revision
# (given in the row name)


# Input: all.possible.path

# A list that provides all the possible clinical paths;
# the i^th element of the list provides all posible clinical
# paths in case of a Probability matrix associated to a set
# of candidate records with i revisions

# Mode Approch
assign.path.mode=function(candidate.records,
                          T.max,
                          f.param.estimate,
                          prior.measures,
                          all.possible.path){

# Count the total number of hospitalizations due to a revision
# intervention among the candidate records
n=sum(candidate.records$revision)

# If there is not any revision, set the output to NA
if(n==0) out=cbind(NA,NA)

# Let consider the cases for which revision interventions exist
# among candidate records
if(n>0){

        ## Set the case of two primary interventions exist among
        ## candidate records
        case="double"

        ## Define which records are related to primary interventions
        primaries=which(candidate.records$primary==T)

        ## Define which records are related to revision interventions
        revisions=which(candidate.records$revision==T)

        ## Count the total number of hospitalizations due to a parimary
        ## intervention among the candidate records
```

```
k=sum(candidate.records$primary)

## Count the total number of candidate records
N=nrow(candidate.records)

## Consider the case with one observed primaries
if(k==1){

        ### Compute the time elapsing bitween the intervention time
        ### and the last observation time
        t0=abs(candidate.records$t[primaries]-T.max)

        ### Compute f() at t0: here it is exponential shaped by
        ### hypothesis (section simulation study)
        p=pexp(t0,rate = f.param.estimate)

        ### Define if an unobserved primary must be considered
        ### in the candidate records according to estimated f
        case=sample(c("double","single"),
        size = 1,
        replace = F,
        prob = c(1-p,p))
}

## Consider the case of an unobserved primary
if(k==1 & case=="double"){

        ### Initialize a matrix M[i,j]
        ### with element M_i,j = time elapsing
        ### between interventions i and j
        M=matrix(0,N,N)

        ### Fill the matrix M
        for(i in 1:N) M[1:i,i]=abs(candidate.records$t[i]-
                                   candidate.records$t[1:i])

        ### Adapt the matrix M to the case of an unobserved primary
        M=matrix(ceiling(M[-nrow(M),revisions]),ncol=n)
        colnames(M)=candidate.records$caur[revisions]
        M=rbind(T.max,M)

        ### Initialize the operational probability matrix
        P=matrix(NA,nrow=(N),ncol=n)

        ### Fill the Operational Probability Matrix with prior measures
        for (i in 1:n) P[,i]=prior.measures[colnames(M)[i],(M[,i]+1)]

        ### Define the probabilities of every possible clinical path
        ### according to the Operational Probability Matrix
        foo=c()
        for(i in 1:(2^n)){
                foo[i]=prod(P[cbind(as.matrix(
                   all.possible.path[[n]]
                   )[i,],1:n)])
        }
        ### Choose the most probable path
        chosen.path=all.possible.path[[n]][which.max(foo),]
        options(warn=-1)

        ### Set the candidate records id to which revisions
        ### are linked to
        ref.id=as.vector(c(0,candidate.records$myid))
        options(warn=0)

        ### Define the output as two colums: the candidate
        ### records id and the id of the hospitalizations
        ### to which they are linked to
        out=cbind(id=candidate.records$myid[revisions],
                  ref.id=ref.id[as.vector(as.numeric(chosen.path))])
}

## Consider the case with one observed primary and
## no unobserved primaries
if(k==1 & case=="single"){
        ## Define the output as two colums: the candidate records
        ## id and the id of the hospitalizations to which they
        ## are linked to
        out=cbind(id=candidate.records$myid[revisions],
                  ref.id=candidate.records$myid[-N])

        ## Consider the case with two observed primaries
        if(k==2){

        ### Initialize a matrix M[i,j]
        ### with element M_i,j = time elapsing
        ### between interventions i and j
        M=matrix(0,N,N)

        ### Fill the matrix M
        for(i in 1:N) M[1:i,i]=abs(candidate.records$t[i]-
                                   candidate.records$t[1:i])
        M=matrix(ceiling(M[-nrow(M),revisions]),ncol=n)
        colnames(M)=candidate.records$caur[revisions]

        ### Initialize the operational probability matrix
        P=matrix(NA,nrow=(N-1),ncol=n)

        ### Fill the Operational Probability Matrix
        ### with prior measures
        for (i in 1:n) P[,i]=prior.measures[colnames(M)[i],(M[,i]+1)]
        foo=c()
```

```
                ### Define the probabilities of every possible clinical path
                ### according to the Operational Probability Matrix
                for(i in 1:(2^n)){
                        foo[i]=prod(P[cbind(as.matrix(
                            all.possible.path[[n]])[i,],1:n)]
                            )
                }
                ### Choose the most probable path
                chosen.path=all.possible.path[[n]][which.max(foo),]
                options(warn=-1)

                ### Set the candidate records id to which revisions
                ### are linked to
                ref.id=as.vector(candidate.records$myid)
                options(warn=0)

                ### Define the output as two colums: the candidate
                ### records id and the id of the hospitalizations
                ### to which they are linked to
                out=cbind(id=candidate.records$myid[revisions],
                          ref.id=ref.id[as.vector(as.numeric(chosen.path))])

        }

        # Return the output
        return(out)
}


# Median Approch
assign.path.median=function(candidate.records,
                            T.max,
                            f.param.estimate,
                            prior.measures,
                            all.possible.path){

# Count the total number of hospitalizations due to a revision
# intervention among the candidate records
n=sum(candidate.records$revision)

# If there is not any revision, set the output to NA
if(n==0) out=cbind(NA,NA)

# Let consider the cases for which revision interventions exist
# among candidate records
if(n>0){

        ## Set the case of two primary interventions exist among
        ## candidate records
        case="double"
```

```
## Define which records are related to primary interventions
primaries=which(candidate.records$primary==T)

## Define which records are related to revision interventions
revisions=which(candidate.records$revision==T)

## Count the total number of hospitalizations due to a parimary
## intervention among the candidate records
k=sum(candidate.records$primary)

## Count the total number of candidate records
N=nrow(candidate.records)

## Consider the case with one observed primaries
if(k==1){

        ### Compute the time elapsing bitween the intervention time
        ### and the last observation time
        t0=abs(candidate.records$t[primaries]-T.max)

        ### Compute f() at t0: here it is exponential shaped by
        ### hypothesis (section simulation study)
        p=pexp(t0,rate = f.param.estimate)

        ### Define if an unobserved primary must be considered
        ### in the candidate records according to estimated f
        case=sample(c("double","single"),
        size = 1,
        replace = F,
        prob = c(1-p,p))
}

## Consider the case of an unobserved primary
if(k==1 & case=="double"){

        ### Initialize a matrix M[i,j]
        ### with element M_i,j = time elapsing
        ### between interventions i and j
        M=matrix(0,N,N)

        ### Fill the matrix M
        for(i in 1:N) M[1:i,i]=abs(candidate.records$t[i]-
        candidate.records$t[1:i])

        ### Adapt the matrix M to the case of an unobserved primary
        M=matrix(ceiling(M[-nrow(M),revisions]),ncol=n)
        colnames(M)=candidate.records$caur[revisions]
        M=rbind(T.max,M)

        ### Initialize the operational probability matrix
        P=matrix(NA,nrow=(N),ncol=n)
        ### Fill the Operational Probability Matrix with prior measures
        for (i in 1:n) P[,i]=prior.measures[colnames(M)[i],(M[,i]+1)]

        ### Define the probabilities of every possible clinical path
        ### according to the Operational Probability Matrix
        foo=c()
        for(i in 1:(2^n)){
                foo[i]=prod(P[cbind(as.matrix(
                all.possible.path[[n]]
                )[i,],1:n)])
        }
        ### Choose the most probable path
        chosen.path=all.possible.path[[n]][
        sample(1:(2^k),size = 1,replace = T,prob = foo),
        ]
        options(warn=-1)

        ### Set the candidate records id to which revisions
        ### are linked to
        ref.id=as.vector(c(0,candidate.records$myid))
        options(warn=0)

        ### Define the output as two colums: the candidate
        ### records id and the id of the hospitalizations
        ### to which they are linked to
        out=cbind(id=candidate.records$myid[revisions],
                ref.id=ref.id[as.vector(as.numeric(chosen.path))])
        }

## Consider the case with one observed primary and
## no unobserved primaries
if(k==1 & case=="single"){
        ## Define the output as two colums: the candidate records
        ## id and the id of the hospitalizations to which they
        ## are linked to
        out=cbind(id=candidate.records$myid[revisions],
        ref.id=candidate.records$myid[-N])

        ## Consider the case with two observed primaries
        if(k==2){

        ### Initialize a matrix M[i,j]
        ### with element M_i,j = time elapsing
        ### between interventions i and j
        M=matrix(0,N,N)

        ### Fill the matrix M
        for(i in 1:N) M[1:i,i]=abs(candidate.records$t[i]-
        candidate.records$t[1:i])
        M=matrix(ceiling(M[-nrow(M),revisions]),ncol=n)
```

```
                colnames(M)=candidate.records$caur[revisions]

                ### Initialize the operational probability matrix
                P=matrix(NA,nrow=(N-1),ncol=n)

                ### Fill the Operational Probability Matrix
                ### with prior measures
                for (i in 1:n) P[,i]=prior.measures[colnames(M)[i],(M[,i]+1)]
                foo=c()

                ### Define the probabilities of every possible clinical path
                ### according to the Operational Probability Matrix
                for(i in 1:(2^n)){
                        foo[i]=prod(P[cbind(as.matrix(
                        all.possible.path[[n]])[i,],1:n)]
                        )
                }
                ### Choose the most probable path
                chosen.path=all.possible.path[[n]][
                sample(1:(2^k),size = 1,replace = T,prob = foo),
                ]
                options(warn=-1)

                ### Set the candidate records id to which revisions
                ### are linked to
                ref.id=as.vector(candidate.records$myid)
                options(warn=0)

                ### Define the output as two colums: the candidate
                ### records id and the id of the hospitalizations
                ### to which they are linked to
                out=cbind(id=candidate.records$myid[revisions],
                ref.id=ref.id[as.vector(as.numeric(chosen.path))])

        }

        # Return the output
        return(out)
}

# Produce the automatic linkage by sets of candidate records
# belonging to the same patient (Data$pseudo)

# Data: the whole dataset

# A data.frame with the variables required by assign.path.mode
#  and assign.path.median

# Set assign.path equal to assign.path.mode or assign.path.median

matching.revisions = function(Data,
                              T.max,
                              f.param.estimate,
                              prior.measures,
                              all.possible.path){
        out=by(Data,Data$pseudo,assign.path,
        T.max = T.max,
        f.param.estimate = f.param.estimate,
        prior.measures = prior.measures,
        all.possible.path = all.possible.path)
        out=lapply(out,t)
        out=matrix(as.vector(as.numeric(do.call(c,lapply(out,unlist)))),
            ncol=2,byrow = T)
        return(out)
}

# Output: a matrix with two columns: the candidate records id's and
# the corresponding linked record id's
```

Bibliography

[1] Danielle GT Arts, Nicolette F De Keizer, and Gert-Jan Scheffer. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. *Journal of the American Medical Informatics Association*, 9(6):600–611, 2002.

[2] Australian Orthopaedic Association National Joint Replacement Registry (AOANJRR). *Hip, Knee & Shoulder Arthroplasty: 2019 Annual Report*. AOA, 2019.

[3] Ileana Baldi, Antonio Ponti, Roberto Zanetti, Giovannino Ciccone, Franco Merletti, and Dario Gregori. The impact of record-linkage bias in the cox model. *Journal of evaluation in clinical practice*, 16(1):92–96, 2010.

[4] Yoav Ben-Shlomo, Ashley Blom, Chris Boulton, Robin Brittain, Emma Clark, Richard Craig, Sebastian Dawson-Bowling, Kevin Deere, Colin Esler, Andy Goldberg, et al. *National Joint Registry for England and Wales: 16th annual report, 2019*. Available from: www.njrcentre.org.uk, 2020.

[5] Nicholas M Bernthal, Paul C Celestre, Alexandra I Stavrakis, John C Ludington, and Daniel A Oakes. Disappointing short-term results with the depuy asr xl metal-on-metal total hip arthroplasty. *The Journal of Arthroplasty*, 27(4): 539–544, 2012.

[6] David J Biau and Moussa Hamadouche. Estimating implant survival in the presence of competing risks. *International orthopaedics*, 35(2):151–155, 2011.

[7] David Jean Biau, Aurélien Latouche, and Raphaël Porcher. Competing events influence estimated survival probability: when is kaplan-meier analysis appropriate? *Clinical Orthopaedics and Related Research®*, 462:229–233, 2007.

[8] Fraser Birrell, Olof Johnell, and Alan Silman. Projecting the need for hip

replacement over the next three decades: influence of changing demography and threshold for surgery. *Annals of the rheumatic diseases*, 58(9):569–572, 1999.

[9] Nick Black. Patient reported outcome measures could help transform healthcare. *Bmj*, 346, 2013.

[10] I Bormann. Digitizeit. *DigitizeIt, Braunschweig, Germany*, 2016. URL `https://www.digitizeit.xyz/`.

[11] Nicolae Ciprian Bota, Dan-Viorel Nistor, Sergiu Caterev, and Adrian Todor. Historical overview of hip arthroplasty: From humble beginnings to a high-tech future. *Orthopedic Reviews*, 13(1), 2021.

[12] AR Britton, DW Murray, CJ Bulstrode, K McPherson, and RA Denham. Pain levels after total hip replacement: their use as endpoints for survival analysis. *The Journal of bone and joint surgery. British volume*, 79(1):93–98, 1997.

[13] E Ciminello, P Laricchiuta, and M Torre. Appendice 2a. interventi di artroprotesi: analisi dei dati sdo nazionali 2016 e 2017. In M Torre, E Carrani, S Ceccarelli, A Biondi, M Masciocchi, and A Cornacchia, editors, *Registro Italiano ArtroProtesi. Report Annuale 2019*, pages 75–106. Il Pensiero Scientifico Editore, 2020. ISBN 978-88-490-0693-3.

[14] E Ciminello, S Madi, P Laricchiuta, and M Torre. Appendice 2a. interventi di artroprotesi: analisi dei dati sdo nazionali 2018 e 2019. In M Torre, S Ceccarelli, A Biondi, E Carrani, M Masciocchi, and A Cornacchia, editors, *Registro Italiano ArtroProtesi. Report Annuale 2020*, pages 91–122. Il Pensiero Scientifico Editore, 2021. ISBN 978-88-490-0714-5.

[15] Kelly L Corbett, Elena Losina, Akosua A Nti, Julian JZ Prokopetz, and Jeffrey N Katz. Population-based rates of revision of primary total hip arthroplasty: a systematic review. *PloS one*, 5(10):e13520, 2010.

[16] Paul A Dieppe and L Stefan Lohmander. Pathogenesis and management of pain in osteoarthritis. *The Lancet*, 365(9463):965–973, 2005.

[17] Direzione Generale della Programmazione sanitaria - Ufficio 6, Ministero della Salute. Rapporto annuale sull'attività di ricovero ospedaliero. Dati SDO 2019, 2020. URL `https://www.salute.gov.it/portale/documentazione/p6_2_2_1.jsp?lingua=italiano&id=3002`.

[18] Halbert L Dunn. Record linkage. *American Journal of Public Health and the Nations Health*, 36(12):1412–1416, 1946.

[19] D Ellams, O Forsyth, A Mistry, et al. *National Joint Registry for England and Wales: 7th annual report, 2010*. Available from: www.njrcentre.org.uk, 2010.

[20] European Union EC. REGULATION (EU) 2017/745 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC. *Official Journal of the European Union L117/1*, 60, 2017.

[21] Romain Forestier, Alain Francon, Valérie Briole, Céline Genty, Xavier Chevalier, and Pascal Richette. Prevalence of generalized osteoarthritis in a population with knee osteoarthritis. *Joint Bone Spine*, 78(3):275–278, 2011.

[22] C Germinario, M Torre, S Palmieri, PL Lopalco, R Prato, D Martinelli, and working group: Regione Puglia. Registro regionale delle protesi d'anca. *Rapporti ISTISAN*, 05(18):46–53, 2005.

[23] Leicester Gill, Michael Goldacre, Hugh Simmons, Glenys Bettley, and Myfanwy Griffith. Computerised linking of medical records: methodological guidelines. *Journal of Epidemiology & Community Health*, 47(4):316–319, 1993.

[24] Marianne H Gillam, Philip Ryan, Stephen E Graves, Lisa N Miller, Richard N de Steiger, and Amy Salter. Competing risks survival analysis applied to data from the australian orthopaedic association national joint replacement registry. *Acta orthopaedica*, 81(5):548–555, 2010.

[25] Terence J Gioe, Kathleen K Killeen, Katherine Grimm, Susan Mehle, and Karen Scheltema. Why are total knee replacements revised?: analysis of early revision in a community knee implant registry. *Clinical Orthopaedics and Related Research®*, 428:100–106, 2004.

[26] Harvey Goldstein, Katie Harron, and Angie Wade. The analysis of record-linked data using multiple imputation with data value priors. *Statistics in medicine*, 31(28):3481–3493, 2012.

[27] Ted A Gooley, Wendy Leisenring, John Crowley, and Barry E Storer. Estimation of failure probabilities in the presence of competing risks: new representations of old estimators. *Statistics in medicine*, 18(6):695–706, 1999.

[28] Shaun J Grannis, J Marc Overhage, and Clement J McDonald. Analysis of identifier performance using a deterministic linkage algorithm. In *Proceedings of the AMIA Symposium*, page 305. American Medical Informatics Association, 2002.

[29] Stephen E Graves. The value of arthroplasty registry data. *Acta orthopaedica*, 81(1):8–9, 2010.

[30] Roee Gutman, Christopher C Afendulis, and Alan M Zaslavsky. A bayesian procedure for file linking to analyze end-of-life medical costs. *Journal of the American Statistical Association*, 108(501):34–47, 2013.

[31] Patricia Guyot, AE Ades, Mario JNM Ouwens, and Nicky J Welton. Enhanced secondary analysis of survival data: reconstructing the data from published kaplan-meier survival curves. *BMC medical research methodology*, 12(1):1–13, 2012.

[32] WH Harris. Last decade in THA: unsettling and disappointing. *J Arthroplasty*, 29(3):648–649, 2014.

[33] Trevor J Hastie and Robert J Tibshirani. *Generalized additive models*. Routledge, 2017.

[34] Peter Herberts and Henrik Malchau. Long-term registration has improved the quality of hip replacement: a review of the swedish thr register comparing 160,000 cases. *Acta Orthopaedica Scandinavica*, 71(2):111–121, 2000.

[35] Michel H Hof, Anita C Ravelli, and Aeilko H Zwinderman. A probabilistic record linkage model for survival data. *Journal of the American Statistical Association*, 112(520):1504–1515, 2017.

[36] Kevin T Hug, Tyler S Watters, Thomas P Vail, and Michael P Bolognesi. The withdrawn ASR™ THA and hip resurfacing systems: how have our patients fared over 1 to 6 years? *Clinical Orthopaedics and Related Research®*, 471(2): 430–438, 2013.

[37] EC Huskisson, PA Dieppe, AK Tucker, and LB Cannell. Another look at osteoarthritis. *Annals of the rheumatic diseases*, 38(5):423–428, 1979.

[38] Joseph G Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, pages 46–60, 2000.

[39] Joseph G Ibrahim, Ming-Hui Chen, Yeongjin Gwon, and Fang Chen. The power prior: theory and applications. *Statistics in medicine*, 34(28):3724–3749, 2015.

[40] Thorvaldur Ingvarsson, Gunnar Hagglund, Halldor Jonsson, and L Stefan Lohmander. Incidence of total hip replacement for primary osteoarthrosis in Iceland 1982–1996. *Acta orthopaedica Scandinavica*, 70(3):229–233, 1999.

[41] Italy. Art. 57 Legge Finanziaria n.289 del 27.12.2002. *Gazzetta Ufficiale Serie Generale n.305 del 31.12.2002*, 2002.

[42] Italy. Decreto del Ministro della Salute del 20.02.2007. Nuove modalità per gli adempimenti previsti dall'articolo 13 del Decreto Legislativo 24.02.1997, n. 46 e successive modificaioni e per la registrazione dei dispositivi impiantabili attivi nonché per l'iscrizione nel Repertorio dei dispositivi medici. *Gazzetta Ufficiale Serie Generale n.63 del 16.03.2007*, 2007.

[43] Italy. DL n.179 del 18.10.2012 (Supplemento ordinario n.194/L alla Gazzetta Ufficiale del 19.10.2012, n.245) convertito in legge con Legge di conversione n.221 del 17.12.2012 "Ulteriori misure urgenti per la crescita del Paese". *Gazzetta Ufficiale Serie Generale n.294 del 18.12.2012*, 2012.

[44] Italy. Decreto del Presidente del Consiglio dei Ministri del 03.03.2017 - "Identificazione dei sistemi di sorveglianza e dei registri di mortalità, di tumori e di altre patologie". *Gazzetta Ufficiale Serie Generale n. 109 del 12.05.2017*, 2017.

[45] S Mehdi Jafari, Catelyn Coyle, SM Javad Mortazavi, Peter F Sharkey, and Javad Parvizi. Revision hip arthroplasty: infection is the most common cause of failure. *Clinical Orthopaedics and Related Research®*, 468(8):2046–2051, 2010.

[46] Matthew A Jaro. Probabilistic linkage of large public health data files. *Statistics in medicine*, 14(5-7):491–498, 1995.

[47] John D Kalbfleisch and RL Prentice. Survival analysis, 1980.

[48] Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

[49] Theofilos Karachalios, George Komnos, and Antonios Koutalos. Total hip arthroplasty: survival and modes of failure. *EFORT open reviews*, 3(5):232–239, 2018.

[50] Gunky Kim and Raymond Chambers. Regression analysis under incomplete linkage. *Computational Statistics & Data Analysis*, 56(9):2756–2770, 2012.

[51] Stephen Richard Knight, Randeep Aujla, and Satya Prasad Biswas. Total hip arthroplasty-over 100 years of operative history. *Orthopedic reviews*, 3(2), 2011.

[52] Christoph Kolling, Beat R Simmen, G Labek, and Jörg Goldhahn. Key factors for a successful national arthroplasty register. *The Journal of bone and joint surgery. British volume*, 89(12):1567–1573, 2007.

[53] Steven Kurtz, Fionna Mowat, Kevin Ong, Nathan Chan, Edmund Lau, and Michael Halpern. Prevalence of primary and revision total hip and knee arthroplasty in the United States from 1990 through 2002. *JBJS*, 87(7):1487–1497, 2005.

[54] G Labek, M Thaler, W Janda, M Agreiter, and B Stöckl. Revision rates after total joint replacement: cumulative results from worldwide joint register datasets. *The Journal of Bone and Joint Surgery. British Volume*, 93(3):293–297, 2011.

[55] Sarah Lacny, Todd Wilson, Fiona Clement, Derek J Roberts, Peter D Faris, William A Ghali, and Deborah A Marshall. Kaplan-meier survival analysis overestimates the risk of revision arthroplasty: a meta-analysis. *Clinical Orthopaedics and Related Research®*, 473(11):3431–3442, 2015.

[56] Sarah Lacny, Todd Wilson, Fiona Clement, Derek J Roberts, Peter Faris, William A Ghali, and Deborah A Marshall. Kaplan–meier survival analysis overestimates cumulative incidence of health-related events in competing risk settings: a meta-analysis. *Journal of clinical epidemiology*, 93:25–35, 2018.

[57] Partha Lahiri and Michael D Larsen. Regression analysis with linked data. *Journal of the American statistical association*, 100(469):222–230, 2005.

[58] Jerald F Lawless. *Statistical models and methods for lifetime data*, volume 362. John Wiley & Sons, 2011.

[59] Ian D Learmonth, Claire Young, and Cecil Rorabeck. The operation of the century: total hip replacement. *The Lancet*, 370(9597):1508–1519, 2007.

[60] Thoralf R Liebs, Farina Splietker, and Joachim Hassenpflug. Is a revision a revision? an analysis of national arthroplasty registries' definitions of revision. *Clinical Orthopaedics and Related Research®*, 473(11):3421–3430, 2015.

[61] Zhihui Liu, Benjamin Rich, and James A Hanley. Recovering the raw data behind a non-parametric survival curve. *Systematic reviews*, 3(1):1–10, 2014.

[62] Henrik Malchau, Peter Herberts, Thomas Eisler, Göran Garellick, and Peter Söderman. The swedish total hip replacement register. *JBJS*, 84(suppl_2): S2–S20, 2002.

[63] Henrik Malchau, Göran Garellick, Daniel Berry, William H Harris, Otto Robertson, Johan Kärrlholm, David Lewallen, Charles R Bragdon, Lars Lidgren, and Peter Herberts. Arthroplasty implant registries over the past five decades: development, current, and future impact. *Journal of Orthopaedic Research®*, 36(9):2319–2330, 2018.

[64] Christopher T Martin, John J Callaghan, Yubo Gao, Andrew J Pugely, Steve S Liu, Lucian C Warth, and Devon D Goetz. What can we learn from 20-year followup studies of hip replacement? *Clinical Orthopaedics and Related Research®*, 474(2):402–407, 2016.

[65] Victoria K Matharu and Gulraj S Matharu. Metal-on-metal hip replacements: implications for general practice. *British Journal of General Practice*, 67(665): 544–545, 2017.

[66] Michael H McGlincy. A bayesian record linkage methodology for multiple imputation of missing links. In *ASA Proceedings of the Joint Statistical Meetings*, pages 4001–4008. Citeseer, 2004.

[67] Medical Device Clinical Evaluation Working Group IMDRF. Post-Market Clinical Follow-Up Studies, 2021. URL `http://www.imdrf.org/documents/documents.asp#technical`.

[68] Tom Melvin and Marina Torre. New medical device regulations: the regulator's view. *EFORT open reviews*, 4(6):351–356, 2019.

[69] Andrew J Metcalfe, Maria LE Andersson, Rhian Goodfellow, and Carina A Thorstensson. Is knee osteoarthritis a symmetrical disease? analysis of a 12 year prospective cohort study. *BMC musculoskeletal disorders*, 13(1):1–8, 2012.

[70] CG Moran and TC Horton. Total knee replacement: the joint of the decade: A successful operation, for which there's a large unmet need. *Bmj*, 320(7238): 820, 2000.

[71] DW Murray and SJD Frost. Pain in the assessment of total knee replacement. *The Journal of bone and joint surgery. British volume*, 80(3):426–431, 1998.

[72] John Neter, E Scott Maynes, and R Ramanathan. The effect of mismatching on the measurement of response errors. *Journal of the American Statistical Association*, 60(312):1005–1027, 1965.

[73] Beat Neuenschwander, Gorana Capkun-Niggli, Michael Branson, and David J Spiegelhalter. Summarizing historical information on controls in clinical trials. *Clinical Trials*, 7(1):5–18, 2010.

[74] Audrey Neuprez, Arnaud Henri Neuprez, Jean-François Kaux, William Kurth, Christophe Daniel, Thierry Thirion, Jean-Pierre Huskin, Philippe Gillet, Olivier Bruyère, and Jean-Yves Reginster. Total joint replacement improves pain, functional quality of life, and health utilities in patients with late-stage knee and hip osteoarthritis for up to 5 years. *Clinical rheumatology*, 39(3):861–871, 2020.

[75] OECD(2021). Health at a Glance 2021: OECD Indicator, 2021. URL `https://www.oecd-ilibrary.org/sites/8b492d7a-en/index.html?itemId=/content/component/8b492d7a-en`.

[76] Alma B Pedersen, Søren P Johnsen, Søren Overgaard, Kjeld Søballe, Henrik T Sørensen, and Ulf Lucht. Total hip arthroplasty in denmark: incidence of primary operations and revisions during 1996–2002 and estimated future demands. *Acta Orthopaedica*, 76(2):182–189, 2005.

[77] Roberto Picus, Dieter Randeu, Mirko Bonetti, and Carla Melani. Registro protesico della pa di bolzano. tassi di revisione a 1 e 2 anni degli interventi primari di protesi d'anca e di ginocchio - primi confronti tra gli ospedali provinciali. report riap 2019. In M Torre, E Carrani, S Ceccarelli, A Biondi, M Masciocchi, and A Cornacchia, editors, *Registro Italiano ArtroProtesi. Report Annuale 2019*, page 15. Il Pensiero Scientifico Editore, 2020. ISBN 978-88-490-0693-3.

[78] Silvano Piffer, Cristiana Armaroli, Maria Antonella D'Alpaos, Sergio Mezzina, Eugenio Carrani, and Marina Torre. Progetto di recupero dei dati storici sugli interventi di anca per ampliare la completezza del registro provinciale di artroprotesi della provincia autonoma di trento. report riap 2019. In M Torre,

E Carrani, S Ceccarelli, A Biondi, M Masciocchi, and A Cornacchia, editors, *Registro Italiano ArtroProtesi. Report Annuale 2019*, page 14. Il Pensiero Scientifico Editore, 2020. ISBN 978-88-490-0693-3.

[79] Anne Postler, Cornelia Lützner, Franziska Beyer, Eric Tille, and Jörg Lützner. Analysis of total knee arthroplasty revision causes. *BMC musculoskeletal disorders*, 19(1):1–6, 2018.

[80] Catherine Quantin, Christine Binquet, Karima Bourquard, Ronny Pattisina, Béatrice Gouyon-Cornet, Cyril Ferdynus, Jean-Bernard Gouyon, and François-André Allaert. Which are the best identifiers for record linkage? *Medical informatics and the Internet in medicine*, 29(3-4):221–227, 2004.

[81] Jonas Ranstam and Otto Robertsson. Statistical analysis of arthroplasty register data. *Acta orthopaedica*, 81(1):10–14, 2010.

[82] Jonas Ranstam, Johan Kärrholm, Pekka Pulkkinen, Keijo Mäkelä, Birgitte Espehaug, Alma Becic Pedersen, Frank Mehnert, Ove Furnes, and NARA Study Group. Statistical analysis of arthroplasty data: Ii. guidelines. *Acta orthopaedica*, 82(3):258–267, 2011.

[83] Sowmya R Rao and David A Schoenfeld. Survival methods. *Circulation*, 115 (1):109–113, 2007.

[84] Otto Robertsson. Knee arthroplasty registers. *The Journal of bone and joint surgery. British volume*, 89(1):1–4, 2007.

[85] Otto Robertsson, Michael J Dunbar, Kaj Knutson, and Lars Lidgren. Past incidence and future demand for knee arthroplasty in Sweden: a report from the Swedish Knee Arthroplasty Register regarding the effect of past and future population changes on the number of arthroplasties performed. *Acta Orthopaedica Scandinavica*, 71(4):376–380, 2000.

[86] O Rolfson, J Kärrholm, LE Dahlberg, and G Garellick. Patient-reported outcomes in the swedish hip arthroplasty register: results of a nationwide prospective observational study. *The Journal of bone and joint surgery. British volume*, 93(7):867–875, 2011.

[87] Ola Rolfson, Eric Bohm, Patricia Franklin, Stephen Lyman, Geke Denissen, Jill Dawson, Jennifer Dunn, Kate Eresian Chenok, Michael Dunbar, Søren Overgaard, et al. Patient-reported outcome measures in arthroplasty registries:

report of the Patient-Reported Outcome Measures Working Group of the International Society of Arthroplasty Registries Part II. Recommendations for selection, administration, and analysis. *Acta orthopaedica*, 87(sup1):9–23, 2016.

[88] Ola Rolfson, Kate Eresian Chenok, Eric Bohm, Anne Luebbeke-Wolff, Geke Denissen, Jennifer Dunn, Stephen Lyman, Patricia Franklin, Michael Dunbar, Søren Overgaard, et al. Patient-reported outcome measures in arthroplasty registries. *Acta orthopaedica*, 87:3–8, 2016.

[89] E Romanini, M Torre, V Manno, G Baglio, S Conti, et al. Chirurgia protesica dell'anca: la mobilità interregionale. *Giornale Italiano di Ortopedia e Traumatologia*, 34, 2008.

[90] E Romanini, V Manno, S Conti, G Baglio, S Di Gennaro, M Masciocchi, and M Torre. Mobilità interregionale e chirurgia protesica del ginocchio. *Ann Ig*, 21:329–336, 2009.

[91] Emilio Romanini, Francesco Decarolis, Ilaria Luzi, Gustavo Zanoli, Michele Venosa, Paola Laricchiuta, Eugenio Carrani, and Marina Torre. Total knee arthroplasty in Italy: reflections from the last fifteen years and projections for the next thirty. *International orthopaedics*, 43(1):133–138, 2019.

[92] Leslie L Roos Jr, Andre Wajda, and J Patrick Nicol. The art and science of record linkage: methods that work with few identifiers. *Computers in Biology and Medicine*, 16(1):45–57, 1986.

[93] Adrian Sayers, Jonathan T Evans, Michael R Whitehouse, and Ashley W Blom. Are competing risks models appropriate to describe implant failure? *Acta orthopaedica*, 89(3):256–258, 2018.

[94] Michael Schemper, Samo Wakounig, and Georg Heinze. The estimation of average hazard ratios by weighted cox regression. *Statistics in medicine*, 28 (19):2473–2489, 2009.

[95] Guido Schwarzer, Martin Schumacher, Thomas B Maurer, and Peter E Ochsner. Statistical analysis of failure times in total joint replacement. *Journal of clinical epidemiology*, 54(10):997–1003, 2001.

[96] Leonard Shan, B Shan, David Graham, and Akshat Saxena. Total hip replacement: a systematic review and meta-analysis on mid-term quality of life. *Osteoarthritis and cartilage*, 22(3):389–406, 2014.

[97] Eun-Kyoo Song, Jong-Keun Seon, Jae-Young Moon, Yim Ji-Hyoun, and P Kinov. The evolution of modern total knee prostheses. *Rijeka, Croatia: InTech*, 10: 54343, 2013.

[98] S Stea, B Bordini, M De Clerico, K Petropulacos, and A Toni. First hip arthroplasty register in Italy: 55,000 cases and 7 year follow-up. *International orthopaedics*, 33(2):339–346, 2009.

[99] Alesio Tarantino, Emilio Romanini, Michele Venosa, Marina Torre, Irene Schettini, Remo Goderecci, Giandomenico Logroscino, and Vittorio Calvisi. Registro italiano artroprotesi: curva di apprendimento e ottimizzazione delle procedure di immissione dei dati. *Recenti Progressi in Medicina*, 111(5):327–330, 2020.

[100] M Torre, I Luzi, E Carrani, L Leone, E Romanini, and G Zanoli. *Progetto Registro Italiano Artroprotesi. Idea, sviluppo e avvio. Primo Report*. Il Pernsiero Scientifico Editore, 2014. ISBN 978-88-490-0513-4.

[101] M Torre, E Carrani, S Ceccarelli, A Biondi, M Masciocchi, and A Cornacchia. *Registro Italiano ArtroProtesi. Report Annuale 2019*. Il Pernsiero Scientifico Editore, 2020. ISBN 978-88-490-0693-3.

[102] M Torre, S Ceccarelli, A Biondi, E Carrani, M Masciocchi, and A Cornacchia. *Registro Italiano ArtroProtesi. Report Annuale 2020*. Il Pernsiero Scientifico Editore, 2021. ISBN 978-88-490-0714-5.

[103] Marina Torre, Emilio Romanini, Gustavo Zanoli, Eugenio Carrani, Ilaria Luzi, Luisa Leone, and Stefania Bellino. Monitoring outcome of joint arthroplasty in italy: implementation of the national registry. *Joints*, 5(02):070–078, 2017.

[104] Marina Torre, Ilaria Luzi, Fiorino Mirabella, Martina Del Manso, Gustavo Zanoli, Gabriele Tucci, and Emilio Romanini. Cross-cultural adaptation and validation of the Italian version of the Hip disability and Osteoarthritis Outcome Score (HOOS). *Health and quality of life outcomes*, 16(1):1–9, 2018.

[105] Marina Torre, Eugenio Carrani, Michela Franzò, Enrico Ciminello, Iuliia Urakcheevaa, Duilio Luca Bacocco, Riccardo Valentini, Simona Pascucci, Saif Madi, Carla Ferrara, Virgilia Toccaceli, Letizia Sampaolo, Stefania Ceccarelli, Alessia Biondi, and Paola Laricchiuta. Il registro italiano delle protesi impiantabili: una nuova realtà per la sicurezza del paziente. *Bollettino epidemiologico nazionale*, 2(2):16–23, 2021.

[106] Keith Tucker. How registry data can improve outcomes from joint replacement–a seminal paper. *Acta orthopaedica*, 91(3):230–231, 2020.

[107] Keith Tucker, Paul Gregg, Peter Kay, Martyn Porter, Peter Howard, Martin Pickford, and Crina Cacou. Monitoring the introduction and performance of a joint replacement: the united kingdom metal-on-metal alert. *JBJS*, 93 (Supplement_3):37–42, 2011.

[108] I Urakcheeva, A Biondi, and M Torre. *Italian Arthroplasty Registry. Annual Report 2019 – Addendum*. Il Pensiero Scientifico Editore, 2020.

[109] Claus Varnum, Alma Bečić Pedersen, Per Hviid Gundtoft, and Søren Overgaard. The what, when and how of orthopaedic registers: an introduction into register-based research. *EFORT open reviews*, 4(6):337–343, 2019.

[110] Ines Vielgut, Norbert Kastner, Karin Pichler, Lukas Holzer, Mathias Glehr, Gerald Gruber, Andreas Leithner, Gerold Labek, and Patrick Sadoghi. Application and surgical technique of total knee arthroplasties: a systematic comparative analysis using worldwide registers. *International orthopaedics*, 37 (8):1465–1469, 2013.

[111] Theresa Weldring and Sheree MS Smith. Article commentary: patient-reported outcomes (PROs) and patient-reported outcome measures (PROMs). *Health services insights*, 6:HSI–S11093, 2013.

[112] Ian Wilson, Eric Bohm, Anne Lübbeke, Stephen Lyman, Søren Overgaard, Ola Rolfson, Annette W-Dahl, Mark Wilkinson, and Michael Dunbar. Orthopaedic registries with patient-reported outcome measures. *EFORT Open Reviews*, 4 (6):357–367, 2019.

[113] V Wylde and AW Blom. The failure of survivorship. *The Journal of Bone and Joint Surgery. British volume*, 93-B(5), 2011.