On the Epidemiology of Soft Tissue Sarcoma and Risk of Cancer following Knee prosthesis surgery. Important factors and methodological notes

Wagner, Philippe

2017

Document Version:
Manuskriptversion före sakkunniggranskning

On the Epidemiology of Soft Tissue Sarcoma and Risk of Cancer following Knee prosthesis surgery

On the Epidemiology of Soft Tissue Sarcoma and Risk of Cancer following Knee prosthesis surgery

Important factors and methodological notes

Philippe Wagner

LUND
UNIVERSITY

DOCTORAL DISSERTATION
by due permission of the Faculty Medicine, Lund University, Sweden.
To be defended at Segerfalkssalen, BMC, Lund. Thursday May 4th, 2017 at 09.00.

Faculty opponent
Associate Professor Håkan Jonsson

Organization LUND UNIVERSITY	Document name DOCTORAL DISSERTATION
	Date of issue May 4th, 2017
Author(s) Philippe Wagner	Sponsoring organization

Title and subtitle

On the Epidemiology of Soft Tissue Sarcoma and Risk of Cancer following Knee prosthesis surgery - Important factors and methodological notes

Abstract

The current thesis originates from research efforts in the oncology and orthopedic departments at Lund University Hospital, and treats what initially appeared to be two different subjects.

One subject deals with the epidemiology of adult soft tissue sarcomas (STS), a group of often fatal diseases of unknown cause, treated by both oncologists and orthopedic surgeons. Here, the historical inability to clarify their etiology have resulted in a lack of preventive strategies and a significant loss of years of life. Here, we study factors associated with stature and reproductive events, heredity and tissue trauma.

The other subject treats the risk of cancer following knee prosthesis surgery, a concern among orthopedic researchers and an important public health issue as the number of prostheses is steadily increasing, not least in younger patients.

These two subjects turn out to be related as soft tissue tumors have been identified in locations adjacent to prosthetic implants. Therefore, identifying risk factors of STS may provide clues about the potential carcinogenic effects associated with prosthetic implants and/or the associated surgery.

In the study of both subjects, we work with some of the world's largest and most detailed study populations. A population based case-control study set in South Sweden between 1998 and 2009, entailing almost 1000 cases and the Swedish Knee Arthroplasty Register, including all operated patients from 1975 to present day.

We find that 57% of the STS, and 74% of extremity STS, incidence in our study cohort can be attributed to factors related to stature and reproductive events, heredity and tissue trauma. We also find an excess number of STSs in the knee prosthesis cohort, together with a low but significant excess of more common cancers. Considering the occurrence of STS, we conclude that tissue trauma may be a contributing factor in the increased cancer risk of knee arthroplasty patients.

Key words Soft tissue sarcoma, epidemiology, risk factor, knee arthroplasty, cancer		
Classification system and/or index terms (if any)		
Supplementary bibliographical information		Language
ISSN and key title		ISBN
Recipient's notes	Number of pages 97	Price
	Security classification	

Signature _____ Date _____

On the Epidemiology of Soft Tissue Sarcoma and the Risk of Cancer following Knee prosthesis surgery

Important factors and methodological notes

Philippe Wagner

LUND
UNIVERSITY

Coverphoto by

Faculty
Department

ISBN xxx-xx-xx-x
ISSN xxx-xx-xx-

Printed in Sweden by Media-Tryck, Lund University
Lund 2016

To Marlon, Inez and Sofie

Content

List of papers

1. Wagner P, Alvegård T, Ranstam J, Rydholm A, Vult von Steyern F, Olsson H. Oral contraceptive use, parity, and constitutional characteristics in soft tissue sarcoma: a Swedish population based case-control study 1988-2009. Cancer Causes Control. 2014 Sep;25(9):1167-77.

2. Wagner P, Alvegård T, Rydholm A, Vult von Steyern F, Olsson H. Hereditary cancers and the protective effect of oral contraceptive use in adult soft tissue sarcomas - a Swedish total population study 1988 - 2009.

3. Wagner P, Olsson H. The heredity odds ratio - in a case control setting.

4. Wagner P, Alvegård T, Lidgren L, Robertsson O, Vult von Steyern, F, Olsson, H. Tissue trauma and the subsequent risk of soft tissue sarcoma - a population based case control study in the south of Sweden 1988-2009.

5. Wagner P, Olsson H, Lidgren L, Robertsson O, Ranstam J. Increased cancer risks among arthroplasty patients: 30 year follow-up of the Swedish Knee Arthroplasty Register. Eur J Cancer. 2011 May;47(7):1061-71.

List of Abbreviations

Ah	Aryl hydrocarbon
BMI	Body Mass Index
CLR	Conditional Logistic Regression
DNA	Deoxyribonucleic acid
ELR	Exact Logistic Regression
ER	Estrogen receptor
GH	Growth Hormone
GIST	Gastrointestinal Stromal Tumor
h^2	Narrow sense heritability
HR	Hazard rate ratio
HRT	Hormone replacement therapy
IARC	International Agency For Research on Cancer
IGF-1	Insulin growth factor 1
IGF-1R	IGF-1 Receptor
IRR	Incidence Rate Ratio
LFS	Li-Fraumeni Syndrome
MAR	Missing-At-Random
MFH	Malignant Fibrous Histiocytoma
MICE	Multiple Imputation by Chained Equations
MISS	Melanoma in Southern Sweden
NSAID	Non-Steroidal Anti-Inflammatory Drug
OA	Osteoarthritis
OC	Oral Contraceptive

OR	Odds ratio
OR_g	The heritability odds ratio
PAR	Population Attributable Risk
PR	Progesterone receptor
PS	Propensity Score
RA	Rheumatoid Arthritis
RCS	Restricted Cubic Spline
RR	Relative Risk
SCR	Swedish Cancer Register
SEER	Surveillance, Epidemiology, and End Results Program
SH	Sex hormone
SKAR	Swedish Knee Arthroplasty Register
SPAR	Statens PersonAdressRegister
STS	Soft Tissue Sarcoma
TCDD	2,3,7,8-Tetrachlorodibenzo-p-dioxin
TP53	Tumor Protein p53
UPS	Undifferentiated pleomorphic sarcoma

Introduction

The current thesis originates from intersecting research efforts in the oncology and orthopedic departments at Lund University Hospital, where the author started to work some odd ten years ago. Driven by the authors' different interests within cancer epidemiology and orthopedic research, as well as the availability of large and rich cohorts, it treats what initially appeared to be two different subjects.

One subject deals with the identification of risk factors of adult soft tissue tumors, soft tissue sarcomas (STS), a group of often fatal diseases of unknown cause, treated by both oncologists and orthopedic surgeons. Here, the historical inability to clarify their etiology have resulted in a lack of preventive strategies and a significant loss of years of life, as early stage tumors often lack distinct symptoms and consequently are diagnosed late with a subsequently poor prognosis. As the literature on risk factors of STS is sparse, we took a que from research on common cancers to identify some key areas of interest. These included stature and reproductive events, heredity and tissue trauma.

The other subject treated the risk of cancer following knee prosthesis surgery, a long-standing cause for concern among orthopedic researchers and an important public health issue as the number of prostheses is steadily increasing, not least in younger patients.

As it turns out, these two subjects are related, as STSs have been identified in locations adjacent to prosthetic implants, which has raised subsequent questions of causality [1,2]. Therefore, identifying risk factors of STS may additionally provide clues about the potential carcinogenic effects associated with prosthetic implants and/or the associated surgery, as is illustrated below, in figure 1.

In the study of both subjects, the author had the unique privilege of working with some of the world's largest and most detailed study populations. For the study of STS epidemiology, we worked with what is one of the world's largest population based case-control studies on risk factors. Run by the department of Cancer epidemiology, set in the South Sweden Health Care Region, collecting all cases of STS through the regional tumor registry between 1988 and 2009, the study comprises almost 1000 STS cases and matched controls, each with information corresponding to a seven page questionnaire. For the study of knee prosthesis patients we worked with the Swedish Knee Arthroplasty Register, one of the world's oldest national registries of knee prosthesis patients to date, dating back to 1975 [3].

While these subjects are important in their own right, they also present a series of interesting challenges in terms of epidemiological and statistical methodology. For example, the study of STS as a heterogonous group of rare diseases with potentially different etiologies include issues of subgroup analyses and adjusting for multiple confounders in small sample studies. The rarity of the disease also leads to difficulties in investigating disease heritability. The study of prosthesis surgery related cancers includes challenges of evaluating potential detection bias.

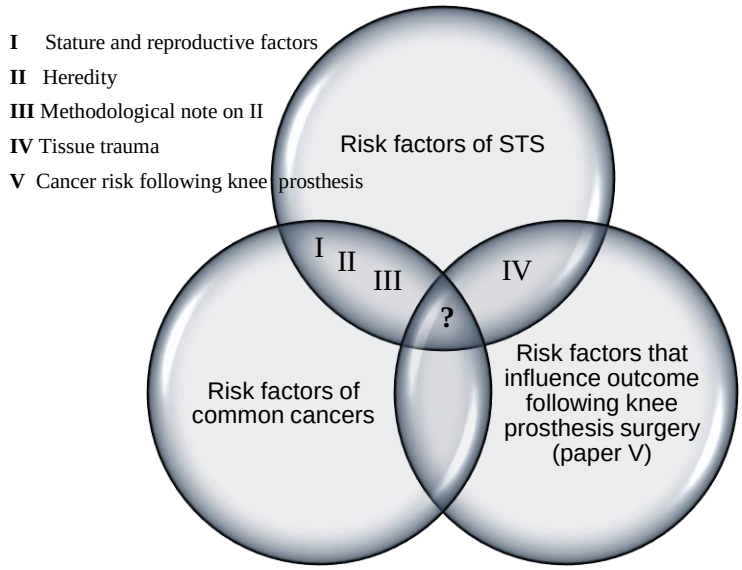


Figure 1 – Grapichal illustration of the focus of papers I-IV and how they potentially relate to the outcomes of paper V of the present thesis. The questionmark indicates factors that may not only help explain the potential increase in STS incidence after knee prosthesis surgery, but also some of the remaining tumour disease outcomes.

Background

What follows is a brief summary of the theoretical underpinnings of some of the thoughts that went into the design and interpretation of the studies included in the present thesis.

Tumor development

The signifying characteristic leading to the development of a tumor is the uncontrolled proliferation of cells. Failing normal response to regulating signals, cancer cells grow, divide, and invade normal tissue and organs. This process can be viewed in stages [4]. Starting during *initiation*, a genetic alteration occur in a single cell which induces abnormal cell proliferation [5] that during *promotion* leads to the outgrowth of a population of derived tumor cells. During *progression* the cell population continues to grow, acquiring additional mutations that accumulate in the genetic material of the cells. Some of these mutations, such those promoting increasing growth rates, provide the cell with a selective advantage over remaining cells, making it dominant in the cell population. This process is referred to as clonal selection and it continues throughout tumor development helping the tumor acquire novel traits and grow more rapidly, becoming increasingly malignant [5]. The process of carcinogenesis can also be viewed in terms of the necessary traits to acquire for a tumor to become malignant. These include self-sufficiency in growth signaling, insensitivity to anti-growth signaling, by-passing apoptosis, ability to replicate without bounds, ability to sustain angiogenesis and to metastasize and invade adjacent tissues, often referred to as the six hallmarks of cancer [6]. In recent years, additional hallmarks have been added, such as; deregulated metabolism, evading the immune system, genome instability and inflammation [7]

Changes in genetic materials that drive this process may be spontaneous, inherited or acquired through carcinogenic or radiation exposure [8]. Significant changes are often those affecting regulatory genes that control cell proliferation, differentiation and survival. There are two main types of genes involved; proto-oncogenes and tumor suppressor genes. A proto-oncogene is a gene that under normal circumstance helps the cell grow but following mutation accelerates growth out of control. In contrast, a tumor suppressor gene is a gene that functions to slow

down cell division, repair DNA or induce apoptosis. A significant difference between the two is that a proto-oncogene is activated, while a tumor suppressor gene is inactivated by mutation [9].

In some hereditary cancer syndromes, tumor initiation has been described as a "two-hit" process [10]. Under ordinary circumstances, normal gene function may be maintained after mutation as long as one allele stays intact. Gene function is not lost until a second mutation occur in the remaining allele. This model explains the earlier age-at-onset and elevated risk of multiple primary tumors of inherited disease, as the first "hit" has already occurred at birth. This has been the dominating theory to explain the actions of tumor suppressor genes.

Some causes and contributors

The major cancer risk factors known today include tobacco smoking, responsible for 19.4% of cancers, then, in men, deficient intake of fruit and vegetables, 6.1%, occupational exposures 4.9%, alcohol consumption, 4.6%, while in women, overweight and obesity, 6.9% and infectious agents 3.7%. Taken together, 42.7% of all cancers can be attributed to the population distribution of a combination of 14 risk factors, in the UK in 2010 [11]. These factors only account for avoidable exposures and subsequently preventable cancers. Highly heritable factors have been estimated to account for only 5% of cancer risks, although general genetic factors have been estimated to account for 33% of variation in cancer outcomes [12].

While some risk factors cause DNA damage and initiate tumor development others may simply promote its growth by providing tools and building blocks, creating a favorable microenvironment. Some, such as smoking habits, can be modified, while others, such as inherited genes and age, cannot. Of specific interest in the present thesis are sex and growth hormones, heritable predisposition and tissue trauma and inflammation, as well as exposure to metals.

Inflammation and tissue trauma

As first suggested by Virchow in 1863 [13], it has long been recognized that inflammation following infection or tissue trauma, may promote tumorigenesis. It may initiate development by the production of free radicals causing DNA damage and mutation. It may also support advancement to later stages of development, because while mutations in tumor suppressor and proto-oncogenes are necessary for tumor development, they are not sufficient. Promotion and progression relies on ancillary processes and cells not necessarily cancerous themselves [14]. In response to major tissue damage, induced by infection or trauma, lost tissue must be replaced. To achieve this goal, the inflammatory processes works to promote survival and expansion of remaining cells, often comprised of tissue stem cells, to repopulate damaged tissue. Suggested by the fact that inflammatory mediators such as

chemokines, cytokines and eicanosoids are known to promote increased proliferation in both normal and tumor cells, together with several other molecules and pathways which are involved in both homeostasis, tissue regeneration and repair, as well as in tumorigenesis, these same inflammatory processes may promote survival and proliferation in initiated cancer cells, thereby contributing to tumor growth and progression. [14]. There are numerous empirical examples of inflammatory disease and events of tissue trauma that are associated with increased risks of tumor disease, and of anti-inflammatory drug treatments that reduce these risks [14]. Evidence of this connection is abundant.

However, the relationship between cancer and inflammation is not straight forward. There is also evidence to suggest that the immune and inflammatory systems prevent tumor development through immunosurveillance, where there may be dedicated mechanisms to identify and eliminate initiated cells, and adaptive immune recognition of cancer specific antigens. Although it has been suggested that the net effect of inflammation is to promote cancer development, and that the relationship between inflammation and cancer cannot be described in one grand unifying theory [14].

Sex hormones

Sex hormones (SH) are steroid hormones, produced in the gonads, integral to the maintenance and development of secondary sex characteristics, to the pubertal growth spurt and to the reproductive process [15,16]. They include androgens, estrogens and progestogens, of which testosterone, estradiol and progesterone are important derivatives.

Estrogens, while regulating important processes in normal tissues, with receptors present in almost all tissue, are known to influence the development and progression of a number of diseases [17], They, for instance, play a vital role in the pathology of hormone-dependent tumors, both as a promotors and initiators of disease, as they induce cell proliferation in addition to exerting genotoxic effects through free radical which are byproducts of estrogen metabolism that damage DNA and induce mutation. Estrogen, and estrogen receptors (ER), have been reported to influence the development of at least four groups of tumors; breast and gynecological (cervical, endometrial and ovarian), endocrine (adrenocortical, ovarian, pancreatic, prostate and thyroid), digestive cancers (colorectal, esophageal, liver and pancreatic) and lung cancer [18].

The effect of estrogen on these tumors depend on two nuclear estrogen receptors, ERα and ERβ. ERα may promote cell proliferation in endocrine gland, breast and gynecological cancers while inhibiting the same in digestive and lung cancer. ERβ, on the other hand functions the other way around and inhibits the former while promoting the latter. [18] The organ based differences in effect are seen in epidemiological studies on cancer risk where hormone replacement therapy based on estrogen alone is observed to increase the risks of endometrial and ovarian cancer

[19,20] but possibly reduce risk of colorectal cancer [21]. This is further validated by observations with respect to prognosis, where the effect of changes in ER vary [18]. An extensively studied organ in relation to estrogen and tumorigenesis is the breast. Here, tumors are observed to express a high level of aromatase enzyme, enabling the biogenesis of estrogen in an autocrine fashion. Further evidence of estrogen involvement in the progression of tumors of the breast are the studies, both in vitro and in vivo, showing a reduction in cancer cell proliferation by use of aromatase inhibitors [18].

Progesterone is a steroid hormone produced in the female ovaries, adrenal gland and placenta of pregnant females. It influences several processes including the menstrual cycle, pregnancy, lactation and breast feeding and pubertal breast development. Its actions are mainly mediated through receptors PRA and PRB, found in tissues of the breast and reproductive organs, where altered receptor function may contribute to tumorigenesis. In the breast, progesterone together with estrogen promotes increased proliferation and cell survival, while in contrast it reduces estrogen induced growth in the uterus and protects ovaries from malignant transformation [22]. Progesterone therefore plays a role in both breast and gynecological cancers. Progestins are prescribed as part of contraceptives or postmenopausal hormone replacement therapy as a mean of counteracting endometrial growth induced by estrogen [23].

Oral contraceptives (OC) combining both estrogen and progesterone have been reported to raise risks of breast, cervix and liver and reduces risk of endometrial and ovarian cancers [24]. Breast cancer risks may be elevated with early use. [25-27]. Hormone replacement therapy (HRT) with estrogen plus progestagen is associated with an increased risk of breast [28 29] and ovarian cancer [20], a reduced risk of colorectal cancer [30]and no apparent effect on endometrial [31] or lung cancer [32].

Additional evidence of the effect of sex hormones on cancer risk is that events that increase the cumulative exposure to endogenous estrogen and progesterone, such as early onset menstruation, late onset of menopause, late first time pregnancy or never having had a child, as well as events that reduces cumulative exposure, such as; pregnancy and breast feeding have all been observed to affect breast cancer risk.[33]. Risks of ovarian and endometrial cancers decline with the number of full-term pregnancies [34-36].

Growth hormones / Insulin-like growth factors

More than 50 years ago, insulin-like growth factors (IGF) were discovered as proteins produced in the liver to mediate effects of growth hormone (GH) on growth and differentiation of skeletal muscle and bone. The extensive research into its role in carcinogenesis, however, did not reach its peak until the 1990s [37]. IGF-1 has since been found to influence each key stage of cancer development, from proliferation, apoptosis, angiogenesis and metastasis to therapeutic resistance. Influences that affect almost every type of cell in the body [38]. The actions of IGF-1 is thought to

promote carcinogenesis through the induction of hyperproliferation, disrupting regular balance between proliferation and cell death. An imbalance thought to favor "initiated" stem cells. As it additionally acts as an anti-apoptotic agent, transformed cells may also experience prolonged survival. Even though these effects may be slight, over time, affecting a large number of cells, the risk of malignant transformation increases. IGF-1 is consequently not an initiator of cancer development, but a powerful promotor. [37]. Indeed, elevated levels of insulin-like growth factor 1 (IGF-1) have been linked to increased risks of prostate, breast, and colorectal cancers [37]. After the extensive research efforts into IGF-1 since the 1990:ies, systemic therapies are now available that block IGF-signaling to the benefit of both cancer and sarcoma patients.

Inherited genetic mutations

There are several inherited genetic (germ line) mutations that manifest as familial cancer syndromes involving common cancers. These include Hereditary breast and ovarian cancer syndrome caused by mutations in the BRCA1/BRCA2 genes, Cowden Syndrome and Bannayan-Riley-Ruvalcaba syndrome, with mutations in the PTEN gene, familial malignant melanoma, CDKN2/p16 genes, Familial Adenomatous Polyposis (FAP), the APC gene , Lynch syndrome (Hereditary nonpolyposis colorectal cancer), including MLH1 and MSH2 genes, Hereditary prostate cancer, the HPC1 gene, Familial gastrointestinal stromal tumor, the KIT gene, along with many other syndromes and mutations [39]. Common to many syndromes are that they are caused by mutations in tumor suppressor genes. Here predisposition may manifest according to Knudsons [10] "two-hit" hypothesis, where mutation may be inherited in one allele of a gene of this kind, leaving a single allele to sustain normal gene function. This may be sufficient until inactivation of the second allele, for instance through accumulated carcinogen exposures, subsequently leading to the development of a tumor.

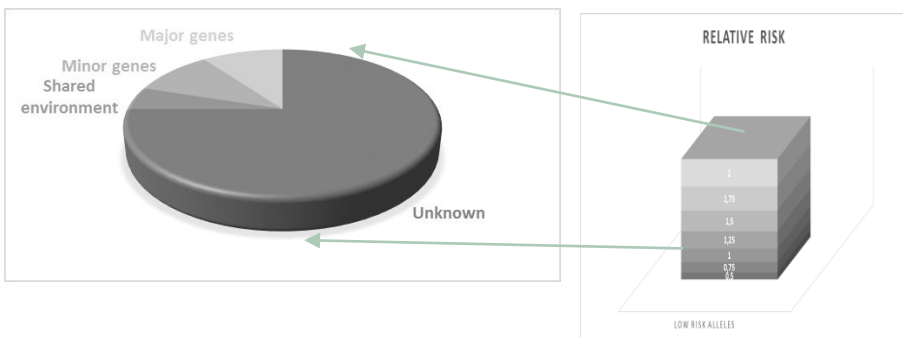
Over time, research focus into inheritance has focused mainly on identifying mutations segregating with outcome in large family pedigrees. Analyses have led to the identification of several of the highly penetrant genes mentioned above to underlay known pre-disposition syndromes. However, studies have shown that while these genes account for most large families with multiple cancer cases, for instance in breast, they only account for a small portion of two or three case families [40]. Moreover, it has also been shown that a large proportion of familial cancer aggregation is due to genetic inheritance [41] and that high penetrance mutations in known genes cannot account for most of this association, at least in the case of breast and colorectal cancers. The majority of familial aggregation of cancers remain unexplained, as shown in figure 2. While the unexplained familial clustering may be due to high penetrance mutations in yet unidentified genes, there has also been a hypothesis of a polygenic model where a large number of genes associated with low excess risks individually, combine to produce a range of susceptibility of different

degrees [40]. We see from table 1 that a relative risk (*RR*) of three, which is roughly what we typically see associated disease in a first degree relative, can be produced either by two to three high penetrance alleles, depending on whether they would work together multiplicatively or additively, or by 30 – 40 up to 300 – 400 genes, depending on their prevalence. Either way, identification of genes underlying this "missing heritability" is important, as it allows for identification, proper monitoring and prophylactic treatment in those under high risk.

Table 1 – Types of dominant alleles needed to produce a familial relative risk of a 1.8 (Adapted from Houlston et al.)

	RR	Frequency of allele in population	Number of alleles required to account for 3-fold familial risk
High- penetrance allele	20	0.001	2 – 3
Low- penetrance allele	1.5	0.01	294 – 400
Low-penetrance allele	1.5	0.1	37 – 50
Low- penetrance allele	1.5	0.3	33 – 44

Figur 2 - Schematic adapted from Houston et al. Major cancer genes: BRCA1/BRCA2, APC, MMR, minor genes: ATM, CHEK2, TP53, MYH, SMAD4, STK11 togheter with shared environmental exposure only account for a small portion of familial cancers.



Exposure to metals

It is not unreasonable to assume that many metals share common mechanisms and pathways for carcinogenicity, as they share important chemical characteristics. Some metals have established carcinogenic effects, such as arsenic, chromium and nickel. Arsenic exposure, for instance, has been associated with lung cancer through occupational and environmental studies, although experimental studies have remained inconclusive. Experimental studies, on the other hand, have clearly shown the carcinogenic potential of hexavalent chromium (the carcinogen responsible for the movie Erin Brockovich). Occupational studies of workers exposed to chromium

chemicals show elevated rates of lung and nasal cancer. Lead, beryllium and cadmium have all been implicated in experimental and/or epidemiological studies, however, results remain inconclusive. Limited findings also point to antimony and cobalt.[42]

There are several pathways for metal carcinogenicity. In experimental settings many metals induce DNA damage. They may also interfere with DNA repair processes. Some metals may affect cellular communication and homeostasis, as well as immune response. The pathways of carcinogenicity may differ also between metals according to metal type, solubility, metal-metal interaction among many other factors. [42]

While there is clear evidence of carcinogenic effects of some metals, important information is still lacking. Further studies are needed to settle issues of dose and response and multiple exposures of several metals in different scenarios which is only partially understood. Initial epidemiological studies in this field were of powerful carcinogens under highly elevated levels of exposure, where confounding bias induced by other exposures appeared unlikely. However, the current challenges call for studies of weaker carcinogens under lower levels of exposure, where it may be difficult to account for effects of individual metals. Improvements are also needed in exposure assessment techniques, as well as their incorporation into epidemiological designs. [42]

Sarcoma

Sarcoma, from the Greek “sarx”, meaning flesh, and “oma”, meaning growth, is the collective name of tumors that originate from transformed cells of mesenchymal origin, i.e. tumors of bone, cartilage, fat, muscle, vascular or hematopoietic tissues are considered sarcomas, as opposed to the more common tumors of epithelial tissue, termed carcinomas. And even though the former constitute the major part of the body tissues, sarcomas are exceedingly rare, barely making up 1% of cancer cases [43].

Soft tissue sarcoma

Sarcomas are subdivided into groups of sarcomas of the bone and of the soft tissues, where the latter is all but tumors of bone or cartilage. Soft tissue sarcomas (STS) make up about 0.6% of all incident cancer cases and 0.7% of cancer deaths. STS develop in any part of the body, although it has been reported to be most frequently in limbs. There are no known precursor lesions for STS, but there are benign forms that develop from the same tissues. These growths resemble normal tissue, do not invade locally and rarely reoccur after excision. In cytogenetic studies certain forms of STS have been shown not to originate from their benign counterpart, such as leiomyosarcomas and leiomyomas, as well as lipomas from liposarcomas. They usually develop as deep masses as opposed to superficial lesions. The expected survival of STS cases is low, approximately 67% in 5 years [9]. Factors associated with prognosis for disease-specific mortality include; age at diagnosis, duration of symptoms, tumor size and anatomical and compartmental location as well as having had radiotherapy [44]. The etiology of STS is poorly understood, but some risk factors have been identified.

Epidemiology

The epidemiological study of STS is complicated by the fact that they are rare and that histopathological classifications are inconsistent [9]. Furthermore, histopathological classification and cancer registry coding are also known to be inconsistent. The Surveillance, Epidemiology and End Result (SEER) registry and the International Association for Research on Cancer (IARC) classify STS according to both anatomical site and histological origin, something that has given rise to some confusion [9]. To make matters worse, there are more than 60 different histological subtypes. The most common are reported to be liposarcomas, leiomyosarcomas and fibrosarcomas, where the latter includes gastrointestinal

stromal tumor (GIST) and malignant fibrous histiocytoma (MFH) / Undifferentiated pleomorphic sarcoma (UPS). MFH/UPS, a diagnostic entity that in the WHO classification of tumors was later reclassified and subdivided into additional categories, of which one is Undifferentiated/Unclassified Sarcoma, a category that recognizes that some tumors simply cannot be classified due to demonstrable line of differentiation or lack of specific histologic, genetic or immunohistochemical characteristics [45]. A classification not used in the present study as data was collected before this reclassification was made.

STS is a heterogeneous group of tumors. Fibrosarcoma can occur in any anatomic site, MFH usually occur in the leg and dermafibrosarcoma occur on the trunk. Liposarcomas are usually large and found in thighs, retroperitoneum and inguinal region. Incidence appears to increase steadily with age. Rhabdomyosarcomas develop in skeletal muscle in the extremities, with peaks in incidence at five years and between 15 and 19 years of age, are most common during childhood, where it constitutes about half of all sarcomas and about 7% of all cancers. They are uncommon after the age of 45. Leiomyosarcomas develop in smooth muscle most frequently of the uterus or digestive tract. They are more common in women then in men and appear to peak during pregnancy [9].

Overall, the risk of adult STS increases with age and while most cases often appear sporadically without any known etiology, some risk factors have been identified.

Risk factors

Putative risk factors of STS are difficult to study due to the rarity of outcomes and the diversity of histological subgroups with potentially different etiologies. This brings with it consequent problems of both validity and statistical precision, with a subsequent lack of reproducible results. The problem is further exacerbated in environmental and occupational studies where accurate measurement of exposure history is difficult, and researchers have to rely on self-reported data with misclassification bias as a major consequence. Therefore, it is not surprising that there are only a handful established risk factors of STS. These include exposure to radiation and genetic susceptibility. Exposure to certain chemicals, such as phenoxy herbicides and chlorinated phenols, have also been implicated, but with conflicting evidential support [43,46]. With respect to additional risk factors, the literature is sparse.

Phenoxy herbicides, dioxins, and pesticides

Phenoxy herbicides, dioxins, and pesticides are probably the most studied risk factors in STS, and also the most controversial. Following initial reports from Sweden [47] of substantial effects of dioxin, or TCDD, an industrial contaminant, several occupational studies, as well as studies of accidental exposures, have

attempted to replicate the findings, with varying degrees of success. Out of approximately forty studies, fifteen showed a statistical significant elevated risk, while the remaining showed insignificant or lowered risks [33]. A complicating factor may be that contaminant exposures cannot be measured directly and has to be inferred through occupation or proximity to a site of an accident, which results in great uncertainty with respect to levels of exposure.

Dioxin, as several other chemical toxins, exerts its effect, at least in part, through binding to the Ah, or aryl hydrocarbon, receptor, also known as the "dioxin receptor" on the cells surface, and does not directly alter DNA and cause mutation. Its carcinogenic action is primarily, but not exclusively, in terms of tumor promotion [48]. Dioxin, and substances like it, are produced as unintentional byproducts of for example herbicides, wood preservatives, waste incineration and metal processing.

While there appears to be little concordance between results regarding the carcinogenicity of dioxins, varying from no effect to a 10-fold increase [9], it remains important to study this association due the toxicity of dioxins and the extent of population exposure. However, even though the World Health Organization (WHO) has already classified dioxin as human carcinogen, this association remains controversial [49].

Vinyl chloride

Vinyl chloride appears to be a carcinogen specific to the development of angiosarcomas. In repeated studies it has been shown that occupational exposure to vinyl chloride is associated with large excess risks, risks that have been clearly related to duration and time since first employment as well as cumulative exposure. [9]

Radiation

One of the few established risk factors of STS is exposure to radiation therapy. High-dose radiation is known to markedly increase the subsequent risk of STS, where tumors are usually secondary to therapy of a primary tumor. The risk appears especially high among those 55 years of age or younger [43] and specifically for angiosarcoma. Reported risks from a study of 194 798 women with invasive breast cancer [9] were *SIR* 26.2 (16.5 – 41.4) for angiosarcomas and 2.5 (1.8 – 3.5) for other STSs. A study using SEER registry data showed angiosarcoma to be responsible for 56.8% of radiation induced STS and MFH 15.9%. Risks were also elevated in those not exposed to radiation therapy, *SIR* 2.1 (1.1 – 4.4) for angiosarcoma and 1.3 (1.0 – 1.7) for other STS, suggesting that risks may be inflated to some extent due to confounding factors. One such factor may be low penetrant effects of inherited genes.

Immunosuppression

The idea of immunosuppression as an etiological factor in STS is not new, dioxins, which have long been associated with STS, have been reported to exhibit immunotoxic effects. In the last decade or two, studies of children with AIDS have shown excess risks of STS in general, *RR* 3.3 (2.6-4.1) and leiomyosarcoma in particular, *RR* 1900. [33]

Sex and growth hormones

As mentioned, hormonal factors, such as SH and GH, have been shown to play an important role in the development of common cancers. SH exposure levels have been linked to changes in risks and/or progression of prostate, breast, gynecological, and colorectal cancers [20,21,50-53]. Results have shown both elevated blood levels of endogenous estrogen, through the study of reproductive events [51,52], as well as exogenous estrogen and progesterone, through the study of OCs and HRT [52], to be associated with an increased risk of breast cancer. In contrast, studies have also shown exogenous estrogen and progesterone exposure to be protective against endometrial, ovarian, and colorectal cancers[20,21,53].

In sarcoma research, few studies have investigated this association, although some observations have been made. Regression of metastasis have been observed following oophorectomy [33], steroid increases risk of angiosarcoma of the liver and uterine leiomyosarcomas seem to increase from oral contraceptive use, OR 1.7 (0.7 – 4.1).[33] A hospital based case-control study reported risk increases for late age at first pregnancy and birth [54], but did not observe an association with menstrual cycle, parity, age at menopause or history of abortion. Moreover, incidence patterns coincide with reproductive events [55] and childhood rhabdomyosarcoma has been linked to sexual maturation and maternal still births [33].

Additionally, elevated levels IGF-1 have been linked to increased risks of prostate, breast, and colorectal cancers [37].

Since GH/IGF-1 and SH levels are important in the development of connective tissues as well as in carcinoma growth [37], it is not unlikely that they also play a role in sarcomagenesis. Indeed, pediatric sarcomas account for over 20% of solid tumors in children, with peaks around puberty, but only 1% of tumors in adults.

Furthermore, some STS subtypes have been found to have estrogen [55-57] and IGF-1 receptors [37,58], while IGF-1R is currently being targeted for therapy [59].

Body mass index

Another possible indication of hormones as a part of the etiology of common cancers is the increased risk in those that are overweight. Most cases involve breast, endometrial, kidney, or colon cancers. The most recognized biological mechanism thought to cause this observed association is related to the endocrine and metabolic effects of obesity [60]. Evidence suggests that insulin resistance is connected to

increased risks of colon and endometrial cancer. High insulin levels lead to decreased blood levels of IGF-1 binding protein 1 and 2, which in turn lead to an increased bioavailability of IGF-1. Moreover, elevated estrogen levels, due to superfluous amounts of adipose tissue, seem to be a mediating mechanism in the association between obesity and postmenopausal breast cancer as well as endometrial cancer.

Studies are limited, but STS risks have been reported to increase in those overweight, although results have varied. [9] In a population based study excess risks were observed for leiomyosarcoma OR 2.5 (1.1 – 5.7), carcinosarcoma 2.9 (1.3 – 6.7) and stromal sarcoma, 3.5 (1.1 – 10.9).

Heredity

Although heredity is an important factor for many tumor diseases, its part in the etiology of STS is not fully understood. While most STS patients do not have a clear family history of tumor disease, association studies have shown increased risks in connection to having a family member with cancer [61], and/or soft tissue tumors [62]. Although, the latter was not statistically significant, possibly due to the rarity of the disease. Moreover, an increased risk of STS has been observed in patients following diagnosis of a first STS [63].

Register studies of connections to other tumor types have pointed to some associations, indicating either common environmental or heritable genetic factors. Primarily in childhood, but also in adult cancer patients, a high incidence of secondary sarcomas have been reported. One study reported a SIR of 9.1 (2.4 – 20.2) [9]. It was irrespective of radiation treatment and most pronounced among women. In breast cancer probands, an increase in maternal sarcomas was detected. In a Swedish study stratified by histological subgroups, fibrosarcomas in parents were associated with endocrine gland and stomach cancers, and parent breast cancers was linked to leiomyosarcomas. In an additional study, MFH/UPS was associated with renal carcinomas.

To explain parts of these associations, there are heritable predisposition syndromes that manifest as STS. The most well-known is the Li–Fraumeni syndrome (LFS), a rare familial cancer syndrome. As can be seen from the diagnostic criteria in table 2, affected family members with LFS are predisposed for early-onset cancers of the breast, brain, adrenal gland, leukemia and for sarcomas of the skeleton and the soft tissues.

Table 2 – Li-Fraumeni syndrome diagnostic criteria based on outcomes in proband, first- and second degree relatives (FDR) and (SDR), repsectively.

Criteria	Individual	Outcome	At age
Classic LFS	Proband	Sarcoma	< 45
	FDR	Any cancer	< 45
	FDR or SDR	Any cancer	< 45
		or sarcoma	Any age
Chrompet Criterion I	Proband	LFS-spectrum tumor	< 46
	FDR or SDR	LFS-spectrum tumor (not breast)	< 56
		or multiple LFS-spectrum tumors (not breast)	Any age
Chrompet Criterion II	Proband	Multiple LFS-spectrum tumors (except ≥ 2 breast)	First < 46
Chrompet Criterion III	Proband	Adrenal cortical carcinoma or tumor in the choroid plexus.	
Birch LFL	Proband	Childhood cancer, sarcoma, brain tumor or adrenal cortical carcinoma	< 45
	FDR or SDR	LFS-spectrum	Any
	FDR or SDR	Any cancer	< 60
Eeles LFL	2 FDR or 2 SDR	LFS-spectrum	Any

The syndrome has often been associated with a germ line mutation in the p53 gene, a mutation found in 50-80% of affected families [64,65]. The p53 gene is an important tumor suppressor, also termed "guardian of the genome", which is mutated in approximately 50% of all cancers. It regulates net cell growth, either by a reduction of cell births or by promoting apoptosis (cell death by suicide)[66]. Germ line p53 mutations have been shown to contributed to approximately 4% of STS [9]

However, recent studies have shown this mutation to be associated with a broader range of cancer sites and age-of-onset than suggested by the LFS criteria, as defined in table 2 [65,67-69]. This may indicate that heredity represents a larger part of STS etiology then is currently known. In fact, it is estimated that p53 germ line mutations may be responsible for 15-20% of all inherited cancers [70]. Other mutations have been identified in a small portion of LFS families, though they do not appear to explain the wide range of clinical phenotypes of LFS [65].

Other heritable syndromes involving STS include hereditary retinoblastoma and neurofibromatosis. In hereditary retinoblastoma, caused by a mutation in the RB-1 tumor suppressor gene, in addition to retinoblastoma, there is an increased risk of secondary sarcomas. Interestingly, mutations in both Rb1 and p53 genes have been observed in a number of tumors possibly indicating the need for inactivation of

multiple tumor suppressor genes in tumor development. Neurofibromatosis type 1 and 2, are caused by mutations in the NF1 and 2 genes, respectively. All functional aspects of NF1 gene product are currently not known , it appears to have multiple functions in different tissues [71]. However, NF1 reportedly takes part in processes that promote cell growth and differentiation and induces an increased lifetime risk of sarcoma, usually neuro- or fibrosarcomas, of 7-14%. [33]. NF2 is a tumor suppressor gene, involved in the contact dependent inhibition of cellular proliferation. Those with neurofibromatosis type 2 is characterized by the occurrence of vestibular schwannomas (acoustic neuromas), which are benign cranial nerve tumors [72].

Given the rarity of known predisposition syndromes and the magnitude of observed familial associations, it seems clear that there is more still to uncover of heredity in STS.

Tissue trauma and repair

In relation to the hip and knee prostheses studies, some research has also been made into trauma related tissue damage and its possible promotion of STS genesis. A case study has reported ongoing chronic inflammation with the presence of suture materials in cells being observed in a tumor located in a surgical wound [73]. And although the process of malignant transformation may not be well understood, the additional observation of soft tissue tumors in surgical wounds, following traumatic injury and in burn scars, questions the balance with wound healing [73-88].

A common explanation for some of these observations appear to be that physical injury at a site of a tumor attracts attention to that site and increases the probability of the tumor being detected [89]. While this explanation is viable, there is mounting evidence of inflammation as both an initiator and promotor of tumorigenesis, as well as of tumor progression [14]. And as the physiological response of tissue repair and regeneration following injury is mediated through inflammation it is not unlikely that physical injury may cause and/or contribute to tumor initiation, promotion and/or progression, not least when the consequent inflammation becomes chronic. Moreover, additional stages of wound healing share biological pathways with tumor development, further pointing to the wound microenvironment as a stimulator of tumor cell growth [90]. Recently, tissue injury have been shown to promote sarcoma formation in animal models [91] in addition to previous models showing wounds as possible tumor growth promotors [92]. It has furthermore been suggested that the risk of STS is increased in chronic repair processes, with one study showing an association with a reduced DNA repair process. [33]

In summary

The only risk factor with a firm connection to STS to date are the risks of secondary STSs subsequent to radiotherapy and some rare heritable cancer syndromes. There have been conflicting evidence for an association with occupation, herbicide and

chlorophenol exposure as well as with place of residence in connection to industrial emissions [43]. There have been insufficient evidence for conclusions about effects of menstrual and reproductive factors [54] and no evidence of associations with DDT or asbestos. There is insufficient information to draw conclusions regarding birth weight, maternal age, pregnancy medications, pregnancy conditions, history of infections and tobacco, alcohol and drug use [43].

Consequently, we still know very little about what causes STS and what to do to prevent it. The need for identifying factors that account for STS risks is apparent, both to yield missing etiological clues and, further down the line, potential prevention targets.

Cancer and joint prosthesis of the lower extremities

Approximately 13 000 patients are fitted with a knee prosthesis every year in Sweden alone [93]. The number has steadily increased since they were introduced six decades ago. The number of operations has even surpassed the number of hip arthroplasties in the high-income countries [94]. In addition, improved long-term prosthetic survival coupled to high patient satisfaction has led to younger patients being operated. It has, therefore, become essential to address the concern raised by several researchers that degradation of prostheses may increase the risk of developing cancer.

The concern is that wear and corrosion releases significant amounts of polymer and metal particles into the tissues, lymph nodes and lungs of knee prosthesis patients. Serum levels of cobalt and chromium are normally up to five times that of the average person and up to 50–300 times greater during prosthesis failure [95]. Moreover, prosthesis materials have been shown to cause cancer in animal and epidemiological studies [96,97]. However, if this contributes to increased cancer risks in prosthesis patients is not known. The evidence has been conflicting.

Today, patient risks are mainly monitored by means of cross checking hospital data with national cancer registries. The organs and systems most likely to be affected, at least in the ten to twenty year time span available in these data, is likely to be hematopoietic systems, the urogenital system and skin. In the long term, twenty to forty years, were data is sparse, we are more likely to see the putative effects on solid tumors.

To date, these studies have repeatedly shown risk increases in hematopoietic tumors, for prostate cancer and melanoma, along with decreased risks for gastro intestinal and airway cancers [98-100]. Other risks have been fluctuating. However, many of these studies have not been stratified by indicating diagnosis, and because it is known that rheumatoid arthritis (RA) patients have an already elevated risk of,

for instance, hematopoietic cancers, it may be difficult to disentangle surgery and prosthesis related risks from those of the underlying disease.

Cancer and inflammatory disease

As was previously mentioned, much research have been made into the connection between inflammation and tumor disease. With respect to inflammatory musculoskeletal conditions and, most of all, lymphatic tumors, several examples are associated with increased cancer risks. These include RA, Sjögren's syndrome and SLE. In RA patients, the risk of lymphoma has additionally been observed to decline by use of anti-inflammatory treatment. Also of interest is that leukemia and lymphoma have been diagnosed in patients with osteonecrosis, although the inflammatory processes associated with this condition are not fully understood. The subsequent question is whether osteoarthritis (OA), which is the leading diagnosis in knee arthroplasty patients, might also drive malignancy, with or without an implant [101]. We have suggested that factors of pre-existing disease that drive malignancy is further enhanced by metal ion exposure following implementation [101] and promote development of for instance hematopoietic malignancies.

Soft tissue sarcoma and joint prosthesis

One study has shown prosthesis debris to induce sarcoma in rats [96]. Soft tissue sarcomas have also been observed in connection to prostheses of the hip in case reports[1,2], but was absent when collecting information in a larger register, in one instance [2], although not when considering metal-on-metal articulation [74]. Because sarcomas are rare, even in large prosthesis register cohorts, the question whether prosthesis debris may drive sarcomagenesis remains difficult to investigate.

However, acquiring further knowledge of risk factors of STS may help elucidate not only its association with joint prostheses, but additionally shed light on potential associations with more common cancers, as etiological factors may be shared.

Aims

The present thesis has three general aims:

- Identify important risk factors of soft tissue sarcomas to help to identify etiological factors that could have a bearing on preventive work.
- Evaluate cancer risks associated with knee prosthesis surgery.
- To determine whether our first aim will improve the understanding of findings by our second aim.

The specific aims were

- To evaluate the influence of stature and growth, as well as reproductive factors on STS risk. (**Paper I**)

- To evaluate the influence of familial STS and cancer on STS risk, as well as the effect of sex and OC-use on this relationship. To further determine whether putative associations are due to inherited genetics and to identify links to other cancer types and their potential pathways. (**Paper II**)

- To develop a method for determining how much of the odds ratio (*OR*) between STS and cancer in a sibling, that is due to inherited genetics, and to extend this method to a case control setting. That is, to adjust this *OR* for environmental factors shared among family members. (**Paper III**)

- To evaluate the influence of tissue trauma on the risk of STS. (**Paper IV**)

- To determine if having knee prosthesis surgery is associated with increased risks of cancer. If so, what types of cancer and what can we say about putative causes? (**Paper V**)

Study design and data collection

What follows is a summary of study design, data sources and collection procedures used in the present thesis. These include the main studies of each paper, but also validation cohorts were important results are replicated.

Data sources

The Swedish Cancer Register

With the objective of producing health care statistics for quality assessment as well as for research purposes, the Swedish National board of Health and Welfare, since the 1958, maintains a register of all of malignant, and certain types of benign, cases of tumor disease, simply referred to as the Swedish Cancer Register (SCR) [102]. It is based on compulsory reporting by clinicians and pathologists working for Swedish healthcare providers. The register contains information concerning tumor site, histological type, basis for, and date of, diagnosis and follow-up data, such as date and cause of death as well as date of migration. The register is reportedly almost complete, containing 96.3% of all cases, although this figure varies with tumor type and patient age. It was worst for soft tissues, nervous system, leukemia and lymphoma and worse in those older than 70 year of age than in those younger.

The Swedish Population Register

The Swedish population register is a register of all people currently living in Sweden, maintained by the Swedish tax agency [103]. It contains names, addresses, Swedish personal identity numbers, places of birth, citizenships, spouses, children, parents, legal guardians, adoptions, migration in and out of the country, addresses abroad and deaths and burial sites of all those registered. Information also includes dates of these events.

SPAR is a public register that contains all persons registered in the population register, with the same information, in addition to earnings and property owned.

The MISS Study

A study cohort originating from the department of Cancer Epidemiology used for studies of different exposures that affect women's health, measured through a variety of outcomes. The study was initiated in 1990 as a prospective study of malignant melanoma risk factors and includes a thousand native Swedish women in each 1-year age group between 25 and 65. In total 40000, randomly selected from SPAR and checked against the cancer register for no prior malignancies, were invited. Women were asked to fill out a standardized questionnaire concerning melanoma risk factors at inclusion and then again for a follow-up in the years 2000 to 2002. Questions pertained to several areas of life style factors, including parity, family history, physical exercise, smoking habits, alcohol consumption, use of combined oral contraceptives, age at menopause, educational level and stature. [104].

The Swedish Knee Arthroplasty Register

In the mid 1970:ies, the surgeons of the Swedish Orthopedic Association realized that each operating center alone could not gain sufficient experience to critically evaluate all emerging implants models and surgical procedures to allow for choosing optimal combinations. They therefore initiated a nationwide multicenter study with the primary purpose of warning against suboptimal techniques and implants. As a result, Sweden currently has the lowest prosthesis revision rate in the world [93].The register started in 1975 and is still ongoing today, making it the world's oldest nationwide knee arthroplasty registry, presently covering all 74 orthopedic clinics that routinely perform knee arthroplasties in Sweden. Reporting is done by means of a one page questionnaire filled out at the time of surgery in the operating theater, collecting information on patient history, prosthesis model and surgery [93]. The questionnaire is then sent to the registry data entry office, where data is validated and entered by registry personnel. It presently contains 96.6% present of all surgeries. Validation is done by means of monitoring visits to participating units by registry personnel. [3,93]

Study design

Study I-IV Soft tissue sarcomas

STS is a heterogeneous group of rare diseases. In studying them, it is therefore inefficient to use the most basic of study designs; the cohort study. This is where

one defines a patient population, the cohort, at a given point in time and simply follows them prospectively through time to observe any outcome that participants may experience [105]. In STS, for the cohort to be large enough to attain sufficient statistical precision, it would likely have to contain a large portion of the population of Sweden. However, to collect information on all these people is of course both financially and logistically infeasible, and consequently, we are forced to adopt a different approach.

The case-control design

The case-control design was seemingly pioneered by Lane-Calypon in 1926 [106] in an investigation into the association between reproductive events and breast cancer. Thereafter, the design was sparsely used, until the 1950ies, when researchers were investigating the connection between smoking and lung cancer. Among the studies produced during this time, the most influential in terms of the modern approach to case-control designs, is likely Doll and Hill, 1952 [107]. In terms of theory, an important contribution was made by Cornhill, who demonstrated the crucial symmetry property of the parameter of main interest in the case-control study, the *OR*. This ultimately increased the relevance and understanding of the case-control study, as it connected its result to that of other study designs. And of course, one cannot discuss the origins of the case-control study without mentioning the contributions of Mantel and Haenzel, who clarified the objectives of this design, systemized their use and presented two now established approaches to their statistical analysis [106].

The case-control study is well suited for studying diseases with long induction periods, such as cancer, as it allows the investigator to look back through extensive periods of time. Periods of time for which maintaining a cohort study would be infeasible. This may help explain some of its past and current popularity [106] in this research area. Most of our current knowledge of carcinogenic exposures are due to case-control designs. And even though it has been widely criticized in the past, some of it fair and some of it due to misconception, to put it in the words of the late great statistician and epidemiologist Norman Breslow: "What would we replace it with?". It is simply paramount to medical sciences.

The basic idea behind the case-control study is simple. In its most basic form, where one is investigating the effect of the presence/absence of a risk factor on a given disease, one collects, during a given period of time, all cases of the disease and estimate the prevalence of the risk factor among these cases, p_{ca}. This prevalence is then compared to the corresponding prevalence in a set of "controls", p_{co}, by calculating the relative odds of exposure between the two, the exposure *OR*

$$OR = \frac{\frac{p_{ca}}{1 - p_{ca}}}{\frac{p_{co}}{1 - p_{co}}}$$

Here, the controls are a population free from the given disease, often chosen to be similar to the cases with respect to certain characteristics. This is usually done in order to limit confounding bias and random variation in the result. The similarity can be achieved in several ways. One common way is through "matching". Here, one selects one or more controls per case based on the characteristics of the case. The process helps to balance characteristics between cases and controls, so that investigators can be reassured that any difference detected between them cannot be due to differences in the distribution of the characteristics, i.e. it helps to avoid confounding effects of important characteristics, such as sex and age, with that of exposure.

The *OR*, because of its symmetry property, as pointed out by Cornwall, when disease prevalence is low, or under certain sampling schemes, corresponds to the *RR* of disease given exposure status. Due to the works of Mantel, Haenzel, Breslow and others [106], one can then test the statistical significance of this association. Consequently, one may estimate and test the same parameter in a case-control study as is of main interest in a cohort study, but with much less effort.

The main draw back in an otherwise brilliant design is in terms of bias. The case-control study, as opposed to the cohort study, is retrospective. This induces issues of recall bias in the data collection process, when, for instance, using self-reported exposure data, as study participants may not recall their entire exposure history in sufficient detail. Other issues include selection bias, which may occur when the sampling frame of cases and controls do not agree and/or does not correspond to the intended source population. The retrospective design can also cause confounding of temporal ordering of events. Effects of these types of bias can usually be contrasted using a prospective study design.

Choice and of controls and sampling method

In a case-control study, cases are to be selected to represent the cases in a well-defined population. Controls are to be chosen as to represent the disease free individuals of that same population. This can be done in several ways, and some different considerations weigh into the way controls are collected. These include what effect measure one wishes to estimate along with considerations of validity and statistical precision. Depending on the way controls are sampled, the estimated *OR* can approximate different effect measures [108]. In the present project, as controls are sampled at the same point in time as cases are diagnosed, which is referred to as incidence density sampling, *OR*s can be most closely described to estimate the incidence rate ratio (*IRR*) [108]. Considering validity and precision, controls are selected by matching, in order to balance potential confounders between cases and control. In doing so one wants to be mindful not to introduce bias by matching on the wrong variables. While matching on confounders increases validity and matching on strong determinants of disease increases precision, matching on factors on the causal path between exposure and outcome, the mediators, may introduce

bias and reduce validity. Additionally, matching on variables strongly associated with exposure, may reduce precision, as cases and controls become too similar and the contrast of exposure, which is the basis for statistical precision, is lost. A problem referred to as overmatching. In the present study, in order to evaluate the potential effect of overmatching, two sets of controls are collected with different degrees of geographical matching; parish, and region. The sets contain two controls each.

Collecting information on cases and controls

The present study is set in the south of Sweden, from which all STS cases 15 years of age or older, except for those with uterine tumors, were collected through the Southern Swedish Regional Tumor Registry, a regional node of the national cancer register. Once reported to the register, a questionnaire was sent to the case, after consulting with the treating physician. Controls were sent the questionnaire at the same time, after being selected from the SPAR register matched for sex, age, birth year and parish or region of residence and checked against the SCR to be without previous tumor disease. For those who failed to respond, a new control was selected. This process was repeated until the selected control responded or selected controls started to repeat. No control was sent the questionnaire a second time and matched to a second case. All control questionnaires used corresponded to unique individuals. This produced a number of cases at the end of the study without a matched control. The use of a population based register as sampling frame for the controls limited the risk of selection issues with respect to controls. Further details of the study is given in paper I and a copy of the questionnaire can be found in the appendix of this thesis.

Flow chart

To illustrate the different stages of participant selection into the studies I, II and IV of the thesis, we present a flow chart in figure 3. Here, region- and parish-matched controls are presented in individual sections. On a separate line in each box of the case section, the number of cases matched to at least one region-, parish-, or either region or parish, control is given. Abbreviations correspond to; Ext – extremities, Retr/Visc – Retroperitoneal/Visceral and Not spec – No location specified. Below we also provide a table describing what subpopulations are used for the analyses in each of the studies I through IV. The different papers and subsequent analyses contain different subpopulations. These are of varying sizes in the interest of striking a balance between validity and statistical precision. The definition and size of each subpopulation are described in table 3.

Table 3 – Description of the different subpopulations, used for the analyses in papers I-IV, derived from the original case-control study desrcibed in figure 3.

Study	Group	Analysis	Matching	Nbr cases	Nbr controls	Control type
I	All	Main analysis of exposure OR	Matched	634	1251	Region
	All	OR sensitivity analysis including all collected data	Unmatched	855	2021	Region
II	All	Main analysis of exposure OR	Matched	629	1231	Parish
	Female NS STS	Heritability in case families	-	91 ǂ ‡ (321)*	-	-
	Female STS	Heritability in case families	-	283 ǂ (938)*	-	-
	Male MFH/UPS ≤ 65	Heritability in case families	-	40 ǂ (163)*	-	-
	Male STS	Heritability in case families	-	341 ǂ (1521)*	-	-
	All	Heritability in case families	-	637 ǂ (2833)*	-	-
III	Female NS STS	Heritability OR	Unmatched	139 ‡ (546)*	183 (718)	Parish
IV	STSE	Main analysis of exposure OR	Matched	249	491	Parish
	All controls	Estimate propensity scores for adjusting ORs	-		2066+2021	Both
	STSE	Location specific exposure OR	Unmatched	253	491+497	Both
	All	Sensitivity analysis to detect recall bias	Unmatched	855	2066+2021	Both
	STSE	Survival in exposed vs unexposed	Unmatched	316	-	-
	STSE	Sensitivity analysis to evaluate detection bias using Survival analysis	Matched	249	491	Parish

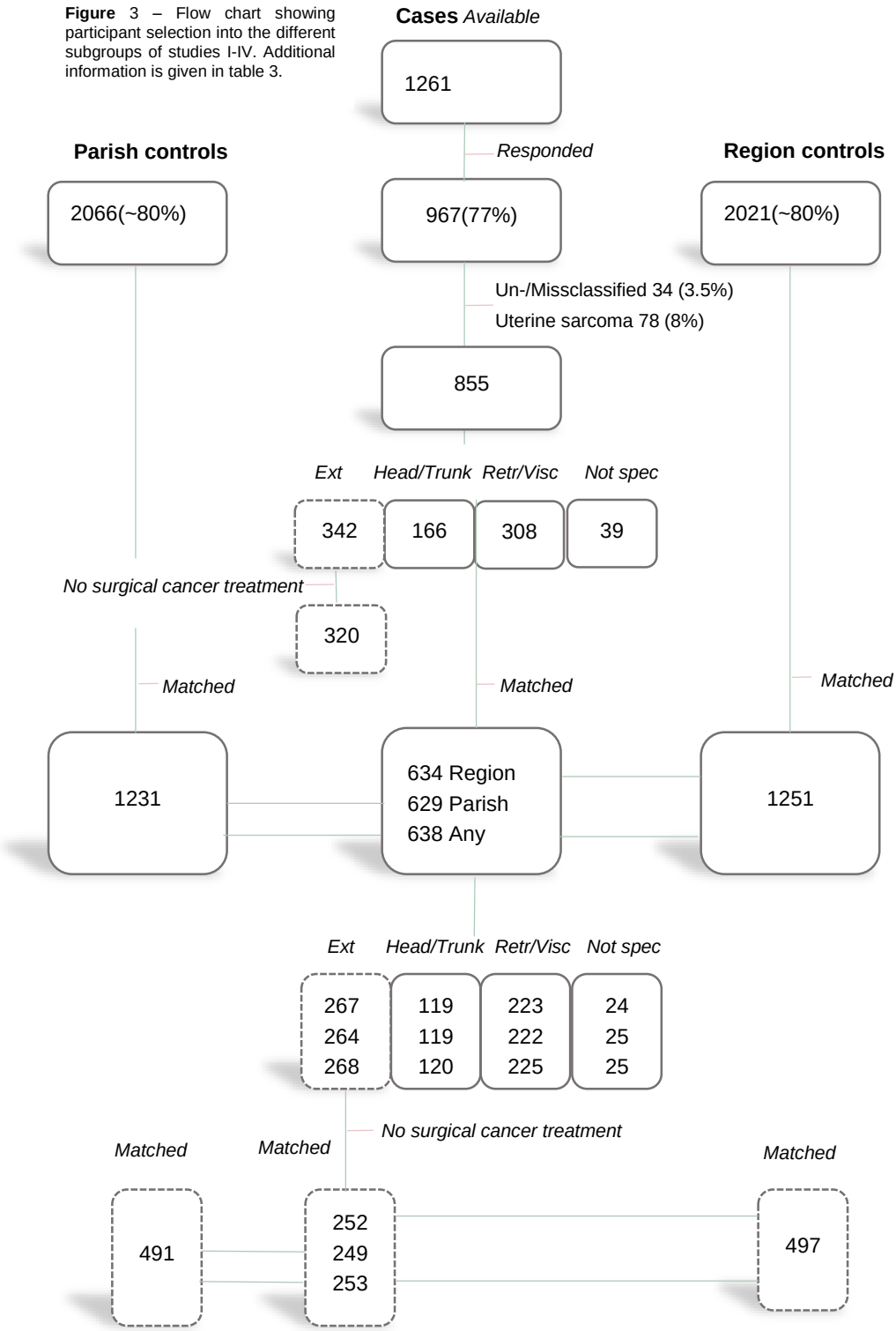
Table 3 – Explanations

** Number of proband family members. Includes only parents and sisters for female cases.*

ǂ Only includes cases collected during the same time period as the controls of the main analysis. GISTs are excluded.

‡ Contain different number of study particiants as the estimation of the heredity OR is based on interaction analyses that demand a larger study size. The latter analysis also allows for adjusting for covariates, making it possible to include cases and controls collected during different time periods. Consequently, there are not two controls per case in this analysis.

Figure 3 – Flow chart showing participant selection into the different subgroups of studies I-IV. Additional information is given in table 3.



Study V - Knee prosthesis

The cohort design

The cohort study is the most basic and straightforward of the epidemiological designs. It consists of defining a group of study participants at a point in time and subsequently following them with respect to outcome for an extended period of time. This can be done in a prospective manor, defining the cohort in present time and following it into the future. It can also be done by identification of a "historical cohort", determining a historical time of initiation of the study and following them through to present time. Study V of the present thesis is of this latter type. Contrasting this design with the case-control design, it is clear that the prospective cohort study demands more effort, not least in terms of funds, logistics and time invested by both investigators and study participants. On the other hand, it offers a wider picture of the effects of exposure because of the ability to study multiple outcomes, as opposed to the case-control study, that can only focus on a single outcome. Therefore, it has been suggested that the case-control study may be the method of choice for investigating causes of a specific disease and cohort studies the choice when investigating health effects of given exposure. It is also known to avoid some typical biases that plague case-control studies [105].

The primary focus of a cohort study is the outcome incidence rate, I. This entails computing two things; the number of incident outcomes, e, and sum of follow-up time, p, for members of the cohort. The estimate of interest is then simply $I=e/p$. The estimation of these two entities in study V is straight forward, by cross-checking the SKAR register with the Swedish Cancer register, counting the number primary tumors diagnosed and years passed, from the date of knee prosthesis surgery to the date of death, migration or end of the study. The dates of death and migration were collected from the SPAR register. In Study V, we study standardized incidence rate ratios (SIRs) for different types of tumor disease. The definition of this measure is given in the chapter that follow.

Epidemiological measures and bias

What follows is a short summary of the epidemiological measures used in the present thesis, their motivation, standard methods of analysis, issues to consider when applying them and the most common types of bias that affect their use. Possible solutions to some of these issues are also provided. At the end of the chapter a table is presented overseeing what measures are used in each papers.

The odds ratio

As mentioned previously, the analysis of a case-control study is focused on the *OR*. We also stated that, in the present study, this *OR* approximates the incidence rate ratio, *IRR*. To give an intuitive idea as to why this is, we will give an example of a pair-matched study (matching one control to each case) of similar design, based on an example due to Greenland [109].

Let us assume that P_r is the proportion of exposed, and N the number of study participants left, in the study at time T. Further assume that incidence rates are I_e and I_{ne} in the exposed and non-exposed groups, respectively. We would then expect the distribution of case-control pairs, concordant/discordant for exposure, to be

Controls	Cases		
		Exposed	Unexposed
	Exposed	$NTI_eP_rP_r$	$NTI_{ne}(1-P_r)P_r$
	Unexposed	$NTI_eP_r(1-P_r)$	$NTI_{ie}(1-P_r)(1-P_r)$

Now, in a matched case-control study, the *OR* is estimated by the ratio of the cells corresponding to pairs discordant for exposure. If we were to do this here, with the current design, we would get

$$\frac{NTI_eP_r(1-P_r)}{NTI_{ne}P_r(1-P_r)} = \frac{I_e}{I_{ne}} = IRR$$

that is; when matching on study time *T*, the *OR* equals the *IRR*. As a side note, this estimation procedure also explains the phenomenon of overmatching, as matching on variables predictive of exposure makes matched pairs similar with respect to exposure and leaves few discordant pairs and little information about the *OR*/*IRR*.

While one can perform the analysis of a matched-pairs case-controls study this way, for instance by using McNemar's test, one is frequently in need of adding more controls to increase statistical precision or to adjust for additional covariates and confounder to increase validity. In order to do so, a common approach is to use logistic regression. If the study is matched, one has to condition on the matched groups using what is referred to as conditional logistic regression (CLR). In the example given above, with

$$P\left[Y_{ij}|X_{ij}\right] = \frac{e^{\alpha_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{X}_{ij}\boldsymbol{\beta}}}$$

for logistic regression, with $\boldsymbol{\beta} = \beta_1, .., \beta_n$ as a vector of parameters with , $\beta_k = \log(OR_k)$ and OR_k is the *OR* corresponding to the effect of covariate *k*, and $\boldsymbol{X}_{ij}$ as a covariate vector of the case, as indicated by j=1, or control, j=2, of pair i respectively, this amounts to

$$P[Y_{i1} = 1, Y_{i2} = 0 \,|X_{i1}, X_{i2}, Y_{i1} + Y_{i2} = 1] =$$

$$= \frac{P[Y_{i1} = 1|X_{i1}]P[Y_{i2} = 0|X_{i1}]}{P[Y_{i1} = 1|X_{i1}]P[Y_{i2} = 0|X_{i1}] + P[Y_{i1} = 0|X_{i1}]P[Y_{i2} = 1|X_{i1}]}$$

$$= \frac{\frac{e^{\alpha_i + \boldsymbol{X}_{i1}\boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{X}_{i1}\boldsymbol{\beta}}} \frac{1}{1 + e^{\alpha_i + \boldsymbol{X}_{i2}\boldsymbol{\beta}}}}{\frac{e^{\alpha_i + \boldsymbol{X}_{i1}\boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{X}_{i1}\boldsymbol{\beta}}} \frac{1}{1 + e^{\alpha_i + \boldsymbol{X}_{i2}\boldsymbol{\beta}}} + \frac{1}{1 + e^{\alpha_i + \boldsymbol{X}_{i1}\boldsymbol{\beta}}} \frac{e^{\alpha_i + \boldsymbol{X}_{i2}\boldsymbol{\beta}}}{1 + e^{\alpha_i + \boldsymbol{X}_{i2}\boldsymbol{\beta}}}}$$

$$= \frac{e^{\boldsymbol{X}_{i1}\boldsymbol{\beta}}}{e^{\boldsymbol{X}_{i1}\boldsymbol{\beta}} + e^{\boldsymbol{X}_{i2}\boldsymbol{\beta}}} \qquad (1)$$

when conditioning on outcome and applying the logistic model above. Standard estimation of $\beta_k = \log(OR_k)$ for the CLR model then proceeds by calculating one term per pair i, as defined above, and multiplying them together for all pairs in the study, to yield what is referred to as the likelihood function. The estimates of the effects OR_k for all covariates are then produced by finding the vector $\boldsymbol{\beta}$ that maximizes this

likelihood function. This is what is referred to as maximum likelihood estimation [110].

Although under certain conditions, such as when the effect of matching variables can be modelled statistically, unconditional logistic regression can be used as an alternative to CLR. It may even be preferred in certain scenarios to counteract potential effects of overmatching [111].

Homogeneity of the odds ratio - subgroup analysis

The standard assumption in both experimental and observational studies is that the effect of exposure, or treatment, is equal over all subgroups. This is usually not a realistic assumption, for example considering responders/non-responders to medical treatments, as well as the occurrence of treatment side effects only in some, but it is usually adopted due to the difficulties that arise otherwise. Subgroup analyses demand large study sizes, are inherently difficult to conduct and interpretation is riddled with pitfalls [112]. When one still opts for these analyses, two important aspects that are usually stressed in order to avoid common pitfalls, are that analyses should be planned a priori and that differences should be analyzed using statistical tests. We consider two different types of subgroup analyses important to the current project, conducted using the same statistical approach.

Heterogeneity of effects between different outcomes

One important aspect of the analyses of STS-studies are the possibility that sarcomas that originate from different tissues and locations may have different etiologies and risk factors associated with them, and therefore, for any given risk factors, the effect on various STS subtypes may be different.

Intuitively, when handling this question statistically, we view each set of cases of a specific subtype, together with their corresponding controls, as a separate study. We then test the difference in observed effects between studies statistically by creating an indicator variable for each study and then pooling all study data in the same statistical model. Differences in effect are then tested using the interaction terms between the exposure and study indicator variables in a CLR-model according to

$$L(\boldsymbol{\beta}) = \prod L_k(\boldsymbol{\beta}) = \prod \prod \frac{e^{\beta_e x_{i1k} + \boldsymbol{\beta} \boldsymbol{I}_k x_{i1k}}}{e^{\beta_e x_{i1k} \boldsymbol{\beta} \boldsymbol{I}_k x_{i1k}} + e^{\beta_e x_{i2k} + \boldsymbol{\beta} \boldsymbol{I}_k x_{i2k}} + e^{\beta_e x_{i3k} + \boldsymbol{\beta} \boldsymbol{I}_k x_{i3k}}}$$

where x_{ijk}, is an exposure indicator for subject j, where j=1 indicates a cases and $j \in \{2,3\}$ indicates controls, in matched group i, and study k and $\boldsymbol{I}_k$ is an indicator vector of length corresponding to the number of studies minus one, with zeros in all places

except for number k. β_e is the effect of exposure, in terms of $log(OR)$, for the study selected as the reference, and $\boldsymbol{\beta}$ is a vector of the same length as $\boldsymbol{I}_k$, containing the differences in effects between remaining studies k and the reference study, in terms of $log(OR_k)$- $log(OR)$. The likelihood is then tested against a likelihood where the study indicator is removed. A significant result is taken as evidence that effects of exposure differ between histological subtypes.

Heterogeneity of effects between different subgroups – effect modification

We also investigate differences in effect of exposure between different subgroups, based on for instance, sex and age. This is done equivalently, by simply replacing the study indicator $\boldsymbol{I}_k$ by a group indicator and selecting a suitable reference group.

Population attributable fraction

The population attributable risk (PAR) is defined as the proportion by which the incidence rate of an outcome in a given population could be reduced if the exposure of interest was eliminated [113]. That is

$$PAR = \frac{I_p - I_e}{I_p}$$

where I_p is the population outcome incidence rate and I_e the incidence rate of the exposed group. We see that the PAR is a simple measure of the potential benefit of a population intervention targeting exposure, e. This can be re-written as

$$PAR = \frac{P_e(RR - 1)}{1 + P_e(RR - 1)}$$

to make explicit its dependence on exposure prevalence, P_e, and associated RR. Here, the RR is often approximated using the OR from, for instance, a case- control study.

Heritability

In the search for inherited genetic causes of disease researchers often study familial aggregation of outcomes. Significant aggregation is often considered a prerequisite for establishing presence of important genetic effects [114]. Although, we note that this largely ignores inheritance due to rare genetic mutations, as can be seen following the reasoning put forth earlier in this thesis. When familial aggregation is established, a natural next step is to determine the pattern of clustering to disentangle whether the association is due to inherited genetic or shared environmental factors, or possibly both. This is mainly done in two different ways [114]; through regression models, estimating parameters corresponding to aggregation specific to certain familial relationship pairs, or through multivariate models, including variance components and path analysis formulations, out of which the ACE model is one of the more well known. The latter is used to partition outcome variance into parts due to additive genetic, shared and individual environmental effects, which can be used to estimate the heritability of outcome. The proportion of variance due to additive genetic effects is sometimes referred to as the narrow sense heritability (h^2) [115].

In the present work, we identify familial aggregation by estimating the *OR* of having a first degree relative (FDR) with cancer using a logistic regression model. We then study the pattern of familial clustering using a multivariate ACE model. However, the ACE model to estimate heritability is applied in a somewhat different manor than what is commonly done. In order to gauge to what part the familial aggregation observed in our study is due to predisposition genes shared between STS and common cancers, we estimate the heritability h^2 of cancer in families of STS cases and compare it to that of the general population. While this is not a rigorous approach, and may in fact boarder on a methodological flaw if attempting to apply it in a general setting, we reason that shared genes may be rare in the general population, and comparatively common among cases. In this scenario, under certain assumptions about gene prevalence and effect, the genetic variance in case families would exceed that of the general public. Therefore, a high narrow sense heritability, compared to that of cancer in the general population, may by a sign that the association is, at least in part, due to shared predisposition genes.

The problems with this approach are several. For one, in a general setting, if our genes and outcome do not fulfill the assumptions about prevalence and effect, the heritability results may be difficult to interpret. If predisposition genes were to be common among STS families, then gene variance, in terms of presence/absence of relevant disease alleles, would be low, and consequently genetic variance and heritability, would be low. That is, in theory, the familial association could be genetic in nature, but the heritability in case families could be lower than that of the general population. This could also be true if the shared genetic mutations were protective for STS. Therefore, it would not be possible to differentiate between these

two situations in a general setting, without further assumptions. Cancer prevalence in case families may also vary due to additional characteristics that set them apart from the general population. One such characteristic may be age, considering the average age of adult STS probands, another could be race.

Heritability odds ratio

In essence, the mentioned issues with the above approach is due mainly to the fact that it was applied to case families only, and consequently, does not consider the difference in cancer prevalence between case and control families. It is done this way because the ACE-model does not extend easily to the case-control setting, and especially not to our specific problem of investigating the association between STS and familial cancers. The problem would have been more tractable had we been investigating the same outcome in both probands and family members [114].

The *OR* of having an FDR with cancer, that we started with to study familial aggregation, does exactly this, compares the cancer prevalence between cases and controls. On the other hand, it also conflates familial aggregation due to heritable predisposition with that due to shared environment. In an attempt to combine the best aspects of both these approaches, we developed a new measure; the heritability odds ratio, OR_g, and apply it to our case-control setting. This approach combines partitioning of familial associations into genetic and environmental parts with comparing cancer prevalence between case and control families, under some common assumptions. The resulting measure is most easily viewed as an *OR* of having a sibling with cancer, OR_{sib}, adjusted for the effects of common environmental exposures. That is the OR_{sib} due only to shared genetic factors.

The measure is built on the basic principal that under assumptions of independence of genetic and shared environmental factors, as well as their effects, the *OR* of a child having cancer given cancer in the parent, OR_{child}, can be partitioned into a product of an *OR* due to genetic factors, OR_g, and an *OR* due to shared environmental factors, OR_c.

$$OR_{child} = OR_g OR_c$$

this extends earlier ideas of partitioning the OR in a twin study setting [116].

We also note that the prevalence of cancer in children can be expressed in terms of the prevalence of cancer in parents and the OR_{child}. We use this observation to predict what the prevalence of cancer would be in proband sibs, had the OR_{sib} been equal to OR_g, that is; if the part of the association due to shared environmental exposures OR_c was equal to one. Comparing the predicted cancer prevalence in sibs between cases and controls, by means of the *OR*, then gives us an approximate

estimate of the association between cancer in a sib and STS in the proband adjusted for shared environmental factors, i.e. the association due to genetic factors such as predisposition genes shared between STS and common cancers.

Hazard rate ratio

As mentioned, in cohort studies, focus is often on incidence of disease. One way of describing this incidence is to look at the proportion of cohort members that have experienced an event before a given point in time *T*. This way of describing data becomes especially relevant when studying mortality, whereby it may not be strange that this method of data summary and its subsequent methods of analysis is referred to survival analysis. In a short, heuristic manner, it works in the following way

First, we note that a useful measure in this context is the hazard rate

$$h(t) = \lim_{\delta \to 0} \frac{P[\mathrm{t} < T \leq t + \delta | T > t]}{\delta}$$

i.e. the event rate at time *t*, conditional on survival until that time *t*. We realize that if we knew this entity for all times *t*, we would also know the proportion that have experienced the event at any time *t*. We call this summary of the proportion that experienced an event at time *t*, the failure function *F*(*t*). The relationship between *h* and *F* is governed by the equation

$$h(t) = \frac{\frac{\partial F(t)}{\partial t}}{F(t)}$$

which is equivalent to

$$\int h(t) = \int \frac{\frac{\partial F(t)}{\partial t}}{F(t)} = -lnF(t) + C$$

for any constant *C*. Adding, that *F*(*0*)=*1*, this further implies that

$$F(t) = e^{-\int h(t)dt}$$

and

$$S(t) = 1 - F(t) = 1 - e^{-\int h(t)dt}$$

if we are interested in survival at time *t*, as opposed to mortality. *S*(*t*) is usually estimated using the Kaplan-Meier method [106]. Now, in order to analyze the impact of different factors on survival in the cohort, a common approach is to apply what is referred to as the proportional hazards model

$$h(t) = h_0(t)e^{\beta_1 x_1 + .. + \beta_n x_n}$$

where, if for instance x_1 is a dichotomous variable, β_1=log(HR_1), and HR_1 is the ratio between the hazard rates corresponding to x_1=1 and x_1=0, i.e. the hazard rate ratio. Here x_1=1 might for example indicate an exposure group. If x_1 is continuous, HR_1 will be the ratio between hazards corresponding to one unit difference in x_1.

Among the proportional hazard models, the Cox proportional hazards model is likely the most well-known. A basis for its appeal is likely that is semi-parametric, in the sense that one does not have to specify the baseline hazard $h_0(t)$ for its application to data. Instead, in estimating the model using the maximum likelihood approach [106], Cox managed to show that the only part of the likelihood relevant in estimating the model parameters $\boldsymbol{\beta} = \beta_1, .., \beta_n$, are the factors

$$l(\boldsymbol{\beta}) = \prod_i \frac{e^{\boldsymbol{\beta X}_{\text{case}\,i}}}{e^{\boldsymbol{\beta X}_{\text{case}\,i}} + e^{\boldsymbol{\beta X}_{\text{control1}\,i}} + .. + e^{\boldsymbol{\beta X}_{\text{controln}\,i}}} \qquad (2)$$

where $\boldsymbol{X}_{case}i$ is the covariate vector for the case, the one who experienced the event, and $\boldsymbol{X}_{control\ 1}\,i, \ldots, \boldsymbol{X}_{control\ n}\,i$ covariate vectors for those event-free at the time *t* when case number *i* experienced its event. We note that (2) has the same form as (1) for estimating the *OR* in the matched case control study above, leading us to the conceptual idea that most epidemiological study designs originate from a cohort study and are only different in terms of what data one decides are convenient to collect from this cohort and the manner in which they are analyzed. Cox regression, as described above, is basically a case-control study nested in the study cohort, with controls collected by incidence density sampling. But instead of a fixed number of controls, it uses everyone in the cohort alive at the time of data collection.

In study IV, Cox regression is used to compare survival between those having been exposed to accidents before STS diagnosis to those who have not.

Standardized incidence rate ratios

The standardized incidence rate ratio is a classical epidemiological tool, which has been used to study disease at least since 1786 [106]. Part of its appeal is likely that it provides a single summary measure for comparing disease incidence, or mortality, between groups, when incidence rates, as well as rate differences between groups, may vary between population strata.

The indirect, as opposed to the direct standardization of rate ratios, makes use of rates external to the cohort and compare the number of events observed in the exposed cohort to that of a fictitious cohort with the same distribution with respect to the standardization variables, but with incidence rates equal to the external rates. Often the population used to generate the external rates is the general population of the country where the study is set. The SIR in this context is basically a comparison of the number of events observed compared to what would have been expected had the cohort not been exposed.

In mathematical terms the indirect standardization is defined as

$$SIR = \frac{\sum_{k=1}^{M} e_k}{\sum_{k=1}^{M} p_k \lambda_k} = \frac{E}{E^e} \qquad (3)$$

where e_k are the number of events, p_k is the number of person-years in strata k of the study cohort and λ_k is the incidence rate of the corresponding strata in the comparison population. The latter may explain an additional part of its appeal as the summing of events over all strata and comparing it to a fixed number, E^e, with no statistical variation, provides power in detecting excess incidence and mortality even when outcome is rare. This of course in addition to it being easy to understand and communicate.

As a side note we observe that if we assume that the effect of exposure is homogenous and proportional over all population strata, the *SIR* corresponds to an incidence rate ratio (*IRR*).

$$SIR = \frac{\sum_{k=1}^{M} e_k}{\sum_{k=1}^{M} p_k \lambda_k} = \frac{\sum_{k=1}^{M} p_k IRR \lambda_k}{\sum_{k=1}^{M} p_k \lambda_k} = IRR$$

where *IRR* is the incidence rate ratio comparing exposed subjects to those in the general population of the same strata.

Table 4 – Epidemiological measures used in papers I-V of the thesis.

Study	Measure
I	OR
II	OR, PAR, h^2
III	Heredity OR; OR_g
IV	OR, PAR, HR
V	SIR

Bias

The following cases of bias can affect any of the treated measures above, but here we will discuss them primarily in the context of the case-control study and its *OR*.

Confounding bias and adjustment

Confounding bias is in "A dictionary of Epidemiology" [117], defined as

> Bias of the estimated effect of an exposure on an outcome due to the presence of common causes of the exposure and the outcome.

In the present thesis we subscribe to the view that prerequisite causal knowledge is necessary for correct confounder adjustment and the subsequent removal of bias. Other approaches have been suggested and are widely used, but, as Hernan et al. clearly demonstrates [118], do not produce the intended bias reduction. These approaches are almost exclusively based on statistical associations, such as automated stepwise selection, comparing adjusted and unadjusted estimates or investigating associations of potential confounders with outcome and exposure.

In the present work little is known about what influences the risk of STS as well as the inter-relationship between covariates, and it is consequently difficult to identify confounders, mediators and colliders. The causal diagram used to summarize causal knowledge, usually illustrated through a directed acyclic graph (DAG), is not, or only partially, known. This makes the number of potential causal diagrams for the present studies numerous for any given exposure. Furthermore, as several exposures are investigated in the same analysis, several DAGs, as well as

adjusted analyses, would likely have to be generated for each exposure. This is not the common approach and adds up to an amount of detail difficult to include in a single manuscript. As such, we argue that the main interest of the present work is in crude estimates and we take a very pragmatic approach to confounder adjustment. While different models for adjustment are evaluated in paper I, in papers II and IV the standard approach of including all potential measured confounders is used, with the motivation that this will hopefully include the true confounders. Furthermore, if some care is taken not to include colliders (causes of both exposure and outcome), but we include some that are mediators, the resulting estimate will still have the interpretation of an indirect effect, which is still an effect of exposure on outcome, albeit possibly with a less obvious interpretation, as this may not be what one would expect to see in an intervention or randomized trial.

Recall bias

Recall bias, defined, in "A dictionary of Epidemiology" [117], as

> Systematic error due to differences in Accuracy or completeness of recall to memory of past events or experiences. For example, a mother whose child has died of leukemia may be more likely than the mother of a healthy living child to remember details of such past experiences as use of x-ray services when the child was in utero. A type of Information or Measurement bias.

In case-control studies using self-reported data, it is commonly known that cases with serious disease are more likely to recall and report events thought to be associated with their condition then their corresponding controls, producing a bias common to most case-control studies, termed recall bias.

Recall bias can be adjusted for in what is referred to as a sensitivity analysis [119] when the probability of reporting exposure conditional on exposure status is known, for instance through a subsample study. However, in most cases, as in the current project, these probabilities are not known. We are then left to venture best guesses, usually a range of possible values, to evaluate how these may have impacted study results.

In the current thesis, potential recall bias is a relevant aspect of all case-control studies I-IV, but particularly important and mainly discussed in paper IV. Here, in order to investigate whether recall bias may explain study findings, we provide some theoretical results to predict the *OR*s that would have been observed under the assumption that recall bias is produced mainly by the failure of controls to report exposure, while also assuming no effect of exposure. The results help us do two things; First, evaluate whether it is possible for recall bias to produce *OR*s of the magnitude observed in study IV. If it cannot, we can rule out recall bias as a possible explanation for the observations of study IV. Second, we note that the predicted *OR*s

under the assumption of recall bias vary with exposure prevalence, an observation we use to construct simple statistical tests for detecting the presence of this bias in the study data. Theory and background for the calculations are given in appendix I of paper IV.

Prospective validation study

Another way to evaluate the possibility of recall bias is to compare the results of the study to that of one with a prospective design that does not suffer from these issues. In study I, we validate the results of OC ever-use in a prospective cohort, the MISS study. We also investigate the risk of STSE of the lower extremities in the prospective knee prosthesis surgery cohort of study V, to potentially validate the surgery related results of study IV.

Selection bias

Selection bias is defined, in "A dictionary of Epidemiology" [117], as

> Bias in the estimated association or effect of an exposure on an outcome that arises from the procedures used to select individuals into the study or the analysis. When the selection involves conditioning on a factor that is affected by the exposure or a cause of the exposure, and also affected by the outcome or a cause of the outcome, selection bias can arise even in the absence of a causal effect of exposure on outcome, i.e., under the Causal Null Hypothesis.

In the present case-control study, issues of selection are reduced by the use of population based registers. However, due to the less than perfect response rates, it is still possible that there are systematic differences between cases and controls associated with the likelihood to respond. We evaluate potential effects of selection bias in studies I and IV, by partial validation of results in an external prospective study. Even if a cohort study may also suffer from this bias, contrasting two different study designs, with different sampling schemes, may still provide valuable information.

Detection bias

Detection bias is defined, in "A dictionary of Epidemiology" [117], as

> Bias due to systematic differences between the study groups in Ascertainment, assessment, diagnosis, or verification of outcomes. As other biases, it has numerous mechanisms and forms. An example is verification of diagnosis by laboratory tests in hospital cases but failure to apply the same tests to cases outside the hospital.

It can be grouped into the broader category of "information bias" and may present in both case-control and cohort study settings. In study IV, this bias may for instance occur when those exposed to injury, focus attention to the injury site and thereby increase probability of detecting a tumor. This may then increase the incidence of detected tumors in the injured study group, creating the illusion that injury may increase tumor risks. We evaluate the possibility of detection bias in study IV and V by examining time from exposure to tumor diagnosis, the latency time [120], with the rationale that excess risks due to detection should be close in time to exposure. Exposure in close connection to diagnosis would also contradict a causal role in tumor initiation, considering the long latency periods reported in connection to solid tumors [121]. We keep in mind, however, that this reasoning pertains mainly to solid tumor, as latency has been observed to vary between tumor types. Hematopoietic malignancies, for instance, have the much shorter latency of 0 to 5 years [122].

The latency analyses serve a particularly important purpose with respect to the SIRs of paper V, because of the study context. This is because, as we can see from equation (3) describing the indirect standardization, that is the SIR, when operation rates are increasing as they have been during recent years [123], the SIR is a weighted average of cohort cancer rates, with the largest weights attached to the most recent years. This shifts focus to short term risks that are not of primary interest as these may involve effects of detection bias, or more secondary effects in terms of tumor promotion. It limits the possibility to study causal risks associated with tumor initiating exposure.

Statistical analyses

What follows is a short summary of the statistical methods used in papers I, II, IV and V, that may not be considered standard. Methods that are explained in detail in the paper appendices are not explained in great detail here. As Paper III is a methods paper, only a short summary of the posed problem and two main points that contribute to its solution are provided. For remaining parts of this and the works of remaining papers, we simply refer to said papers.

Exact logistic regression

In small samples the regression coefficients in ordinary logistic as well as conditional logistic regression are known to be biased [124,125]. The same is true for highly unbalanced or stratified data. One way of avoiding this bias is to change approach and opt for exact tests, such as Fischer's exact or an exact version of the McNeamer test. An exact analogue to the standard regression approach in matched case-control analyses, conditional logistic regression, is exact logistic regression (ELR). ELR is based on exact permutational distributions of the sufficient statistics under the logistic model. To illustrate what this means, we give a short example.

Let us assume that we have a logistic model of binomial data

$$P(Y_j = y_j) = \binom{n_j}{y_j} p_j{}^{y_j} (1 - p_j)^{n_j - y_j}$$

with

$$p_j = \frac{e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}}{1 + e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}}$$

where $\boldsymbol{x}_1 = (x_{11}, x_{12}, .., x_{1J})$, $\boldsymbol{x}_2 = (x_{21}, x_{22}, .., x_{2J})$ and $\boldsymbol{y} = (y_1, y_2, .., y_J)$ are vectors containing information of covariates $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, and outcome $\boldsymbol{y}$, respectively, and we are interested specifically in the effect of $\boldsymbol{x}_2$ on outcome. Under regular conditions we could make use of the likelihood

$$L(y) = \prod_{j=1}^{J} P(Y_j = y_j) = \prod_{j=1}^{J} \binom{n_j}{y_j} p_j{}^{y_j}(1-p_j)^{n_j - y_j}$$

$$= \frac{e^{\beta_0 t_0 + \beta_1 t_1 + \beta_2 t_2}}{\prod_{j=1}^{J}\left(1 + e^{\beta_0 + \beta_1 x_{j1} + \beta_2 x_{j2}}\right)^{n_j}} \prod_{j=1}^{J} \binom{n_j}{y_j}$$

with sufficient statistics

$$t_k = \sum_{j=1}^{J} y_j x_{jk} \text{ and } t_0 = \sum_{j=1}^{J} y_j$$

to construct a likelihood ratio test for β_2. But in situations where this does not work well the idea is to make inferences about β_2 by conditioning on the sufficient statistics t_k in $L(y|t_0,t_1)=P(y|t_0,t_1)$ and thereby generate a permuational distribution of the likelihood ratio, to compare with the observed value and hence produce a p-value. Conditioning yields

$$P(y|t_0,t_1) = \frac{\frac{e^{\beta_0 t_0 + \beta_1 t_1}}{\prod_{j=1}^{J}\left(1 + e^{\beta_0 + \beta_1 x_{j1}}\right)^{n_j}} \prod_{j=1}^{J} \binom{n_j}{y_j}}{\sum_{y \in S} \frac{e^{\beta_0 t_0 + \beta_1 t_1}}{\prod_{j=1}^{J}\left(1 + e^{\beta_0 + \beta_1 x_{j1}}\right)^{n_j}} \prod_{j=1}^{J} \binom{n_j}{y_j}} = \frac{\prod_{j=1}^{J} \binom{n_j}{y_j}}{\sum_{y \in S} \prod_{j=1}^{J} \binom{n_j}{y_j}}$$

where S contains all y for which

$$t_1 = \sum_{j=1}^{J} y_j x_{j1} \text{ and } t_0 = \sum_{j=1}^{J} y_j$$

We can now create a permutational distribution of the likelihood ratio by generating all possible values of y and its corresponding likelihood ratio. By subsequent comparison of the observed likelihood ratio to the permutation distribution, one can generate an exact p-value.

Although elegant, an unfortunate disadvantage of this method is that it does not handle continuous covariates.

Population attributable risk in a case-control setting

Estimating the PAR in a case control setting is a little different from doing so in a cohort setting, as we do not have available to us a random sample from the population from which to estimate the risk factor prevalence. Instead one can use an alternative definition using only quantities available in the case control setting. We observe that the PAR

$$PAR = \frac{P[D] - P[D|\bar{E}]}{P[D]}$$

can be re-written as

$$PAR = \frac{P[E|D](RR - 1)}{(RR)}$$

where *D* is a disease indicator, *E* an exposure indicator, $\bar{E}$ meaning no exposure, *RR* the relative risk and *P[E|D]* is the probability of exposure in cases. The latter identity can, in contrast the earlier one, be used with an adjusted *RR*, something that generates bias with the former [113,117].

In the present study we note that the interpretation of the PAR calculated from data may not be representative for the general population of the South of Sweden during the time period due control selection issues. It may be that responding controls are healthier than the general population, not least because they are cancer free as far back as the 1960ies. However, it is our view that studying the PAR still communicates valuable information.

Heritability

Calculation of the narrow sense heritability h^2 of cancer in STS case families is done by means of a liability threshold model provided by Javars et al [114]. Details are given in appendix I of paper II.

Propensity score

In observational studies, different types of biases are always a concern. Common ways of reducing its influence is by regression adjustment or stratification. Another now quite popular way, is to use propensity score (PS) methods. A PS is a balancing score generated by analyzing the probability of exposure in the study population, for instance by means of regression, and assigning to each study participant a score corresponding to their probability of exposure. When conditioned on, for instance by stratification, the score balances covariates included in the score between exposure groups. When covariates used for constructing the score are confounders, this reduces confounding bias. Balance using the PS can also be achieved by means of regression adjustment or weighting.

As adjustment for the PS, that is; including it in a regression model as a covariate, as opposed to weighting or stratification, necessitates modelling of the functional form of the outcome dependence on the score. To avoid this challenge, we choose to stratify the score by quintiles. Quintiles have been standard with respect PSs, likely due to the fact that it was shown to remove 90% of the bias when studying confounding by continuous confounders and/or linear treatment effects [126].

It has been shown that PS adjustment performs the best, in terms not only of bias reduction, but a combination of reduction and variance, the mean squared error, when including, in addition to confounders, covariates related to outcome only, as this reduces variance, but not to exposure only, as this increases variance [127]. While this speaks clearly against using strategies based on statistical modelling of exposure probabilities, as in the present work of paper IV, we maintain that our foremost interest is in bias reduction and that outcome based modelling approaches would not be feasible in the present study setting, with several very small subgroups of STSs of different histological subtypes needing adjusted analyses including multiple confounders. We use PSs for the sole purpose that they allow for efficient stratification on a large set of variables.

Propensity score in a case-control setting

In case-control samples, as opposed to prospective studies, it may be especially difficult to differentiate between confounders, mediators and colliders, as data collection is retrospective. The basic strategy of prospective studies, is to collect covariates at baseline, and as such, one can be reassured that most covariates will not be mediators or colliders. Even if case-control studies often contain data on the time of exposure, information is of course subject to recall and may not always be precise enough to be of used in the temporal ordering of events.

PS methods were developed and primarily used in a cohort study setting. It has not experienced the same wide spread use in case-control studies, possibly because the estimation of the PS in this setting is less than straight forward. In study IV, we use a method based on a report by Månsson et al. [128], whereby PSs are estimated using the control population only. This is done on the grounds that outcome, STS, is rare, whereby the covariate distribution in controls approaches that of the general population when the study size increases. As our study is matched, we additionally allow for the propensity to vary with the levels of the matching variables. While Månsson et al. report that similar methods are accompanied by some difficulties of artificial effect modification [128] this appears to affect studies primarily of small to medium size, where as we estimate prosperity scores by pooling all 2066 + 2021 = 4087 controls.

As previously mentioned, in this thesis we use PS methods to attempt to avoid small sample issues with maintained validity, in situations where standard regression analyses may not be applicable. As opposed to conducting separate regression analyses for each subgroup, effectively reducing sample size and using up degrees of freedom for each covariate, we estimate prosperity scores by pooling all 4087 controls, and conducting analyses stratified by PS quintiles. Considering the number of controls, predicted scores should be fairly stable. Because men and women have been observed to differ with respect to risk factor profiles, not least in the present work, the analysis to generate PSs was stratified by gender. This approach has the advantage that for any subgroup analysis we only have to account for a single variable to balance potential confounders between cases and controls. We can therefore get an adjusted analysis for each subtype without the drawback of reduced statistical precision due to lost degrees of freedom.

Penalized regression

In order to generate PS we need to predict the probability of exposure in controls. To this end, several different methods of prediction have been used [126]. One way to improve ordinary regression predictions, in terms of mean squared error, is by the introduction of bias. A useful way of introducing bias is by shrinkage of the regression coefficients. One very attractive shrinkage estimator is the L1- penalty estimator, that has the convenient property of shrinking coefficients of unimportant predictors to exactly zero, and thereby performing variable selection [129]. This is a major advantage over other methods in terms of interpretability of results. In order to give an intuitive idea of how this works, we mention that in the case of linear regression and orthonormal predictors the L1-shrinkage estimator is defined as the solution to the optimization problem

$$argmin_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 \; s.t. \; \|\boldsymbol{\beta}\|_1 \leq t$$

or the equivalent

$$argmin_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1$$

which has the solution

$$\widehat{\beta_J} = sign\left(\widehat{\beta_J^{ols}}\right)\left(\widehat{\beta_J^{ols}} - \lambda\right)_+$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_n)$. Here we see that the L1-shrinkage estimator is merely a fixed translation of all regression parameters with a truncation at zero. The constant t has a one-to-one correspondence with λ. In applications, λ can be determined for instance by the use of cross-validation. [130] In the current project, the L1-penalization is used in the context of logistic regression, as we intend to predict the binary exposure indicator. This changes the optimization problem to

$$-\sum_{i=1}^{N}\left(y_i \boldsymbol{x_i}' \boldsymbol{\beta} - \log\left(1 + e^{\boldsymbol{x_i}' \boldsymbol{\beta}}\right)\right) + \lambda \|\boldsymbol{\beta}\|_1$$

The effect of parish of residence on exposure was evaluated before using penalized regression using mixed effects logistic regression with a random intercept for matched pairs. This was because this factor may be complicated to evaluate in a fixed effects setting, and as the analysis showed little effect of parish of residence, we avoid having to solve this problem in the penalized regression framework.

Firth correction

Logistic regression has well known problems with separation, diverging parameter estimates and bias in small samples. One possible solution to these problems is the

use of a special type of penalized regression, called firth correction [131,132]. Firth correction amounts to a penalized likelihood function of

$$L(\boldsymbol{\beta})^{*} = L(\boldsymbol{\beta}) + \frac{1}{2}|I(\boldsymbol{\beta})|$$

where *I* corresponds to the Fisher information matrix. It may look complicated but the penalized likelihood is equivalent to adding 0.5 to each data cell, for instance in the four fold table corresponding to data on a dichotomous variable. We use it in place of exact logistic regression in paper IV much because it is easily combined with multiple imputation. Not least because of available software [133].

Multiple imputation

Multiple imputation has during recent years emerged as the primary method of handling missing data in a wide range of study designs. Its chief advantage lies in the fact that it quantifies the uncertainty in imputations made and incorporates it into the precision estimate for studied statistics, such as regression coefficients. It does this by imputing missing values repeatedly and summarizing results in a pooled estimate, in a clever way. One of its more attractive methods is multiple imputation by chained equations (MICE). It handles imputation of several covariates of different types, continuous/binary/ordinal/categorical and so on, simultaneously. In order to get an intuitive feel for what it does, we introduce a 7 steps schematic here adapted from Azur et al. [134]

1. A single imputation is made of all missing values in all variables, using for instance the mean of the variable where the variable value is missing.

2. For a chosen single variable, *x*, the mean imputed values are removed and corresponding entries are set to missing once more.

3. The chosen variable *x* is regressed on remaining variables in the imputation model.

4. The missing values in the chosen variable *x* are replaced using the estimated model in step 3. These values are then used throughout the imputation process, when the chosen variable *x* is used as independent in the imputation model to impute the values of subsequent variables.

5. Steps 2 – 4 are repeated for each variable that has missing values. Replacing all missing values in all variables constitutes one cycle.

6. Cycles are repeated, a number of times, ten is standard, with the intent to make regression coefficients stable, and subsequently make the procedure independent of the order of imputed variables. At the end of all cycles, one imputed dataset has been created. This is the "chained equations" part.

7. Imputations are then repeated a number of times, 5-10 is commonplace, up to 40, has been suggested tom improve performance, depending on the problem and computational power.

The process is then evaluated using diagnostics to compare observed and imputed values as well as checking whether the algorithm has converged [135].

The use of MICE is in part based on the assumption of missing data being missing-at-random (MAR). MAR is the assumption that the probability of an observation being missing depends only on the observed data.

MICE is used in study IV primarily to lessen the impact of a specific type of recall bias. One may suspect that not only are controls less likely to report exposures, but may do it with a lesser degree of detail, i.e. when reporting for instance injury, one may be less likely to report both location and point in time. Therefore, we impute, for both injury and surgery, exposure, exposure location, age and time to data collection (diagnosis in cases).

Poisson modelling of SIRs

Given the mathematical expression for the SIR above and a Poisson distribution for the observed events E, we can compare the number of observed and expected events, E and E^e, through Poisson regression, to estimate the SIR as well as make inferences, according to

$$SIR = \frac{E}{E^e}$$

$$\log(SIR) = \log(\frac{E}{E^e})$$

$$\log(E) = \log(SIR) - \log(E^e)$$

where the last term is modelled as an offset. Because SIRs cannot be readily compared between different strata, as differences may be due to the distribution of standardizing variables, and not to exposure. In these situations the above regression equation can be used for comparison between strata by conditioning on standardizing variables $\boldsymbol{X}$.

$$\log(E) = \log(\alpha) + \boldsymbol{X\beta} - \log(E^e)$$

where log(α) + $\boldsymbol{X\beta}$ = *SIR*. In the current study the above equation is also used to investigate the change in SIRs relative to time from surgery, both before and after. This is simply done by including time as one of the covariates in $\boldsymbol{X}$ [136]. In order to differentiate time trends from temporal changes in the distribution of standardization variables, sex, age and calendar year, these were also included in $\boldsymbol{X}$.

Restricted cubic splines

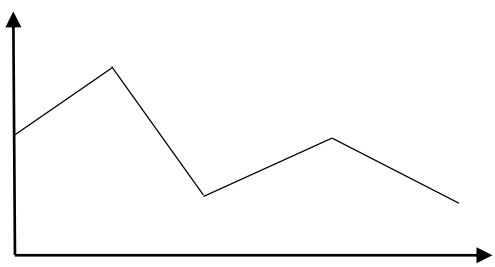
When examining the relationship between a covariate *X* and an outcome *Y* statistically, one finds it to be non-linear in many cases. Regression splines is a convenient way to estimate non-linear functional relationships as well as the associated precision of this estimation.

A regression spline in its most simple form is a linear spline

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 (X - x_1)_+ + \cdots + \beta_n (X - x_n)_+$$

where $(t)_+$ indicates a function that takes on the value 0 when *t* is negative and *t* otherwise, and with knots $x_1 < x_2 < \ldots < x_n$, in the interior of the study interval where *x* is observed. Although a linear spline gives improved freedom to model more complex relationships compared to linear regression, the sharp turns it takes at the knots often appear unrealistic.

Figure IV - Simple graphical illustration of the sharp turn in functional relationships that are rarely observed in practice.



Nature often moves in smooth motions. Modelling can be improved in this respect by allowing linear terms to take on non-linear forms, such as polynomials of a higher degree. The estimated functional form will gain additional smoothness if we force first and second degree derivatives to be equal at the knots, effectively limiting the rate of change that can occur at the knots. A spline of this form can be written as

$$E[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_2 (X - x_1)_+{}^3 + \cdots + \beta_{n+3} (X - x_n)_+{}^3$$

In some instances, as is common in polynomial regression, predictions may be quite unstable and diverge rapidly in the first and last intervals of *X*. In order to avoid this problem one can force the estimated function to be linear at the beginning and end intervals. In order to force linearity at the first interval, β_{II} and β_{II} have to be set to zero. At the end interval, the last two β's become redundant, as they become functions of remaining β's. The remaining terms make up the regression equation

$$E[Y|X] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{n-1} X_{n-1}$$

where $X_1 = X$.

$$X_{j+1} = (X - x_j)_+^3 - \frac{x_n - x_j}{x_n - x_{n-1}} (X - x_{n-1})_+^3 + \frac{x_{n-1} - x_j}{x_n - x_{n-1}} (X - x_n)_+^3$$

and $j = \{1, \ldots, n-2\}$. And even though, the restricted cubic spline (RCS) yields non-linear functional relationships between *X* and *Y*, the RCS regression equation is linear in the regression parameters and RCS can therefore be used in any standard regression setting to estimate and make inferences about possible functional relationships. [137]

In the current study, RCS is used to model the seemingly non-linear relationship between time-from-surgery and SIR. And as time intervals are narrow to gain more detail in time trends, RCS analysis also provides smoothing to otherwise noisy data.

Table 5 – Measures and statsitical methods used in each paper I-V.

Study	Measure	Method
I	OR	Standard, Conditional and Exact logistic regression
	Adjusted OR	Standard, Conditional and Exact logistic regression
II	OR	Conditional logistic regression
	Adjusted OR	Conditional logistic regression
	PAR	PAR for case control studies
	h^2	ACE liability threshold model
III	Heredity OR; OR_g	Method developed in paper III
IV	OR	Conditional logistic regression
	Adjusted OR	Conditional logistic regression stratified by PS
	Location specific OR	Standard logistic regression with Firth correction
	PAR	PAR for case control studies
	HR	Stratified cox regression
V	SIR	Poisson analysis
	SIR change with time	Poisson regression with RCS

Results and Discussion

Soft tissue sarcomas are relatively rare, but often come with a poor prognosis at diagnosis. This is because of delay, as early stage sarcomas lack distinct symptoms. And as STS are common in young adults, they are also responsible for many years of life lost. Meanwhile, there are few known causes [43] and consequently, there is little or no basis for disease prevention.

The current work is a great leap forward in this sense. Although, results for the most part are not directly applicable in a public health prevention setting, they point to specific research areas where causative factors may be found. They also indicate great rewards if these factors are properly understood and managed, as up to an estimated 57 % of all STS incidence could potentially be prevented, and 74% of STSE. Each area and its attributed fraction of STS risk is shown in table 6, as well as the fraction attributed to all areas combined. As is shown here, prior to our work, almost no areas to which we could attribute STS risk were known.

The PAR for the purpose of this result summary and discussion, is not used in what may be considered as its standard context, to gauge how much of the disease incidence may be eradicated had a given risk factor been removed from the population. Instead, we use it to estimate the proportion of incidence attributed to factors associated with the factor under study. For instance, if we consider stature relative to peers at pubertal onset, and the PAR associated with it, it is meaningless to know what proportion of incidence could be prevented had everyone in the population been shorter than their peers at age eleven, which is contradiction by definition. The PAR for this factor is instead meant to be interpreted as the proportion due to involvement of factors associated with pre-pubertal height, such as GH and IGF-1, and that could be prevented had we been able to control the effect of these factors on STS genesis and growth. However, some caution is recommended when viewing the figure in table 6 as one has to also keep in mind that this PAR is calculated for a population that is probably healthier than the general population during the same time period.

Table 6 – STS risks attributable to each risk factor studied in papers I-IV. Results are adjusted for sex and age. The calculations do not include particiants with previous surgical treatment for tumour disease.

Factor type	Factor	Group	Subgroup	PAR
Risk	Tissue damage			
	Surgery	Extremities		40%
	Injury from accident			9%
	Both			43%
Risk	Heredity			
	First degree relative with cancer	Female		9%
			Have never used OCs	18%
			Have ever used OCs	0%
		Male		11%
		All		10%
Protective	Reproductive			
	Ever used oral contraceptives	Female		14%
	Three or more children			22%
	Both			30%
Protective	Stature			
	Shorter than peers at age 11	All		38%
Both	All			
	Surgery, injury, FDR with cancer, OC ever-use, Three or more children, Shorter than peers at age 11	Female		67%
	Surgery, injury, FDR with cancer, Shorter than peers at age 11	Male		61%
	Surgery, injury, FDR with cancer, Shorter than peers at age 11	All		57%
	Surgery, injury, FDR with cancer, OC ever-use, Three or more children, Shorter than peers at age 11	Extremities	Female	90%
	Surgery, injury, FDR with cancer, Shorter than peers at age 11		Male	69%
	Surgery, injury, FDR with cancer, Shorter than peers at age 11		All	74%
Risk	Established			
	Exposure to radiotherapy	All		2%

Study I

We studied the effect of hormone related factors on STS in the Swedish population between 1988 and 2009 using a population based matched case–control design. Our study was the largest on this topic to date, including 634 cases in a primary matched analysis and 855 cases in an unmatched sensitivity analysis. We identified protective effects connected to constitutional characteristics, hormonal and reproductive factors. Being shorter than your peers at age 11 was associated with an *OR* of 0.51 (0.36–0.74). Having used oral contraceptives (OC), *OR* 0.75 (0.49–1.15), and high parity, *OR* 0.16 (0.04–0.63), comparing three or more children to two or less, also appeared to reduce the risk of STS. The risk was further reduced with the duration of OC use ($p = 0.01$), comparing use for 11 years or more to use for 3 years or less yielded an OR of 0.10 (0.02–0.41). No effect was observed for ever having had HRT *OR* 1.02 (0.70–1.47). The effect of BMI varied significantly with subtype ($p = 0.03$) and tumor location ($p<0.001$).

In the present study, an estimated 30% of female STS incidence could be attributed to not ever using OCs or having less than three children. Thirty eight percent could be attributed to being equally tall or taller than peers at age eleven. While the latter two are not modifiable in the sense of an intervention, the figures do illustrate the influential magnitude of sex and growth hormones on STS genesis and the potential long term benefit of studying these associations.

However, one factor that may be a potential prevention target is OC use. Results do not of course imply that all women should take oral contraceptives so that a few can avoid developing STS, as the potential adverse effects of extended OC use, including increased risks of breast cancer, by far outweigh the benefits, especially considering the very low risk of developing STS in the general population. However, if high risk groups could be identified from future research, the number of unnecessarily exposed to OCs could be reduced and those who benefit could be increased. Furthermore, as results indicate that there may be a critical age period for OC driven risk reduction, prevention could potentially be limited to this time period and thereby avoid unnecessary harmful exposures.

A second potential public health implication related to OC use is the risk-lowering effect of parity and OC-use on female STS risk. One can just ponder the implications for treating clinics if women stopped using OCs and had markedly fewer children. This could likely increase STSE incidence by about 40% among females. This is certainly worth further studies, with a focus on projecting different scenarios.

A second prevention target may be growth hormones/IGF-1, which has been shown to be associated with dietary habits [138]. It is possible that a diet focused on controlling cancer risk in general may also apply to STS. Our results indicate that a

successful prevention in this area could have a significant impact on STS incidence in all histological subtypes.

Study II

The risk of STS increased with having a first degree relative with STS, *OR* 4.2 95% CI (0.8 – 21.9). With respect to other tumors, the overall *OR* was low and non-significant. In females who had never used oral contraceptives (OCs), there was a significant effect of cancer in a first degree relative (FDR), *OR* 1.5 (1.1-2.3), *PAR* 18%. It was *OR* 0.7 (0.4-1.2), *PAR* 0%, in OC ever-users. The difference was significant (p=0.021). Among histological subtypes, MFH/UPS in men, 65 years of age or younger, showed an *OR* of 4.5 (1.8 – 11.6), which was significantly greater than in remaining age and subtype strata (p=0.024). Here, 54% of tumor incidence was attributable to familial cancers and heritability among case family members was 54% (10 – 84%). 9% of cases were 45 years of age or younger, the cut-off age for the Li-Fraumeni diagnostic criteria. In females, risks were related primarily to female parents and sibs, *OR* 3.27 (1.43 – 7.48), and not to males 1.36 (0.62 – 2.94). Only 22% of these STS cases were below the age of 45. Here, the largest effect was seen in the rarer subtypes, for which the proportion of incidence attributable to familial cancers was 31% and the heritability was 69% (22 – 90%).

The study suggests that low penetrant inherited mutations, potentially connected to p53 pathways, may be important in the etiology of adult STS, especially in women, and that OC use may potentially limit their effect. In fact, this effect may be considerable, as in a non-OC-using population 18% of female STS incidence is due familial cancers as opposed to 0% in an all-OC-using population. Furthermore, if genes and risk alleles could be identified from future research, prevention could potentially target those who would most benefit and effectively limit unnecessary exposure, making prevention efficient with less harmful impact on public health.

Study III

In study II, we focused on risk associated with a positive family history of cancer measured in terms of an odds *OR*. This approach allowed us to extract valuable information familial aggregation of cancer in STS, but it is in some sense methodologically flawed as interest is exclusively in inherited genetic, as opposed to shared environmental, factors and it fundamentally conflates effects of the two. In contrast to studies of early onset disease, in adult STS we could not readily assume that familial effects were due exclusively to inherited genetics, as adults, as

opposed to children or young adults, have sufficient time to accumulate enough environmental exposures to be at increased risk for malignances. It is therefore vital to the results of the study to be able to distinguish between inherited genetic and shared environmental effects in this contexts.

In order to separate them, we wanted to apply an ACE model to partition the outcome variance into components of additive genetic and shared and unique environmental effects. However, this latent trait model did not easily lend itself to the analysis of case-control data. As an ad hoc solution, we applied the method to case families only and used published data on the general population as a reference to gauge whether cancer in STS case families were more or less "genetic" than cancers "in general". To solve this problem in a more rigorous manor, we extend earlier ideas of decomposing the *OR* to our current setting, with case and control families. The result was that we could produce an *OR* for cancer in a sibling, adjusted for common environmental exposures, OR_g. That is, an *OR* measuring only the association between cancer in a sibling and the risk of STS in a proband due to shared genetic factors, such as shared predisposition genes.

Or, to put it more correctly, in the case-control setting, we could only produce an approximate interval for the adjusted *OR*, as opposed to an exact estimate, due missing information. This interval, however, turned out to be quite useful in our study setting, as it may prove to be in many others.

The method was applied to female STS and cancer in female siblings. Results showed the OR_g was somewhere in the interval (1.68 (0.99 – 2.29) - 1.82 (1.07 – 3.13)) which verified that this association indeed was due to inherited genetic predisposition shared between STS and common cancers, i.e. potentially one or more unknown cancer syndromes.

As crude and adjusted *OR*s are the bread and butter of epidemiology, results from the current method may be easier to digest for the average epidemiologist than that of the ACDE model that pertains to partitioning of variance on a latent scale. In addition, we also note that the regression formulation suggested allows for application to small samples, such as through Firth correction or exact regression methods, and potentially for generalization to other types of outcomes, such as time-to-event. We feel that the method is simple, at least in the cross-sectional study setting, results are easy to communicate and the method is generalizable, and as such a welcome complement to existing methods for studying heritable dichotomous outcomes.

Study IV

We studied the association between accidental injury and/or surgery in 249 cases of STSE, each matched to four controls. We saw that having been exposed to surgery

or injury from accidents to any part of the body both increased the risk of STSE, *OR* 2.1 95% CI (1.5 – 3) and 1.5 (1.1 – 2.1), respectively. Adjusting for potential confounders yielded *OR*s of 2.4 (1.6 – 3.6) and 1.4 (1.00 – 2.00). Surgery related *OR* (*SOR*) varied with STSE histological subtypes, from 1.2 to 4.2. MFH/UPS, leiomyosarcoma and liposarcoma showed significantly elevated *SOR*s that were significantly greater than for remaining subtypes (P=0.029). The accidental injury related odds ratio (*AOR*) varied from 0.8 to 3.5, being significantly elevated for leiomyo- and liposarcomas, and the *AOR* of the two combined was significantly greater than remaining subtypes (P=0.045). Analyzing surgery or accidental injury to a given part of the body and the subsequent risk of STSE in the same location showed the *SOR* to be 4.13 95%CI (1.87 – 9.15) and the *AOR* was 2.29 95%CI (1.21 – 4.32).

The *PAR* for surgery 40% and for accidental injury, 9%. It was and 43% for both combined. This indicates that tissue trauma and, potentially mediating inflammatory response may be an important contributing factor in sarcomagenesis of the extremities.

These results do not necessarily help prevention, as one usually cannot eliminate surgeries from the population, although there may be important exceptions such as breast cancer patients treated with mastectomy, who may be at high risk for STS, and the choice to reconstruct the missing breast/breasts. However, results may also facilitate early detection, which is detrimental to prognosis, as tissue damage-induced tumors are currently not an accepted diagnostic entity. This knowledge may on occasions help clinicians identify STS, as one may otherwise look to other differential diagnoses. Furthermore, if inflammation is a mediating factors, this may raise questions regarding possible prevention using non-steroidal anti-inflammatory drugs (NSAIDs).

Further studies are of course needed to elucidate the biological mechanisms behind these associations and may help identify high risk groups that could be subject to monitoring, or possibly prevention. Although a first step would be to reproduce study results. A natural study setting for such further studies are large prospective registers, where all surgeries, as well as diagnoses are well documented and recall bias is eliminated. Swedish registers are well suited for this purpose.

Study V

We evaluated the long-term cancer risks associated with having a knee prosthesis, a long-term follow-up of all knee patients in Sweden between 1975 and 2006, comparing the cancer incidence of operated individuals to the national incidence in Sweden by means of standardized incidence ratios.

For male and female patients with RA or OA, the overall cancer risks were elevated, ranging from 1.10 95% CI: (1.03 – 1.18) to 1.26 (1.23 – 1.29). The greatest increases in risk were observed for leukemia subtypes, myelodysplastic syndromes (MDS) and essential thrombocytosis (ET), ranging from 3.31 (1.24–8.83) for ET in men with OA to 7.38 (1.85–29.51) for ET in women with RA. Increased risks were also observed for breast cancer, prostate cancer and melanoma. A latency analysis revealed elevated risks late in the study period for both solid and hematopoietic cancers. However, only increases in MDS and possibly prostate cancer and melanoma rates appeared to be connected to the operation. Elevated risks of MDS and possibly prostate cancer and melanoma indicated a potential connection to exposure to metals in the implant. The observed excessive incidence of ET was likely associated with the inflammatory disease.

Recent research

Since the publication of our study, there have been additional reports, mainly dealing with the risk associated with metal-on-metal hip prostheses, for which concerns were raised in the public media [139]. Two studies replicated an overall increased risk [140,141], 1.08 (1.04 – 1.12) and one did not [142] and yet another was close [143] when looking at non-metal-on-metal prostheses, which was the larger subgroup, 1.04 (0.99 – 1.09). Two studies replicated the risk of skin cancer [140,143] and one the risk of prostate cancer [140], while one was close 1.18 (0.97 – 1.41) [143] when looking at metal-on-metal prostheses, and one study the risk of hematopoietic malignancies [140]. Although the latter did not stratify by diagnosis leading to surgery. Another study showed a significant increase in STS comparing metal-on-metal to non-metal hip prostheses [74].

Furthermore, in periprosthetic masses of twenty patients fitted with metal-on-metal hip prostheses, cancer-related genetic alternations were identified that are commonly found in hematological malignancies and bone dysplasia. One periprosthetic mass was a liposarcoma. Genetic alternations were associated with in-situ time of the prosthesis but no significant association with cobolt or chromium levels, in accordance with an additional study. This made the authors hypothesize that alterations may be due to prolonged induced inflammatory response, as opposed to genotoxic effects of metal ions [144].

The discussion regarding causes of elevated risks and discrepancies between studies remain. To complicate matters, there are a myriad of factors that may confound the association between joint prosthesis and cancer and for which the distribution between study populations may differ. These include everything from patient characteristics in term of underlying and comorbid conditions, previous occupational and life style exposures, treatment and diagnostic procedures as well as potentially procedures and willingness to supply information for tumor

registration, especially for rare tumors where registry completeness is poor [102]. As the proportion operated in the general population has increased dramatically [123], in addition to there seemingly being a lack of consensus of the main indications for surgery [123], patient populations are indeed likely to differ between studies, and over time. To what extent these differences matter to cancer risks is unknown.

Another potential explanation for the elevated risks is that of detection bias. Observations that support this hypothesis may be that the particular risks that have been replicated appear to be those that may be affect by increasing visits to the doctor's office with consequent repeated examinations, x-rays and blood tests; hematological malignancies, skin, breast and prostate cancer. Here, the difference between OA and RA patients and the short latency, mainly in OA patients, may be interpreted as support, as RA patients may have had a higher frequency of health care contacts preceding surgery and thereby have less bias in the time period following surgery. It may also be noteworthy that studies using hospital controls, as opposed to the general population, have not resulted in these elevated risks [145].

Here, differences between studies may be due to different routines with regards to patient examinations and diagnostic procedures. The change over time in risks of, for instance, breast cancer may correspond to improved methods of detection [146,147]. This would in part explain the seemingly constant risk increase over follow-up time, as in improvement in diagnostic capabilities during a given calendar period, could possibly affect the detection rate of all study participants irrespective of when during their post-surgery follow-up it may have be introduced.

The hypothesis of detection bias could potentially be evaluated using mortality data, in lines with the reasoning of study IV. As it happens, this has been studied in the present knee cohort previously [148]. Here, mortality was reduced compared to the general population, likely due to healthy patient selection effects. However, this reduction, over time, from surgery, changed to an increase in mortality after about 12 years following operation. This translated into mortality being increased in younger patients, less than 55 years of age at operation. This increase appeared to be mainly due to cardiovascular, gastrointestinal, urogenital causes, but an increasing elevated risk of tumor related deaths was also observed, although not statistically significant, from 0.79 (0.76 – 0.82) in those 55 years of age at operation or older to 1.28 (0.89 – 1.79) in those younger. This may be a sign of a gradual increase in tumor death with follow-up, as older patients die early. Additionally, the proportional mortality ratio for tumor related deaths (the observed proportion relative to the expected proportion) among older patients, 1.03 (0.99 – 1.07), indicating that despite the reduced mortality there may be a slight shift in the causes of death towards tumor diseases. Considering the prognosis following diagnosis of some of the excess cancer types, it may not be unexpected that a diagnosis early during follow-up impacts survival years later. Therefore, the elevated risk of death by tumor disease may be a sign that elevated cancer risks are not, or just in part, due to detection bias. Although, as results were not significant, they need to be

confirmed in a larger population setting for increased precision and potentially ruling out random variation as a cause. The possible increase in risk should be offset against a report of a total population of OA patients in the same region showing a reduced overall mortality compared to the general population [149]. Also, considering that a corresponding hip arthroplasty cohort reported a reduced cancer incidence SIR 0.95 (0.92 – 0.97) and a subsequent reduced overall and as well as tumor related mortality SMR 0.69 (0.67-0.70) and 0.54 (0.50 – 0.57), yielding a PMR of 0.78 and thereby possibly indicating a correlation between cancer incidence and mortality and consequently pointing to a biological cause, as opposed one of study design.

A competing hypothesis, particularly for the short term risks, related to biological processes, is that of inflammation. Inflammation due to the underlying condition, potentially exacerbated by surgery related tissue trauma, and further induced by prosthesis debris [101]. Inflammation has been implicated in the etiology of common cancers [14] Surgery with subsequent tissue trauma induced inflammatory response has been observed to drive progression of micro metastasis [150,151]. Inflammation is thought to have a potential part in all stages of tumor development, from initiation to promotion and progression, being able to act on cancer risk in the short term. Further empirical support for this hypothesis may come from our study IV, where excess lower extremity STSs was observed in our knee prosthesis cohort, 5-10 years following surgery, in accordance with the general results for STSE following accidents or surgery of any kind, where excess risks were observed mainly 3-10 years following operation. Furthermore, about half of STSE tumors observed in the knee cohort were that of MFH/UPS, leiomyo- and liposarcomas, tumors for which tissue trauma appears to be an important factor, as seen from paper IV and table 6 . Furthermore, gastrointestinal tumors, reported to be reduced in several other studies [98,100,152], likely due to the use of NSAIDs, were not reduced in our study. This may question the balance between induced inflammation and the extent of NSAID use for this population.

Differences between study populations following this hypothesis may in part be due to differences in this balance, as NSAIDs are known to act as chemopreventive agents, not least for breast [153], prostate [154] and potentially for skin [155].

Conclusions and future perspectives

Study I – IV

We have identified several areas for further research into understanding the etiology and potential prevention targets of STS. These areas, if properly understood and managed, could lead to potential reductions of STS of up to 57%. This increases to 75% in STSE and up to 90% in female STSEs. This is important, because even though STSs are rare, they are often associated with a poor prognosis and substantial amount of years of life lost.

Several interesting research questions follow the present work, but the most relevant in terms of prevention can probably be grouped into three steps:

- *Validate results*. Studies I, II and IV can all be replicated to some extent using SNBH registers. Considering the different sources of bias, this step is crucial.

- *Identify high risk subjects*. Many of the present risk factors are prevalent in the population, and as STS is rare, any prevention with these as target would affect a large number of persons unnecessarily. If we can identify groups at high risk, this number can be reduced and prevention may be efficient and ultimately feasible. A study that elaborates on this issue is Merlo et al [156].

- *Identify preventive action*. Some research areas connected to potential preventions have been identified, such as diet and exogenous sex hormone exposure, now further studies are needed to better understand the biological underpinnings of these associations and ultimately translate them into prevention strategies for groups of high risk individuals.

Some thoughts along the lines of point two and three, would perhaps be

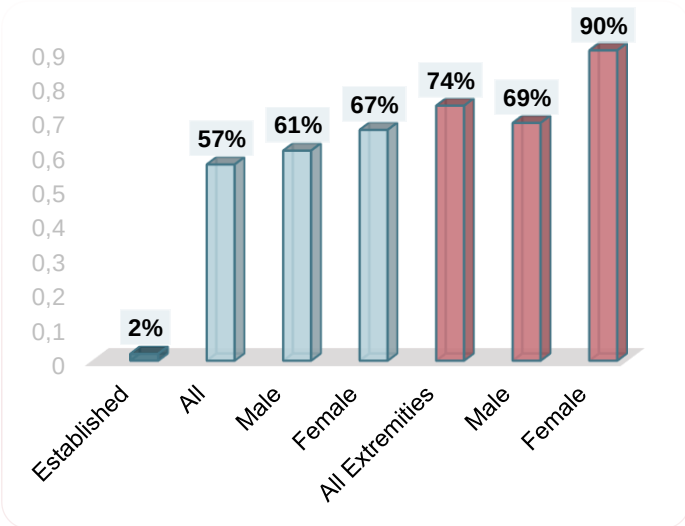
- *Study I* suggested that GH/IGF-1 may play a part in the etiology of a wide verity of STSs of different locations and histological subtypes. Other studies have suggested that IGF-1 levels may be targeted to reduce risks of common cancers through dietary prevention. Perhaps risks of STS are also modified using these same strategies.

- *Study I* also suggests that there may be a critical age for exogenous sex hormone exposure to exert maximum protection. If identifiable, one may limit exposure for high risk individuals to this age period, and thereby limit unnecessary exposure and subsequent side effects.

- *Study II* suggests that a portion of female STS may be due to genetic predisposition, and that exogenous sex hormone exposure may limit or event prevent this genetic risk. A study genotyping candidate genes associated with p53 pathways in female STS cases, possibly focusing on the less frequent histological subtypes, may reveal genetic markers for high risk individuals suitable for prevention.

- *Study IV* suggests that a sizable portion of STSE may be associated with tissue trauma, perhaps mediated by inflammatory response. Considering the potential effect of NSAID in the prevention of common cancers, perhaps they could also be used as chemopreventive agents in this setting. However, as mentioned, the ability to identify high risk subjects is key, as NSAID use is not free from side effects and should not be applied to large populations unnecessarily.

Figure 4 – Fraction of STS (light blue) and STSE (red) incidence attributable to risk factors identified in the pressent thesis, compared to the fraction attributed to risk factors previously established for STS (dark blue). Here we consider only exposure to radiation therapy as a establihed.



Study V

The research into cancer risks associated with prosthesis surgery is still inconclusive. Our study, in concert with others, have indicated excess tumor incidence following surgery. A few processes have been identified to explain these findings, in study IV we identified tissue trauma as an additional factor. However, findings are still inconsistent and warrant continued studies and surveillance. Perhaps with increasing study sizes, follow-up and specific focus on associated tumor diseases, results may be less heterogeneous, especially in terms of short term risks. The latter could possibly be achieved by nesting a more focused study within the population based registers, to collect information on potential confounders, such as inflammatory markers and NSAID use, tumor staging at diagnosis and survival, to further evaluate the possibility of confounding and detection bias, getting closer to identifying, or refuting, causal factors. And perhaps with time, register studies will yield the needed information on long term risks, lacking today.

Populärvetenskaplig sammanfattning

Den aktuella avhandlingen härrör ur forskningsinsatser inom klinikerna för onkologi och ortopedi vid Universitetssjukhuset i Lund, och behandlar vad som ursprungligen verkade vara två skilda ämnen.

Ena ämnet behandlar epidemiologin för mjukdelssarkom hos vuxna (MDS), en grupp av ofta dödliga sjukdomar av okänd orsak, som behandlas av både onkologer och ortopeder. Här har den historiska oförmågan att kartlägga etiologin för dessa tumörer resulterat i en brist på förebyggande åtgärder med en betydande förlust av levnadsår som följd. I det här avhandlingsarbetet studerar vi faktorer förknippade med kroppskonstitution, reproduktiva händelser, ärftlighet och vävnadstrauma.

Det andra ämnet behandlar risken för cancer efter knäproteskirurgi, en oro bland ortopedforskare och en viktig folkhälsofråga, då antalet proteser stadigt ökar, inte minst hos yngre patienter.

Dessa två ämnen har visat sig vara relaterade då mjukdelstumörer har identifierats i anslutning till protesimplantat, vilket fört med sig att kunskap kring riskfaktorer för MDS potentiellt skulle kunna ge ledtrådar om carcinogena effekter hos protesimplantat och dess associerade faktorer, så som underliggande sjukdom och behandling.

I studien av båda ämnena, arbetar vi med några av världens största och mest detaljerade studiepopulationer. En är en populationsbaserad fall-kontrollstudie med alla fall av MDS från södra sjukvårdsregionen mellan åren 1998 och 2009, vilket inkluderar nästan 1000 fall. En annan är Svenska Knäprotesregistret, världens äldsta register i sitt slag, som inkluderar alla opererade patienter från 1975 till nutid.

Vi fann att 57% av MDS-fallen, och 74% av de MDS-fall som uppstod i extremiteter, i vår studiekohort kunde tillskrivas faktorer relaterade till kroppskonstitution och reproduktiva händelser, ärftlighet och vävnadstrauma. Vi finner också ett överskott av MDS-tumörer i knäproteskohorten, tillsammans med ett lågt, men statistiskt signifikant, överskott av mer vanliga cancerformer. Dessa inkluderade prostata cancer, malignt melanom och hematologiska maligniteter. Med tanke på förekomsten av MDS, drar vi slutsatsen att vävnadstrauma kan vara en bidragande faktor i till den ökade cancerrisken.

Även om dessa ämnen i sig är viktiga, presenterar de också en rad intressanta utmaningar när det gäller epidemiologisk och statistisk metodik. Till exempel utgör studiet av MDS, som en heterogen grupp av sällsynta tumörsjukdomar med potentiellt skilda etiologier, problem i termer av subgruppsanalyser samt justering

för multipla störfaktorer i små studier. Den sällsynta sjukdomen leder också till svårigheter att utreda effekten av ärftlighet, då etablerade analysmetoder för detta inte på ett enkelt sätt går att förena med den etablerade studiedesign som ofta används i dessa fall. Studien av proteskirurgirelaterad cancer inkluderar utmaningar i att skilja mellan kausala effekter och den bias som uppstår när sjukhuspatienter undersöks på grund av sin sjukdom och cancerdiagnoser uppkommer som ett bifynd.

Acknowledgements

This work is by no means the work of a single person. There are several persons I would like to thank and whose contributions I would like to acknowledge here.

First and foremost, to my supervisor and guide into the amazing world of Cancer epidemiology, Professor *Håkan Olsson*, thank you for imparting invaluable wisdom and knowledge. For your confidence in my work and for great advice and immense patience with a novice researcher and his somewhat odd ideas. Thank you for giving me a framework for understanding epidemiology. You should know that it has meant a lot.

To my mentor, Professor *Jonas Ranstam*, thank you for introducing me into the exciting world of biostatistics. Thank you for numerous constructive discussions and immense patience with someone who was yet to understand the purpose and foundation of statistical science. Watching you work has taught me more than most books. Thank you for generously opening many doors. It has been crucial.

To my co-supervisor and co-authors; Professor *Lars Lidgren*, for your support in this and many other projects, and for giving me a push in the right direction when needed. To *Otto Robertsson* for always interesting discussions and ideas. They have taught me a lot. (Even though I still cannot explain the cumulative hazard in a meaningful way.) To Professor *Anders Rydholm*, *Fredrik Vult von Steyern* and *Thor Alvegård*, thank you for a great collaboration. This would not have worked without your support and input.

To *Mats Enlund*, *Professor Jerzy Leppert*, Professor *Kent Nilsson* and everyone at CKF. This would certainly not have been possible without you. Thank you all for the wonderful support, understanding and for creating a great scientific work environment. It is what helps people grow.

To *Elisabet Rodby-Bosquet* for introducing me to a wonderful workplace, for advice on big things and small, for support, encouragement and scientific collaborations. You, my friend, are a force of nature.

Professor *Gunnar Hägglund* for support and great collaborations. For giving me the opportunity to take my first small steps into science. Thank you for your immense patience in explaining orthopedics to a math geek.

Professor *Juan Merlo* and the Department of Social Epidemiology for great collaborations on interesting and challenging questions. For never settling. For pushing the boundaries. We have only just begun.

To my former co-workers *Aleksandra Turkiewicz* and *Rebecca Rylance*, thank you for lively discussions about statistical methodology, life, politics, religion and other important things. (In that order.) A special thanks to Rebecca for working intensely to improve the horrible excuse for English language that is this authors writing.

To Professor *Marie Reilly* for graciously hosting me at KI and MEB, making me part of a great work environment. Thank you. It has meant a lot. To my MEB roommates; *Johan Zetterqvist* and *Hanna Bower*, and everyone at MEB, for great discussions on causality and most other things. Thank you for making me feel welcome.

To *Monika Andersson* and *Ingrid Mårtesson* at the Department of Cancer Epidemiology for great help with the administrative aspects of research. For helping me carry tons and tons of binders. A sincere, thank you. To *Per Broberg* and *Harald Andersson* for many interesting discussions.

To my parents *Zoltán* and *Eva* for loving support and encouragement. For being there. Not just for the years of this thesis, but for all the years. There really are no words to thank you. I love you. To my sister *Bim*, for being there for me growing up, and her loving family. I miss you.

Last, but certainly not least, to my family; *Marlon*, *Inez* and *Sofie*. My life. It is all for you. Sorry for all the work. This is the only thesis I will write, promise.

And to all the patients and people that put time and effort into answering the study questionnaires that led to this research. Your contributions have not gone unnoticed. I have read much of what you wrote, in detail.

References

1. Mallick A, Jain S, Proctor A, Pandey R. Angiosarcoma around a revision total hip arthroplasty and review of literature. J Arthroplasty 2009;24(2):323.e17-20.
2. Visuri T, Pulkkinen P, Paavolainen P. Malignant tumors at the site of total hip prosthesis. Analytic review of 46 cases. J Arthroplasty 2006;21(3):311-23.
3. Robertsson O, Ranstam J, Sundberg M, W-Dahl A, Lidgren L. The Swedish Knee Arthroplasty Register: a review. Bone Joint Res 2014;3(7):217-22.
4. Pitot HC. The molecular biology of carcinogenesis. Cancer 1993;72(3 Suppl):962-70.
5. GM. C. The Developement and Causes of Cancer. In: GM. C, editor. The Cell: A Molecular Approach. 2nd edition. Sunderland (MA): Sinauer Associates; 2000.
6. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100(1):57-70.
7. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell 2011;144(5):646-74.
8. IARC. World Cancer Report 2014. Lyon, France: IARC; 2014. 79-84 p.
9. Schottenfeld D. Cancer Epidemiology and Prevention. New York: Oxford University Press; 2006.
10. Knudson A. Alfred Knudson and his two-hit hypothesis. (Interview by Ezzie Hutchinson). Lancet Oncol 2001;2(10):642-5.
11. Parkin DM, Boyd L, Walker LC. 16. The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010. Br J Cancer 2011;105 Suppl 2:S77-81.
12. Mucci LA, Hjelmborg JB, Harris JR, Czene K, Havelick DJ, Scheike T, Graff RE, Holst K, Möller S, Unger RH and others. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. JAMA 2016;315(1):68-76.
13. Okada F. Inflammation and free radicals in tumor development and progression. Redox Rep 2002;7(6):357-68.
14. Rakoff-Nahoum S. Why cancer and inflammation? Yale J Biol Med 2006;79(3-4):123-30.
15. Rogol AD, Roemmich JN, Clark PA. Growth at puberty. J Adolesc Health 2002;31(6 Suppl):192-200.
16. Noriega NC. Evolutionary Perspectives on Sex Steroids in the Vertebrates. InTech; 2012.
17. Burns KA, Korach KS. Estrogen receptors and human disease: an update. Arch Toxicol 2012;86(10):1491-504.
18. Chen GG, Zeng Q, Tse GM. Estrogen and its receptors in cancer. Med Res Rev 2008;28(6):954-74.

19. Weiderpass E, Adami HO, Baron JA, Magnusson C, Bergstrom R, Lindgren A, Correia N, Persson I. Risk of endometrial cancer following estrogen replacement with and without progestins. J Natl Cancer Inst 1999;91(13):1131-7.
20. Collaborative Group On Epidemiological Studies Of Ovarian C, Beral V, Gaitskell K, Hermon C, Moser K, Reeves G, Peto R. Menopausal hormone use and ovarian cancer risk: individual participant meta-analysis of 52 epidemiological studies. Lancet 2015;385(9980):1835-42.
21. Kennelly R, Kavanagh DO, Hogan AM, Winter DC. Oestrogen and the colon: potential mechanisms for cancer prevention. Lancet Oncol 2008;9(4):385-91.
22. Diep CH, Daniel AR, Mauro LJ, Knutson TP, Lange CA. Progesterone action in breast, uterine, and ovarian cancers. J Mol Endocrinol 2015;54(2):R31-53.
23. Lange CA, Yee D. Progesterone and breast cancer. Womens Health (Lond) 2008;4(2):151-62.
24. Burkman R, Schlesselman JJ, Zieman M. Safety concerns and health benefits associated with oral contraception. Am J Obstet Gynecol 2004;190(4 Suppl):S5-22.
25. Olsson H, Moller TR, Ranstam J. Early oral contraceptive use and breast cancer among premenopausal women: final report from a study in southern Sweden. J Natl Cancer Inst 1989;81(13):1000-4.
26. Collaborative Group on Hormonal Factors in Breast C. Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies. Lancet 1996;347(9017):1713-27.
27. Jernstrom H, Loman N, Johannsson OT, Borg A, Olsson H. Impact of teenage oral contraceptive use in a population-based series of early-onset breast cancer cases who have undergone BRCA mutation testing. Eur J Cancer 2005;41(15):2312-20.
28. Olsson HL, Ingvar C, Bladstrom A. Hormone replacement therapy containing progestins and given continuously increases breast carcinoma risk in Sweden. Cancer 2003;97(6):1387-92.
29. Chlebowski RT, Kuller LH, Prentice RL, Stefanick ML, Manson JE, Gass M, Aragaki AK, Ockene JK, Lane DS, Sarto GE and others. Breast cancer after use of estrogen plus progestin in postmenopausal women. N Engl J Med 2009;360(6):573-87.
30. Johnson JR, Lacey JV, Lazovich D, Geller MA, Schairer C, Schatzkin A, Flood A. Menopausal hormone therapy and risk of colorectal cancer. Cancer Epidemiol Biomarkers Prev 2009;18(1):196-203.
31. Brinton LA, Felix AS. Menopausal hormone therapy and risk of endometrial cancer. J Steroid Biochem Mol Biol 2014;142:83-9.
32. Bae JM, Kim EH. Hormonal Replacement Therapy and the Risk of Lung Cancer in Women: An Adaptive Meta-analysis of Cohort Studies. J Prev Med Public Health 2015;48(6):280-6.
33. Schottenfeld D, Fraumeni JF. Cancer epidemiology and prevention. Oxford ; New York: Oxford University Press; 2006. xviii, 1392 p. p.
34. La Vecchia C, Negri E, Franceschi S, Parazzini F. Long-term impact of reproductive factors on cancer risk. Int J Cancer 1993;53(2):215-9.

35. Hankinson SE, Colditz GA, Hunter DJ, Willett WC, Stampfer MJ, Rosner B, Hennekens CH, Speizer FE. A prospective study of reproductive factors and risk of epithelial ovarian cancer. Cancer 1995;76(2):284-90.
36. Titus-Ernstoff L, Perez K, Cramer DW, Harlow BL, Baron JA, Greenberg ER. Menstrual and reproductive factors in relation to ovarian cancer risk. Br J Cancer 2001;84(5):714-21.
37. Maki RG. Small is beautiful: insulin-like growth factors and their role in growth, development, and cancer. J Clin Oncol 2010;28(33):4985-95.
38. Jenkins PJ, Mukherjee A, Shalet SM. Does growth hormone cause cancer? Clin Endocrinol (Oxf) 2006;64(2):115-21.
39. Garber JE, Offit K. Hereditary cancer predisposition syndromes. J Clin Oncol 2005;23(2):276-92.
40. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. Oncogene 2004;23(38):6471-6.
41. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, Koskenvuo M, Pukkala E, Skytthe A, Hemminki K. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343(2):78-85.
42. Hayes RB. The carcinogenicity of metals in humans. Cancer Causes Control 1997;8(3):371-85.
43. Burningham Z, Hashibe M, Spector L, Schiffman JD. The epidemiology of sarcoma. Clin Sarcoma Res 2012;2(1):14.
44. Maretty-Nielsen K. Prognostic factors in soft tissue sarcoma. Dan Med J 2014;61(11):B4957.
45. Doyle LA. Sarcoma classification: an update based on the 2013 World Health Organization Classification of Tumors of Soft Tissue and Bone. Cancer 2014;120(12):1763-74.
46. Olsson H. An updated review of the epidemiology of soft tissue sarcoma. Acta Orthop Scand Suppl 2004;75(311):16-20.
47. Hardell L. [Malignant mesenchymal tumors and exposure to phenoxy acids -- a clinical observation]. Lakartidningen 1977;74(33):2753-4.
48. Schwab M. Encyclopedia of Cancer Berlin, Heidelberg: Springer; 2012.
49. Cole P, Trichopoulos D, Pastides H, Starr T, Mandel JS. Dioxin and cancer: a critical review. Regul Toxicol Pharmacol 2003;38(3):378-88.
50. Folkerd EJ, Dowsett M. Influence of sex hormones on cancer progression. J Clin Oncol 2010;28(26):4038-44.
51. Britt K, Ashworth A, Smalley M. Pregnancy and the risk of breast cancer. Endocr Relat Cancer 2007;14(4):907-33.
52. Yager JD, Davidson NE. Estrogen carcinogenesis in breast cancer. N Engl J Med 2006;354(3):270-82.
53. Mueck AO, Seeger H, Rabe T. Hormonal contraception and risk of endometrial cancer: a systematic review. Endocr Relat Cancer 2010;17(4):R263-71.

54. Fioretti F, Tavani A, Gallus S, Negri E, Franceschi S, La Vecchia C. Menstrual and reproductive factors and risk of soft tissue sarcomas. Cancer 2000;88(4):786-9.
55. Mastrangelo G, Coindre JM, Ducimetiere F, Dei Tos AP, Fadda E, Blay JY, Buja A, Fedeli U, Cegolon L, Frasson A and others. Incidence of soft tissue sarcoma and beyond: a population-based prospective study in 3 European regions. Cancer 2012;118(21):5339-48.
56. Chaudhuri PK, Walker MJ, Beattie CW, Das Gupta TK. Presence of steroid receptors in human soft tissue sarcomas of diverse histological origin. Cancer Res 1980;40(3):861-5.
57. Li XQ, Hisaoka M, Hashimoto H. Expression of estrogen receptors alpha and beta in soft tissue sarcomas: Immunohistochemical and molecular analysis. Pathol Int 2003;53(10):671-9.
58. Sekyi-Otu A, Bell RS, Ohashi C, Pollak M, Andrulis IL. Insulin-like growth factor 1 (IGF-1) receptors, IGF-1, and IGF-2 are expressed in primary human sarcomas. Cancer Res 1995;55(1):129-34.
59. Rikhof B, de Jong S, Suurmeijer AJ, Meijer C, van der Graaf WT. The insulin-like growth factor system and sarcomas. J Pathol 2009;217(4):469-82.
60. Calle EE, Kaaks R. Overweight, obesity and cancer: epidemiological evidence and proposed mechanisms. Nat Rev Cancer 2004;4(8):579-91.
61. McDuffie HH, Pahwa P, Karunanayake CP, Spinelli JJ, Dosman JA. Clustering of cancer among families of cases with Hodgkin Lymphoma (HL), Multiple Myeloma (MM), Non-Hodgkin's Lymphoma (NHL), Soft Tissue Sarcoma (STS) and control subjects. BMC Cancer 2009;9:70.
62. Ji J, Eng C, Hemminki K. Familial risk for soft tissue tumors: a nation-wide epidemiological study from Sweden. J Cancer Res Clin Oncol 2008;134(5):617-24.
63. Fernebro J, Bladström A, Rydholm A, Gustafson P, Olsson H, Engellau J, Nilbert M. Increased risk of malignancies in a population-based study of 818 soft-tissue sarcoma patients. Br J Cancer 2006;95(8):986-90.
64. Ognjanovic S, Olivier M, Bergemann TL, Hainaut P. Sarcomas in TP53 germline mutation carriers: a review of the IARC TP53 database. Cancer 2012;118(5):1387-96.
65. Wu CC, Shete S, Amos CI, Strong LC. Joint effects of germ-line p53 mutation and sex on cancer risk in Li-Fraumeni syndrome. Cancer Res 2006;66(16):8287-92.
66. Vogelstein B, Prives C. p53: The Most Frequently Altered Gene in Human Cancers. Nature Education 2010;3(9):6.
67. Cho Y, Kim J, Kim Y, Jeong J, Lee KA. A case of late-onset Li-Fraumeni-like syndrome with unilateral breast cancer. Ann Lab Med 2013;33(3):212-6.
68. Nichols KE, Malkin D, Garber JE, Fraumeni JF, Li FP. Germ-line p53 mutations predispose to a wide spectrum of early-onset cancers. Cancer Epidemiol Biomarkers Prev 2001;10(2):83-7.
69. Ruijs MW, Verhoef S, Rookus MA, Pruntel R, van der Hout AH, Hogervorst FB, Kluijt I, Sijmons RH, Aalfs CM, Wagner A and others. TP53 germline mutation testing in 180 families suspected of Li-Fraumeni syndrome: mutation detection rate and

relative frequency of cancers in different familial phenotypes. J Med Genet 2010;47(6):421-8.

70. Palmero EI, Achatz MI, Ashton-Prolla P, Olivier M, Hainaut P. Tumor protein 53 mutations and inherited cancer: beyond Li-Fraumeni syndrome. Curr Opin Oncol 2010;22(1):64-9.
71. Upadhyaya M. Neurofibromatosis Type 1. Berlin Heidelberg: Springer; 2012.
72. Evans DG. Neurofibromatosis type 2. Handb Clin Neurol 2015;132:87-96.
73. Sherlock DJ, Rickards H, Gardecki TI, Hamer JD. Development of a sarcoma in a surgical scar. Postgrad Med J 1987;63(746):1097-8.
74. Makela KT, Visuri T, Pulkkinen P, Eskelinen A, Remes V, Virolainen P, Junnila M, Pukkala E. Cancer incidence and cause-specific mortality in patients with metal-on-metal hip replacements in Finland. Acta Orthop 2014;85(1):32-8.
75. Can Z, Yilmaz S, Riza A, Apaydin EI, Kuzu I. Sarcoma developing in a burn scar: case report and review of the literature. Burns 1998;24(1):68-71.
76. Green JJ, Heymann WR. Dermatofibrosarcoma protuberans occurring in a smallpox vaccination scar. J Am Acad Dermatol 2003;48(5 Suppl):S54-5.
77. Ozercan IH, Okur MI, Coskun F, Yildirim AM. Malignant fibrous histiocytoma and squamous carcinoma derived from a burn scar. Acta Chir Belg 2004;104(6):745-7.
78. Cohen S, Ad-El D, Benjaminov O, Gutman H. Post-traumatic soft tissue tumors: case report and review of the literature a propos a post-traumatic paraspinal desmoid tumor. World J Surg Oncol 2008;6:28.
79. Kowal-Vern A, Criswell BK. Burn scar neoplasms: a literature review and statistical analysis. Burns 2005;31(4):403-13.
80. Ozyazgan I, Kontaş O. Burn scar sarcoma. Burns 1999;25(5):455-8.
81. Cocke WM, Tomlinson JA. Malignant fibrous histiocytoma developing in burn scar of the ear. Burns 1993;19(3):241-3.
82. Kim GI, Lee JH, Kim HK, Park SH, Kim CH. Malignant fibrous histiocytoma in a chronic burn scar: a rare case report and review of the literature. Burns 2004;30(7):742-5.
83. Yoon PW, Jang WY, Yoo JJ, Yoon KS, Kim HJ. Malignant fibrous histiocytoma at the site of an alumina-on-alumina-bearing total hip arthroplasty mimicking infected trochanteric bursitis. J Arthroplasty 2012;27(2):324.e9-324.e12.
84. Min WK, Kim SY, Oh CW, Kim SJ, Park TI, Koo KH. Malignant fibrous histiocytoma arising in the area of total hip replacement. Joint Bone Spine 2008;75(3):319-21.
85. Gomez P, Morcuende J. High-grade sarcomas mimicking traumatic intramuscular hematomas: a report of three cases. Iowa Orthop J 2004;24:106-10.
86. Webster-Cyriaque J. Development of Kaposi's sarcoma in a surgical wound. N Engl J Med 2002;346(16):1207-10.
87. Asahi T, Kurimoto M, Kawaguchi M, Yamamoto N, Sato S, Endo S. Malignant fibrous histiocytoma originating at the site of a previous fronto-temporal craniotomy. J Clin Neurosci 2002;9(6):704-8.

88. Inoshita T, Youngberg GA. Malignant fibrous histiocytoma arising in previous surgical sites. Report of two cases. Cancer 1984;53(1):176-83.
89. Society AC. 2016-09-20. What are the risk factors for soft tissue sarcomas? <http://www.cancer.org/cancer/sarcoma-adultsofttissuecancer/detailedguide/sarcoma-adult-soft-tissue-cancer-risk-factors>. 2016-09-20.
90. Hofer SO, Molema G, Hermens RA, Wanebo HJ, Reichner JS, Hoekstra HJ. The effect of surgical wounding on tumour development. Eur J Surg Oncol 1999;25(3):231-43.
91. Van Mater D, Añó L, Blum JM, Webster MT, Huang W, Williams N, Ma Y, Cardona DM, Fan CM, Kirsch DG. Acute tissue injury activates satellite cells and promotes sarcoma formation via the HGF/c-MET signaling pathway. Cancer Res 2015;75(3):605-14.
92. Pendergrast WJ, Futrell JW. Biologic determinants of tumor growth in healing wounds. Ann Surg 1979;189(2):181-8.
93. Robertsson O. Annual Report 2016 Swedish Knee Arthroplasty Register. Helsingborg: Swedish Knee Arthroplasty register (SKAR); 2016.
94. Robertsson O. Annual Report 2008 Swedish Knee Arthroplasty Register. Helsingborg: Swedish Knee Arthroplasty register (SKAR)´; 2008.
95. Lidgren L. Chronic inflammation, joint replacement and malignant lymphoma. J Bone Joint Surg Br 2008;90(1):7-10.
96. Memoli VA, Urban RM, Alroy J, Galante JO. Malignant neoplasms associated with orthopedic implant materials in rats. J Orthop Res 1986;4(3):346-55.
97. McGregor DB, Baan RA, Partensky C, Rice JM, Wilbourn JD. Evaluation of the carcinogenic risks to humans associated with surgical implants and other foreign bodies - a report of an IARC Monographs Programme Meeting. International Agency for Research on Cancer. Eur J Cancer 2000;36(3):307-13.
98. Onega T, Baron J, MacKenzie T. Cancer after total joint arthroplasty: a meta-analysis. Cancer Epidemiol Biomarkers Prev 2006;15(8):1532-7.
99. Paavolainen P, Pukkala E, Pulkkinen P, Visuri T. Cancer incidence after total knee arthroplasty: a nationwide Finnish cohort from 1980 to 1996 involving 9,444 patients. Acta Orthop Scand 1999;70(6):609-17.
100. Visuri T, Pukkala E, Pulkkinen P, Paavolainen P. Decreased cancer risk in patients who have been operated on with total hip and knee arthroplasty for primary osteoarthrosis: a meta-analysis of 6 Nordic cohorts with 73,000 patients. Acta Orthop Scand 2003;74(3):351-60.
101. Wagner P, Olsson H, Ranstam J, Robertsson O, Zheng MH, Lidgren L. Metal-on-metal joint bearings and hematopoetic malignancy. Acta Orthop 2012;83(6):553-8.
102. Barlow L, Westergren K, Holmberg L, Talbäck M. The completeness of the Swedish Cancer Register: a sample survey for year 1998. Acta Oncol 2009;48(1):27-33.
103. Skatteverket. 2017 18th March. Statens personadressregister. Skatteverket <https://www.statenspersonadressregister.se/>. Accessed 2017 18th March.

104. Lindqvist PG, Epstein E, Olsson H. The relationship between lifestyle factors and venous thromboembolism among women: a report from the MISS study. Br J Haematol 2009;144(2):234-40.
105. Breslow NE, Day, N.E. Statistical Methods in Cancer Research Volume II - The Design and Analysis of Cohort Studies. Oxford, UK: International Agency for Cancer Research; 1987.
106. Breslow NE, Day, N.E. Statistical Methods in Cancer Research Volume I - The analysis of case-control studies. Oxford, UK: International Agency for Cancer Research; 1980.
107. Doll R. The causes of death among gas-workers with special reference to cancer of the lung. Br J Ind Med 1952;9(3):180-5.
108. Pearce N. What does the odds ratio estimate in a case-control study? Int J Epidemiol 1993;22(6):1189-92.
109. Greenland S, Thomas DC. On the need for the rare disease assumption in case-control studies. Am J Epidemiol 1982;116(3):547-53.
110. Aldrich J. R.A. Fisher and the making of maximum likelihood 1912 - 1922. Statistical science 1997;12(3):162 - 176.
111. Brookmeyer R, Liang KY, Linet M. Matched case-control designs and overmatched analyses. Am J Epidemiol 1986;124(4):693-701.
112. Stallones RA. The use and abuse of subgroup analysis in epidemiological research. Prev Med 1987;16(2):183-94.
113. Rockhill B, Newman B, Weinberg C. Use and misuse of population attributable fractions. Am J Public Health 1998;88(1):15-9.
114. Javaras KN, Hudson JI, Laird NM. Fitting ACE structural equation models to case-control family data. Genet Epidemiol 2010;34(3):238-45.
115. Neale MC, Cardon LR. Methodology for Genetic Studies of Twins and Families. Netherlands: Springer; 1992.
116. Ramakrishnan V, Thacker LR. POPULATION ATTRIBUTABLE FRACTION AS A MEASURE OF HERITABILITY IN DICHOTOMOUS TWIN DATA. Commun Stat Simul Comput 2012;41(3).
117. Association IE. A dictionary of Epidemiology. Oxford; New York: Oxford University Press; 2001.
118. Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. Am J Epidemiol 2002;155(2):176-84.
119. Greenland S. Basic methods for sensitivity analysis of biases. Int J Epidemiol 1996;25(6):1107-16.
120. Olsson H, Baldetorp B, Ferno M, Perfekt R. Relation between the rate of tumour cell proliferation and latency time in radiation associated breast cancer. BMC Cancer 2003;3:11.
121. Thompson DE, Mabuchi K, Ron E, Soda M, Tokunaga M, Ochikubo S, Sugimoto S, Ikeda T, Terasaki M, Izumi S and others. Cancer incidence in atomic bomb survivors. Part II: Solid tumors, 1958-1987. Radiat Res 1994;137(2 Suppl):S17-67.

122. Preston DL, Kusumi S, Tomonaga M, Izumi S, Ron E, Kuramoto A, Kamada N, Dohy H, Matsuo T, Matsui T and others. Cancer incidence in atomic bomb survivors. Part III. Leukemia, lymphoma and multiple myeloma, 1950-1987. Radiat Res 1994;137(2 Suppl):S68-97.
123. Culliford DJ, Maskell J, Kiran A, Judge A, Javaid MK, Cooper C, Arden NK. The lifetime risk of total hip and knee arthroplasty: results from the UK general practice research database. Osteoarthritis Cartilage 2012;20(6):519-24.
124. Greenland S, Schwartzbaum JA, Finkle WD. Problems due to small samples and sparse data in conditional logistic regression analysis. Am J Epidemiol 2000;151(5):531-9.
125. Nemes S, Jonasson JM, Genell A, Steineck G. Bias in odds ratios by logistic regression modelling and sample size. BMC Med Res Methodol 2009;9:56.
126. Austin PC. An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies. Multivariate Behav Res 2011;46(3):399-424.
127. Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. Am J Epidemiol 2006;163(12):1149-56.
128. Månsson R, Joffe MM, Sun W, Hennessy S. On the estimation and use of propensity scores in case-control and case-cohort studies. Am J Epidemiol 2007;166(3):332-9.
129. Vidaurre D, Bielza, C. , Larranaga, P. A survey of L1 Regression. International Statistical Review 2013;81(3):361-387.
130. Tibshirani R. Regression shrinkage and selection via the lasso. . Journal of the Royal Statistical Society 1996;58(1):267-288.
131. Firth D. Bias reduction of maximum likelihood estimates. Biometrika 1993;80:27-38.
132. Heinze G, Schemper M. A solution to the problem of separation in logistic regression. Stat Med 2002;21(16):2409-19.
133. Heinze G. logistf: Firth's bias reduced logistic regression. 1.06. Vienna, Austria: R Foundation for statistical computing; 2006.
134. Azur MJ, Stuart EA, Frangakis C, Leaf PJ. Multiple imputation by chained equations: what is it and how does it work? Int J Methods Psychiatr Res 2011;20(1):40-9.
135. van Buuren S. G-OCG. MICE: Multivariate Imputation by Chained Equations in R. Journal of statistical software 2011;45(3).
136. Atkinson E.J. CCS, Pedersen R.A., Therneau T.M. Poisson models for person-years and expected rates. Mayo Clinic; 2008. Report nr 81.
137. Durrleman S, Simon R. Flexible regression models with cubic splines. Stat Med 1989;8(5):551-61.
138. Levine ME, Suarez JA, Brandhorst S, Balasubramanian P, Cheng CW, Madia F, Fontana L, Mirisola MG, Guevara-Aguirre J, Wan J and others. Low protein intake is associated with a major reduction in IGF-1, cancer, and overall mortality in the 65 and younger but not older population. Cell Metab 2014;19(3):407-17.
139. MHRA. 2012 20th March. Press release: MHRA updates advice for metal-on-metal hip replacements. MHRA - Regulating Medicines and Medical Devices <http://webarchive.nationalarchives.gov.uk/20141205150130/http://www.mhra.gov.uk/NewsCentre/Pressreleases/CON143784>. Accessed 2017 20th March.

140. Brewster DH, Stockton DL, Reekie A, Ashcroft GP, Howie CR, Porter DE, Black RJ. Risk of cancer following primary total hip replacement or primary resurfacing arthroplasty of the hip: a retrospective cohort study in Scotland. Br J Cancer 2013;108(9):1883-90.
141. Dybvik EH. Cancer and Total Hip Replacement. Cancer as a risk factor for prosthesis and prosthesis as a risk factor for cancer. Bergen: Univeristy of Bergen; 2015.
142. Smith AJ, Dieppe P, Porter M, Blom AW, National Joint Registry of E, Wales. Risk of cancer in first seven years after metal-on-metal hip replacement compared with other bearings and general population: linkage study between the National Joint Registry of England and Wales and hospital episode statistics. BMJ 2012;344:e2383.
143. Makela KT, Visuri T, Pulkkinen P, Eskelinen A, Remes V, Virolainen P, Junnila M, Pukkala E. Risk of cancer with metal-on-metal hip replacements: population based study. BMJ 2012;345:e4646.
144. Sarhadi VK, Parkkinen J, Reito A, Nieminen J, Porkka N, Wirtanen T, Laitinen M, Eskelinen A, Knuutila S. Genetic alterations in periprosthetic soft-tissue masses from patients with metal-on-metal hip replacement. Mutat Res 2015;781:1-6.
145. Goldacre MJ, Wotton CJ, Seagroatt V, Yeates D. Cancer following hip and knee arthroplasty: record linkage study. Br J Cancer 2005;92(7):1298-301.
146. Lewold S, Olsson H, Gustafson P, Rydholm A, Lidgren L. Overall cancer incidence not increased after prosthetic knee replacement: 14,551 patients followed for 66,622 person-years. Int J Cancer 1996;68(1):30-3.
147. Wagner P, Olsson H, Lidgren L, Robertsson O, Ranstam J. Increased cancer risks among arthroplasty patients: 30 year follow-up of the Swedish Knee Arthroplasty Register. Eur J Cancer 2011;47(7):1061-71.
148. Robertsson O, Stefansdottir A, Lidgren L, Ranstam J. Increased long-term mortality in patients less than 55 years old who have undergone knee replacement for osteoarthritis: results from the Swedish Knee Arthroplasty Register. J Bone Joint Surg Br 2007;89(5):599-603.
149. Turkiewicz A, Neogi T, Bjork J, Peat G, Englund M. All-cause Mortality in Knee and Hip Osteoarthritis and Rheumatoid Arthritis. Epidemiology 2016;27(4):479-85.
150. El Saghir NS, Elhajj, II, Geara FB, Hourani MH. Trauma-associated growth of suspected dormant micrometastasis. BMC Cancer 2005;5:94.
151. Coffey JC, Wang JH, Smith MJ, Bouchier-Hayes D, Cotter TG, Redmond HP. Excisional surgery for cancer cure: therapy at a cost. Lancet Oncol 2003;4(12):760-8.
152. Olsen JH, McLaughlin JK, Nyren O, Mellemkjaer L, Lipworth L, Blot WJ, Fraumeni JF, Jr. Hip and knee implantations among patients with osteoarthritis and risk of cancer: a record-linkage study from Denmark. Int J Cancer 1999;81(5):719-22.
153. Yiannakopoulou E. Aspirin and NSAIDs for breast cancer chemoprevention. Eur J Cancer Prev 2015;24(5):416-21.
154. Vidal AC, Freedland SJ. Aspirin and prostate cancer prevention. Aging (Albany NY) 2015;7(5):292-3.
155. Goodman JR, Grossman D. Aspirin and other NSAIDs as chemoprevention agents in melanoma. Cancer Prev Res (Phila) 2014;7(6):557-64.

156. Merlo J, Wagner P, Mulinari S, Wemrella M, SVe S, Hedblad B. The tyranny of the averages and the indiscriminate use of risk factors in Public Health: The case of coronary heart disease. Recently submitted2017.