Aus dem Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie
Institut der Universität München

Direktor: Prof. Dr. rer. nat. Ulrich Mansmann

Analysis of adverse events with modern statistical methods

Dissertation
zum Erwerb des Doktorgrades der Humanbiologie
an der Medizinischen Fakultät der
Ludwig-Maximilians-Universität zu München

vorgelegt von

Kirsi Marjaana Manz

aus

Parikkala, Finnland

Jahr

2023

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	PD Dr. rer. biol. hum. Markus Pfirrmann, M.Sc.
Mitberichterstatter:	PD Dr. Gerald Bastian Schulz Prof. Dr. Sven Schmiedl
Mitbetreuung durch den promovierten Mitarbeiter:	
Dekan:	Prof. Dr. med. Thomas Gudermann
Tag der mündlichen Prüfung:	26.09.2023

„Ein Arzneimittel, von dem behauptet wird, daß es keine Nebenwirkungen habe, steht im dringenden Verdacht, auch keine Hauptwirkung zu besitzen.“

– Gustav Kuschinsky

Contents

Zusammenfassung

Diese Dissertation beschäftigt sich mit zeitgemäßer Analyse unerwünschter Ereignisse in klinischen Studien. Unerwünschte Ereignisse sind ungünstige Vorkommnisse, wie zum Beispiel ein abnormaler Laborwert, ein Symptom oder eine Erkrankung, die während der Einnahme von Medikation auftreten. Dieser Zusammenhang ist koinzident und kann, muss aber nicht, im direkten Bezug zu der Einnahme der Medikation stehen. Nach dieser Definition zählen alle ungünstigen Vorkommnisse während einer Behandlung, zum Beispiel im Rahmen einer klinischen Studie, als unerwünschtes Ereignis, wie zum Beispiel auch ein Verkehrsunfall. Daher beinhalten die Sicherheitsdaten einer klinischen Prüfung mehr Ereignisse, als man im Allgemeinen erwarten könnte.

Die Wirksamkeit in klinischen Studien wird heutzutage mit Methoden analysiert, die die unterschiedlichen Beobachtungsdauern der Patienten berücksichtigen. Dazu zählen die Kaplan-Meier Überlebenskurven oder auch das Cox Model. Allerdings ist es leider immer noch sehr verbreitet, dass unerwünschte Ereignisse als einfache Prozentangaben angegeben werden, obwohl diese Vorgehensweise mit gewissen Problemen verknüpft ist. Auf diese Problematik wurde schon in den 1980er Jahren hingewiesen. In klinischen Studien werden die Probanden meistens nicht an einem festen Kalendertag in die Studie eingeschlossen, sondern gestaffelt. Ebenso ist die individuelle Beobachtungsdauer der Probanden unterschiedlich: manche verlassen die Studie vorzeitig aufgrund aufgetretener Ereignisse, manche verbleiben in der Studie bis zum Ende und werden zensiert, zu manchen Probanden verliert sich der Kontakt und es ist unklar, was mit ihnen geschehen ist. Diese unterschiedlichen Beobachtungsdauern müssen auch in der Analyse von unerwünschten Ereignissen berücksichtigt werden, um unverfälschte Ergebnisse zu erhalten.

Das Ziel der vorliegenden Arbeit ist zu untersuchen, wie die Analyse unerwünschter Ereignisse verbessert werden könnte. Es existieren Methoden, die die Zeit unter Risiko für ein Ereignis berücksichtigen. Auch konkurrierende Ereignisse bis zum Auftreten eines unerwünschten Ereignisses sollen berücksichtigt werden. Als ein Beispiel für ein konkurrierendes Ereignis gilt der Tod vor Auftreten eines jeglichen unerwünschten Ereignisses: Nach dem Tod kann kein unerwünschtes Ereignis mehr beobachtet werden.

Die vorliegende Arbeit analysiert Daten aus einer großen hämato-onkologischen Studie an Patienten mit chronischer myeloischen Leukämie. Verschiedene Methoden zur Analyse unerwünschter Ereignisse werden vorgestellt und angewandt. Als erstes wird die „Standard“-Analyse für das Auftreten des ersten unerwünschten Ereignisses ohne Berücksichtigung der Beobachtungsdauern durchgeführt. Dazu werden die Anteile der Patienten mit unerwünschten Ereignissen (Inzidenzproportionen) in zwei Therapiegruppen ins Verhältnis miteinander gesetzt (relatives Risiko). Berücksichtigt man die Zeit unter Risiko für ein Ereignis, gelangt man zu Inzidenzraten und deren Verhältnis (*incidence rate ratio*).

Das Konzept von konkurrierenden Ereignissen wird anhand der Daten der in dieser Arbeit analysierten Studie erläutert. Die korrekte Methode um Wahrscheinlichkeiten für unerwünschte Ereignisse zu berechnen ist die kumulative Inzidenzfunktion. Dazu bietet sich der Aalen-Johansen Schätzer an, der als Goldstandard gilt, da er das Auftreten von konkurrierenden Ereignissen und unterschiedlichen Beobachtungsdauern mitberücksichtigt. Zwei andere Schätzer der kumulativen Inzidenzfunktion werden mit dem Goldstandard verglichen.

Die sogenannten Multi-State Modelle erlauben das gleichzeitige Schätzen der Hazards und Wahrscheinlichkeiten zum Auftreten von unerwünschten Ereignissen oder konkurrierenden Ereignissen. Diese Modelle haben den zusätzlichen Vorteil, dass der Übergang von unerwünschtem Ereignis zum konkurrierenden Ereignis, wie zum Beispiel Krankheitsprogress, untersucht werden kann. Multi-State Modelle werden kurz eingeführt und exemplarisch für verschiedene Schweregrade von unerwünschten Ereignissen geschätzt.

Danach werden Methoden vorgestellt und angewandt, die für die Analyse wiederkehrender Ereignisse geeignet sind: Das Andersen und Gill (AG) Modell, das Prentice, Williams und Peterson (PWP) Modell und das Wei, Lin und Weissfeld (WLW) Modell. Wiederkehrende Ereignisse sind Ereignisse, die nach einer ereignisfreien Periode wieder auftreten. Es ist möglich, einen Gesamttherapieeffekt oder einzelne Effekte für jedes Wiederauftreten zu schätzen. Die mittlere kumulative Funktion (*mean cumulative function*) bietet eine Möglichkeit, wiederkehrende unerwünschte Ereignisse graphisch darzustellen. Diese Methode hat ähnlich dem Aalen-Johansen Schätzer den Vorteil, dass unterschiedliche Beobachtungsdauern und konkurrierende Ereignisse mitberücksichtigt werden. Der zusätzliche Vorteil ist, dass auch wiederkehrende Ereignisse analysiert werden können.

Eine zeitgemäße Analyse von unerwünschten Ereignissen sollte die individuelle Beobachtungsdauer der Patienten mitberücksichtigen. Der einfachste Weg, dieses zu erreichen, ist Inzidenzraten auszurechnen. Für diese wird für jeden Patienten die Zeit unter Risiko für ein Ereignis ermittelt. Das Verhältnis der Inzidenzraten in zwei Gruppen kann zu Hilfe gezogen werden, um Unterschiede in Inzidenzen pro Therapiegruppe zu finden. Eine geeignetere Methode, Wahrscheinlichkeiten für das Auftreten von unerwünschten Ereignissen zu ermitteln, stellt die kumulative Inzidenzfunktion mit dem Aalen-Johansen Schätzer dar. Diese Methode ist der Goldstandard, da konkurrierende Ereignisse und individuelle Beobachtungsdauern mitberücksichtigt werden. Für wiederkehrende Ereignisse wird das Darstellen der mittleren kumulativen Funktionen empfohlen. Modelle für wiederkehrende Ereignisse können zusätzliche Einsichten in die Sicherheitsdaten gewähren.

List of abbreviations

AE	Adverse event
AG	Andersen-Gill (model)
ALAT	Alanine aminotransferase
AP	Accelerated phase
BC	Blast crisis
CI	Confidence interval
CIF	Cumulative incidence function
CML	Chronic myeloid leukemia
CR	Competing risk
CTCAE	Common Terminology Criteria for Adverse Events
EAIR	Exposure-adjusted incidence rate
ELN	European LeukemiaNet
eCRF	Electronic case report form
HLGT	High level group term
HLT	High level term
HR	Hazard ratio
IFN-α	Interferon alpha
IP	Incidence proportion
IR	Incidence rate
IRR	Incidence rate ratio
ITT	Intention to treat
KM	Kaplan-Meier
LLN	Lower limit of normal
MCF	Mean cumulative function
MedDRA	Medical Dictionary for Regulatory Activities
MMR	Major molecular response
MR4	Deep molecular response
Ph	Philadelphia chromosome
PT	Preferred term
PWP	Prentice, Williams and Peterson (model)
RR	Relative risk
SAE	Serious adverse event
SCT	Stem cell transplantation
SOC	System organ class
TKI	Tyrosine kinase inhibitor
WLW	Wei, Lin and Weissfeld (model)

1. Introduction

1.1 Motivation

Safety of patients in clinical trials is of utmost importance. Also in drug approval it is essential to show that the benefits of the drug outweigh its risks and that the drug meets the required standards for efficacy and safety. Nowadays it is a common practice to comprehensively evaluate drug efficacy using sophisticated methods, but limited progress has been made in reporting safety, although this problem was highlighted as early as in the 1980s [1]. In recent decades, survival methods that take into account patient observation time in clinical trials have found their place in efficacy analysis, but unfortunately safety reporting still relies largely on reporting crude percentages of patients experiencing adverse events. This may be intuitive and easy for everyone to understand, but it is prone to bias and not sufficient in the most cases. Another drawback is that clinical trials are usually conducted to demonstrate efficacy. At least theoretically, it is straightforward to evaluate the efficacy of the investigational drug, because the trial is explicitly designed for this purpose: Already before the study starts, the design of the trial is pre-specified in a manner that one or several null hypotheses of (usually) no difference in the efficacy between the investigational drug and control (*i.e.* standard care, placebo, …) can be rejected based on the data created during the trial. Therefore, most trials are powered for efficacy rather than for safety related outcomes. In addition, the trials usually are relatively short and it is questionable whether the entire safety profile of the drug can be observed. Given these limitations, more attention should be paid to safety analysis to better understand and characterize safety issues related to the drugs under investigation.

Safety in clinical trials is mostly related to adverse events. An adverse event is any unfavorable and unintended sign, including an abnormal laboratory finding, symptom, or disease, temporally associated with the use of a medicinal product, whether or not considered related to the medicinal product [2]. To distinguish pre-existing conditions from adverse events that occur during the use of a medicinal product, the term “treatment emergent” may be added for clarification [3].

Two problematic scenarios are briefly introduced below: In oncological trials, study medication may be given until disease progression, at which time the medication is changed. This leads to a dependence of the follow-up times for adverse events on the time to disease progression. This in turn means that the treatment group with longer progression-free survival has a higher likelihood of observing adverse events [4]. The group that remains progression-free for longer appears to have a higher incidence of adverse events when the different follow-up times between the two groups are not taken into account. Another example concerns a scenario in which the drug under investigation causes fatal events. When the number of adverse events is pooled, it is not possible to say whether the low number of adverse events is due to a better adverse event profile or to higher mortality from the drug. Death in this example is called a competing risk for the occurrence of an adverse event. Competing events are events that prevent observation future of adverse events: If a patient dies, adverse events cannot be observed anymore. Therefore, merely counting adverse events in such situations is misleading and introduces bias.

There are methods correctly accounting for the time until an event occurs and methods accounting for competing events, which are suitable for the analysis of safety data. These methods are presented and examined in this work. Initially, only the first adverse event is considered. Then, the analyses are extended to recurrent events, *i.e.*, adverse events that recur after an event-free period. For this work, adverse event data from a large hemato-oncological study in patients with chronic myeloid leukemia are used. The focus of this work is on the methods rather than the medical interpretation of the results. The aim of this work is to investigate how the reporting of adverse events in clinical trials could be improved.

1.2 Chronic myeloid leukemia (CML)

Chronic myeloid leukemia (CML) is a form of leukemia in which the bone marrow produces too many immature (myeloid) cells that cannot work properly. CML can be characterized by the Philadelphia chromosome found in the blood of approximately 95% of CML patients (Philadelphia positive/Ph+ CML) [5]. This abnormal Philadelphia chromosome creates a new gene called BCR-ABL1 by fusion of genes from chromosome 9 and chromosome 22. The BCR-ABL1 gene causes too much of a protein called tyrosine kinase to be produced, which in turn promotes cancer cell growth. Tyrosine kinase causes too many white blood cells to grow. These diseased, immature white blood cells do not grow like normal cells and are produced in huge numbers. At the same time, not enough healthy white blood cells are produced. Main symptoms of CML at diagnosis are fatigue, anemia, splenomegaly (enlarged spleen), abdominal discomfort and infections [6]. Some patients are also asymptomatic.

1.2.1 Stages

CML can be divided into three different phases: chronic phase, accelerated phase and blast crisis. Whereas the chronic phase is the (more or less) stable phase of the disease, the latter two phases are considered progression of the disease and are associated with poorer outcome.

Chronic phase:

Most newly diagnosed patients are diagnosed in the chronic phase of the disease. This phase is characterized by a low count of blast cells in the blood and/or bone marrow and accompanied by no or only mild symptoms. The chronic phase can be sufficiently treated with the so-called tyrosine kinase inhibitors (TKIs) in most cases.

Accelerated phase:

The accelerated phase is one of the two progressive phases. It is mainly defined by increasing blast counts in the blood/bone marrow. In the clinical study investigated in this work the accelerated phase is defined by [7]

- ≥ 15% blasts in the peripheral blood or bone marrow, but <30% blasts in both the peripheral blood and bone marrow,
- ≥ 30% blasts plus promyelocytes in peripheral blood or bone marrow,
- ≥ 20% basophils in the peripheral blood, or
- thrombocytopenia ($<100\times10^9$/L) that is unrelated to therapy.

Blast crisis:

The second progressive and most severe phase of CML is the blast crisis. The blast crisis is associated with health deterioration including infections, thrombosis and anemia due to the bone marrow failure (high production of immature blast cells) [8]. In the present study blast crisis is defined as ≥ 30% blasts in peripheral blood or bone marrow or appearance of extramedullary biopsy-proven involvement other than hepatosplenomegaly (enlarged liver or spleen) [7].

1.2.2 Treatment

Nowadays, CML can effectively be treated with TKIs. Currently there are four first-line TKIs (imatinib, dasatinib, nilotinib and bosutinib) available in several countries for treatment of newly diagnosed CML patients [9]. TKIs inhibit the BCR-ABL1's kinase activity and reduce the frequency of progression to blast crisis and eliminate the symptoms of chronic phase [10]. In general, survival of treated CML patients is very good. For example, Hochhaus *et al.* found an overall survival rate of 83% after 10 years of imatinib treatment [11] and the CML IV study reported a 10-year overall survival of 82% [12]. After progression to accelerated phase or blast crisis, stem cell transplantation is performed in many patients.

1.2.3 Response to treatment

During TKI therapy, the response to treatment should be closely monitored. In most cases, a peripheral blood sample is sufficient for this purpose and an invasive bone marrow biopsy is no longer required. Response is quantified by determining the amount of BCR-ABL1 in the blood. The principle of "the less, the better" applies. For this, real-time quantitative polymerase chain reaction (RT-qPCR) technique is used to amplify the BCR-ABL1 gene along with a reference gene (BCR, ABL1 or GUSB), with the reference gene serving as a control for both the quantity and the quality of the sample [13]. For both BCR-ABL1 and the reference gene, the number of amplified transcripts is determined. The ratio between the detected BCR-ABL1 genes and the number of detected reference genes is then calculated. If no BCR-ABL1 transcripts were detected, the number of reference gene transcripts provides an indication of the sensitivity with which residual disease can be excluded in that sample [14]. To obtain comparable results between different laboratories, the BCR-ABL1/control gene ratios are normalized to the so-called international scale (IS). The international scale originates from the IRIS CML study and represents a standardized baseline from that study [15]. The IRIS study was the first to use a common baseline value to make results from different laboratories comparable, and this scale is still used today to normalize molecular response levels. The BCR-ABL (IS) value is the percentage of BCR-ABL1/control gene converted to the IS scale. A value of 100% is the baseline value from the IRIS study to which all values are compared to it. Different "depths" of response can be defined [7,16]:

- MMR: Major molecular response (MMR) is defined as ≤ 0.1% BCR-ABL (IS) or equivalently as ≥ 3 log reduction in BCR-ABL transcripts compared to the standardized baseline
- MR4: <0.01% BCR-ABL (IS), corresponding to a 4-log reduction from IRIS baseline or undetectable BCR-ABL1 transcripts with at least 10 000 detected ABL1 control gene transcripts (or 24 000 GUSB transcripts)
- MR4.5: <0.0032% BCR-ABL (IS), corresponding to a ≥ 4.5-log reduction from IRIS baseline or undetectable BCR-ABL1 transcripts with at least 32 000 detected ABL1 reference transcripts (or 77 000 GUSB transcripts).

The BCR-ABL (IS) value is the key parameter for evaluating response to treatment and is closely monitored. The latest recommendations of the European LeukemiaNet (ELN) include monitoring BCR-ABL (IS) values at least every three months and determine a scheme for optimal response, "warning" and treatment failure [9]. Optimal response means that BCR-ABL (IS) values decrease over time in a predetermined manner such that current treatment can usually be safely continued. Failure occurs when BCR-ABL (IS) values remain high. In these cases, the current treatment should be changed. If the BCR-ABL (IS) values are too high for optimal response and too low for failure, the so-called warning level is reached. According to Hochhaus *et al.*, in these cases, “careful consideration should be given to treatment continuation or change, depending on the patients’ characteristics, comorbidities and tolerability” [9].

1.2.4 Prognosis

Thirty years ago, CML was a fatal disease with a median survival of 4-6 years [14], meaning that half of CML patients die within 4-6 years of diagnosis. After the introduction of imatinib as the first TKI, the prognosis of CML has improved significantly. Nowadays, the relative survival of CML patients is comparable to that of the general population [14].

There are four established prognostic scores for CML patients: Sokal, Euro, EUTOS and the ELTS score [17-20]. Prognostic factors included in the scores are age and spleen size, as well as the following laboratory parameters: Platelet count, blasts, eosinophils, and basophils from a peripheral blood sample. Bone marrow puncture is not required to calculate the scores. The scores are calculated at the time of diagnosis, and based on their values, patients can be divided into low, intermediate, and high-risk groups. The Sokal and Euro scores were developed to predict overall survival. The EUTOS score was developed to predict the probability of achieving complete cytogenetic response (CCyR) at 18 months. The CCyR was considered the best early surrogate for overall survival available at that time. The latest score, the ELTS (EUTOS long time survival) score, predicts the probabilities of dying from CML, because with the great success of TKI treatment, CML patients are living longer and the chances of dying for reasons unrelated to CML continue to increase. The ELN 2020 recommendations suggest using the ELTS prognostic score over other previously published prognostic scores [9].

1.2.5 Treatment-free remission

A relatively new concept in the treatment of CML is the possibility of discontinuing any CML-specific treatment for patients who have responded to TKI treatment and maintained this response over an extended period of time. To date, there are no established criteria for treatment discontinuation, so the safest option for CML patients is to discontinue CML treatment as part of clinical trials where patients are closely monitored. Currently, there are few clinical trials evaluating the safety and feasibility of treatment discontinuation. The ongoing European wide discontinuation trial EURO-SKI (European Stop Tyrosine Kinase Inhibitor Study) aims to define the precise conditions for safe TKI discontinuation. The results of the interim analysis suggest that longer treatment duration and longer duration of deep molecular response are favorable for maintaining molecular response, but the final results are not yet available [21]. A part of the TIGER trial investigated in this work also includes a treatment-free phase for those patients who have maintained a deep molecular response for a pre-specified period of time.

1.3 Aims

The aim of this work is to investigate how the reporting of adverse events in clinical trials could be improved. For this purpose, data from a large hemato-oncological study in CML patients, which included more than 7000 adverse events, are used. Based on the results, suggestions are made for an analysis that is as simple as possible, but also methodically as appropriate as needed, leading to comprehensive, state-of-the-art reporting of safety data.

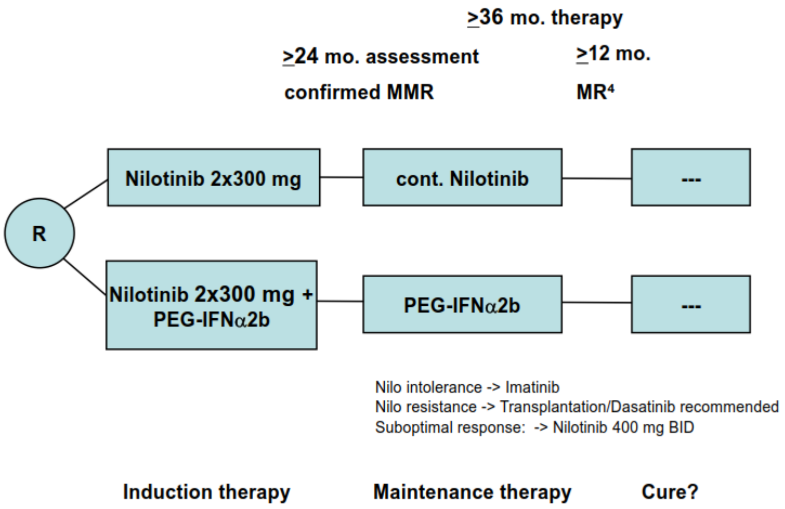
2. Data

2.1 CLM V Study –the TIGER study

The TIGER study is a hemato-oncological open-label, two-arm, randomized, phase III trial designed to optimize treatment in newly diagnosed Philadelphia chromosome positive (Ph+) or Ph negative/BCR-ABL positive patients with CML in chronic phase. The trial compares nilotinib 300 mg twice daily to nilotinib 300 mg twice daily plus Peg-IFNα2b (pegylated interferon α, IFN-α) [7].

The first of the two main objectives of the study is to evaluate the major molecular response (MMR) rate at 18 months of nilotinib monotherapy and to compare the results with nilotinib plus IFN-α therapy. The second main objective is to evaluate the feasibility of discontinuing drug therapy in the setting of stable deep molecular response (MR4) after nilotinib vs. IFN-α maintenance therapy. Secondary objectives of the study are to evaluate the efficacy and tolerability of IFN-α added to nilotinib 2x300 mg/day and to evaluate the efficacy and tolerability of a maintenance therapy with nilotinib vs. IFN-α after stable MMR and after at least 24 months of nilotinib therapy [7].

The two therapies are administered as shown in Figure 1: After randomization, nilotinib treatment is started in both therapy groups. IFN-α will be added in the combination arm after recovery of the blood counts (complete hematologic response), but not earlier than 6 weeks after initiation of nilotinib-therapy. After confirmed MMR at least 24 months after start of nilotinib therapy, maintenance therapy will be administered (nilotinib monotherapy is continued while nilotinib is discontinued in the combination arm). After a continuous deep molecular response (MR4) for at least 12 months (but after a total of at least 3 years of treatment), all therapies may be discontinued and patients continue to be followed and monitored. If a molecular relapse (loss of MMR) occurs, patients will be treated again with nilotinib 2x300 mg/day. Patients will continue to be followed up until termination of the trial, withdrawal of consent, or death.


≥36 mo. therapy
≥24 mo. assessment
confirmed MMR
≥12 mo.
MR⁴
R
Nilotinib 2x300 mg
cont. Nilotinib

Nilotinib 2x300 mg + PEG-IFNα2b
PEG-IFNα2b

Nilo intolerance -> Imatinib
Nilo resistance -> Transplantation/Dasatinib recommended
Suboptimal response: -> Nilotinib 400 mg BID
Induction therapy
Maintenance therapy
Cure?

Figure 1: Study design of the TIGER study. After an induction therapy with either nilotinib or nilotinib with pegylated interferon α (PEG-IFNα2b) and sufficient response to treatment (MMR=major molecular response), the maintenance therapy with nilotinib or interferon α only starts. After reaching a deep molecular response (MR4) for at least 12 months the treatment is discontinued. Source: [7].

With reference to the study protocol, the efficacy analysis will be performed in accordance with randomization to either nilotinib or nilotinib+IFN-α therapy (intention-to-treat analysis). *For the purpose of this work, which is to show and discuss the different methods suitable for analysis of adverse events, rather than medical interpretation, the treatment arms of the study will be referred to as A and B.* For the safety analysis, all patients who received at least one dose of either study drug will be included.

2.1.1 Patients

As of December 14th, 2018, a total of 750 patients had been entered into the database. Of these, 692 patients were randomized to either treatment A (n=353) or B (n=339), 27 patients were screening failures, 25 patients were from the pilot phase of the study, and for 6 patients the patient number was created in error. Of the 692 randomized patients, 353 received treatment A and 336 patients received treatment B (total n=689); no drug intake was reported for 3 patients. A flow chart of patient selection can be seen in Figure 2. Analyses included 689 patients, 353 in treatment arm A and 336 in treatment arm B.

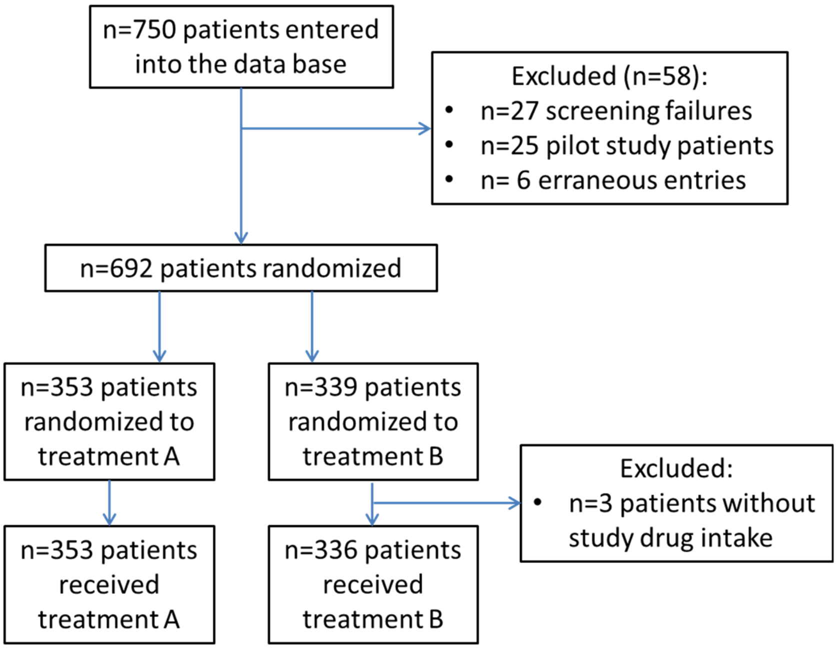

n=750 patients entered into the data base
Excluded (n=58):
• n=27 screening failures
• n=25 pilot study patients
• n= 6 erraneous entries
n=692 patients randomized
n=353 patients randomized to treatment A
n=339 patients randomized to treatment B
Excluded:
• n=3 patients without study drug intake
n=353 patients received treatment A
n=336 patients received treatment B

Figure 2: Flowchart of the patients in the TIGER study.

2.1.2 Adverse events reporting in the TIGER study

2.1.2.1 Definition of an adverse event

For the TIGER study, an adverse event (AE) was defined as “the appearance of (or worsening of any pre-existing) undesirable sign(s), symptom(s), or medical condition(s) occurring after taking the first dose of study drug even if the event is not considered to be related to the study drug(s)” [7]. Unlike an adverse drug reaction, no association with the study drug is required. Besides worsening of pre-existing conditions or adverse events resulting from concomitant medications, other causes such as accidental injuries should also be reported as adverse events as they occur during the study.

*Serious adverse event SA*E

A Serious Adverse Event (SAE) is defined [7] as an adverse event that

- is fatal or life-threatening
- results in persistent or significant disability/incapacity
- constitutes a congenital anomaly/birth defect
- constitutes a new cancer diagnosis
- constitutes overdosing of the drugs investigated
- requires inpatient hospitalization or prolongation of existing hospitalization
- is medically significant, *i.e.*, defined as an event that jeopardizes the patient or may require medical or surgical intervention to prevent one of the outcomes listed above.

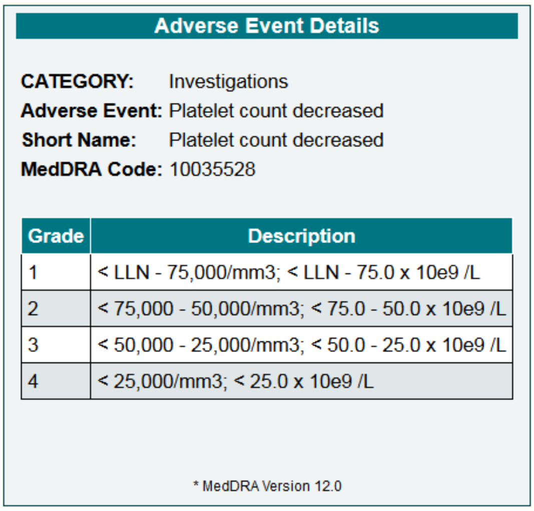
Extensive reporting was required for SAEs, including rapid reporting to the trial office within 24 hours of becoming aware of the occurrence using detailed SAE reporting forms. Information recorded on these forms was to be entered into the electronic case report forms (eCRF) from which the data for this work were extracted.

2.1.2.2 Standardized coding of the adverse events

CTCAE

All adverse events were assessed using the Common Terminology Criteria for Adverse Events (CTCAE) version 4.0. for which a comprehensive file can be downloaded from https://ctep.cancer.gov/protocolDevelopment/electronic_applications/docs/CTCAE_4.03.xlsx (accessed on February 20th, 2023). The CTCAE criteria include five different severity grades that can be assigned to adverse events. The grades are defined as follows: 1 –mild, 2 –moderate, 3 –severe, 4 –life-threatening, and 5 –death related to AE. For an example of the CTCAE criteria for thrombopenia (platelet count decreased), see Figure 3. The corresponding MedDRA Code and the System Organ Class (SOC, called CATEGORY in the figure) for this AE can be seen along with the platelet count cutoff values that define each severity grade. Note that grade 5 is missing here because death due to thrombopenia is not considered applicable.

CTCAE coding is compatible with MedDRA starting with Version 4.0 (Version 4.0 is compatible with MedDRA Version 12.0). The MedDRA coding system is introduced below.



Adverse Event Details

CATEGORY: Investigations
Adverse Event: Platelet count decreased
Short Name: Platelet count decreased
MedDRA Code: 10035528

Grade	Description
1	< LLN - 75,000/mm3; < LLN - 75.0 x 10e9 /L
2	< 75,000 - 50,000/mm3; < 75.0 - 50.0 x 10e9 /L
3	< 50,000 - 25,000/mm3; < 50.0 - 25.0 x 10e9 /L
4	< 25,000/mm3; < 25.0 x 10e9 /L

* MedDRA Version 12.0

Figure 3: Example of coding „platelet count decreased” using the Common Terminology Criteria for Adverse Events (CTCAE) version 4.0. Source: Screenshot of the interactive web-based CTCAE application (https://safetyprofiler-ctep.nci.nih.gov/CTC/CTC.aspx). MedDRA=Medical Dictionary for Regulatory Activities, LLN=Lower Limit of Normal.

MedDRA

MedDRA stands for Medical Dictionary for Regulatory Activities and provides a standardized coding system for adverse events (www.meddra.org). A unique, 8-digit code can be assigned for each adverse event. The MedDRA system has a hierarchy so that reported codes can be grouped into medically meaningful groups for adverse event analysis/reporting. The MedDRA coding system is also used in the TIGER study.

Table 1 shows the hierarchy of MedDRA Codes for the adverse event “platelet count decreased”. It is possible to group adverse events at each of the reported levels, depending on the intended purpose of the reporting. For example, many reported abnormal laboratory results fall into the system organ class “investigations”, “headache” would be assigned to “nervous system disorders”, and “hypertension” would be assigned to “vascular disorders”.

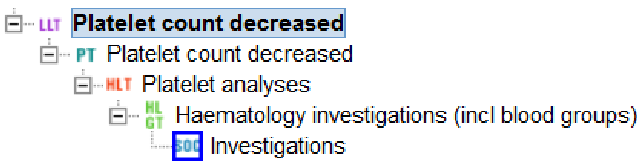
<table>
<tr><th colspan="3">MedDRA PT Code 10035528</th></tr>
<tr><td>Lowest Level Term</td><td>LLT</td><td rowspan="5"> LLT Platelet count decreased
PT Platelet count decreased
HLT Platelet analyses
HLGT Haematology investigations (incl blood groups)
SOC Investigations</td></tr>
<tr><td>Preferred Term</td><td>PT</td></tr>
<tr><td>High Level Term</td><td>HLT</td></tr>
<tr><td>High Level Group Term</td><td>HLGT</td></tr>
<tr><td>System Organ Class</td><td>SOC</td></tr>
</table>

Table 1: MedDRA coding and its hierarchical structure for the adverse event “platelet count decreased”. MedDRA= Medical Dictionary for Regulatory Activities, PT=Preferred Term, LLT=Lowest Level Term, HLT=High Level Term, HLGT=High Level Group Term, SOC=System Organ Class. The graphic on the right is a screenshot from the MedDRA Browser, which can be accessed via https://tools.meddra.org/wbb/ (MedDRA User ID and password required).

In the TIGER study, the Preferred Term (PT) should be reported according to MedDRA Version 12.0. For the recurrent event analysis, all adverse events coded with the same PT will be grouped together and considered as the same type of adverse event.

2.1.2.3 Data reported

The following data were entered for each adverse event that occurred during the study (the asterisk denotes questions which were mandatory to fill out):

- reporting date of the AE*
- description of the AE*
- MedDRA code* (preferred term, PT) and severity grade* (CTCAE grading)
- start date*, whether AE is ongoing (yes/no), end date
- relation to study drugs: not applicable/not known, not done, no relation, unlikely, possible, definite, cannot be assessed
- action taken regarding study drugs: not applicable/not known, not done, no change in study medication, study medication temporarily withheld, study medication permanently discontinued
- medication required for the AE (yes/no) and if yes, type of medication
- whether the AE was an SAE* (yes/no)

In the case that the adverse event was reported as serious (*i.e.* SAE), additional questions were asked with binary answer categories (yes/no): Was the SAE fatal, life-threatening, leading to disability, leading to congenital anomaly, required hospitalization, and was the SAE medically significant.

2.1.2.4 Data available for analysis

As of December 14th, 2018, which was chosen as the cut-off date for this work, a total of 7864 adverse events of any grade had been reported. An overview of the adverse events data is shown in Figure 4.

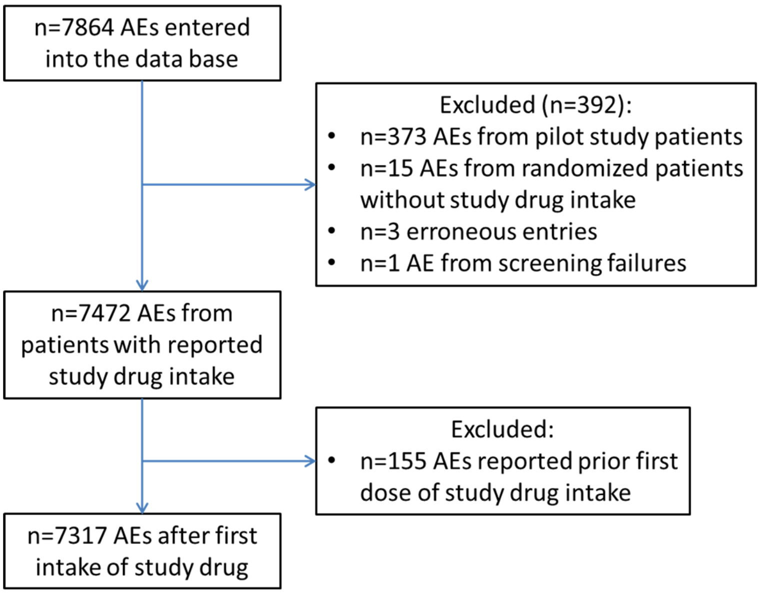

n=7864 AEs entered into the data base
Excluded (n=392):
• n=373 AEs from pilot study patients
• n=15 AEs from randomized patients without study drug intake
• n=3 erroneous entries
• n=1 AE from screening failures
n=7472 AEs from patients with reported study drug intake
Excluded:
• n=155 AEs reported prior first dose of study drug intake
n=7317 AEs after first intake of study drug

Figure 4: Flowchart of the adverse events (AE) reported in the TIGER study.

Of these 7864 AEs, 392 were excluded (n=373 from pilot study patients, n=15 from patients who were randomized but did not receive any study drug, n=3 erroneous entries, and n=1 AE from a patient reported as a screening failure). The remaining 7472 AEs were reported among the 689 patients who had received one or more doses of the study drug. For combination treatment arm B, all patients who had received at least one dose of either drug were included. To assess the impact of drug therapy on adverse events, all adverse events reported before the first dose of study drug were also excluded, too (n=155). Thus, a total of 7317 AEs reported after the first intake of study drug were available for analysis (see Figure 4).

To be included in the analyses, the reported adverse events had to have a valid start date. A reporting date (*i.e.* the date the AE was entered) and a start and end date of the AE are entered in the database, as mentioned earlier. Because the reporting date coincides with the start date of the AE in most cases, the data entry date was used for any missing start date. If this resulted in a negative duration of the AE, the start date was set to one day before the end date of the AE. In these few cases, the AE had an artificial duration of one day, but this had the advantage of allowing the AE to be used for analysis. By preparing the data in this way, all adverse events had a start date and could be included in the time-to-event analysis. In summary, the safety data set contains 7317 adverse events in 689 patients, of whom 353 patients received drug A and 336 patients received drug B.

Please note that although data from the TIGER study is used, the focus of this thesis is on the methods used to analyze adverse event data and not on the medical interpretation of the results.

3. Methods

In this chapter, methods for safety data analysis will be introduced. Chapter 3.1 deals with methods for analyzing the first adverse event. First, simple percentage-based methods such as incidence proportion and incidence ratio are introduced (Chapters 3.1.1 to 3.1.3). These methods are the most commonly encountered methods in the literature to date. Problems with the application of these methods arise as soon as individual observation times differ. Chapters 3.1.4 to 3.1.10 present time-to-event methods that are more suitable for analyzing adverse event data. These methods include the Kaplan-Meier survival curve, the Cox proportional hazard model, and the gold standard for estimating adverse event probabilities, the Aalen-Johansen estimator of the cumulative incidence function. Multi-state models are also briefly introduced. The second part of this Chapter (Chapter 3.2) goes beyond the first adverse event and presents methods for recurrent events analysis. Repeated events analysis allows for more efficient use of data and, depending on the model, all available data can be used. In Chapters 3.2.1 to 3.2.4 four different methods for repeated adverse events analysis are introduced. The advantages and disadvantages of the models are discussed throughout the chapter. The chapter closes with a short remark on the software used to analyze the data.

3.1 Methods for analysis of first adverse events

3.1.1 Incidence proportion and relative risk

Incidence proportion (IP) is the percentage of patients with at least one adverse event in a given time interval

$$\mathrm{IP}_i = \frac{\#\mathrm{AE}_i}{\mathrm{n}_i}$$

with $\#AE_i$ as the number of patients in treatment group *i* with one or more adverse events and n_i as the total number of patients in the treatment group *i* [22]. Thus, this proportion represents the probability of experiencing an adverse event. The incidence proportion is a valid estimator of the adverse event probability in so-called complete data sets. These are data in which patients are followed up for exactly the same time (no censoring occurs) and it is known for each patient whether or not an adverse event occurred during the follow-up period. In the presence of censoring, the incidence proportion is known to underestimate the probability of an adverse event [23].

The confidence interval for the incidence proportion was calculated using the Wilson (score) method with continuity correction (method 4 in Ref. [24]), which is the default method for calculating the confidence interval for a single proportion in the R function "prop.test" of the package "stats" used in this work. The lower boundary (L) of the confidence interval is calculated as

$$L = \frac{2np + z^2 - 1 - z * \sqrt{z^2 - 2 - 1/n + 4p(n(1-p)+1)}}{2(n+z^2)},$$

and the upper boundary (U) as

$$U = \frac{2np\ +\ z^2 + 1 + z * \sqrt{z^2 + 2 - 1/n + 4p(n(1-p) + 1)}}{2(n + z^2)},$$

respectively. Here, *n* is the sample size, *p* the incidence proportion, and *z* the *1-α/2* quantile of the standard normal distribution.

The incidence proportions of two treatments (*e.g. i*=1,2) can be compared using the relative risk (RR), which is the ratio of the two incidence proportions:

$$\mathrm{RR} = \frac{\mathrm{IP}_1}{\mathrm{IP}_2}.$$

A higher incidence proportion of adverse events in group 1 compared to group 2 leads to a relative risk of >1 and vice versa. By calculating the confidence interval around the point estimate of relative risk, a significant effect can be found if the confidence interval does not include the value 1. With regard to the bias involved in calculating relative risk for groups with different observational times, it is not easy to say whether relative risk overestimates or underestimates true risk: Both the numerator and denominator underestimate true risk, but it is not possible to prognosticate how the ratio of the two risks will behave.

The confidence interval for the relative risk was calculated using the Wald method (method 1 in Ref. [24]), which is the default method for calculating the confidence interval for a risk ratio in the R function "riskratio " of the package "epitools" used in this work. The confidence interval is calculated as

$$p\ \pm\ z * \sqrt{\frac{p(1-p)}{n}} \approx 1.96 * \sqrt{\frac{p(1-p)}{n}}$$

where *p* is the proportion (or, in this case, the relative risk), *z* the *1-α/2* quantile of standard normal distribution and *n* the sample size. With the commonly used *α=0.05* and thus a 95% confidence interval, *z* is approximately 1.96.

3.1.2 Incidence rates and incidence rate ratios

The incidence rate (IR) of an adverse event for therapy group *i*,

$$\mathrm{IR}_i = \frac{\#\mathrm{AE}_i}{\text{Population time at risk}_i},$$

is calculated as the number of patients with at least one adverse event in group *i* (#AE$_i$) divided by the population risk time for an adverse event in group *i* [22]. The population risk time for an AE is calculated as the sum of the individual risk times for each patient. Each patient contributes to the risk time as long as they are followed within the study and no adverse event occurs. The incidence rate estimator is valid for constant hazards [23]. The constant hazards assumption means that the risk of experiencing an adverse event remains the same for all times. This assumption can be examined by plotting the event-specific hazards over time.

The ratio of the two incidence rates, the incidence rate ratio (IRR), is used to compare incidence rates between two groups [22]

$$\mathrm{IRR} = \frac{\mathrm{IR}_1}{\mathrm{IR}_2}.$$

Incidence rates estimate event-specific hazards when the hazard can be assumed to be constant [23]. Similar to the calculation of incidence rates for any adverse event, incidence rates can also be calculated for a certain type of event *h* [25]:

$$\mathrm{IR}_{event\ h} = \frac{\#\text{type h events}}{\text{Population time at risk for event h}}\,.$$

This is the ratio of the number of patients with at least one type *h* event to the time the population is at risk for that event.

Similarly to the incidence proportion and relative risk, confidence intervals for incidence rates and incidence rate ratios are calculated using the Wilson score method with continuity correction (method 4 in Ref. [24]) and the Wald method (method 1 in Ref. [24]), respectively. The corresponding formulae were given in Section 3.1.1. Again, the R functions “prop.test” and “riskratio” were used.

3.1.3 Exposure-adjusted incidence rate

Similarly to the incidence rate the so-called exposure-adjusted incidence rate (EAIR) can be calculated. The exposure-adjusted incidence rate

$$\mathrm{EAIR}_i = \frac{\#\mathrm{AE}_i}{\text{Population exposure time at risk}_i}$$

is defined as the number of patients in treatment group *i* with one or more adverse events during their exposure time divided by the total exposure time of patients in treatment group *i* at risk for an adverse event [26]. Exposure time refers to the time patients are exposed to the drug. Typically, exposure time is defined in the study protocol as the time from the first intake of the study drug to the last intake plus a specified period of time. After this period, adverse events are no longer considered to be associated with the study drug and, depending on the study protocol, may not be reported. For the purpose of calculating EAIR, the time from first intake of study drug to the occurrence of an adverse event is used as the exposure time for all patients with an adverse event. For patients without an adverse event, the exposure time at risk terminates at therapy end plus the predefined time period thereafter. It should be noted that any adverse events occurring after this period are not counted.

For the present work, exposure-adjusted incidence rates are not further considered.

3.1.4 1-Kaplan-Meier estimator

Kaplan-Meier (KM) survival curves are probably the best-known example of time-to-event analysis [27]. The Kaplan-Meier curve estimates probabilities of an event (such as death) or probabilities to be without this event over time. Patients without an event are considered until the last observation at which they are known to be event-free (*i.e.* alive) and censored at that time. In estimating survival probabilities, the KM curve starts at 1 (all patients are alive) and decreases with each time point at which a patient dies. This step function eventually reaches zero when all patients have either died or been censored (and the person with the longest observation time had an event and was not censored). The same approach can be applied to adverse events: The time to occurrence of an adverse event is counted for each patient, and patients without an adverse event are censored on their last observation date. To yield event probabilities, the 1-KM curve is drawn starting at zero and reaching the value of one when all patients have either experienced an adverse event or been censored without an adverse event (and the person with the longest observation time had an event and was not censored). However, this simplified approach is not realistic, and one should also count death and perhaps other events besides the AE (competing events). The concept of competing events is discussed below. The 1-KM curve tends to overestimate the probabilities of adverse events when multiple types of competing events are present and therefore should not be used [22,28]. Overestimation occurs “as soon as the first event of interest directly after the first detection of a competing risk has been recorded” (see *e.g.* [28]).

3.1.5 Cox proportional hazards model

In addition to the Kaplan-Meier method, the Cox proportional hazards model [29] (“Cox model”) is commonly used in the analysis of time-to-event data. It can be used to estimate a treatment effect between two treatment groups for an individual *i* according to the following formula (*e.g.* [30])

$$\lambda_i(t) = \lambda_0(t) exp(\beta X_i), \qquad i = 1,2,\ldots,n.$$

Here $\lambda_i(t)$ is the hazard for a person *i* to experience an event at time *t* and $\lambda_0(t)$ is a common baseline hazard. Beta represents the treatment effect estimate. X_i stands for the individual’s treatment group and is equal to 1 if the patient is assigned to the experimental group and 0 if the patient is assigned to the control group. The aim is to estimate exp(β), the hazard ratio for the treatment effect. The hazard ratio is expected to be independent of time, implying that the hazards between the two groups are proportional over time. The Cox model considers only the first event and neglects all subsequent events.

3.1.6 Competing risks scenario and censoring

In a competing risk setting, a patient may experience multiple possible events, but only the time to the first event and its type is evaluated. An intuitive example of two competing risks would be the occurrence of an adverse event and death: No adverse event can be observed after death, so death is a competing risk to AEs. In the analysis, the time to occurrence of an AE or the time to death without an AE is considered, whichever occurs first.

The occurrence of a competing event may not lead to the censoring of the observation time: Suppose a clinical trial ends on a specific calendar day. At the end of the trial, the event of interest may not have occurred in all trial participants. Subjects for whom the event has not occurred are still at risk for the event, but because the trial has ended, all that is known is that the event has not occurred by the end of the trial (and thus is supposed to occur after the trial has ended). This type of patient is referred to as censored at the time the study ends. If the follow-up had been longer, the theoretical assumption for survival analysis without the presence of competing events is that the event would have been eventually observed, and if the follow-up period had been infinite, the event would certainly have occurred in every participant. Thus, censoring means that the event was not observed until the end of the study but would have been observed if the study had lasted longer, whereas a competing risk hinders the event of interest from occurring (if one dies without AE, one cannot experience AE after death).

It should be noted that the Kaplan-Meier estimator cannot handle competing risks in an appropriate way: Only the composite endpoint “adverse event or death” could be correctly estimated with the Kaplan-Meier estimator. In many publications, when analyzing the event of interest in the presence of competing risks with the Kaplan-Meier estimator, the competing events are censored, which is not the correct way to handle the situation. This leads to an overestimation of the probability of AEs because not every patient experiences an AE as the first event (because some patients experience the competing event) [23].

In our study, the following competing risks were considered: Disease progression (*i.e.* occurrence of accelerated phase or blast crisis), death, allogeneic stem cell transplantation (SCT) and switch to another CML treatment (“TKI switch”). These events were treated as competing risks to the occurrence of AEs. For an illustration of this scenario, see Figure 5. Shown are three patients, who all start at time 0 (in the present study, the start of randomized CML therapy). Patients are followed until an event occurs or until the individual data cutoff date which could coincide with the end of the study. Patient 1 had an adverse event after 2 years of therapy (AE), patient 2 was followed for 4 years without an event and then censored (X), and patient 3 had one of the competing risks after 1 year of therapy (CR). For the calculation of the population time at risk for an adverse event, these three patients would contribute the time of 7 (patient) years.

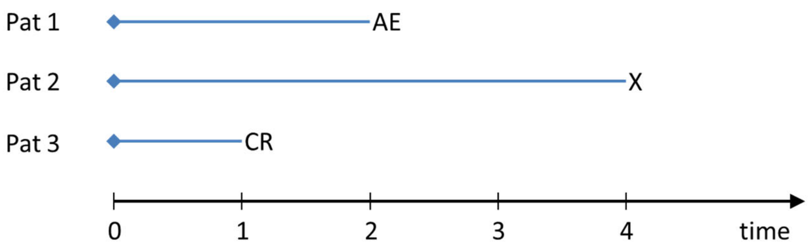


Figure 5: Competing risk scenario for three patients. AE=adverse event, X=last observation, CR=competing risk.

3.1.7 Multi-state models

The competing risk scenario may be illustrated by corresponding multi-state models. A simple multi-state model is shown in Figure 6.

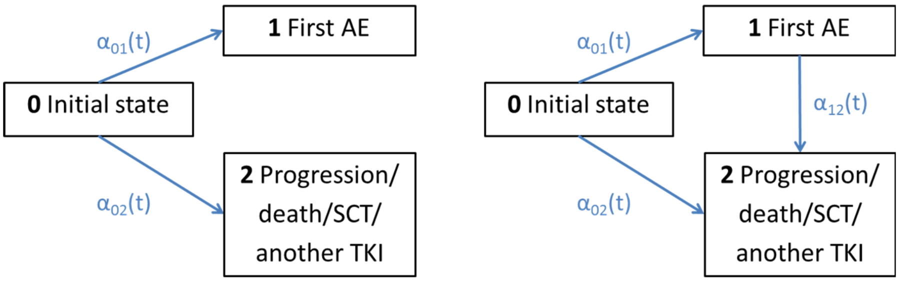


Figure 6: Simple multi-state model suitable for the analysis of an first adverse event (AE) under competing risks of disease progression, death, stem cell transplantation (SCT), or treatment switch to another TKI (tyrosine kinase inhibitor). α_{xy} denotes the hazard of transition from state x to state y, *i.e.* α_{01} is the hazard for experiencing a first AE.

The model on the left describes a situation where only the first transition is considered. All patients start in the initial state 0 and move to state 1 after an adverse event has occurred. If disease progression, death, stem cell transplantation or treatment switch occurs without a prior AE, patients move from initial state 0 to state 2. However, this model does not account for the transition from state 1 to state 2. Such a transition is shown in the model on the right-hand side of Figure 6, where patients who experience disease progression, death, stem cell transplantation or treatment switch after an AE move from state 1 to state 2. In this work, the model with the additional transition after an AE is used.

The final states (state 1 and 2 Figure 6 left, states 2 Figure 6 right) are called absorbing states, since no further transitions from these states can occur. The AE state in the model on the right-hand side of Figure 6 (state 1) is referred to as transient state, since a transition to and from this state is possible.

Multi-state models are fully characterized by transition hazards α_{xy} and transition probabilities P_{xy} between two states *x* and *y*. The transition hazard describes the instantaneous hazard of moving from state *x* to state *y* and the transition probability describes the probability of this transition. The hazard can be calculated with the Nelson-Aalen estimator and the transition probabilities with the Aalen-Johansen estimator. Both estimators are introduced on the next pages.

The multi-state model of Figure 6 is called an illness-death model without recovery. One can extend the models to include, for example, a second adverse event or to allow recovery from adverse events, but for simplicity and to demonstrate multi-state models, the model shown on the right-hand side in Figure 6 is used in the present work. A notable advantage of multi-state models is that all estimates are obtained simultaneously and thus (informative) censoring is not an issue here.

3.1.8 Aalen-Johansen estimator of the cumulative incidence function

The correct method, the gold standard, for estimating the probability of an adverse event in the presence of competing risks is the Aalen-Johansen estimator of the cumulative incidence function (CIF) [23,31]. It corresponds to the expected proportion of patients who experience an adverse event over time. Both time to first event and the type of that event are taken into account. The Aalen-Johansen estimator can be expressed as [23]

$$\text{Aalen-Johansen}_h = \sum_u P(T > u-) \frac{\text{number of type-}h\text{-events at u}}{\text{number of patients at risk before u}},$$

where T is the event time of the event of type h, *P(T>u-)* is the Kaplan-Meier estimator for the probability of having had no event just before time *u,* and *h* describes the type of event, *i.e.* AE or death without AE. The right-hand side of the equation includes the probability of not having failed from any cause before (*P(T>u-)*) and the cause-specific hazard for the event of interest (Nelson-Aalen estimator, see Section 3.1.10). Thus, the CIF estimator depends on all event-specific hazards, not just the hazard for AE itself, and takes into account the probability of remaining at risk for cause *h* at time *t* (*i.e.* not having failed before *t* due to other causes) [32]. Confidence intervals for the CIF estimates are calculated using the method introduced by Choudhury [33] and the Gray's test can be used to test for equality between two cumulative incidence functions, for example, between two different therapies [34].

The Aalen-Johansen estimator for AE probability is considered the gold standard because it accounts for both competing risks and censoring.

3.1.9 Parametric incidence rated based estimator of the cumulative incidence function

Assuming a time constant hazard for the occurrence of an adverse event, a parametric estimator for the cumulative incidence function for therapy group *i* can be derived, as performed by Grambauer *et al.* [25] and addressed, for example, by Allignol *et al.* [23] and Proctor and Schumacher [22]:

$$\text{Parametric CIF estimator } (t) \ = \frac{\text{IR AE}_i}{\text{All event IR}_i} * (1 - \exp(-t * \text{All event IR}_i)),$$

where IR AE_i is the incidence rate for adverse events, all event IR_i the sum of all incidence rates for the event of interest and all competing events, and *t* is the observation time. In the present work, adverse events and the composite endpoint of competing events/risks (CR) are used, which means that the "all event IR" is the sum of "IR AE" and "IR CR": All event IR_i = IR AE_i + IR CR_i. It can be seen that the cumulative incidence function depends on all hazards/events, not just the actual event of interest. This estimator is the parametric analog of the non-parametric Aalen-Johansen estimator of the cumulative incidence function and accounts for competing events.

3.1.10 Nelson-Aalen estimator of event-specific hazards

When considering competing risks, the hazard for each event should be considered. These hazards are referred to as event-specific hazards, such as the hazard for an adverse event or the hazard for death without AE, and can be understood as the instantaneous risk of experiencing the event in question.

Cumulative (event-specific) hazards are often plotted for illustrative purposes. These can be done using the non-parametric Nelson-Aalen estimator [31]. The Nelson-Aalen estimator is defined as (see *e.g.* [25])

$$\text{Nelson-Aalen}_h = \sum_{u \le t} \frac{\text{number of type-}h\text{-events at } u}{\text{number of patients at risk before } u},$$

where the sum over all observed event times u of any event type h, and $u \le t$ is calculated. This estimator is used in the multi-state models for estimation of hazard “alpha” (α_{xy}, see Figure 6).

3.2 Methods for analysis of recurrent adverse events

In the following, models are presented that are suitable for recurrent events analysis. In the previous chapter, methods suitable for analysis of the first adverse event were introduced but depending on the situation valuable data might thus be disregarded. Recurrent adverse events are adverse events that recur after an adverse event-free episode, and depending on the type of the AE, its recurrence may or may not be relevant to the analysis. It may be of interest to analyze multiple types or different severities of AEs to gain more insight into occurrence of specific AEs in different therapy groups. However, the research question should be formulated before analyzing the data, and the decision on the model(s) to be used should be based solely on the purpose of the analysis.

The simplest analysis of recurrent events is to count the number of events in a given period of time and then use Poisson, or in the case of larger variance, the negative binomial distribution to analyze the data. However, these methods fully work only when all patients are followed for the same period of time (no censoring). Since the present work deals with censored data, the Poisson and negative binomial distribution-based methods are not further considered. This work focuses on time-to-event methods introduced below.

3.2.1 Andersen-Gill (AG) model

The Andersen-Gill model [35] is probably the most widely used model for recurrent events. This model is based on the previously introduced Cox proportional hazard model [29]. The model assumes that the events are independent of each other, *i.e.* the occurrence of a first AE does not affect the occurrence of a second AE. Also, it does not matter if the event times originate from the same patient or from different patients. This is a very strong assumption. A global parameter for the treatment effect is estimated, and since the events are assumed to be independent, a common baseline hazard for all events is assumed.

The AG model can be estimated with standard statistical software using the Cox model. The data set needs to include all events for an individual and not just the first event, as for the standard Cox model. Basically, then, the hazard ratio from the Cox model with additional events is estimated as the treatment effect for the AG model. The advantage over the standard Cox model is that the complete data can be used.

The AG model can also be used for non-independent events (the risk of a second AE is higher/lower after a first AE occurs, etc.), but to account for this within-individual correlation, robust standard errors should be estimated [36]. In this work, the robust sandwich variance estimator provided in the software SAS is used.

If the above strong assumption of independency of events does not hold, then the AG model estimates the so-called total effect, which is the sum of the “expected” direct effect and an indirect effect on later disease risk that can be only observed when the events are dependent on each other [37,38]. This means that there could be an effect of the first AE on the first recurrence of the same AE, for example. In this case, the effect of treatment on the first AE is the "expected" direct effect. For the subsequent same AE, the treatment effect is then composed of the direct effect of treatment plus the effect of treatment by the (earlier) first AE. As stated above, in this case the AG model estimates the total treatment effect.

3.2.2 Prentice, Williams and Peterson (PWP) model

The Prentice, Williams and Peterson (PWP) model is also a Cox model based conditional model for recurrent events [39]. In the PWP model, the ordered adverse events are analyzed. Here, a separate Cox model is fitted for each recurrent event, and for this reason, the baseline risk for each event may also vary. All patients are at risk to experience a first AE, but only those patients with an AE are at risk to experience a second AE. Accordingly, only the patients who experienced a second AE are also at risk for a third AE, and so on. Therefore, the number of patients at risk decreases as the number of recurrent AEs increases. This results in the number of patients at risk being so small for a given recurrence that the model can no longer be reliably estimated. Therefore, not all data can be used for this model.

In addition to fitting separate effects for each AE occurrence, it is also possible to estimate a common overall treatment effect. When separate effects are estimated, the interpretation for each recurrence must be conditional on the previous event(s): “Given the first AE, the risk for the second AE is significantly increased/decreased.”

The PWP model differs from the AG model in that the patients at risk for the (*k+1*)th recurrence consist of those who have already had the first *k* recurrences and that the risks for different recurrences are allowed to vary [40]. One might prefer the PWP model to the AG model if the treatment effect can be expected to differ for subsequent events, for example, if the occurrence of the first event increases the risk of recurrence. This type of conditional risk set might better reflect true disease progression, but this is certainly disease-specific. The PWP model is also suitable if the effects are to be estimated separately for each AE.

Regarding the time considered, the PWP model offers two different analysis options: One can analyze the *total time* or the *gap time*, as discussed in the original publication of the model [39]. While the total time model considers the time since the study entry, the gap time model examines the events since the last event. The total time model evaluates the effect of therapy on the *k*th event since study entry and the gap time model since the time after the last AE. It should be noted that patients are only at risk of experiencing an event during the “gap times” and not during an ongoing event. Because some adverse events, such as laboratory abnormalities, may persist for years, the actual time period at risk for a next AE may become small. This assumption of the model can be reasonable, but it can also impose some limitations that depend on the type of analysis: For example, while it may be reasonable that one cannot suffer a second episode of a decreased platelet count while the platelet count is still decreased (persisting), this could be a limitation when considering a combined serious adverse event end point, for example: While suffering from life-threatening anemia, one is not at risk for other serious adverse events until the condition has resolved.

The PWP model estimates a direct treatment effect on the risk of an event, and in the presence of competing risks, the model is always limited to the analysis of cause-specific hazards [38].

3.2.3 Wei, Lin and Weissfeld (WLW) model

The Wei, Lin and Weissfeld (WLW) model is another recurrent event regression model [40]. Each recurrence time (“time to event”) is measured from the study start to the time of occurrence of an event. Thus, the times are calculated similarly to the PWP total time model. The difference with the PWP model is that in the WLW model, a person is at risk for all events as long as he or she is observed in the study. The order of occurrence of events is ignored, meaning that patients are considered at risk for the next event even if they have not experienced a previous event. More neutrally, “no particular structure of dependence among distinct failure times on each subject is imposed” [40]. However, this flexibility comes with some limitations, as discussed below.

The WLW model estimates one treatment effect for each time-ordered event: When the first n recurrencies are considered, n effects are estimated. A combined treatment effect can be calculated as a simple linear combination of the event-by-event effect estimates and considered an “average treatment effect”.

The conceptualization of the risk set of the WLW model could be problematic because all patients are at risk for all events as long as they are observed within the study. This means that a patient may be at risk for the third occurrence of an event even if he never experienced a first and a second event. Such conceptualization of the risk set often seems to lead to a violation of the proportional hazards assumption, which prevents a straightforward interpretation of the model estimates [38].

Another disadvantage of the method is that not all data can be used in the analyses. In principle, all patients still in the study remain at risk for all events, but the models cannot be estimated if the number of patients experiencing an event becomes too small. For example, it could be that only one patient had 9 occurrences of the same AE and all other patients had only 2 or 3. In this constellation, only a maximum of 1 to 2 recurrences of the AE could be analyzed. However, the sample size is still larger than in the case of the PWP model, where patients without events are excluded from the risk set for the first second recurrence (second same AE) and so on.

For this work, the maximum number of events for which the model could still be estimated was used. The maximum number was identified by starting with the full data set and gradually reducing the number of estimated recurrences until the model converged. In this way, data loss was kept to a minimum, while making the most reasonable usage of the data.

3.2.4 Mean cumulative function (MCF)

The mean cumulative function (MCF) is a nonparametric approach for analyzing the cumulative number of events per patient over the entire observation time. Measurements are assumed to be independent across subjects (one patient’s AEs are independent of another patient’s AEs), but not necessarily the times between AEs (an initial AE could influence the occurrence of subsequent AEs).

The mean cumulative function can be estimated by *M(t)*, which is the mean cumulative number of AEs up to time *t* [41]. This function is a step function that starts at zero and increases each time an AE has occurred and remains constant between times with no observed AEs.

An illustration of the mean cumulative function can be found in Figure 7. The individual cumulative number of AEs is shown for three patients (yellow, blue and green lines). In one patient, an AE occurs one year after the therapy start (blue line). In a second patient, a total of 3 AEs occur: the first after 1.5 years, the second after 2 years, and the third AE after 3 years (yellow line). A third patient remains free of AEs for 3 years and then experiences an AE (green line). Considering all individual AE profiles and taking the average at each time point, one obtains the mean cumulative function represented by the black dashed line. Please note that Figure 7 is just an illustration and the mean cumulative function shown there is not based on the exact estimations of the 3 hypothetical patients.

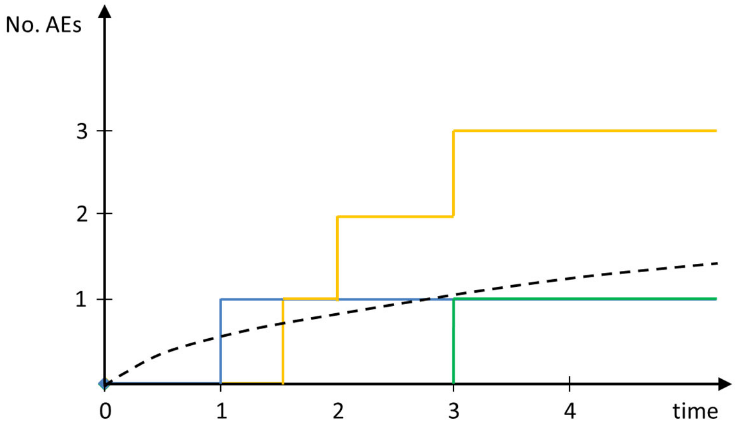


Figure 7: Illustration of the mean cumulative function. The black dashed line shows the mean cumulative function over time, and the blue, yellow, and green lines show the cumulative number of adverse events (AEs) for three example patients.

From plots of the MCF, the number of adverse events expected by time *t* can be obtained. Confidence intervals around the estimates can be calculated, and there is a test for comparing two MCFs over time [41-43]. The confidence intervals are calculated point-wise using a normal approximation (see [41] and technical appendix of [42]). To compare two MCFs, the differences between the two MCFs are first calculated. Then a pointwise confidence interval is calculated around each difference. If the confidence interval does not include the value of zero, point-wise statistical significance can be claimed. If all pointwise confidence intervals do not include the value zero, statistically significantly different MCFs can be asserted. However, this seems quite cumbersome, and there are certainly many cases where some of the pointwise confidence intervals include the value zero. Cook and Lawless ([43] Section 3.7.5) proposed an approach similar to the "weighted log-rank test" to test whether two MCFs are significantly different (over time). They introduced two different weighting functions: a constant one that should be used in cases of proportional MCFs, and a linear one that should preferably be used in cases of non-proportional (but still non-crossing) MCFs [43]. These two tests are included in the "proc reliability" procedure in SAS used in this work, and in this work the test based on a constant weighting function is preferred.

The use of the MCF and its associated test provides an advantage over Kaplan-Meier analysis and the log-rank test because not only the first AE but also recurrent AEs are included. Depending on the research interest, one may want to plot the MCF for single (recurrent) AEs, for all AEs, or, for example, for AEs related to a particular organ system such as cardiovascular AEs.

3.2.5 Types of recurrent data analyzed

The recurrent event methods introduced above are formulated to model the same type of recurrent events such as adverse events. There is an interesting systematic comparison by Ozga *et al.* investigating the suitability of the recurrent events models for modeling composite endpoints [30]. An example for such a composite endpoint could be adverse event, progression or death, whichever occurs first. In this work, only one type of recurrent event is considered: adverse events. Even when multiple AE severities have been combined, such analyses are being considered as belonging to the category of “single” endpoints. For more details, see Chapter 4.6.1.

3.3 Software used for the analysis

All analyses in the present work were carried out using SAS 9.4 and R version 3.5.0 [44].

4. Results

First, a brief summary of the main events during the TIGER study is given to provide an overview (Chapter 4.1). The safety analysis is presented starting with a “conventional” analysis without consideration of the time to an adverse event (Chapter 4.2). The analysis is then extended to include time to first adverse event, together with different models to analyze the data: Chapter 4.3 deals with cumulative incidences, Chapter 4.4 introduces a multi-state model, and Chapter 4.5 assesses the extend of bias introduced by using other estimators for the cumulative incidence function than the gold standard. At the end of this chapter (Chapter 4.6), the methods are extended to recurrent adverse events, presenting different models for such data. The chapter concludes with a brief summary of the results (Chapter 4.7).

4.1 Brief summary of the main events in the TIGER study

As of December 14th 2018, 689 (99.6%) of the 692 randomized patients had received at least one dose of study drug (A in group A, or either A or B in group B). A total of 34 patients (4.9%) had withdrawn informed consent. Only data up to the date of withdrawal were used. Disease progression had occurred in 15 (2.2%) patients (4 accelerated phases, 8 blast crises, and 3 accelerated phased followed by a blast crisis). Allogeneic stem cell transplantation had been performed in 16 patients, in most cases after disease progression. A total of 23 patients (3.3%) had died: 5 patients died of CML (4 of them due to blast crisis), 14 due to other causes, and in 4 patients the cause of death was unknown.

The median time under treatment was similar in both groups: 30 months (range: 5 days – 70 months) in treatment group A and 30 months (range: 7 days – 68 months) in treatment group B, respectively. Time was calculated from first to last reported intake of study drug(s), without consideration of short-term interruptions in treatment.

Regarding the different study phases (see also Figure 1), 411 (59.4%) patients had completed induction therapy and entered the maintenance phase. A total of 158 of these 411 patients (22.8% of all patients) had completed the maintenance phase and entered the treatment-free phase of the study.

4.2 Conventional analysis of adverse events

Firstly, an analysis of adverse events was performed as commonly found in the literature. The analysis was based solely on the number of patients with adverse events without consideration of event times. Raw (crude) percentages were calculated and compared between two groups using relative risks. As mentioned earlier, this type of analysis would be valid in a complete data set where all patients were observed for the same time period and there were no missing data. Including times under risk for an adverse event in the analysis also allows incidence rates and incidence rate ratios to be calculated.

4.2.1 Data preparation

The number of patients with adverse events was counted. For the analysis of any adverse event, all adverse events were considered regardless of the reported severity. For the analysis of severe and serious adverse events, only those adverse events with a reported value for the severity grade could be included; adverse events with unknown or missing severity grade were excluded.

The time under risk for an adverse event was calculated in a similar way as in Figure 5: For patients with reported adverse events, the time was counted from the start date of therapy to the start date of the adverse event. For patients with a competing risk to the occurrence of adverse events (disease progression/death/stem cell transplantation/TKI treatment switch), the time from the start of therapy to the reported date of the occurrence of one of the competing risks was counted. For patients without adverse events and competing risks, while under therapy, the time from therapy the start of therapy to the patient's last observation was counted. If a patient paused therapy for a short period and an adverse event occurred during this pause, the adverse event was counted. The population risk time was then calculated as the sum of the individual risk times. The inclusion of competing risks was chosen for better comparability with the other methods used in the present work.

The calculations were performed with exact values, even if the intermediate results are given with three decimal places.

4.2.2 Times under risk for adverse events

The times under risk for different severities and types of AEs are shown in Table 2.

Therapy group	Median (total) risk time					
	All severities	Severe AEs	SAEs	Fatigue	Thrombopenia	Neutropenia
A	1 (1768)	21 (8319)	29 (10586)	21 (8092)	29 (10386)	31 (11325)
B	1 (1167)	18 (7155)	27 (9999)	18 (7007)	25 (9180)	30 (10354)

Table 2: Population times under risk for different adverse event severities and different types of adverse events for both therapy groups A and B. Median risk time is in months, total risk time in parentheses in patient-months. (S)AE = (serious) adverse event.

While the median time under risk for any adverse event was one month in both therapy groups, the corresponding times for severe AEs were 21 and 18 months for therapies A and B, respectively. Regarding the risk times for SAEs, the median times were 29 and 27 months in therapy groups A and B, respectively. Because almost all patients experienced at least one AE, the median and total risk times were small compared with the risk times for SAEs, which were observed only in a minority of patients. For fatigue, thrombopenia and neutropenia median risk times ranged from 21 to 31 months in therapy group A and 18 to 30 months in group B.

4.2.3 Any adverse event

Of the 7317 adverse events, one or more events were reported in 636 patients. A bar plot of the frequency of reported adverse events per patient is shown in Figure 8. The median number of AEs per patient in therapy groups A and B is 6 and 8, respectively (range 0 to 79 events).

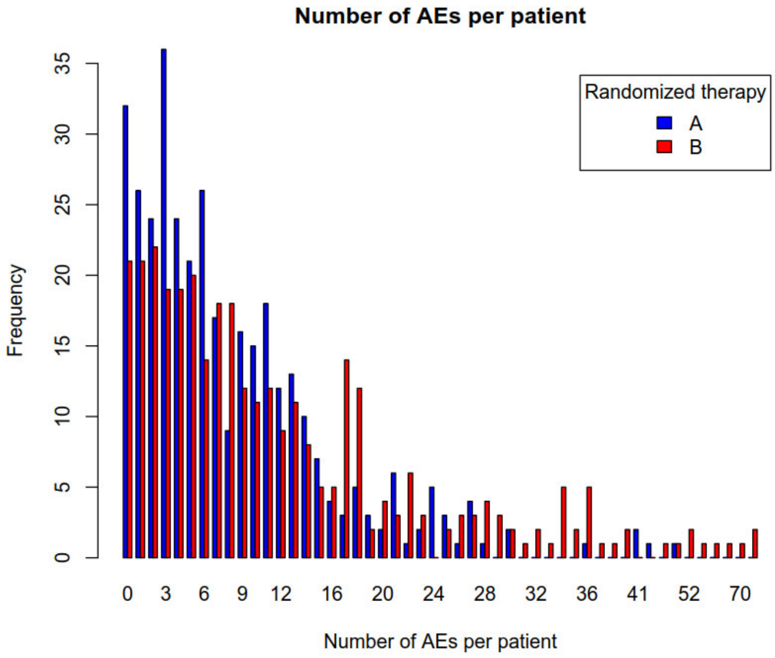


Figure 8: Frequency of reported adverse events (AEs) per patient and therapy group.

From this, the crude incidence proportion (IP) can be calculated: IP_{cr}=636/689=0.923 (95% confidence interval (CI): [0.900, 0.941]). The interpretation of the incidence proportion is as follows: An adverse event had occurred in 92.3% of the patients. For therapies A and B separately, the crude incidence proportions were $IP_{cr,A}$=321/353=0.909, CI: [0.873, 0.936] and $IP_{cr,B}$=315/336=0.938, CI: [0.904, 0.960], respectively. The crude incidence proportion ratio (*i.e.* relative risk) RR_{cr}= $IP_{cr,B}$/$IP_{cr,A}$=1.031, CI: [0.988, 1.076] means that the risk for AEs was 1.03-fold higher for therapy B compared with therapy A. However, since the confidence interval included the value 1, the effect was not statistically significant.

The population times under risk for an adverse event were 1768 patient-months in therapy group A and 1167 patient-months in therapy group B, respectively. These times could be used to calculate the incidence rates (IR) for occurrence of adverse events: IR_A=321/1768.411=0.182 and IR_B= 315/1166.752=0.270. An incidence rate ratio (IRR) of IRR=IR_B/IR_A=0.270/0.182=1.487, CI: [1.297, 1.705] was obtained, which showed a statistically significant higher AE incidence in therapy group B compared to group A.

4.2.4 Severe adverse events (grades 3 to 5)

A total of 832 severe adverse events (*i.e.* adverse events of severity grade 3, 4 or 5) were entered into the data base. 792 of these were reported after initiation of therapy and included in the analysis of the observation of a first severe adverse event. Of these 792 severe adverse events, 343 (43.3%) were reported in therapy arm A and 449 (56.7%) were reported in therapy arm B. A bar plot of the frequency of reported severe adverse events per patient is shown in Figure 9. The median number of severe AEs per patient was zero in both therapy groups (range 0 to 35 events).

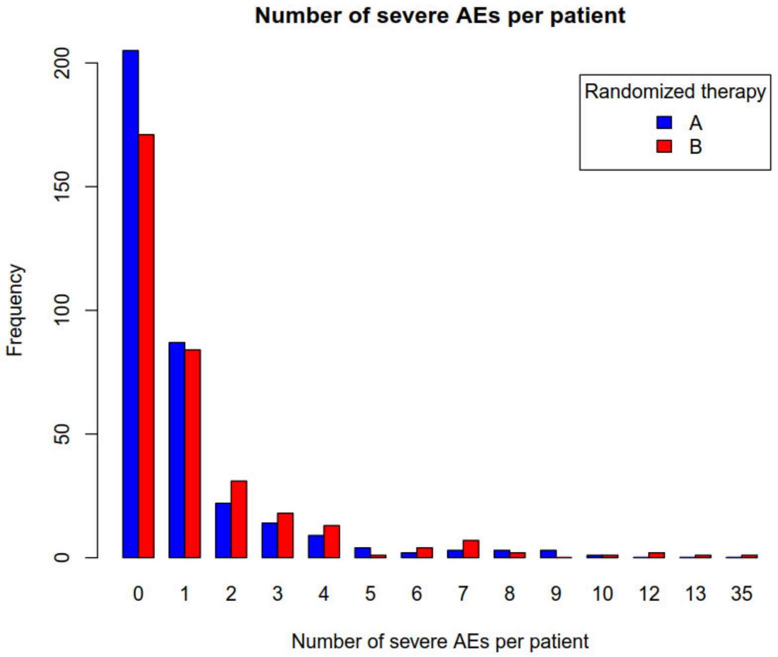


Figure 9: Frequency of reported severe adverse events (AEs) per patient and therapy group.

One or more adverse events were reported in 311 patients. This corresponds to a crude incidence proportion of IP_{cr}=311/689=0.451, CI: [0.414, 0.489]. For therapies A and B separately, the crude incidence proportions were $IP_{cr,A}$=147/353=0.416, CI: [0.365, 0.470] and $IP_{cr,B}$=164/336=0.488, CI: [0.434, 0.543], respectively. The crude relative risk of RR_{cr}= $IP_{cr,B}$/$IP_{cr,A}$=1.172, CI: [0.994, 1.382] corresponds to a 1.2-fold higher risk for severe AEs in therapy group B compared to therapy group A. However, the effect was not statistically significant.

The population times under risk for a severe adverse event were 8319 patient-months in therapy group A and 7155 patient-months in therapy group B, respectively. These times could be used to calculate the incidence rates for occurrence of severe adverse events: IR_A=147/8319.310=0.018 and IR_B=164/7154.628=0.023. The incidence rate ratio of IRR=IR_B/IR_A=0.023/0.018=1.297, CI: [1.041, 1.617] was obtained, which showed a statistically significant higher severe AE incidence in therapy group B compared to group A.

4.2.5 Serious adverse events only

The database contained 201 reported serious adverse events (SAEs). Of these, 187 had a valid start date after therapy start and were included in the analysis. A bar plot of the frequency of reported serious adverse events per patient is shown in Figure 10. The median number of SAEs per patient was zero in both therapy groups (range 0 to 8 events).

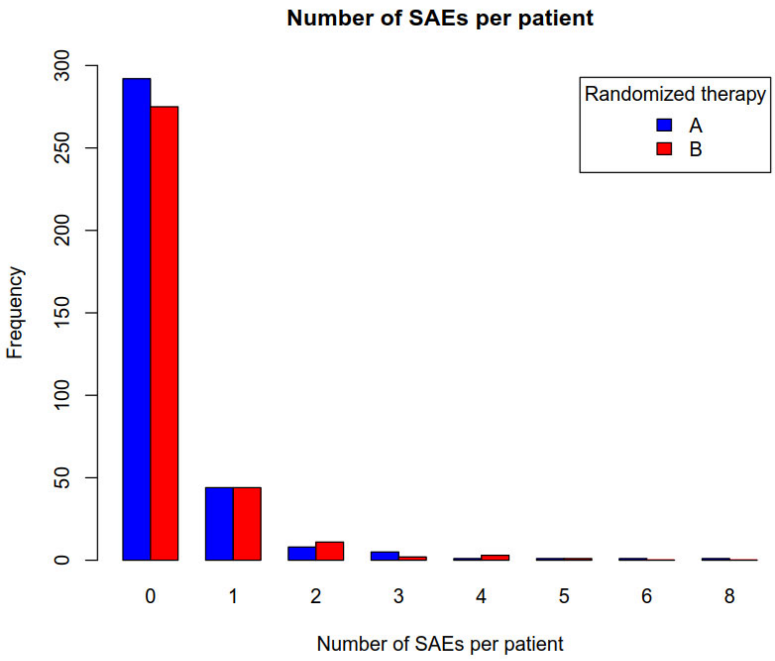


Figure 10: Frequency of reported serious adverse events (SAEs) per patient and therapy group.

These SAEs were distributed as follows: in both groups A and B, one or more SAEs were reported in 61 patients, corresponding to a crude incidence proportion of IP_{cr}=(61+61)/689=0.163, CI: [0.136, 0.193] in all patients and $IP_{cr,A}$=61/353=0.173, CI: [0.136, 0.217] and $IP_{cr,B}$=61/336=0.182, CI: [0.143, 0.228] in patients in treatment groups A and B, respectively. The crude relative risk of RR_{cr}= $IP_{cr,B}$/$IP_{cr,A}$=1.051, CI: [0.761, 1.450] indicates that the risk for the first SAE was 1.1-fold higher for treatment B compared with treatment A. However, the result was not statistically significant.

The population times under risk for SAE were 10586 patient-months in therapy group A and 9999 patient-months in therapy group B, respectively. Using these times, the incidence rates for occurrence of adverse events could be calculated: IR_A=61/10585.791=0.006 and IR_B=61/9998.752=0.006. The incidence rate ratio of IRR=IR_B/IR_A=0.006/0.006=1.059, CI: [0.743, 1.508] did not point towards higher incidence of SAEs in one of the therapy groups.

4.2.6 Single adverse events of any grade

A few of the most commonly reported adverse events were analyzed separately. The most commonly reported adverse events in the TIGER study were fatigue, thrombopenia, increased alanine aminotransferase (ALAT), headache, flu-like symptoms and increased blood bilirubin (Table 3). The numbers refer to the number of adverse events reported, not to the number of patients in whom these events occurred. The two most frequent AEs fatigue and thrombopenia were selected to demonstrate the statistical analysis methods. In addition, with 87 reported occurrences, neutropenia was analyzed as a rare AE. It was included in the present thesis to investigate whether rare adverse events are suitable for recurrent events analysis or not.

Adverse event	Frequency	Percentage
Fatigue	392	5.38
Thrombopenia (platelet count decreased)	315	4.32
Alanine aminotransferase (ALAT) increased	241	3.31
Headache	213	2.92
Flu-like symptoms	178	2.44
Blood bilirubin increased	170	2.33

Table 3: The most frequently reported adverse events among 7317 adverse events in the TIGER study. Frequency refers to the number of adverse events, not the number of patients in whom the adverse event occurred.

4.2.6.1 Fatigue

Among the 689 patients who received the study drug(s), fatigue was reported 392 times. This AE was reported once or more in 109 patients randomized to treatment A and in 128 patients randomized to treatment B. The corresponding crude incidence proportion among all patients was IP_{cr}=(109+128)/689=0.344, CI: [0.309, 0.381] and among patients in treatment arm A and B $IP_{cr,A}$=109/353=0.309, CI: [0.262, 0.360] and $IP_{cr,B}$=128/336=0.381, CI: [0.392, 0.435], respectively. For fatigue, the relative risk was RR_{cr}= $IP_{cr,B}$/$IP_{cr,A}$=1.234, CI: [1.003, 1.518]. Thus, the risk for fatigue was 1.2-fold higher in treatment arm B than in treatment arm A, and was statistically significant.

The population times under risk for an adverse event were 8092 patient-months in therapy group A and 7007 patient-months in therapy group B, respectively. These times could be used to calculate incidence rates for the occurrence of fatigue: IR_A=109/8091.893=0.013 and IR_B= 128/7007.080=0.018. This resulted in the incidence rate ratio IRR=IR_B/IR_A=0.018/0.013= 1.356, CI: [1.053, 1.747], showing a statistically significantly higher incidence of fatigue in therapy group B compared to group A.

4.2.6.2 Thrombopenia

Thrombopenia was reported 315 times in the database. One or more episodes of thrombopenia were reported in 45 and 61 patients in treatment groups A and B, respectively. The crude incidence proportion of thrombopenia was IP_{cr}=(45+61)/689=0.154, CI: [0.128, 0.183]. For both therapy groups separately, the crude incidence proportions were $IP_{cr,A}$=45/353=0.127, CI: [0.095, 0.168] and $IP_{cr,B}$=61/336=0.182, CI: [0.143, 0.228], respectively. The relative risk was RR_{cr}= $IP_{cr,B}$/$IP_{cr,A}$ =1.424, CI: [0.999, 2.031], showing a trend toward a higher risk of thrombopenia in treatment group B compared to group A.

The population times under risk for thrombopenia were 10386 patient-months in therapy group A and 9180 patient-months in therapy group B, respectively. Using these times, the incidence rates for occurrence of thrombopenia could be calculated: IR_A=45/10386.497=0.004 and IR_B=61/9180.025=0.007. This resulted in the incidence rate ratio of IRR=IR_B/IR_A=0.007/0.004= 1.534, CI: [1.044, 2.252], showing a statistically significantly higher incidence of thrombopenia in therapy group B compared to group A.

4.2.6.3 Neutropenia

Neutropenia was a not so common adverse event and reported 87 times. However, it was classified as severe or life-threatening in 44 cases (50.6%). Neutropenia occurred once or more frequently in 8 patients randomized to treatment A and 24 patients randomized to treatment B. The overall crude incidence proportion was IP_{cr}=(8+24)/689=0.046, CI: [0.032, 0.066]. For treatment arms A and B, crude incidence proportions were $IP_{cr,A}$=8/353=0.023, CI: [0.011, 0.046] and $IP_{cr,B}$=24/336=0.071, CI: [0.047, 0.106]. The relative risk of neutropenia, RR_{cr}= $IP_{cr,B}$/$IP_{cr,A}$=3.152, CI: [1.436, 6.917], showed a statistically significantly increased risk for neutropenia in treatment group B compared to treatment group A.

The population times under risk for neutropenia were 11325 patient-months in therapy group A and 10354 patient-months in therapy group B, respectively. Using these times, the incidence rates for the occurrence of neutropenia could be calculated: IR_A=8/11324.88=0.001 and IR_B=24/ 10354.10=0.002. This resulted in the incidence rate ratio IRR=IR_B/IR_A=0.002/0.001=3.281, CI: [1.475, 7.300], showing statistically significantly higher incidence of neutropenia in therapy group B compared to group A.

4.2.6.4 Tabular overview of the results

The calculated relative risks and incidence rate ratios are reported Table 4. *Please note that no exact p-values are reported in this thesis, only the distinction between significant and non-significant results (p<0.05 vs. "n.s."), as intermediate results of the study are reported.*

Accounting for different follow-up times in both therapy groups led to increased effect sizes. Using incidence rate ratios and thus accounting for different observation times in both groups resulted in statistically significantly higher incidence rates for any, any severe adverse event and thrombopenia compared to the relative risks without accounting for follow-up time. Regardless of the method used, there was no statistically significant difference in the incidence of SAEs between the two therapy groups.

Adverse event	$RR_{B\text{ vs. }A}$ [95% CI]	RR p-value	$IRR_{B\text{ vs. }A}$ [95% CI]	IRR p-value
Any	1.031 [0.988, 1.076]	n.s.	1.487 [1.297, 1.705]	**<0.05**
Any severe	1.172 [0.994, 1.382]	n.s.	1.297 [1.041, 1.617]	**<0.05**
Any SAE	1.051 [0.761, 1.450]	n.s.	1.059 [0.743, 1.508]	n.s.
Fatigue	1.234 [1.003, 1.518]	**<0.05**	1.356 [1.053, 1.747]	**<0.05**
Thrombopenia (platelet count decreased)	1.424 [0.999, 2.031]	n.s.	1.534 [1.044, 2.252]	**<0.05**
Neutropenia	3.152 [1.436, 6.917]	**<0.05**	3.281 [1.475, 7.300]	**<0.05**

Table 4: Relative risks (RR), incidence rate ratios (IRR), their 95% confidence intervals (CI), and the corresponding p value for the comparison of treatment B versus treatment A for the adverse events listed in the left column. SAE = serious adverse event.

4.2.7 Summary of the conventional adverse event analysis

A simple analysis using incidence proportions and relative risks to compare adverse event proportions between therapy groups ignores any temporal effects and does not take into account the different observation times between the two groups. Such analyses are common in the literature and showed a significantly higher incidence of fatigue and neutropenia in the present thesis. A minor improvement over completely ignoring time in the analysis of adverse events is the use of incidence rates and incidence rate ratios that take into account the duration of follow-up in both groups by including the time to the occurrence of the adverse event in the denominator. In the present analysis, both the time to occurrence of an adverse event and the time to occurrence of a competing risk were used to calculate incidence rate ratios. This approach made the differences in adverse events between the two therapy groups more apparent: significantly higher incidence rates were additionally found for any, any severe adverse event, and thrombopenia. Regardless of the method used, there was no statistically significant difference in the incidence of SAEs between the two therapy groups.

In the following, the time to first adverse event is included in the analyses by using the time-to-event methods known from survival analysis. The results of this improvement are compared with the results presented in section 4.2.

4.3 First adverse event: Cumulative incidences

4.3.1 Data preparation

The following competing risk scenario was chosen for the analysis of time to first AE: Disease progression (*i.e.,* occurrence of accelerated phase or blast crisis, n=15), death (n=21), allogeneic stem cell transplantation (SCT, n=16), and switch to another CML treatment ("TKI switch", n=109) were treated as competing risks to occurrence of AEs. The composite endpoint of these four events (*i.e.,* time to first occurrence of any of these events, n=161) was used in the analysis. This approach of combining the event competing to the event of interest is the standard proceeding seen in the literature (*e.g.* [45,46]), regardless of the sample size. Since a relation to randomized treatment should be possible, only the time of treatment with the study drug was considered. Thus, for each patient, we had 3 possible paths (see also Figure 5): A patient could either experience an AE (event), have a progression/death/SCT/TKI switch without a prior AE (competing event) or be followed until their last observation date without an AE or competing event (censoring). Patients without events who withdrew their written informed consent were censored on the date of withdrawal of consent.

4.3.2 First adverse event of any grade

For the analysis of a first adverse event of any grade, all 7313 adverse events were considered. The time to the first event (AE or competing risk, whichever occurred first) and its type were determined. Patients in whom no event occurred were censored at their last observation date. An adverse event was reported as the first event in 93.2% (636/689) of all patients, a competing event in 0.4% (3/689) patients and 7.3% (50/689) of the patients were censored without an event. The cumulative incidence function (CIF) of time to first adverse event, as estimated by the Aalen-Johansen estimator, was plotted in Figure 11. It can be seen that the first AE occurred rapidly after the start of therapy and the incidences were approximately the same in both groups. After about 2 months, differences in the cumulative incidences between the two therapies could be seen: the cumulative incidence of AEs for therapy B became greater and remained so until the last observation. However, Gray's test of equality of the two CIFs showed no statistically significant differences in the cumulative incidence of AEs of any grade.

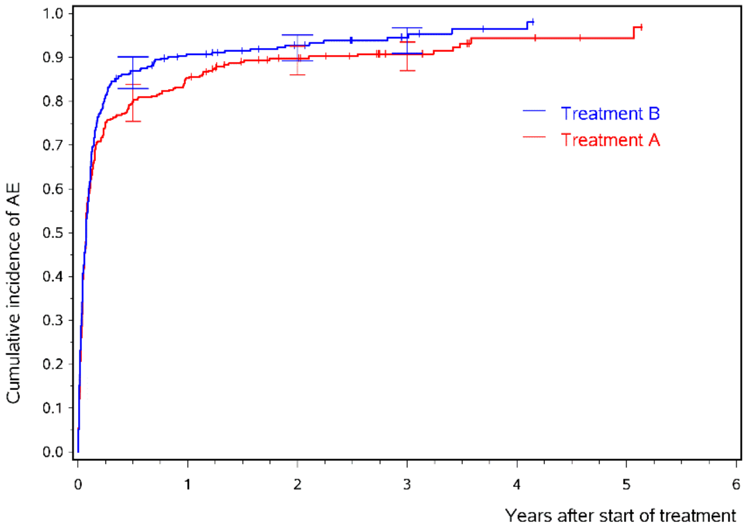


Figure 11: Cumulative incidence functions of a first adverse event (AE) of any severity for therapies A and B. Gray's test p=n.s.. Also shown are the 95% confidence intervals at 6 months, and at 2 and 3 years after therapy start.

4.3.3 First severe adverse event

For the analysis of the first severe adverse event, a total of 792 severe AEs with a valid start date were considered. A severe adverse event was reported as the first event in 44.0% (303/689) of all patients, a competing event in 5.2% (36/689), and 50.8% (350/689) of patients were censored without an event. The cumulative incidence function of time to first severe adverse event is shown in Figure 12. It can be seen that the incidence of a severe AE up to about 2 months was approximately the same in both therapy groups. Thereafter, the incidence became greater in therapy group B and remained so. Gray's test for equality of the two CIFs showed a statistically significant difference in the cumulative incidences of severe AEs (p<0.05). The cumulative incidence of severe adverse events was statistically significantly higher in therapy arm B than in therapy arm A.

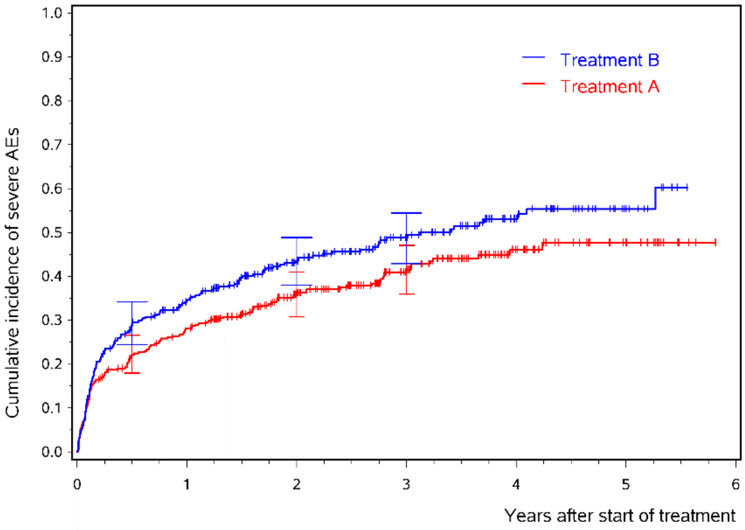


Figure 12: Cumulative incidence functions of a first severe adverse event (AE) of severity grades 3 to 5 for therapies A and B. Gray's test p<0.05. Also shown are the 95% confidence intervals at 6 months, and at 2 and 3 years after therapy start.

4.3.4 First serious adverse event

For the analysis of the first SAE, a total of 187 SAEs with a valid start date after therapy start were included. An SAE was reported as the first event in 16.1% (111/689) of all patients, a competing event in 11.8% (81/689), and 72.1% (498/689) of patients were censored without any event. The cumulative incidence function of the time to a first SAE is shown in Figure 13. For both therapies A and B, no clear difference could be seen between the cumulative incidences of SAEs. The incidence remained low and the curves crossed a few times. The non-significant p value from Gray's test supported the conclusion that there was no statistically significant difference in the cumulative incidence of SAEs between the two therapies.

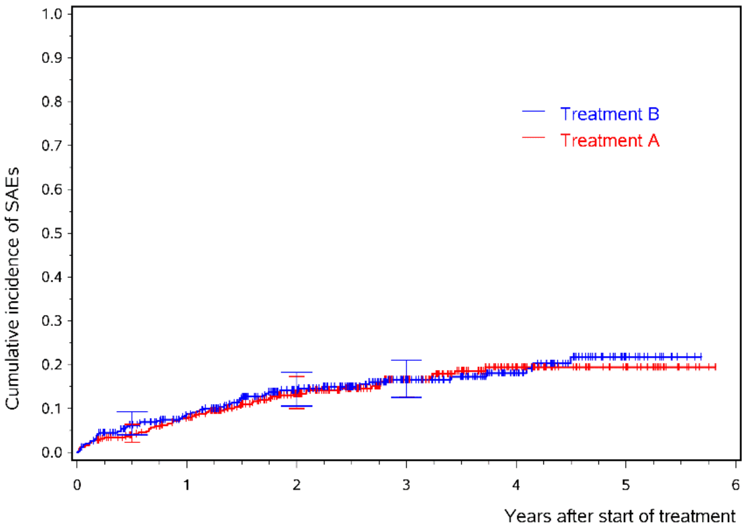


Figure 13: Cumulative incidence functions of a first serious adverse event (SAE) for therapies A and B. Gray's test p= n.s.. Also shown are the 95% confidence intervals at 6 months, and at 2 and 3 years after therapy start.

4.3.5 First single adverse event of any grade

4.3.5.1 Fatigue

Fatigue was reported as the first event in 33.5% (231/689) of all patients, as a competing event in 12.5% (86/689) of patients, and 54.0% (372/689) of patients were censored without any event. The cumulative incidence function for the first episode of fatigue is shown in Figure 14. Fatigue was reported significantly more frequently in therapy group B compared to group A right from the start of therapy (p<0.05).

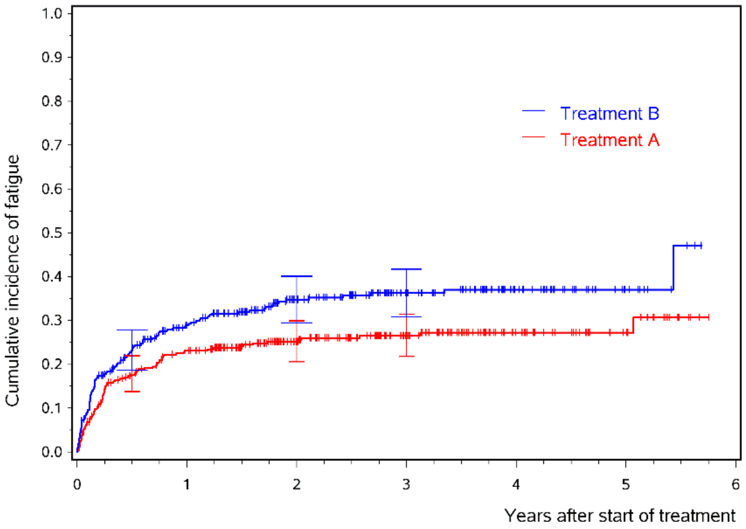


Figure 14: Cumulative incidence functions of fatigue for therapies A and B. Gray's test p<0.05. Also shown are the 95% confidence intervals at 6 months, and at 2 and 3 years after therapy start.

4.3.5.2 Thrombopenia

Thrombopenia was reported as the first event in 14.9% (103/689) of all patients, as a competing event in 12.6% (87/689) of patients, and 72.4% (499/689) of patients were censored without any event. The cumulative incidence function for the first reported thrombopenia is shown in Figure 15. Thrombopenia was reported more frequently in therapy group B than in group A, and the result was statistically significant (p<0.05). The first episode of thrombopenia occurred shortly after the start of therapy in both groups, and almost no new thrombopenias were reported after few months of therapy.

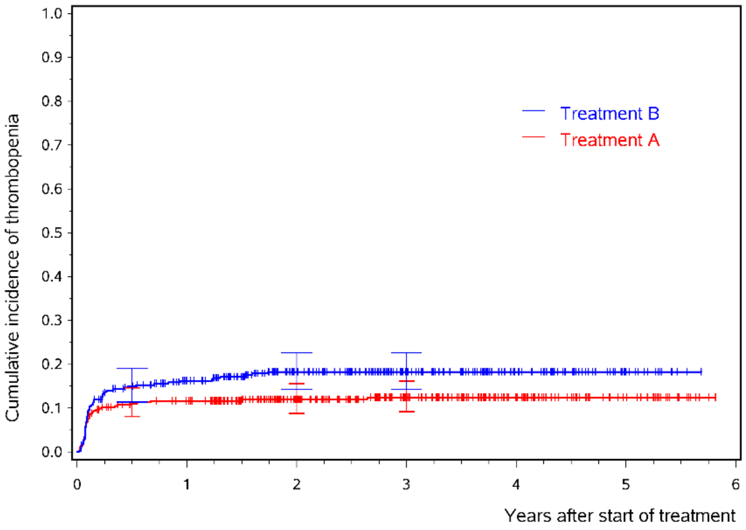


Figure 15: Cumulative incidence functions of thrombopenia for therapies A and B. Gray's test p<0.05. Also shown are the 95% confidence intervals at 6 months, and at 2 and 3 years after therapy start.

4.3.5.3 Neutropenia

Neutropenia was reported as the first event in 4.2% (29/689) of all patients, as a competing event in 16.5% (114/689) of patients, and 79.2% (546/689) of patients were censored without events. The cumulative incidence function for the first reported neutropenia is shown in Figure 16. There was a statistically significant increase in the incidence of first reported neutropenia in therapy group B compared to group A (p<0.05). Of note, only 8 neutropenias were reported in treatment group A and the cumulative incidence curve is almost flat.

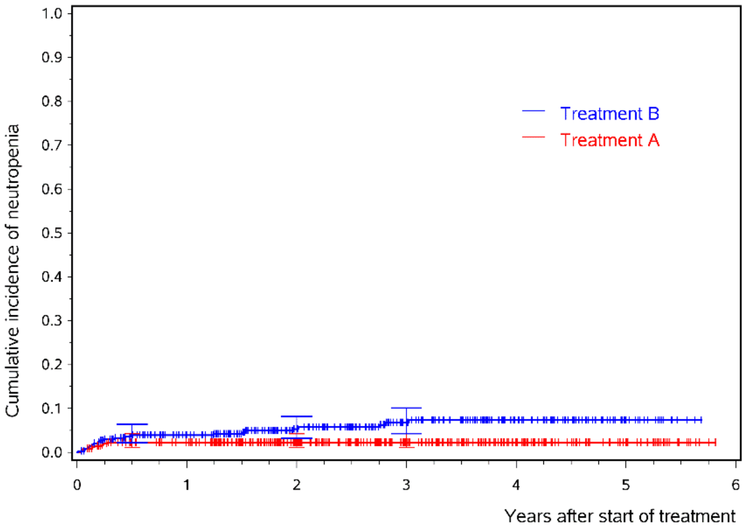


Figure 16: Cumulative incidence functions of neutropenia for therapies A and B. Gray's test p<0.05. Also shown are the 95% confidence intervals at 6 months, and at 2 and 3 years after therapy start.

4.3.6 Summary of cumulative incidences of first adverse events

The cumulative incidence of any, severe and serious adverse events between the two therapy groups was examined using the Aalen-Johansen estimator and Gray's test. A statistically significantly higher incidence of severe adverse events was found in therapy group B compared with group A. There were no statistically significant differences in the cumulative incidence of any or serious adverse events between the two therapy groups. Regarding the single selected adverse events, the cumulative incidence of fatigue, thrombopenia, and neutropenia was statistically significantly higher in therapy group B compared to group A.

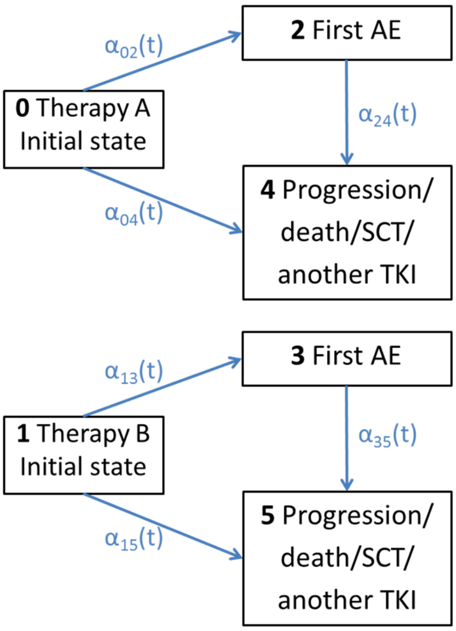
Comparing the conventional analysis and the effect of including time to first AE, competing risks, and censoring, only few differences were found: the incidence of severe adverse events and thrombopenia differed significantly between the two groups when the time-to-event approach was used. Nevertheless, the inclusion of time to event and competing risks are important for a more appropriate analysis of adverse events, and in other studies the differences could be much more pronounced. Of note, in the present study the median observation time was similar in both groups (30 months, see Chapter 4.1). In addition, the graphs of the cumulative incidence functions provide a convenient visualization of the data. For example, it can be seen that any first adverse event occurred rapidly after initiation of therapy or that severe adverse events kept occurring throughout the study. Such an observation would not have been possible in a purely "conventional" analysis.

4.4 First adverse event: Multi-state modelling

Three different adverse event types were selected to demonstrate multi-state modelling: any adverse event, severe adverse events of CTCAE grades 3 to 5, and serious adverse events (SAE) only. As already seen, an adverse event of any type had occurred in almost every patient, whereas severe adverse events and SAEs were reported less frequently. The multi-state modelling is based on the cumulative cause-specific hazards and transition probabilities between different states (see also Chapter 3.1.7). Based on such graphs, conclusions can be drawn about whether a prior adverse event has an effect on the combined endpoint of disease progression, death, stem cell transplantation, or change in TKI medication. The model used is not suitable for the study of treatment effects, as no formal testing was performed. The R package *msm* [47] can estimate multi-state models with covariates such as therapy, but these models are beyond the scope of this thesis.

4.4.1 Model and data preparation

To analyze the data, an illness-death model without recovery was used, which is shown in Figure 17.



$\alpha_{02}(t)$
2 First AE
0 Therapy A Initial state
$\alpha_{24}(t)$
$\alpha_{04}(t)$
4 Progression/ death/SCT/ another TKI
$\alpha_{13}(t)$
3 First AE
1 Therapy B Initial state
$\alpha_{35}(t)$
$\alpha_{15}(t)$
5 Progression/ death/SCT/ another TKI

Figure 17: Multi-state models for the analysis of first adverse event (AE) for therapies A (top) and B (bottom). The cause-specific hazards between two states x and y are denoted as α_{xy}. SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor. States 4 and 5 are final states where patients remain after they have reached the state.

In more detail, after start of therapy A or B (initial states 0 and 1, respectively) patients may experience an adverse event (transient states 2 and 3) or the competing composite endpoint of disease progression, death, stem cell transplantation, or TKI switch (final states 4 and 5). The states 4 and 5 can be reached without or after an adverse event. Patients remaining in the initial state were censored at their last observation date. One multi-state model was created per therapy group.

Event-specific hazards $\alpha_{xy}(t)$ were estimated using the Nelson-Aalen estimator, and probabilities of transitions between the states, $P_{xy}(t)$, are estimated using the Aalen-Johansen estimator. For the modelling the R packages `mvna` [48], `cmprsk` [49] and `etm` [50] were used.

4.4.2 Multi-state model for any adverse event

The cumulative cause-specific hazards for the first adverse event of any grade (α_{02}(t) and α_{13}(t) in Figure 17), for the composite endpoint (progression/death/SCT/TKI switch) without prior adverse event (α_{04}(t) and α_{15}(t)), and for the composite endpoint after prior adverse event (α_{24}(t) and α_{35}(t)) for both therapies are shown in the left column of Figure 18. The associated transition probabilities are shown in the right column of Figure 18. For ease of comparison, all hazards and probabilities are presented on the same scale.

The cumulative event-specific hazards for the first AE (Figure 18 top left) increased non-linearly at the beginning of treatment and were higher in therapy group B compared to group A starting at about 2 months. The corresponding probability of experiencing an AE in both groups (Figure 18 top right) increased from the start of therapy, reached its maximum after about 2 to 3 months, and then decreased again. The cumulative hazard for the composite endpoint without prior AE (Figure 18 middle left) and the corresponding transition probability (Figure 18 middle right) were very small. The cumulative hazard for the composite endpoint after prior AE (Figure 18 bottom left) was much higher than without prior AE and increased more or less linearly over time. The probability of reaching the composite endpoint after an AE increased with time, eventually reaching 1 (Figure 18 bottom right). The value of 1 (100%) basically means that any patient who had experienced their first AE after about 5 years of therapy also moved forward to the final composite endpoint state.

In summary, the hazard of experiencing an adverse event was higher in therapy group B than in A and the probability of the first AE was highest after about 2 to 3 months of therapy and then decreased. The hazard for the endpoint and the corresponding transition probability were much higher if a patient had already experienced an AE than if he or she had not.

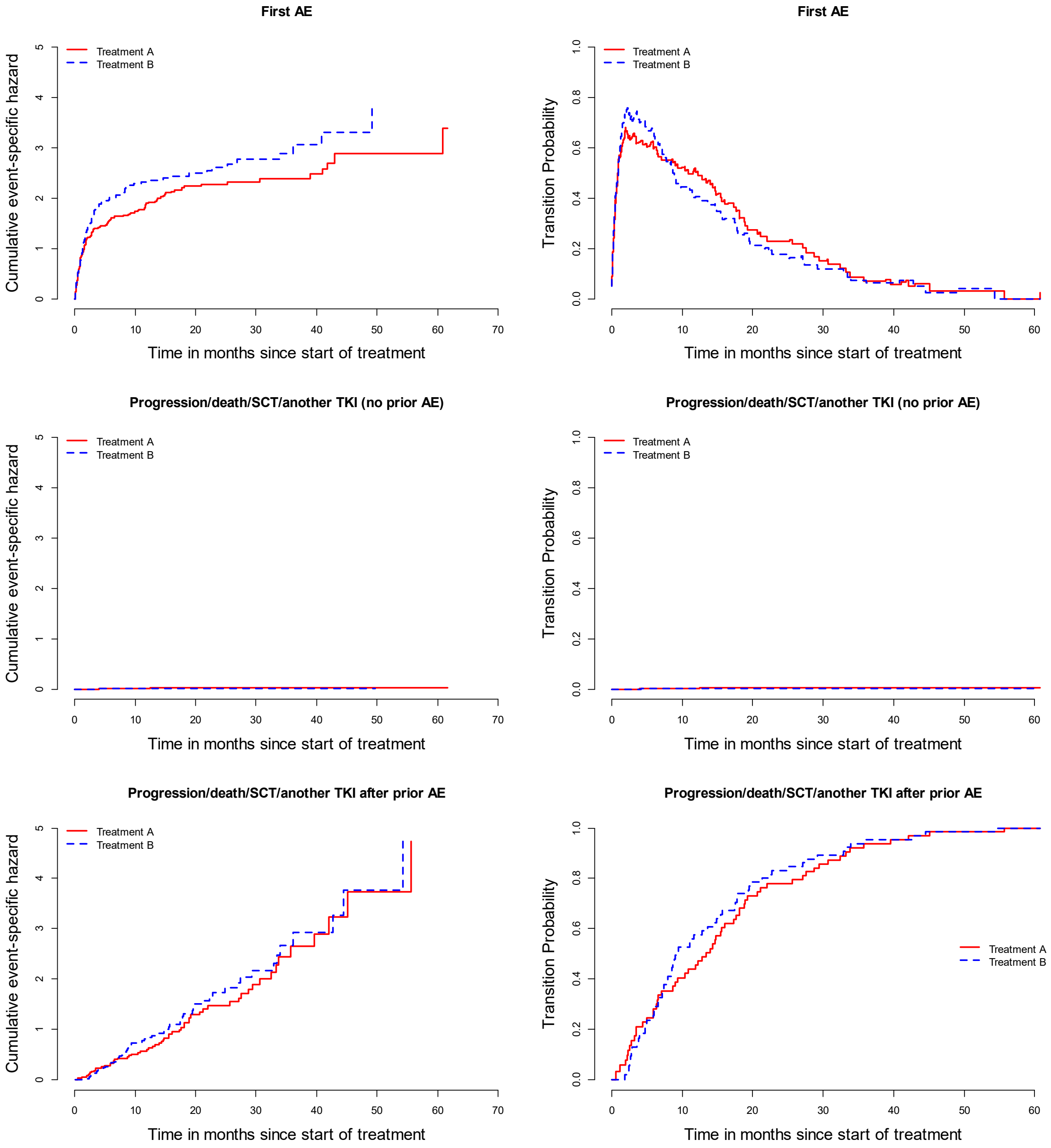


Figure 18: Cumulative event-specific hazards (left) and transition probabilities (right) between the initial state and adverse event state (top), initial state and final state (middle), and adverse event state and final state (bottom) for any adverse event in both therapy groups A and B. AE=adverse event, SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor.

4.4.3 Multi-state model for severe adverse events

The cumulative cause-specific hazards for the first severe adverse event ($\alpha_{02}(t)$ and $\alpha_{13}(t)$), for the composite endpoint without prior adverse event ($\alpha_{04}(t)$ and $\alpha_{15}(t)$), and for the composite endpoint after prior adverse event ($\alpha_{24}(t)$ and $\alpha_{35}(t)$) for both therapies are shown in the left column of Figure 19. The corresponding transition probabilities are shown in the right column of Figure 19. All hazards and all probabilities are shown on the same scale, respectively.

The cumulative hazards for the first severe AE (Figure 19 top left) and for the endpoint without prior severe AE (Figure 19 middle left) were low compared to the hazards for the endpoint after the occurrence of an AE (Figure 19 bottom left). The probability of experiencing a severe AE is generally low (less than 20%), and the highest probabilities for both therapy groups were found within the first year of therapy (Figure 19 top right). The transition probability of reaching the endpoint without prior severe AE also remained low, but increased slightly with time (Figure 19 middle right). Patients could experience progression, die, receive a stem cell transplantation, or switch TKI treatment also without severe AEs. As also with any AE, the transition probability to the final endpoint state increased with time, reaching its maximum value of 1 at approximately 4 years of therapy (Figure 19 bottom right). Over time, patients transition from the severe AE state to the final endpoint state.

The hazard and probability of severe AE remained low in both therapy groups. The hazard and probability of the endpoint after the occurrence of severe AE were much higher than in patients without severe AE.

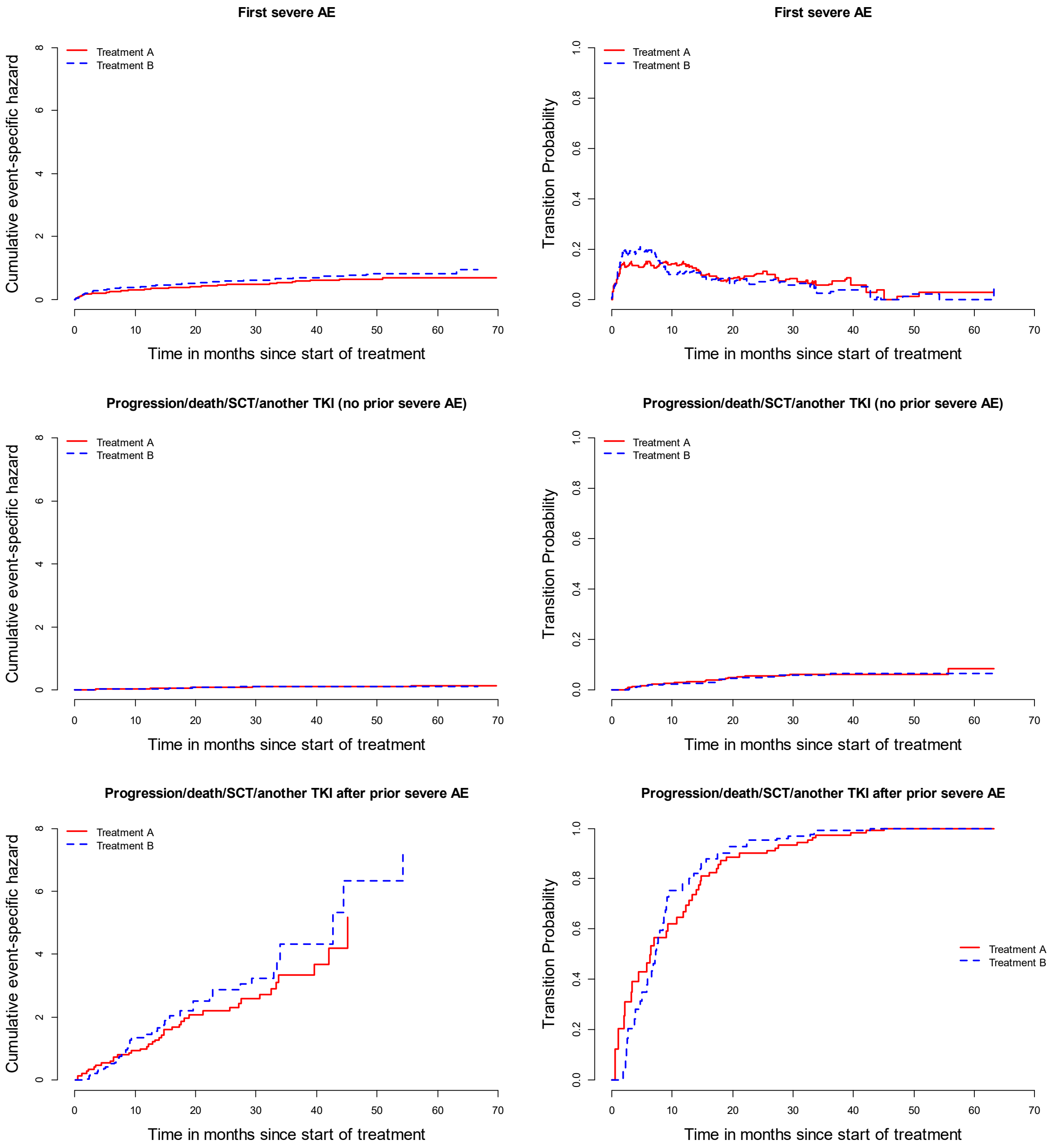


Figure 19: Cumulative event-specific hazards (left) and transition probabilities (right) between initial state and adverse event state (top), initial state and final state (middle), and adverse event state and final state (bottom) for severe adverse events in both therapy groups A and B. AE=adverse event, SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor.

4.4.4 Multi-state model for serious adverse events

The cumulative cause-specific hazards for the first serious adverse event ($\alpha_{02}(t)$ and $\alpha_{13}(t)$), for the composite endpoint without prior SAE ($\alpha_{04}(t)$ and $\alpha_{15}(t)$), and for the composite endpoint after prior SAE ($\alpha_{24}(t)$ and $\alpha_{35}(t)$) for both therapies are shown in the left column of Figure 20. The associated transition probabilities are shown in the right column of Figure 20. All hazards and all probabilities are shown on the same scale, respectively.

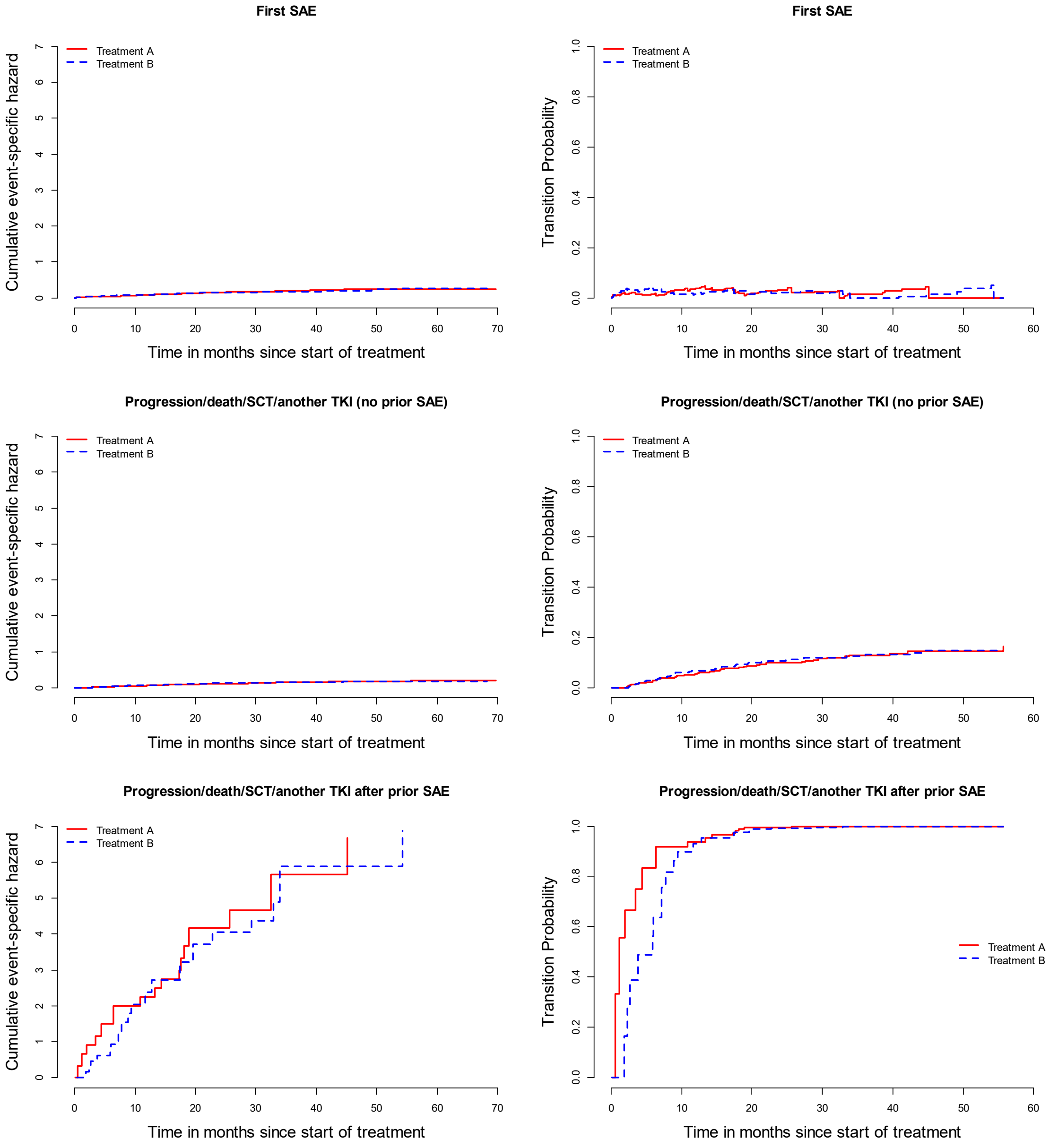


Figure 20: Cumulative event-specific hazards (left) and transition probabilities (right) between initial state and adverse event state (top), initial state and final state (middle), and adverse event state and final state (bottom) for serious adverse events (SAE) in both therapy groups A and B. SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor.

The hazard of SAE remained low (Figure 20 top left) in both therapy groups. The transition probabilities to the SAE state were also low and remained below 5%. The cumulative hazard for the endpoint after prior SAE (Figure 20 bottom left) was much higher than in patients without prior

SAE (Figure 20 middle left). In terms of probabilities, the probability of SAE was low (Figure 20 top right). The probability for the endpoint without SAE was low but increased with time (Figure 20 middle right). After the occurrence of an SAE, patients rapidly transitioned to the final state (Figure 20 bottom right).

4.4.5 Summary of multi-state modelling of adverse events

Cumulative cause-specific hazards and transition probabilities were estimated for any, for severe, and for serious adverse events. Cumulative hazards were found to decrease with increasing severity of AEs ("any AE occurs more frequently than a severe or serious event"). The probability of experiencing an adverse event was highest for any AE and continued to decrease with increasing AE severity. In addition, the highest probability for any and severe adverse events was seen after a few months of therapy. In general, the hazard of entering the final state was much higher after a prior (S)AE than without one, and once the patient experienced an AE, the probability of entering the final state of progression, death, SCT or treatment switch was also high. The more severe the AE was, the earlier the transition occurred. It may be interesting for future work to separate the composite competing risks into progression or death and SCT or treatment switching and repeat the analyses.

4.5 First adverse event: Bias assessment

The cumulative incidence of a first adverse event can be estimated using the semi-parametric Aalen-Johansen estimator, the 1-Kaplan-Meier estimator, or the parametric version of the Aalen-Johansen estimator. The Aalen-Johansen estimator is considered the gold standard because it accounts for competing risks and censoring. The Kaplan-Meier estimator cannot adequately handle competing risks, and the parametric estimator of cumulative incidence is based on incidence rates and thus on the assumption of constant hazards. In this chapter, the three estimators are compared to examine the extent of bias caused by using the Kaplan-Meier or the parametric estimator instead of the gold standard.

4.5.1 Comparison of the Aalen-Johansen estimator and the 1- Kaplan-Meier cumulative incidence function estimator

To show the importance of the gold standard, the cumulative incidence function was plotted using the Aalen-Johansen estimator and the biased 1-Kaplan Meier (KM) estimator for different amounts of competing events and censored data. In Table 5 the percentages of events (AE), competing events (progression/death/SCT/TKI switch), and censored data (no AE, no competing event) are shown for the analysis of a first AE of any grade, severe AE, and SAE.

Type of AEs	Events (%)			Competing events (%)			Censored (%)			Total (%)
	Therapy A	Therapy B	Total	Therapy A	Therapy B	Total	Therapy A	Therapy B	Total	
Any AE	46.6	45.7	92.3	0.3	0.1	0.4	4.3	2.9	7.3	100
Severe AEs	20.6	23.4	44.0	3.0	2.2	5.2	27.6	23.2	50.8	100
SAEs	8.1	8.0	16.1	6.2	5.5	11.8	36.9	35.3	72.1	100

Table 5: Distribution of events, competing events, and censoring for analysis of any AE, severe AEs and SAEs. (S)AE=(serious) adverse event.

The three following graphs (Figure 21 to Figure 23) show the cumulative incidence functions using the Aalen-Johansen estimator (left) and the 1-KM estimator (right) for the three types of AE data studied. The cumulative incidences after 3 years and 5 years of therapy are shown in Table 6. Of note, no 5-year cumulative incidence was calculated for the any AE analysis. This is because either a 5-year observation period was not reached (therapy B) or too few patients were still under observation to obtain reliable estimates at this time point (therapy A). The differences of the point estimates of the cumulative incidence functions using both estimators are shown in Table 7. This is referred to as “magnitude of bias”.

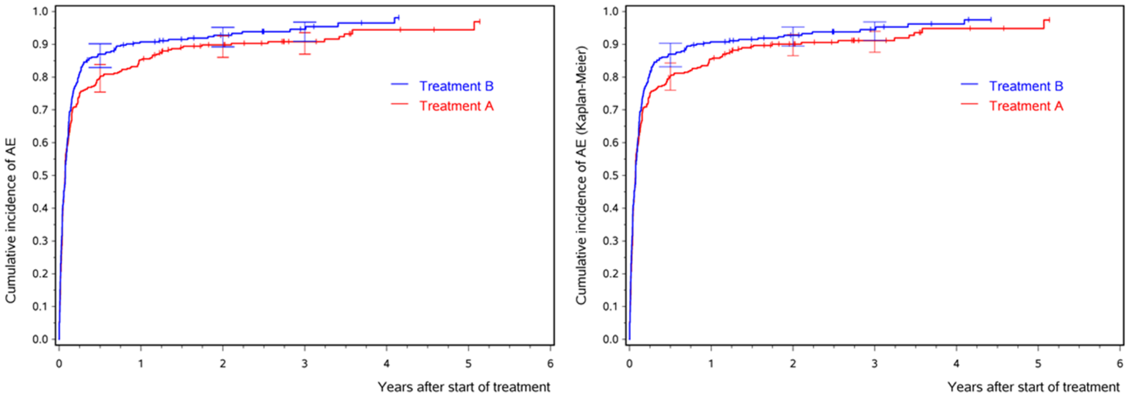


Figure 21: Cumulative incidences of adverse events (AE) of any severity using the Aalen-Johansen estimator (left) and the 1 - Kaplan-Meier estimator (right).

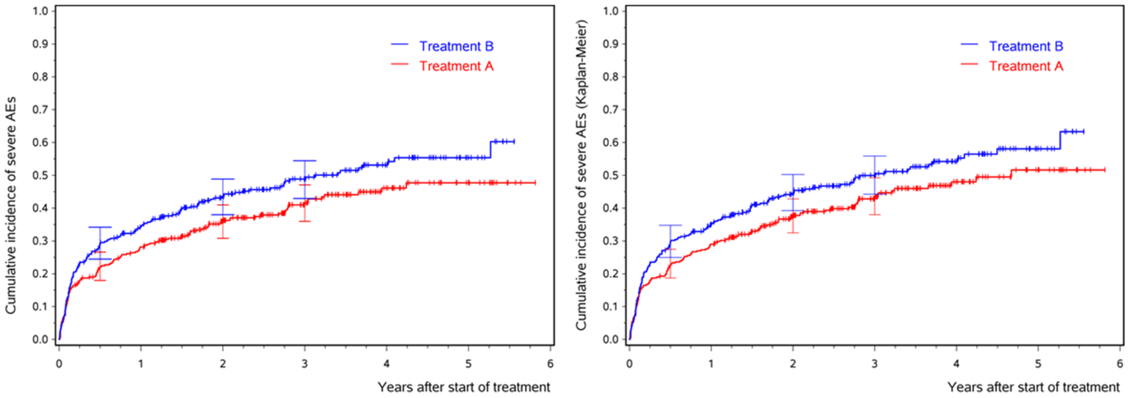


Figure 22: Cumulative incidences of severe adverse events (AE) of grades 3 to 5 using the Aalen-Johansen estimator (left) and the 1 - Kaplan-Meier estimator (right).

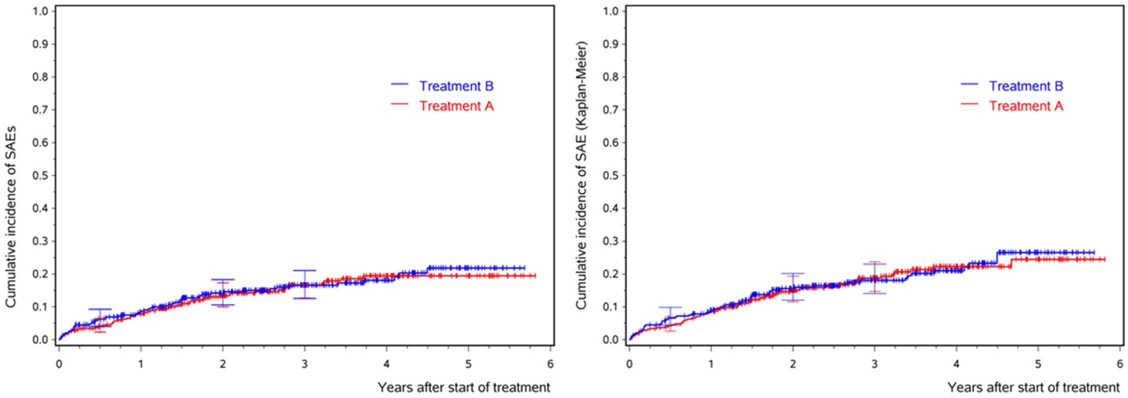


Figure 23: Cumulative incidences of serious adverse events (SAE) using the Aalen-Johansen estimator (left) and the 1 - Kaplan-Meier estimator (right).

Type of AEs	**CIF (Aalen-Johansen)**				**CIF (1-KM)**			
	At 3 years		**At 5 years**		**At 3 years**		**At 5 years**	
	Therapy A	Therapy B	Therapy A	Therapy B	Therapy A	Therapy B	Therapy A	Therapy B
Any AE	0.908 [0.870, 0.935]	0.946 [0.909, 0.968]	NA	NA	0.911 [0.876, 0.940]	0.945 [0.912, 0.968]	NA	NA
Severe AEs	0.415 [0.359, 0.470]	0.488 [0.429, 0.545]	0.477 [0.408, 0.543]	0.553 [0.485, 0.617]	0.434 [0.380, 0.493]	0.499 [0.443, 0.559]	0.516 [0.442, 0.595]	0.580 [0.510, 0.652]
SAEs	0.166 [0.127, 0.211]	0.166 [0.126, 0.211]	0.195 [0.148, 0.246]	0.218 [0.161, 0.281]	0.187 [0.147, 0.237]	0.181 [0.141, 0.231]	0.245 [0.186, 0.318]	0.265 [0.201, 0.345]

Table 6: Cumulative incidence of all adverse events (AE), severe AEs, and serious AE (SAE) 3 and 5 years after initiation of therapy for both therapy groups A and B. Shown are the point estimates together with their 95% confidence intervals (in parentheses). CIF=cumulative incidence function, KM=Kaplan-Meier.

Magnitude of bias: Difference in CIFs (1-KM – Aalen-Johansen)				
Type of AEs	**At 3 years**		**At 5 years**	
	Therapy A	Therapy B	Therapy A	Therapy B
Any AE	0.3%	-0.1%	NA	NA
Severe AEs	1.9%	1.1%	3.9%	2.7%
SAEs	2.1%	1.5%	5.0%	4.7%

Table 7: Differences in cumulative incidence functions (CIFs) of any adverse event (AE), severe AEs, and serious AE (SAE) 3 and 5 years after initiation of therapy for both therapy groups A and B. KM=Kaplan-Meier.

The 3-year cumulative incidences for any adverse event for the Aalen-Johansen estimators in therapy groups A and B were 90.8% and 94.6%, respectively. The corresponding cumulative incidences using the Kaplan-Meier estimator were 91.1% and 94.5%, respectively. In this case, hardly any differences were seen (CIF-differences were 0.3% and -0.1%). However, the difference of -0.1% indicates a higher value for the Aalen-Johansen estimator than for the Kaplan-Meier estimator. The small difference is likely due to a rounding error of the program routine used to calculate the cumulative incidences. Based on the formulae, the values of the Aalen-Johansen estimator should not be higher than the values of the 1-Kaplan-Meier estimator.

When considering severe AEs and the Aalen-Johansen estimator, the 3-year cumulative incidences for therapy groups A and B were 41.5% and 48.8%, and the 5-year cumulative incidences were 47.7% and 55.3%, respectively. Using the Kaplan-Meier estimator, the 3-year cumulative incidences in therapy groups A and B were 43.4% and 49.9%, respectively, and the 5-year cumulative incidences were 51.6% and 58.0%, respectively. Thus, at 3 years, the Kaplan-Meier estimator yielded 1.9% and 1.1% higher estimates for therapies A and B, respectively. After 5 years, the differences were 3.9% and 2.7%, respectively.

For SAEs alone, the 3-year Aalen-Johansen estimates for therapy groups A and B were 16.6% each. Five years after therapy initiation, the corresponding estimates were 19.5% and 21.8%. Looking at the Kaplan-Meier estimates, cumulative incidences after 3 years were 18.7% and 18.1% for therapies A and B, respectively. After another 2 years, the estimates were 24.5% and 26.5%, respectively. Thus, in sum, the cumulative incidences estimated by the Kaplan-Meier method were 2.1% and 1.5% higher than the Aalen-Johansen estimates in therapy groups A and B, respectively, 3 years after therapy initiation. At 5 years, the corresponding differences were 5.0% and 4.7%, respectively.

In summary, as expected from the theoretical formulae, the Kaplan-Meier estimator overestimated the cumulative incidence of events when competing risks were present. In the present work, with many events, very few competing risks, and only a small amount of censoring (any AE scenario), both estimators produced essentially the same estimates. As the rate of competing events and censoring increased, more pronounced differences were found between the two estimators. The greatest difference between the Aalen-Johansen and Kaplan-Meier estimates was 5%, which was found for SAEs.

4.5.2 Comparison of the non-parametric Aalen-Johansen estimator and the parametric incidence rate based cumulative incidence function estimator

As suggested by Grambauer *et al.* [25], a parametric approach of the cumulative incidence function based on incidence rates and thus constant hazards over time was calculated for the first AE and the composite endpoint progression/death/SCT/TKI switch without prior AE. This parametric estimator is presented together with the nonparametric Aalen-Johansen estimator to examine the performance of both estimators. Figure 24 to Figure 27 show both estimators for any AE, severe AEs and SAEs only.

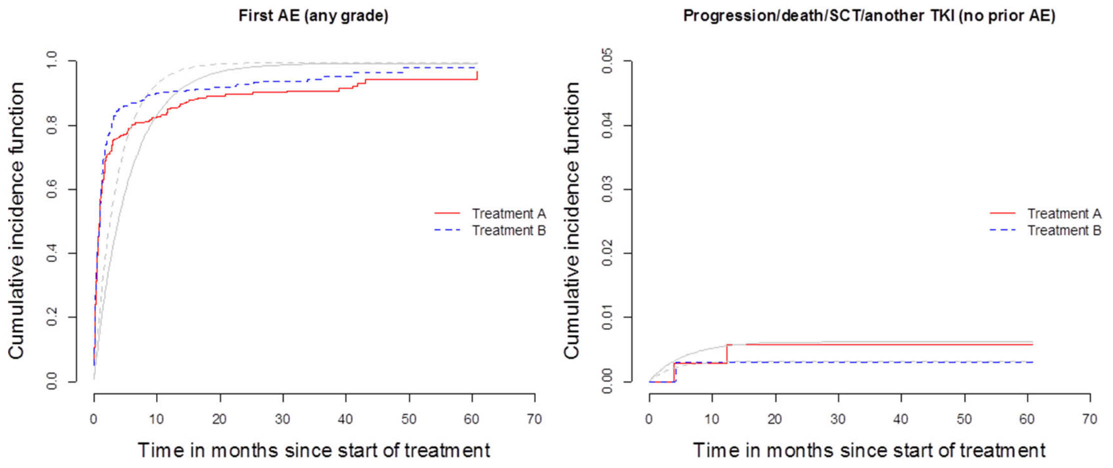


Figure 24: Aalen-Johansen estimator of cumulative incidence function (CIF) and parametric CIF estimator based on constant hazards for the first adverse event (AE) of any grade (left) and for the composite endpoint of progression/death/SCT/another TKI (right). The red solid line and blue dashed line show the Aalen-Johansen estimators for treatments A and B, respectively, and the grey solid line and grey dashed line show the parametric CIF estimates for therapy groups A and B, respectively. SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor.

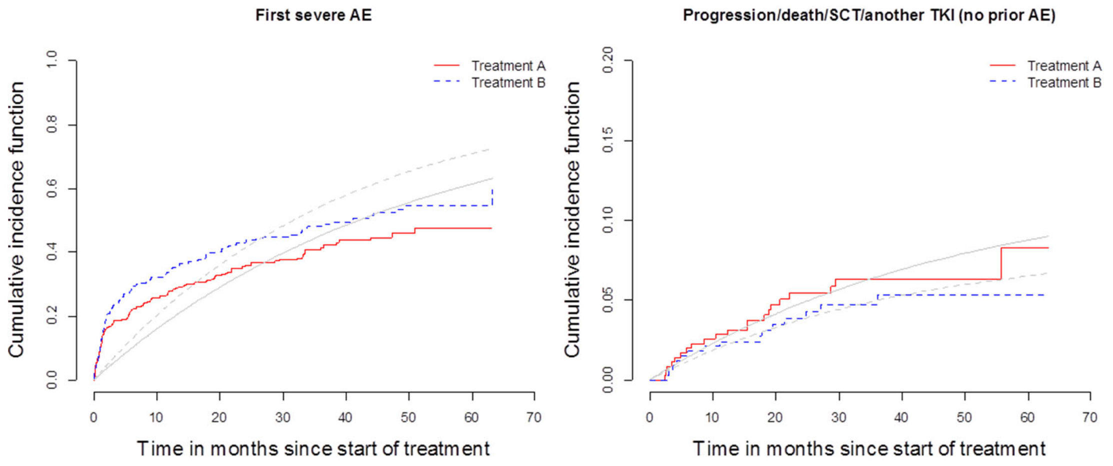


Figure 25: Aalen-Johansen estimator of cumulative incidence function (CIF) and parametric CIF estimator based on constant hazards for the first severe adverse event (AE) (left) and for the composite endpoint of progression/death/SCT/another TKI (right). The red solid line and blue dashed line show the Aalen-Johansen estimators for treatments A and B, respectively, and the grey solid and grey dashed line show the parametric CIF estimates for therapy groups A and B, respectively. SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor.

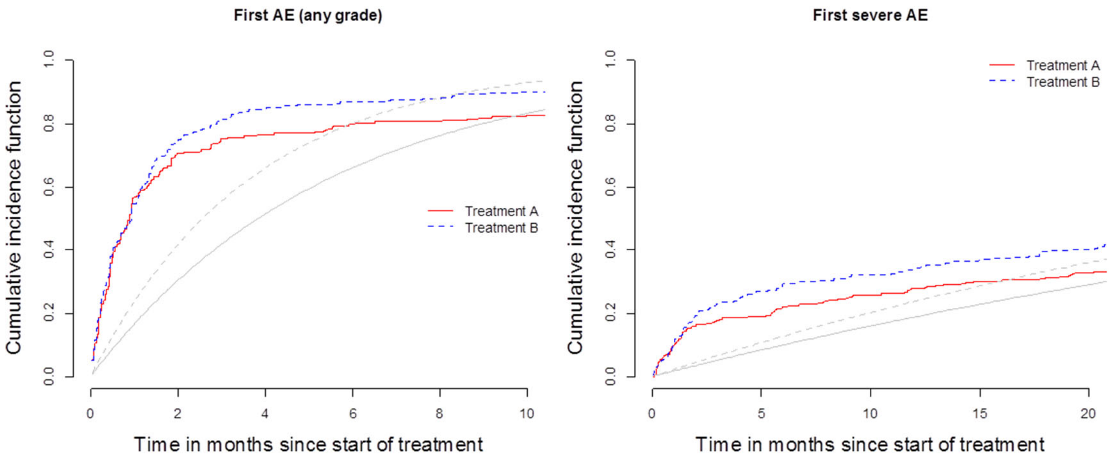


Figure 26: A closer look at the cumulative incidence functions (CIF) for the first AE of any grade (Figure 24) during the first 10 months (left) and the first severe AE (Figure 25) during the first 20 months (right) with the Aalen-Johansen estimator of the CIF and the parametric CIF estimator based on constant hazards. The red solid line and blue dashed line show the Aalen-Johansen estimators for treatments A and B, and the grey solid line and grey dashed line show the parametric CIF estimates for therapy groups A and B, respectively.

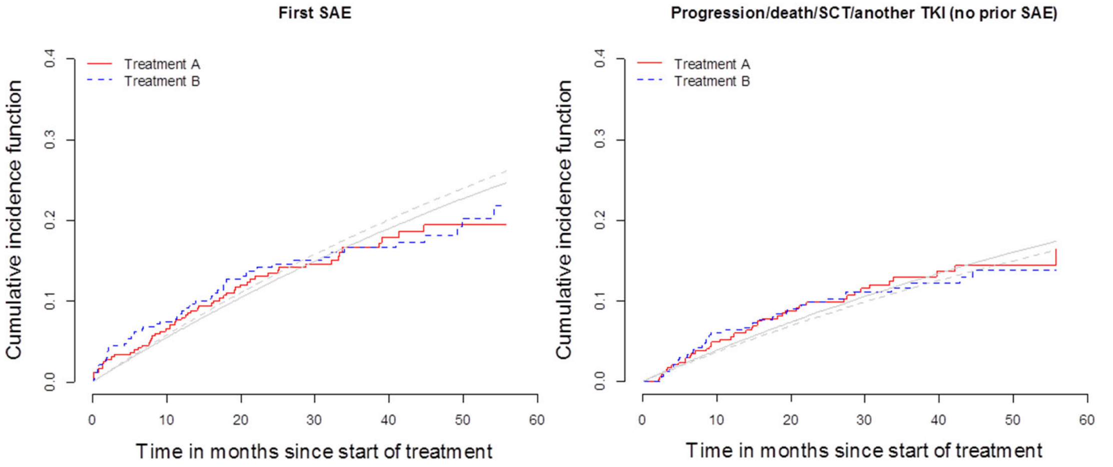


Figure 27: Aalen-Johansen estimator of the cumulative incidence function (CIF) and parametric CIF estimator based on constant hazards for the first serious adverse event (SAE) (left) and for the composite endpoint of progression/death/SCT/another TKI (right). The red solid line and blue dashed line show the Aalen-Johansen estimators for treatments A and B, respectively. The grey solid line and grey dashed line show the parametric CIF estimates for therapy groups A and B, respectively. SCT=stem cell transplantation, TKI=tyrosine kinase inhibitor.

In the case of any AE (Figure 24, left), the parametric CIF seriously underestimated the cumulative incidence of AEs during the first months of therapy and slightly overestimated it thereafter. Figure 26 (left graph) shows a closer look at the cumulative incidence curves for the first AE during the first 10 months of therapy. During the first months the differences between the two estimators were particularly large: Two months after start of therapy, the Aalen-Johansen estimator showed AE probabilities of 71% and 75% for therapy groups A and B, respectively, whereas the parametric estimator yielded values of 30% and 42%. The parametric estimator should not be used to estimate the probability of any AE. With respect to the composite endpoint, both estimators estimated essentially the same (Figure 24, right). This is likely due to the fact that only 2 and 1 competing risks to AE were reported in therapy groups A and B, respectively. Indeed, as shown in Figure 18, the cumulative hazard for the first AE (top left graph) does not follow straight line, which would correspond to a constant hazard. The hazard for the endpoint (Figure 18, middle left graph) is very low, almost zero, which goes along with the constant hazard assumption.

For severe adverse events (Figure 25, left) the parametric CIF estimator again underestimated the AE incidence at the beginning of the therapy, but not as much as in case of any AE. The cumulative incidence curves up to 20 months after therapy are shown in Figure 26 (right graph). For example, 3 months after the start of therapy, the Aalen-Johansen estimator showed AE probabilities of 18% and 23% in therapy groups A and B, respectively, whereas the parametric estimator yielded the values of 5% and 7%, respectively. These differences were still large so that the use of the parametric estimator of the cumulative incidence function for severe adverse events cannot be recommended in the present study. When considering the composite endpoint (Figure 25, right), both estimators appear to perform similarly. At the time of the last event (after 56 months in therapy group A and 37 months in therapy group B), both estimates were very similar. However, it is not recommended to extrapolate the parametric estimator beyond the last event time [23]. The underlying cumulative hazards for the AE and endpoint are shown in Figure 19 (top left graph and middle left graph, respectively).

In the SAE analysis, the cumulative incidences of the first SAE (Figure 27, left) were (very) slightly underestimated during the first 3 years of therapy and slightly overestimated thereafter. When considering the incidences of the competing events (Figure 27, right), the parametric estimator slightly underestimated the incidences, but at the time of the last event, both estimators estimated the same. The cumulative hazards for both event types are shown in Figure 20 (top left graph for SAE, middle left graph for endpoint).

In summary, in the present work, the incidence rate-based parametric estimator of the cumulative incidence function tended to underestimate the incidences of adverse events at the beginning of the therapy and overestimated it over time. In contrast, the incidences of the composite endpoint were estimated similarly with both estimators. This may be different in other studies with different hazard profiles. Based on the results for the competing risks, the assumption of constant hazards might be true, but this was not the case for adverse events. It is recommended to plot the cumulative hazards and use the nonparametric Aalen-Johansen estimator to estimate the cumulative incidences.

4.5.3 Summary of the bias assessment of analysis of first adverse event

For assessing the amount of bias caused by estimating the cumulative incidence function of the time to a first adverse event using methods other than the gold standard of the Aalen-Johansen estimator, the following two biased estimators were evaluated: the 1 – Kaplan-Meier estimator and the parametric incidence rate-based estimator. Both estimators were compared against the gold standard. The Kaplan-Meier estimator tended to overestimate the cumulative incidences of adverse events up to 5% after 5 years of therapy compared to the gold standard. As the amount of competing risks and censoring increased, the differences between the two estimators increased. Both estimators were appropriate for the “any adverse events scenario”, in which the majority of patients had an adverse event and very few competing risks occurred. For severe and serious adverse events, the Aalen-Johansen estimator would be preferable. When comparing the Aalen-Johansen estimator and the parametric cumulative incidence estimator, the parametric estimator tended to underestimate the incidence of adverse events after initiation of therapy and overestimate it thereafter. The incidence of competing events was similar using both estimators.

For the “any adverse event scenario”, for which the constant hazard assumption does not hold, the differences were large, making the parametric estimator unsuitable for the analysis. For severe adverse events and SAEs, the differences decreased, but judged clinically, they could still be too large. The Aalen-Johansen estimator remains the best option for estimating cumulative incidences because no constant hazard assumption is required and competing risks are accounted for.

4.6 Recurrent adverse events analysis

So far, only the first AE was considered. While it might be sufficient to analyze only the first serious adverse event, the picture may change when analyzing events that recur frequently. One might hypothesize that for SAEs that do not occur so frequently, analyzing only the first SAE might be sufficient to capture the AE pattern. Such analysis decisions need to be done by a team of experts and are or not further detailed here.

In the following, recurrent events analysis is performed and compared with the results of the analysis of the first event. The same AEs were chosen that were analyzed for the first adverse event: Fatigue, thrombopenia, and neutropenia. In addition, any severe adverse event and any serious adverse event (SAE) were analyzed. It was assumed that the recurrence of any AE would not be of great interest, so such a scenario was not considered further. The majority of "any AEs" are mild AEs that are likely to be common (headache, runny nose, etc.) even if unrelated to the study drugs. For this reason, the focus was laid to the recurrence of severe or serious AE. This is, of course, debatable, depending on the clinical question underlying the analysis, but for the present thesis and to demonstrate the approach of recurrent events analysis, these two composite endpoints were also considered.

4.6.1 Data preparation

Data for recurrent event analysis was prepared as follows: Of single AEs coded with the same MedDRA Preferred Term (*e.g.* thrombopenia, 10035528), only those events separated by an event-free episode of at least one day from the previous event were treated as recurrent events. Otherwise, *i.e.* if only the severity grade of the AE had changed, the AE was considered ongoing and not a “new” event.

For AEs with a reported end (the question “ongoing” was answered with “no”) but without a reported end date, an artificial duration of one day was set. All AEs reported as ongoing without an end date were treated as ongoing until the last observation date of the corresponding patient and then censored.

To estimate the effect of treatment on AEs, the competing risks of progression/death/stem cell transplantation/TKI switch were included in the analysis. Thus, all observations from the competing event onward were censored, and all AEs occurring before censoring time were analyzed. All AEs occurring after censoring were thus neglected. When an AE and a competing event occurred on the same day, which was occasionally the case for SAEs, the competing event was treated as it would have had occurred one day after the AE. This allowed the corresponding AE to be included in the analysis. This seemed plausible and preserved the natural order of events: most likely, the (serious) adverse event was reported and the corresponding action occurred on the same day. Of note, a change in treatment was the most common competing risk to the AEs in this work.

4.6.2 Recurrent severe adverse events

Table 8 shows the number of severe adverse events reported (middle column) and the episodes reported before a competing event occurred (right column). Due to censoring from the occurrence of a competing event, AEs from the right column of Table 8 were included in the analysis. The highest possible number that still allowed convergence of the models was analyzed, which was the case for the first 4 severe adverse events.

Number of reported severe adverse event occurrences	Total frequency	Frequency prior competing event
1	313	299
2	94	79
3	40	31
4	13	11
5	1	0

Table 8: Number of reported occurrences of severe adverse events. All events separated by an event-free episode of at least one day after the previous event were treated as recurrent events.

The results using different models for recurrent events are shown in Table 9 (assuming a common baseline hazard for all subsequent AEs) and Table 10 (assuming that the baseline hazard differs between subsequent AEs). For better comparability, also the crude relative risk estimate (see Chapter 4.2, Table 4) and the Cox model estimate, both of which apply only to the first AE only, are shown. The AG model refers to the AG model with robust (sandwich) variance estimate. The WLW model has weights included in the results (last four rows of Table 10). These weights were used to calculate a mean hazard ratio (shown in the last row of Table 9). The highest weight (74%) was assigned to the first AE, the second AE contributes 18%, the third 6%, and the fourth AE 2% to the mean effect.

Applicability	Model	HR* (B vs. A)	95% CI	p value
First AE only	Crude RR	1.172	0.994 – 1.382	n.s.
	Conventional Cox model	1.349	1.113 – 1.635	**<0.05**
Recurrent AEs	AG model	1.350	1.071 – 1.704	**<0.05**
	PWP total time model with common effects	1.260	1.037 – 1.531	**<0.05**
	PWP gap-time model with common effects	1.258	1.036 – 1.529	**<0.05**
	WLW model (mean effect)	1.312	1.076 – 1.600	**<0.05**

Table 9: Estimates of the different models for severe adverse events (AE). Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *For the "Crude RR" model, relative risk (RR) is shown instead of HR. AG=Andersen and Gill, PWP=Prentice, Williams and Peterson, WLW = Wei, Lin and Weissfeld.

Model	AE	HR (B vs. A)	95% CI	p value	Weight*
PWP total time model	1.	1.255	1.000 – 1.576	**<0.05**	
	2.	1.339	0.857 – 2.092	n.s.	
	3.	1.344	0.614 – 2.944	n.s.	
	4.	0.520	0.111 – 2.450	n.s.	
PWP gap-time model	1.	1.576	1.000 – 1.576	**<0.05**	
	2.	2.044	0.839 – 12.044	n.s.	
	3.	3.078	0.650 – 3.078	n.s.	
	4.	2.226	0.155 – 2.226	n.s.	
WLW model	1.	1.255	1.001 – 1.575	**<0.05**	0.744
	2.	1.258	0.809 – 1.956	n.s.	0.178
	3.	2.330	1.088 – 4.989	**<0.05**	0.056
	4.	1.149	0.317 – 4.168	n.s.	0.021

Table 10: Estimates of the different models for the first, second, third, and fourth severe adverse event (AE). Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *Weight applies only to the WLW model and indicates how much weight is given to the corresponding AE in the estimation of the mean effect shown in Table 9. PWP=Prentice, Williams and Peterson, WLW =Wei, Lin and Weissfeld.

Recurrent events models assuming the same hazard for the first and all subsequent events all showed similar and statistically significantly increased hazard of severe AE for therapy group B compared with group A (Table 9). Assuming that all AEs could have a different risk of occurrence, a therapy effect was found for the first severe AE with statistically significantly increased hazard in therapy group B compared with group A (Table 10). However, the p values were just below the significance threshold of 0.05. It seems more intuitive to allow the hazards to vary, as in this case all adverse events of CTCAE grades 3 to 5 were included regardless of the exact type of the AEs (the severe AE could be thrombopenia, diarrhea, or any other grade 3 to 5 AE). Interestingly, the WLW model showed a significant therapy effect on the occurrence of the third severe AE. This may be because this model allows a patient to be at risk for all recurrent events, even if no AEs were previously observed. For example, a patient may be at risk for the third AE even if he or she did not experience the first AE. The conditional risk set of the PWP model does not allow for this. Thus, this difference with regard to the third severe AE is likely due to the somewhat problematic risk set of the WLW model and should not be given too much attention.

Mean frequency of severe adverse events

The mean frequency of severe adverse events in the two therapy groups A and B over time can be seen in Figure 28. It is presented as a mean cumulative function and interpreted as the mean cumulative number of adverse events per patient (and therapy group) up to a certain time *t* after the start of treatment. When competing events occurred, patients were censored for the mean cumulative function. Note that the censored observations are not indicated by small vertical lines as in the Kaplan-Meier/Aalen-Johansen estimator curves. Please also note that all mean cumulative frequency plots show therapy group A in blue and group B in red. The colors are reversed compared to all other plots in this thesis. The strict structure of the plot of mean cumulative function in SAS software does not seem to support custom plot colors per group.

A statistically significantly higher frequency of severe adverse events was found for therapy group B compared to group A (p<0.05).

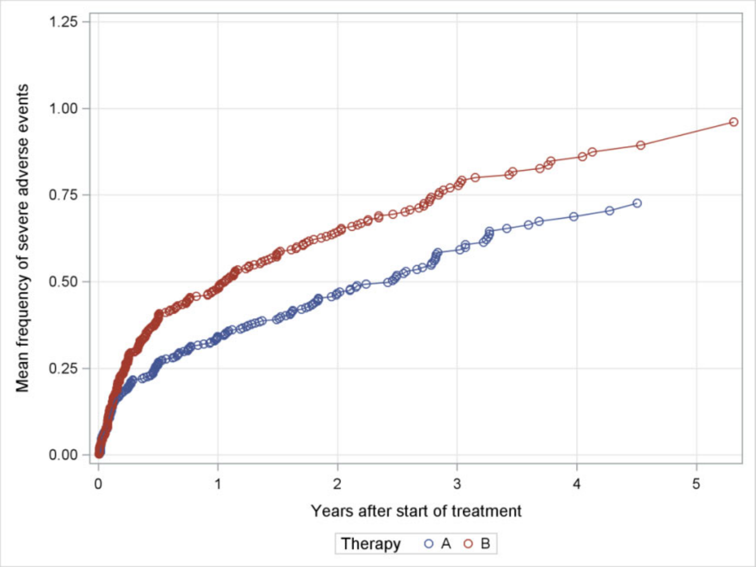


Figure 28: Mean frequency of severe adverse events for therapy groups A (blue line) and B (red line). P-value for the test of equality of the two mean cumulative functions: p<0.05.

4.6.3 Recurrent serious adverse events

During the TIGER study, a total of 188 SAEs were reported after therapy initiation (through the data cutoff date of December 14th, 2018). Table 11 shows the number of reported occurrences of severe adverse events. The first two SAEs per patient were analyzed.

Number of reported SAE occurrences	Total frequency	Frequency prior competing event
1	121	110
2	24	20
3	7	5
4	4	3
5	1	1

Table 11: Number of reported occurrences of serious adverse events (SAE). All events separated by an event-free episode of at least one day after the previous event were treated as recurrent events.

The results of the different models for recurrent events are shown in Table 12 (assuming a common baseline hazard for all subsequent AEs) and Table 13 (allowing the baseline hazard to differ between subsequent AEs). For ease of comparison, the crude relative risk estimate and the Cox model estimate are also shown, both for the first AE only.

Applicability	Model	HR* (B vs. A)	95% CI	p value
First SAE only	Crude RR	1.051	0.761 – 1.450	n.s.
	Conventional Cox model	1.058	0.749 – 1.495	n.s.
Recurrent SAEs	AG model	0.994	0.646 – 1.377	n.s.
	PWP total time model with common effects	1.059	0.750 – 1.496	n.s.
	PWP gap-time model with common effects	1.058	0.749 – 1.495	n.s.
	WLW model (mean effect)	0.996	0.683 – 1.367	n.s.

Table 12: Estimates of the different models for serious adverse events (SAE). Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *For the "Crude RR" model, relative risk (RR) is shown instead of HR. AG=Andersen and Gill, PWP=Prentice, Williams and Peterson, WLW = Wei, Lin and Weissfeld.

Model	SAE	HR (B vs. A)	95% CI	p value	Weight*
PWP total time model	1.	1.026	0.706 – 1.491	n.s.	
	2.	0.578	0.228 – 1.467	n.s.	
PWP gap-time model	1.	1.026	0.706 – 1.491	n.s.	
	2.	0.583	0.230 – 1.480	n.s.	
WLW model	1.	1.026	0.707 – 1.489	n.s.	0.860
	2.	0.598	0.242 – 1.479	n.s.	0.140

Table 13: Estimates of the different models for the first and second serious adverse event (SAE). Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *Weights apply only to the WLW model and indicate how much weight is given to the corresponding SAE in the estimation of the mean effect reported in Table 12. PWP=Prentice, Williams and Peterson, WLW =Wei, Lin and Weissfeld.

Regardless of the assumption of equal or unequal risks for each SAE, no therapy effect on the occurrence of SAEs was found (Table 12 and Table 13). When considering the estimates for the occurrence of the second SAE (Table 13), a lower hazard was observed for therapy group B compared to group A (HR=0.6 for all models). However, this observation was also not statistically significant.

Mean frequency of SAEs

The mean frequency of SAEs in the two treatment groups A and B over time is shown in Figure 29. There were no clear differences between the two therapy groups, as also shown by the non-significant test result of equality of the two mean frequency functions (p=n.s.). SAEs seem to occur with a more or less constant rate in both therapy groups. This assumption is consistent with a linear increase in mean SAE frequency over time.

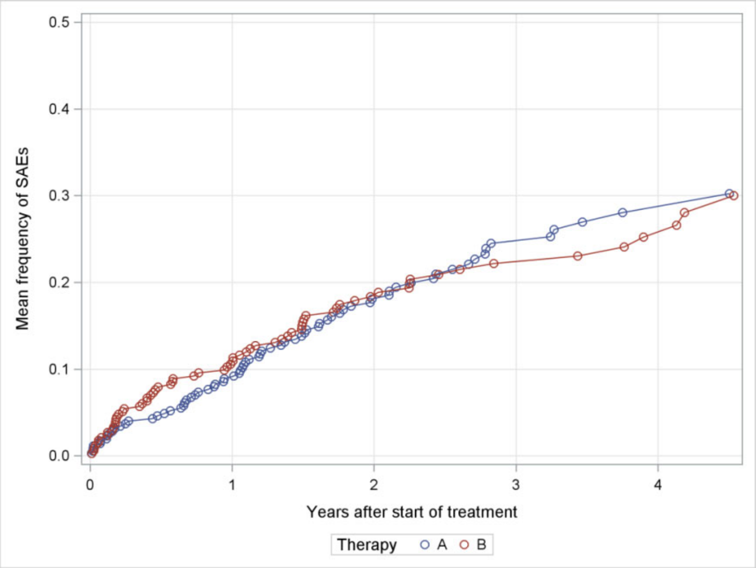


Figure 29: Mean frequency of serious adverse events (SAEs) for therapy groups A (blue line) and B (red line). P-value for the test of equality of the two mean cumulative functions: p=n.s..

4.6.4 Recurrent fatigue

Table 14 shows the number of fatigue episodes reported. The first two episodes of fatigue were included in the analysis of recurrent events.

Number of reported episodes of fatigue	Total frequency	Frequency prior competing event
1	235	228
2	33	32
3	6	6
4	2	2

Table 14: Number of reported fatigue episodes. All events separated by an event-free episode of at least one day from the previous event were treated as recurrent events.

The results using different models for recurrent events are shown in Table 15 (assuming a common baseline hazard for all subsequent AEs) and Table 16 (allowing the baseline hazard to differ between subsequent AEs). For ease of comparison, also the crude relative risk estimate and the Cox model estimate, both applying to first AE only, are shown.

Applicability	Model	HR* (B vs. A)	95% CI	p value
First AE only	Crude RR	1.234	1.003 – 1.518	**<0.05**
	Conventional Cox model	1.427	1.121 – 1.818	**<0.05**
Recurrent AEs	AG model	1.362	1.050 – 1.764	**<0.05**
	PWP total time model with common effects	1.332	1.042 – 1.704	**<0.05**
	PWP gap-time model with common effects	1.320	1.033 – 1.686	**<0.05**
	WLW model (mean effect)	1.353	1.058 – 1.729	**<0.05**

Table 15: Estimates of the different models for recurrent fatigue. Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *For the “Crude RR” model, relative risk (RR) is shown instead of HR. AE=adverse event, AG=Andersen and Gill, PWP=Prentice, Williams and Peterson, WLW = Wei, Lin and Weissfeld.

Model	AE	HR (B vs. A)	95% CI	p value	Weight*
PWP total time model	1.	1.344	1.036 – 1.745	**<0.05**	
	2.	1.243	0.600 – 2.575	n.s.	
PWP gap-time model	1.	1.344	1.036 – 1.745	**<0.05**	
	2.	1.156	0.570 – 2.342	n.s.	
WLW model	1.	1.344	1.036 – 1.743	**<0.05**	0.887
	2.	1.420	0.698 – 2.891	n.s.	0.113

Table 16: Estimates of the different models for the first and second reported episode of fatigue. Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p-values. *Weights apply only to the WLW model and indicate how much weight is given to the corresponding adverse event (AE) in estimating the mean effect shown in Table 15. PWP=Prentice, Williams and Peterson, WLW =Wei, Lin and Weissfeld.

A statistically significant therapy effect for recurrent fatigue was found with models that assume the same hazard for the first and all subsequent AEs (Table 15): The hazard for fatigue was increased in therapy group B compared to group A (HR=1.3 to 1.4). The Cox model considering only the first episode of fatigue showed a similar hazard ratio (HR=1.4), and also the 95% confidence intervals appeared similar. When the first and second episodes of fatigue were considered separately (Table 16), the results for the first reported fatigue episode remained similar to the common baseline hazard models, and no therapy effect was found on the second fatigue episode.

Mean frequency of fatigue

The mean frequency of fatigue in the two therapy groups A and B is shown in Figure 30. Fatigue was reported more frequently in group B than in group A, and this result was statistically significant (p-value for the test of equality of the two mean cumulative functions: $p<0.05$). An outlier in therapy group A can be seen about 6 years after the start of therapy. This is due to the fact that only a very small number of patients were observed in the study 6 years after the start of therapy and a single event can lead to extreme results.

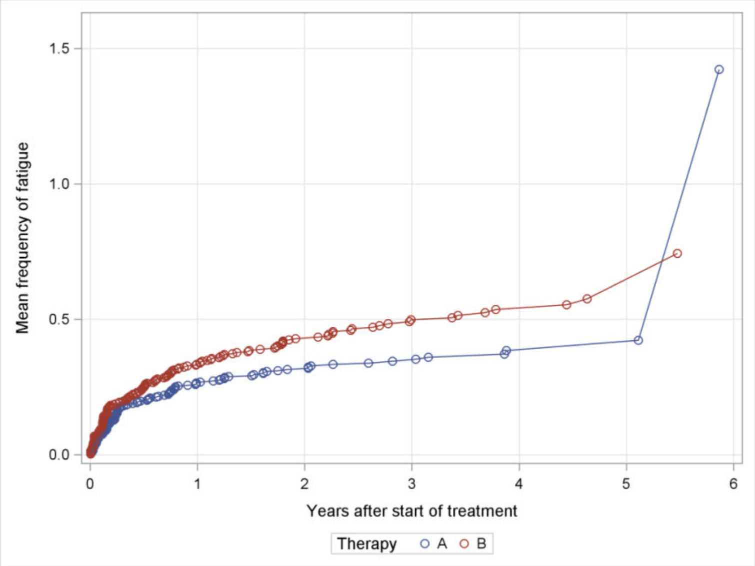


Figure 30: Mean frequency of fatigue for therapy groups A (blue line) and B (red line). P-value for the test of equality of the two mean cumulative functions: $p<0.05$.

4.6.5 Recurrent thrombopenia

Table 17 shows the number of reported episodes of thrombopenia. The first 3 episodes of thrombopenia were included in the recurrent events analysis.

Number of reported episodes of thrombopenia	Total frequency	Frequency prior competing event
1	106	103
2	33	30
3	8	6
4	6	3
5	3	2
6	3	2
7	1	2
8	1	0

Table 17: Numbers of reported episodes of thrombopenia. All events separated by an event-free episode of at least one day from the previous event were treated as recurrent events.

The results using different models for recurrent events are shown in Table 18 (assuming a common baseline hazard for all subsequent AEs) and Table 19 (allowing the baseline hazard to differ between subsequent AEs). For ease of comparison, the crude relative risk estimate and the Cox model estimate are also shown, both applying to the first AE only.

Applicability	Model	HR* (B vs. A)	95% CI	p value
First AE only	Crude RR	1.424	0.999 – 2.031	n.s.
	Conventional Cox model	1.641	1.178 – 2.288	**<0.05**
Recurrent AEs	Robust AG model	1.651	1.088 – 2.506	**<0.05**
	PWP total time model with common effects	1.696	1.205 – 2.387	**<0.05**
	PWP gap-time model with common effects	1.580	1.124 – 2.222	**<0.05**
	WLW model (mean effect)	1.473	1.036 – 2.063	**<0.05**

Table 18: Estimates of the different models for thrombopenia. Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *For the "Crude RR" model, relative risk (RR) is shown instead of HR. AE=adverse event, AG=Andersen and Gill, PWP=Prentice, Williams and Peterson, WLW = Wei, Lin and Weissfeld.

Model	AE	HR (B vs. A)	95% CI	p value	Weight*
PWP total time model	1.	1.493	1.009 – 2.209	**<0.05**	
	2.	2.467	1.172 – 5.194	**<0.05**	
	3.	2.927	0.229 – 28.644	n.s.	
PWP gap-time model	1.	1.493	1.009 – 2.209	**<0.05**	
	2.	1.818	0.863 – 3.829	n.s.	
	3.	2.491	0.284 – 21.824	n.s.	
WLW model	1.	1.493	1.009 – 2.210	**<0.05**	0.768
	2.	1.268	0.603 – 2.666	n.s.	0.209
	3.	2.643	0.314 – 22.268	n.s.	0.023

Table 19: Estimates of the different models for the first, second and third reported episodes of thrombopenia. Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p values. *Weights apply only to the WLW model and indicate how much weight is assigned to the corresponding adverse event (AE) in the estimation of the mean effect shown in Table 18. PWP=Prentice, Williams and Peterson, WLW =Wei, Lin and Weissfeld.

Assuming common baseline hazards (Table 18), all models yielded similar results with statistically significant influence of therapy on AE occurrence: a higher hazard for AEs was observed in therapy group B compared to group A. The hazard ratios varied between 1.5 and 1.7. The results of the recurrent event models with a common baseline hazard for each AE were similar to the results of the conventional Cox model, which considers only the first episode of thrombopenia.

Considering the models in which the risk for subsequent AEs varies (Table 19), all models showed a statistically significant therapy effect with the same HR of 1.5 for the first AE. These models can be used to investigate whether the therapy has an effect, for example, only for the first AE or also for subsequent AEs. For thrombopenia, the PWP total-time model showed a therapy effect for the first and for the second AE, with the hazard increasing from HR=1.5 for the first to HR=2.5 for the second AE: If a patient randomized to therapy B already had a first episode of thrombopenia, he or she was at significantly increased risk for the second episode of thrombopenia, compared with patients in therapy group A. Of note, the risk for the third episode of thrombopenia was even higher (HR=2.9), but because of the small number of patients in this stratum, the confidence intervals were wider and no statistically significant effect was found. On the other hand, using the PWP gap-time model showed a significant treatment effect only for the first episode of thrombopenia. Of note, all models showed an increase in the hazards for subsequent AEs regardless of statistical significance.

Mean frequency of thrombopenia

The mean frequency of thrombopenia over time in both therapy groups is shown in Figure 31. A higher frequency of thrombopenia was observed for therapy group B. This effect result was statistically significant (p<0.05 from the test for equality of the two mean frequency functions).

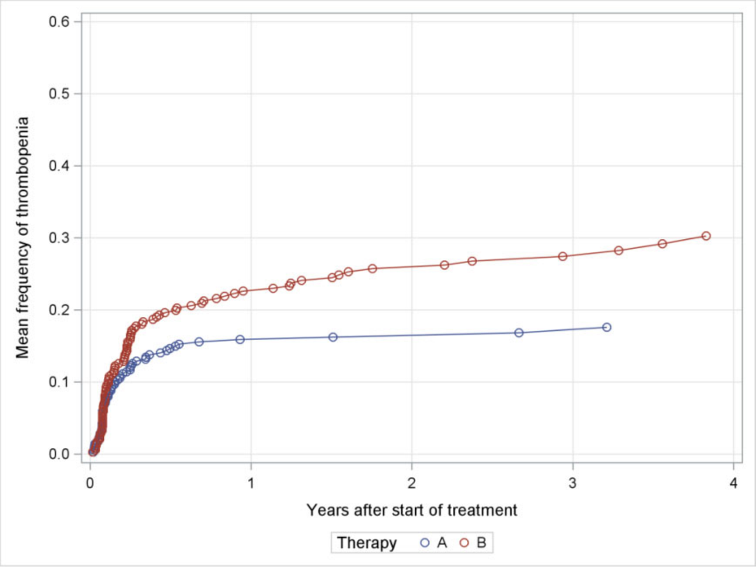


Figure 31: Mean frequency of thrombopenia for therapy groups A (blue line) and B (red line). P-value for the test of equality of the two mean cumulative functions: p<0.05.

4.6.6 Recurrent neutropenia

Table 20 shows the number of reported episodes of neutropenia. The first two episodes were analyzed to obtain a sufficient number of events in the last stratum.

Number of reported episodes of neutropenia	**Total frequency**	**Frequency prior competing event**
1	32	29
2	9	7
3	4	3
4	2	1
5	1	1

Table 20: Number of reported episodes of neutropenia. All events separated by an event-free episode of at least one day from previous event were treated as recurrent events.

The results using different models for recurrent events are shown in Table 21 (assuming a common baseline hazard for all subsequent AEs) and Table 22 (allowing the baseline hazard to vary between subsequent AEs). For ease of comparison, the crude relative risk estimate and the Cox model estimate are also shown, both for the first AE only.

Applicability	**Model**	**HR* (B vs. A)**	**95% CI**	**p value**
First AE only	Crude RR	3.152	1.436 – 6.917	**<0.05**
	Conventional Cox model	2.571	1.312 – 5.051	**<0.05**
Recurrent AEs	AG model	3.211	1.368 – 7.519	**<0.05**
	PWP total time model with common effects	2.865	1.339 – 6.135	**<0.05**
	PWP gap-time model with common effects	2.790	1.304 – 5.988	**<0.05**
	WLW model (mean effect)	2.685	1.250 – 5.767	**<0.05**

Table 21: Estimates of the different models for neutropenia. Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p-values. *For the “Crude RR” model, relative risk (RR) is shown instead of HR. AE=adverse event, AG=Andersen and Gill, PWP=Prentice, Williams and Peterson, WLW = Wei, Lin and Weissfeld.

Model	AE	HR (B vs. A)	95% CI	p value	Weight*
PWP total time model	1.	2.821	1.250 – 6.370	**<0.05**	
	2.	3.172	0.372 – 27.083	n.s.	
PWP gap-time model	1.	2.821	1.250 – 6.370	**<0.05**	
	2.	2.585	0.309 – 21.596	n.s.	
WLW model	1.	2.821	1.244 – 6.399	**<0.05**	0.873
	2.	1.911	0.220 – 16.637	n.s.	0.127

Table 22: Estimates of the different models for the first and second reported episode of neutropenia. Shown are the hazard ratios (HR) for therapy group B vs. A, their corresponding 95% confidence intervals (CI), and the p-values. *Weights apply only to the WLW model and show how much weight is assigned to the corresponding adverse event (AE) in the estimation of the mean effect in Table 20. PWP=Prentice, Williams and Peterson, WLW =Wei, Lin and Weissfeld.

All models with common baseline hazard assumption showed a statistically significantly increased hazard for neutropenia in therapy group B compared to group A (Table 21). The AG model showed the highest hazard ratio of HR=3.2, while the hazard ratios of the PWP models and the WLW model were slightly lower (HR=2.7 to 2.9). Analyzing only the first AE, the hazard ratio was HR=2.6 (Cox model). The difference between the AG and the PWP models can probably be explained by the different risk sets: The AG model uses the full data, whereas in the PWP models, only those patients who already experienced a previous AE are at risk for the next AE.

When analyzing the first and second episode of neutropenia separately (Table 22) a statistically significant effect was observed only for the first AE with a HR of 2.8. Note that the confidence intervals for the second AE were very wide due to the small sample size in that stratum.

Mean frequency of neutropenia

The mean frequency of neutropenia in both therapy groups A and B over time can be seen in Figure 32. Neutropenias in group B were reported up to about 3.5 years after the start of therapy, in group A neutropenia occurred only in the first few months of therapy.

Without taking into account the observation times (see Chapter 4.2.6), the relative risk for neutropenia was RR=3.152, CI: [1.436, 6.917], indicating a 3-fold increased risk for neutropenia in treatment group B compared to treatment group A. The mean frequency functions seem misleading. Neutropenia was reported as either severe or life-threatening in 51% of all cases. The cumulative incidence function of the first episode of neutropenia (Figure 16) did not show the very different observation times between the two groups as clearly as the mean frequency function does. Of note, both functions account for competing risks.

Further investigation of neutropenias revealed that a total of 29 patients with neutropenias were included in the recurrent event analysis. In these 29 patients, a total of 41 neutropenias were reported before a competing risk had occurred (see Table 20), of which were 8 in therapy group A (27.6%) and 21 were in group B (72.4%). Specifically, neutropenia was reported once in 22 patients, twice in 4 patients, three times in 2 patients and five times in 1 patient.

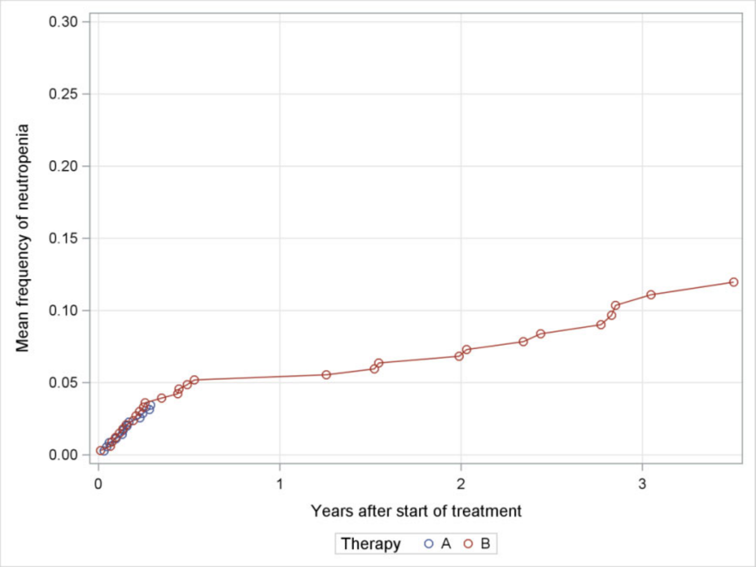


Figure 32: Mean frequency of neutropenia for therapy groups A (blue line) and B (red line). P-value for the test of equality of the two mean cumulative functions: p<0.05. Neutropenia occurred in therapy group A only up to a few months after the therapy start, in group B up to a few years after the start of therapy.

4.6.7 Summary of recurrent events analysis

Using models that account for the recurrence of adverse events, increased hazards of recurrence of fatigue, thrombopenia, neutropenia, and any grade 3 to 5 adverse events for therapy group B compared to group A were found. No statistically significant differences in the occurrence of serious adverse events between both therapies were found. In general, using the models that allowed each AE to have a different hazard, significant effects were limited to the first adverse event, with the following two exceptions: According to the PWP total time model, there was a statistically significant therapy effect on the occurrence of the second episode of thrombopenia and similarly on the occurrence of the third severe adverse event when the WLW model is used. While the case of thrombopenia seems plausible, the significant third severe adverse event might be due to the questionable assumptions of the model used. If there is a therapy effect for the occurrence of first episode of thrombopenia, there might be an increased hazard for a second episode of thrombopenia in those patients who had already experienced thrombopenia once (conditional risk set of the PWP model). On the other hand, the risk set of the WLW model allows a patient to be at risk for the third adverse event even if he or she has not experienced either the first or the second adverse event. Using the models that assume the same baseline hazard for the first and subsequent AEs, only an effect on the recurrence of the AEs can be detected, but it is not possible to distinguish exactly for which adverse events the effects are found. In general, the results from all models were similar to the conventional Cox model, which is limited to analysis of the first adverse event. In addition to the models, the mean cumulative function was shown to be useful in visualizing the mean number of adverse events occurring up to time *t*.

4.7 Summary of the results

In the present work, an analysis of adverse events was performed. First, a simple analysis based on incidence proportions (*i.e.,* percentage of patients who experienced one or more AEs) was carried out, and the risk for AEs between the two therapy groups was compared using relative risks. Taking into account the different observation times in the two therapy groups by calculating the population time at risk for an adverse event led to incidence rates and incidence rate ratios for comparing adverse events between therapy groups. Of these simple methods, incidence rates (and rate ratios) are the more appropriate method for adverse events analysis because times to adverse event or competing event are accounted for. Although it is a simple method, incidence rates are hard to find in the literature. Most papers report only incidence proportions.

Subsequently, the competing risks scenario was introduced with disease progression, death, stem cell transplantation and CML treatment switch as competing events to the first adverse event. The Aalen-Johansen estimator, which accounts for competing events, was used to estimate the cumulative incidence of adverse events. The cumulative incidence of adverse events can be interpreted as the expected proportion of patients who experience an adverse event over time. Gray's test was used to determine whether the cumulative incidence functions differ significantly between the two therapy groups.

Using multi-state models for the time to the first adverse event allowed the investigation of incidence of adverse events, competing events, and the additional path of occurrence of a competing event *after* an adverse event. Multi-state modelling is based on the analysis of cause-specific hazards and transition probabilities between the states. The cumulative event-specific hazards were plotted using the Nelson-Aalen estimator and the transition probabilities were plotted using the Aalen-Johansen estimator. Summing the transition probabilities over time yields the cumulative incidence function previously examined. The main finding was that the hazards and the transition probabilities for the competing event after an adverse event were very high compared with all other hazards. It seems plausible that, for example, CML medication was changed once a severe adverse event had occurred (the majority of the competing events were treatment switches). In addition, the hazard and the probability of experiencing a first adverse event of any severity increased sharply after initiation of therapy and then decreased. For SAEs, the hazard and probability of an SAE were low, but once an SAE occurred, most patients also rapidly experienced a subsequent competing event.

In addition to the Aalen-Johansen estimator, there are other approaches to estimate the cumulative incidence function, including an incidence rate-based parametric estimator or the 1-Kaplan-Meier estimator. The performance of these three estimators was compared. The 1-Kaplan-Meier estimator overestimates probability of adverse events in the presence of competing risks. This overestimation was found to be up to 5% in the present work. The parametric estimator for the cumulative incidences of adverse events was severely biased when all events were analyzed: after two months of therapy, a 41% difference was observed between the two estimators (cumulative incidence: 30% by the parametric estimator, 71% by the Aalen-Johansen estimator). The parametric estimator assumes a constant hazard for the occurrence of adverse events. The event-specific hazards calculated with the multi-state model can help in assessing whether the assumption of constant hazard assumption holds. In particular, in the case of any adverse event, the multi-state model showed that the hazard of an AE cannot be assumed to be constant.

Subsequently, the analysis was extended to recurrent adverse events. Two events with the same MedDRA code separated by an event-free episode of at least one day from the previous event were treated as recurrent events. Depending on the model used, a different number of recurrent adverse events could be included in the analysis. To minimize data loss, the maximum number of adverse events for which the models still could be estimated was used. It is possible to assume either a common hazard or non-common hazard for the first and all subsequent adverse events. The first assumption allows investigating whether there is a therapy effect on recurrent adverse events; the latter additionally allows distinguishing whether the therapy effect is for the first or subsequent adverse events. The mean frequency function was introduced, which provides a nice visualization of recurrent adverse events and also takes into account competing risks. Similar to testing the equality of two cumulative incidence functions with the Gray’s test, a test is available to investigate whether the mean frequency functions are differing significantly between two therapies.

In summary, a comprehensive state-of-the-art analysis of adverse events includes an estimation of the event-specific hazards and the cumulative incidences of the event of interest and of all competing events. Extension of the analysis to recurrent events is possible and its feasibility depends on the frequency and nature of the recurrent events. The mean frequency function provides an easy-to-understand visualization of the recurrence of adverse events.

5. Discussion

The aim of this work was to investigate how the reporting of adverse events could be improved. In the literature, usually only crude incidence proportions of patients with adverse events are reported. Results can be very misleading if different observation times and censoring of the patients are not taken into account.

In the present study (Chapter 4.2), median treatment times did not differ in the two therapy groups (30 months). However, individual treatment times still differed considerably. In the beginning, a simple analysis using incidence proportions and relative risks to compare adverse event proportions between therapy groups was carried out. This ignores any temporal effects and does not take into account the different observation times between the two groups. Significantly higher incidence of fatigue and neutropenia was found in the therapy group B compared to group A. A minor improvement over completely ignoring time in the analysis of adverse events is the use of incidence rates and incidence rate ratios. Both measures take into account the duration of follow-up by including the time to the occurrence of the adverse event in the denominator. This approach made the differences in adverse events between the two therapy groups more apparent: significantly higher incidence rates were additionally found for any and any severe adverse event, and thrombopenia. Regardless of the method used (incidence proportions or incidence rate ratios), no statistically significant difference in the incidence of SAEs between the two therapy groups was found.

The correct method for estimating the incidence of adverse events in case of different follow-up times and censoring is the cumulative incidence function using the Aalen-Johansen estimator. It can be considered the gold standard because it takes into account different observation times and also competing events, which are usually present in adverse events analysis. An example of a competing event is death, after which no adverse event can be observed. For the present study, disease progression, death, stem cell transplantation and CML treatment switch were chosen as competing events to adverse events. These events were summarized into a composite endpoint, for which the first occurrence of any of these events was analyzed. Thus, using the cumulative incidence function, the time to occurrence of an adverse event or any of the competing events was analyzed, whichever occurred first. Gray's test was used to test, whether there were statistically significant differences in the cumulative incidence functions between the two treatment groups. The analyses were carried out for the first of any, any severe and any serious adverse event, and additionally for the selected single adverse events such as fatigue.

As shown in Chapter 4.3, a statistically significantly higher incidence of severe adverse events was found in therapy group B compared with group A, but there were no statistically significant differences in the cumulative incidence of any or serious adverse events between the two therapy groups. Regarding the single selected adverse events, the cumulative incidence of fatigue, thrombopenia, and neutropenia was statistically significantly higher in therapy group B compared to group A.

Comparing the conventional analysis and the effect of including time to first AE, competing risks, and censoring, only few differences were found: the incidence of severe adverse events and thrombopenia differed significantly between the two groups when the time-to-event approach was used. Nevertheless, the inclusion of time to event and competing risks are important for a more appropriate analysis of adverse events, and in other studies the differences could be much more pronounced.

Graphs of cumulative incidence functions provide a convenient visualization of the data. For example, it was seen that any first adverse event occurred rapidly after initiation of therapy, or that for flu-like symptoms, presumably the addition of the second drug in therapy group B led to a sharp increase in the incidence in therapy group B. Such observations would not have been possible in a purely "conventional" analysis.

By means of a simple multi-state model, see Chapter 4.4, cumulative cause-specific hazards and transition probabilities were estimated for any, for severe, and for serious adverse events. Cumulative hazards were found to decrease with increasing severity of AEs ("any AE occurs more frequently than a severe or serious event"). The probability of experiencing an adverse event was highest for any AE and continued to decrease with increasing AE severity. In addition, the highest probability for any and severe adverse events was seen after a few months of therapy. In general, the hazard of entering the final state was much higher after a prior (S)AE than without one, and once the patient experienced an AE, the probability of entering the final state of progression, death, SCT or treatment switch was also high. The more severe the AE was, the earlier the transition occurred.

To explore the magnitude of bias caused by estimating the cumulative incidence functions of the time to a first any, severe, and serious adverse event using methods other than the gold standard of the Aalen-Johansen estimator, a comparison using three different estimators was carried out: the Aalen-Johansen estimator, the 1 – Kaplan-Meier estimator and the parametric incidence rate-based estimator. The 1 – Kaplan-Meier estimator is known to overestimate the cumulative incidence from the time point on, where the first event of interest directly after the first detection of a competing risk has been recorded (see *e.g.* [28]) and the parametric estimator assumes constant hazards [25].

The results from Chapter 4.5 show that the Kaplan-Meier estimator tended to overestimate the cumulative incidences of adverse events up to 5% after 5 years of therapy compared to the gold standard. As the amount of competing risks and censoring increased, the differences between the two estimators increased. Both estimators performed similarly in the "any adverse event scenario", in which the majority of patients had an adverse event and very few competing risks occurred. For severe and serious adverse events, the Aalen-Johansen estimator was found to be preferable over the 1-Kaplan-Meier estimator. Both estimators can be estimated with standard statistical software, therefore, the use of the Aalen-Johansen estimator should be preferred.

When comparing the Aalen-Johansen estimator and the parametric cumulative incidence estimator, the parametric estimator tended to underestimate the incidence of adverse events after initiation of therapy and overestimate it thereafter. For the "any adverse event scenario", for which the constant hazard assumption does not apply, the differences were large, making the parametric estimator unsuitable for the analysis. For severe adverse events and SAEs, the differences decreased, but from a clinical perspective, they might still be too large. Because of the use of the parametric estimator led to highly biased results, its use is very limited, at least for the data of the present study. The parametric estimator was first introduced in the analysis of incidence of bloodstream infections and HIV data [25]. While the estimator performed similarly to the Aalen-Johansen estimator for the HIV data, it was biased for the bloodstream infection data, similar to this work, because the required constant hazards assumption appeared to be violated [25]. The authors argued that although the parametric estimator did not capture the temporal dynamics of the

Aalen-Johansen estimates, it still reflected the empirical time course and was easy to use [25]. Since the Aalen-Johansen estimator can also be easily estimated with standard statistical software, the use of the parametric estimator does not seem feasible.

The Aalen-Johansen estimator remains the best option for estimating cumulative incidences because no constant hazard assumption is required and competing risks are accounted for.

In some cases, it may be important to go beyond the first occurrence of an adverse event. This led to the analysis of recurrent events, for which an overview is given by Hengelbrock *et al*. [38]. There are few models that are suitable for the analysis of events that recur after a certain event-free period. In this work, three such models were considered: Andersen and Gill (AG) model [35], Prentice, Williams and Peterson (PWP) model [39] and Wei, Lin and Weissfeld (WLW) model [40]. These models are extensions of the conventional Cox model with different assumptions. The AG model assumes independency of events, that is, the risk of experiencing subsequent adverse events is independent of the occurrence of the first adverse event. An advantage of the AG model is that all data can be used. One effect is estimated as an overall measure of the treatment effect. The PWP model uses a conditional risk set. Only those patients who experienced a first adverse event are considered at risk for experiencing a second adverse event. This results in a decrease in the number of patients at risk for increasing number of recurrencies, but it may still be the more biologically plausible model. One can estimate an overall effect, but also an effect for each recurrent event. The WLW model has the counterintuitive definition of the risk set that all patients still followed up within the study are at risk for the next recurrence, even if they have not experienced a previous recurrence. Also, in this model, the number of patients at risk decreases from recurrence to recurrence as patients are censored when their individual follow-up time is reached. The WLW model estimates an effect for each recurrence, and an overall effect can be calculated as a linear combination of the individual effects. The choice of the appropriate model should depend on the purpose of the study. Here, all three models were estimated to assess their potential because the focus of the work was on the models themselves and not on medical findings.

Recurrent adverse events analysis in this work (see Chapter 4.6) showed increased hazards of recurrence of fatigue, thrombopenia, neutropenia, and any grade 3 to 5 adverse events for therapy group B compared to group A. No statistically significant differences in the occurrence of and serious adverse events between both therapies were found. In general, the results from all models were similar to the conventional Cox model, which is limited to analysis of the first adverse event. As pointed out in Hengelbrock *et al.* [38], in case of competing events, the recurrent event models are limited to estimation of cause-specific hazards. As pointed out in Pfirrmann *et al*. [45], absolute event probabilities can no longer be calculated from cause-specific hazards, and authors have called for including estimates of the treatment effect on cause-specific hazards and on cumulative incidences (which can be considered as event probabilities) [51].

Therefore, Hengelbrock *et al*. suggested accompanying the estimates of recurrent event analyses with estimates of the mean cumulative function [38], which was done in the present work. In addition to allowing the comparison of incidences of adverse events when competing risks are present, the mean cumulative function was also shown to be useful in visualizing the mean number of adverse events occurring up to time *t*.

Hengelbrock *et al.* [38] analyzed adverse event data from a study of patients with renal cell carcinoma randomized to sunitinib (treatment group) or interferon alpha (control group) using recurrent event methods. In the trial, median treatment duration differed significantly between the two treatment groups (11 months vs. 4 months), resulting in different exposure times to adverse events per group. Recurrent events were treated similarly to the present work, and the number of patients with recurrent adverse events was also low. Because the crude percentages and relative risks of adverse events per group were difficult to interpret because of the different exposure times to treatment, mean cumulative functions were presented that showed a higher number of adverse events in the treatment group compared with the control group. In their study, the crude relative risks overestimated the risk of adverse events in the treatment group because of the difference in treatment duration. This was not a problem in the TIGER study presented in this thesis because the median treatment duration was the same in both treatment groups (30 months). The results of the recurrent event models were roughly comparable to those of the Cox model, which considers only the first adverse event, which is also consistent with the results in this thesis. It appears that if adverse events do not recur frequently, the Cox model may be sufficient for their analysis.

It is more or less standard procedure to analyze the data and then rely on p-values for testing for statistical significance. Statistically significant results are considered important and have a higher chance of being published. In well-designed trials, statistical significance should also be consistent with clinically relevant results. However, this is usually true only for efficacy analyses, for which most trials are designed. Often, safety analyses accompany the primary analyses and serve as secondary endpoint of the study.

It is worth reflecting on the interpretation of safety data and the differences in the assessment of efficacy and safety in clinical trials. The considerations follow the introduction in Chapter 1.3 of [52]. Efficacy is easy to assess, at least in theory, because the criteria for assessing efficacy must be explicitly stated in the trial protocol. Studies are generally designed to show a difference in efficacy. For that, pre-specified null hypotheses that there are no differences between the two therapies are rejected based on observations made during the trial. A CML-specific example would be the comparison of the rate of deep molecular response achieved after, say, two years of therapy. If more than one hypothesis is to be tested, an adjustment for multiple testing is specified in the study protocol so that the probability of the new therapy being falsely classified as effective when it is not is kept at an acceptable level (typically 5%). Thus, efficacy is known or assumed before the trial is conducted. The sample size of the study is determined to demonstrate this efficacy.

The evaluation of safety is different from the evaluation of efficacy. Most adverse events are not identified before the trial, but are observed during the trial. This also means that it is not possible to determine whether a therapy is safe or not before the trial is conducted. This leads to the situation where the hypotheses for testing safety are being made during/after the trial, so the same observations are used to generate and test the hypotheses. It would be much better to create the hypotheses before the data are collected.

One problem with p-values in this context is that no p-value adjustment is usually made for safety data testing. All p-values below 0.05 are considered statistically significant. This, in turn, can lead to “false alarms”, *i.e.* significant results resulting from multiple testing even when in reality there is no statistically significant effect. Also, statistical significance does not equal clinical relevance. Careful interpretation of the safety data by medical experts is required. For example, in the present work, no statistically significant difference in the incidence of serious adverse events was observed between the two therapy groups. However, it could be the case that the overall incidence of SAEs is unacceptably high and should lead to discontinuation of the study drug(s) in both treatment arms. Such a conclusion would not be possible if only differences between the groups were tested. Statistically significant results may be considered as “hints” regarding the safety and serve as potential research questions worth addressing in future studies.

Another point worth considering is the “healthy subjects” effect: Clinical trials have strict inclusion and exclusion criteria that allow only relatively healthy subjects to participate in the trials. This could mean that subjects at higher risk for adverse events are excluded from trial participation. Safety data from such subjects can only be obtained in post-approval studies that evaluate the safety of approved drugs under real-world conditions.

This work has many limitations, the most important of which are described below.

5.1 Limitations

Data quality

Although the data were obtained from a large study with ongoing data monitoring and data cleaning to ensure good quality data, some of the adverse events could not be included in the analyses presented in this work. Reasons included missing severity grade, missing or implausible start date, or missing description of the adverse event itself.

Some of the inconsistencies in the data were resolved by manual data cleaning, *e.g.* changing ongoing adverse events with reported end date to “not-ongoing”. This is because the AE was likely ongoing initially and when it resolved and the end date was entered, the question about duration was forgotten to be changed to “no”. Another example of data cleaning would be the assignment of an artificial duration of one day for AEs that were marked as not ongoing but for which no actual end date was provided. In this way, such AEs could be included in the analysis of recurrent events. Special care was taken not to report any data after the patient withdrew informed consent.

Recurrent events

Although more than 7000 adverse events from 689 patients were included in the analysis, the analysis of recurrent events was somewhat disappointing: Many of the adverse events were not recurrencies, but often only the severity had changed between two subsequent AEs. Such adverse events were counted as a single adverse event. This was very common for hematological adverse events like thrombopenia or leukopenia, which were among the most common adverse events during the study. It was common for such laboratory value-related AEs to persist for years. This largely reduced the size of the original data set of over 7000 AEs. It is a matter of opinion how recurrent events should be defined, but for this work with a focus on the methods, all adverse events with at least one day break between two events were counted as recurrent. Because some recurrent event models count as at-risk for the next adverse event only those individuals who have experienced a previous adverse event, the risk sets became very small with increasing number of recurrent adverse events. For this reason, a cut-off value was set for each adverse event to allow inclusion of a maximum number of adverse events so that the models still converged. All subsequent adverse events could not be included in the analysis. These limits were determined experimentally by increasing the number of recurrent events until the models no longer converged. Depending on the type of the adverse event, 2 to 4 subsequent adverse events could be included in the recurrent event analysis.

No covariates

No other covariates besides the therapy were included in the analyses. Due to the randomized study design, most covariates such as age or sex should be balanced between both therapy groups (for age and sex this was the case with median age of 51 years in both therapy groups (p=n.s.) and proportion of males of 58.1% and 60.8% (p=n.s.) in treatment groups A and B, respectively). However, it cannot be excluded that, for example, age has a significant effect on the occurrence of a first adverse event but not on the occurrence of subsequent adverse events. Another example it could be that adverse events occur more frequently in women than in men. Such studies involving covariates are left for future analyses.

Study medication

Another limitation of the present work is the study medication itself. Since the purpose of this work was to perform a comprehensive adverse event analysis and not to focus on the therapies themselves, both therapy arms were labeled as A and B, respectively. Therapy A consists of a CML-specific TKI drug nilotinib that is compared to a combination of this drug with interferon alpha. This means that B is a combination of A and interferon alpha. Statistically significant differences in the incidence of adverse events between therapies A and B could be found. These differences were in most cases likely caused by interferon alpha, because nilotinib was given in both therapy groups. It seems pretty plausible that giving two drugs instead of one would lead to more adverse events, which was observed for many of the adverse events studied. However, an important finding is that the addition of interferon alpha to nilotinib (the TKI used in therapy group A) did not increase the number of serious adverse events. Synergetic effects, such as additive or multiplicative effects of the drugs used in therapy group B, could also not be reasonably considered due to the lack of data.

All adverse events were analyzed, regardless of their reported relation with the study drug(s). In the present study, it was intended that the relation of the adverse event to the study drug should be reported. This information is highly subjective and missing in many cases. Because an adverse event is defined as any untoward medical event during the study, regardless of the relation to the study drug(s), no special attention was paid to a possible association between the adverse event and the study drug(s).

The study design provides three phases of the study: in the induction phase, both treatment groups receive the drugs according to randomization. In the maintenance phase, group B receives only one of the two combination drugs used in the induction phase (interferon alpha), and in the discontinuation phase, no study medication was to be administered at all. All analyses performed in this work were intention-to-treat analyses that did not distinguish between the different phases of the study. It would be better to perform separate analyses for each study phase to see how the incidences of adverse events change between the different phases. This distinction might also be more suitable to characterize the adverse events attributable to one of the drugs. However, as pointed out previously, the focus of this thesis lies on the methods.

Censoring

For all time-to-event models, the assumption of independent censoring was made. This means that the event and censoring were supposed not to be related, which could be questionable: In many cases where treatment was changed, it happened shortly after the occurrence of a severe AE/SAE. Treatment switching was considered a competing risk to AEs, and therefore patients were censored when the treatment was changed. In these cases, the censoring mechanism does not appear to be independent of the occurrence of AEs. However, it remains unclear how this might affect the results.

Resolving adverse events

Multi-state models can be used to model the probability of an adverse event, but also used to model the probability of transition from the AE back to the “healthy” state. This application might be of interest for single AEs, where it might be important to know whether the AE persists over a long period of time and what the probability is that the AE will resolve after it has occurred. This type of transition was not considered in the present work.

6. Conclusion

In general, analysis of adverse events in clinical trials is limited to reporting the percentage of patients with adverse events. Adverse events in different groups or between different therapies are often compared using relative risks or other simple measures. This does not take into account that adverse events can occur at any time during the study and that patients are not usually followed for the same duration of time. Failure to account for time in the safety analysis can substantially bias the results and, in the worst case, result in approval of a drug that causes more harm than is acceptable or, the other way around, approval of an effective drug might be delayed for safety reasons simply because the safety analysis was conducted in an inappropriate manner.

This work focused on introducing different methods for analyzing adverse events data and the way carrying out these analyses. First, the "conventional" proportion-based analyses were shown. Then, time to event and type of event were included in the analyses, recognizing the fact that the time until the event and the type of event matter. Competing risks were introduced, which are in competition with adverse events and should also be considered. The Aalen-Johansen estimator is the gold standard for analyzing the cumulative incidence of adverse events in the presence of competing risks and censoring. However, cumulative incidences can be estimated using other estimators, such as the Kaplan-Meier or the parametric estimator, too. To assess to amount of bias using these "other" estimators, the performance of different cumulative incidence estimators was compared for different amount of events and censored data. This bias assessment confirmed the importance of using the Aalen-Johansen estimator for an appropriate analysis of a first adverse event. Analyses were then extended for recurrent adverse events, but in the absence of sufficient data, no large differences between first and recurrent adverse events were found. However, depending on the research question, it might be useful to use recurrent event methods. For example, investigating recurrent skin or bladder cancer might be good situations in which such models could be of great use. In the study investigated in this thesis, extending the analysis of adverse events to include recurrent adverse events did not appear to result in a substantial information gain regarding adverse events.

In summary, a state-of-the-art analysis of adverse events should include the time to occurrence of an event. The simplest way is to calculate the times at risk for an adverse event, report the incidence rates, and compare different groups using incidence rate ratios. A more sophisticated and better option is to show the cumulative incidence function of adverse events using the Aalen-Johansen estimator, which also takes into account competing events for the occurrence of an adverse event and different follow-up times of patients. The differences between two cumulative incidence functions can be tested using the Gray's test. This presented approach is limited to the first adverse event. For recurrent adverse events analysis one could plot the mean cumulative function visualizing the expected number of adverse events by time t. Then, a test of equality of two mean cumulative functions can be used to test if the mean number of expected adverse events differs between different therapy groups. In this way, time to event and individual patient times are properly accounted for, resulting in estimates that are less susceptible to bias than when time is neglected.

Hopefully, adverse event reporting in future studies will incorporate some of the sophisticated methods presented in this work.

References

1. O'Neill RT. Statistical analyses of adverse event data from clinical trials. Special emphasis on serious events. Drug Inf J. 1987; 21:9-20.

2. International Conference on Harmonisation. 1994. ICH 2EA: Clinical Safety Data Management: Definitions and Standards for Expedited Reporting. CPMP/ICH/377/95. https://www.ema.europa.eu/documents/scientific-guideline/international-conference-harmonisation-technical-requirements-registration-pharmaceuticals-human-use_en-15.pdf. Accessed on August 30th, 2020.

3. International Conference on Harmonisation. 1998. ICH E9: Statistical Principles for Clinical Trials. CPMP/ICH/363/96. https://www.ema.europa.eu/documents/scientific-guideline/ich-e-9-statistical-principles-clinical-trials-step-5_en.pdf. Accessed on August 30th, 2020.

4. Unkel S, Amiri M, Benda N, Beyersmann J, Knoerzer D, Kupas K, Langer F, Leverkus F, Loos A, Ose C, Proctor T, Schmoor C, Schwenke C, Skipka G, Unnebrink K, Voss F, Friede T. On estimands and the analysis of adverse events in the presence of varying follow-up times within the benefit assessment of therapies. Pharmaceutical Statistics. 2019;18:166–183.

5. Rowley JD. A new consistent chromosomal abnormality in chronic myelogenous leukemia identified by quinacrine fluorescence and Giesma staining. Nature 1973; 243: 290–293.

6. Quintas-Cardama A, Cortes JE. Chronic myeloid leukemia: diagnosis and treatment. Mayo Clin Proc 81(7):973-988.

7. Treatment optimization of newly diagnosed Ph/BCR-ABL positive patients with chronic myeloid leukemia (CML) in chronic phase with nilotinib vs. nilotinib plus interferon alpha induction and nilotinib or interferon alpha maintenance therapy. TIGER clinical trial study protocol Version 1.6 of December 17th 2017.

8. Ilaria RL. Pathobiology of lymphoid and myeloid blast crisis and management issues. Hematology Am Soc Hematol Educ Program. 2005;1(1):188-194.

9. Hochhaus A, Baccarani M, Silver RT, Schiffer C, Apperley JF, Cervantes F, Clark RE, Cortes JE, Deininger MW, Guilhot F, Hjorth-Hansen H, Hughes TP, Janssen JJWM, Kantarjian HM, Kim DW, Larson RA, Lipton JH, Mahon FX, Mayer J, Nicolini F, Niederwieser D, Pane F, Radich JP, Rea D, Richter J, Rosti G, Rousselot P, Saglio G, Saußele S, Soverini S, Steegmann JL, Turkina A, Zaritskey A, Hehlmann R. European LeukemiaNet 2020 recommendations for treating chronic myeloid leukemia. Leukemia. 2020;34(4):966-984.

10. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, Deininger MW, Silver RT, Goldman JM, Stone RM, Cervantes F, Hochhaus A, Powell BL, Gabrilove JL, Rousselot P, Reiffers J, Cornelissen JJ, Hughes T, Agis H, Fischer T, Verhoef G, Shepherd J, Saglio G, Gratwohl A, Nielsen JL, Radich JP, Simonsson B, Taylor K, Baccarani M, So C, Letvak L, Larson RA; IRIS Investigators. Five-year follow-up of patients receiving imatinib for chronic myeloid leukemia. N Engl J Med. 2006 Dec 7;355(23):2408-17.

11. Hochhaus A, Larson RA, Guilhot F, Radich JP, Branford S, Hughes TP, Baccarani M, Deininger MW, Cervantes F, Fujihara S, Ortmann CE, Menssen HD, Kantarjian H, O'Brien SG, Druker BJ; IRIS Investigators. Long-Term Outcomes of Imatinib Treatment for Chronic Myeloid Leukemia. N Engl J Med. 2017 Mar 9;376(10):917-927. doi: 10.1056/NEJMoa1609324.

12. Hehlmann R, Lauseker M, Saußele S, Pfirrmann M, Krause S, Kolb HJ, Neubauer A, Hossfeld DK, Nerl C, Gratwohl A, Baerlocher GM, Heim D, Brümmendorf TH, Fabarius A, Haferlach C,

Schlegelberger B, Müller MC, Jeromin S, Proetel U, Kohlbrenner K, Voskanyan A, Rinaldetti S, Seifarth W, Spieß B, Balleisen L, Goebeler MC, Hänel M, Ho A, Dengler J, Falge C, Kanz L, Kremers S, Burchert A, Kneba M, Stegelmann F, Köhne CA, Lindemann HW, Waller CF, Pfreundschuh M, Spiekermann K, Berdel WE, Müller L, Edinger M, Mayer J, Beelen DW, Bentz M, Link H, Hertenstein B, Fuchs R, Wernli M, Schlegel F, Schlag R, de Wit M, Trümper L, Hebart H, Hahn M, Thomalla J, Scheid C, Schafhausen P, Verbeek W, Eckart MJ, Gassmann W, Pezzutto A, Schenk M, Brossart P, Geer T, Bildat S, Schäfer E, Hochhaus A, Hasford J. Assessment of imatinib as first-line treatment of chronic myeloid leukemia: 10-year survival results of the randomized CML study IV and impact of non-CML determinants. Leukemia. 2017 Nov; 31(11):2398-2406.

13. Beillard E, Pallisgaard N, van der Velden VH, Bi W, Dee R, van der Schoot E, Delabesse E, Macintyre E, Gottardi E, Saglio G, Watzinger F, Lion T, van Dongen JJ, Hokland P, Gabert J. Evaluation of candidate control genes for diagnosis and residual disease detection in leukemic patients using 'real-time' quantitative reverse-transcriptase polymerase chain reaction (RQ-PCR) - a Europe against cancer program. Leukemia. 2003 Dec;17(12):2474-86.

14. Hehlmann R. Chronic myeloid leukemia. Springer. 2017

15. Hughes TP, Kaeda J, Branford S, Rudzki Z, Hochhaus A, Hensley ML, Gathmann I, Bolton AE, van Hoomissen IC, Goldman JM, Radich JP; International Randomised Study of Interferon versus STI571 (IRIS) Study Group. Frequency of major molecular responses to imatinib or interferon alfa plus cytarabine in newly diagnosed chronic myeloid leukemia. N Engl J Med. 2003 Oct 9;349(15):1423-32.

16. Cross NC, White HE, Colomer D, Ehrencrona H, Foroni L, Gottardi E, Lange T, Lion T, Machova Polakova K, Dulucq S, Martinelli G, Oppliger Leibundgut E, Pallisgaard N, Barbany G, Sacha T, Talmaci R, Izzo B, Saglio G, Pane F, Müller MC, Hochhaus A. Laboratory recommendations for scoring deep molecular responses following treatment for chronic myeloid leukemia. Leukemia. 2015 May;29(5):999-1003.

17. Sokal JE, Cox EB, Baccarani M, Tura S, Gomez GA, Robertson JE, Tso CY, Braun TJ, Clarkson BD, Cervantes F, et al. Prognostic discrimination in "good-risk" chronic granulocytic leukemia. Blood. 1984;63:789-99.

18. Hasford J, Pfirrmann M, Hehlmann R, Allan NC, Baccarani M, Kluin-Nelemans JC, Alimena G, Steegmann JL, Ansari H. A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. Writing Committee for the Collaborative CML Prognostic Factors Project Group. J Natl Cancer Inst. 1998;90:850-8.

19. Hasford J, Baccarani M, Hoffmann V, Guilhot J, Saussele S, Rosti G, Guilhot F, Porkka K, Ossenkoppele G, Lindoerfer D, Simonsson B, Pfirrmann M, Hehlmann R. Predicting complete cytogenetic response and subsequent progression-free survival in 2060 patients with CML on imatinib treatment: the EUTOS score. Blood. 2011;118(3):686-92.

20. Pfirrmann M, Baccarani M, Saussele S, Guilhot J, Cervantes F, Ossenkoppele G, Hoffmann VS, Castagnetti F, Hasford J, Hehlmann R, Simonsson B. Prognosis of long-term survival considering disease-specific death in patients with chronic myeloid leukemia. Leukemia. 2016; 30(I):48-56.

21. Saussele S, Richter J, Guilhot J, Gruber FX, Hjorth-Hansen H, Almeida A, Janssen JJWM, Mayer J, Koskenvesa P, Panayiotidis P, Olsson-Strömberg U, Martinez-Lopez J, Rousselot P, Vestergaard H, Ehrencrona H, Kairisto V, Machová Poláková K, Müller MC, Mustjoki S, Berger

MG, Fabarius A, Hofmann WK, Hochhaus A, Pfirrmann M, Mahon FX, EURO-SKI investigators. 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 Jun;19(6):747-757.

22. Proctor T, Schumacher M. Analysing adverse events by time-to-event models: The CLEO-PATRA study. Pharmaceut. Statist. 2016;15;306-412.

23. Allignol A, Beyersmann J, Schmoor C. Statistical issues in the analysis of adverse events in time-to-event data. Pharmaceut. Statist. 2016;15;297-305.

24. Newcombe RG. Two.sided confidence intervals for the single proportion: Comparison of seven methods. Statistics in Medicine. 1998;17:857-872.

25. Grambauer N, Schumacher M, Dettenkofer M, Beyersmann J. Incidence Densities in a Competing Events Analysis. Am J Epidemiol. 2010;172:1077–1084.

26. Siddiqui O. Statistical methods to analyze adverse events data of randomized clinical trials. J Biopharm Stat. 2009;19(5):889-899.

27. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. J Am Stat Assoc. 1958; 53(282):457–481

28. Pfirrmann M, Hochhaus A, Lauseker M, Saussele S, Hehlmann R, Hasford J. Recommendations to meet statistical challenges arising from endpoints beyond overall survival in clinical trials on chronic myeloid leukemia. Leukemia. 2011; 25(9):1433-8.

29. Cox DR. Regression models and life-tables. J R Stat Soc Series B (Methodological). 1972;34(2):187-220.

30. Ozga A-K, Kieser M, Rauch G. A systematic comparison of recurrent event models for application to composite endpoints. BMC Medical Research Methodology.2018; 18:2.

31. Aalen O, Borgan O, Gjessing H. Survival and Event History Analysis: A Process Point of View. Berlin, Germany: Springer; 2008.

32. Dignam JJ, Kocherginsky MN. Choice and interperation of statistical tests used when competing risks are present. J Clin Oncol. 2008 Aug 20;26(24):4027-34.

33. Choudhury JB. Non-parametric confidence interval estimation for competing risks analysis: application to contraceptive data. Stat Med. 2002 Apr 30;21(8):1129-44.

34. Gray, RJ. A class of K-sample tests for comparing the cumulative incidence of a competing risk. The Annals of Statistics 1988; 16: 1141-1154.

35. Andersen PK, Gill RD. Cox's regression model for counting processes: A large sample study. Ann. Stat. 1982;10(4):1100-20.

36. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. J Am Stat Assoc. 1989;84(408):1074-8.

37. Cheung YB, Xu Y, Tan SH, Cutts F, Milligan P. Estimation of intervention effects using first of multiple episodes in clinical trials: The Andersen-Gill model re-examined. Statist. Med. 2010;29:328-336.

38. Hengelbrock J, Gillhaus J, Kloss S, Leverkus F. Safety data from randomized controlled trials: applying models for recurrent events. Pharm Stat. 2016 Jul;15(4):315-23.

39. Prentice RL, Williams BJ, Peterson AV. On the regression analysis of multivariate failure time data. Biometrika. 1981;68(2):373-9.

40. Wei LJ, Lin DY, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distributions. Journal of the American Statistical Association. 1989; 84 (408): 10651073.

41. Nelson W. Confidence Limits for Recurrence Data - Applied to Cost or Number of Product Repairs. Tehnometrics. 1995;3(2):147-157.

42. Doganaksoy N and Nelson W. A Method to Compare Two Samples of Recurrence Data. Lifetime Data Analysis 1998;4:51–63.

43. Cook RJ and Lawless JF. The statistical Analysis of Recurrent Events. New York: Springer; 2007.

44. R Core Team 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

45. Pfirrmann M, Lauseker M, Hoffmann VS, Hasford J. Prognostic scores for patients with chronic myeloid leukemia under particular consideration of competing causes of death. Ann Hematol. 2015; 94 Suppl 2:S209-18.

46. Brioli A, Manz K, Pfirrmann M, Hänel M, Schwarzer AC, Prange-Krex G, Fabisch C, Knop S, Illmer T, Krammer-Steiner B, Hochhaus A, von Lilienfeld-Toal M, Mügge LO. Frailty impairs the feasibility of induction therapy but not of maintenance therapy in elderly myeloma patients: final results of the German Maintenance Study (GERMAIN). J Cancer Res Clin Oncol. 2020 Mar;146(3):749-759.

47. Jackson JH. Multi-State Models for Panel Data: The msm Package for R. Journal of Statistical Software. 2011;38(8):1-29.

48. Allignol A, Beyersmann J, Schumacher M. mvna: An R Package for the Nelson-Aalen Estimator in Multistate Models. R News. 2008;8(2):48-50.

49. Gray B. cmprsk: Subdistribution Analysis of Competing Risks. R package version 2.2-9. 2019. https://CRAN.R-project.org/package=cmprsk

50. Allignol A, Schumacher M, Beyersmann J. Empirical Transition Matrix of Multi-State Models: The etm Package. Journal of Statistical Software. 2011;38(4): 1-15.

51. Latouche A, Allignol A, Beyersmann J, Labopin M, Fine JP. A competing risks analysis should report results on all cause-specific hazards and cumulative incidence functions. J Clin Epidemiol 2013; 66(6):648–653.

52. Gould AL. Statistical Methods for Evaluation Safety in Medical Product Development. John Wiley and Sons Ltd; 2015.

Acknowledgement

Ich möchte mich bei allen Personen bedanken, die mich während der Promotionszeit unterstützt haben.

Als erstes möchte ich mich bei Herrn Prof. Dr. Ulrich Mansmann bedanken, diese Dissertation in seinem Institut schreiben zu dürfen. Mein besonderer Dank gilt meinem Doktorvater PD Dr. Markus Pfirrmann, der mich in meinem Vorhaben stets unterstützt hat und mit seiner umfassenden Expertise auf meinem Weg begleitet hat. Dr. Michael Lauseker danke ich für die Aufrechterhaltung der Kaffeeversorgung im IBE, die neben manchen Gesprächen in der Kaffeeküche einen wichtigen Beitrag zu meiner Lebensqualität während der Arbeit geleistet hat. Danke auch an Denise Kohn und Verena Loidl, die immer ein offenes Ohr für mich hatten, auch über die große Entfernung zwischen München und Greifswald hinweg.

Am neuen Ort in Greifswald möchte ich mich bei Dr. Kerstin Weitmann und Gabriele Robers für ihre Unterstützung und den Glaube daran danken, dass ich die Dissertation fertigstellen werde.

Als letztes möchte ich mich bei meiner Familie, bei Peter, Antti und Colin bedanken. Für die Liebe, für die Geduld, und auch für die Hinweise, mal den Laptop zuzuklappen und was mit euch zu unternehmen.

Affidavit

LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Promotionsbüro
Medizinische Fakultät

Eidesstattliche Versicherung

Manz, Kirsi Marjaana

Name, Vorname

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Dissertation mit dem Titel:

Analysis of adverse events with modern statistical methods

selbständig verfasst, mich außer der angegebenen keiner weiteren Hilfsmittel bedient und alle Erkenntnisse, die aus dem Schrifttum ganz oder annähernd übernommen sind, als solche kenntlich gemacht und nach ihrer Herkunft unter Bezeichnung der Fundstelle einzeln nachgewiesen habe.

Ich erkläre des Weiteren, dass die hier vorgelegte Dissertation nicht in gleicher oder in ähnlicher Form bei einer anderen Stelle zur Erlangung eines akademischen Grades eingereicht wurde.

Greifswald, 04.10.2023 Kirsi Marjaana Manz

Ort, Datum Unterschrift Doktorandin bzw. Doktorand

Publications

Peer reviewed

1. **Manz KM**, Clowes P, Kroidl I, Kowuor DO, Geldmacher C, Ntinginya NE, Maboko L, Hoelscher M, Saathoff E: Trichuris trichiura infection and its relation to environmental factors in Mbeya region, Tanzania: A cross- sectional, population-based study. PLoS ONE (2017); 12(4): e0175137.

2. Rupprecht TA*, **Manz KM***, Fingerle V, Lechner C, Klein M, Pfirrmann M, Koedel U: Diagnostic value of cerebrospinal fluid CXCL13 for acute Lyme neuroborreliosis. A systematic review and meta-analysis. Clinical Microbiology and Infection (2018); 24: 1234-1240.

3. Rinaldetti S, Pfirrmann M, **Manz K**, Guilhot J, Dietz C, Panagiotidis P, Spiess B, Seifarth W, Fabarius A, Müller M, Pagoni M, Dimou M, Dengler J, Waller CF, Brümmendorf TH, Herbst R, Burchert A, Janben C, Goebeler ME, Jost PJ, Hanzel S, Schafhausen P, Prange-Krex G, Illmer T, Janzen V, Klausmann M, Eckert R, Büschel G, Kiani A, Hofmann W-K, Mahon F-X, Saussele S: Effect of ABCG2, OCT1, and ABCB1 (MDR1) Gene Expression on Treatment-Free Remission in a EURO-SKI Subtrial. Clinical Lymphoma, Myeloma & Leukemia (2018); 18(4): 266-71.

4. **Manz KM**: New Advances in the Treatment of Trichuriasis. Curr Treat Options Infect Dis (2018); 10: 362-372.

5. Thaller PH, Fürmetz J, Chen F, Degen N, **Manz KM**, Wolf F : Bowlegs and intensive football training in children and adolescents -a systematic review and meta-analysis. Dtsch Arztebl Int (2018); 115: 401-8.

6. Krause D, Warnecke M, Schuetz CG, Soyka M, **Manz KM**, Proebstl L, Kamp F, Chrobok AI, Pogarell O, Koller G: The Impact of the Opioid Antagonist Naloxone on Experimentally Induced Craving in Nicotine-Dependent Individuals. Eur Addict Res (2018); 24: 255-265.

7. Aklan B, Zilles B, Paprottka P, **Manz K**, Pfirrmann M, Santl M, Abdel-Rahman S & Lindner LH: Regional deep hyperthermia: quantitative evaluation of predicted and direct measured temperature distributions in patients with high-risk extremity soft-tissue sarcoma. Int J Hyperthermia. 2019; 36(1): 170-185.

8. Proebstl L, Kamp F, **Manz K**, Krause D, Adorjan K, Pogarell O, Koller G, Soyka M, Falkai P, Kambeitz J: Effects of stimulant drug use on the dopaminergic system: A systematic review and meta-analysis of in vivo neuroimaging studies. European Psychiatry (2019); 59: 15-24.

9. Kamp F, Proebstl L, Hager L, Schreiber A, Riebschläger M, Neumann S, Straif M, Schacht-Jablonowsky M, **Manz K**, Soyka M, Koller G: Effectiveness of metamphetamine abuse treatment: Predictors of treatment completion and comparison of two residential treatment programs. Drug and Alcohol Dependence (2019); 201: 8-15.

10. Proebstl L, Kamp F, Hager L, Krause D, Riebschläger M, Neumann S, Schacht Jablonowsky M, Schreiber A, Straif M, **Manz K**, Soyka M, Koller G: Associations between methamphetamine use, psychiatric comorbidities and treatment outcome in two inpatient rehabilitation centers. Psychiatry Research (2019); 280: 112505.

11. Proebstl L, Krause D, Kamp F, Hager L, **Manz K**, Schacht-Jablonowsky M, Straif M, Riebschläger M, Neumann S, Schreiber A, Soyka M, Koller G: Methamphetamine withdrawal and the

restoration of cognitive functions -a study over a course of 6 months abstinence. Psychiatry Research (2019); 281: 112599.

12. Brioli A, **Manz K**, Pfirrmann M, Hänel M, Schwarzer AC, Prange-Krex G, Fabisch C, Knop S, Illmer T, Krammer-Steiner B, Hochhaus A, von Lilienfeld-Toal M, Mügge L-O: Frailty impairs the feasibility of induction therapy but not of maintenance therapy in elderly myeloma patients: final results of the German Maintenance Study (GERMAIN). Journal of Cancer Research and Clinical Oncology (2020); 146:749-759.

13. **Manz K**, Fenchel K, Eilers A, Morgan J, Wittling K, Dempke WCM: Efficacy and Safety of Approved First-Line Tyrosine Kinase Inhibitor Treatments in Metastatic Renal Cell Carcinoma: A Network Meta-Analysis. Adv Ther (2020); 37:730-744.

14. Degen N, Suero E, Bogusch M, Neuerburg C, **Manz K**, Becker CA, Befrui N, Kammerlander C, Böcker W, Zeckey C: Intraoperative use of cortical step sign and diameter difference sign: Accuracy, inter-rater agreement and influence of surgical experience in subtrochanteric transverse fractures. Orthop Traumatol Surg Res. (2020); Apr 9. pii: S1877-0568(20)30103-1.

15. Kamp F, Hager L, Proebstl L, Schreiber A, Riebschläger M, Neumann S, Straif M, Schacht-Jablonowsky M, Falkai P, Pogarell O, **Manz K**, Soyka M, Koller G: 12- And 18-month follow-up after residential treatment of methamphetamine dependence: Influence of treatment drop-out and different treatment concepts. Journal of Psychiatric Research (2020); doi: https://doi.org/10.1016/j.jpsychires.2020.05.029.

16. **Manz KM**, Kroidl I, Clowes P, Gerhardt M, Nyembe W, Maganga L, Assisya W, Ntinginya NE, Berger U, Hoelscher M, Saathoff E: Schistosoma haematobium infection and environmental factors in Southwestern Tanzania: A cross-sectional, population-based study. PLoS Negl Trop Dis (2020);14(8):e0008508.

17. Sharaf K, Grueninger I, Hilpert A, Polterauer D, Volgger V, **Manz K**, et al: Stapes and stapes revision surgery: Preoperative air bone gap is a prognostic marker. Otol Neurotol (2021) 42(7):985-993.

18. Masouris I, **Manz K**, Pfirrmann M, Dreyling M, Angele B, Straube A, et al: CXCL13 and CXCL9 CSF levels in central nervous system lymphoma - diagnostic, therapeutic, and prognostic relevance. Front. Neurol. (2021) 12:654543.

19. **Manz KM**, Mansmann U: Inequality indices to monitor geographic differences in incidence, mortality and fatality rates over time during the COVID-19 pandemic. PLoS ONE (2021) 16(5):e0251366.

20. Franke AG, Koller G, Krause D, Proebstl L, Kamp F, Pogarell O, Jebrini T, **Manz K** et al: Just “Like Coffee” or Neuroenhancement by Stimulants? Front. Public Health (2021) 9:640154.

21. **Manz K**, Mansmann U: Regionales Monitoring von Infektionen mittels standardisierter Fallfatalitätsraten am Beispiel von SARS-CoV-2 in Bayern. Bundesgesundheitsblatt (2021) 64:1146-1156 (2021).

22. Franke AG, Koller G, Neumann S, Proebstl L, **Manz K**, Krause D, et al: Psychopathology and Attention Performance in Methamphetamine Users with ADHD Symptomology in Childhood. Int J Ment Health Addiction (2021). https://doi.org/10.1007/s11469-021-00682-0

23. Jebrini T, **Manz K**, Koller G, Krause D, Soyka M, Franke AG: Psychiatric Comorbidity and Stress in Medical Students Using Neuroenhancers. Front Psychiatry (2021) 12:771126. doi: 10.3389/fpsyt.2021.771126.

24. **Manz K**, Batcha AMN, Mansmann U: Regionale und zeitliche Trends der SARS-CoV-2 assoziierten Sterblichkeit in Bayern: Eine altersstratifizierte Analyse über 5 Quartale f¨ur Personen ab 50 Jahren. Gesundheitswesen 2022; 84:e2-e10.

25. Krause D, Chrobok A, Karch S, Keeser D, **Manz KM**, Koch W et al.: Binding potential changes of SERT in patients with depression are associated with remission: A prospective [123I]β-CIT-SPECT study. Exp Clin Psychopharmacol (2022) doi: 10.1037/pha0000566.

26. **Manz KM**, Schwettmann L, Mansmann U, Maier W.: Area Deprivation and COVID-19 Incidence and Mortality in Bavaria, Germany: A Bayesian Geographical Analysis. Front Public Health (2022) 10:927658. doi:10.3389/fpubh.2022.927658

27. Behle N, Kamp F, Proebstl L, Hager L, Riebschläger M, Schacht-Jablonowsky M, Hamdorf W, Neumann S, Krause D, **Manz KM**, et al.: Treatment outcome, cognitive function, and psychopathology in methamphetamine users compared to other substance users. World J Psychiatry (2022) 12(7):944-957. doi:10.3389/fpubh.2022.927658

28. Loidl V, Koller D, Mansmann U, **Manz KM**$^{\#}$: Mapping Regional Differences in Infection Rates for the Coronavirus (COVID-19): Results of a Bayesian Approach to Administrative Districts of Bavaria. Gesundheitswesen (2022). German. doi:10.1055/a-1830-6796.

29. Chen F, Wolf F, **Manz KM**, Fürmetz J, Gonser S, Thaller PH: Quality of Long Standing Radiographs Assessment of the Patella Positon. The Knee (2023). Accepted.

* Shared first authorship

$^{\#}$ last authorship

Submitted/under work

Jansen L, Schwettmann L, Behr C, Eberle A, Holleczek B, Kajüter H, **Manz K**, Peters F, Pritzkuleit R, Schmidt-Pokrzywniak A, Sirri E, Tetzlaff F, Voigtländer S, Arndt V: Trends in cancer incidence by socioeconomic deprivation in Germany in 2007-2018: An ecological registry-based study. International Journal of Cancer (2023). *In revision.*

Warmbein A, Rathgeber I, Hübner L, Mehler-Klamt AC, Huber J, Schroeder I, Scharf C, Gutmann M, Biebl J, **Manz K**, Kraft E, Eberl I, Zoller M, Fischer U: Robot-assisted early mobilization for intensive care unit patients: Feasibility and first-time clinical use. Journal of Intensive Care (2023). *Submitted.*

Manz K, Mansmann U, Nennstiel U, Marzi C, Brockow I. Refer rate for two-stage newborn hearing screening: Systematic review with a Bayesian meta-analysis. *Under work.*

Manz K, Bahr J, Ittermann T, Döhner K, Koschmieder S, Griesshammer M, Greinacher A, Völzke H, Heidel FH. Validation of MPN associated risk factors RDW and lymphocyte ratio as predictors of thromboembolic complications in healthy individuals: analysis on 6853 participants of the SHIP-study. *Under work.*

Talks

03/2021: 67th Biometrical Colloquium. Talk with title “Simultanes regionales Monitorieren von SARS-CoV-2 Infektionen und COVID-19 Sterblichkeit in Bayern durch die standardisierte Infektionsmortalitätsrate (sIFR)”. Authors: K Manz, U Mansmann.

08/2020: FORUM COVID-19, LMU Munich. Talk with title “Challenges regarding COVID-19 data on death and incidences”. Authors: K Manz, U Mansmann.

08/2020: GMDS & CEN-IBS 2020. Talk with title “Modelling spatially heterogeneous *Schistosoma haematobium* infection in Southwestern Tanzania using generalized additive mixed models”. Authors: K Manz, U Berger, I Kroidl, P Clowes, M Gerhardt, W Nyembe, L Maganga, W Assisya, N Ntinginya, M Hoelscher, E Saathoff.

Posters

07/2021: ISCB 42, virtual meeting

07/2019: ISCB 40 Leuven, Belgium

08/2018: ISCB 39 Melbourne, Australia

07/2017: ISCB 38 Vigo, Spain

08/2016: HEC, Munich

03/2009: DPG Frühjahrstagung Greifswald