

From the Clinical Epidemiology Division at the Department  
of Medicine, Solna  
Karolinska Institutet, Stockholm, Sweden

# AGNOSTIC STUDIES IN EPIDEMIOLOGY

Torsten Dahlén



**Karolinska  
Institutet**

Stockholm 2023

All previously published papers were reproduced with permission from the publisher.

Published by Karolinska Institutet.

Printed by Universitetservice US-AB, 2023

© Torsten Dahlén, 2023

ISBN 978-91-8016-888-5

Cover illustration: Illustration of many associations with different strengths between multiple factors. Torsten Dahlén.

# Agnostic studies in Epidemiology

## Thesis for Doctoral Degree (Ph.D.)

By

**Torsten Dahlén**

The thesis will be defended in public at Inghesalen in Solna, 31st March 2023 at 9 am

**Principal Supervisor:**

Docent Gustaf Edgren  
Karolinska Institutet  
Department of Medicine  
Division of Clinical Epidemiology

**Co-supervisors:**

Docent Mark Clements  
Karolinska Institutet  
Department of Medical Epidemiology and  
Biostatistics

Professor Martin L Olsson  
Lund University  
Department of Laboratory Medicine  
Division of Hematology and Transfusion  
Medicine

Docent Patrik Magnusson  
Karolinska Institutet  
Department of Medical Epidemiology and  
Biostatistics

**Opponent:**

Professor Henrik Toft Sørensen  
Aarhus University  
Department of Clinical Epidemiology

**Examination Board:**

Professor Jesper Lagergren  
Karolinska Institutet  
Department of Molecular Medicine and Surgery

Docent Mattias Rantalainen  
Karolinska Institutet  
Department of Medical Epidemiology and  
Biostatistics

Professor Anske Van der Bom  
Leiden University  
Department of Immunohematology and Blood  
Transfusion and Department of Clinical  
Epidemiology



To my amazing wife Sofia and our three kids,  
Flynn, Eyvind and Edith,  
who came along during this thesis.



## Popular science summary of the thesis

Scientific questions have in epidemiology evolved around the urge to understand and describe the relationship between a single factor and a specific medical condition. In this thesis we explore a less specific line of questioning. Instead of asking a limited number of questions we try to address all possible questions related to a disease or many diseases. The rationale to conduct research in this manner is primarily to reduce bias of the sometimes narrow questions asked by the individual researcher or research groups. The topics explored in this thesis range from identifying relationships between ABO blood group, RhD status and disease, the existence of disease transmission through blood transfusion, adverse events in patients treated with tyrosine kinase inhibitors in chronic phase chronic myeloid leukemia and lastly, the existence of cancerous disease aggregation within families. All studies are conducted using population-based health-registers available in Sweden. The applied methodologies are similar and based on defining a large set of diseases or factors and performing a large set of tests to identify possible interrelationships. In order to avoid the risk of identifying false relationships that may be appearing by chance, we use additional methods to reduce this risk.

In the first study we find 49 relations between an ABO blood group and a disease. For RhD status, we identify one such relation. In the study of transfusion-transmitted disease, we were able to confirm the known transfusion-transmitted disease, e.g. hepatitis virus. In terms of adverse events in patients with chronic myeloid leukemia, we also find increased risks in multiple disease as compared to the general population. Reassuringly for treating physicians and patients, there were no new unknown severe adverse events. In the last study of familial cancer aggregation, we do find such aggregation in families, most commonly in breast- or prostate cancer.

With the agnostic approach we were thus able to detect and confirm previous findings. We believe the methods used are interesting and have a potential to fill an unmet need in times of increasing amounts of data. The methods can be further enhanced to reduce the risks of finding relationships that are random and to be able to investigate more relationships concurrently.

# Abstract

In epidemiology there has been a consistent effort to construct refined methods aiming towards the ability to draw causal inference between an exposure and an outcome. This thesis, partly inspired by the genome-wide association studies, has on the contrary strived towards exploration of data. The fundamental idea has been to decipher associations between one or many exposures with one or many outcomes.

Using population-based register data from Sweden, this thesis explored the association between ABO blood group and RhD status in 1,217 disease categories, the occurrence of transfusion-transmitted disease examining 1,155 disease categories, the spectrum of adverse events in tyrosine kinase-inhibitor treated patients with chronic phase chronic myeloid leukemia in 670 disease categories, and lastly the occurrence of familial aggregation of 60 cancerous disease. To account for multiple testing we use previously employed methods of adjustment.

In Paper I, using the large-scale donation-transfusion database, SCANDAT-3S with 8 million individuals, we identified 49 associations with ABO blood group and disease. Many associations were previously known but we identified a novel association of a protective role for blood group B, as compared to O, in suffering kidney stones. For RhD status, we identified only one disease after adjustment for multiple testing, namely pregnancy-induced hypertension.

In Paper II, which used the same database and the unique connection between blood donor, the blood product and the recipients of blood, we identified 15 disease categories that seemed to be transfusion-transmitted. Among them there were strong signals suggesting transmission in hepatitis virus and HIV. For most other findings, the effect sizes were small. A general conclusion was that the current practice in Sweden, regarding transfusion safety, seems acceptable in terms of the risk of transfusion-transmission of disease.

In Paper III, we used a database covering the full Swedish chronic myeloid leukemia-population diagnosed since 2002. In this study, also consisting of a matched control cohort, we identified 142 disease categories with increased incidence as compared to the control cohort. We also found 41 associations between tyrosine kinase inhibitors, used for the treatment of chronic myeloid leukemia, and a disease category. No unknown severe adverse events were found.

In Paper IV, we created 3.5 million pedigrees using the Multi-generation Register and explored cancerous disease clustering within pedigrees. We identified multiple cancer syndromes, e.g. BRCA1/2 and hereditary colon cancer.



The approach, agnostic in that we strive towards testing all possible hypotheses without prejudice, has the advantage of removing or at least reducing the researcher bias – where the research hypothesis is constrained by the environment of the individual researcher. The approach is mainly limited by the problems of misclassification of exposures and outcomes, the inability to optimally construct modelling for a large set of hypotheses, and by false discoveries. We have proposed some solutions to overcome these issues. A main aspect is that the method should not be used to draw inference but rather to generate hypothesis for future refined studies in a world with increasing amounts of high-resolution data.

## List of scientific papers in this thesis

- I. **Dahlén T**, Clements M, Zhao J, Olsson M L, Edgren G. An agnostic study of associations between ABO and RhD blood group and phenome-wide disease risk. *Elife* 10, e65658 (2021).
- II. **Dahlén T**, Zhao J, Busch PB, Edgren G. Searching for unknown transfusion-transmitted disease: a retrospective, nationwide cohort study using agnostic methods. Manuscript.
- III. **Dahlén T**, Edgren G, Ljungman P, Flygt H, Richter J, Olsson-Strömberg U, Wadenvik H, Dreimane A, Myhr-Eriksson K, Zhao J, Sjölander A, Höglund M, Stenke L. Adverse outcomes in chronic myeloid leukemia patients treated with tyrosine kinase inhibitors: Follow-up of patients diagnosed 2002–2017 in a complete coverage and nationwide agnostic register study. *Am. J. Hematol.* 97, 421–430 (2022).
- IV. **Dahlén T**, Magnusson PKE, Edgren G. Identification of cancer syndromes using family structure data. Manuscript.

## Scientific papers not included in this thesis

- I. **Dahlén T**, Kalin M, Cederlund K, Nordlander A, Björkholm M, Ljungman P, Blennow O. Decreased invasive fungal disease but no impact on overall survival by posaconazole compared to fluconazole prophylaxis: a retrospective cohort study in patients receiving induction therapy for acute myeloid leukaemia/myelodysplastic syndromes. *Eur J Haematol.* 2016 Feb;96(2):175–80.
- II. **Dahlén T**, Edgren G, Lambe M, Höglund M, Björkholm M, Sandin F, Själander A, Richter J, Olsson–Strömberg U, Ohm L, Bäck M, Stenke L; Swedish CML Group and the Swedish CML Register Group. Cardiovascular Events Associated With Use of Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia: A Population–Based Cohort Study. *Ann Intern Med.* 2016 Aug 2;165(3):161–6.
- III. **Dahlén T**, Zhao J, Magnusson PKE, Pawitan Y, Lavröd J, Edgren G. The frequency of misattributed paternity in Sweden is low and decreasing: A nationwide cohort study. *J Intern Med.* 2022 Jan;291(1):95–100.
- IV. **Dahlén T**, Li H, Nyberg F, Edgren G. A population–based, retrospective cohort study of the association between ABO blood group and risk of COVID–19. *J Intern Med.* 2022 Nov 13.
- V. Zhao J, **Dahlén T**, Brynolf A, Edgren G. Risk of hematological malignancy in blood donors: A nationwide cohort study. *Transfusion.* 2020 Nov;60(11):2591–2596.
- VI. Zhao J, **Dahlén T**, Edgren G. Costs associated with transfusion therapy in patients with myelodysplastic syndromes in Sweden: a nationwide retrospective cohort study. *Vox Sang.* 2021 May;116(5):581–590.
- VII. Shahim B, Redfors B, Lindman BR, Chen S, **Dahlen T**, Nazif T, Kapadia S, Gertz ZM, Crowley AC, Li D, Thourani VH, Kodali SK, Zajarias A, Babaliaros VC, Guyton RA, Elmariah S, Herrmann HC, Cohen DJ, Mack MJ, Smith CR, Leon MB, George I. Neutrophil-to-Lymphocyte Ratios in Patients Undergoing Aortic Valve Replacement: The PARTNER Trials and Registries. *J Am Heart Assoc.* 2022 Jun 7;11(11):e024091.
- VIII. Söderlund S, **Dahlén T**, Sandin F, Olsson–Strömberg U, Creignou M, Dreimane A, Lübking A, Markevärn B, Själander A, Wadenvik H, Stenke L, Richter J, Höglund M. Advanced phase chronic myeloid leukaemia (CML) in the tyrosine kinase inhibitor era – a report from the Swedish CML register. *Eur J Haematol.* 2017 Jan;98(1):57–66.
- IX. Flygt H, Sandin F, **Dahlén T**, Dreimane A, Lübking A, Markevärn B, Myhr–Eriksson K, Olsson K, Olsson–Strömberg U, Själander A, Söderlund S, Wennström L, Wadenvik H, Stenke L, Höglund M, Richter J. Successful tyrosine kinase inhibitor discontinuation outside clinical trials – data from the population–based Swedish chronic myeloid leukaemia registry. *Br J Haematol.* 2021 Jun;193(5):915–921.

- X. Hirt C, Iannazzo S, Chirolì S, McGarry LJ, le Coutre P, Stenke L, **Dahlén T**, Lipton JH. Cost Effectiveness of the Third-Generation Tyrosine Kinase Inhibitor (TKI) Ponatinib, vs. Second-Generation TKIs or Stem Cell Transplant, as Third-Line Treatment for Chronic-Phase Chronic Myeloid Leukemia. *Appl Health Econ Health Policy*. 2019 Aug;17(4):555–567.
- XI. Zhang H, **Dahlén T**, Khan A, Edgren G, Rzhetsky A. Measurable health effects associated with the daylight saving time shift. *PLoS Comput Biol*. 2020 Jun 8;16(6):e1007927.
- XII. Jia G, Li Y, Zhang H, Chattopadhyay I, Boeck Jensen A, Blair DR, Davis L, Robinson PN, **Dahlén T**, Brunak S, Benson M, Edgren G, Cox NJ, Gao X, Rzhetsky A. Estimating heritability and genetic correlations from large health datasets in the absence of genetic data. *Nat Commun*. 2019 Dec 3;10(1):5508.

# CONTENTS

1	INTRODUCTION.....	4
2	BACKGROUND.....	5
2.1	Blood donation and transfusion.....	5
2.2	Blood groups and their relation to disease.....	9
2.2.1	ABO blood group and disease.....	9
2.2.2	RhD status and relation to disease.....	10
2.3	Blood safety and transfusion-transmitted disease.....	11
2.3.1	Blood safety within a historical context.....	11
2.3.2	Emerging transfusion-transmitted disease.....	14
2.4	Chronic myeloid leukemia.....	14
2.4.1	Adverse events related to tyrosine kinase inhibitors.....	16
2.5	Hereditary cancer syndromes.....	16
2.6	The agnostic approach.....	17
2.6.1	The issue of multiple testing.....	17
3	RESEARCH AIMS.....	21
4	MATERIALS AND METHODS.....	22
4.1	Paper I.....	22
4.1.1	Data sources.....	22
4.1.2	Study design.....	24
4.1.3	Statistical approach.....	25
4.2	Paper II.....	26
4.2.1	Data sources.....	26
4.2.2	Study design.....	26
4.2.3	Statistical approach.....	28
4.3	Paper III.....	28
4.3.1	Data sources.....	28
4.3.2	Study design.....	29
4.3.3	Statistical approach.....	30
4.4	Paper IV.....	31
4.4.1	Data sources.....	31
4.4.2	Study design.....	31
4.4.3	Statistical approach.....	32
5	ETHICAL CONSIDERATIONS.....	33
6	RESULTS.....	35
6.1	Paper I.....	35
6.1.1	Study population and baseline characteristics.....	35
6.1.2	Main findings in regard to ABO and disease.....	35
6.1.3	Main findings in regard to RhD status and disease.....	35

6.2	Paper II.....	39
6.2.1	Study population and baseline characteristics .....	39
6.2.2	Main findings in terms of possible transfusion-transmitted disease.....	39
6.3	Paper III.....	45
6.3.1	Study population and baseline characteristics .....	45
6.3.2	Main findings in terms of adverse events in comparison to the control cohort .....	45
6.3.3	Main findings in terms of adverse events within the CML cohort in terms of different TKIs.....	45
6.4	Paper IV .....	48
6.4.1	Study population and baseline characteristics .....	48
6.4.2	Main findings in terms of familial disease clustering.....	48
7	DISCUSSION.....	53
7.1	Random error .....	54
7.1.1	Statistical estimation.....	54
7.1.2	Multiple testing and P-values .....	54
7.2	Selection bias, information bias and confounding.....	55
7.2.1	The healthy-donor effect.....	55
7.2.2	Misclassification and measurement error of outcomes and exposure information.....	55
7.2.3	Selection of a control cohort and detection bias .....	57
7.2.4	The increased mortality of transfused patients.....	57
7.2.5	Allocation of blood units for transfusion.....	57
7.2.6	Confounding by indication.....	58
7.3	Computational limitations.....	58
8	CONCLUSIONS.....	60
9	POINTS OF PERSPECTIVE.....	61
9.1	Future studies.....	61
9.1.1	The association between blood groups and disease.....	61
9.1.2	Transfusion-transmitted disease .....	61
9.1.3	Adverse events in CP-CML patients.....	61
9.1.4	Familial disease clustering.....	61
9.2	Development of the agnostic approach.....	61
10	ACKNOWLEDGEMENTS .....	62
11	REFERENCES .....	63

## List of abbreviations

AP-CML	Accelerated phase chronic myeloid leukemia
BP-CML	Blastic phase chronic myeloid leukemia
CML	Chronic myeloid leukemia
CP-CML	Chronic phase chronic myeloid leukemia
CPU	Central processing unit
DES	Disease excess score
FDR	False-discovery rate
GPU	Graphics processing unit
HBV	Hepatitis B virus
HCV	Hepatitis C virus
HR	Hazard ratio
NAT	Nucleic acid test
IRR	Incidence rate ratio
ICD	International Classification of Diseases
TKI	Tyrosine kinase inhibitor
SCANDAT	Scandinavian donations and transfusion database
SCANDAT-3S	Scandinavian donations and transfusion database Swedish part of iteration 3.
WBD	Whole-blood donation

# 1 INTRODUCTION

In the field of epidemiology, there has been a consistent aim to carefully construct studies with a single hypothesis using retrospective or prospective data to draw inference relating an exposure to a specific outcome. Utilizing a constantly refined conceptual framework, together with robust and fitting statistical modelling, the strength of any such association has become more precise with methodology that even allows for casual inference.<sup>1</sup> However, the classical approach, investigating a limited and pre-set number of possible associations to derail their relationship to a disease or condition, has some obvious limitations. With increasing amounts of data, and the possibility of a near infinite number of associations, there is need for a new complementary approach that can ask all questions and estimate the likelihood of an actual association. The findings can then be further investigated in a more precise modelling framework.

The genetical sciences have advanced in this direction, due to the enormous amount of data generated from genome expression studies. These studies test 1,000, 100,000 or even millions of single gene mutations against a single hypothesis.<sup>2</sup> This development introduces the idea for the *agnostic* approach in the non-genetic epidemiological setting, where an undefined number of possible factors with associations to a specific or unspecific number of diseases are investigated without any preconceived ideas. In addition to the increasing amount of high-resolution data available today, a significant limitation of the classical epidemiological approach is the inherent *researcher bias* that limits the questions to be asked by the sometimes narrow environment of the individual researcher. Further, strong associations may exist that would never be investigated as a consequence of an unexpected or complex relationship that may far outreach any plausible current biological or clinical understanding of a disease.

This thesis explores the agnostic approach, in a deep-fishing environment, using multiple sources of data with varying resolution to study factors and their possible association with a large set of disease. In detail, we study the blood groups and their association with disease, the possibility of an unknown transfusions-transmitted agent, adverse events in patients treated with tyrosine-kinase inhibitors in chronic myeloid leukemia in chronic phase and lastly, we try to decipher familial aggregation of malignant disease using non-genetic relational data – all using similar or an inherently conceptual alike agnostic approach.



## 2 BACKGROUND

Due to the heterogenous nature of this thesis, in terms of the underlying topics of investigation, the background has been divided into multiple parts to give an introductory orientation to all topics. In section 2.1 through 2.3 the reader is introduced to blood transfusions and blood groups related to Paper I and II. In section 2.4 there is information on CML and its treatment related to Paper III. Section 2.5 introduces hereditary, mostly mono-genic, familial cancer syndromes explored in Paper IV. Lastly, section 2.6, gives some details regarding the multiple testing issue which is central to the agnostic approach.

### 2.1 Blood donation and transfusion

*Tab. 1.*



*Figure 1. Figure from James Blundell's publication in The Lancet regarding early observations in human-to-human blood transfusion. Reprinted under creative common licensing.*

The origin of the practice of human-to-human blood transfusion is debated. Many regard a reported transfusion in 1818 by the obstetrician James Blundell to be the first.<sup>3,4</sup> The transfused patient was moribund in gastric carcinoma and died two days after receiving the transfusion. Earlier, non-human to human transfusions, xenotransfusions, were conducted with little benefit and multiple complications.<sup>5</sup> In 1901, Karl Landsteiner identified one of the major obstacles of transfusion medicine, namely the ABO blood group antigens, paving the way to ABO compatible transfusions.<sup>6</sup> Since then, much has happened, and blood transfusion has become a cornerstone in modern medicine and one of the more common medical interventions. Nearly 20% of the population can expect to receive a transfusion at any time during their lifespan. This translates into

more than 400,000 units of red blood cells transfused annually in Sweden and more than 250 million blood units are transfused globally on an annual basis.<sup>7-10</sup>

Since the discovery of the agglutination, in the case of ABO blood group incompatibility, many aspects of transfusion medicine have been studied and optimized to reduce risks related to donating and receiving blood – from the selection of donors, phlebotomy technique, separation, storage and handling, to matching blood components with the recipients and evaluation of early blood transfusion reactions. Given the amount of blood donated, and hence transfused, inconsistencies in a few transfused units may have severe impact on the individual receiving a blood transfusion. This renders all aspects of blood transfusion safety important.

In Sweden blood is collected by two separate methods, whole-blood donation (WBD) and apheresis. WBD is the most common source and represents 93 % of all donated blood.<sup>11</sup> During WBD, blood is drawn and collected in a bag and then processed by centrifugation to separate red-blood cells from plasma and platelets. During centrifugation a layer with leukocytes and platelets accumulates on top of the erythrocytes, called the “Buffy coat”.<sup>12</sup>

In apheresis donation, blood is directly processed and centrifuged, allowing the extraction of one specific component at the time and leaving the remaining components returned to the donor. This, in turn, enables the possibility to donate platelets and plasma more frequently than when using WBD.

To increase storage times, limit transfusion-transmitted disease and contamination of bacteria (usually from skin bacteria at venepuncture), some blood components, including platelets, are processed with pathogen-inactivation strategies utilising different chemical substances and ultraviolet irradiation.<sup>13</sup> Today, leukoreduction, the removal or reduction of the “Buffy-coat” layer, is conducted in all blood collected using chemical and/or mechanical filtration with the intent to decrease alloantibody formation, decrease the risk of transfusion reactions and to possibly remove some viruses.<sup>14,15</sup> However, the benefit of leukoreduction has not been uniformly demonstrated for all types of blood components. The process also renders a not so insignificant proportion on the blood unusable depending on the specific method used.<sup>15,16</sup>

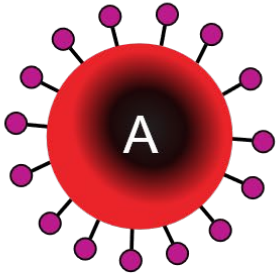
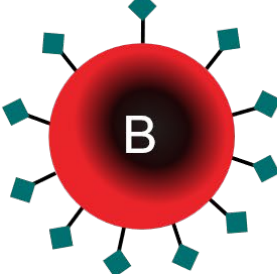
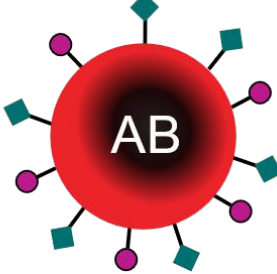
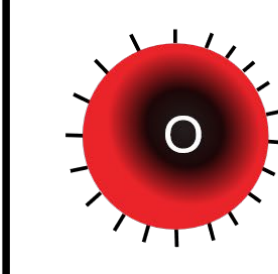
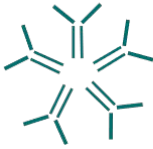

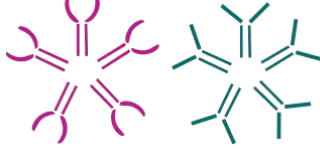



	Group A	Group B	Group AB	Group O
Red blood cell type				
Antibodies in plasma	 Anti-B	 Anti-A	None	 Anti-A and Anti-B
Antigens in red blood cell	 A antigen	 B antigen	 A and B antigens	None

Figure 2. ABO blood group system with the corresponding antigens. Blood group incompatibility arises when recipient antibodies are directed against transfused red blood cells blood group.

An uncommon, but highly fatal complication of blood transfusion is transfusion-associated graft-versus-host disease, in where it's believed that donor infused T-lymphocytes engraft and cause graft-versus-host disease in severely immunocompromised recipients.<sup>17</sup> To hinder this, units of blood transfused to such individuals are irradiated before use in order to destruct these lymphocytes.<sup>18</sup> In patients with repeated acute transfusion reactions of allergic and anaphylactic origin, IgA deficient individuals or patients at risk of severe hyperkalaemia, blood products are "washed" to remove the plasma components and unknown allergic stimuli, such as IgA and potassium, with clinically demonstrated efficiency.<sup>19</sup>

Red blood cells can be stored 0 to 42 days in low temperature in preservative solutions. The optimal length of storage of transfused blood has been widely debated and tested in randomized clinical trials.<sup>20-24</sup> However, there is no clear evidence favouring fresh blood to older blood other than mechanistic storage lesions.<sup>25,26</sup> Platelets can be stored for a significantly shorter period, usually ranging from 5-7 days, depending on collection technique and processing due to the short survival of platelets.<sup>27,28</sup> Platelets must also be stored at no less than 20 degrees Celsius with a constant movement rendering them especially prone to bacterial contamination.<sup>29</sup> Plasma can be frozen and stored up to a year, or longer.<sup>25</sup>

Blood usage has steadily declined in Sweden since late 1990s, from 600,000 units to around 400,000 units per year.<sup>8</sup> The decrease may partly be explained by decreased amounts of red blood cells transfusions due to increasing evidence of comparable effectiveness and safety in most patient populations of more restrictive as compared to liberal transfusion thresholds.<sup>30</sup> Red blood cells are indicated to treat haemorrhage, anaemia due to disease or due to therapy in symptomatic or prophylactic situations. Platelet transfusions are indicated for major haemorrhages, when large quantities of all blood components are lost, or more commonly due to disease or drug-induced severe thrombocytopenia or platelet dysfunction, both as prophylactic or therapeutic measures. Plasma is also indicated in major haemorrhage, and also in cases of deficiencies in coagulation factors or in select autoimmune conditions. Granulocyte transfusions can also be administered, but donation and transfusions are conducted using other principles and are therefore not covered in more detail in this thesis.

To monitor all parts of the process, from "vein-to-vein", modern national blood services implement hemovigilance systems to continuously evaluate and act on threats due to problems of the donation-transfusion chain.<sup>31,32</sup> These systems typically monitor short-term donor and recipient health parameters, such as emerging infectious threats and transfusion reactions, and are in Sweden a part of the national regulatory laws related to blood transfusions.

## 2.2 Blood groups and their relation to disease

### 2.2.1 ABO blood group and disease

Since the recognition of A, B and O individuals, the knowledge surrounding the ABO blood grouping system has markedly evolved. Further understanding has led to the identification of the A and B blood group antigens, yielding the four distinct blood group phenotypes depending on the expression of A and B surface antigens on red blood cells with the concomitant occurrence of the reciprocal anti-A or anti-B antibodies in non-new-borns.<sup>33-35</sup> The antibodies, typically of IgM type, are developed and produced in measurable quantities within the first year of life. The mechanism responsible for their development is disputed but the main hypothesis involve exposure to food or invading organisms, proposed by studies demonstrating highly elevated antibody titres after exposure to e.g., microbes has occurred.<sup>35,36</sup> Importantly, exposure to mismatched blood leads to antibody attachment and agglutination. Depending on exposed amount and reactivity this may in turn cause fatal intravascular haemolysis after complement fixation.<sup>36</sup>

In the early 20<sup>th</sup> century, a mendelian inheritance pattern was proposed by Epstein and Ottenberg and later described by Bernstein such as A and B follow a co-dominant inheritance with autosomal recessive inheritance of O.<sup>37</sup> Genotypic inheritance of AA or AO, BB or BO, OO and lastly AB thus produces the corresponding phenotypes, A, B, O and AB, respectively. The O gene product is non-functional as compared to A and B.<sup>38</sup> The genes coding for the ABO blood group antigens are located within a single locus on chromosome 9 and were first cloned in 1990.<sup>38,39</sup> Molecular aspects have been studied extensively and since then more than 200 distinct alleles have been discovered accounting for phenotypic differences in factors such as antigen expression level and weak and strong antigenic reactions.<sup>37</sup> However, most individuals express the same antigens with minorities demonstrating other alleles, e.g. 99% of the A expression is explained by either A<sub>1</sub> or A<sub>2</sub> allele.

The ABO surface antigens are a construct of carbohydrate chains and proteins residing on the surface of red blood cells but also in many other tissues. The AB genes code for glycosyltransferases that in term produce the antigen chains anchored to the surface proteins from the substrate H located on chromosome 19.<sup>37</sup> This dependence on substrate H for further processing has led to the discovery of very rare instances of individuals missing genes coding for substrate H which result in an O phenotype independent on ABO genotype, but with sera that also contain anti-H antibodies which can lead to strong haemolytic reactions if transfused with normal O blood. Since its discovery in Bombay a particular such blood group phenotype was named Bombay.<sup>38</sup>

The ABO blood group system has a variable distribution in different ethnic and geographical regions of the world.<sup>40–46</sup> One reason for the diversity in distribution of blood groups may be accounted for by phenotypic properties of different blood groups in relation to acquiring and harbouring disease-causing agents. As such this may ultimately impact the clinical disease severity, resulting in differences in blood group distributions depending on endemic disease patterns around the world. As such, for multiple diseases, the occurrence and modification of disease severity has been investigated to be related to blood group.

A first study linking blood group distribution to disease was published in 1962, demonstrating a relation between ABO blood group and ischemic heart disease.<sup>47</sup> Multiple associations have thereafter been revealed covering numerous diseases.<sup>48–54</sup> A strong association has been observed for individuals with blood group O, who have a decreased risk of thromboembolic events and increased risks of selected hemorrhagic events.<sup>48,55,56</sup> This difference in risk, concerning thrombotic and hemorrhagic events, has been studied mechanistically and has, at least partly, been attributed to variability in levels of *Factor VIII* and *von Willebrand factor*, where ABO blood group may explain as much as 30% of this variability.<sup>56–60</sup> Another association, possibly accounting for differences in blood group distribution, includes associations with the infectious diseases *Plasmodium falciparum* malaria, *Helicobacter pylori* and *Vibrio cholera*.<sup>61</sup> In these infectious agents, multiple mechanistic aspects of pathogenesis, from microbe attachment and entry into cells to subsequent disease development and severity of disease, have been postulated or demonstrated in relation to ABO blood group.<sup>53,61–63</sup> Recently, as the SARS-CoV-2 pandemic has uncovered, an association between blood group A and an increased susceptibility to severe infection was described in early reports.<sup>52,64</sup> Since then, we, among others, have conducted large-scale studies of representative cohorts to investigate this initial finding confirming an increased risk for non-O as compared to O individuals in some settings.<sup>50,65</sup> Interestingly, this effect was attenuated in vaccinated individuals.<sup>50</sup> For multiple other conditions associations between blood group and disease occurrence or severity have been proposed or demonstrated. Examples include pancreatic cancer, gastric cancer, leptospirosis, acne vulgaris, cholesterol metabolism and ARDS after major trauma or sepsis.<sup>51,54,55,66–69</sup> However, several of these studies are conducted using retrospective data or using small study populations, sometimes yielding conflicting results.<sup>70,71</sup>

### **2.2.2 RhD status and relation to disease**

The Rh blood group system was given its name after experiments on *Macacus rhesus* monkey.<sup>72</sup> The blood group system consists of multiple antigens named according to the International Society of Blood Transfusion CDE nomenclature. Whereas phenotype

and genotype relations are easily determined in the ABO system the Rh system is genetically complex. With many polymorphisms of the genes involved, RHD and RHCE, genotype–phenotype relations cannot be derived without errors.<sup>73</sup> However, regarding the specific RhD antigen within the Rh blood group system, it follows a Mendelian inheritance with RhD positive phenotype consisting of DD or Dd and RhD negative dd.

Clinically, the D antigen is central due to its inherent immunogenicity. The RhD antigen was discovered to cause sensitisation in RhD negative mothers with RhD positive fetus or fetuses. A subsequent pregnancy with RhD positive fetus may then induce a haemolytic disease in the infant and newborn.<sup>74</sup> Some interest has also been devoted to the possibility of multiple populations of red blood cells expressing different RhD phenotypes in patients with myeloproliferative neoplasms.<sup>73</sup>

As in the case of ABO blood group there is also a diversity in the frequency distribution of RhD positivity and negativity. Contrary to the ABO blood group system, a clear connection between RhD and specific non immunoreactive–related disease has not been identified. The belief is that RhD status is impacted by migration and genetic drift rather than by environmental selection due to endemic disease.<sup>75</sup> However, the possibility of connections between disease and RhD status has not been studied in large databases in a systematic manner.

## **2.3 Blood safety and transfusion–transmitted disease**

Besides the immunogenic properties of transfused blood, one of the major obstacles in transfusion safety has been, and is still, the possibility of transmission of disease from the donor to the recipient. Donor screening, by specific blood testing and questionnaires outlining current health and recent geographical exposures, is a measure conducted to reduce such possibilities. Currently, two of the most problematic issues are the possibility of a donor with pre–symptomatic or silent disease carrier state or during outbreaks of an unknown or new disease.<sup>76,77</sup> Below is an outline of historical, as well as current, aspects of transfusion–transmitted disease with a Swedish emphasis.

### **2.3.1 Blood safety within a historical context**

Syphilis was one of the first diseases to be identified as transfusion–transmitted.<sup>78</sup> Donor screening began in the 1940s in U.S., however, the screening tools suffered from low sensitivity in mainly early phases of the disease. However, when blood was collected and stored the bacteria responsible for syphilis, *Treponema pallidum*, seldom survived as compared to the direct transfusion conducted before *World War II* where the donor and recipient were connected, and transfusion was conducted *ad hoc*.<sup>79,80</sup> All donors in

Sweden are screened for syphilis at every donation occasion, because of the latency period before antibody reaction certain cases could possibly be missed. However, due to the low prevalence of syphilis, the low temperature storage of red blood cells and the oxygen-rich storage of platelets, survival and transmission remains unlikely.<sup>81</sup>

Transmission of hepatitis b virus and hepatitis c virus, causing hepatitis, cirrhosis, hepatocellular carcinoma and multiple non-liver related conditions, was until 1993 one of the greatest obstacles in transfusion medicine.<sup>82-84</sup> Screening for HBV was introduced in Sweden between 1970-1972. HCV screening, using serology testing, was introduced between 1989 and 1991, and has effectively prevented transmission. There is, however, a strong suspicion that there are individuals alive today with an undiagnosed transfusion-transmitted HCV infection, and possible hepatitis, because of blood received before identification of the virus or fully developed screening tools were in place.<sup>85</sup> Most developed countries utilise sensitive nucleic acid testing for HCV donor screening.<sup>77,85</sup> In the US, where NAT tests have been used since 1999, the risk of HCV transmission is estimated to be 1 in every 2 million blood units transfused.<sup>86</sup> A similar risk was detected in a UK surveillance study.<sup>87</sup> In Italy, also utilising NAT, a more recent study estimated the risk to be as low as 1 in every 13 million donations.<sup>88</sup>

Human immunodeficiency virus, consisting of HIV-1 and HIV-2, causing the acquired-immunodeficiency syndrome with a rich spectrum of infectious complications and multiple mostly haematological malignancies, was first identified in 1983.<sup>89,90</sup> The virus was early identified to be transfusion-transmitted, however, affecting mostly individuals with bleeding disorders as factor concentrates were constructed using pooled blood from up to 100 to 1,000 different donors increasing the risk of receiving blood from a HIV positive donor.<sup>91</sup> In Sweden, serological testing is conducted at every donation beginning in 1985. Since first introduced, no single case of transmission has occurred in Sweden.<sup>11</sup> As in the case of HBV and HCV, most other developed countries have implemented NAT decreasing the undetectable window from 14-19 days to 3-6 days since timepoint of infection and depending on specific technique and cut-off levels employed.<sup>92</sup> A small number of cases of transmission within the NAT "undetectable window" have been reported world-wide emphasizing the need to educate and question donors before each donation.

Human T-cell leukaemia virus, the first retrovirus discovered before HIV, is a group of two oncogenic viruses with long latency period and responsible for development of leukaemia, lymphoma, multiple inflammatory conditions, myelopathy and various immunodeficiency syndromes.<sup>93</sup> Serology testing is performed only before the first donation, motivated by the low frequency of the virus in blood donors and in the general population. Only 18 new cases were found upon testing 550,000 individuals and screening is thus estimated to identify 1 individual each 7th year ultimately resulting in 1 death each 200 years.<sup>94,95</sup>



West-Nile virus, carried by mosquitos and most commonly causing asymptomatic disease but in select cases leads to fatal meningoencephalitis, has been demonstrated in the transfusion-transmission setting alarming the transfusion-medicine society.<sup>96</sup> However, outbreaks are highly seasonal and restricted to endemic areas. Hence, screening is performed in accordance with the local situation. In Sweden, with import cases from areas with outbreaks, no routine testing is conducted.<sup>97</sup> There are multiple other vector-carried viruses, bacteria's and protozoans were transfusion-transmission has been documented or is theoretically possible, including diseases such as Dengue virus, Chikungunya virus, Tick-borne encephalitis, *Borrelia burgdorferi* and malaria spp.<sup>98</sup> However, as of today, with a lack of throughout knowledge, known protective strategies and a generally low prevalence of these infections screening is not conducted. Screening is indicated for the protozoa *Trypizomana cruzi* causing Chagas disease, with possible cardiac arrhythmias as a consequence, in select cases for epidemiological reasons.<sup>99</sup>

Among the human herpesviruses it is, in some parts, unknown to what extent transfusion-transmission occurs. A partial explanation is the in general high seroprevalence of the viruses and unknown effect of pathogen reduction measures. No screening is performed regarding Human Herpesvirus simplex 1/2, Varicella-Zoster virus or Human Herpesvirus 8.<sup>98</sup> Cytomegalovirus is of concern for immunosuppressed individuals who can develop aggressive generalized disease. However, given the high seroprevalence, ranging from 50% to nearly 90%, depending mainly on age, the benefit of current insensitive serology testing remains unknown.<sup>100,101</sup> NATs are not possible to use because the virus resides in monocytes with questionable shedding of virus into the blood. Leukoreduction removes some but not all the virus.<sup>102</sup> Similarly, this also applies to Epstein-Barr virus where possible transmission is suggested in highly immunocompromised individuals, seroprevalence is high and leukoreduction removes the greater amount of virus particles.<sup>103,104</sup>

Because the very nature of infectious diseases they deserve a central position in transfusion safety. However, blood products could possibly transmit any disease-causing factor generating disease states such as haemolytic febrile reactions, haemolytic transfusion reactions and transfusion-induced acute lung injury (TRALI).<sup>105</sup> The prion disease, variant Creutzfeldt-Jakob disease, causing an outbreak in the UK in the 1990s, has led to several cases of documented transfusion-transmission.<sup>106</sup> Based on these observations taken together with the similar protein misfolding that occurs in multiple neurodegenerative disease, a recent observational study investigated the possibility of transfusion-transmission of such diseases. This study found no evidence of possible transfusion-transmission of Alzheimer disease, Parkinson's disease or dementia.<sup>107</sup> Previously, using the same data and principally the same method, chronic

lymphocytic leukaemia showed no evidence of being transmissible through blood transfusion.<sup>108</sup>

### **2.3.2 Emerging transfusion-transmitted disease**

There have been several emerging threats of transfusion-transmitted infections after the hepatitis viruses and HIV. Recently, the Zika virus, after the discovery of the congenital Zika syndrome in Brazil, has been debated in terms of what preventive measures should be conducted given the risk of disease in recipients and costs of blood screening.<sup>109–111</sup> Regarding Covid-19, albeit initial worrying reports, respiratory human-to-human transmissible viruses have not been clearly demonstrated to be transfusion-transmissible, however, the pandemic has posed entirely different concerns in terms of blood safety.<sup>112–114</sup>

Another discussion that again has surfaced is the possibility of transfusion-transmission of human papillomaviruses and its impact in recipients given the serotype-dependent oncogenic potential.<sup>115</sup> The case for other arboviruses, after the discovery of transfusion-transmission of West-Nile virus, such as Dengue and Chikungunya has also been discussed. The preventive measures have in some cases been to utilise NATs but more generally led to epidemiological measures limiting the possibility of donors with known recent exposure to donate blood.<sup>116</sup>

## **2.4 Chronic myeloid leukemia**

Chronic myeloid leukemia is a hematological malignancy that in its initial chronic phase is dependent on a continuously activated tyrosine kinase.<sup>117</sup> In most cases, this is a result of a translocation between chromosome 9 and 22, resulting in the gene product BCR::ABL1 (Figure 3).<sup>118</sup> Since the discovery of the translocation in 1960, dubbed the Philadelphia chromosome, more fine-tuned molecular methods could in 1970s and 1980s characterize the resulting tyrosine kinase.<sup>117</sup> The disease manifests as asymptomatic in half of the patients at diagnosis in chronic phase, discovered usually by a blood sample demonstrating peripheral leukocytosis with leukocytes in earlier development stages visible in blood, various degrees of anemia and thrombocytosis or thrombocytopenia.<sup>119</sup> Before the introduction of targeted therapy disease progression from chronic phase to more advanced phases, accelerated phase or the highly fatal acute leukemic blast phase, would inevitably occur within a few years.<sup>120</sup> The symptoms from the disease, extramedullary hematopoiesis with markedly enlarged spleen and liver, and bone-marrow depression due to the loss of healthy hematopoiesis, with severe anemia, thrombocytopenia and granulocytopenia, would gradually develop. Increased blood viscosity can occur from the highly proliferative large white blood cells, blastic

cells or bone-marrow pre-cursors, resulting in life-threatening symptoms with increased bleeding, stroke or deteriorating dyspnea.<sup>121</sup>

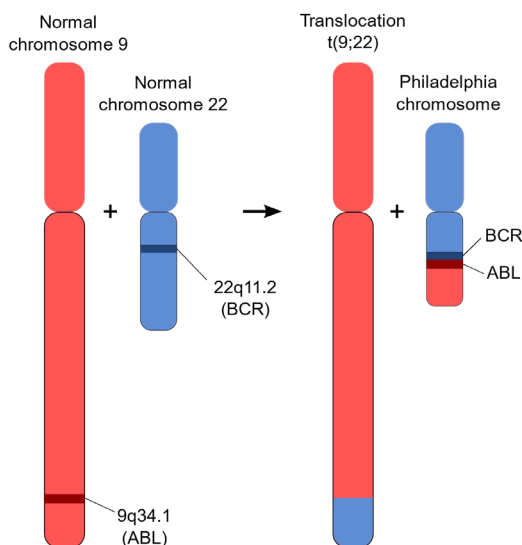


Figure 3. Schematic representation of the common CML-driving translocation generating the Philadelphia chromosome. Reused with permission under the creative-common license (CC-BY-SA-3.0, <https://creativecommons.org/licenses/by-sa/3.0/>).

In the late 1990s the first clinical trial with a tyrosine kinase inhibitor was conducted, as a consequence of the knowledge of the disease-driving translocation in chronic phase.<sup>122–126</sup> The drug, initially called ST1571 and later imatinib, targeted the BCR::ABL1 activated tyrosine kinase revolutionizing treatment and survival in CP-CML.<sup>119,127</sup> Register-based data from Sweden demonstrates relative patient survival rates matching the general population for most age groups.<sup>128</sup>

Due to drug intolerance and occurrence of specific mutations less sensitive or even resistant to imatinib, multiple other TKIs have been developed. The TKIs primarily differs in kinase-binding domain, potency and specificity for kinase inhibition.<sup>129</sup> The newer TKIs, dasatinib, nilotinib, bosutinib and ponatinib, have in clinical trials been demonstrated to generally result in earlier and deeper molecular response, measured by standardized real-time reverse-transcription polymerase chain reaction of BCR::ABL1.<sup>130–133</sup> However, treatment using these 2nd and 3rd line TKIs have not clearly demonstrated increased efficacy in terms of overall survival.<sup>126,134,135</sup> Despite the success of these drugs, there are still considerable issues in the management of specific patients, allowing for continued development of new drugs and treatment strategies.<sup>136</sup>

Besides the introduction of new drugs, it has also been discovered that a selected group of patients can discontinue treatment with TKI therapy after maintaining a highly reduced disease burden for a considerable time.<sup>137</sup> The success for this selected

group of patients, maintaining a small but constant molecular disease, was initially demonstrated in the STIM-study and later in the large pan-European collaboration, the EURO-SKI.<sup>138,139</sup> For patients starting TKI therapy today, however, only a minority will become eligible for TKI discontinuation in the future, and as such long term TKI safety is a key element in the management of CP-CML.<sup>140</sup>

#### **2.4.1 Adverse events related to tyrosine kinase inhibitors**

Since the introduction of TKIs, a main aspect in management of CP-CML patients has been to reduce the negative impact of continuous TKI therapy. From adverse event data from clinical trials and a close co-operation within the CML community, guidelines have been developed to limit the toxicity of TKI treatment. They are mainly focused on selecting the best TKI for a specific patient and comorbid factors together with management recommendations for patients experiencing toxicities.<sup>141,142</sup>

However, due to the relatively short follow-up conducted within randomized clinical trials, together with small study populations in clinical trials of CML, long-term adverse events have not been determined in a systematic manner. A concern of cardiovascular adverse events arose in 2013 after reports of such events in patients treated with nilotinib.<sup>143,144</sup> Later the phase 3 trial of ponatinib was stopped early because of discovery of a high rate of vascular events in a previous phase 2 trial.<sup>131,133,145</sup>

Our research group, utilizing the unique Swedish CML register, demonstrated clearly increased risk of cardiovascular events related mainly to nilotinib as compared to imatinib treatment.<sup>146</sup> The 5-year data from the randomized controlled trial on nilotinib treatment, when retrospectively analyzed for cardiovascular events, also demonstrated an increased risk of cardiovascular events in patients receiving nilotinib as compared to imatinib.<sup>134</sup> This may be a partial answer as to why the new generation TKIs have not demonstrated superiority in terms of overall survival. However, further research is needed to systematically study large patient populations to determine possible uncommon and late effects of TKI treatment in the real-world setting.

## **2.5 Hereditary cancer syndromes**

With increasing life expectancy and advances in treatment of cardiovascular disease, cancer is becoming one of the leading causes of death and morbidity in the world.<sup>147-149</sup> Most cases of cancer are believed to be acquired and attributed to multiple alterations in genes with key functions in cell survival, apoptosis and proliferation.<sup>150</sup> There are, however, a number of cancer syndromes where individual families harbor highly penetrant mutations and exhibit familial risks that are several orders of magnitude

higher than one would expect from estimates of average heritability. Examples of such genes are the BRCA 1 and 2, which are associated with highly elevated risks of breast and ovarian cancer.<sup>151,152</sup> If individuals carrying these genes are identified early in life, it may be sufficient with interventions such as prophylactic mastectomy and even prophylactic hysterectomy and increased screening rather than therapeutic interventions.<sup>153</sup> Other examples of cancer syndromes include Lynch syndrome, primarily giving rise to colorectal and endometrial cancer, Li–Fraumeni syndrome, arising from germline TP53 mutations with a vast clinical representation of different cancer subtypes, and Multiple endocrine neoplasia 1 and 2, with mainly endocrine gland malignancies. In total, there are approximately 100 cancer syndromes identified.<sup>154,155</sup> As proposed by the Two-hit hypothesis of oncogenesis, in the case of familial cancer versus sporadic cancer, a major difference between the two is that a single alteration already has occurred in germline cells. As such, a single further hit is needed to generate a cancerous cell, explaining the in general younger age of disease onset in familial cancer syndromes.<sup>156</sup>

Previously, highly penetrant cancer syndromes have been discovered by clinical identification of high-risk families. However, in the development of the genomic era, new syndromes have been discovered in disease groups where molecular genetic methods are employed in clinical practice. For example, in the field of hematology hereditary syndromes such as germline mutations in GATA2, RUNX1 and CEBPA, all yielding a range of hematological diseases such as myelodysplastic syndromes, acute leukemia and aplastic anemia, have more recently been discovered due to the availability of these methods.<sup>157,158</sup>

For some cancer regions, such as lung, liver and cervix, our etiologic understanding is so advanced, that their incidence could be substantially reduced by removing tobacco use, reduce alcohol abuse, and by implementing vaccination against hepatitis B and oncogenic human papillomaviruses.<sup>159–162</sup> However, in the case for other cancer diseases there is substantial differences in incidence that are beyond our knowledge. One example is testicular cancer, which has increased several-fold since the mid-1900's without any clear explanation, or the unexplained and dramatic sex disparities seen in relation to cancer risk.<sup>163,164</sup> As such, it is likely that unidentified cancer syndromes exist.

## **2.6 The agnostic approach**

### **2.6.1 The issue of multiple testing**

Performing multiple hypothesis tests, which is central to the agnostic approach, increase the risk of false positive findings, that is type I errors. The number of expected type I

errors depend mainly on the pre-set alpha level and the number of tests conducted. The probability of finding at least one false positive result can be modelled by  $1 - (1 - \alpha)^{N_{tests}}$ , as seen in Figure 4.<sup>165,166</sup>

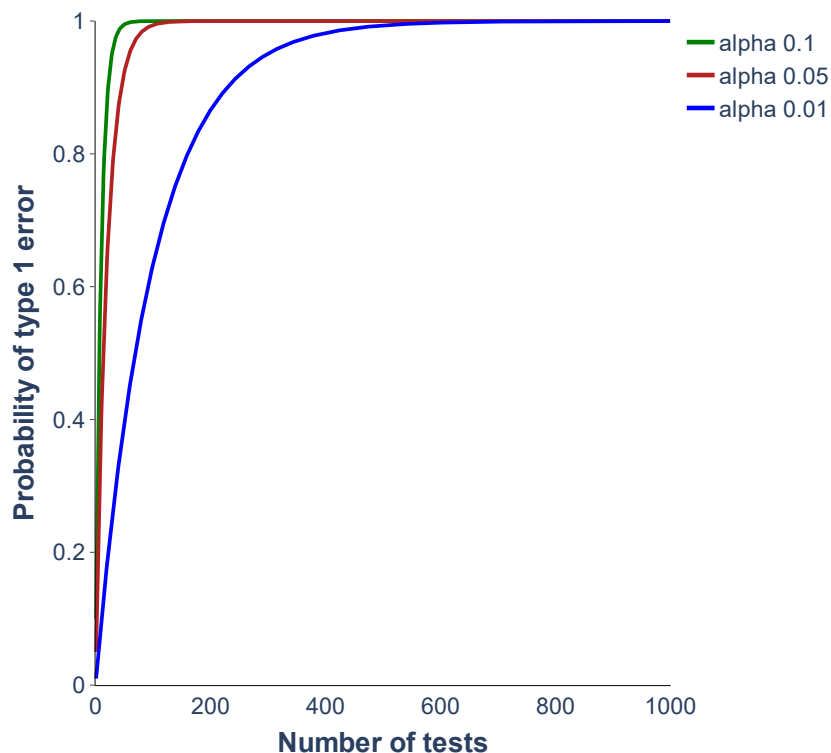


Figure 4. Probability of type 1 errors depending on the alpha level and number of tests conducted.

In genomic studies, it is not uncommon to perform hundreds, thousands, or even millions of tests with increasing number of type I errors.<sup>167</sup> Methods to decrease the probability of type I errors have been developed to counteract this problem. Visually, the P-values can be displayed in a distribution plot, and if there is truly no difference between the distributions of data within each individual test, then all P-values would be uniformly distributed as the first panel in Figure 5. However, in cases with actual differences there will be an enriched area or skewness to lower values demonstrating a true difference between data distributions as visualized in the second panel of Figure 5. Current methods to adjust for type 1 error are aimed at determining the true finding among all these findings.<sup>126</sup>

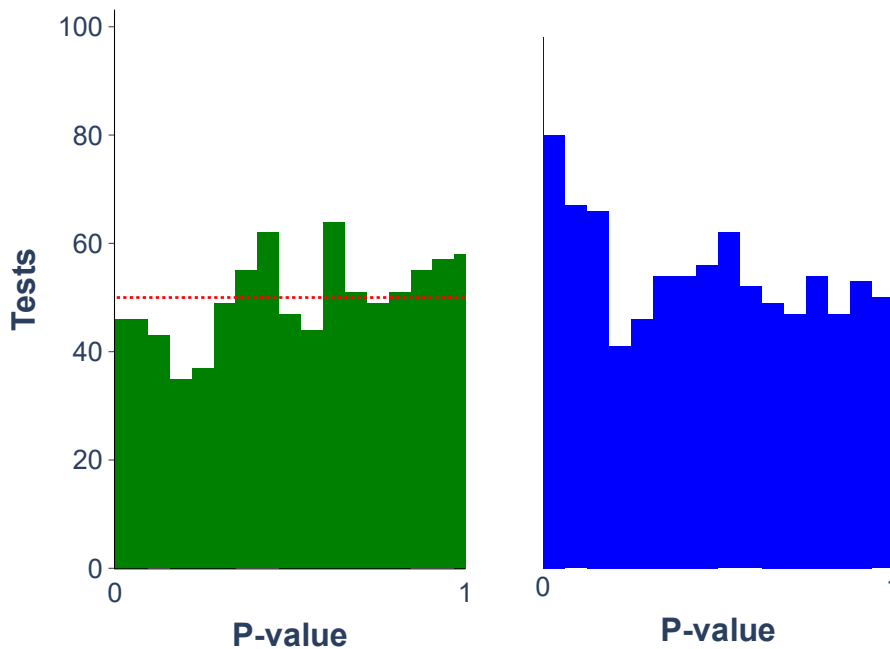


Figure 5. Distribution of P-values of tests left panel with a uniform distribution, right panel with a left skew.

There are several methods for handling multiple testing, but two distinct groups of approaches are widely used. The first one called the Familywise Error-Rate, includes the Bonferroni-method which by far is the most commonly employed although others exist.<sup>168</sup> The objective of the Bonferroni-method is to control the type I error rate and does so by setting the acceptance rate for significance at  $\alpha/n_{tests}$ . Bonferroni adjustment reduces the probability of type I error but rejects many true positives, thus reducing the power to detect possible relevant findings.<sup>169</sup>

The other approach that has become more commonly applied is instead to control the false discovery rate with the intention to control the expected proportion of type I errors among all rejected hypothesis.<sup>170</sup> The original FDR method is slightly more mathematically demanding and consists of multiple steps:

- a. Sort significant  $P$  -values  $p_1, p_2, \dots, p_n$  according to the smallest value.
- b. Find  $k$  such it is the largest index  $i$  for which  $p_i \leq d \times i/n$  for all  $i$  where  $d$  is the chosen rejection level.
- c. Declare all tests  $p_1, p_2, \dots, p_k$  significant.

The strengths of this FDR approach, developed by Benjamini and Hochberg, is that it addresses some of the shortcomings of the Bonferroni method. Most importantly it increases the power to detect a true positive among false positives.<sup>171</sup> A dilemma when encountering the problems of multiple tests is on what proportion of tests adjustment should be applied. A first analysis may be conducted with 800 independent tests, then a few days later or much later 800 further tests may be performed on the same data. The Bonferroni method would produce much different findings in this scenario, if

performed on 800 or the 1,600 tests. However, the FDR method would produce the same proportion of false discoveries independent of the number of tests conducted.<sup>171</sup> A pitfall with the FDR method is that it assumes independence between tests, which may not in an agnostic situation, be the case. Another issue of the approach is the methods ability to discriminate false positives and true positives when most hypotheses are truly false. Extensions of the models have been conducted, one such extension is the adaptive FDR, however resulting in less power when the opposite is true.<sup>172</sup> A more computationally intensive FDR method, but requiring no assumptions, has been developed but has not met widespread adoption. Instead of determining a fixed proportional error rate, which is set in the Benjamini and Hochberg method usually applying the “standard” 0.05 alpha, and an estimated rejection region this method instead determines a fixed rejection region and estimates the error rate.<sup>173</sup>



### 3 RESEARCH AIMS

The overall research aim of this thesis is to explore associations between factors and disease using an agnostic approach, more specifically:

In Paper I, we investigate the association between ABO blood group and RhD status and disease occurrence

In Paper II, we investigate the occurrence of transfusion-transmissible disease

In Paper III, we investigate if there are associations between TKI-treated CML patients and a large set of disease outcomes as compared to a matched control population

In Paper IV, we investigate the existence and clustering of hereditary cancer in the Swedish population using register-based data

## 4 MATERIALS AND METHODS

### 4.1 Paper I

#### 4.1.1 Data sources

In Paper I we utilized the Swedish part of the vein-to-vein transfusion database, the Scandinavian Donations and Transfusions database (SCANDAT-3S).<sup>8</sup> This database is depicted in Figure 6, in terms of linkages with other health care registers, and in Figure 7, in terms of timepoints of partial- or full coverage. The compiled database contains information on approximately 8 million unique Swedish individuals, comprising individuals who have undertaken a blood group antigen test before a blood donation, transfusion or for any other reason. In terms of data used for this particular study, in regards to the linked health care and populations statistical resources, we utilized the National Patient Register, the Swedish Cancer Register, and the Cause of Death Register to identify information on outcomes.<sup>174-176</sup> The blood donations and transfusions data was used for exposure information and definition of cohorts together with the Multi-generation register. Population statistics for baseline and follow-up information was retrieved from Total Population Register.<sup>177</sup> The blood group data was originally obtained, together with all other transfusion-related data, from the regional blood banks in Sweden.

To reduce the heterogeneity in ICD codes we limited the database to follow all events from 1<sup>st</sup> January 1997, when ICD revision 10 was used throughout Sweden. End of follow-up for all outcomes was 31<sup>st</sup> December 2017.

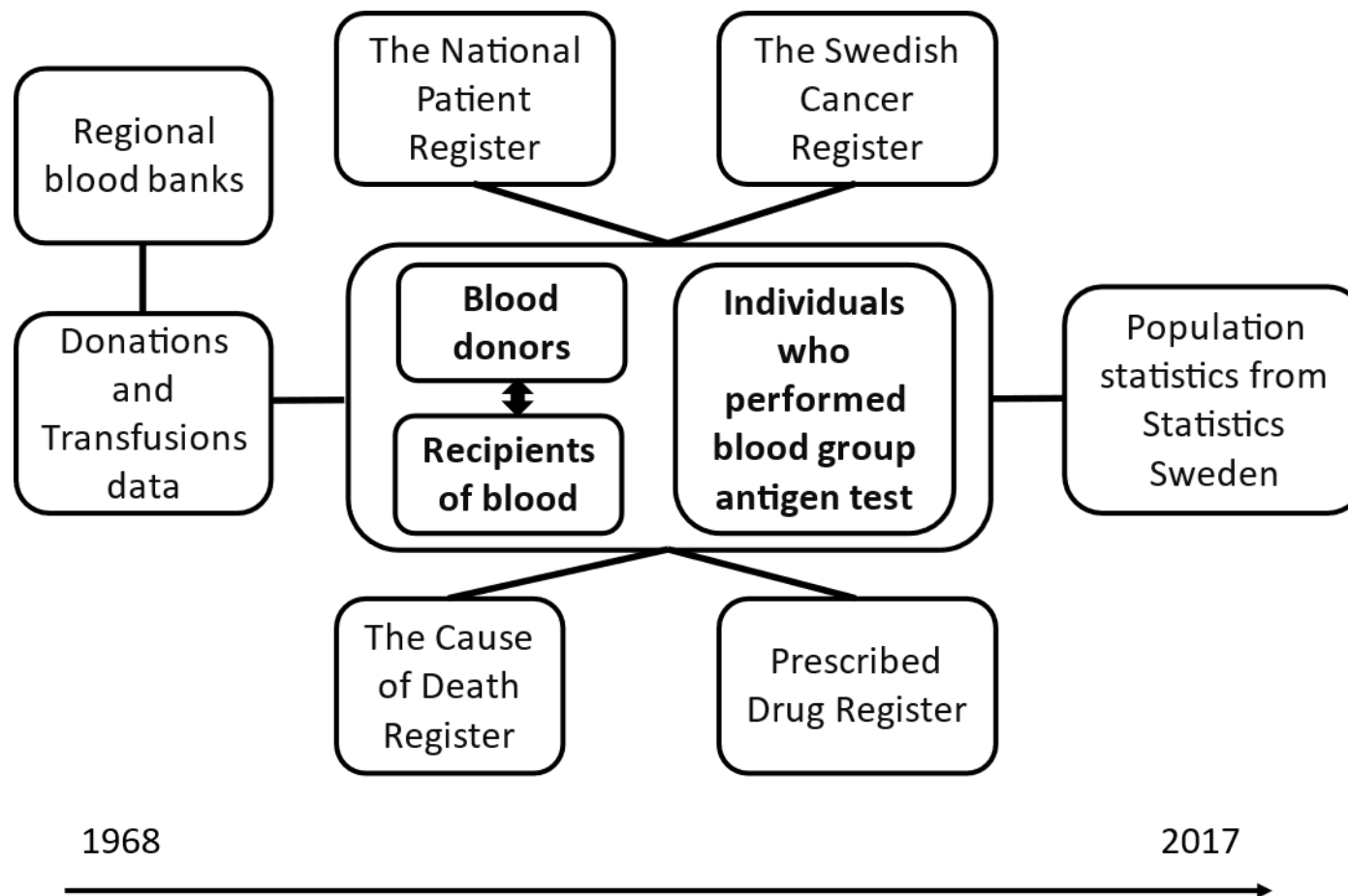


Figure 6. Overall structure of the SCANDAT-3S database.

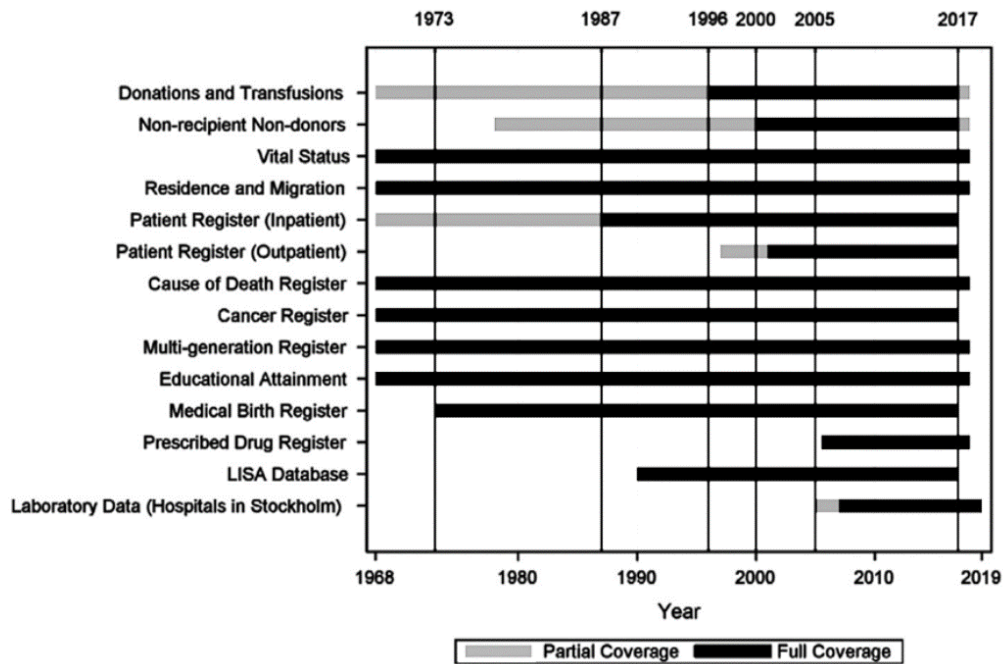


Figure 7. Coverage dates of the SCANDAT-3S database, published with permission from Jingcheng Zhao.<sup>8</sup>

#### 4.1.2 Study design

In Paper I we created a retrospective cohort study using a two-pronged approach. As such, we generated two separate study cohorts. Firstly we conducted an exploratory analysis and then a validation was conducted of the findings from the exploration, as show in Figure 8. The population in the exploratory cohort consisted of all individuals born in Sweden with at least 1 parent also born in Sweden and who had undergone a blood group antigen test and who did not donate blood within 90 days of that test. Any individual in the exploratory cohort, who donated blood after 90 days, was censored from this cohort at the time of blood donation. The validation cohort consisted of all blood donors. In the exploratory cohort, follow-up was started at the time of the first blood antigen test and for the validation cohort, at the time of the first blood group antigen test within 90 days from blood donation, as recorded in the SCANDAT-3S database, at the 18<sup>th</sup> birthday or 1<sup>st</sup> of January 1997, whichever occurred last. Individuals where then followed until emigration, 31<sup>st</sup> December 2017, death, or each incident event case for each disease category, whichever occurred first.

Disease outcome categories were created using disease categories from the ICD 10 for non-malignant disease, consisting of the 3 letter ICD-codes recorded in the National Patient Register. For malignant disease categories, a separate classification was used based on a previous publication.<sup>178</sup> In total, we constructed 1,217 disease categories but limited the analysis to disease categories with 50 or more events during follow-up.

In the exploratory cohort we performed regression analysis calculating incidence rate ratios comparing ABO blood group, using O as reference, and RhD status, with RhD negative as reference, in relation to each outcome. These findings were then adjusted for multiple testing. Significant outcomes from the exploration were then validated using the same approach in the validation cohort. Adjustment for multiple testing was again performed. All findings were presented, both before and after adjustment for multiple testing.

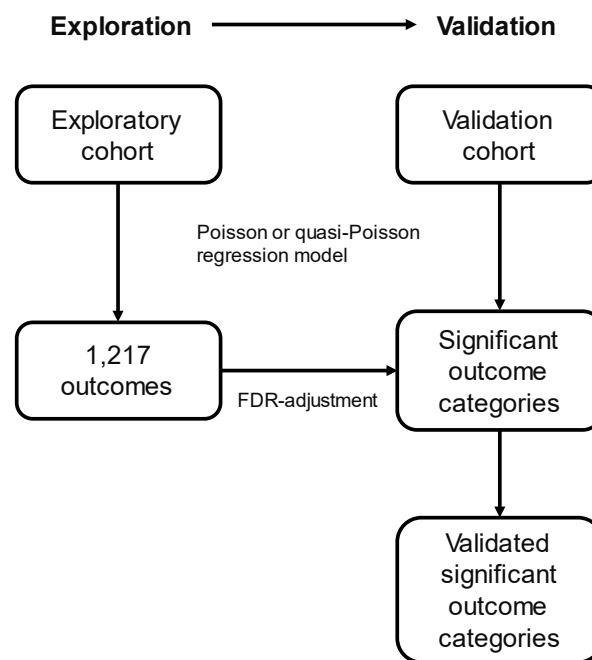


Figure 8. Overview of study design used in Paper I.

#### 4.1.3 Statistical approach

For the analysis, in both the exploratory and validation cohort, we constructed time-event tables using the SAS Stratify macro for each outcome category investigated.<sup>179</sup> Using these data we then initially fitted a Poisson regression-model for each outcome category, incorporating age (fitted using a restricted cubic spline function with 5 knots placed according to Harrell's method), calendar year (also fitted with the same restricted cubic spline function with 5 knots), sex (categorical) and ABO blood group (categorical with O as reference) and RhD status (categorical with RhD negative as reference) as covariates.<sup>180</sup> To account for Poisson regression assumptions, we used the Lagrange multiplier test to estimate occurrence of significant deviation from equi-dispersion. In models with significant under- or over-dispersion, we initially reduced the number of knots from 5 to 4 in the spline functions and again tested for under- or over-dispersion. In models with significant dispersion, we instead performed the analysis within a quasi-Poisson model. To account for multiple testing, we used FDR-adjustment in the exploratory analysis and Bonferroni-adjustment in the validation analysis.<sup>181</sup>

## 4.2 Paper II

### 4.2.1 Data sources

In Paper II, we used the same data sources as in Paper I. However, to capture all possible events we used the full scope of follow-up in the database for both donors and recipients of blood. Figure 6 and Figure 7, together with the information available in Section 4.1.1 above, a full description of the multiple parts of this database is provided. For this particular study, we thus used the linked health care and populations statistical resources to retrieve information on outcomes from the National Patient Register, the Swedish Cancer Register and the Cause of Death Register. The information on recipients of blood was used to identify the study cohort and the transfusion data, with linkages of each blood transfusion to a date and a specific donor, as exposure information. Data from the Total Population Register was used to obtain dates regarding baseline characteristics and follow-up.

### 4.2.2 Study design

The approach used to identify possible transfusion-transmitted disease was based on a similar agnostic framework that was used in Study I together with a model used in previous studies to identify transfusion-transmission of specific diseases, namely hepatitis c, neurodegenerative disease and chronic lymphocytic leukemia.<sup>85,107,108</sup> Figure 9 gives a brief overview of the study design. The study design was based on the rationale that there is evidence of disease transmission if recipients of blood from donors who develop a particular disease carry an increased risk of disease development as compared to recipients of blood from donors who do not develop that particular disease.

The study cohort included all patients in the SCANDAT-3S database who received allogeneic blood transfusion of either erythrocytes, platelets, plasma or whole blood. This meant that any individual with a recorded transfusion between 1968 and 2017 was included in the main cohort. Individuals were followed from the date of the first transfusion, death, emigration, until the last day of data in the database linkage, December 31<sup>st</sup> 2017 or until an event occurred in the investigated disease category.

As in Paper I, we constructed disease outcome groups using the ICD 10 disease categories. However, during the study period ICD 8 through 10 was used during follow-up. As such, non-malignant disease categories were created mapping disease coding from ICD 8 and ICD 9 to ICD 10. This mapping involved extensive manual translation of mainly ICD 8 to 10 together with the available translation of ICD 9 to 10 from the National Board of Health and Welfare.<sup>182</sup> To generate the mapping, an inclusive search was conducted using PubMed, including all terms relating to the Swedish

National Patient Register. All articles were then retrieved and information regarding ICD coding and disease groups were extracted. This was later used in parts to generate translations using previously used definitions. There is an ongoing validation of this ICD translation, however, the SAS code to generate the disease grouping is available upon request but is currently un-commented. For malignant disease, we used the same disease grouping as in Paper I. In total, we constructed 1,155 disease categories.

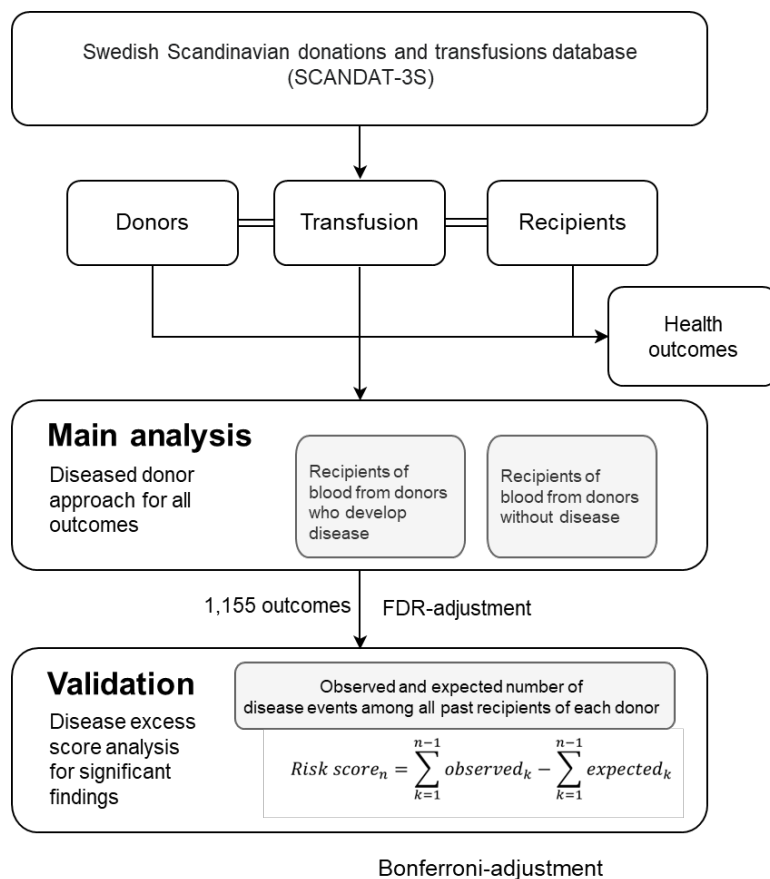


Figure 9. Study II flowchart.

As such, we modelled hazard ratios (HRs) for each disease group comparing recipients exposed to blood from diseased donors as compared to non-diseased donors. Significant HR both before and after adjustment for multiple testing was visualized. In a further step to validate the findings, we also investigated, in disease groups of significant outcomes after adjustment for multiple testing, using another principal approach. This approach was based on the rationale that an increased risk of disease transmission could also be identified if there was an increased disease occurrence in recipients of blood from the same donor, as compared to the expected disease occurrence in that group. In this approach, designated the Disease Excess Score (DES), we compared the actual observed number of disease cases in recipients from the same donor to the expected amount of disease cases of a particular disease. An

significantly increased DES may give similar indications of the occurrence of transfusion–transmission.

### **4.2.3 Statistical approach**

In the first analysis, modelling each outcome category and comparing HRs for exposed and non–exposed recipients, we performed a stratified, within calendar year of first transfusion and the hospital of the transfusion, Cox proportional hazard regression. Time–event tables were created using the SAS Stratify macro.<sup>179</sup> Aside from the main exposure variable (exposed and non–exposed) we included the following covariates: sex (as a categorical variable), time since the most recent transfusion (as a continuous variable), number of transfusions (as a restricted cubic spline with 5 equally placed knots) and age at first transfusion (as a restricted cubic spline with 3 equally placed knots). P–values for all results were then adjusted according to FDR. For the second analysis, conducted among the significant findings, we also utilized Cox regression – both to estimate the predicted number of diseases among each recipient of each donor and also to model the maximum disease excess score among all donors for a specific recipient. Again, the SAS Stratify macro was used to create event–time tables. The same covariates and strata as in the first analysis were used.

## **4.3 Paper III**

### **4.3.1 Data sources**

In Paper III, we utilize data from the national quality of care register in CML. This register was established in 2002. The information in the quality of care register is manually entered from physicians or study–nurses. CML subjects are either reported spontaneously at the time of diagnosis or identified and reported using the Cancer Register. The Cancer Register is built by compulsory reporting from all health care providers of all suspected and confirmed cancer cases.<sup>119</sup> These two methods of identifying CML cases ensures near complete coverage. For this study we utilized a pseudonymized database, the CMLbase 2, constructed using the CML register and record–linkage to multiple health care registers (Figure 10). In addition to all CML cases, the linked database also contains a control cohort with 5 randomly sampled controls based on age– and place of residency at the time of diagnosis and sex. The controls were sampled from the general population using replacement by Statistics Sweden.<sup>183</sup> The data that was specifically used for the study was from the quality of care register, to identify CML patients, disease phase and disease characteristics. Regarding TKI exposure, information was taken from the Prescribed Drug Register and from the



national quality of care register (to find information on drugs for patients receiving drugs directly from a sponsor due to enrollment in a clinical trial). Outcome information was retrieved from the National Patient Register, with information on in-patient and specialized out-patient care.

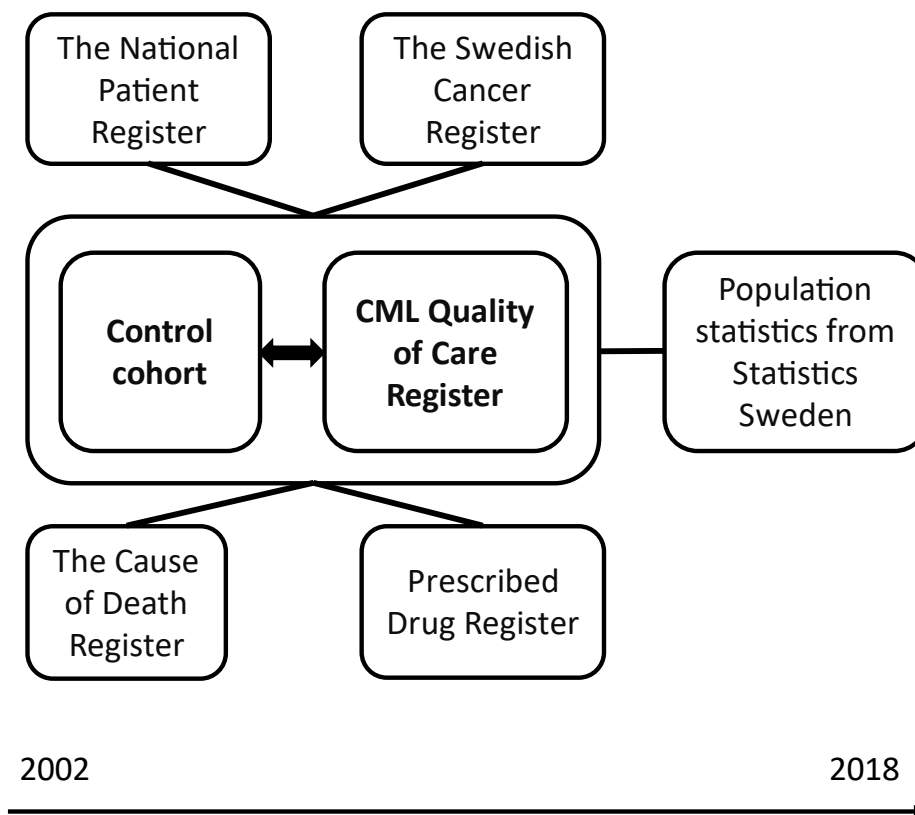


Figure 10. Linkages within the CMLBase 2 database.

#### 4.3.2 Study design

In Paper III, we utilized an approach that we had previously conducted but in a limited setting investigating cardiovascular outcomes in CP-CML patients.<sup>146</sup> In a main analysis we compared the incidence of a large set of outcomes to that of a control cohort to elucidate if there was an increased risk for any of the investigated disease outcomes. In principal we performed a matched cohort study, with CP-CML-cases and controls. The CP-CML population consisted of individuals diagnosed with CP-CML from 2002 to 2018. CP-CML patients and controls were followed from the date of diagnosis to emigration, death or the incident case of the investigated disease. We also censored CP-CML patients if they progressed to AP- or BP-CML or 5 months before allogeneic hematopoietic stem cell transplantation. To limit the impact of the abundant care when initiating treatment and also due to possible complications of the CP-CML itself we conducted a sensitivity analysis. We performed a delayed-entry analysis initiating follow-up at 6 months after diagnosis investigating the significant disease categories from the main analysis.

In a secondary analysis, we further studied the outcomes identified as significant after adjustment for multiple testing, from the main analysis within the CP-CML cohort. Here, rather than comparing CP-CML to controls, we stratified the events by the time-dependent variable of type of TKI-treatment (imatinib, dasatinib, nilotinib, bosutinib or ponatinib) within the CP-CML population. We then studied possible associations between TKI treatment and a specific outcome limiting the analysis to the CP-CML cohort.

The outcomes used were derived from the National Patient Register using ICD codes truncated at 3 points corresponding to the ICD categories. Multiple ICD chapters were removed as they were unlikely to be related to the disease (e.g. congenital malformations, pediatric conditions). We also did not include cancer diagnosis as it has previously been addressed in detail in the CP-CML population using the same data.<sup>184</sup> In total, the study was addressing 670 disease categories.

### 4.3.3 Statistical approach

In the main analysis we used Poisson regression to estimate IRRs between the CP-CML and control cohort using the control-cohort as reference for each investigated outcome. The analysis was adjusted for sex, age and calendar period (where age and calendar period were fitted as restricted cubic splines using three knots with placements according to Harrell's method).<sup>180</sup> To account for under- or over-dispersion, we applied the Lagrange multiplier test to each investigated outcome.<sup>185</sup> For outcomes where the test signaled significant dispersion we instead performed Poisson regression with empirical variance estimation to estimate confidence limits and P values. To account for multiple testing, we performed FDR-adjustment according to Benjamini and Hochberg.<sup>170</sup> Both adjusted (in main manuscript) and un-adjusted findings (in supplementary materials) are presented. For the delayed-entry model, the exact same statistical approach was used but instead of FDR-adjustment we utilized Bonferroni-adjustment as we performed the analysis on a subset of the initial outcome strata.

For the secondary analysis, we utilized a similar approach but also incorporated a specific CML disease progression risk score called the Sokal score (as a categorical variable) and the time-dependent TKI treatment variable (time-dependent and categorical). IRRs were estimated using imatinib-treatment as reference. In models that signaled significant under- or over-dispersion, we utilized quasi-Poisson model to allow for the time-dependent treatment variable. Adjustment for multiple testing was conducted by the Bonferroni-method and results displayed both before and after adjustment.

## 4.4 Paper IV

### 4.4.1 Data sources

In Paper IV; we compiled another database with information on all individuals in the Swedish Multi-generation Register together with information from the Swedish Cancer Register and Population Statistics from *Statistics Sweden*. The resulting register-linked database was based on all individuals recorded in the Multi-generation Register since its introduction together with information on cancer diagnosis. The Multi-generation Register contains information on maternity and paternity for more than 95% of individuals registered in Sweden 1961 or later and born 1932 or later.<sup>186</sup> For adopted individuals, there is information on non-biological parents. The register has been validated, in terms of faulty recordings on biological fathers, using ABO blood group data. Faulty registered fathers occurred in 1.6% of all offspring-father registrations during the whole register period, with slightly elevated occurrences in early periods and less than 1% in recent decades.<sup>187</sup>

Using this data we constructed pedigrees with 1 top individual in each tree using all relation information, as seen in Figure 11. As such, each individual could exist in multiple pedigrees. Regarding outcomes, the cancer disease diagnosis, we utilized the same grouping system as utilized in Paper I and II with slight adaptations, generating 60 distinct cancer disease groups.

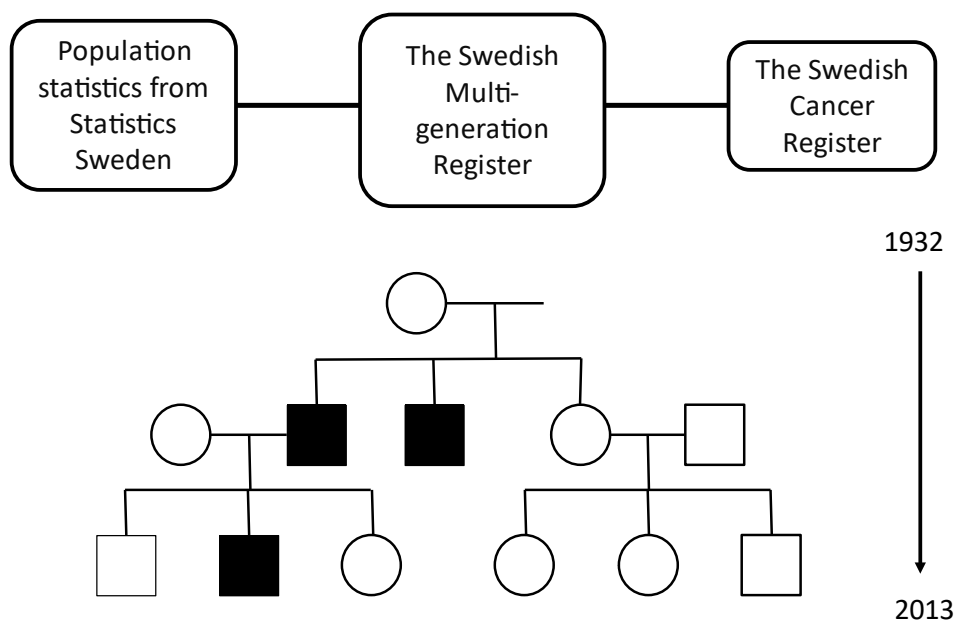


Figure 11. Register-linkages and creation of pedigrees.

### 4.4.2 Study design

The study utilized a multi-step process to identify excessive malignant disease occurrences in a family. This multi-step process consisted of firstly estimating the expected probability of each malignant disease in an individual. We then summarized

the expected and observed number of malignant disease cases of each disease for each family. To investigate the likelihood of identifying an observed number of cases given the expected number of cases in a family, we ran a set of 1,000 simulations randomly allocating the observations based on the expected count in a family. A family was defined as an outlier if A) the observed count was greater than the 99<sup>th</sup> percentile given the expected number of cases in a family and B) there was more than 1 observed case of the disease in the family. We then identified outlier families within multiple malignant disease groups to study more complex clustering patterns.

To estimate the expected number of malignant disease cases for each malignancy investigated we began follow-up at the later of birth or 1<sup>st</sup> of January 1958 (the start of the Swedish Cancer Register). Follow-up ended at emigration, death, 31<sup>st</sup> December 2013, or at the diagnosis of a malignant disease, whichever occurred first.

#### **4.4.3 Statistical approach**

To estimate the expected count for each individual and for each malignant disease we used Poisson regression after constructing time-event tables using the SAS Stratify macro.<sup>179</sup> In the expected model we included the following co-variables: age (as a restricted cubic spline with 5 knots placed according to the percentiles described by Harrell's), calendar period (also as a restricted cubic spline with the same principles and knots), and sex (as a categorical variable).<sup>180</sup>

After summarizing the expected and observed counts for each family and for each malignant disease, we ran 1,000 simulations randomly allocating the observed counts assuming a Poisson process based on the expected counts for each family. We then identified the 99<sup>th</sup> percentile and defined any observed count higher than this as an outlier based on the expected value.

## 5 ETHICAL CONSIDERATIONS

There are multiple ethical aspects in the Papers contributing to this thesis. A first aspect is that in general there is no informed consent obtained in health-register participants. A second aspect is that we may identify individuals who could carry a disease which has not yet been diagnosed. Third, and lastly, the data management and storage of the sensitive information in the databases is used to conduct the research. Importantly, however, is that all studies have been approved by either local ethics committees or the new central review authority.

The individuals in the SCANDAT-3S, the Cancer Register or the Multi-generation Register, to give a few examples, have not given an informed consent to allow our usage of their personal data. The cost, in terms of personal integrity, is therefore high. However, the other aspect is that we are able to construct sufficiently large databases, also involving individuals who are not alive and where informed consent for obvious practical reasons is impossible, to perform relevant research. To address the issue of personal integrity the database has been pseudonymized when linked, meaning we have only a number sequence for each patient and the key to unlock this sequence is stored in secure facilities at the National Board of Health and Welfare to make further future linkage with other registries possible. The possible gains from the current research, and other research conducted utilising the same database, are possibly large as we may demonstrate issues in donor or recipient health that may impact donor screening, recipient screening or how we monitor transfusion-transmission. This thesis work does not cover all planned parts of the research utilising these databases and other studies. There are a multitude of important research questions that will be answered using the data. As such, the possible personal integrity considerations arising from the missing informed consent are well balanced with the quality and the quantity of research that can be conducted. The long list of previous publications from earlier versions of the SCANDAT or the CMLBase database demonstrates the impact that this type of data can have. Also, to conduct the agnostic studies we rely on having *all* data, especially for small disease entities where “missingness” may play a crucial role in how to interpret possible findings. The type of findings the current studies could generate are of importance whatever the results are. E.g., findings of disease transmission would impact the transfusion medicine society depending on what disease is found, on the other hand, if no association of disease transmission is found, then this is a important signal that tells us that current practice guidelines are safe.

The second potentially ethically problematic situation and dilemma, is if we identify individuals with possible disease but who have not been diagnosed. As of today there are no means to “contact and trace”. This is exemplified by an ongoing effort to trace individuals with a high suspicion of hepatitis c, a condition where there is effective treatment available today. However, because of the delicate situation with personal data

and integrity the laws guarding the data have prevented this. This is problematic from an ethical viewpoint, when a treatable but possible cancerous condition is discovered, and probably unknown to the patient, but identified using register data.

As to the third point, for the storage and management of this kind of sensitive data there are rigid security protocols in place and only on-site access to the data (or access via remote computers to remote servers for analysis without possibility of file transfer etc). The server security is managed by top-tier security experts who have applied multiple levels of data security.

## 6 RESULTS

### 6.1 Paper I

#### 6.1.1 Study population and baseline characteristics

In the main explorative cohort, we identified 4,204,234 individuals in the SCANDAT-3S database with a recorded blood group and who had undertaken a blood antigen test without donating blood within 90 days of that test. In total, this cohort accrued 50 million years of person-time of follow-up. The median age at the first blood antigen test was 52 (IQR, 30–71) years. There were 60% females in the cohort. The validation cohort, consisting of blood donors, included 1,197,522 individuals with a total of 22 million years of follow-up. The median age at the first blood antigen test was in this cohort 30 (IQR, 23–41) years and 49% of individuals were female.

The blood group distribution was similar in the two cohorts with 47% and 45%, 5% and 5%, 10% and 11%, and 38% and 39% for A, AB, B and O, in the main and validation cohort, respectively.

#### 6.1.2 Main findings in regard to ABO and disease

In the main exploratory cohort, we identified 348 associations between ABO blood group and a disease group before adjustment for multiple testing. After adjustment, 143 associations remained significant. Of these 143 disease categories, at least one association remained in the analysis in the validation cohort, however, associations were fewer in terms of number of blood groups and disease groups. After adjustment for multiple testing in the validation cohort, 49 associations in 27 disease categories remained. The adjusted results in the validation cohort are summarized in Figure 12, Figure 13 and Figure 14. In general, these results are concordant with previous publications. However, the association of a lower IRR for kidney stones in individuals with blood group B as compared to O has not been previously described.

#### 6.1.3 Main findings in regard to RhD status and disease

There were fewer associations between RhD status and disease as compared to ABO blood group and disease. In the analysis, before adjustment for multiple testing, 98 associations were found. After adjustment only 13 remained significant. In the validation cohort, 5 of these were replicated. After adjustment for multiple testing only pregnancy-induced hypertension remained, being more common in RhD positive individuals as compared to RhD negative.

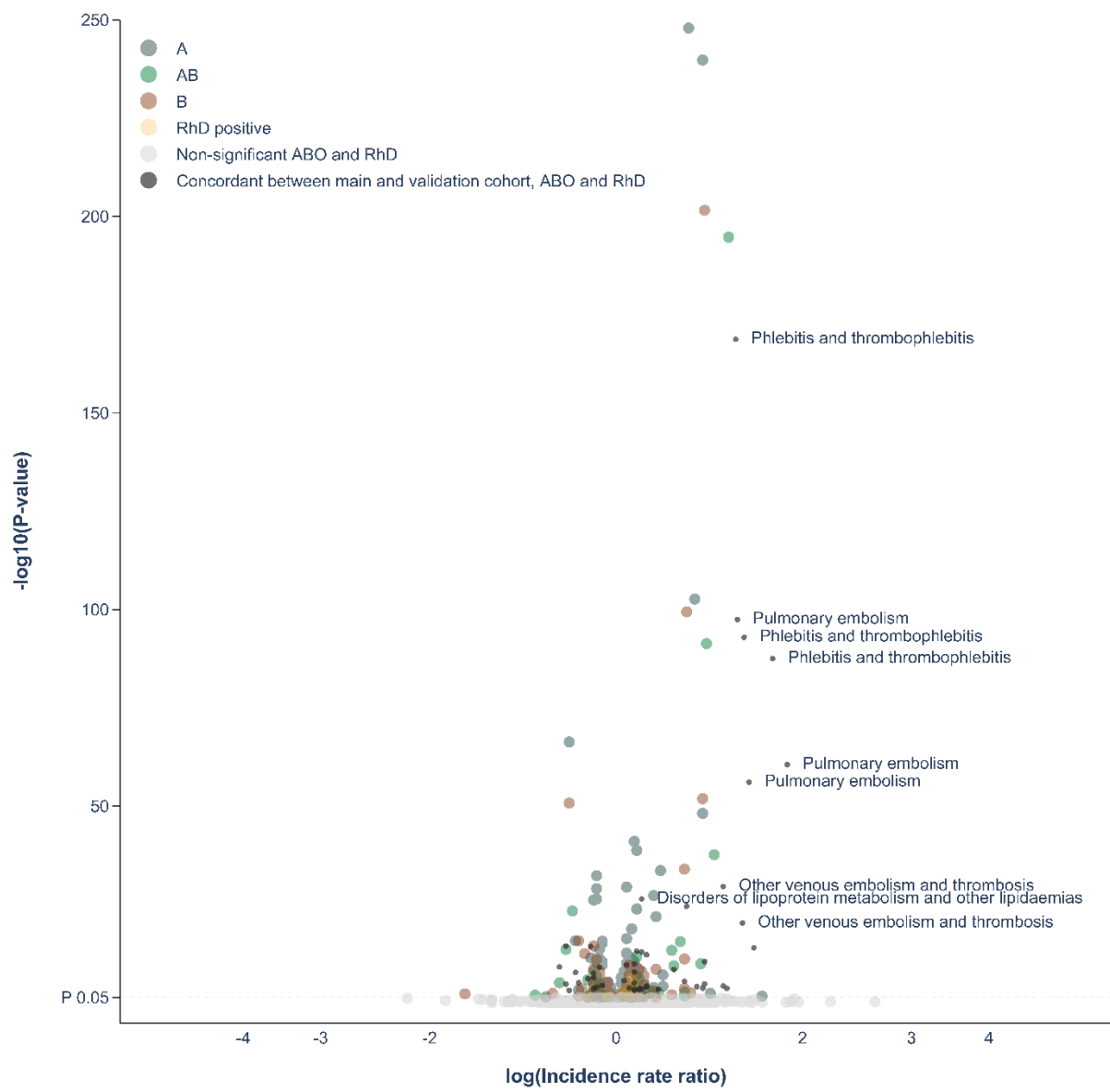


Figure 12. Manhattan-plot according to disease category with findings in terms of association between ABO blood group and RhD status and disease. Larger dots indicate an increased IRR.



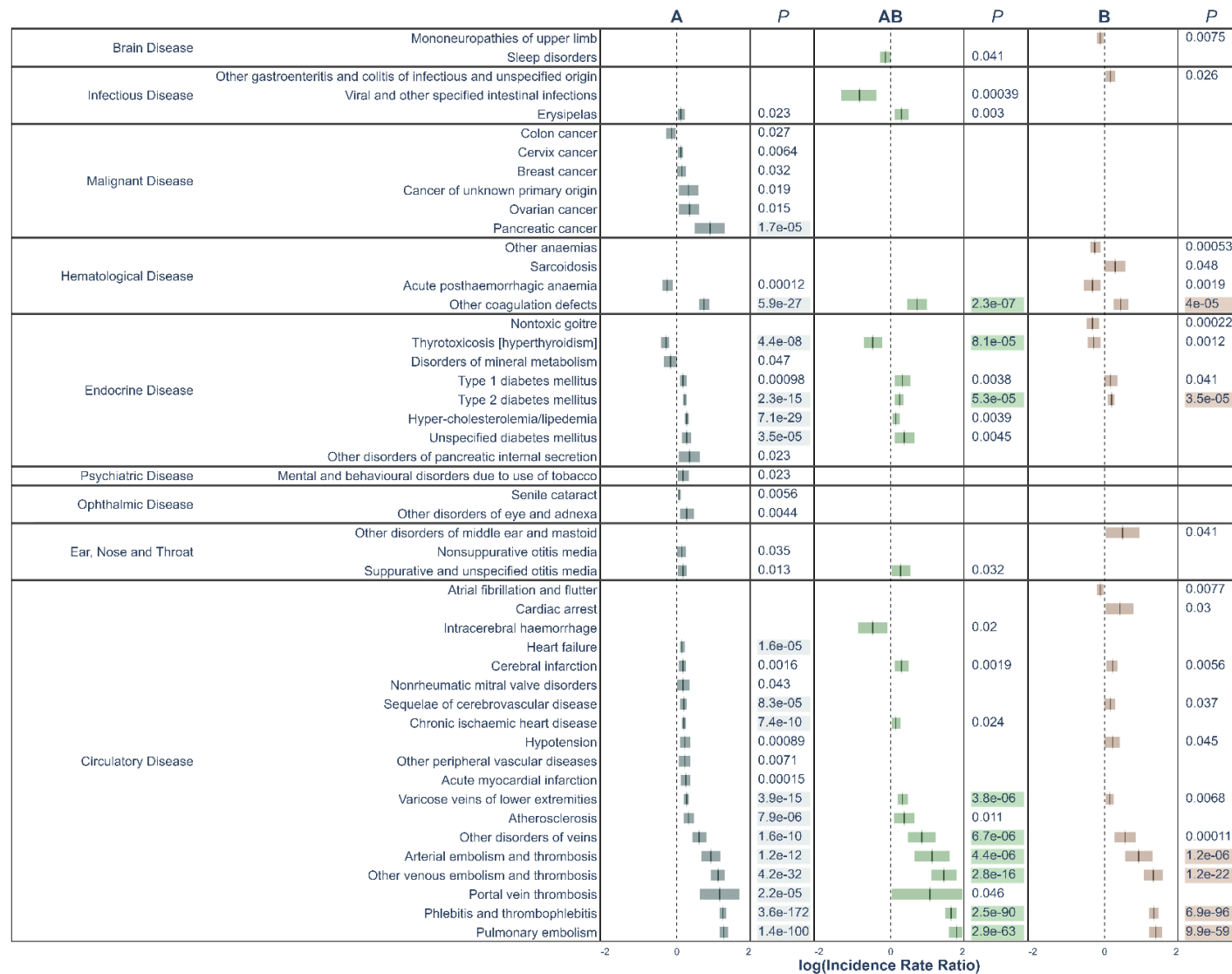


Figure 13. All significant un-adjusted findings from the validation cohort (Part 1). Significant disease categories in the validation cohort. Blood group as compared to blood group O and log10(IRR) displayed with 95% confidence bands. All P values are raw, highlighted P value indicates associations that remained statistically significant also after Bonferroni-adjustment.

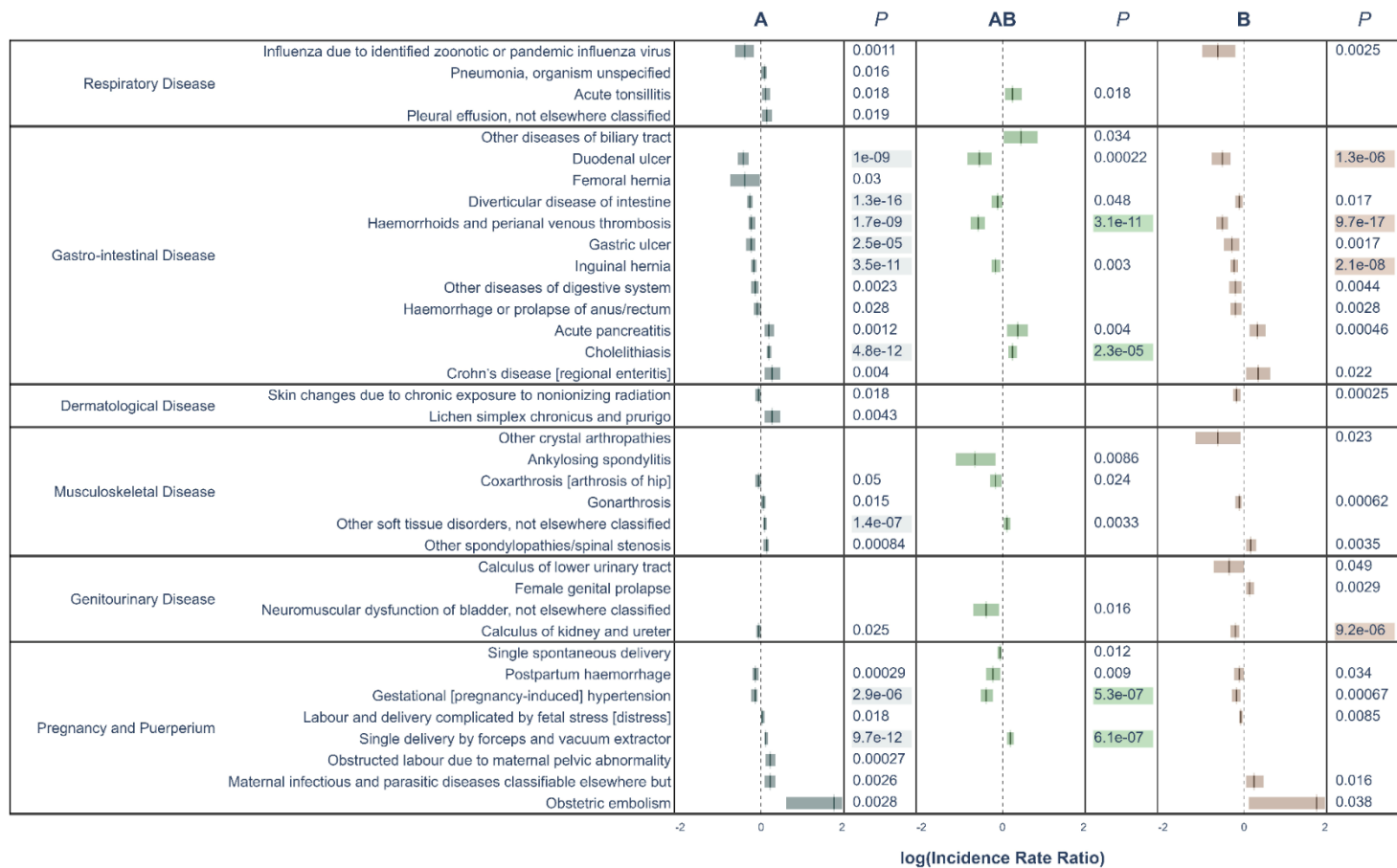


Figure 14. All significant un-adjusted findings from the validation cohort (Part 2). Significant disease categories in the validation cohort. Blood group as compared to blood group O and log<sub>10</sub>(IRR) displayed with 95% confidence bands. All P values are raw, highlighted P value indicates associations that remained statistically significant also after Bonferroni-adjustment.

## **6.2 Paper II**

### **6.2.1 Study population and baseline characteristics**

The study cohort consisted of the 1,729,165 recipients of blood in the SCADAT-3S database, receiving allogeneous erythrocytes, platelets, plasma or whole-blood transfusions between 1968 and 2017. The follow-up time in the recipient cohort consisted of 13.4 million person-years and 56% of individuals were female. In total 19 million blood products were transfused in these recipients from 1,051,744 blood donors with a total follow-up of 20 million person-years. Of the donor cohort 46% were female.

### **6.2.2 Main findings in terms of possible transfusion-transmitted disease**

In the first main analysis to study transfusion-transmitted disease, we identified a total of 65 associations. After adjustment for multiple testing, 15 of the disease categories remained. Of these, 13 could be replicated using the second principal approach detailed above to study transfusion-transmitted disease. The full scope of the findings is depicted in Figure 15 and the significant findings, after adjustment for multiple testing, are found in Figure 16. In general, we identify, with strength, the known transfusion-transmitted disease and their complications (hepatitis and esophageal varices). An unexpected outcome found in the initial analysis was the outcome group abnormal findings in specimens from male genital organs. This outcome could not be replicated in the second analysis due to few events. This category also contains a wide array of findings that were not further specified in the database due to truncation of ICD codes within this ICD category.

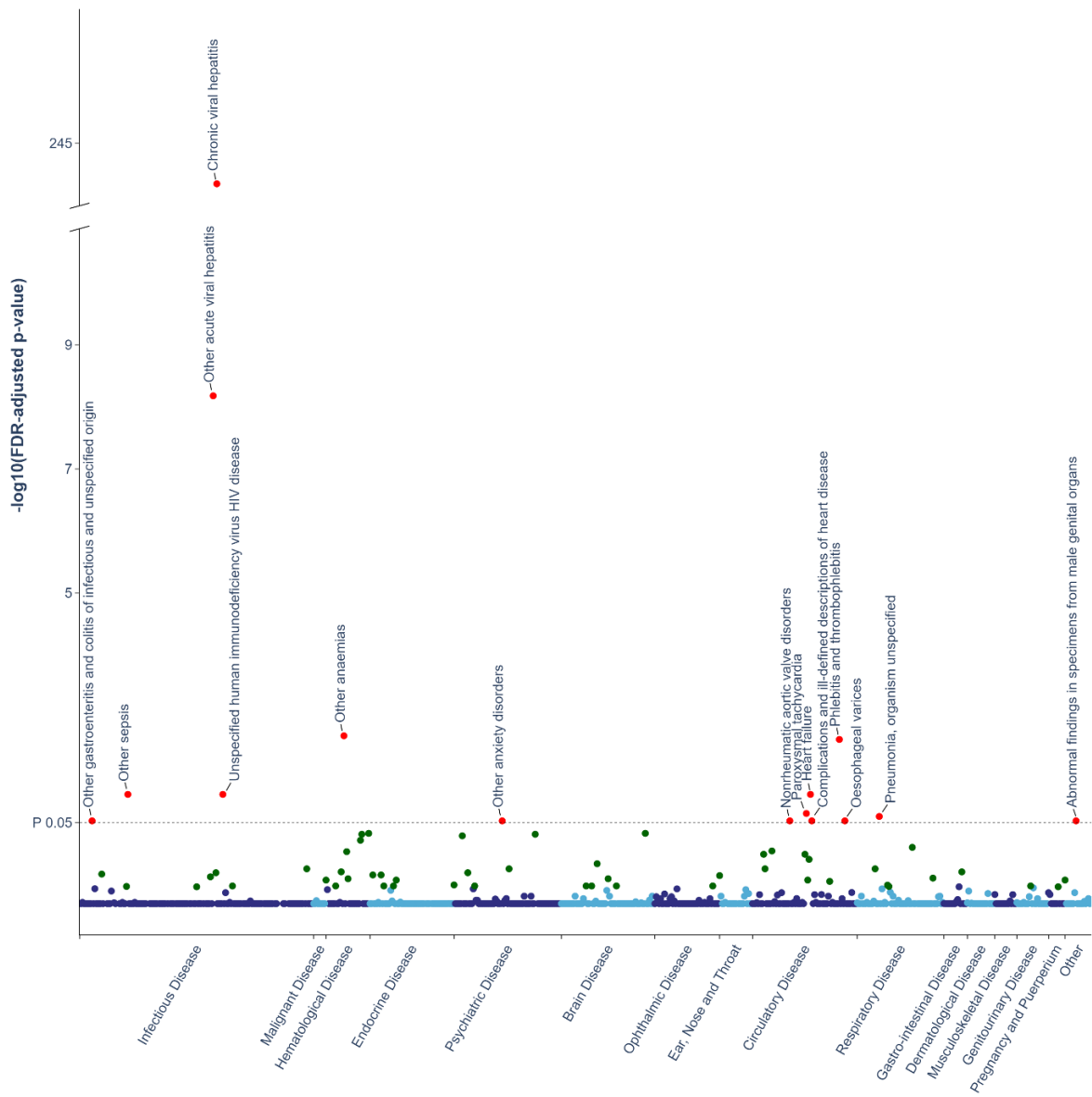


Figure 15. Manhattan plot with each disease category represented by a dot along the x-axis and  $-\log_{10}(\text{FDR-adjusted } P\text{-values})$  along the y-axis. FDR-significant findings in red (details in Figure 16), findings in green are significant before but not after FDR-adjustment.

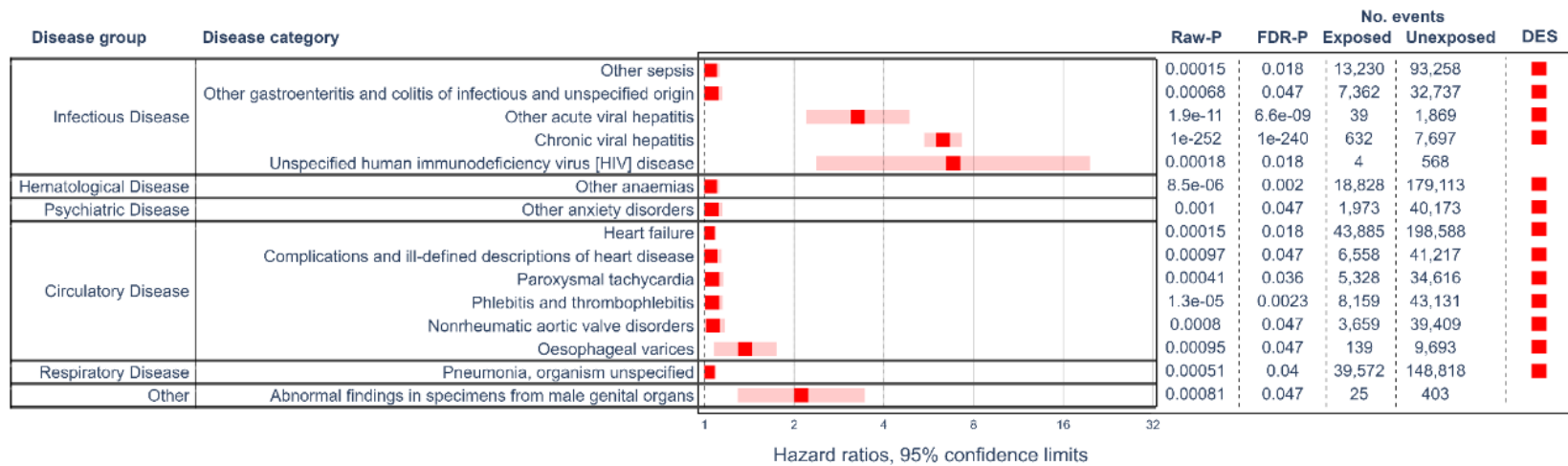


Figure 16. Significant findings, after adjustment for multiple testing, in Paper II. Red dots in DES column signals a finding also significant in DES analysis.

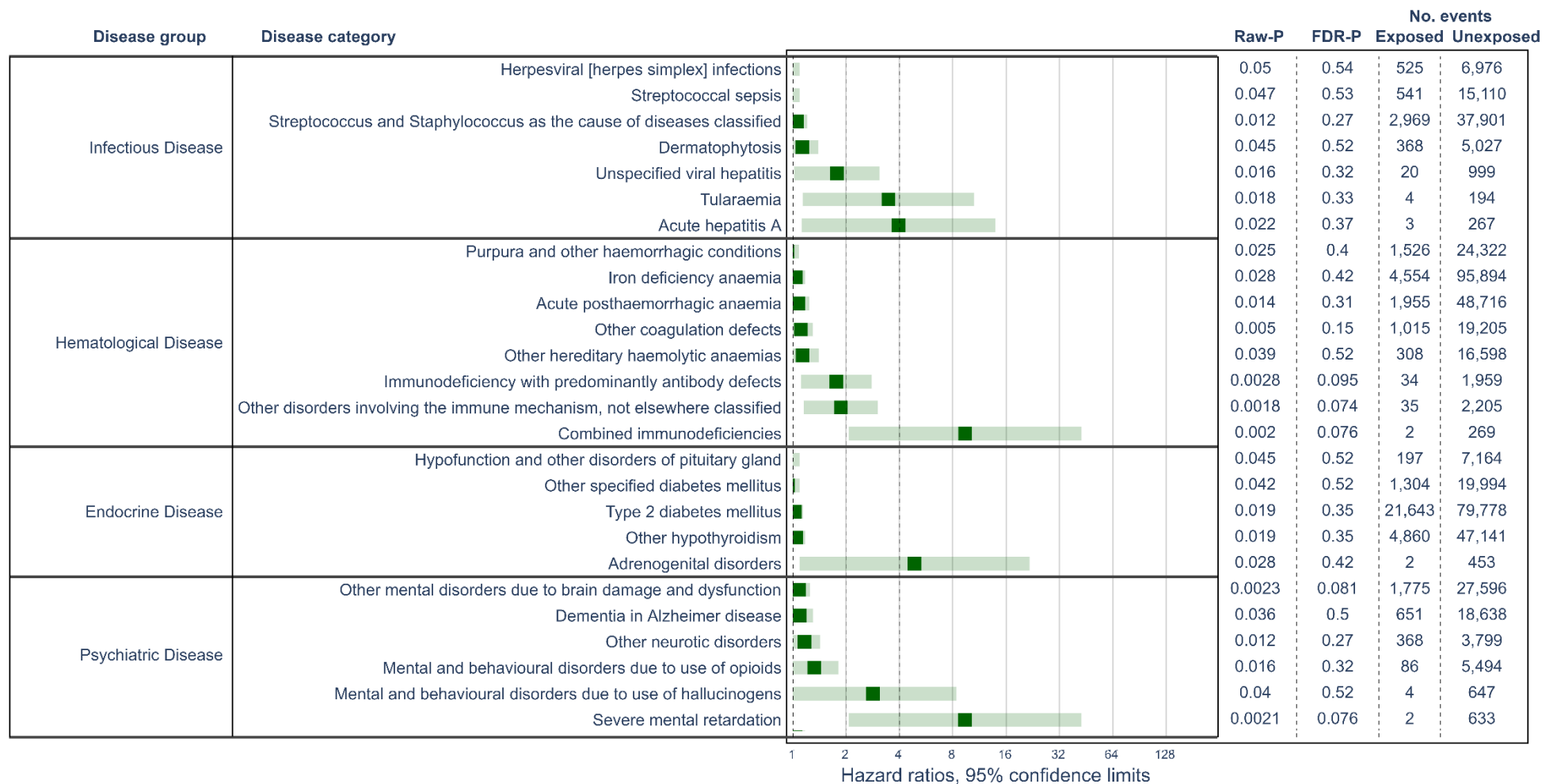


Figure 17. Summary of diseases that did not remain statistically significant after adjustment for multiple testing. (Part 1)

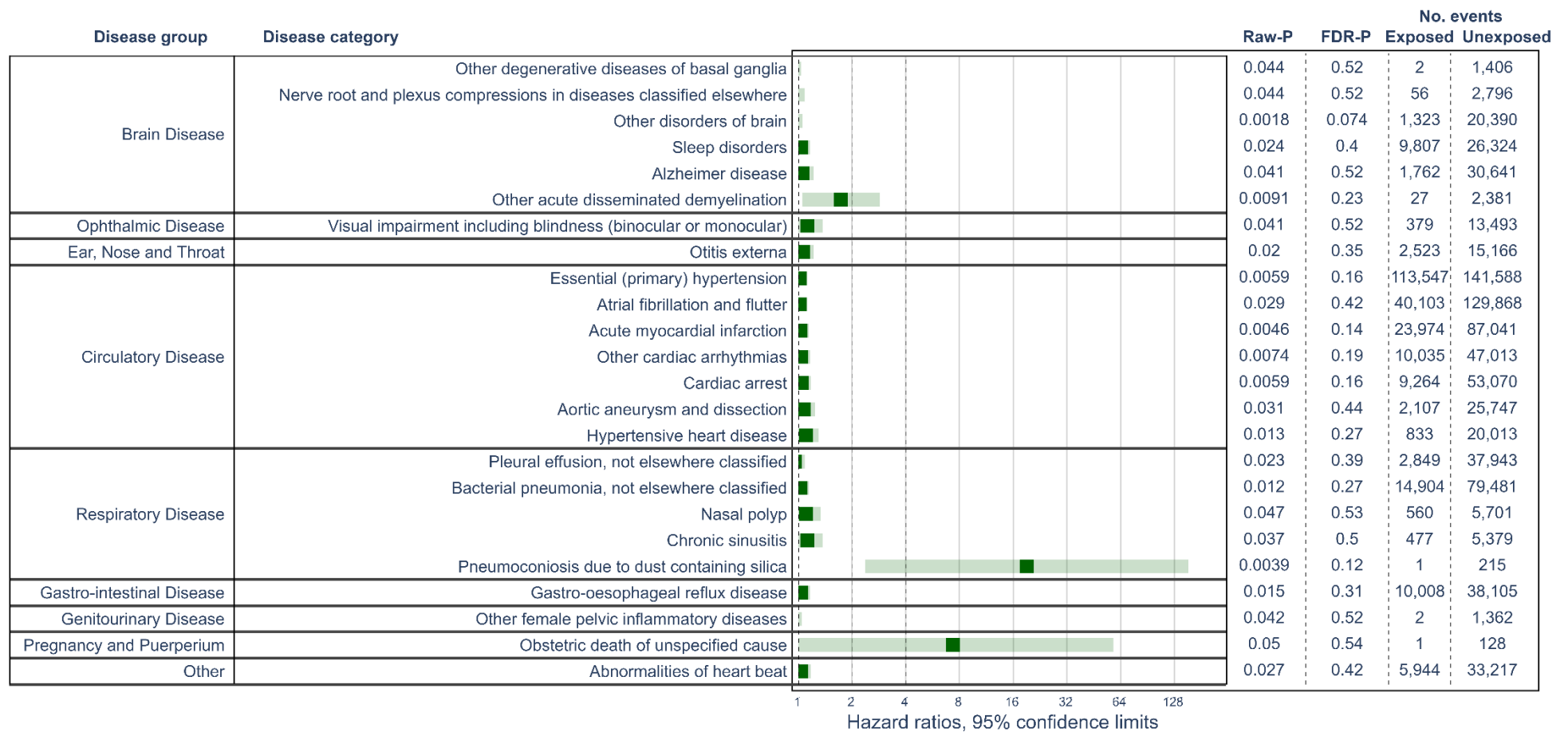


Figure 18. Summary of diseases that did not remain statistically significant after adjustment for multiple testing. (Part 2)

## **6.3 Paper III**

### **6.3.1 Study population and baseline characteristics**

Between 2002 and 2017 there were 1,312 adult patients diagnosed with CML in chronic phase receiving a TKI in Sweden. The total follow-up time in the cohort was 8,510 patient-years and 46% of patients were female. Median age at diagnosis was 51 (IQR, 46–71). The control cohort consisted of 6,640 individuals matched by age, sex and place of residency at diagnosis with similar baseline characteristics.

Further deciphering of the CML cohort demonstrated that 91%, 29%, 29%, 4% and 2% ever received imatinib, dasatinib, nilotinib, bosutinib or ponatinib at any time-point during follow-up.

### **6.3.2 Main findings in terms of adverse events in comparison to the control cohort**

In 405 disease categories of the total 670 disease categories investigated there were at least 1 event in the CP-CML population. Before adjustment for multiple testing, we identified 169 disease categories with associations with increased risk in the CML cohort as compared to the control cohort. After adjustment for multiple testing, 142 disease categories remained significant. These are depicted in Figure 19 and Figure 20, in terms of number of events and strengths of the associations. In the sensitivity analysis, where start of follow-up was initiated 6 months after diagnosis, there were 54 disease categories that still remained significant, also summarized in the same Figures. As a summary, no new severe morbidities with increased risks as compared to the control cohort were identified.

### **6.3.3 Main findings in terms of adverse events within the CML cohort in terms of different TKIs**

In terms of associations between individuals TKIs and possible adverse events within the disease categories, we found associations in 41 disease categories and a specific TKI as compared to imatinib treatment. After Bonferroni adjustment, 3 associations remained. This analysis was, however, hampered by imperfect modelling due to very few events in multiple disease groups. In the analysis of within the CML cohort, before adjustment for multiple testing, we identify that there is an increased risk for patients receiving nilotinib, as compared to imatinib, to experience a myocardial infarction.



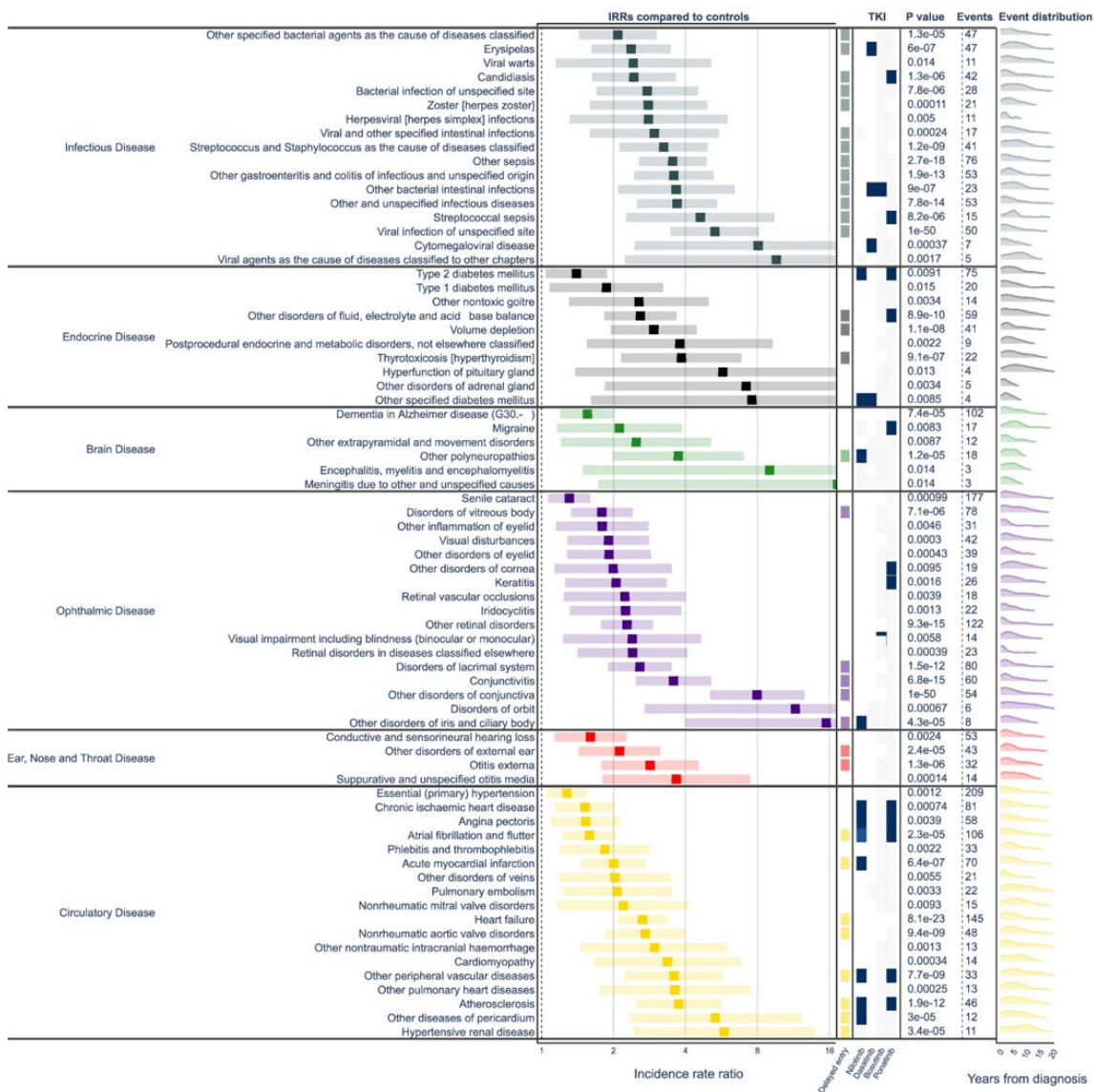


Figure 19. All findings significant findings after adjustment for multiple testing (Part 1). First column demonstrates IRRs for significant finding after FDR adjustment for the CML population as compared to the control population: strong color demonstrating the point estimate and lighter color demonstrating 95% confidence intervals. Second column demonstrates the same analysis but findings significant after Bonferroni-adjustment in a delayed entry model with 6 months from diagnosis of the significant findings from the first analysis. Third column demonstrates an un-adjusted analysis, as to multiple testing, for significant associations between a specific TKI as compared to imatinib, in terms of the outcome. Lighter blue color demonstrates a lower IRR as compared to stronger color. Fourth and fifth column present raw p-values for the analysis in column 1 and events, respectively. Sixth column presents event distribution during follow-up from date of diagnosis.

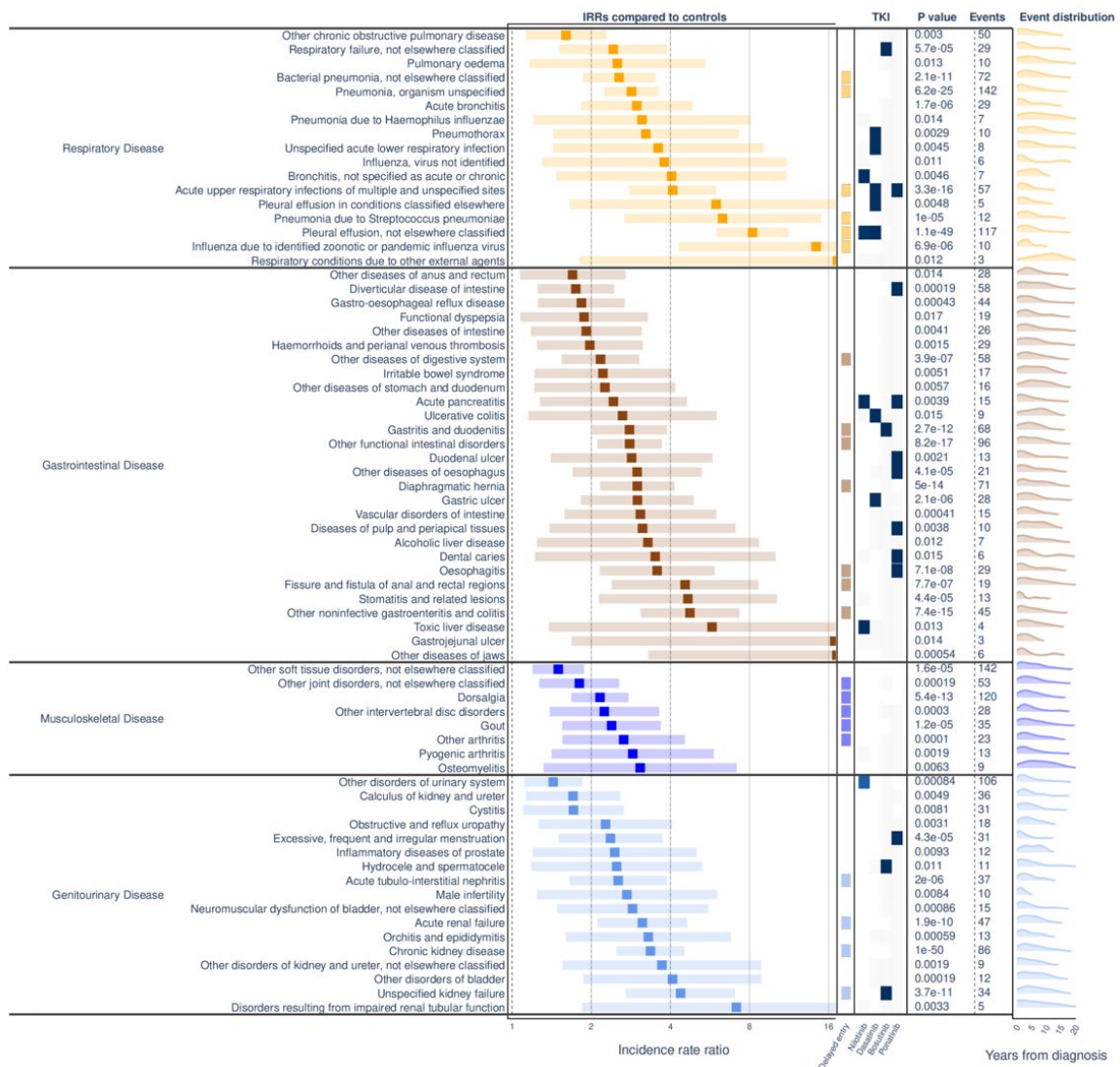


Figure 20. All findings significant findings after adjustment for multiple testing (Part 2). First column demonstrates IRRs for significant finding after FDR adjustment for the CML population as compared to the control population: strong color demonstrating the point estimate and lighter color demonstrating 95% confidence intervals. Second column demonstrates the same analysis but findings significant after Bonferroni-adjustment in a delayed entry model with 6 months from diagnosis of the significant findings from the first analysis. Third column demonstrates an un-adjusted analysis, as to multiple testing, for significant associations between a specific TKI as compared to imatinib, in terms of the outcome. Lighter blue color demonstrates a lower IRR as compared to stronger color. Fourth and fifth column present raw p-values for the analysis in column 1 and events, respectively. Sixth column presents event distribution during follow-up from date of diagnosis.

## **6.4 Paper IV**

### **6.4.1 Study population and baseline characteristics**

In the Multi-generation Register, there were 12,744,444 individuals, either as index persons (born 1932 to 2013 and registered as a resident after 1961) or as a registered parent to any of the index persons. From these individuals, we were able to construct 3,379,218 pedigrees. The pedigrees spanned over a median of 4 generations (IQR, 2 to 6), encompassing a median of 8 individuals (IQR, 4 to 14) with 373 million individual person-years of follow-up.

### **6.4.2 Main findings in terms of familial disease clustering**

After running the simulations we identified 33,721 pedigrees that we defined as outliers, based on the definition of having more observed cancer cases than one could expect from the 99-percentile of the simulations. The vast majority, 33,182, were outliers in a single malignant disease category, and 523 and 16 pedigrees were outliers in 2 or 3 disease categories, respectively. Figure 23 demonstrates pedigrees co-occurring within multiple disease clusters. For unknown cancer, or cancer of the placental, nasal sinuses or tonsillar location, no clustering was observed. In Table 1 all findings are summarized. Notably, breast-, prostate-, low-differentiated thyroid-, colon cancer and melanoma displayed the highest degree of clustering with the percentage of all cancer cases in pedigrees defined as outliers of 9.6%, 7.8%, 7.5%, 5.8% and 7.8%, respectively. Figure 22 demonstrates observed as compared to expected cancer cases for all types of cancer within families. In the same Figure we can identify families with <1 expected case but with 10 observed cases. Figure 21 and Figure 22 display the same information but for separated for each cancerous disease.

Anatomical region	Cancer type	Pedigrees with increased occurrence	Individuals with cancer in pedigrees with increased risk	Unique Cancer cases in full cohort during follow-up	Percentage of all cancer cases with clustering
Head-Neck cancer	Lip cancer	12	25	5662	0.44
	Tongue cancer	14	22	4118	0.53
	Salivary gland cancer	3	6	3169	0.19
	Other oral cancer	15	26	5546	0.47
	Pharyngeal cancer	16	25	3670	0.68
Digestive tract cancer	Esophageal adenocarcinoma	8	12	3336	0.36
	Esophageal squamous-cell carcinoma	20	34	5468	0.62
	Gastric cancer	371	631	36744	1.72
	Hepatocellular cancer	35	58	5962	0.97
	Biliary tract cancer	2	2	2729	0.07
	Pancreatic cancer	363	594	27480	2.16
	Other upper-digestive cancer	85	146	18183	0.80
	Small-intestine cancer	50	71	5867	1.21
	Colon cancer	3030	5364	92775	5.78
	Rectal cancer	895	1506	48662	3.09
	Anal cancer	3	6	2294	0.26
	Other lower-digestive cancer	3	4	2840	0.14
	Respiratory cancer	Larynx cancer	19	32	6152
Any lung cancer		67	129	30811	0.42
Lung squamous-cell carcinoma		89	123	10388	1.18
Lung adenocarcinoma		394	492	18264	2.69
Small-cell lung cancer		55	72	6915	1.04
Other non-small cell lung cancer		117	152	12210	1.24
Pleural mesothelioma		16	24	2614	0.92
Urinary organ cancer	Urothelial cancer	1155	1992	59494	3.35
	Kidney cancer	480	753	30028	2.51
	Other urinary organ cancer	2	4	3959	0.10
Skin cancer	Melanoma	2956	4231	54038	7.83
	Non-melanoma cancer	1191	2432	65763	3.70
CNS, PNS, Thyroid cancer	Nervous WHO III-IV cancer	207	271	13231	2.05
	Meningioma	173	214	11072	1.93
	Thyroid well-differentiated cancer	192	239	9187	2.60
	Thyroid low-differentiated cancer	105	121	1614	7.50
	Other CNS cancer	407	482	16131	2.99
	Other endocrine cancer	459	674	22100	3.05
Hematological cancer	Non-Hodgkin lymphoma	820	1262	38867	3.25
	Chronic lymphocytic leukemia	165	245	10198	2.40
	Hodgkin lymphoma	179	177	6977	2.54
	Multiple myeloma	128	209	15160	1.38
	Acute lymphoblastic leukemia	105	99	4225	2.34
	Acute myeloid leukemia	43	58	8449	0.69
	Chronic myeloid leukemia	1	2	3232	0.06
	Other hematological cancer	112	183	13183	1.39
Female cancer	Breast cancer	8730	17183	179580	9.57
	Cervix cancer	357	536	21653	2.48
	Corpus cancer	742	1202	35683	3.37
	Ovarian cancer	609	967	28627	3.38
	Vulvar/vaginal cancer	18	32	5549	0.58
	Other female cancer	2	2	3667	0.05
Male cancer	Prostate cancer	8331	14823	189260	7.83
	Testis cancer	346	348	8526	4.08
	Penis cancer	6	8	2207	0.36
All Eye cancers	Eye cancer	105	81	4113	1.97
Cancer of unknown primary	Cancer of unknown primary	382	628	37835	1.66
Bone and connective tissue cancer	Bone cancer	17	20	2820	0.71
	Connective tissue cancer	69	99	9016	1.10

Table 1. Familial disease clustering by malignant disease.

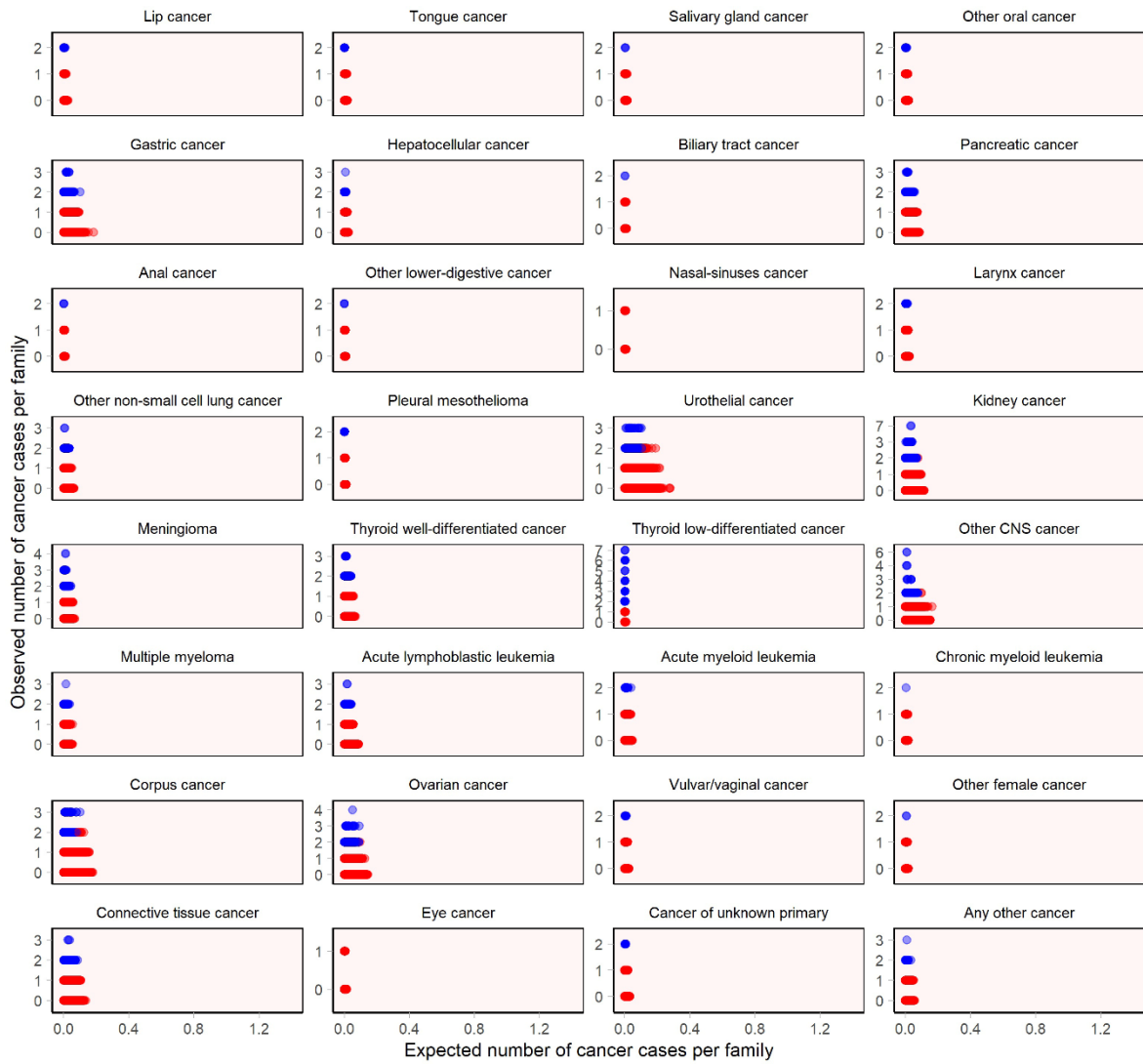


Figure 21. Expected and observed number of cancer cases for each disease for all pedigrees (Part 1). Blue dots indicate pedigrees with more cancer cases than the 99<sup>th</sup>-percentile of 1,000 simulations with Poisson process based on the expected count in each pedigree. Red dots indicate pedigrees with observed cases within the 99-percentile. As such, there are pedigrees with less than 1 expected cancer case but with 10 observed cases.



Figure 22. Expected and observed number of cancer cases for each disease for all pedigrees (Part 2). Blue dots indicate pedigrees with more cancer cases than the 99<sup>th</sup>-percentile of 1,000 simulations with Poisson process based on the expected count in each pedigree. Red dots indicate pedigrees with observed cases within the 99-percentile. As such, there are pedigrees with less than 1 expected cancer case but with 10 observed cases.

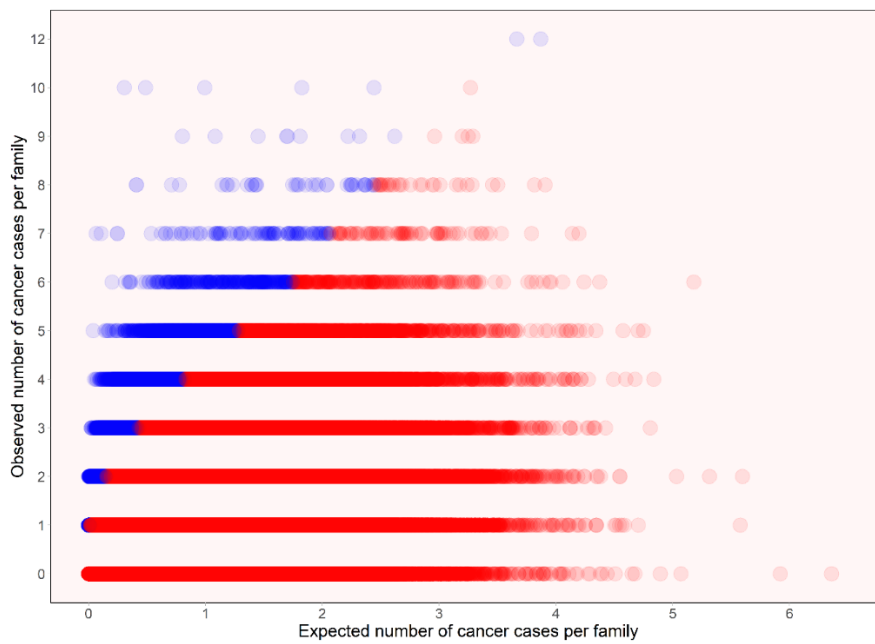


Figure 23. Expected and observed number of cancer cases for all cancerous disease for all pedigrees. Blue dots indicate pedigrees with more cancer cases than the 99<sup>th</sup>-percentile of 1,000 simulations with Poisson process based on the expected count in each pedigree. Red dots indicate pedigrees with observed cases within the 99-percentile. As such, there are pedigrees with less than 1 expected cancer case but with 10 observed cases.

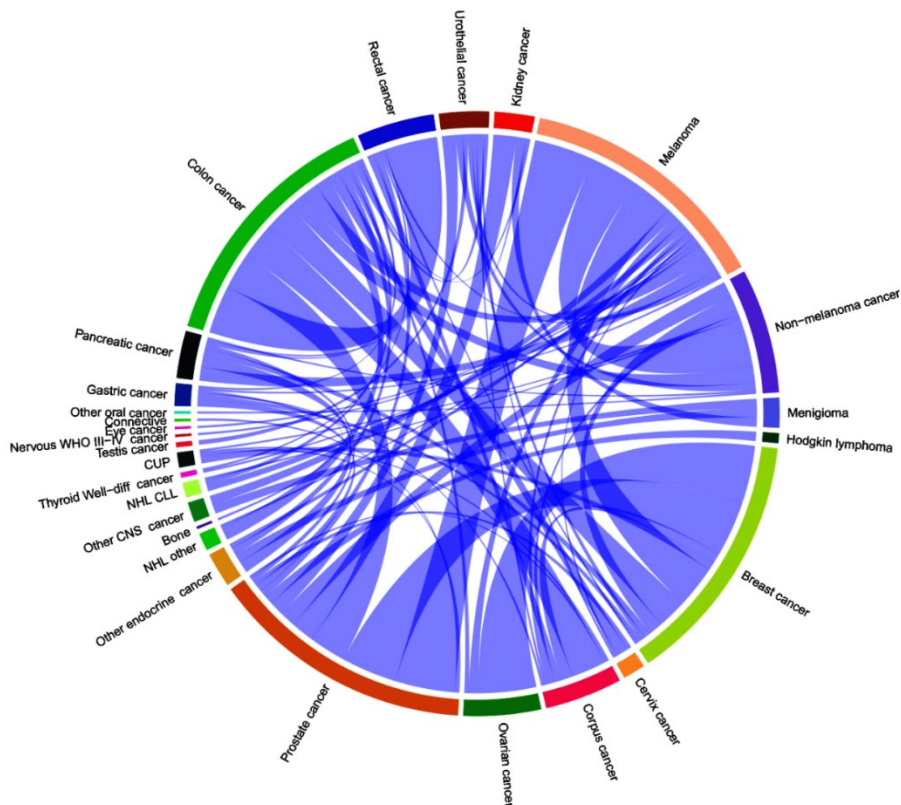


Figure 24. Multiple disease clusters within the same pedigrees. A circular map displaying all disease with clustering in multiple malignant disease occurring concurrently in the same pedigree. The size of each cancer site (base) is determined by the relative number of pedigrees co-occurring in multiple diseases. Connecting arrows (blue) display the co-occurring disease connections with width of the arrow explaining the relative extent of the relationship

## 7 DISCUSSION

The clinical research aims in this thesis are diverse in terms of the individual subjects studied. Instead, the thesis is united by the general methodological approach. The principal framework involves studying associations between either one or multiple factors and one or multiple outcomes. In Paper I and II, we utilize one of the largest epidemiological databases available in Sweden with a combined cohort of 8 million individuals. In Paper IV we also used a large database, including 12 million individuals from the Swedish population. In Paper III, however, the approach is the same but the cohort size is much reduced. Thus, a main differences between the Papers is the statistical power available to detect associations. This impacts the agnostic approach as the scope of the agnostic search is limited to the study of outcomes with enough events to generate meaningful analysis. Also, one must be aware that the approach used is limited by imprecise and generic models of analysis. There is no possible way to carefully consider all available aspects of an association in this setting. The agnostic approach is a tool aimed at screening and to initiate further analysis, most suitable in areas where one does not intuitively understand an association.

In each Paper (Appendix 1-4), there is a thorough discussion relating to specific strengths and limitations within each study. In the following section, the focus is instead that of classical epidemiological caveats in relation to the approach in specific situations in each of the individual studies. The main sources of error within epidemiological studies, depicted in Figure 24, are discussed in the following sections.

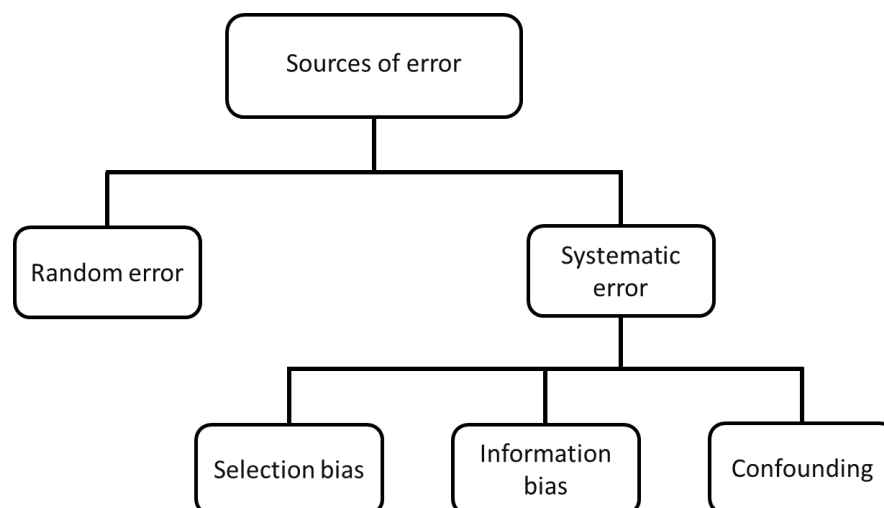


Figure 25. General sources of error in epidemiological studies.



## 7.1 Random error

### 7.1.1 Statistical estimation

For Paper I, III and IV we used Poisson regression as the primary model of investigation to estimate risk ratios using count event data of time-intervals. When constructing the large number of models it is not feasible to accurately assess the model fit for each model and hence to accurately account for the random error. To address the issue of model fit we applied a scheme where, in Paper I (limiting the definition of the agnostic approach) we arbitrarily choose to limit the analysis to disease categories with at least 50 events. In a second attempt to find a fitting model, also utilized in Paper III, we tested the Poisson regression assumption of the conditional mean being equal to the variance.<sup>188</sup> With this assumption, in over-dispersed data, significant findings would be overestimated and confidence limits narrower. In under-dispersed data we would have the reverse situation. For models where testing indicated deviance from equi-dispersion, we therefore applied models that could fit count data within a similar framework and also account for under- or over-dispersion. Further, model misspecification due to unmeasured or residual confounding, is in this registry-based setting with 1,000s of models impossible to account for. This is more likely to influence Paper II and Paper III. Again, this must be thought of as an inherent problem to the agnostic approach and counteracted by further and more precise studying of relevant findings.

### 7.1.2 Multiple testing and P-values

The critique of the interpretation and usage of P-values, with emphasis on pre-defined alpha, larger sample size and its implications on P-value as well as lack of relationship with effect size, renders a more complicated discussion of the usage of adjustment for multiple testing.<sup>189-193</sup> The P-value is a way to represent a measure of probability in obtaining a certain or even more extreme result assuming a true null hypothesis. However, in the studies conducted, one of the main assumptions of the P-value - the assumption that the study is free from systematic error - most definitely is violated in at least one but more likely a large portion of the investigated hypotheses. As such, there are multiple levels of confusion when looking at the findings from the individual studies. In the presentation of results, our focus has been to visualize all findings before and after adjustment and to display relevant measures of effect to place the P-values into a greater context. This allows the reader to see and the discussion to address the full spectrum of the data and to allow some interpretation and leniency in the type I error rate.

As to the overall objective with multiple testing for the current Papers, it is also important to consider that we want both the possibility to limit false discoveries (type I error rate) but also limit the possibility to reject true discoveries (type II error rate). Here the strength of the FDR approach contra the Bonferroni-method is superior since FDR and Bonferroni will produce the same rejections if all hypotheses are true whereas FDR will generate less rejections in all other circumstances. An issue in multiple testing using the FDR approach, however, is the assumption the independence of samples, that may have a large impact on the rejection rate. We again, have counteracted this problem by either browsable figures or supplementary tables with individual P-values and also by giving information on the estimated effect sizes. Future studies could apply strategies for multiple testing without assumptions in this regard, e.g. the Benjamini and Yekutieli method, and also allow for construction of adjusted confidence intervals.<sup>181</sup>

## **7.2 Selection bias, information bias and confounding**

### **7.2.1 The healthy-donor effect**

The healthy-donor effect is an important bias in transfusion-medicine. It may affect any situation where donors with different donation characteristics, such as donation-frequencies or donors during different timepoints of follow-up or donors as compared to the general population, may be vastly different in their health-status allowing for these characteristics to exist but also reducing their inter-comparability. As such, selection bias may be introduced which cannot be handled adequately in the modelling.<sup>194-196</sup> This selection bias is important to recognize but is not a problem per se in these studies. This could indirectly theoretically have an impact on Paper I reducing the power to validate the findings in the main exploratory cohort, as donors would be of reduced risk, as compared to the high-risk group of recipients of blood or individuals who have undertaken a blood group test for some reason other than to donate blood. However, this was not the case as the number of events in all disease groups were at least 50 – which could be related to the long follow-up recorded in donors.

### **7.2.2 Misclassification and measurement error of outcomes and exposure information**

One of the major issues with all of the studies is the construction of outcome or exposure categories allowing for misclassification of outcomes. As the outcome categories are based on hospital-reported ICD information – which is correlated to health care reimbursement – this may introduce multiple levels of misclassification.<sup>174</sup> A further issue, is the translation and bundling of diagnosis codes used at different time-

periods when some medical conditions were not even discovered or impossible to diagnose without the fine-tuned instruments available in more recent times.

In Paper I and III, the problem of misclassification was addressed by limiting the analysis to use only recent outcome data, and utilizing the most recent ICD iteration. In Paper I, however, it would also be unlikely that there was differential misclassification based on ABO blood group and RhD status. The misclassification is in this instance rather an measurement error based on the underlying definitions of disease leading to a diagnosis code according to the ICD within the hospital coding system. In all studies, this is somewhat accounted for by adjusting for calendar period. Further, the exposure may be misclassified in Paper III as the risk-increase for a certain event may be delayed and may also be the reason for switching to another treatment. E.g. the risk of myocardial infarction may not become markedly reduced until a washout-period of several months or years for a specific TKI has passed. Hence if treatment switch occurs, that event would be counted towards the new TKI rather than the former. A possible way by which this misclassification could be tackled and further elucidated, would be to generate analyses with covariates such as ever-treated or defining groups treated with a certain sequence of TKI.

In Paper II the misclassification may have a larger impact as it relies on follow-up from 1968 until 2017 with the manual custom translation of mainly ICD 8 to ICD 10 but also a translation from ICD 9 to 10. Again, however, this misclassification is unlikely to be differential but may reduce the power to detect disease transmission, especially in distant times as coding was more sparse, and not reflecting heterogeneity of the disease spectrum. The same limitation applies to Paper IV, but is constrained to the usage of cancer diagnosis data. However there may also exist another form of misclassification where families with apparent increased incidence of disease with generally low mortality may indeed become increasingly screened as compared to other families. As a corollary, there may seem to be clustering or an unexpectedly high incidence of these diseases, when instead there is an increased tendency to screen for the particular disease. An example would be that of prostate cancer, where many low-grade instances may not require treatment or only hormonal treatment, but could potentially be detected with increasing incidence in families with a family history of prostate cancer. This would then be depicted as a reduced mortality in these families related to prostate cancer as compared to other families with the same characteristics of individuals. However, the reverse may also be true in families with strong hereditary syndromes, where there is also generally more aggressive and advanced disease with increased mortality.<sup>197-200</sup>

### **7.2.3 Selection of a control cohort and detection bias**

The study in Paper III is partly conducted with a matched control cohort with the matching based on attained age at diagnosis, sex and place of residency at the time of diagnosis of the cases. A general note of caution when interpreting the results as such lies in the nature of detection bias, in which there is a difference in the probability of becoming diagnosed with some of the outcomes due to the exposure – a diagnosis of CML.<sup>201–203</sup> At CML diagnosis there is a vast array of screening, vigorous blood sampling and cardiac monitoring for example. This is continued during follow-up and patients are informed to alert a physician if new symptoms are revealed. To account for some of this bias, we performed a lag-time, or delayed-entry model, to account for some of the outcomes that appeared in close proximity to the diagnosis. One must however also consider, that in the case of CP-CML, roughly 50% of patients are asymptomatic at diagnosis but have indeed acquired health care leading to the suspicion of CML – as such, the underlying cause for visiting the general practitioner or emergency department, may be another un-diagnosed disease. As such, matching does not account for this initial event. These biases may have different impact depending on severity and acuteness of the investigated outcome, and matching does not control for the bias of this initial event.

### **7.2.4 The increased mortality of transfused patients**

As blood transfusions are given to individuals with a high risk of death and a generally short overall survival (depicted in Study II by the median follow-up time), there may be insufficient follow-up to develop a possible transfusion-transmitted disease if the latency period is long.<sup>204</sup> An important note when interpreting the data in Study II is the need of sufficient follow-up to strongly reject all null-findings. A similar aspect related to the mortality of recipients of blood is the fact that the number of transfusions is highly related to mortality and death. This is a competing-risk in terms of receiving further transfusions. An individual transfused with multiple units of blood is in the same sense also more likely to have received a unit of blood from a donor who later developed a specific disease. To control for this, we have adjusted the analysis for the number of units transfused, as described in 4.2.2.

### **7.2.5 Allocation of blood units for transfusion**

Allocation of blood is in general a structured process when the oldest blood product with matching blood group is distributed to the patient from the local blood bank. This process is, however, disturbed by a few factors that need to be considered. One such factor is the few frequently and chronically transfused patients where products are matched on further immunological factors to reduce alloimmunization.<sup>205–207</sup> This needs to be considered in the analysis as transfused patients characteristics, as well as donor

characteristics, vary between different regions and hospitals. This is relevant to Paper II and we have consequently adjusted for hospital by performing the regression analysis with strata of the hospital of transfusion.

### **7.2.6 Confounding by indication**

An interesting discussion, regarding the results from Paper III, is that confounding by indication may reduce the point estimates for some outcomes. This can be illustrated of by the fact that a patient who suffers from a medical condition, or carries a risk-factor for a condition, may be less likely to receive a drug with the already known side effect.<sup>208</sup> As such, risk ratios for other drugs may be slightly inflated whereas for a drug with a known specific side effect, the risk ratio may be reduced. In Paper III, this could be highlighted by the risk of suffering cardiovascular events in relation to TKI treatment with nilotinib or ponatinib. These drugs were reported and discussed intensely in 2014 regarding their risk for contributing to vascular disease. Since then physicians were instructed not to prescribe these drugs to patients with known cardiovascular risk factors which in turn, using this population-based observational data, could reduce the risk estimates as compared to imatinib since individuals with high cardiovascular risk would be more likely to receive this non-cardiotoxic drug. In addition, clinicians would also screen individuals who were considered to change treatment to any of nilotinib or ponatinib identifying cardiovascular risk factors that were then accounted for in another TKI. This time-dependent confounding effect, called the treatment-confounder feedback, is difficult to account for in a non-randomized setting. Marginal structural models could in part be used to generate unbiased estimates.<sup>202</sup>

## **7.3 Computational limitations**

One of the main issues and limitations in the agnostic approach has been the scarcity of computational resources within the secure environment that is available for storage and analysis of data. The agnostic approach generates the possibility to perform a very large number of variations in input parameters in the models generating even more data allowing for further customization and levels of depth in the analysis. An example would be to allow defining an array of separate definitions of exposure status in Paper II, depending on the disease development and time after donation in the donor. The current model, using highly parallelized data processing using multiple computational servers with roughly 100 CPU cores and array of fast storage and usage of RAM to allocate temporary storage, takes roughly 2 weeks to complete. Another layer of input variables would, for each change in additional input variable, add another 2 weeks using the same setup. As such, to allow for high-scale processing of similar high-resolution

data, there is a need to develop strategies that could be employed within available cloud computational platforms to allow for sensible and scalable computational usage. The same applies for Paper IV, where computational power of high-dimensional data limits the possibility to perform clustering analysis. The production of graphs and visualization of data is also limited by data power and, for example, Figure 1 in Paper IV, contains 210 million datapoints. The solution has been to allow for rasterization of figures and to use GPU-powered processing.

## 8 CONCLUSIONS

Using the agnostic approach in large-scale, as well as smaller scale, population-based registers some conclusions can be made. However, most importantly, this method allow us to generate further hypotheses that should be explored in a refined and focused framework. Also, in all studies, we find what we expect to find which is comforting and supporting of the methodology.

- From Paper I we can, based on the assumption that ABO blood group and RhD status is not affected by any particular confounding factor, make firm conclusions regarding the associations with disease. Strong relationships, in terms of low P-values and varying effect size, exist for 49 ABO blood group and disease and 1 RhD status and disease. Many of the findings have been previously described, a few findings are new and deserve further investigation and interpretation.
- From Paper II we can identify known transfusion-transmitted disease. Identified are hepatitis virus and HIV, which clearly have been transmitted in the Swedish donation-transfusion population. We also identify some conditions with small effect sizes but, significant P-values after adjustment for multiple testing. These need further investigation in more fine-tuned environments.
- From Paper III we identify the expected known severe complications in the CP-CML population. Further exploration of data is needed within the ophthalmic complications, as they are frequent and diverse and not previously well described.
- From Paper IV we identify established malignant disease with known familial clustering, e.g. BRCA1/2 and hereditary colon cancer. However, the methodology needs further tuning and allocation of computational resources to allow for more complex modelling.

## 9 POINTS OF PERSPECTIVE

### 9.1 Future studies

#### 9.1.1 The association between blood groups and disease

- Specific epidemiological studies addressing some of the novel associations identified, e.g. calculus of the kidney and ureter, pregnancy-induced hypertension, well-differentiated thyroid cancer, and sarcoidosis.
- In cases where more fine-tuned modelling confirm the findings further construct plausible mechanistic studies.

#### 9.1.2 Transfusion-transmitted disease

- Develop the approach to allow for variation in exposure definition.
- Validate model and general framework in other similar cohorts.
- Run the model in the upcoming fourth iteration of the SCANDAT database.

#### 9.1.3 Adverse events in CP-CML patients

- Specifically address the ophthalmic complications in fine-tuned models.
- Address newer generation TKIs in new register-linkages with more follow-up (specifically ponatinib, in the future also asciminib).

#### 9.1.4 Familial disease clustering

- Develop the approach to allow also non-malignant disease data as well as usage of clustering algorithms by using distributed computing.

### 9.2 Development of the agnostic approach

- Integrate distributed computing with high computational power to allow for more model input variation.
- Apply modeling concepts from machine-learning, e.g. brute-force methodology comparing a large number of models for fit using a metric that can be compared in all models tested.
- Allow publishing of live graphs and tables to be able to interpret the data.
- Evaluation of the translations of ICD coding



# 10 ACKNOWLEDGEMENTS

There is no single sentence that can accurately describe the gratefulness that I have towards my main supervisor, Gustaf. Thank you for allowing me to fail and succeed and learn under your supervision, thank you for the endless discussions with prompt responses at any time of the day or night. You have helped me to become independent – independent in its best possible meaning.

My co-supervisors, Mark, Martin and Patrik, you have all given me support in many ways on the journey – Thank you!

Leif, for being the best possible mentor one could wish for. Your kindness, empathy, and knowledge is inspiring and something to strive for. Thank you for introducing me to the national and international CML community.

The research group, Jing (thanks for everything, this thesis would not be possible without you, are you superhuman?), Anne, Cecilia, Lucas and Torkel – Thanks for all discussions regarding research and non-research.

Joel, my clinical mentor (and possibly friend) during my residency – thanks for all the structured mentor meetings (all according to the protocol of the National Board of Health and Welfare). Keep on being Joel, you're good at it!

Kristina and Maria, for being my clinical superiors during this thesis – You are both tremendous leaders and have generously given me time to conduct business. Thanks, and I'm especially grateful for the continued extensive support you have promised me!

Colleagues at the Hematology Department, thanks to everyone for the last 11 years – I love going to work everyday.

Jonatan, thanks for being an excellent friend. Thanks for telling me that, "You only miss the things you didn't buy", which can be extrapolated to any situation.

Mom and Dad, for possibly inspiring me to do what I do and for all the help and discussions on the way and for all of life. Extra thanks to my mom for proof-reading my thesis. Ragnar, with his family, for being my bigger brother always helping me when I'm in need. My extended Ericsson/Eriksson-family for all the generous support during the way!

My family, my truly amazing wife Sofia, our kids – you make my mind occupied with joy and release me from all stressors in life.

# 11 REFERENCES

1. Pearl J. Causal inference in statistics: An overview. *Statistics Surv.* 2009;3.
2. Uffelmann E, Huang QQ, Munung NS, Vries J de, Okada Y, Martin AR, et al. Genome-wide association studies. *Nat Rev Methods Primers.* 2021;1(1):59.
3. Learoyd P. The history of blood transfusion prior to the 20th century—part 2. *Transfusion Med.* 2012;22(6):372–6.
4. Blundell. OBSERVATIONS ON TRANSFUSION OF BLOOD. *Lancet.* 1829;12(302):321–4.
5. Roux FA, Sai P, Deschamps J. Xenotransfusions, past and present. *Xenotransplantation.* 2007;14(3):208–16.
6. Boulton FE. Blood transfusion; additional historical aspects. Part 1. The birth of transfusion immunology. *Transfusion Med.* 2013;23(6):375–81.
7. Kamper-Jørgensen M, Edgren G, Rostgaard K, Biggar RJ, Nyrén O, Reilly M, et al. Blood transfusion exposure in Denmark and Sweden. *Transfusion.* 2009;49(5):888–94.
8. Zhao J, Rostgaard K, Hjalgrim H, Edgren G. The Swedish Scandinavian donations and transfusions database (SCANDAT3-S) – 50 years of donor and recipient follow-up. *Transfusion.* 2020;60(12):3019–27.
9. Carson JL, Guyatt G, Heddle NM, Grossman BJ, Cohn CS, Fung MK, et al. Clinical Practice Guidelines From the AABB: Red Blood Cell Transfusion Thresholds and Storage. *Jama.* 2016;316(19):2025.
10. Roberts N, James S, Delaney M, Fitzmaurice C. The global need and availability of blood products: a modelling study. *Lancet Haematol.* 2019;6(12):e606–15.
11. SweBA. Swedish Blood Alliance 2021 [Internet]. 2021 [cited 2023 Feb 1]. Available from: [http://www.kitm.se/wp-content/uploads/2022/05/Blodverksamheten-i-Sverige-2021-v1\\_O.pdf](http://www.kitm.se/wp-content/uploads/2022/05/Blodverksamheten-i-Sverige-2021-v1_O.pdf)
12. Wikman A, Larsson S, Storry J, Schött U, Folatre JGS. [Blood components, special components and whole blood – what and to whom?]. *Lakartidningen.* 2021;118.
13. Dean CL, Wade J, Roback JD. Transfusion–Transmitted Infections: an Update on Product Screening, Diagnostic Techniques, and the Path Ahead. *J Clin Microbiol.* 2018;56(7):e00352–18.
14. Bassuni WY, Blajchman MA, Al-Moshary MA. Why implement universal leukoreduction? *Hematology Oncol Stem Cell Ther.* 2008;1(2):106–23.

15. Simancas-Racines D, Osorio D, Martí-Carvajal AJ, Arevalo-Rodriguez I. Leukoreduction for the prevention of adverse reactions from allogeneic blood transfusion. *Cochrane Db Syst Rev*. 2015;2015(12):CD009745.
16. Dzik WH. Leukoreduction of blood components. *Curr Opin Hematol*. 2002;9(6):521–6.
17. Manduzio P. Transfusion-associated graft-versus-host disease: A concise review. *Hematology Reports*. 2018;10(4):7724.
18. Bahar B, Tormey CA. Prevention of Transfusion-Associated Graft-Versus-Host Disease With Blood Product Irradiation: The Past, Present, and Future. *Arch Pathol Lab Med*. 2018;142(5):662–7.
19. Tobian AAR, Savage WJ, Tisch DJ, Thoman S, King KE, Ness PM. Prevention of allergic transfusion reactions to platelets and red blood cells through plasma reduction. *Transfusion*. 2011;51(8):1676–83.
20. Cook RJ, Heddle NM, Lee KA, Arnold DM, Crowther MA, Devereaux PJ, et al. Red blood cell storage and in-hospital mortality: a secondary analysis of the INFORM randomised controlled trial. *Lancet Haematol*. 2017;4(11):e544–52.
21. E. SM, M. NP, F. AS, J. TD, R. SS, Meghan D, et al. Effects of Red-Cell Storage Duration on Patients Undergoing Cardiac Surgery. *New Engl J Med*. 2015;372(15):1419–29.
22. Fergusson DA, Hébert P, Hogan DL, LeBel L, Rouvinez-Bouali N, Smyth JA, et al. Effect of Fresh Red Blood Cell Transfusions on Clinical Outcomes in Premature, Very Low-Birth-Weight Infants: The ARIPI Randomized Trial. *Jama*. 2012;308(14):1443–51.
23. Lacroix J, Hébert PC, Fergusson DA, Tinmouth A, Cook DJ, Marshall JC, et al. Age of Transfused Blood in Critically Ill Adults. *New Engl J Med*. 2015;372(15):1410–8.
24. Dhabangi A, Ainomugisha B, Cserti-Gazdewich C, Ddungu H, Kyeyune D, Musisi E, et al. Effect of Transfusion of Red Blood Cells With Longer vs Shorter Storage Duration on Elevated Blood Lactate Levels in Children With Severe Anemia: The TOTAL Randomized Clinical Trial. *Jama*. 2015;314(23):1–10.
25. Storch EK, Custer BS, Jacobs MR, Menitove JE, Mintz PD. Review of current transfusion therapy and blood banking practices 11 This article reflects the views of the author and should not be construed to represent FDA's views or policies. *Blood Rev*. 2019;38:100593.
26. García-Roa M, Vicente-Ayuso MDC, Bobes AM, Pedraza AC, González-Fernández A, Martín MP, et al. Red blood cell storage time and transfusion: current practice, concerns and future perspectives. *Blood Transfus Trasfusione Del Sangue*. 2016;15(3):222–31.
27. Thomas S. Platelets: handle with care. *Transfusion Med*. 2016;26(5):330–8.
28. Ohto H, Nollet KE. Overview on platelet preservation: Better controls over storage lesion. *Transfus Apher Sci*. 2011;44(3):321–5.

29. Carson JL, Stanworth SJ, Roubinian N, Fergusson DA, Triulzi D, Doree C, et al. Transfusion thresholds and other strategies for guiding allogeneic red blood cell transfusion. *Cochrane Db Syst Rev.* 2016;2016(10):CD002042.
30. Wikman A, Egenvall M, Jansson KÅ, Jeppsson A, Lindgren S, Nilsson M, et al. [Patient blood management – to transfuse blood on appropriate indications]. *Lakartidningen.* 2020;117.
31. Chung K, Harvey A, Basavaraju SV, Kuehnert MJ. How is national recipient hemovigilance conducted in the United States? *Transfusion.* 2015;55(4):703–7.
32. Rieux C, Brittenham G, Bachir D, Meyer ED, Boudjedir K, network F hemovigilance. Delayed hemolytic transfusion reaction in the French hemovigilance system. *Transfus Clin Biol.* 2019;26(2):109–11.
33. Daniels G. (2013). ABO, H, and Lewis Systems. In *Human Blood Groups*. In: *Human Blood Groups*. Wiley–Blackwell.
34. Reid ME, Lomas–Francis C, Olsson ML.(2012). *The Blood Group Antigen FactsBook* (3rd Edition). Academic Press.
35. Hosoi E. Biological and clinical aspects of ABO blood group system. *J Medical Investigation.* 2008;55(3,4):174–82.
36. Simmons DP, Savage WJ. Hemolysis from ABO Incompatibility. *Hematology Oncol Clin North Am.* 2015;29(3):429–43.
37. Storry JR, Olsson ML. The ABO blood group system revisited: a review and update. *Immunohematol.* 2009;25(2):48–59.
38. Harmening DM. (2018). *Modern blood banking & transfusion practices.* (7th Edition). Philadelphia, PA: F.A. Davis Company.
39. Yamamoto F, Marken J, Tsuji T, White T, Clausen H, Hakomori S. Cloning and characterization of DNA complementary to human UDP–GalNAc: Fuc alpha 1----2Gal alpha 1----3GalNAc transferase (histo–blood group A transferase) mRNA. *J Biological Chem.* 1990;265(2):1146–51.
40. Ahmed M, Memon A, Iqbal K. Distribution pattern of ABO and Rhesus blood groups among different ethnic population of Karachi. *Jpma J Pak Medical Assoc.* 2019;69(10):1474–8.
41. Apecu RO, Mulogo EM, Bagenda F, Byamungu A. ABO and Rhesus (D) blood group distribution among blood donors in rural south western Uganda: a retrospective study. *Bmc Res Notes.* 2016;9(1):513.
42. Canizalez–Román A, Campos–Romero A, Castro–Sánchez JA, López–Martínez MA, Andrade–Muñoz FJ, Cruz–Zamudio CK, et al. Blood Groups Distribution and Gene Diversity of the ABO and Rh (D) Loci in the Mexican Population. *Biomed Res Int.* 2018;2018:1925619.

43. Golassa L, Tsegaye A, Erko B, Mamo H. High rhesus (Rh(D)) negative frequency and ethnic-group based ABO blood group distribution in Ethiopia. *Bmc Res Notes*. 2017;10(1):330.
44. Lialiaris T, Digkas E, Kareli D, Pouliliou S, Asimakopoulos B, Pagonopoulou O, et al. Distribution of ABO and Rh blood groups in Greece: an update. *Int J Immunogenet*. 2011;38(1):1–5.
45. Liu J, Zhang S, Wang Q, Shen H, Zhang Y, Liu M. Frequencies and ethnic distribution of ABO and RhD blood groups in China: a population-based cross-sectional study. *Bmj Open*. 2017;7(12):e018476.
46. Volken T, Crawford RJ, Amar S, Mosimann E, Tschaggelar A, Taleghani BM. Blood Group Distribution in Switzerland – a Historical Comparison. *Transfus Med Hemoth*. 2017;44(4):210–6.
47. Bronte-Stewart B, Botha MC, Krut LH. ABO Blood Groups in Relation to Ischaemic Heart Disease. *Brit Med J*. 1962;1(5293):1646.
48. Vasan SK, Rostgaard K, Majeed A, Ullum H, Titlestad KE, Pedersen OBV, et al. ABO Blood Group and Risk of Thromboembolic and Arterial Disease. *Circulation*. 2016;133(15):1449–57.
49. Goel R, Bloch EM, Pirenne F, Al-Riyami AZ, Crowe E, Dau L, et al. ABO blood group and COVID-19: a review on behalf of the ISBT COVID-19 working group. *Vox Sang*. 2021;116(8):10.1111/vox.13076.
50. Dahlén T, Li H, Nyberg F, Edgren G. A population-based, retrospective cohort study of the association between ABO blood group and risk of COVID-19. *J Intern Med*. 2023 Mar;293(3):398–402.
51. Reilly JP, Meyer NJ, Shashaty MGS, Feng R, Lankester PN, Gallop R, et al. ABO Blood Type A Is Associated With Increased Risk of ARDS in Whites Following Both Major Trauma and Severe Sepsis. *Chest*. 2014;145(4):753–61.
52. Fan Q, Zhang W, Li B, Li DJ, Zhang J, Zhao F. Association Between ABO Blood Group System and COVID-19 Susceptibility in Wuhan. *Front Cell Infect Mi*. 2020;10:404.
53. Cserti CM, Dzik WH. The ABO blood group system and *Plasmodium falciparum* malaria. *Blood*. 2007;110(7):2250–8.
54. Terzi E, Türsen B, Dursun P, Erdem T, Türsen Ü. The Relationship between ABO Blood Groups and Acne Vulgaris. *Saudi J Medicine Medical Sci*. 2016;4(1):26–8.
55. Edgren G, Hjalgrim H, Rostgaard K, Norda R, Wikman A, Melbye M, et al. Risk of Gastric Cancer and Peptic Ulcers in Relation to ABO Blood Type: A Cohort Study. *Am J Epidemiol*. 2010;172(11):1280–5.
56. Franchini M, Favaloro EJ, Targher G, Lippi G. ABO blood group, hypercoagulability, and cardiovascular and cancer risk. *Crit Rev Cl Lab Sci*. 2012;49(4):137–49.

57. Orstavik KH, Magnus P, Reisner H, Berg K, Graham JB, Nance W. Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. *Am J Hum Genet.* 1985;37(1):89–101.
58. Germain M, Chasman DI, Haan H de, Tang W, Lindström S, Weng LC, et al. Meta-analysis of 65,734 Individuals Identifies TSPAN15 and SLC44A2 as Two Susceptibility Loci for Venous Thromboembolism. *Am J Hum Genetics.* 2015;96(4):532–42.
59. Lindström S, Brody JA, Turman C, Germain M, Bartz TM, Smith EN, et al. A large-scale exome array analysis of venous thromboembolism. *Genet Epidemiol.* 2019;43(4):449–57.
60. O'Donnell J, Boulton FE, Manning RA, Laffan MA. Amount of H Antigen Expressed on Circulating von Willebrand Factor Is Modified by ABO Blood Group Genotype and Is a Major Determinant of Plasma von Willebrand Factor Antigen Levels. *Arteriosclerosis Thrombosis Vasc Biology.* 2002;22(2):335–41.
61. Franchini M, Bonfanti C. Evolutionary aspects of ABO blood group in humans. *Clin Chim Acta.* 2015;444:66–71.
62. Degarege A, Gebrezgi MT, Ibanez G, Wahlgren M, Madhivanan P. Effect of the ABO blood group on susceptibility to severe malaria: A systematic review and meta-analysis. *Blood Rev.* 2019;33:53–62.
63. Stowell CP, Stowell SR. Biologic roles of the ABH and Lewis histo-blood group antigens Part I: infection and immunity. *Vox Sang.* 2019;(5).
64. Gérard C, Maggipinto G, Minon J. COVID-19 and ABO blood group: another viewpoint. *Brit J Haematol.* 2020;190(2):e93–4.
65. Ray JG, Schull MJ, Vermeulen MJ, Park AL. Association Between ABO and Rh Blood Groups and SARS-CoV-2 Infection or Severe COVID-19 Illness. *Ann Intern Med.* 2020;174(3):M20–4511.
66. Davoodi L, Razavi A, Jafarpour H, Heshmati M, Soleymani E, Ghasemian R. Relationship Between the Prevalence of Blood Groups and Severity of Leptospirosis: A Case-Control Study. *Infect Dis Res Treat.* 2020;13:1178633720936273.
67. Wolpin BM, Chan AT, Hartge P, Chanock SJ, Kraft P, Hunter DJ, et al. ABO Blood Group and the Risk of Pancreatic Cancer. *Jnci J National Cancer Inst.* 2009;101(6):424–31.
68. Amundadottir L, Kraft P, Stolzenberg-Solomon RZ, Fuchs CS, Petersen GM, Arslan AA, et al. Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nat Genet.* 2009;41(9):986–90.
69. Li S, Xu RX, Guo YL, Zhang Y, Zhu CG, Sun J, et al. ABO blood group in relation to plasma lipids and proprotein convertase subtilisin/kexin type 9. *Nutrition Metabolism Cardiovasc Dis.* 2015;25(4):411–7.

70. Nakao M, Matsuo K, Ito H, Shitara K, Hosono S, Watanabe M, et al. ABO Genotype and the Risk of Gastric Cancer, Atrophic Gastritis, and Helicobacter pylori Infection. *Cancer Epidem Biomar*. 2011;20(8):1665–72.
71. Yu H, Xu N, Li ZK, Xia H, Ren HT, Li N, et al. Association of ABO Blood Groups and Risk of Gastric Cancer. *Scand J Surg*. 2018;109(4):309–13.
72. Klein HG, Anstee DJ. *Mollison's blood transfusion in clinical medicine*. 12th ed. Chichester, West Sussex, UK: Wiley–Blackwell; 2014. 1 p.
73. Avent ND, Reid ME. The Rh blood group system: a review. *Blood*. 2000;95(2):375–87.
74. Hendrickson JE, Delaney M. Hemolytic Disease of the Fetus and Newborn: Modern Practice and Future Investigations. *Transfus Med Rev*. 2016;30(4):159–64.
75. Anstee DJ. The relationship between blood groups and disease. *Blood*. 2010;115(23):4635–43.
76. Busch MP. Transfusion-transmitted viral infections: building bridges to transfusion medicine to reduce risks and understand epidemiology and pathogenesis. *Transfusion*. 2006;46(9):1624–40.
77. Busch MP, Bloch EM, Kleinman S. Prevention of transfusion–transmitted infections. *Blood*. 2019;133(17):1854–64.
78. Schryver A, Meheus AZ. Syphilis and blood transfusion: a global perspective. *Transfusion*. 1990;30(9):844–7.
79. Schmidt PJ. Syphilis, a disease of direct transfusion. *Transfusion*. 2001;41(8):1069–71.
80. Tamrakar P, Bett C, Molano RD, Ayub A, Asher DM, Gregori L. Effect of storage on survival of infectious *Treponema pallidum* spiked in whole blood and platelets. *Transfusion*. 2021;61(11):3181–9.
81. Jayawardena T, Hoad V, Styles C, Seed C, Bentley P, Clifford V, et al. Modelling the risk of transfusion-transmitted syphilis: a reconsideration of blood donation testing strategies. *Vox Sang*. 2019;114(2):107–16.
82. Seto WK, Lo YR, Pawlotsky JM, Yuen MF. Chronic hepatitis B virus infection. *Lancet*. 2018;392(10161):2313–24.
83. Pol S, Lagaye S. The remarkable history of the hepatitis C virus. *Genes Immun*. 2019;20(5):436–46.
84. Yi Z, Yuan Z. Infectious Agents Associated Cancers: Epidemiology and Molecular Biology. *Adv Exp Med Biol*. 2017;1018:129–46.

85. Dahl V, Majeed A, Wikman A, Norda R, Edgren G. Transmission of viral hepatitis through blood transfusion in Sweden, 1968 to 2012. *Eurosurveillance*. 2020;25(29):1900537.
86. Stramer SL, Glynn SA, Kleinman SH, Strong DM, Caglioti S, Wright DJ, et al. Detection of HIV-1 and HCV Infections among Antibody-Negative Blood Donors by Nucleic Acid-Amplification Testing. *New Engl J Medicine*. 2004;351(8):760-8.
87. Jarvis LM, Dow BC, Cleland A, Davidson F, Lycett C, Morris K, et al. Detection of HCV and HIV-1 antibody negative infections in Scottish and Northern Ireland blood donations by nucleic acid amplification testing. *Vox Sang*. 2005;89(3):128-34.
88. Velati C, Romanò L, Piccinini V, Marano G, Catalano L, Pupella S, et al. Prevalence, incidence and residual risk of transfusion-transmitted hepatitis C virus and human immunodeficiency virus after the implementation of nucleic acid testing in Italy: a 7-year (2009-2015) survey. *Blood Transfusio Trasfusione Del Sangue*. 2018;16(5):422-32.
89. Lucas S, Nelson AM. HIV and the spectrum of human disease. *J Pathology*. 2015;235(2):229-41.
90. Ghosn J, Taiwo B, Seedat S, Autran B, Katlama C. HIV. *Lancet*. 2018;392(J Infect Dis 215 2017):685-97.
91. Graaf M van der, Diepersloot RJA. Transmission of human immunodeficiency virus (HIV/HTLV-III/LAV): A review. *Infection*. 1986;14(5):203-11.
92. Cappy P, Barlet V, Lucas Q, Tinard X, Pillonel J, Gross S, et al. Transfusion of HIV-infected blood products despite highly sensitive nucleic acid testing. *Transfusion*. 2019;59(6):2046-53.
93. Tagaya Y, Matsuoka M, Gallo R. 40 years of the human T-cell leukemia virus: past, present, and future. *F1000research*. 2019;8:F1000 Faculty Rev-228.
94. Malm K, Ekermo B, Hillgren K, Britton S, Fredlund H, Andersson S. Prevalence of human T-lymphotropic virus type 1 and 2 infection in Sweden. *Scand J Infect Dis*. 2012;44(11):852-9.
95. Tynell E, Andersson S, Lithander E, Arneborn M, Blomberg J, Hansson HB, et al. Screening for human T cell leukaemia/lymphoma virus among blood donors in Sweden: cost effectiveness analysis. *Bmj*. 1998;316(7142):1417.
96. Holt E. West Nile virus spreads in Europe. *Lancet Infect Dis*. 2018;18(11):1184.
97. Betsem E, Kaidarova Z, Stramer SL, Shaz B, Sayers M, LeParc G, et al. Correlation of West Nile Virus Incidence in Donated Blood with West Nile Neuroinvasive Disease Rates, United States, 2010-2012 - Volume 23, Number 2-February 2017 - Emerging Infectious Diseases journal - CDC. *Emerg Infect Dis*. 2017;23(2):212-9.
98. Stramer SL, Hollinger FB, Katz LM, Kleinman S, Metzger PS, Gregory KR, et al. Emerging infectious disease agents and their potential threat to transfusion safety. *Transfusion*. 2009;49(s2):1S-29S.



99. Pérez-Molina JA, Molina I. Chagas disease. *Lancet*. 2018;391(10115):82–94.
100. Olsson J, Kok E, Adolfsson R, Lövheim H, Elgh F. Herpes virus seroepidemiology in the adult Swedish population. *Immun Ageing*. 2017;14(1):10.
101. Svahn A, Berggren J, Parke A, Storsaeter J, Thorstensson R, Linde A. Changes in seroprevalence to four herpesviruses over 30 years in Swedish children aged 9–12 years. *J Clin Virol*. 2006;37(2):118–23.
102. Committee A Clinical Transfusion Medicine, Heddle NM, Boeckh M, Grossman B, Jacobson J, Kleinman S, et al. AABB Committee Report: reducing transfusion-transmitted cytomegalovirus infections. *Transfusion*. 2016;56(6pt2):1581–7.
103. Qu L, Xu S, Rowe D, Triulzi D. Efficacy of Epstein-Barr virus removal by leukoreduction of red blood cells. *Transfusion*. 2005;45(4):591–5.
104. Trottier H, Buteau C, Robitaille N, Duval M, Tucci M, Lacroix J, et al. Transfusion-related Epstein-Barr virus infection among stem cell transplant recipients: a retrospective cohort study in children. *Transfusion*. 2012;52(12):2653–63.
105. Delaney M, Wendel S, Bercovitz RS, Cid J, Cohn C, Dunbar NM, et al. Transfusion reactions: prevention, diagnosis, and treatment. *Lancet*. 2016;388(10061):2825–36.
106. Hewitt PE, Llewelyn CA, Mackenzie J, Will RG. Three reported cases of variant Creutzfeldt–Jakob disease transmission following transfusion of labile blood components. *Vox Sang*. 2006;91(4):348–348.
107. Edgren G, Hjalgrim H, Rostgaard K, Lambert P, Wikman A, Norda R, et al. Transmission of Neurodegenerative Disorders Through Blood Transfusion. *Ann Intern Med*. 2016;165(5):316.
108. Hjalgrim H, Rostgaard K, Vasan SK, Ullum H, Erikstrup C, Pedersen OBV, et al. No evidence of transmission of chronic lymphocytic leukemia through blood transfusion. *Blood*. 2015;126(17):2059–61.
109. Musso D, Stramer SL, Committee ATTD, Busch MP, Diseases IS of BTWP on TTI. Zika virus: a new challenge for blood transfusion. *Lancet*. 2016;387(10032):1993–4.
110. Paixao ES, Cardim LL, Costa MCN, Brickley EB, Carvalho–Sauer RCO de, Carmo EH, et al. Mortality from Congenital Zika Syndrome — Nationwide Cohort Study in Brazil. *New Engl J Med*. 2022;386(8):757–67.
111. Russell WA. Estimating the Effect of Discontinuing Universal Screening of Donated Blood for Zika Virus in the 50 U.S. States. *Ann Intern Med*. 2021;174(5):728–30.
112. Cho HJ, Koo JW, Roh SK, Kim YK, Suh JS, Moon JH, et al. COVID-19 transmission and blood transfusion: A case report. *J Infect Public Heal*. 2020;13(11):1678–9.
113. Chang L, Yan Y, Wang L. Coronavirus Disease 2019: Coronaviruses and Blood Safety. *Transfus Med Rev*. 2020;34(2):75–80.

114. He B, Shi Y, Li B, Duan X, Wang Q. Severe Acute Respiratory Syndrome Coronavirus 2 Infection and Blood Safety. *Acta Haematol-basel.* 2022;145(4):347–53.
115. Cladel NM, Jiang P, Li JJ, Peng X, Cooper TK, Majerciak V, et al. Papillomavirus can be transmitted through the blood and produce infections in blood recipients: Evidence from two animal models. *Emerg Microbes Infec.* 2019;8(1):1108–21.
116. Petersen LR, Busch MP. Transfusion-transmitted arboviruses. *Vox Sang.* 2010;98(4):495–503.
117. Goldman JM. Chronic Myeloid Leukemia: A Historical Perspective. *Semin Hematol.* 2010;47(4):302–11.
118. Wan TSK. Cancer Cytogenetics: Methodology Revisited. *Ann Lab Med.* 2014;34(6):413–25.
119. Höglund M, Sandin F, Hellström K, Björemann M, Björkholm M, Brune M, et al. Tyrosine kinase inhibitor usage, treatment outcome, and prognostic scores in CML: report from the population-based Swedish CML registry. 2013;15;122(7):1284–92 122(7).
120. Cortes JE, Talpaz M, Kantarjian H. Chronic myelogenous leukemia: A review. *Am J Medicine.* 1996;100(5):555–70.
121. Gertz MA. Acute hyperviscosity: syndromes and management. *Blood.* 2018;132(13):1379–85.
122. ROWLEY JD. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature.* 1973;243(5405):290–3.
123. Kantarjian H, Sawyers C, Hochhaus A, Guilhot F, Schiffer C, Gambacorti-Passerini C, et al. Hematologic and Cytogenetic Responses to Imatinib Mesylate in Chronic Myelogenous Leukemia. *New Engl J Medicine.* 2002;346(9):645–52.
124. Sawyers CL. Chronic Myeloid Leukemia. *New Engl J Medicine.* 1999;340(17):1330–40.
125. O’Brien SG, Guilhot F, Larson R a, Gathmann I, Baccarani M, Cervantes F, et al. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. 2003;348(11).
126. Hochhaus A, Larson RA, Guilhot F, Radich JP, Branford S, Hughes TP, et al. Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. *New Engl J Medicine.* 2017;376(10):917–27.
127. Druker BJ, Talpaz M, Resta DJ, Peng B, Buchdunger E, Ford JM, et al. Efficacy and Safety of a Specific Inhibitor of the BCR-ABL Tyrosine Kinase in Chronic Myeloid Leukemia. *New Engl J Medicine.* 2001;344(14):1031–7.

128. Bower H, Björkholm M, Dickman PW, Höglund M, Lambert PC, Andersson TML. Life Expectancy of Patients With Chronic Myeloid Leukemia Approaches the Life Expectancy of the General Population. *J Clin Oncol*. 2016;34(24):2851–7.
129. Jabbour E, Kantarjian H, Cortes J. Use of Second- and Third-Generation Tyrosine Kinase Inhibitors in the Treatment of Chronic Myeloid Leukemia: An Evolving Treatment Paradigm. 2015;15(6).
130. Cortes JE, Gambacorti-Passerini C, Deininger MW, Mauro MJ, Chuah C, Kim DW, et al. Bosutinib Versus Imatinib for Newly Diagnosed Chronic Myeloid Leukemia: Results From the Randomized BFORE Trial. *J Clin Oncol*. 2017;36(3):JCO.2017.74.716.
131. Cortes J, Kim DW, Pinilla-Barz J, Coutre P, Paquette R, Chuah C, et al. A phase 2 trial of ponatinib in Philadelphia chromosome-positive leukemias. 2013;369(19).
132. Kantarjian H, Shah NP, Hochhaus A, Cortes J, Shah S, Ayala M, et al. Dasatinib versus Imatinib in Newly Diagnosed Chronic-Phase Chronic Myeloid Leukemia. 2010;362(24).
133. Lipton JH, Chuah C, Guerci-Bresler A, Rosti G, Simpson D, Assouline S, et al. Ponatinib versus imatinib for newly diagnosed chronic myeloid leukaemia: an international, randomised, open-label, phase 3 trial. *Lancet Oncol*. 2016;17(5):612–21.
134. Hochhaus A, Saglio G, Hughes TP, Larson RA, Kim DW, Issaragrisil S, et al. Long-term benefits and risks of frontline nilotinib vs imatinib for chronic myeloid leukemia in chronic phase: 5-year update of the randomized ENESTnd trial. *Leukemia*. 2016;30(5):1044–54.
135. Cortes JE, Saglio G, Kantarjian HM, Baccarani M, Mayer J, Boqué C, et al. Final 5-Year Study Results of DASISION: The Dasatinib Versus Imatinib Study in Treatment-Naïve Chronic Myeloid Leukemia Patients Trial. *J Clin Oncol*. 2016;34(20):2333–40.
136. Hughes TP, Mauro MJ, Cortes JE, Minami H, Rea D, DeAngelo DJ, et al. Asciminib in Chronic Myeloid Leukemia after ABL Kinase Inhibitor Failure. *New Engl J Med*. 2019;381(24):2315–26.
137. Ross DM, Hughes TP. Treatment-free remission in patients with chronic myeloid leukaemia. *Nat Rev Clin Oncol*. 2020;17(8):493–503.
138. Mahon FX, Réa D, Guilhot J, Guilhot F, Huguet F, Nicolini F, et al. Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre Stop Imatinib (STIM) trial. *Lancet Oncol*. 2010;11(11):1029–35.
139. Saussele S, Richter J, Guilhot J, Gruber FX, Hjorth-Hansen H, Almeida A, et al. Discontinuation of tyrosine kinase inhibitor therapy in chronic myeloid leukaemia (EURO-SKI): a prespecified interim analysis of a prospective, multicentre, non-randomised, trial. *Lancet Oncol*. 2018;19(6):747–757
140. Flygt H, Sandin F, Dahlén T, Dremaine A, Lübking A, Markevörn B, et al. Successful tyrosine kinase inhibitor discontinuation outside clinical trials — data from the

population-based Swedish chronic myeloid leukaemia registry. *Brit J Haematol.* 2021;193(5):915–21.

141. Steegmann J, Baccarani M, Breccia M, Casado L, García-Gutiérrez V, Hochhaus A, et al. European LeukemiaNet recommendations for the management and avoidance of adverse events of treatment in chronic myeloid leukaemia. 2016;30(8).

142. Hochhaus A, Baccarani M, Silver RT, Schiffer C, Apperley JF, Cervantes F, et al. European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. *Leukemia.* 2020;34(4):966–84.

143. Giles F, Mauro M, Hong F, Ortmann CE, McNeill C, Woodman R, et al. Rates of peripheral arterial occlusive disease in patients with chronic myeloid leukemia in the chronic phase treated with imatinib, nilotinib, or non-tyrosine kinase therapy: a retrospective cohort analysis. *Leukemia.* 2013;27(6):1310–5

144. Kim T, Rea D, Schwarz M, Grille P, Nicolini F, Rosti G, et al. Peripheral artery occlusive disease in chronic phase chronic myeloid leukemia patients treated with nilotinib or imatinib. *Leukemia.* 2013;27(6):1316–21.

145. Cortes JE, Kim DW, Pinilla-Ibarz J, Coutre PD, Paquette R, Chuah C, et al. Ponatinib efficacy and safety in Philadelphia chromosome-positive leukemia: final 5-year results of the phase 2 PACE trial. *Blood.* 2018;132(4):393–404.

146. Dahlén T, Edgren G, Lambe M, Höglund M, Björkholm M, Sandin F, et al. Cardiovascular Events Associated With Use of Tyrosine Kinase Inhibitors in Chronic Myeloid Leukemia: A Population-Based Cohort Study. *Ann Intern Med.* 2016;165(3):161.

147. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *Ca Cancer J Clin.* 2018;68(6):394–424.

148. Collaboration GB of DC, Fitzmaurice C, Abate D, Abbasi N, Abbastabar H, Abd-Allah F, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-Years for 29 Cancer Groups, 1990 to 2017. *Jama Oncol.* 2019;5(12):1749–68.

149. Collaboration GB of D 2019 C, Kocarnik JM, Compton K, Dean FE, Fu W, Gaw BL, et al. Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life Years for 29 Cancer Groups From 2010 to 2019. *Jama Oncol.* 2022;8(3):420–44.

150. M. CC. Oncogenes and Cancer. *New Engl J Med.* 2008;358(5):502–11.

151. Wooster R, Neuhausen SL, Mangion J, Quirk Y, Ford D, Collins N, et al. Localization of a Breast Cancer Susceptibility Gene, BRCA2, to Chromosome 13q12–13. *Science.* 1994;265(5181):2088–90.

152. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, et al. Linkage of Early-Onset Familial Breast Cancer to Chromosome 17q21. *Science.* 1990;250(4988):1684–9.

153. Tung NM, Boughey JC, Pierce LJ, Robson ME, Bedrosian I, Dietz JR, et al. Management of Hereditary Breast Cancer: American Society of Clinical Oncology, American Society for Radiation Oncology, and Society of Surgical Oncology Guideline. *J Clin Oncol*. 2020;38(18):2080–106.
154. Gargallo P, Yáñez Y, Segura V, Juan A, Torres B, Balaguer J, et al. Li–Fraumeni syndrome heterogeneity. *Clin Transl Oncol*. 2020;22(7):978–88.
155. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. *Nat Rev Cancer*. 2015;15(3):181–94.
156. Issa JP. The Two–Hit Hypothesis Meets. *Cancer Res*. 2022;82(7):1167–9.
157. Kohlmann W, Schiffman JD. Discussing and managing hematologic germ line variants. *Hematology*. 2016;2016(1):309–15.
158. Brown JR. Inherited predisposition to chronic lymphocytic leukemia. *Expert Rev Hematol*. 2008;1(1):51–61.
159. Lei J, Ploner A, Elfström KM, Wang J, Roth A, Fang F, et al. HPV Vaccination and the Risk of Invasive Cervical Cancer. *New Engl J Med*. 2020;383(14):1340–8.
160. Haaf K ten. Confronting the burden of tobacco–related lung cancer in Europe in the next decades. *Lancet Regional Health – Europe*. 2021;4:100085.
161. Chang MH, Chen CJ, Lai MS, Hsu HM, Wu TC, Kong MS, et al. Universal Hepatitis B Vaccination in Taiwan and the Incidence of Hepatocellular Carcinoma in Children. *New Engl J Medicine*. 1997;336(26):1855–9.
162. Collaboration GB of DLC, Akinyemiju T, Abera S, Ahmed M, Alam N, Alemayohu MA, et al. The Burden of Primary Liver Cancer and Underlying Etiologies From 1990 to 2015 at the Global, Regional, and National Level: Results From the Global Burden of Disease Study 2015. *Jama Oncol*. 2017;3(12):1683–91.
163. Richiardi L, Bellocco R, Adami HO, Torráng A, Barlow L, Hakulinen T, et al. Testicular cancer incidence in eight northern European countries: secular and recent trends. *Cancer Epidemiology Biomarkers Prev Publ Am Assoc Cancer Res Cosponsored Am Soc Prev Oncol*. 2004;13(12):2157–66.
164. Edgren G, Liang L, Adami HO, Chang ET. Enigmatic sex disparities in cancer incidence. *Eur J Epidemiol*. 2012;27(3):187–96.
165. Curran–Everett D. Multiple comparisons: philosophies and illustrations. *Am J Physiology–regulatory Integr Comp Physiology*. 2000;279(1):R1–8.
166. Marino MJ. How often should we expect to be wrong? Statistical power, P values, and the expected prevalence of false discoveries. *Biochem Pharmacol*. 2017;151:226–33.
167. Oord EJCG van den. Controlling false discoveries in genetic studies. *Am J Medical Genetics Part B Neuropsychiatric Genetics*. 2008;147B(5):637–44.

168. HOCHBERG Y. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika*. 1988;75(4):800–2.
169. Korthauer K, Kimes PK, Duvallet C, Reyes A, Subramanian A, Teng M, et al. A practical guide to methods controlling false discoveries in computational biology. *Genome Biol*. 2019;20(1):118.
170. Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society: Series B (Methodological)*. 1995;57(1):289–300.
171. Glickman ME, Rao SR, Schultz MR. False discovery rate control is a recommended alternative to Bonferroni-type adjustments in health studies. *Journal of Clinical Epidemiology*. 2014;67(8):850–7.
172. Blanchard G, Roquain E. Adaptive FDR control under independence and dependence. *Arxiv*. 2007.
173. Storey JD. A direct approach to false discovery rates. *J Royal Statistical Soc Ser B Statistical Methodol*. 2002;64(3):479–98.
174. Ludvigsson JF, Andersson E, Ekblom A, Feychting M, Kim JL, Reuterwall C, et al. External review and validation of the Swedish national inpatient register. *Bmc Public Health*. 2011;11(1):450.
175. Brooke HL, Talbäck M, Hörnblad J, Johansson LA, Ludvigsson JF, Druid H, et al. The Swedish cause of death register. *Eur J Epidemiol*. 2017;32(9):765–73.
176. 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.
177. Ludvigsson JF, Almqvist C, Bonamy AKE, Ljung R, Michaëlsson K, Neovius M, et al. Registers of the Swedish total population and their use in medical research. *Eur J Epidemiol*. 2016;31(2):125–36.
178. Radkiewicz C, Johansson ALV, Dickman PW, Lambe M, Edgren G. Sex differences in cancer risk and survival: A Swedish cohort study. *Eur J Cancer*. 2017;84:130–40.
179. Rostgaard K. Methods for stratification of person-time and events – a prerequisite for Poisson regression and SIR estimation. *Epidemiologic Perspectives Innovations Ep*. 2008;5(1):7–7.
180. Harrell FE. (2001). *Regression modeling strategies : with applications to linear models, logistic and ordinal regression, and survival analysis*. Cham: Springer.
181. Goeman JJ, Solari A. Multiple hypothesis testing in genomics. *Stat Med*. 2014;33(11):1946–78.
182. National Board of Health and Welfare. [Classification ICD10 – Swedish National Board of Health and Welfare] [Internet]. [cited 2023 Jan 4]. Available from: <https://www.socialstyrelsen.se/statistik-och-data/klassifikationer-och-koder/icd-10/>

183. Heide-Jørgensen U, Adelborg K, Kahlert J, Sørensen HT, Pedersen L. Sampling strategies for selecting general population comparison cohorts. *Clin Epidemiology*. 2018;10:1325–37.
184. Gunnarsson N, Stenke L, Höglund M, Sandin F, Björkholm M, Dreimane A, et al. Second malignancies following treatment of chronic myeloid leukaemia in the tyrosine kinase inhibitor era. 2015;
185. Johansson Per. Tests for serial correlation and overdispersion in a count data regression model. *J Stat Comput Sim*. 1995;53(3–4):153–64.
186. Ekblom A. Methods in Biobanking. *Methods Mol Biology*. 2010;675:215–20.
187. Dahlén T, Zhao J, Magnusson PKE, Pawitan Y, Lavröd J, Edgren G. The frequency of misattributed paternity in Sweden is low and decreasing: A nationwide cohort study. *J Intern Med*. 2022;291(1):95–100.
188. Coxe S, West SG, Aiken LS. The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives. *J Pers Assess*. 2009;91(2):121–36.
189. Trafimow D, Marks M. Editorial. *Basic Appl Soc Psych*. 2015;37(1):1–2.
190. Leek JT, Peng RD. Statistics: P values are just the tip of the iceberg. *Nature*. 2015;520(7549):612–612.
191. Vyas D, Balakrishnan A, Vyas A. The Value of the P Value. *Am J Robotic Surg*. 2015;2(1):53–6.
192. Goodman S. A Dirty Dozen: Twelve P-Value Misconceptions. *Semin Hematol*. 2008;45(3):135–40.
193. Greenland S, Senn SJ, Rothman KJ, Carlin JB, Poole C, Goodman SN, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. *Eur J Epidemiol*. 2016;31(4):337–50.
194. Rahman MM, Karki S, Hayen A. A methods review of the “healthy donor effect” in studies of long-term health outcomes in blood donors. *Transfusion*. 2022;62(3):698–712.
195. Edgren G, Hjalgrim H, Rostgaard K, Dahl V, Titlestad K, Erikstrup C, et al. Searching for unknown transfusion-transmitted hepatitis viruses: a binational cohort study of 1.5 million transfused patients. *J Intern Med*. 2018;284(1):92–103.
196. Atsma F, Vegt F de. The healthy donor effect: a matter of selection bias and confounding. *Transfusion*. 2011;51(9):1883–5.
197. Barber LE, Gerke T, Markt SC, Peisch SF, Wilson KM, Ahearn TU, et al. Family history of breast or prostate cancer and prostate cancer risk. *Clin Cancer Res*. 2018;24(23):clincanres.0370.2018.

198. Zhen JT, Syed J, Nguyen KA, Leapman MS, Agarwal N, Brierley K, et al. Genetic testing for hereditary prostate cancer: Current status and limitations. *Cancer*. 2018;124(15):3105–17.
199. Frank C, Fallah M, Sundquist J, Hemminki A, Hemminki K. Population Landscape of Familial Cancer. *Sci Rep-uk*. 2015;5(1):12891.
200. Pilarski R. The Role of BRCA Testing in Hereditary Pancreatic and Prostate Cancer Families. *Am Soc Clin Oncol Educ Book*. 2019;39(39):79–86.
201. Delgado-Rodríguez M, Llorca J. Bias. *J Epidemiol Commun H*. 2004;58(8):635.
202. Gerhard T. Bias: Considerations for research practice. *Am J Health-syst Ph*. 2008;65(22):2159–68.
203. Kopec JA, Esdaile JM. Bias in case-control studies. A review. *J Epidemiol Commun H*. 1990;44(3):179.
204. Kamper-Jørgensen M, Ahlgren M, Rostgaard K, Melbye M, Edgren G, Nyrén O, et al. Survival after blood transfusion. *Transfusion*. 2008;48(12):2577–84.
205. Edgren G, Rostgaard K, Hjalgrim H. Methodological challenges in observational transfusion research: lessons learned from the Scandinavian Donations and Transfusions (SCANDAT) database. *Isbt Sci Series*. 2017;12(1):191–5.
206. Westhoff CM. Blood group genotyping. *Blood*. 2019;blood-2018-11-833954.
207. Linder GE, Chou ST. Red cell transfusion and alloimmunization in sickle cell disease. *Haematologica*. 2021;106(7):1805–15.
208. Kyriacou DN, Lewis RJ. Confounding by Indication in Clinical Research. *Jama*. 2016;316(17):1818–9.