Die Entwicklung und Validierung eines Prognosescores für Patienten mit chronischer myeloischer Leukämie unter Einbeziehung der zytogenetischen Remission als einer zeitabhängigen Kovariablen

Markus Pfirrmann

Aus dem

Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie

der Ludwig-Maximilians-Universität München

Direktor: Prof. Dr. U. Mansmann

Die Entwicklung und Validierung eines Prognosescores für Patienten mit chronischer myeloischer Leukämie unter Einbeziehung der zytogenetischen Remission als einer zeitabhängigen Kovariablen

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

vorgelegt von

Markus Pfirrmann

aus
Landau in der Pfalz

2007

Mit Genehmigung der Medizinischen Fakultät
der Universität München

Berichterstatter:	Prof. Dr. med. J. Hasford
Mitberichterstatter:	Prof. Dr. rer. nat. H. Schmetzer Priv. Doz. Dr. med. F. Oduncu
Mitbetreuung durch promovierten Mitarbeiter:	Keine
Dekan:	Prof. Dr. med. D. Reinhardt
Tag der mündlichen Prüfung:	09. Mai 2007

Inhaltsverzeichnis

Kapitel 1

Einleitung

1.1 Motivation dieser Arbeit

Seit Beginn der achtziger Jahre wurde zur Behandlung von Patienten mit chronischer myeloischer Leukämie (CML) neben einer Chemotherapie auch Interferon-α (IFN-α) als medikamentöse Therapie in Betracht gezogen. Hinsichtlich einer Verlängerung der Überlebenszeit haben sich Therapien mit IFN-α gegenüber einer reinen Chemotherapie schließlich in mehreren randomisierten Studien als statistisch signifikant überlegen gezeigt [2, 48, 57, 80].

Allerdings stellte man bei mit IFN-α behandelten Patienten auch fest, dass selbst mit dem damals in der CML anerkanntesten Prognosesystem, dem Sokal-Score [105], die Aufteilung der Überlebenswahrscheinlichkeiten in klar unterscheidbare Risikogruppen nicht zufriedenstellend möglich war [41, 80], wodurch bei einem individuellen Patienten nicht ausreichend verlässlich gesagt werden konnte, ob er mit großer Wahrscheinlichkeit von IFN-α profitieren würde oder nicht. Die unbefriedigenden Ergebnisse erklären sich vermutlich zum einen dadurch, dass man sich bei der Entwicklung von Prognosesystemen Mitte der achtziger Jahre nur auf Daten von Patienten stützen konnte, die mit konventionellen Chemotherapien behandelt worden waren. Zum anderen wurden zur Definition der Risikogruppen keine datengestützten statistischen Methoden verwendet. So definierten z.B. Sokal et al. ihre Risikogruppen durch Teilung „into 3 subgroups of roughly similar size, using hazard ratios of 0.8 and 1.2 as boundaries“ [44, 88, 105].

Aus der Notwendigkeit eines neuen Prognosesystems in Form eines validen, kompetenten Entscheidungshelfers für die Anwendung von IFN-α wurde 1994 auf dem Treffen der „European Investigators on Chronic Myeloid Leukaemia“ (E.I.C.M.L.) das „Collaborative CML Prognostic Factors Project“ (C.P.F.P) geboren. Als erstes Ziel dieses Projekts wurde die Entwicklung und Validierung eines Prognosesystems für das Überleben Philadelphia-Chromosom positiver CML-Patienten in chronischer Phase, die mit IFN-α behandelt werden sollen, definiert. Das retrospektiv anhand der Daten von 908 mit IFN-α behandelten Patienten identifizierte und erfolgreich validierte Prognosesystem wurde 1998 von Hasford et al. als „New Prognostic Score“ veröffentlicht [42]. Das Modell erlaubt die Differenzierung dreier Risikogruppen mit statistisch signifikant unterschiedlichen Überlebenswahrscheinlichkeiten im zeitlichen Verlauf. Wie schon der Sokal-Score stützt sich auch der „New Prognostic Score“ auf zum Diagnosezeitpunkt erhobene Patientendaten. Von Bonifazi et al. [18, 90] erstmals anhand externer Patientendaten

validiert, hat sich Hasfords New CML-Score[1] [42] inzwischen mit statistisch signifikanter Risikogruppentrennung bewährt [13, 64].

Aufbauend auf dieser zum Diagnosezeitpunkt erfolgreichen prognostischen Diskriminierung von Überlebenswahrscheinlichkeiten, versprach die Einbeziehung des wichtigen therapeutischen Erfolgskriteriums „zytogenetische Remission“ die Berücksichtigung wertvoller Zusatzinformationen über den Therapieverlauf. Dies zu untersuchen war die Motivation vorliegender Arbeit. In mehreren Studien hatte sich gezeigt, dass die zytogenetische Remission unter IFN-α einen statistisch signifikanten Einfluss auf die Überlebenszeit besitzt [2, 34, 57, 60, 65, 66, 74, 107]. Als zeitabhängiger Faktor einem erweiterten Prognosesystem beigefügt, versprach der zu medizinisch relevanten Zeiten beobachtete Remissionsgrad eine durch aktuelle Informationen adjustierte, noch exaktere Risikogruppendifferenzierung. In diesem Zusammenhang hatte die Internet-Recherche über „Pubmed Medline“ mit den Begriffen „CML - Interferon - prognosis - cytogenetic remission“ zwar Landmarkmodelle in Abhängigkeit vom zytogenetischen Remissionsgrad (z.B. Kloke et al. [66]) oder die „zytogenetische Remission nach einem Jahr“ als eine signifikante Variable im Cox-Modell (Steegmann et al. [107]) angezeigt, aber es war kein Prognosesystem zu identifizieren, welches über die prognostische Information zu Therapiebeginn und die zytogenetische Remission gemeinsam zu mehreren Therapieverlaufszeitpunkten signifikant unterschiedliche Risikogruppen definierte.

Die erfolgreiche Validierung eines erweiterten Prognosesystems vorausgesetzt, gäbe es im Therapieverlauf eine statistisch gestützte Entscheidungshilfe über die Beibehaltung von IFN-α oder die Suche nach einer therapeutischen Alternative. Der Aspekt der Therapieentscheidungshilfe für oder gegen IFN-α hat inzwischen allerdings maßgeblich an Bedeutung verloren. Spätestens mit der Veröffentlichung der Ergebnisse der IRIS-Studie über den randomisierten Vergleich von Imatinib versus IFN-α + niedrigdosiertes Arabinosyl-Cytosin (Ara-C) [79] hat Imatinib, aufgrund signifikant besserer zytogenetischer Remissionserfolge, signifikant geringerer Progressionswahrscheinlichkeiten und eines günstigeren Nebenwirkungsprofils, IFN-α als wichtigste medikamentöse Therapie abgelöst. Nach wie vor wird jedoch auf das Fehlen von Daten zur Langzeitwirkung von Imatinib und die bei einer beträchtlichen Patientenzahl ungebrochene Wirksamkeit von IFN-α hingewiesen [9, 108].

Aus medizinischer Sicht soll vorliegende Arbeit einen Beitrag zu nachfolgenden, unverändert aktuellen Gesichtspunkten liefern:

- Für die erfolgreich (weiter) mit IFN-α behandelten Patienten [9, 108] bleibt ein für den Therapieverlauf in der Prognosegenauigkeit verbessertes Prognosesystem zur Diskriminierung von Risikogruppen mit unterschiedlichen Überlebenswahrscheinlichkeiten unter IFN-α-Therapie nach wie vor interessant.[2]

- Mit Hilfe eines erweiterten Prognosesystems sollte eine für das Überleben unter IFN-α-Therapie besonders günstige Risikogruppe identifiziert werden. Deren geschätzten Überle-

[1] Der Name „New Prognostic Score“ wurde mittlerweile zugunsten der eindeutigeren Bezeichnung „New CML-Score“ geändert. Alternativ wird auch „European Score“ oder „Hasford-Score“ verwendet.

[2] Im Rahmen der jüngsten Analysen vor dem Studientreffen der deutschen CML-Studiengruppe im November 2005 wurde bei der Studie CML III [109] festgestellt, dass von 324 lebenden Patienten 49 (15%) zuletzt IFN-α und kein Imatinib erhielten. Von 90 Patienten, die in der im Juli 2002 begonnenen Studie CML IV [111] während der Pilotphase in den Arm mit IFN-α als Primärtherapie randomisiert wurden, hatten nach einem Jahr Beobachtungszeit 46 von 90 Patienten (51%) die IFN-α-Therapie beibehalten.

benswahrscheinlichkeiten könnten als Gradmesser für Überlebenswahrscheinlichkeiten unter Imatinib dienen.[3]

Aus methodischer Sicht sollten exemplarisch und detailliert die Suche, Entwicklung und Beurteilung der Leistung eines Prognosesystems mit zeitabhängiger Kovariablen beschrieben werden. Dabei wurde u.a. folgenden Aspekten Rechnung getragen:

- Das methodische Vorgehen bei der Suche und Entwicklung eines Prognosesystems mit zeitabhängiger Kovariablen wird ausführlich erläutert. Weil bisher (Ende 2005) weder für IFN-α noch für Imatinib eine international einheitliche minimale Therapiedauer bis zur abschließenden Beurteilung des zytogenetischen Remissionserfolges festgelegt wurde, wird insbesondere die Entwicklung eines im Therapieverlauf zeitlich möglichst flexibel einsetzbaren Prognosesystems diskutiert. Methoden zur Überprüfung des zugrundeliegenden statistischen Modells und zur Beurteilung der Leistung des Prognosesystems in Lern- und Validierungsstichprobe werden vorgestellt (Kapitel 2).

- Die unregelmäßige Datenerhebung der zeitabhängigen Variablen „zytogenetische Remission“ und daraus resultierende mögliche Ergebnisverzerrungen wurden ausführlich untersucht. Es wurde diskutiert, inwiefern die Entwicklung eines Prognosesystems trotz unvollständiger Daten Sinn macht. Nach Überprüfung auf mögliche Störparameter wurde die Analysestichprobe definiert (Kapitel 3).

- Der Weg zur Gewinnung prognostischer Faktoren aus dem multiplen statistischen Modell und die Bildung der Risikogruppen des endgültigen Prognosesystems werden beschrieben. Ohne wesentliche Informations- oder Genauigkeitsverluste sollten die Risikogruppen des Prognosesystems leicht berechenbar und ihre Überlebenswahrscheinlichkeiten mit den verbreiteten statistischen Methoden darstellbar sein (Kapitel 4).

- Anhand der vorliegenden Lern- und Validierungsstichprobe werden in Kapitel 5 Möglichkeiten und Probleme bei der Überprüfung der Leistungsfähigkeit des neuen Prognosesystems dargelegt.

- In Kapitel 6 wird die klinische Bedeutung des identifizierten Prognosesystems diskutiert.

- Kapitel 7 bietet eine Zusammenfassung der Entwicklung des neuen Prognosesystems sowie seiner Einschränkungen und Leistungen.

Alle Analysen wurden mit Unterstützung des Programmpaketes SAS [96] vorgenommen. In der gegebenen SAS Version nicht angebotene methodische Verfahren wurden auf Basis des Zusatzmodules „SAS IML“ programmiert.

Im übrigen Teil von Kapitel 1 werden der wissenschaftliche Kenntnisstand und Definitionen zur chronischen myeloischen Leukämie vorgestellt sowie die Bedeutung von prognostischen Faktoren und Prognosesystemen beschrieben.

[3] Es ist allerdings vorstellbar, dass gerade die für eine IFN-α-Behandlung besonders geeignete Patientengruppe auch unter Imatinib überdurchschnittlich günstige Überlebenswahrscheinlichkeiten haben wird.

1.2 Chronische myeloische Leukämie

1.2.1 Definition und Krankheitsphasen

Die chronische myeloische Leukämie ist eine klonale myeloproliferative Erkrankung, deren Ätiologie wissenschaftlich nicht gesichert ist [84]. Sie entsteht durch eine maligne Transformation der pluripotenten hämatopoetischen Stammzelle. Ihre Inzidenz beträgt 2:100000; in Deutschland treten jährlich etwa 1600 Neuerkrankungen in allen Altersklassen auf [51].

Die Diagnose der CML wird gestellt bei [25, 51, 110]:

- Leukozytose im peripherem Blut (Anzahl der Leukozyten $> 30 \times 10^9/l$)
- Auftreten von myeloischen Vorstufen im peripheren Blut (Myeloblasten, Promyelozyten, Myelozyten, Metamyelozyten)
- Auftreten von Basophilen und Eosinophilen
- hyperzellulärem Knochenmark vereinbar mit einem chronischen myeloproliferativen Syndrom
- Fehlen der Kriterien für das Vorliegen einer akuten Leukämie
- Fehlen der Kriterien für das Vorliegen anderer myeloproliferativer Erkrankungen
- Nachweis des Philadelphia (Ph)-Chromosoms (Patient ist Ph-positiv) oder der BCR-ABL-Translokation (Patient ist BCR-ABL-positiv)

Maligne transformierte Stammzellen enthalten das CML-typische Ph-Chromosom, ein verkürztes Chromosom 22, entstanden durch die reziproke Translokation von distalen Teilen der langen Arme der Chromosomen 9 und 22, t(9;22)(q11;q34). Die Bruchpunkte liegen auf Chromosom 9 im Bereich des ABL-Protoonkogens und auf Chromosom 22 im Bereich des BCR-Gens [51, 69]. Aus der Zusammenlagerung von Teilen der Gene BCR und ABL auf Chromosom 22, der molekularbiologisch nachweisbaren BCR-ABL-Translokation [34, 51, 84, 99], entsteht ein BCL-ABL-mRNA-Transkript. Dieses kodiert ein BCR-ABL-Protein mit erhöhter Tyrosinkinaseaktivität, welche in Kombination mit der Lokalisation des Proteins Signalübertragungsprozesse auslöst, die mit den pathologischen Effekten der CML-typischen Zellen in Zusammenhang gebracht werden [119]. Die Pathogenese der Erkrankung ist nicht vollständig geklärt [51, 84, 119, 120]. Mehr als 93% der Patienten sind Ph- oder BCR-ABL-positiv, nur bei etwa 7% kann weder das eine noch das andere festgestellt werden [84]. Ph- und BCR-ABL-negative Patienten haben einen prognostisch ungünstigeren Krankheitsverlauf [69]. Es wurde inzwischen vorgeschlagen, diese Patienten zukünftig nicht mehr der CML zuzuordnen [30].

Die chronische Phase umfasst die nicht bedarfsgesteuerte Hyperplasie der Zellen der Granulopoese und teilweise der Megakarypoese mit Vermehrung der Zellzahl im peripheren Blut und im Knochenmark sowie das Auftreten einer Splenomegalie [25, 110]. Am Ende der chronischen Phase steht der Übergang in eine instabile, sog. akzelerierte Phase. Sie hat verschiedene Verlaufsformen [25] und wird klinisch häufig durch den Beginn einer Therapieresistenz erkannt. Die terminale Phase der CML besteht aus der Blastenphase (Blastenkrise) [25], welche sowohl im Anschluss an die akzelerierte Phase auftritt als auch das relativ plötzliche Ende der chronischen Phase bedeuten kann. Wie bei der Entwicklung des New CML-Scores [42], basierte die

Abgrenzung der chronischen Phase zu den beiden progredienten Phasen in vorliegender Arbeit auf Kriterien der italienischen Studiengruppe [57] (siehe Abschnitt 3.3).

1.2.2 Die Remissionskriterien

Die Remissionskriterien, an welchen der Erfolg einer Therapie im Krankheitsverlauf gemessen wird, sind die hämatologische Remission, die zytogenetische Remission und die molekularbiologische Remission [25]. Im Rahmen dieser Arbeit wird nur auf die zytogenetische Remission näher eingegangen.

Die zytogenetische Remission wird über den Anteil von Ph-positive Metaphasen an den untersuchten Metaphasen des Knochenmarks definiert [47, 113]:

- Komplette Remission: Eliminierung aller Ph-positive Metaphasen
- Partielle Remission: 1-35% Ph-positive Metaphasen
- Geringe Remission: 36-65% Ph-positive Metaphasen
- Minimale Remission: 66-95% Ph-positive Metaphasen
- Keine Remission: 96-100% Ph-positive Metaphasen

Im Sinne der Vergleichbarkeit wurde - wie in den meisten Publikationen über Studien zu Imatinib [30, 31, 61, 79] üblich - 35 statt 34% [47, 113] Ph-positive Metaphasen als Grenze der partiellen Remission gewählt. Die in der englischsprachigen Fachliteratur [31, 61, 74, 79, 107] zuletzt gemeinsam mit „major cytogenetic remission“ bezeichneten Kategorien „komplette“ und „partielle Remission“ werden hier unter „deutliche Remission“ zusammengefasst. Das Erreichen einer partiellen oder kompletten zytogenetischen Remission unter IFN-α führte zu signifikant verlängerten Überlebenszeiten [2, 34, 57, 60, 74, 107].

Begriffsklärung

Der Begriff „Remission“ beinhaltet an sich bereits die Reduzierung oder Rückbildung krankheitsindizierender Parameter. Zur begrifflichen Differenzierung wurde für vorliegende Arbeit folgende Konvention gewählt: Unter „zytogenetischer Remission (ZR)“ wird immer die als möglicher prognostischer Faktor analysierte Variable verstanden. Zur Bezeichnung einer Remission im eigentlichen Sinne wird der Remissionsgrad mit angegeben: z.B. komplette zytogenetische Remission oder komplette ZR.

1.2.3 Stand der Therapieentwicklung

Seit 1994 wurden mehrere große randomisierte Studien veröffentlicht, in welchen eine statistisch signifikant längere Überlebenszeit von IFN-α gegenüber Hydroxyurea (HU)- und / oder Busulfan (BU)-Monotherapie vorgelegen hatte [2, 47, 57, 80]. Eine Meta-Analyse obiger und weiterer Studien bestätigte einen statistisch signifikanten Überlebensvorteil von IFN-α gegenüber BU wie gegenüber HU [26].
Während von der „Benelux CML Study Group“ [15] für IFN-α + HU und HU-Monotherapie ähnliche Überlebenswahrscheinlichkeiten beobachtet wurden, konstatierten Hehlmann et al. [48] beim Vergleich ihrer Patienten statistisch signifikant höhere Überlebenswahrscheinlichkeiten bei der Kombinationstherapie.

Gemäß der 1997 veröffentlichten Ergebnisse einer französischen Studie scheint die Kombination aus IFN-α und niedrigdosiertem Ara-C der IFN-α-Monotherapie hinsichtlich der Überlebenszeit signifikant überlegen zu sein [38]. Baccarani et al. [13] konnten dieses Ergebnis bei demselben Therapievergleich in ihrer Studie jedoch nicht bestätigen.
Kluin-Nelemans et al. [64] stellten in einer randomisierten Studie nahezu identische Überlebenswahrscheinlichkeiten zwischen Patienten mit niedrigdosiertem und hochdosiertem IFN-α fest.

IFN-α gehört zur Klasse der antiviralen Zytokine. Zytokine sind körpereigene Proteine, die als Vermittler die Kommunikation zwischen Zellen ermöglichen [82]. Eine eindeutige Erklärung der Wirkweise von IFN-α bei CML existiert bis dato nicht [99, 120]. Im Gegensatz zu den Chemotherapien, konnte in IFN-α-Armen bei 6-23% der Patienten größerer randomisierter multizentrischer Studien eine komplette zytogenetische Remission erreicht werden. [2, 13, 15, 37, 38, 47, 48, 58, 64, 80]. Auch bei Patienten mit dauerhafter kompletter zytogenetischer Remission wird IFN-α bisher keine kurative Wirkung zugesprochen. Mediane Überlebenszeiten lagen zwischen 60 und 89 Monaten [2, 15, 37, 47, 48, 58, 60, 64, 83, 114].

Die einzige anerkannt kurative Therapie ist die allogene Stammzelltransplantation (SZT). Zwischen 40-80% der transplantierten Patienten können von CML geheilt werden [50]. Hansen et al. [39] erreichten in ihrer Studie bei einer SZT mit einem HLA-kompatiblen Fremdspender gleich gute Ergebnisse wie bei einer SZT mit einem Verwandtenspender. Hehlmann [49] kalkulierte, dass für etwa 86% der bis 50-jährigen entweder ein Verwandtenspender (30%) oder ein HLA-kompatibler Fremdspender (56%) zu finden sein müßte. Inzwischen nähert man sich der Altersobergrenze von 70 Jahren [16]. Allerdings lag noch Ende der neunziger Jahre die Sterblichkeit innerhalb der ersten Jahre mit bis zu einem Drittel der Patienten im Vergleich zur IFN-α-Therapie relativ hoch [49]. In der deutschen CML-Studie III wurde 2002 nach einer SZT mit Verwandtenspender eine mit der SZT assozierte Mortalität von 27% und nach einer SZT mit Fremdspender von 23% beobachtet [92].

Für Imatinib wurde in der randomisierten Studie von O'Brien et al. [79] bei 87% der Patienten eine deutliche ZR und dabei für 76% aller Patienten eine komplette ZR festgestellt. Außer im höheren Anteil an Patienten mit deutlicher ZR, war der Imatinib-Arm der Kombination IFN-α + Ara-C durch die geringeren Progressionswahrscheinlichkeiten statistisch signifikant überlegen. Jedoch wird auch bei Imatinib-Patienten Therapieresistenz beobachtet [30, 81]. Nach den neuesten Erkenntnissen in vitro verspricht man sich von den zuletzt entwickelten BCR-ABL-Kinaseinhibitoren AMN107 und BMS-354825 einen noch größeren Therapieerfolg als mit Imatinib. Aktuell (2006) werden klinische Studien zu beiden Präparaten durchgeführt [81].

1.3 Bedeutung prognostischer Faktoren - Prognosesysteme

In der Medizin ist ein prognostischer Faktor zumeist ein bei Patienten erhebbarer, zuverlässiger klinischer Parameter, dessen Merkmalsausprägungen einen statistisch signifikanten Zusammenhang mit dem zukünftigen Ergebnis eines interessierenden Zielparameters aufweisen. Dabei wird der Zielparameter i.d.R. so gewählt, dass ein identifizierter prognostischer Faktor einen Erkenntnisgewinn über den zukünftigen Verlauf einer bestimmten Krankheit bei Anwendung einer oder mehrerer dafür vorgesehener Therapien liefert. Für die CML ist der Hauptzielparameter bisher die Überlebenszeit, hier die Überlebenszeit bei einer Therapie mit IFN-α. Prognostische Faktoren dienen einer Vielzahl klinisch wichtiger Aufgabenstellungen und sind aus der modernen Medizin nicht mehr wegzudenken.

Zu den wesentlichen Zielen prognostischer Faktoren gehören [3, 23, 41, 98, 102]:

- Das bessere Verständnis des Krankheitsverlaufes
- Die genauere Vorhersage individueller Krankheitsverläufe mittels verschiedener Risikogruppen
- Die Entwicklung und Anwendung risikoadaptierter Therapien
- Die präzisere Analyse, da die Vergleichbarkeit der Behandlungsgruppen innerhalb einer kontrollierten Studie besser überprüft und die Schätzungen entsprechend adjustiert werden können
- Die Erhöhung der Validität bei vergleichender Analyse und Bewertung der Ergebnisse verschiedener Studien
- Die Erklärung von Abweichungen im Krankheitsverlauf und die Identifikation von Wechselwirkungen zwischen Behandlung und klinischen Parametern
- Die Ermöglichung (mangels Alternative) eine nicht randomisierte Kontrollgruppe zu wählen
- Die Unterstützung bei der Planung neuer Studien, z.B. anhand identifizierter Stratifikationskriterien
- Der Beitrag zur Sicherung der Qualität der Krankenversorgung

In der Regel werden mehrere prognostische Faktoren in einem Prognosesystem kombiniert. Anerkannte, weltweit verbreitete Prognosesysteme sind z.B. der Apgar-Score [12] und der APACHE-Score [67], in der CML der Sokal-Score [105] und der New CML-Score [42].

Im folgenden Kapitel werden Kriterien für die klinische Akzeptanz eines Prognosesystems und das methodische Vorgehen für seine Entwicklung und Validierung beschrieben.

Kapitel 2

Methoden zur Entwicklung und Validierung von Prognosesystemen

Die Bedeutung prognostischer Faktoren wurde in Abschnitt 1.3 herausgestrichen. Nun werden Richtlinien für deren Analyse und für die Entwicklung und Validierung von Prognosesystemen vorgestellt. Bei Darstellung des methodischen Vorgehens wird exemplarisch auf diese Arbeit Bezug genommen.

Mit der Entwicklung eines Prognosesystems verbindet sich die Hoffnung, dass es im vorgesehenen Bereich Anwendung findet und sich bewährt. Die Wahrscheinlichkeit der Anwendung erhöht sich, wenn bei der Modellentwicklung auf die Anforderungen und Bedürfnisse potenzieller Nutzer eingegangen wurde. Nach einer kurzen Begriffsklärung werden daher zunächst die Kriterien für die klinische Akzeptanz eines Prognosesystems angeführt.

Die Stichprobe aller Patienten, deren Daten die Ein- und Ausschlusskriterien für die Analyse prognostischer Faktoren erfüllen, wird fürderhin als „Analysestichprobe“ bezeichnet und sei für Kapitel 2 als gegeben vorausgesetzt.[1]

2.1 Richtlinien zur Gewinnung valider Prognosesysteme

Begriffsklärung

Eine aus prognostischen Faktoren nach einer bestimmten Formel berechnete Zahl wird in vorliegender Arbeit „Risikowert“[2] genannt. Über den Risikowert findet man das individuelle, durch Wahrscheinlichkeiten ausgedrückte Risiko eines Patienten, ein bestimmtes Ereignis zu vermeiden oder zu erfahren. Das individuelle Risiko definiert sich über seine Relation zu den Risiken der übrigen Patienten und hängt von den bei einem Patienten beobachteten Merkmalsausprägungen der in der Prognoseformel enthaltenen prognostischen Faktoren ab. Der Risikowert kann eine metrische oder eine kategoriale Skalierung besitzen. Besitzt er wie beim Sokal-Score [105] oder New CML-Score [42] eine metrische Skalierung, so wird der Risikowert durch die Angabe von Gruppengrenzen zumeist in kategoriale Risikogruppen unterteilt.

[1] Speziell die Berücksichtigung zeitabhängiger Variablen als mögliche Prognosefaktoren erfordert eine komplexe Qualitätsüberprüfung der erhobenen Daten. Diesem Thema ist Kapitel 3 gewidmet.

[2] Im Englischen meist als „score“ bezeichnet.

Unter dem Begriff „Prognosesystem“ sollen in vorliegender Arbeit alle Formeln und Algorithmen verstanden werden, die angewandt werden müssen, um aus den Merkmalsausprägungen identifizierter prognostischer Faktoren die zur Prognose verwendete Risikogruppe eines Patienten zu erhalten. Beim Sokal-Score wie beim New CML-Score umfasst diese Prognosesystemdefinition damit die Berechnung von Risikowerten sowie ihre anschließende Kategorisierung in drei Risikogruppen.

Zur Unterscheidung von einem „Prognosesystem“ werden die bei Anwendung eines statistischen Modells[3] mit Hilfe sog. Selektionsverfahren bestimmten prognostischen Faktoren gemeinsam als „prognostisches Modell“ bezeichnet.

2.1.1 Kriterien für die klinische Akzeptanz eines Prognosesystems

Die exaktesten mathemathischen Prognosesysteme nützen in der Medizin nichts, wenn sie im klinischen Alltag keine Anwendung finden. Wyatt und Altman [121] setzen für den Erfolg eines Prognosesystems die Erfüllung dreier Hauptkriterien voraus: seine klinische Glaubwürdigkeit, die Genauigkeit seiner Ergebnisse und seine Allgemeingültigkeit. In Anlehnung an ihre Arbeit [121], an Laupacis et al. [70], Peduzzi et al. [85] sowie Simon und Altman [103] werden hier die wesentlichen Punkte aufgelistet:

1. **Klinische Glaubwürdigkeit**

   (a) Die klinische Relevanz der hinter dem Prognosesystem stehenden Hypothese sollte verständlich erklärt sein

   (b) Alle klinisch relevanten Parameter sollten als potenzielle prognostische Faktoren bei der Modellentwicklung berücksichtigt worden sein

   (c) Die potenziellen prognostischen Faktoren sollten in Unkenntnis der Merkmalsausprägung des zu prognostizierenden Parameters erhoben worden sein

   (d) Die in Frage kommenden Parameter sollten für den Arzt leicht und mit vertretbarem Zeitaufwand zugänglich sowie reliabel messbar sein, um Vorhersagen und Entscheidungen mit gebotener Schnelligkeit und Verlässlichkeit treffen zu können

   (e) Im Modell sollten willkürliche Grenzsetzungen bei metrischen Parametern möglichst vermieden worden sein

   (f) Das Prognosesystem sollte unmissverständlich beschrieben sein, damit es leicht und fehlerlos angewandt werden kann

   (g) Die Modellvorhersagen sollten aus der Warte des Arztes Sinn machen

2. **Genauigkeit der Ergebnisse**

   (a) Die statistischen Modellannahmen müssen bei der Modellentwicklung überprüft worden sein

   (b) Das Prognosesystem sollte dem Arzt einen Erkenntnisgewinn bieten, mindestens aber so genaue Ergebnisse liefern, wie sie der Arzt auch ohne Modellanwendung hätte erhalten können

[3] Z.B. Cox-Modell [28] oder logistische Regression [53].

(c) Das Prognosesystem sollte möglichst selten Ereignisse vorhersagen, die nicht eintreten (geringe falsch-positiv Rate) und genauso wenig ein Ereignis nicht antizipieren, welches später eintritt (geringe falsch-negativ Rate)

(d) Ohne wesentliche Einschränkung der Genauigkeit, sollten für das Prognosesystem gute Interpretierbarkeit und leichte Anwendbarkeit angestrebt werden

3. **Allgemeingültigkeit**

(a) Wurde ein Prognosesystem auf Basis von Patienten verschiedener Studien entwickelt, sollte auf die Relevanz unterschiedlicher Therapieansätze eingegangen worden sein

(b) Ein- und Ausschlusskriterien für die Zulassung von Studien zur Analysestichprobe sollten beschrieben sein

(c) Die medizinischen Parameter sollten gemäß international üblicher Konventionen definiert worden sein.[4] Der Zeitpunkt, zu welchem Parameterwerte zu erheben (waren) sind, muss einheitlich und eindeutig festgelegt werden (worden) sein

(d) Ein- und Ausschlusskriterien der Patienten, mit deren Daten das Prognosesystem entwickelt wurde und damit auch der Patienten, für welche das Prognosesystem künftig relevant sein soll, müssen unmissverständlich angegeben (worden) sein

(e) Das Prognosesystem sollte prospektiv in Übereinstimmung mit einem Protokoll entwickelt worden sein, nicht retrospektiv anhand bereits existierender Datensätze mit deren möglichen Verzerrungen der Ergebnisse

(f) Das Prognosesystem sollte vor seiner Veröffentlichung in einer weiteren, neuen Patientenstichprobe getestet worden sein - vorzugsweise von der Institution eines anderen Landes (lokale Verallgemeinbarkeit) und zu einem anderen Zeitpunkt (zeitliche Verallgemeinbarkeit)

(g) Mit Hilfe kontrollierter klinischer Studien sollte prospektiv der Effekt der Modellprognosen auf den klinischen Alltag und die Konsequenzen für den Patienten untersucht worden sein

Die aufgeführten Punkte sollten in der statistischen Vorgehensweise bei Entwicklung und Validierung eines Prognosesystems ihre Entsprechung finden. In manchen Situationen berechtigte Abweichungen sollten begründet werden (worden sein).

2.1.2 Statistische Methoden zur Entwicklung und Validierung eines Prognosesystems

Unter Berücksichtigung der beschriebenen Anforderungen für die spätere klinische Akzeptanz eines Prognosesystems, wird in den nachstehenden Abschnitten auf folgende statistische Gesichtspunkte eingegangen (vgl. Altman und De Stavola [5], Simon [102], Simon und Altman [103]):

- Arbeitshypothese und Kriterien für den Vorschlag eines neuen Prognosesystems
- Definition des Hauptzielparameters

[4] So gilt z.B. speziell beim Alter zu beachten, dass üblicherweise immer nach unten abgerundet wird, d.h. nur die vollendeten Lebensjahre angegeben werden. Falsches Aufrunden könnte über den Risikowert u.U. zu einer falschen Risikogruppe führen.

- Studiendesign
- Aufteilung der Daten in Lern- und Validierungsstichprobe
- Umgang mit fehlenden Werten
- Wahl des statistischen Modells zur Identifikation von Prognosefaktoren
- Univariate Analysen in der Lernstichprobe
- Zusammenhänge zwischen den Kovariablen
- Selektion des besten prognostischen Modells in der Lernstichprobe
- Überprüfung auf Einhaltung der Modellannahmen des statistischen Modells
- Untersuchung der Anpassung des prognostischen Modells an die Daten
- Der Weg vom prognostischen Modell zum Prognosesystem
- Beurteilung des Prognosesystems in der Lernstichprobe
- Beurteilung des Prognosesystems in einer unabhängigen Validierungsstichprobe

Die Berücksichtigung der Kriterien 2.1.1 und einer nach 2.1.2 sorgfältig ausgearbeiteten, der Aufgabenstellung angemessenen Methodik fördern die Validität und Reliabilität eines Prognosesystems, garantieren diese aber nicht. Da Prognosesysteme für sehr unterschiedliche Situationen und Zwecke konstruiert werden, kann es auch wohlbegründete Abweichungen von den hier vorgeschlagenen Richtlinien geben.

2.2 Arbeitshypothese und Kriterien für den Vorschlag eines neuen Prognosesystems

Hasford et al. [42] identifizierten bei der Entwicklung ihres New CML-Scores für das Überleben IFN-α-behandelter Patienten sechs statistisch signifikante Faktoren: Alter, Milzgröße, Thrombozyten sowie die Anteile von Blasten, Eosinophilen und Basophilen im peripheren Blut. Den Risikowert des Scores berechnet man nach der Formel

Risikowert =

$$
\begin{aligned}
1000 \times (\quad & 0,6666 \times \text{Alter [1, falls Alter in vollendeten Jahren} \geq 50 \text{ Jahre; 0, sonst]} \\
+ \; & 0,0420 \times \text{Milzgröße [cm unter dem Rippenbogen]} \\
+ \; & 0,0584 \times \text{Blasten [\%]} \\
+ \; & 0,0413 \times \text{Eosinophile [\%]} \\
+ \; & 0,2039 \times \text{Basophile [1, falls Basophile} \geq 3\%\text{; 0, sonst]} \\
+ \; & 1,0956 \times \text{Thrombozyten [1, falls Thrombozyten} \geq 1500 \times 10^9/\text{L; 0, sonst]}).
\end{aligned}
$$

Patienten gehören mit Risikowerten ≤ 780 zur Niedrigrisikogruppe, mit Risikowerten > 780 und ≤ 1480 zur mittleren Risikogruppe und mit Risikowerten > 1480 zur Hochrisikogruppe.

Nun hatte sich in mehreren Studien gezeigt, dass auch das Ergebnis der zytogenetischen Remission unter IFN-α einen statistisch signifikanten Einfluss auf die Überlebenszeit besitzt [2, 34, 57, 60, 74, 107]. Mit dem New CML-Score als dem besten bekannten Prognosesystem, welches sich ausschließlich auf zum Diagnosezeitpunkt erhobene Daten stützt, ergab sich daraus folgende

Arbeitshypothese:

> Unter Verwendung der zeitabhängigen Variablen „zytogenetische Remission" zusätzlich zu den Baselinevariablen lässt sich ein Prognosesystem finden, auf dessen Basis zu verschiedenen, medizinisch relevanten Verlaufszeitpunkten statistisch signifikant unterschiedliche Risikogruppen bzgl. der Überlebenszeit definiert werden können. Dabei führt das neue Prognosesystem im Vergleich zum New CML-Score zu einem erkennbaren Informationsgewinn.

Die Variable „zytogenetische Remission" ist „zeitabhängig", weil im Prognosesystem ihr Variablenwert in Abhängigkeit vom Beobachtungszeitpunkt berücksichtigt werden sollte. Die Merkmalsausprägung eines zeitabhängigen prognostischen Faktors steht (zumindest z.T.) zeitlich parallel zum Hauptzielparameter Überlebenszeit unter Beobachtung. Im Gegensatz dazu, geht bei den zeitunabhängigen, zum Diagnosezeitpunkt erhobenen Variablen nur ein vor Beobachtungsbeginn der Überlebenszeit erhobener, von der weiteren Beobachtungszeit „unabhängiger" Wert in das Prognosesystem ein, weswegen sie auch als „Baselinevariablen" bezeichnet werden.
Als „medizinisch relevant" wurden vorab die Verlaufszeitpunkte 12, 15, 18, 21 und 24 Monate nach Therapiebeginn erachtet.[5] Zwölf Monate seit Beginn einer IFN-α-Therapie wurde als Minimum gewählt, um für die Beurteilung eines prognostischen Einflusses ausreichend Patienten mit deutlicher ZR beobachtet zu haben (Fallzahl). Dagegen wird später als 24 Monate nach Start einer IFN-α-Therapie nicht mehr bei vielen Patienten eine erste deutliche Remission registriert. Zudem galten auch vor der Zulassung von Imatinib zwei Jahre als eine lange Zeit, um einen deutlichen Remissionserfolg von IFN-α abzuwarten und dann eine Therapieentscheidung zu treffen.[6]
Unter einem „erkennbaren Informationsgewinn" wird z.B. die Identifikation einer höheren Anzahl von Niedrig- und Hochrisikopatienten verstanden oder die berechtigte Etablierung einer vierten Risikogruppe.

Kriterien für den Vorschlag eines neuen Prognosesystems

a) Das neue Prognosesystem in der Lernstichprobe

Die Entwicklung eines Prognosesystems ist ein exploratives Vorgehen. Die Patientenstichprobe, welche als Datenbasis für die Identifikation eines Prognosesystems dient, wird hier als „Lernstichprobe" bezeichnet. Noch vor der Bildung von Risikogruppen, können bereits auf der Basis der aus dem identifizierten Modell errechneten Risikowerte prognostizierte Überlebenswahrscheinlichkeiten (Formel (2.14), s.u.) untersucht werden. Bei zufriedenstellenden Prognoseergebnissen

[5] Ohne Einschränkung seiner prognostischen Differenzierungsqualität, sollte das Prognosesystem natürlich auch zu beliebigen anderen Zeitpunkten zwischen zwölf und 24 Monaten verwendet werden können - und möglichst natürlich auch vor und bis zu einem Jahr nach diesem Zeitraum.

[6] Die im Rahmen dieser Arbeit gewählte Definition von „medizinisch relevant" dient einer sinnvollen kritischen Betrachtung eines Prognosesystems zu besonders wichtigen Verlaufszeitpunkten und hat keinen Allgemeingültigkeitsanspruch. Natürlich ist für einen Patienten eine deutliche ZR auch außerhalb des zweiten Therapiejahres „medizinisch relevant".

wird das Prognosesystem durch die Definition von Risikogruppen komplettiert. Um die Relevanz und die Reliabilität der prognostizierten Risikogruppen zu unterstützen, sollte jede Risikogruppe eines identifizierten Prognosesystems ab dem gewählten Prognosezeitpunkt wenigstens 10% aller Patienten umfassen. Auch für die enstandenen Risikogruppen können prognostizierte Überlebens- (Formel (2.14)) oder Sterbewahrscheinlichkeiten (Formel (2.16)) betrachtet werden. Hinsichtlich der Überlebenswahrscheinlichkeiten ab jedem der fünf medizinisch besonders relevanten Prognosezeitpunkte sollten für die verschiedenen Risikogruppen des in der Lernstichprobe identifizierten Prognosesystems in vorliegender Arbeit folgende Bedingungen erfüllt sein:

- Die Patienten einer höheren Risikogruppe sollten über den zeitlichen Verlauf erkennbar geringere Überlebenswahrscheinlichkeiten als Patienten einer niedrigeren Risikogruppe besitzen und die Kaplan-Meier-Kurven [63] sollten sich nicht überschneiden[7]
- Der p-Wert zum Logrank-Test [76] über alle Risikogruppen sollte zu jedem Zeitpunkt $\leq 0,005$ betragen
- Die p-Werte zum Logrank-Test für die paarweisen Vergleiche der Risikogruppen sollten zu jedem Zeitpunkt $\leq 0,05$ betragen

Das Untersuchen der Überlebenswahrscheinlichkeiten ab einem Zeitpunkt t für die zu t noch unter Beobachtung stehenden Patienten bezeichnet man als Landmark-Analyse mit dem Zeitpunkt t als Landmark [8]. Die Überlebenswahrscheinlichkeiten ab dem Zeitpunkt t werden (in Abhängigkeit von der Risikogruppenzugehörigkeit) nach der Kaplan-Meier-Methode berechnet [63]. In die Kaplan-Meier-Kurven wurden zur Beschreibung der Schätzgenauigkeit und der Einschätzung der Kurvenabstände zu medizinisch sinnvollen Zeitpunkten Konfidenzintervalle eingezeichnet. Die Auswahl einzelner Zeitpunkte wurde Konfidenzbändern vorgezogen, da deren Darstellung beim Vergleich mehrerer Überlebenskurven schnell zu unübersichtlichen Graphiken führt. Die Berechnung der Standardabweichung für 95%-Konfidenzintervalle (95%-K.I.) basierte auf Greenwoods Formel [36, 40, 96].[8] In vorliegender Arbeit wurden Konfidenzintervalle zu den Zeitpunkten drei, sechs und neun Jahre eingezeichnet. Die für die endgültigen Kurvendarstellungen getroffene Wahl ergab sich aus der Äquidistanz der Zeitpunkte ab Therapiebeginn, einer medianen Überlebenszeit von sechs Jahren bei der am Ende von Kapitel 3 aufbereiteten Lernstichprobe und weil mit 91 Patienten nach Jahr 9 fast noch doppelt so viele Patienten für die (stabilere) Schätzung der späten Überlebenswahrscheinlichkeiten „auf die Untergruppen verteilt werden konnten“ als nach 10 Jahren ($n = 50$).[9]
Der zugehörige Test für den Vergleich von Überlebenswahrscheinlichkeiten verschiedener Risikogruppen ist der Logrank-Test [76]. Als Voraussetzung für die Anwendung des Logrank-Tests dürfen sich die ab t berechneten Kaplan-Meier-Kurven nicht kreuzen. Die Logrank-Tests wurden

[7] Wie alle hier aufgestellten Bedingungen, sind diese Forderungen nur bei ausreichender Fallzahl und Beobachtungsdauer zu gewährleisten. Wenn nur (noch) wenige Patienten unter Beobachtung stehen und der am längsten beobachtete Patient verstarb, ist eine Kurvenüberschneidung mit Kurven höherer Risikogruppen und darin länger beobachteten Patienten nicht vermeidbar, aber i.d.R. von keiner statistischen Bedeutung.

[8] Die K.I. nach Greenwood können nur zu Ereigniszeitpunkten (neu) berechnet werden. Wird in einer Kaplan-Meier-Kurve der vorliegenden Arbeit ein 95%-K.I. zwischen zwei Ereigniszeitpunkten angegeben, so basiert seine Berechnung auf dem früheren Ereigniszeitpunkt. Mit jeder Zensierung die zwischen dem angegebenen Zeitpunkt und dem vorangegangenen Ereigniszeitpunkt liegt, wird das eingezeichnete 95%-K.I. das tatsächliche 95%-K.I. ein Stückchen mehr unterschätzen, was aber bei ausreichender Zahl beobachteter Ereignisse sowie weiter unter Beobachtung stehender Patienten i.d.R. nur geringe Unterschätzungen zur Folge hat. Solange zum Berechnungszeitpunkt noch wenigstens 20 Patienten unter Beobachtung stehen, besitzten die K.I. ausreichende asymptotische Genauigkeit [40].

[9] Vgl. Abschnitt 4.1.1, insbesondere Abbildung 4.1.

vorgeschlagen, um die Diskriminierungsqualität eines identifizierten Prognosesystems mit weiteren statistischen Kriterien beurteilen zu können. Für jeden der sieben Tests über alle Risikogruppen wurde vorab das Signifikanzniveau $\alpha = 0,005$ gewählt, da man für die Risikogruppen eines guten Prognosesystems sehr unterschiedliche Überlebenswahrscheinlichkeiten erwarten durfte.[10] Im Falle der paarweisen Vergleiche wurde $\alpha = 0,05$ festgesetzt, weil bei guter Diskriminierung der Überlebenswahrscheinlichkeiten zwischen zwei benachbarten Risikogruppen auch ein - aufgrund kleiner Fallzahlen - nicht extrem niedriger p-Wert akzeptabel sein konnte.

Das Signifikanzniveau $\alpha = 0,05$ wurde generell für die in der Lernstichprobe durchgeführten Tests gewählt. Wird in ein und derselben Stichprobe ohne p-Wert-Adjustierung mehrfach getestet, so erhöht sich das Signifikanzniveau und damit die Wahrscheinlichkeit zufällig signifikanter Testergebnisse. Da wegen des explorativen Vorgehens in der Lernstichprobe die p-Werte i.d.R. nicht adjustiert wurden, sind die Testergebnisse in der Lernstichprobe als deskriptiv oder „Hypothesen generierend“ zu verstehen [98]. Wurde eine p-Wert-Adjustierung oder eine Änderung des Signifikanzniveaus vorgenommen, ist dies nachfolgend explizit angegeben.
Entsprechend besitzen auch die nicht adjustierten p-Werte zu den obigen Logrank-Tests nur beschreibenden Charakter, zumal man von jedem Prognosesystem ohnehin annehmen sollte, dass es unterschiedliche Risiken in der Stichprobe, in der es entwickelt wurde, deutlich zu erkennen vermag. Es bedurfte also weiterer und strengerer Kriterien.

b) Das neue Prognosesystem im Vergleich mit einem früher etablierten Prognosesystem

Das einzige bei Patienten mit CML in chronischer Phase für die Überlebenswahrscheinlichkeiten unter IFN-α-Therapie entwickelte und in unabhängigen, zweiten Stichproben validierte Prognosesystem war der New CML-Score [13, 18, 42, 44, 64, 90].[11] Hielt ein neues Prognosesystem den ersten Prüfungen in der Lernstichprobe stand, sollte es daher anschließend in gemeinsamen Stichproben mit dem New CML-Score verglichen werden. Maßgeblich waren die Risikogruppen und Überlebenswahrscheinlichkeiten ab den fünf gewählten, medizinisch relevanten Verlaufszeitpunkten.[12] Ein neues Prognosesystem sollte unter Ausnutzung der Remissionsvariablen im Vergleich zum New CML-Score

- einen zusätzlichen Informationsgewinn durch die Identifikation einer höheren Anzahl von Niedrig- und Hochrisikopatienten oder einer zusätzlichen Risikogruppe bieten

- und Überlebenswahrscheinlichkeiten über den zeitlichen Verlauf stärker diskriminieren

[10] Vgl. z.B. Hasford et al. [42].

[11] Thomas et al. [117] konnten zwar eine deutliche Trennung der Überlebenswahrscheinlichkeiten zwischen mittlerer Risikogruppe und Hochrisikogruppe entdecken und sprachen auch von einer Validierung des New CML-Scores, bezogen sich aber auf nur 82 Patienten unter 60 Jahren, wovon lediglich sechs Patienten zur Hochrisikogruppe gehörten.
Anstatt die Kaplan-Meier-Kurven zu allen drei Risikogruppen zu zeigen, verglichen Huntly et al. [55] Hochrisikogruppe und Nicht-Hochrisikogruppe in 210 Patienten, von welchen jedoch nur 119 mit IFN-α behandelt worden waren. Die beiden Gruppen besaßen statistisch signifikant unterschiedliche Überlebenswahrscheinlichkeiten. Aufgrund der willkürlich erscheinenden Gruppenzusammenfassung und der unterschiedlichen Therapien konnte allerdings nicht von einer Validierung des New CML-Scores gesprochen werden.

[12] Weder konnte zum Diagnosezeitpunkt oder zu Therapiebeginn eine durch IFN-α induzierte deutliche ZR vorliegen, noch machte es Sinn, an einem für diese Zeitpunkte etablierten, mehrfach validierten Prognosesystem ohne wohlbegründete Veranlassung Veränderungen vorzunehmen.

Der Vergleich der beiden Systeme war hier zunächst in einer gemeinsamen Lernstichprobe geplant.[13] Neben einer Beschreibung der Risikogruppen beider Prognosesysteme hinsichtlich der Patientenzahlen und Überlebenswahrscheinlichkeiten, wurde ein Kriterium für die χ^2-verteilten Teststatistiken der Logrank-Tests überlegt. Informationsgewinn und die stärkere Diskriminierung der Überlebenswahrscheinlichkeiten sollten sich beim neuen Prognosesystem in einer Teststatistik niederschlagen, welche im Vergleich zum New CML-Score im Falle derselben Risikogruppenzahl um einen Wert ≥ 4 erhöht war. Die von 3,84 auf die natürliche Zahl 4 aufgerundete Erhöhung wurde gewählt, weil damit die Differenz der beiden χ^2-verteilten Teststatistiken verglichen mit der χ^2-Verteilung zum Freiheitsgrad 1 gerade jenseits des 95%-Perzentils liegt und die aufgerundete Zahl 4 eine relevante Erhöhung der Teststatistik bei gleichbleibender Risikogruppenzahl „auf einen Blick“ erkennen lässt. Als Hinweis auf eine relevante Erhöhung der Teststatistik unter gleichzeitiger Berücksichtigung einer zusätzlichen Risikogruppe diente mit einer Differenz von ≥ 6 (aufgerundet von 5,99) ein Wert direkt über dem 95%-Perzentil der χ^2-Verteilung mit 2 Freiheitsgraden. Die so definierten „relevanten Erhöhungen“ sollten zu allen fünf Prognosezeitpunkten zwischen 12 und 24 Monaten beobachtet werden. In Anbetracht der vorliegenden Ereigniszahlen und Beobachtungszeiten waren diese Anforderungen an ein neues Prognosesystem durchaus vertretbar.

c) Das neue Prognosesystem in der Validierungsstichprobe

Überzeugte das Prognosesystem in der gemeinsamen Lernstichprobe auch im Vergleich mit dem bisher besten Prognosesystem, so gab es Anlass, das neue System der notwendigen Überprüfung in einer unabhängigen Validierungsstichprobe zu unterziehen. Dabei wurden an das neue Prognosesystem prinzipiell dieselben Anforderungen wie in der Lernstichprobe gestellt. Die Einschränkung „prinzipiell“ weist darauf hin, dass auch in einer Validierungsstichprobe - speziell für die paarweisen Vergleiche der Risikogruppen - ausreichende Fallzahlen und Beobachtungsdauern erforderlich waren. Konnte unter diesen Voraussetzungen das Prognosesystem gemäß der prognostizierten Überlebens- (2.14) und Sterbewahrscheinlichkeiten (2.16) sowie der unter a) angegebenen Kriterien zufriedenstellen, so stand der nach b) durchzuführende Vergleich mit dem bisher etablierten Prognosesystem auf dem Programm. Ein überzeugendes Argument für das neue Prognosesystem wäre insbesondere, wenn die frühere Lernstichprobe des etablierten Prognosesystems zugleich eine unabhängige Validierungsstichprobe des neuen Prognosesystems wäre und letzteres im Vergleich trotzdem erheblich besser abschneiden würde.

Hatte das neue Prognosesystem den Überprüfungen nach a), b) und c) standgehalten, konnte man daran denken, der wissenschaftlichen Gemeinde seine Anwendung in den dafür vorgesehenen Situationen und Patientenstichproben vorzuschlagen. Mit seiner Anwendung in immer wieder neuen Stichproben wird ein Prognosesystem fortgesetzte Validierung erfahren oder auch Anlass bieten, über seine Verbesserung nachzudenken, z.B. weil sich die Behandlung geändert hat oder wichtige neuere Parameter nicht bei seiner Entwicklung berücksichtigt werden konnten. Wegen fortschreitenden Erkenntnisgewinns ist die Aktualität von Prognosesystemen bei vielen Krankheiten zeitlich begrenzt und wiederholtes Arbeiten an einer Verbesserung von Prognosen unabdingbar.

[13] Damit das neue Prognosesystem gegenüber dem Herkömmlichen keinen offensichtlichen Vorteil besaß, sollten die Patienten zur Lernstichprobe beider Prognosesysteme gehören. Lägen nur Patienten der Lernstichprobe des neuen Prognosesystems vor, erhielte man immerhin ein verwertbares Ergebnis, wenn das neue Prognosesystem *trotzdem* eine unbefriedigende Leistung zeigte und verbessert oder verworfen werden müsste.

2.3 Definition des Hauptzielparameters

Der Hauptzielparameter Überlebenszeit berechnete sich aus der Anzahl der Tage zwischen dem Datum der ersten IFN-α-Gabe und entweder dem Todestag oder dem Datum des letzten Kontaktes zum Patienten. Bei Patienten, welche eine allogene Knochenmarktransplantation erhalten hatten, wurde zwischen zwei Fällen unterschieden. Befand sich der Patient vor der KMT in erster chronischer Phase, wurde seine Überlebenszeit zum KMT-Zeitpunkt zensiert. In den übrigen Fällen, d.h. bei KMT in Blastenphase, akzelerierter, zweiter oder späterer chronischer Phase wurde zum Zeitpunkt der KMT nicht zensiert. Hier hatte die IFN-α-Behandlung als Primärtherapie versagt. Sie hatte den Patienten nicht in der ersten chronischen Phase halten können und war mitverantwortlich an den ungünstigeren Überlebenswahrscheinlichkeiten bei einer KMT in fortgeschrittener Phase. Ungeachtet eines Erfolges von IFN-α wurde nur die allogene KMT in erster chronischer Phase angewandt. Der Wechsel zur reinen Chemotherapie oder autologen Transplantation geschah aufgrund unbefriedigender Ergebnisse von IFN-α.[14] Die Überlebenszeiten der betroffenen Patienten wurden daher nicht zensiert.

2.4 Studiendesign

Wyatt und Altman [121] schlagen für Entwicklung und Validierung eines Prognosesystems die Durchführung einer prospektiven Studie vor.[15] Prospektive Studien sind jedoch sehr zeitaufwändig, speziell wenn der Median des Hauptzielkriterium Überlebenszeit über fünf Jahre beträgt.[16] Stehen retrospektiv Studien mit guter Datenqualität und längerem Follow-up des Hauptzielparameters zur Verfügung, so dass man damit ein reliables Prognosesystem finden müßte, könnte man einen längeren Aufschub des Projektes auch als ethisch unverantwortlich betrachten. Entsprechend wurden retrospektiv die Daten bereits vorliegender Studien genutzt.

Um die Gefahr einer systematischen Verzerrung der Ergebnisse („Bias“) einzuschränken, wurden nur prospektiv geplante Studien mit einheitlichem Studienprotokoll berücksichtigt. Für die Patienten wurden gemeinsame Ein- und Ausschlusskriterien festgelegt.[17] Die Daten zu den potenziellen prognostischen Faktoren wurden in Unkenntnis sowohl des später zu beobachtenden Hauptzielparameters als auch der Absicht, ein Prognosesystem zu entwickeln oder zu validieren, erhoben.

Ansari et al. [11] wiesen beim Therapievergleich der deutschen CML-Studie I auf das Problem einer allzu vorzeitigen Absetzung von IFN-α hin. Mit dem Ziel, ein Prognosesystem für Patienten unter IFN-α-Therapie zu entwickeln, war es sinnvoll, dazu auch nur Daten von Patienten zu verwenden, die IFN-α erhalten hatten. Entsprechend wurden Patienten, die nie mit IFN-α behandelt worden waren von der Analysestichprobe ausgeschlossen und insofern vom „Intention-to-treat“-Prinzip abgewichen.

[14] Die Datenbank wurde im Herbst 1999 geschlossen. Bei den vorliegenden Patienten spielte bis dahin die Autotransplantation als geplante Primärtherapie ebenso wenig eine Rolle wie der spätere Einsatz von Tyrosinkinaseinhibitoren.

[15] Vgl. Abschnitt 2.1.1, 3 (e) und 3 (g).

[16] Vgl. Abschnitt 3.4.

[17] Vgl. Kapitel 3.

2.5 Aufteilung der Daten in Lern- und Validierungsstichprobe

Die überzeugendste Methode, den Erfolg eines Prognosesystems zu demonstrieren, ist der Beweis seiner Diskriminierungsfähigkeit in einer von seiner Entwicklung unabhängigen Patientenstichprobe [42, 56, 70, 88, 102, 103]. Ohne diese Fähigkeit zur Identifikation deutlich unterscheidbarer Risikogruppen in unabhängigen Patientenstichproben macht ein Prognosesystem keinen Sinn und bedarf, im günstigsten Fall, einer Überarbeitung. Allerdings ist eine große Analysestichprobe vonnöten, um ein Prognosesystem mit Hilfe des einen Teils der Patientendaten zu entwickeln (Lernstichprobe) und anhand des anderen Teils der Patientendaten zu überprüfen (Validierungsstichprobe).

Die prozentuale Zuteilung von Patienten der Analysestichprobe an die Validierungsstichprobe hängt von der Fallzahl der ersteren und den von Lern- und Validierungsstichprobe zu erfüllenden Aufgaben ab. Zunächst müssen genügend Patienten in der Lernstichprobe verbleiben, um ein reliables Prognosesystem entwickeln zu können. Simon und Altman [103] empfahlen bei der Modellentwicklung im Vergleich zur Anzahl der potenziellen prognostischen Faktoren wenigstens die zehnfache Anzahl von Ereignissen (z.B. Todesfällen) als „vernünftigen Standard“. Auf Basis der für die Entwicklung des New CML-Score gesammelten Daten [42], wurde für die Aufteilung der Daten in Lern- und Validierungsstichprobe von folgenden Annahmen ausgegangen:

- Nach Schließen der Datenbank im Herbst 1999 würde die Analysestichprobe aus etwa 1000 Patienten bestehen mit
- vollständigen Daten zu 10 interessierenden, potenziell prognostischen Variablen,
- mindestens 440 beobachteten Todesfällen und
- einer medianen Überlebenszeit von ungefähr sechs Jahren

Dann entspräche eine Validierungsstichprobe von 20-30% aller Patienten und Todesfälle einer methodisch sinnvollen Patientenaufteilung. Durch den Verbleib von 70-80% der Patienten und Todesfälle in der Lernstichprobe würde mit hoher Wahrscheinlichkeit sichergestellt, dass alle tatsächlich relevanten prognostischen Faktoren identifiziert [56] und die Standardabweichungen geschätzter Modellkoeffizienten möglichst klein gehalten werden können. Auch dem von Simon und Altman [103] vorgeschlagenen Mindestverhältnis von Ereignissen zu untersuchten Variablen würde Rechnung getragen. Andererseits wäre die Validierungsstichprobe groß genug, um zumindest ab jedem Verlaufszeitpunkt innerhalb der ersten beiden Jahre nach Therapiebeginn eine zuverlässige Beurteilung des in der Lernstichprobe identifizierten Prognosesystems zu erlauben.

Bei Überlegungen, ob und wie man seine Analysestichprobe aufteilt, sollte das Hauptaugenmerk immer auf die Lernstichprobe gerichtet sein. Ohne die berechtigte Annahme, unter den gegebenen Variablen alle prognostischen Faktoren identifiziert und die für ein Prognosesystem bedeutungsvollen Koeffizienten mit ausreichender Genauigkeit geschätzt zu haben, ist jedes Prognosesystem obsolet und damit auch seine Überprüfung durch eine wie immer geartete Validierungsstichprobe. In Abhängigkeit der Fallzahl der Analysestichprobe, des Anteils an beobachteten Ereignissen, der Anzahl der zu untersuchenden Variablen und der Stichprobenreduzierung durch fehlende Werte zu diesen Variablen ist es von Vorteil, ggf. auf das „Beiseitelegen“ einer Validierungsstichprobe zu verzichten. Auch könnte sich für die Beurteilung eines Prognosesystems die Validierungsstichprobe zu späteren Landmarkzeitpunkten als zu klein erweisen. Das entwickelte Prognosesystem kann dann mittels einer zu einem späteren Zeitpunkt zugänglichen,

adäquaten Stichprobe überprüft werden.

Zwischen den Patienten verschiedener Studien existiert biologische Heterogenität. Da sich ein Prognosesystem später dieser Heterogenität in bei seiner Entwicklung unbeteiligten Studien stellen muss, empfiehlt es sich, bei der Aufteilung der Analysestichprobe Studien als Stichprobeneinheit zu wählen. Idealerweise werden einige Studien zufällig aus der Grundgesamtheit aller vorliegenden Studien gezogen, bis die Validierungsstichprobe einen Umfang von 20-30% der Analysestichprobe erreicht hat [42, 88]. Im Gegensatz dazu würde das zufällige Ziehen einzelner Patienten aus der Analysestichprobe zwei einander zu ähnliche Stichproben erzeugen und bei Überprüfung des Prognosesystems in der Validierungsstichprobe einen ersten möglichen Hinweis auf seine Allgemeingültigkeit sofort in Frage stellen [88]. Mit dem Ziel, in das Prognosesystem vorliegender Arbeit ein gewisses Maß an biologischer Heterogenität zwischen verschiedenen Studien mit einzubeziehen, sollten in der Lernstichprobe wenigstens fünf verschiedenen Studien mit jeweils mehr als 50 Patienten und bereits erreichter medianer Überlebenszeit verbleiben.

Die Voraussetzungen für die Zulassung von Studien und Patientendaten zur Analysestichprobe und deren Aufteilung in Lern- und Validierungsstichprobe werden in Kapitel 3 beschrieben.

2.6 Umgang mit fehlenden Werten

Zu allen Analysen unter Beteiligung der zeitunabhängigen Variablen mit ihren zum Zeitpunkt der Diagnose erhobenen Werten wurden nur Datensätze ohne fehlende Werte zugelassen. Dieses Verfahren entspricht der gängigen Praxis [5]. Aufgrund der Datenerhebung zum Diagnosezeitpunkt konnte für solche zeitunabhängigen Variablen ein Zusammenhang zwischen dem Fehlen von Werten und der Überlebenszeit ausgeschlossen werden. Ob die Annahme des nichtzufälligen Fehlens von Daten berechtigt war, wurde durch den Vergleich von Überlebenskurven untersucht.[18]

Zur Sicherstellung einer genügend großen Fallzahl in der Lernstichprobe wurden für die multiple Analyse nur solche zeitunabhängigen Variablen zugelassen, zu welchen zumindest für 90% der Patienten Daten vorlagen. Bei der Entwicklung des New CML-Scores hatte sich bereits gezeigt, dass man damit auf keine der von den Klinikern de facto als wichtig erachteten Variablen würde verzichten müssen [42].[19] Die hohe Prozentzahl war erforderlich, da die Aufnahme all dieser Variablen in ein gemeinsames multiples Modell den Anteil der Patienten mit kompletten Daten weiter verringern würde. Bei der auf jeden Fall zu untersuchenden zeitabhängigen Variablen „zytogenetische Remission“ war ohnehin schon mit einem Erhebungsgrad von weniger als 90% zu rechnen. Neben einer ausreichenden Fallzahl für die Lernstichprobe, versprach das bzgl. der zeitunabhängigen Variablen gewählte Vorgehen die Anwendbarkeit eines identifizierten Prognosesystems auf die Mehrzahl der in den neunziger Jahren erfassten Patienten.

Im Falle der zytogenetischen Remission wurde zu Therapiebeginn für alle Patienten grundsätzlich vom Zustand „keine Remission“ ausgegangen. Für die prognostischen Analysen sollten nur Daten von Patienten berücksichtigt werden, bei welchen im Therapieverlauf verlässliche Ergebnisse zu den Remissionsvariablen festgehalten worden waren. Kapitel 3 befasst sich mit der be-

[18] Vgl. Kapitel 3.

[19] In Anbetracht einer Erhebung bei nur 56% aller Patienten [42] wurden z.B. die Blasten im Knochenmark früher offensichtlich nicht als *de facto* wichtig erachtet. Eine vollständigere Erhebung hätte damals möglicherweise zur Aufnahme des Parameters in das finale prognostische Modell geführt.

sonderen Problematik der Daten zu der Remissionsvariablen. Dort wird u.a. beschrieben, ob ein zufälliges Fehlen von Remissionsdaten vorlag, ob ein Zusammenhang zwischen Erhebungshäufigkeit und Remissionsergebnis bestand und welchen Einfluss Störparameter wie „Studie“ und „Art der IFN-α-Therapie“ besaßen.

2.7 Wahl des statistischen Modells zur Identifikation von Prognosefaktoren

2.7.1 Vorüberlegungen zur zeitabhängigen Variablen „zytogenetische Remission“

Die Remissionsstadien

Die zytogenetische Remission (ZR) wird in bis zu fünf Remissionsstadien eingeteilt.[20] Aufgrund signifikant höherer Überlebenswahrscheinlichkeiten werten die Kliniker eine deutliche Remission und dabei insbesondere die angestrebte komplette Remission als therapeutischen Erfolg [2, 34, 57, 60, 74, 107]. Die Kategorien „keine ZR“, „minimale ZR“ und „geringe ZR“ werden oft als „keine deutliche Remission“ zusammengefasst. Für die Daten der Lernstichprobe wurde vermutet, dass sie statistisch signifikant günstigere Überlebenswahrscheinlichkeiten einerseits von Patienten mit partieller ZR gegenüber Patienten ohne deutliche ZR und andererseits von Patienten mit kompletter ZR gegenüber Patienten mit partieller ZR zeigen würden.
Unter diesen Annahmen schienen für ein prognostisches Modell zwei Ereignisse von wesentlicher Bedeutung: das Erreichen einer partiellen ZR und das Erreichen einer kompletten ZR. In den Studienprotokollen war die Häufigkeit der zytogenetischen Diagnostik bis hin zu einer vierteljährlichen Durchführung vorgesehen [48]. Hätte man die zytogenetischen Ergebnisse in der geplanten Qualität und Häufigkeit erhalten, so wäre eine gute Datenbasis auch für die Auswertung weiterer, nachgeordneter Ereignisse wie z.B. „Verlust einer deutlichen ZR“ oder „Wiedererlangung einer partiellen ZR“ vorhanden gewesen. Da die tatsächliche Häufigkeit des Vorliegens verwertbarer Daten zur Zytogenetik jedoch geringer und z.T. sehr unterschiedlich war[21], empfahl sich in Anbetracht erwartbarer Verzerrungen durch die uneinheitliche Datenlage, auf die Modellierung mehrerer Stadienwechsel zu verzichten und sich auf die wichtigsten Ereignisse zu beschränken. Vom Ereignis „deutliche Remission“ konnte angenommen werden, dass es mit hoher Wahrscheinlichkeit „irgendwann“ im Therapieverlauf regristiert werden würde. Die Wahrscheinlichkeit, dass bei Patienten, deren Erreichen einer zunächst partiellen ZR bemerkt worden war, auch der Eintritt einer späteren kompletten ZR bemerkt wurde, läßt sich schwerer abschätzen. Eine Unterscheidung zwischen partieller und kompletter ZR erschien allerdings unverzichtbar.

Zu Therapiebeginn wurde bei allen Patienten vom Stadium „keine ZR“ ausgegangen. Zur Analysestichprobe wurden nur Patienten zugelassen, wenn für sie wenigstens eine auf 20 Metaphasen gestützte Zytogenetik im Therapieverlauf vorlag. Soweit die Ergebnisse der univariaten Analyse der zytogenetischen Remission keine Modifikation anraten ließen, sollten im prognostischen Modell nur die drei Stadien „keine deutliche ZR“, „partielle ZR“ und „komplette ZR“ berücksichtigt werden. Zur Modellierung der drei Stadien waren zwei dichotome Variablen, eine für die partielle und eine für die komplette ZR vorgesehen. Sobald die erste deutliche ZR beobachtet wurde, war im Modell entweder die Variable zur partiellen ZR oder die zur kompletten ZR von 0 auf 1 zu

[20] Vgl. Abschnitt 1.2.2.
[21] Mehr dazu in Kapitel 3.

setzen. Folgte auf eine partielle ZR später eine komplette ZR, so wurde zur entsprechenden Zeit der Faktor zur partiellen ZR zurück auf 0 und der zur kompletten ZR auf 1 gesetzt. Rezidive in ein ungünstigeres Stadium wurden nicht berücksichtigt.
Wegen zu unvollständig erhobener Daten wäre es im Hinblick auf die Fallzahl für die Entwicklung eines prognostischen Modells unbefriedigend gewesen, zur Beurteilung von Remissionsstadien zu einem bestimmten Protokollzeitpunkt, z.B. 12 Monate nach Therapiebeginn, nur mit den Daten derjenigen Patienten zu arbeiten, für die zu diesem Zeitpunkt ein aktuelles Evaluationsergebnis vorlag. Zum betrachteten Zeitpunkt wurde daher immer das bisher günstigste beobachtete Remissionsstadium angenommen.

Neben einer ausreichend hohen Fallzahl und einer Verminderung der Problematik unterschiedlicher Datenerhebungsintensitäten zwischen verschiedenen Studien, boten die Beschränkung auf zwei Ereignisse sowie die Nichtberücksichtigung von Rückfällen in ein ungünstigeres Remissionsstadium weitere Vorzüge: Wenn auch nicht ausdrücklich modelliert, so sind in den Risikogruppen eines neuen Prognosesystems die Rezidive und ihre möglichen Auswirkungen auf die Überlebenszeit indirekt doch enthalten. Erreicht ein Patient durch die Erzielung eines besseren zytogenetischen Ergebnisses eine günstigere Prognosegruppe, liegen auch für seine neue Prognosegruppe Überlebenwahrscheinlichkeiten vor, die u.a von Rezidiven beeinflusst wurden. Würde dagegen „das Erreichen eines Rezidives“ modelliert, würden im resultierenden Prognosesystem Patienten mit Rezidiv sehr wahrscheinlich wieder in eine ungünstigere Risikogruppe zurückfallen. Dann (komplett?) ohne Rezidivpatienten existierende Risikogruppen würden vielleicht zwar noch höhere Überlebenswahrscheinlichkeiten aufweisen, doch müssten die aktuell zugehörigen Patienten mit einer nur temporären Zugehörigkeit rechnen. Demgegenüber besaß die hier gewählte prognostische Konstanz, die nur durch das Erreichen einer günstigeren Prognosegruppe durchbrochen werden konnte, einen psychologischen Vorteil. Andere Pluspunkte waren die leichtere Anwendbarkeit und die bessere Interpretierbarkeit (der Überlebenswahrscheinlichkeiten) eines identifizierten Prognosesystems.

Die Zeit bis zur ersten partiellen / kompletten ZR

Nach der Klärung, welche Ereignisse und Remissionsstadien zu modellieren waren, erhob sich die Frage, wie die Zeit bis zur Beobachtung einer partiellen oder kompletten ZR einbezogen werden sollte.
Eine Möglichkeit wäre gewesen, eine Landmark von z.B. 12 Monaten zu setzen. In einem prognostischen Modell hätte man dann alle mindestens bis dahin beobachteten Patienten und deren besten bis zur Landmark verzeichneten Remissionsstatus berücksichtigen können. Formell wäre das prognostische Modell ohne zeitabhängige Kovariablen zu modellieren. Bei den ab 12 Monaten nach Therapiebeginn gemessenen Überlebenszeiten würden die Einflüsse der Baselinevariablen und von Remissionsfaktoren mit im weiteren Therapieverlauf nun ebenfalls unveränderlichen Werten untersucht. Im Grunde aber bliebe das Modell zeitabhängig: Es wäre ein Modell, welches immer im Zusammenhang mit dem Zeitpunkt „12 Monate nach Therapiebeginn“ zu sehen wäre. Sein Einsatz bei späteren Zeitpunkten mit ihren höheren Anteilen an deutlichen Remissionen wäre mit der Verwendung unangemessener Effektschätzer verbunden.[22] Das „Landmark-Modell“ würde daher nur in einer Situation als idealer Ansatz betrachtet: wenn es bei progressionsfrei gebliebenen Patienten einen von der wissenschaftlichen Allgemeinheit einhellig akzeptierten Entscheidungszeitpunkt für eine vom Remissionserfolg abhängig gemachte Weiterbehandlung mit

[22] Speziell bei einem signifikanten Einfluss partieller oder kompletter Remissionen würden die zugehörigen Effekte eher unter- und die Effekte der Baselinevariablen eher überschätzt.

IFN-α gäbe. Doch auch für Imatinib bestand noch keine Einigkeit über einen solchen Zeitpunkt. Favorisiert wurde daher ein von einer vorher festgelegten Landmark unabhängiges prognostisches Modell, welches alle partiellen und kompletten Remissionen einbezog. Ein solches Modell vermied auch Informationsverlust, weil keine kürzer als bis zu einer Landmark beobachteten Patienten ausgeschlossen werden mussten. Stattdessen konnten die Überlebenszeiten und Ereignisse vor einer Landmark ins prognostische Modell mit eingehen. Die prognostische Nutzung der Ergebnisse zur zytogenetischen Remission war die entscheidende Motivation für diese Arbeit. Bei der Modellentwicklung alle beobachteten deutlichen Remissionen ohne Landmarkbeschränkung einzubeziehen reduzierte die Gefahr, den Einfluss der Variablen „zytogenetische Remission“ falsch einzuschätzen.[23] Um für eine „Time-to-event“-Variable (Überlebenszeit, Ereignis: Tod) prognostische Faktoren unter zeitunabhängigen Baselinevariablen und einer „Time-to-event“-Variablen (Zeit bis zum Erreichen einer günstigeren Remissionsstufe, Ereignis: das Erreichen) zu identifizieren, eignet sich die Verwendung des Cox-Modells [27, 28].

2.7.2 Das Cox-Modell mit zeitabhängigen Kovariablen

Das 1972 von D.R. Cox [28] vorgeschlagene Regressionsmodell war ein Meilenstein für die Analyse des Einflusses von Kovariablen auf die Überlebenszeiten. Wegen seiner Annahme proportionaler Hazardfunktionen (PH-Annahme) ist im englischsprachigen Raum auch der Name „proportional hazards model“ geläufig. Software zur Modellberechnung ist allgemein zugänglich (z.B. SAS [96]).

Eine Hazardfunktion $h_i(t)$ repräsentiert das augenblickliche Risiko eines noch nicht eingetretenen Ereignisses / Zustandswechsels (z.B. Tod) für den Patienten i zum Zeitpunkt t. Analog zur Überlebenszeit in Abschnitt 2.3, entspricht t der ab einem bestimmten Zeitpunkt $t = 0$ gemessenen Zeit. Im Cox-Modell ohne zeitabhängige Kovariablen ist die Hazardfunktion für den i-ten Patienten, $i = 1, 2, \ldots, n$ gegeben durch

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ij}\right\} h_0(t), \tag{2.1}$$

wobei x_{ij} beim Patienten i die beobachtete Merkmalsausprägung der unabhängigen Kovariablen $X_j, j = 1, 2, \ldots, p$ darstellt, β_j den zu X_j gehörenden Regressionskoeffizienten und $h_0(t)$ die allen n Patienten gemeinsame sog. Baselinehazardfunktion zum Zeitpunkt t. Bei (2.1) wird davon ausgegangen, dass sämtliche Werte der p Kovariablen zu Beginn des Beobachtungszeitraumes erhoben werden.

Die zugrundeliegende PH-Annahme wird deutlich, wenn man die Hazardfunktionen zweier Patienten A und B zueinander ins Verhältnis setzt:

$$\frac{h_A(t)}{h_B(t)} = \exp\left\{\sum_{j=1}^{p} \beta_j (x_{Aj} - x_{Bj})\right\}. \tag{2.2}$$

Die zeitabhängige Baselinehazardfunktion $h_0(t)$ hat sich herausgekürzt; für das proportionale

[23] Erwähnenswert ist in diesem Zusammenhang das Resultat der italienischen Studiengruppe [58], die alle 23 kompletten ZR bei 218 untersuchten Patienten erstmals nach über einem Therapiejahr entdeckte, davon 5 nach 3 Jahren und 8 nach 5 oder mehr Jahren. Leider konnte die Studie wegen des Fehlens der Variablen „Anzahl aller ausgewerteten Metaphasen“ nicht für die Analysestichprobe berücksichtigt werden.

Verhältnis beider Hazardfunktionen steht eine zeitunabhängige Konstante, deren Größe durch die β_j und die Differenz der bei A und B erhobenen Kovariablenwerte gegeben ist.

Beim Cox-Modell mit zeitabhängigen Variablen hat die Hazardfunktion eines Patienten i zum Zeitpunkt t die Form

$$h_i(t) = \exp\left\{\sum_{j=1}^{p} \beta_j x_{ij}(t)\right\} h_0(t). \tag{2.3}$$

Im Unterschied zum Cox-Modell mit Beschränkung auf unveränderliche Werte x_{ij}, wird in (2.3) mit $x_{ij}(t)$ die Veränderlichkeit der Variablenwerte über die Zeit berücksichtigt. Damit ist die Relation $h_i(t)/h_0(t)$ nicht mehr zeitunabhängig und eine gleichbleibende Proportionalität der Hazardfunktion $h_i(t)$ weder zur Baselinehazardfunktion noch zur Hazardfunktion eines anderen Patienten garantiert. Während die Annahme gleichbleibend proportionaler Hazardfunktionen entfällt, wird auch beim Cox-Modell mit zeitabhängigen Variablen von zeitunabhängigen, konstanten Modellparametern β_j ausgegangen.

Seien die maximalen Beobachtungszeiten der n Patienten mit t_i, $i = 1, \ldots, n$ bezeichnet. Mit $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ ist die zur Berechnung der Schätzer für die β_j zu maximierende partielle Likelihood-Funktion gegeben durch

$$L(\beta) = \prod_{i=1}^{n} \left[\frac{\exp\left\{\sum_{j=1}^{p} \beta_j x_{ij}(t_i)\right\}}{\sum_{l \in R(t_i)} \exp\left\{\sum_{j=1}^{p} \beta_j x_{lj}(t_i)\right\}} \right]^{\delta_i}. \tag{2.4}$$

$R(t_i)$ enthält die Patienten, welche einschließlich Patient i zum Zeitpunkt t_i noch dem Risiko eines Ereignisses unterlagen und δ_i ist eine Indikatorvariable, die den Wert 1 annimmt, falls bei Patient i zum Zeitpunkt t_i ein Ereignis eintrat und 0, falls die Beobachtungszeit zu diesem Zeitpunkt zensiert wurde. Bei den Kovariablen aller verbliebenen Patienten werden die zum Zeitpunkt t_i beobachteten Merkmalsausprägungen berücksichtigt. Im Falle der beiden Remissionsfaktoren nahmen die $x_{ij}(t_i)$ dann den Wert 1 an, wenn die Zeit bis zur ersten partiellen bzw. kompletten Remission kleiner gleich t_i war. Beim Auftreten mehrerer Ereignisse (Todesfälle) zum selben Zeitpunkt t_i wurde für (2.4) die Approximation von Breslow [22] gewählt. Die Modellparameter β_j wurden mit Hilfe des Programmpaketes SAS [96] und der dabei verwendeten Newton-Raphson-Methode geschätzt.

2.8 Univariate Analysen in der Lernstichprobe

Nachdem ein der Datensituation gerechtes, prinzipielles Modell zur Analyse von Prognosefaktoren gewählt wurde, folgte zunächst eine univariate Analyse des Zusammenhangs zwischen interessierenden Kovariablen und dem Hauptzielparameter.

Um einen Einblick in die Datenstruktur zu gewinnen, wurden alle Daten zunächst deskriptiv betrachtet. Bei kategorialen Variablen waren Häufigkeitsverteilung, Minimum, Median, Maximum und die Anzahl fehlender Werte von Interesse. Bei metrischen Größen wurden zusätzlich Mittelwert und Standardabweichung bzw. Konfidenzintervall angegeben. Für den Hauptzielparameter Überlebenszeit und die zytogenetische Remission als „Time-to-event“-Variablen erfolgte eine besonders ausführliche Beschreibung.

Mit der univariaten Analyse schätzt man den direkten Zusammenhang zwischen dem einzelnen Parameter und der Überlebenszeit, ohne die adjustierende Einflussnahme einer anderen Kovariablen auf die Teststatistik zu berücksichtigen. Die univariate Analyse erlaubt für jede Kovariable den Einschluss aller zu ihr vorliegenden Daten und spielt daher speziell für ungenügend erhobene Kovariablen eine wichtige Rolle: Statistisch signifikante Ergebnisse konnten einen Hinweis darauf geben, welchen unzureichend erfassten klinischen Parametern bei der Datenerhebung in Zukunft mehr Beachtung geschenkt werden sollte.

2.8.1 Die zeitunabhängigen Kovariablen

Der Zusammenhang einer kategorialen Kovariablen mit der Überlebenszeit wurde mit Hilfe von Kaplan-Meier-Kurven und Logrank-Test untersucht. Zur Beurteilung des Einflusses einer metrischen Kovariablen auf die Überlebenszeit empfahlen sich Cox-Modell und Wald-Test (2.7). Die Überprüfung der PH-Annahme als Anwendungsvoraussetzung erfolgte gemäß Abschnitt 2.11. Mit Rücksicht auf bessere Interpretierbarkeit und leichtere Anwendbarkeit eines späteren Prognosesystems wurde auf die Transformation von Variablenwerten verzichtet. Allerdings wurden bei den metrischen Variablen, die für die multiple Analyse in Frage kamen, geeignete Skalierungalternativen durch Kategorisierung überprüft. Dabei fanden sowohl früher vorgeschlagene Gruppengrenzen [42] wie über den „Minimum p-value approach“ mit korrigierten p-Werten [6][24] neu identifizierte Unterteilungen Beachtung. Auch wenn bei Verwendung des „Minimum p-value approach“ [6] der für das multiple Testen empfohlenen Adjustierung des p-Wertes Rechnung getragen wurde, galt zu bedenken, dass die Kategorisierung einer metrischen Variablen Informationsverlust bedeutet und allgemein nicht empfohlen wird (vgl. Abschnitt 2.1.1, 1 (e) [121]). Um Artefakte zu vermeiden, wurde eine Kategorisierung nur dann in Betracht gezogen, wenn schon für die metrische Originalvariable ein statistisch signifikanter Zusammenhang ($\alpha = 0,05$) mit der Überlebenszeit beobachtet werden konnte. Außerdem wurde eine durch Kategorisierung zusätzlich gewonnene Skalierung alternativ nur dann als „geeignet“ zugelassen, wenn a) Patientengruppen mit benachbarten Merkmalsausprägungen ähnliche Überlebenswahrscheinlichkeiten besaßen, b) ab einer bestimmten Merkmalsausprägung jedoch eine merkliche Veränderung der Überlebenswahrscheinlichkeiten auftrat und c) jenseits dieser Grenze Patientengruppen mit benachbarten Merkmalsausprägungen wiederum vergleichbare Überlebenswahrscheinlichkeiten aufwiesen. Anstatt bei solchen sprunghaften, nichtlinearen Veränderungen der Überlebenswahrscheinlichkeiten den prognostischen Einfluss einer Kovariablen durch ein Cox-Modell mit demselben (linearen) Anstieg des relativen Risikos über alle Merkmalsausprägungen des metrischen Parameters zu schätzen, erschien es sinnvoller, die sprunghaften Veränderungen des relativen Risikos zwischen Kategorien mit jeweils homogenen Überlebenswahrscheinlichkeiten zu modellieren und dabei über Dummykodierungen unterschiedliche Sprunghöhen zu berücksichtigen.

2.8.2 Die zeitabhängige Kovariable

Unter Berücksichtigung eines Gruppenwechsels im Therapieverlauf wurden mit Hilfe des Mantel-Byar-Tests [8, 75] die Überlebenswahrscheinlichkeiten verschiedener, aus den fünf Remissionsstadien[25] gebildeter Kategorien miteinander verglichen. Dabei gehörten üblicherweise alle Patienten zum Baselinezeitpunkt zur ungünstigsten Kategorie, z.B. „keine deutliche ZR“. Mit dem Zeitpunkt der Beobachtung des Ereignisses „deutliche ZR“ erfolgte der Wechsel in die entspre-

[24] Das Prinzip dieses Verfahrens wird in Abschnitt 2.9.1 erläutert.
[25] Vgl. Abschnitt 1.2.2.

chende Kategorie, in welcher der Patient hinsichtlich der Überlebenszeit dann ab sofort unter Risiko stand.

Graphisch wurden die Überlebenswahrscheinlichkeiten zweier Gruppen anhand von Simon-Makuch-Kurven beschrieben [104]. Um bei der graphischen Darstellung einer Verzerrung zu ungunsten der ursprünglichen Kategorie („time-till-response bias“ [104]) vorzubeugen, gingen in die Berechnung der Simon-Makuch-Kurven nur solche Patienten ein, die eine bestimmte Zeit[26] bis zu dem Ereignis, welches den Wechsel in die neue Kategorie bedingte, überlebt hatten.

Nach Identifikation von Kategorien mit medizinsch relevant unterschiedlichen Überlebenswahrscheinlichkeiten, sollte der Zeitpunkt nach Therapiebeginn gefunden werden, zu welchem nach Maßgabe der vorliegenden Daten eine Entscheidung für oder gegen eine Fortsetzung der applizierten Therapie am sinnvollsten erscheint. Dazu wurden die in Abschnitt 2.2 angeführten fünf Verlaufszeitpunkte miteinander verglichen. Die Bestimmung des besten Entscheidungszeitpunktes gemäß der gegebenen Daten wurde einerseits am Ergebnis der Landmark-Analyse der Überlebenszeit in Abhängigkeit des bis zur Landmark verbuchten Remissionsstatus' festgemacht, andererseits an dem Verhältnis Informationsgewinn versus dem Risiko für die Patienten, mit der Entscheidung bis einem späteren Zeitpunkt zu warten.

Zur zusätzlichen Überprüfung eines Zusammenhangs zwischen Remissionsstatus und Überlebenszeit, ohne sich zur Beurteilung des Remissionsergebnisses vorher auf einen Zeitpunkt nach Therapiebeginn festzulegen, wurde das Cox-Modell mit zeitabhängigen Kovariablen ((2.3), (2.4)) verwendet.

2.9 Zusammenhänge zwischen den Kovariablen

Die Untersuchung von Korrelationen zwischen den Kovariablen konnte Zusammenhänge aufdecken, die später als Wechselwirkungsterme in einem multiplen Cox-Modell analysierbar waren. Auch war mit einer starken Korrelation u.U. erklärbar, warum in sich in einem Modell jeweils nur eine von zwei Variablen als statistisch signifikant erwiesen hatte. Bei vergleichbarer Prognosequalität empfahl sich, die klinisch sinnvollere bzw. verlässlicher erhebbare Variable im Modell zu belassen.

2.9.1 Korrelationen zwischen zeitunabhängigen Variablen

Um die Stärke der Korrelation zwischen zwei zeitunabhängigen Variablen feststellen zu können, wurden - je nach Variablenskalierung und Berechtigung der Normalverteilungsannahme - Punktewolken betrachtet, Pearsons oder Spearmans Korrelationskoeffizient berechnet, t-Test, U-Test, Kruskal-Wallis oder χ^2-Test durchgeführt.[27]

Die Anwendung von Klassifikationsbäumen (CART: classification and regression trees) [10, 21, 29, 71, 72] bot einen zusätzlichen Einblick in die Datenstruktur und konnte wichtige Hinweise auf Zusammenhänge zwischen Subgruppen von Patienten liefern. Die explorative Datenanalyse mit CART unterliegt keinen einschränkenden Verteilungsannahmen wie andere Regressionsmodelle [72]. Aus gegebenen Einflussgrößen wird zur Konstruktion eines Klassifikationsbaumes zunächst

[26] Meistens wählt man die mediane Zeit.

[27] Basis für einfache deskriptive und induktive Statistiken waren die Bücher von Rüger [94] und Sachs [95].

die Variable ausgewählt, welche einen sog. „Cutpoint“ [10] besitzt, der einen n Fälle umfassenden Datensatz bzgl. der Zielgröße in zwei sich maximal unterscheidende Gruppen unterteilt. Im Falle der Zielgröße Überlebenszeit werden alle möglichen Aufteilungen mit der χ^2-Statistik des Logrank-Tests beurteilt. Nach einem statistisch signifikanten „Split“ [10] an dem Cutpoint mit dem größten χ^2-Wert (und damit kleinstem p-Wert) wird diese Partitionierungsprozedur auf die entstehenden Untergruppen solange rekursiv angewandt, bis eins von zwei Abbruchkriterien erreicht ist: entweder wird der p-Wert größer als ein vorgegebenes Signifikanzniveau oder eine Gruppe kleiner als $\sqrt{n}$ Fälle. Als Signifikanzniveau wurde $\alpha = 0,1$ gewählt. Um eine Überschätzung des Einflusses maximal selektierter Cutpoints von Variablen mit vielen Merkmalsausprägungen zu vermeiden [6, 29, 72, 73, 77], wurden die p-Werte zur Teststatistik korrigiert [73, 77]. Die durch das SAS Macro von Brandmeier et al. [20] nach Lausen und Schumacher [73] in Weiterführung des Vorschlages von Miller und Siegmund [77] adjustierten p-Werte werden im folgenden als p_{ad} bezeichnet. Mit Hilfe von CART identifizierte Zusammenhänge sollten als Wechselwirkungsterme im Cox-Modell untersucht werden.
Der zuvor erwähnte „Minimum p-value approach“ [6] verläuft nach dem gleichen Prinzip wie CART, mit dem Unterschied, dass nur bei **einer** Variablen nach statistisch signifikanten Splits mit kleinsten p_{ad}-Werten gesucht wird.

2.9.2 Zusammenhang zwischen zeitunabhängigen Variablen und zytogenetischer Remission

In Übereinstimmung mit Abschnitt 2.7.1 wurde die abhängige Variable „zytogenetische Remission“ als „Time-to-event“-Variable modelliert. Als Ereignis sollten alle drei voraussichtlich wesentlichen Situationen untersucht werden: das Erreichen einer partiellen ZR, einer kompletten ZR und - beide zusammengenommen - einer deutlichen ZR. Mit der erwarteten Bedeutung der zytogenetischen Remission als sehr wichtigem prognostischen Faktor und der relativ geringen Wahrscheinlichkeit, in den ersten Monaten zu versterben, empfahl es sich, die Prognose der Überlebenswahrscheinlichkeiten auf das tatsächlich beobachtete Remissionsergebnis zu stützen anstatt auf prognostizierte Resultate. Im Hinblick auf das für die Überlebenszeit zu suchende Prognosemodell wurde daher zwar der multiple Einfluss der Baselinevariablen auf die zytogenetische Remission mit dem Cox-Modell untersucht, jedoch weder eine Risikogruppenbildung vorgenommen noch ein Prognosesystem vorgeschlagen. Entsprechend dem Vorgehen bei der Überlebenszeit wurden zuvor die Einflüsse der Baselinevariablen auf die zytogenetische Remission mittels univariater Analyse und CART untersucht.

2.10 Selektion des besten prognostischen Modells in der Lernstichprobe

Um das beste prognostische Modell bestimmen zu können, mussten vorab Kriterien für die Auswahl seiner prognostischen Faktoren einschließlich möglicherweise bedeutsamer Wechselwirkungen festgelegt werden. Dazu wurde ein von Collett [27] vorgeschlagenes Selektionsverfahren adaptiert. Kriterium beim Vergleich zweier Modelle war der Likelihood-Quotienten-Test [54, 94]

$$Q = -2\ ln\left[\frac{L(\hat{\beta}) \text{ im Modell mit } p-q \text{ Kovariablen}}{L(\hat{\beta}) \text{ im Modell mit } p \text{ Kovariablen}}\right], \tag{2.5}$$

mit $1 \leq q \leq p$. Die Freiheitsgrade der q getesteten Kovariablen bestimmen die χ^2-Verteilung, mit welcher die Teststatistik Q hinsichtlich ihres p-Wertes verglichen wird. Als Voraussetzung, die

$-2\ln L(\hat{\beta})$-Statistik verschiedener prognostischer Modelle miteinander vergleichen zu können, sind die Modellberechnungen auf Basis derselben n Patienten und desselben Datenstandes durchzuführen. Entsprechend müssen zu allen n Patienten für alle in Frage kommenden Kovariablen vollständige Daten vorliegen.

Das Selektionsverfahren

1. Schritt: Zunächst wird für jeden klinischen Parameter univariat untersucht, ob seine Aufnahme ins Cox-Modell die $-2\ln L(\hat{\beta})$-Statistik im Vergleich zum Modell ohne Kovariablen statistisch signifikant reduziert, d.h., äquivalent, zu einem statistisch signifikanten Ergebnis des Likelihood-Quotienten-Tests (2.5) führt.[28] $L(\hat{\beta})$ berechnet man durch Einsetzen des geschätzten $\hat{\beta}$ in (2.4).[29] Das Signifikanzniveau für den 1. Schritt wird auf $0,1$ erhöht, um die Wahrscheinlichkeit zu vergrößern, dass ein evtl. im Zusammenwirken mit anderen Kovariablen signifikanter Parameter nicht von der multiplen Modellanalyse ausgeschlossen bleibt.

2. Schritt: Die univariat signifikanten und gemäß Abschnitt 2.6 zugelassenen Kovariablen werden dann in ein gemeinsames Modell gesteckt. Das Weglassen einiger Kovariablen wird evtl. zu keiner signifikanten Erhöhung der $-2\ln L(\hat{\beta})$-Statistik führen. Die Kovariable, deren nicht-signifikanter Beitrag den geringsten Wert besitzt, wird als erstes aus dem Modell entfernt. Durch wechselseitiges Entfernen je einer der übrigen Kovariablen wird die erneute Erhöhung von $-2\ln L(\hat{\beta})$ untersucht und ggf. eine weitere Kovariable aus dem Modell ausgeschlossen. Dieser Vorgang wird so lange wiederholt, bis das Weglassen jeder der verbliebenen Kovariablen eine signifikante Erhöhung der Statistik zur Folge hätte.

3. Schritt: Kovariablen, die nicht univariat signifikant waren und daher im 2. Schritt keine Berücksichtigung finden, könnten sich in Gegenwart anderer als statistisch signifikant erweisen. Nach und nach wird jeweils eine dieser Kovariablen dem nach Schritt 2 gefundenen Modell hinzugefügt. Die zu einer signifikanten Reduktion von $-2\ln L(\hat{\beta})$ führenden Kovariablen werden beibehalten. Dasselbe Vorgehen wird anschließend auf die Wechselwirkungsterme angewandt. Dabei ist zu beachten, dass die beteiligten Haupteffekte gemäß des hierarchischen Prinzips ebenfalls Bestandteil des untersuchten Modells sein sollten.

4. Schritt: Für das nach Schritt 3 identifizierte Modell wird abschließend überprüft, ob einer der Terme entfernt werden kann, ohne $-2\ln L(\hat{\beta})$ signifikant zu erhöhen bzw. ob die Hinzunahme eines ausgeschlossenen Terms $-2\ln L(\hat{\beta})$ noch signifikant reduziert. Ist beides nicht der Fall, hat man das nach dem Selektionsverfahren beste Modell identifiziert.

Durch zusätzliche Anwendung der in SAS [96] beschriebenen „Stepwise Selection Procedure“ wurde überprüft, ob die Wahl eines alternativen Selektionsverfahrens zum selben „besten“ Modell führt. Diskrepanzen zwischen beiden Ergebnissen konnten ggf. ein klinisch sinnvolleres Alternativmodell ins Licht rücken - neben statistischen Signifikanzen dürfen die klinischen Aspekte bei der Wahl des endgültigen Modells nicht vergessen werden.

„Bootstrap-resampling“ und „Shrinkage“

[28] Wenn man die $-2\ln L(\hat{\beta})$-Statistik des kleineren Modells von der des größeren Modells abzieht, erhält man denselben χ^2-verteilten Wert wie beim Q des Likelihood-Quotienten-Tests (vgl. z.B. Collett [27]).

[29] Im univariaten Fall ist $p = 1$, beim Modell ohne Kovariablen setzt man die $\beta_j = 0$.

Die Stabilität des erhaltenen Cox-Modells könnte nun mit Hilfe eines „Bootstrap-resampling"-Verfahrens überprüft werden [4]. Dabei werden aus der gegebenen Lernstichprobe wiederholt neue Stichproben gebildet und zu jeder dieser Stichproben wird das beste prognostische Modell ermittelt. Die Ergebnisse erlauben eine Beurteilung der Konsistenz hinsichtlich der in das jeweilige beste Modell eingeschlossenen Kovariablen und der zugehörigen Koeffizientenschätzer.
Altman und Andersen [4] verglichen mit „Bootstrap-resampling" geschätzte Cox-Modelle mit dem ohne das Verfahren identifizierten „Originalmodell". Obwohl Altman und Andersen [4] nur 216 Patienten untersuchten und immerhin 17 Kovariablen zur Auswahl standen, ergaben sich durch die Ergebnisse des „Bootstrap-resampling"-Verfahrens allerdings keine Zweifel an der Validität des Originalmodells. Mit diesem Resultat vor Augen, dem Rechenaufwand für das „Bootstrap-resampling" aus z.B. 100 verschieden zusammengesetzten Stichproben immer wieder neu das beste Modell zu bestimmen und wegen der umfangreichen Stichprobe in vorliegender Arbeit war die Anwendung eines „Bootstrap-resampling"-Verfahrens primär nicht vorgesehen. Auf „Bootstrap-resampling" sollte nur dann nachträglich zurückgegriffen werden, wenn sich nach Weglassung der nach Barlow und Prentice [14] (vgl. Abschnitt 2.12) identifizierten Patienten mit extremen Residuen eine andere Kovariablenzusammensetzung des Endmodells ergeben würde oder wenn sich die Koeffizientenschätzer in einem Verhältnis zueinander ändern würden, welches möglicherweise erheblichen Einfluss auf die Definition der späteren Risikogruppen hätte. Da sich die Stichproben beim „Bootstrap-resampling" alle aus den Patienten derselben Lernstichprobe zusammensetzen [4, 85], wurde es zur Beurteilung der Validität eines Prognosesystems vorgezogen, die in einer unabhängigen Patientenstichprobe beobachtete Diskriminierungsfähigkeit seiner Risikogruppen zu betrachten [85].

Auch das von Verweij und van Houwelingen [118] vorgeschlagenen Verfahren, jedes $\hat{\beta}_j$ mit einem „Shrinkage factor" zu multiplizieren, wird zur Korrektur der geschätzten Koeffizienten in einer Lernstichprobe benutzt. Ob und inwieweit die um den „Shrinkage factor" korrigierten Koeffizienten zu einer verbesserten Prognose in einer unabhängigen Validierungsstichprobe führen ist vorab nicht bekannt. Neben dem nicht unerheblichen Rechenaufwand, jeden Koeffizienten entsprechend dem Umfang der Lernstichprobe n-mal zu schätzen, sollte hier - entsprechend der Argumentation bzgl. „Bootstrap-resampling" - eine Berücksichtigung von „Shrinkage" nur im Verdachtsfalle eines instabilen Endmodells überlegt werden.

Das endgültige Modell

Hat man sich auf das beste Modell und seine Variablen festgelegt, so schließt man zur Berechnung der Parameterschätzer $\hat{\beta}_j$ des endgültigen Modells solche Patienten, die nur fehlende Daten zu nichtselektierten Kovariablen hatten, nicht mehr aus. Die dadurch erhöhte Patientenzahl führt zu einer Verringerung der Standardabweichung der $\hat{\beta}_j$ und somit zu einer genaueren Schätzung der „wahren" β_j.

Eine Schätzung für die Standardabweichungen zu den $\hat{\beta}_j$ erhält man über die beobachtete $p \times p$ Informationsmatrix $I(\beta)$, die Matrix der negativen zweiten Ableitungen der logarithmierten Likelihood-Funktion (2.4). Das (j,k)-te Element von $I(\beta)$ ist gegeben durch

$$-\frac{\partial^2 \ln L(\beta)}{\partial \beta_j \partial \beta_k}.$$

Die geschätzte Varianz-Kovarianz-Matrix $\hat{V}(\hat{\beta})$ wird durch das Einsetzen der $\hat{\beta}_j$ in die Inverse der Informationsmatrix berechnet [27]:

$$\hat{V}(\hat{\beta}) = I^{-1}(\hat{\beta}). \tag{2.6}$$

Die Wurzel aus dem j-ten Diagonalelement von (2.6), $\sqrt{\hat{v}_{jj}}$, ergibt die geschätzte Standardabweichung zu $\hat{\beta}_j$.

Vor dem ausschließlichen Konzentrieren auf das identifizierte beste Modell, wurden nun noch diejenigen Variablen betrachtet, die aufgrund ihrer zu geringen Erhebungsrate (hier: $< 90\%$) nicht für das multiple Modell in Frage gekommen waren. Jeweils eine dieser Variablen wurde den Kovariablen des extrahierten endgültigen Modells hinzugefügt. Ziel war es, Variablen zu identifizieren, die sich auch in Gegenwart der gefundenen prognostischen Faktoren als statistisch signifikant erweisen und in Zukunft - ein höherer Erhebungsgrad vorausgesetzt - Kandidaten für ein Alternativmodell sein könnten. Für solche bisher möglicherweise zu Unrecht als nicht relevant betrachteten klinischen Parameter konnte alsbald eine konsequentere Datenerhebung angeregt werden.

Um in einem Modell mit p-dimensionalem β die Bedeutung ein oder mehrerer Kovariablen gemeinsam zu überprüfen, testet man die globale Nullhypothese $H_0 : \beta^q = 0$. Dabei sei β^q, mit $1 \leq q \leq p$, der $(1 \times q)$-Parametervektor, der aus den β_j der Kovariablen, die gemeinsam getestet werden sollen, zusammengesetzt wird. Ein geeigneter Test für H_0 ist gegeben durch

$$X^2_W = \hat{\beta}^q \left[\hat{V}(\hat{\beta}^q)\right]^{-1} \hat{\beta}^{q\prime}. \tag{2.7}$$

Er wird zumeist als Wald-Test bezeichnet [27, 96] und besitzt unter H_0 approximativ eine χ^2_q-Verteilung. Der $(1 \times q)$-Vektor $\hat{\beta}^q$ wird nicht neu berechnet, sondern setzt sich nach gemeinsamer Schätzung aus denjenigen $\hat{\beta}_j$ des $(1 \times p)$-Vektors $\hat{\beta}$ zusammen, deren β_j in β^q zur Nullhypothese gehören. Ebenso wird die $(q \times q)$-Matrix $\hat{V}(\hat{\beta}^q)$ aus den Teilen der nach (2.6) berechneten Varianz-Kovarianz-Matrix $\hat{V}(\hat{\beta})$ gebildet, welche Varianzen oder Kovarianzen der q zu testenden Parameter enthalten. Die Nullhypothese wird abgelehnt, wenn $P(X^2_W) < \alpha$. Als Signifikanzniveau wurde $\alpha = 0,1$ gewählt. Es gilt zu beachten, dass der p-Wert nicht unabhängig von den $p - q$ nicht getesteten β_j interpretiert werden kann, denn durch die gemeinsame Schätzung aller β_j der Likelihood-Funktion (2.4) erfahren alle Komponenten des $(1 \times p)$-Vektors $\hat{\beta}$ einen adjustierenden Einfluss durch alle p beteiligten Kovariablen des geschätzten Modells.

Während der Likelihood-Quotienten-Test bzw. die äquivalente Differenz zweier $-2 \ln L(\hat{\beta})$-Statistiken unter den oben beschriebenen Voraussetzungen zwei Modelle miteinander vergleicht, werden mit dem Wald-Test die Parameter innerhalb eines Modells beurteilt. Wenn das beste Modell schon feststeht, eignet sich daher der Wald-Test zur Beurteilung der prognostischen Relevanz einer unzureichend erhobenen Variablen bei gemeinsamer Schätzung mit den Variablen des besten Modells.

2.11 Überprüfung der Modellannahmen des statistischen Modells

2.11.1 Überprüfung der PH-Annahme im Cox-Modell mit zeitunabhängigen Variablen

Für das Cox-Modell ohne zeitabhängige Kovariable ist die Annahme proportionaler Hazardfunktionen Anwendungsvorraussetzung. Diese Proportionalität sollte für die Hazardfunktionen zweier Patienten A und B über den gesamten Zeitverlauf durch einen konstanten, zeitunabhängigen Wert ausgedrückt werden können (2.2). Ist die Proportionalitätsannahme nicht gerechtfertigt, sollte eine Transformation der betroffenen Variablen oder gar ein alternatives statistisches Modell erwogen werden.

Um die Zeitunabhängigkeit und damit die konstante Proportionalität einer kategorialen / kategorisierten zeitunabhängigen Kovariablen X zu überprüfen, erweitert man das univariate Modell mit X um den Wechselwirkungsterm $X \times \ln t$, mit t für die Überlebenszeit [27].[30] Erweist sich die Wechselwirkung als signifikant (Wald-Test, $\alpha = 0,05$), so ist die Annahme proportionaler Hazardfunktionen verletzt [27].

Zur graphischen Prüfung der PH-Annahme werden für die K Kategorien einer diskreten Kovariablen die Kurven $\ln(-\ln \hat{S}_k(t))$ versus $\ln t$ betrachtet. Dabei steht $\hat{S}_k(t)$ für die mit der Kaplan-Meier-Methode geschätzte Überlebenswahrscheinlichkeit der Kategorie k, $k = 1, \ldots, K$ zum Zeitpunkt t. Verlaufen die Kurven annähernd parallel, darf man von berechtigter PH-Annahme ausgehen [40].

Ein Verfahren um Abweichungen von der PH-Annahme in einem multiplen Modell zu untersuchen, besteht in der Betrachtung des interessierenden Modells im zeitlichen Verlauf [5, 97]. Mit den Daten jeweils aller n Patienten bildet man Datensätze, die sich nur durch die Zensierung der Überlebenszeit zu verschiedenen, vorher festgelegten Verlaufszeitpunkten unterscheiden. Dadurch stehen die Datensätze zeitlich hierarchisch miteinander in Beziehung. Die Zahl der Ereignisse wächst von Zeitpunkt zu Zeitpunkt und die Modelle nähern sich immer mehr dem eigentlichen Modell mit den vollständigen Beobachtungszeiten und Ereignissen an. Zu jedem der Verlaufszeitpunkte wird der Parametervektor $\hat{\beta}$ geschätzt. Sind für eine Variable Unterschiede bzgl. der geschätzten Koeffizienten $\hat{\beta}_j$ und ihrer statistischen Signifikanz oder ein Trend über den Zeitverlauf erkennbar, deutet dies für die Variable im untersuchten Modell auf eine Abweichung von der PH-Annahme hin.

Andersen et al. [7] schlagen zwei graphische Verfahren vor, um in einem multiplen Modell mit p Variablen die Berechtigung der PH-Annahme für eine diskrete / kategorisierte Variable X_p mit K Kategorien zu beurteilen.[31] Um bei der kategorialen Variablen die Effekte der einzelnen Kategorien im Cox-Modell zu schätzen, wird X_p durch $X_{p+1}, \ldots, X_{p+k}$ ersetzt, wobei die Werte der neuen Vektoren durch

$$x_{i,p+k} = I(x_{ip} \in k),\ k = 1, \ldots, K-1$$

[30] Wegen der zumeist schiefen Verteilung von t empfiehlt sich die Wahl von $\ln t$ [5].

[31] Ohne Beschränkung der Allgemeinheit sei der Laufindex j, $j = 1, \ldots, p$ so definiert, dass die untersuchte Variable den Laufindex $j = p$ besitze.

gebildet werden. I steht für die Indikatorfunktion, die den Wert 1 annimmt, wenn die Merkmalsausprägung x_{ip} des Patienten i zur Variablen X_p der Kategorie k angehört und 0, sonst. Für die Referenzkategorie K gilt $X_{p+K} \equiv 0$. Im Cox-Modell entspricht die Verwendung der X_{p+k}, $k = 1, \ldots, K$ der Umformung der Hazardfunktion (2.1) in

$$h_i(t) = \exp\left\{\sum_{\substack{j=1 \\ j \neq p}}^{p+K-1} \beta_j x_{ij}\right\} h_0(t). \tag{2.8}$$

Um nun die Berechtigung der PH-Annahme für (2.8) graphisch überprüfen zu können, wird für die Hazardfunktion ein Modell eingeführt, bei welchem jede der K Kategorien der zu betrachtenden kategorialen Variable X_p eine eigene Baselinehazardfunktion, $h_{01}, \ldots, h_{0K}$ besitzt:

$$h_i(t) = \exp\left\{\sum_{j=1}^{p-1} \beta_j x_{ij}\right\} h_{0k(i)}(t). \tag{2.9}$$

Im Exponenten sind die übrigen $p-1$ Kovariablen des multiplen Modells enthalten. Andersen et al. [7] bezeichnen (2.9) als nach X_p stratifiziertes Modell. Je nachdem, zu welchem Stratum der Patient i gehört, nimmt $k(i)$ den entsprechenden Index k zwischen 1 und K an.

Beim stratifizierten Modell werden die $p-1$ Koeffizienten aus der partiellen Likelihoodfunktion

$$L(\beta_1, \ldots, \beta_{p-1}) = \prod_{i=1}^{n} \prod_{k=1}^{K} \left[\frac{\exp\left\{\sum_{j=1}^{p-1} \beta_j x_{ij}\right\}}{\sum_{l \in R_k(t_i)} \exp\left\{\sum_{j=1}^{p-1} \beta_j x_{lj}\right\}} \right]^{\delta_{ki}} \tag{2.10}$$

geschätzt. Im Unterschied zur zeitunabhängigen Version von (2.4) besteht $R_k(t_i)$ nur aus Patienten des Stratums k und auch die Indikatorvariable δ_{ki} nimmt bei einem zum Zeitpunkt t_i beobachteten Ereignis nur dann den Wert 1 an, falls Patient i zum Stratum k gehört.

Nach Schätzung der Koeffizienten berechnet man die kumulierte Baselinehazardfunktion für Stratum k,

$$H_{0k}(t) = \int_0^t h_{0k}(s)\, ds,$$

z.B. wie in vorliegender Arbeit mit Hilfe der Approximation von Breslow [7, 22]

$$\hat{H}_{0k}(t) = \sum_{i=1}^{n} \frac{\delta_{ki}\ I(t_i \leq t)}{\sum_{l \in R_k(t_i)} \exp\left\{\sum_{j=1}^{p-1} \hat{\beta}_j x_{lj}\right\}}. \tag{2.11}$$

Die Indikatorfunktion I garantiert, dass nur über Ereigniszeiten bis einschließlich Zeitpunkt t aufsummiert wird.

Die erste graphische Prüfung besteht in einer Betrachtung der Kurven von $\log \hat{H}_{01}(t), \ldots,$ $\log \hat{H}_{0K}(t)$ versus t. Auch wenn die beiden Modelle (2.8) und (2.9) nicht mit Hilfe eines LQ-Tests direkt miteinander vergleichbar sind, sollten die Kurven bei berechtigter Annahme von

(2.8) annähernd parallel verlaufen [7]. Zudem sollte die konstante vertikale Distanz zwischen log $\hat{H}_{0k}(t)$ und log $\hat{H}_{0K}(t)$ approximativ $\hat{\beta}_{p+k}$ betragen, $k = 1, \ldots, K-1$.

Alternativ trägt man die Kurven $\hat{H}_{0k}(t)$, $k = 1, \ldots, K-1$ versus $\hat{H}_{0K}(t)$ auf. Unter Modell (2.8) sollten die $k-1$ Kurven approximativ aus Geraden durch den Nullpunkt bestehen, deren Steigung in etwa dem Wert $\exp(\hat{\beta}_{p+k})$, $k = 1, \ldots, K-1$ entspricht. Ein konvexer (konkaver) Kurvenverlauf $\hat{H}_{0k}(t)$ versus $\hat{H}_{0K}(t)$ weist auf ein mit der Zeit zunehmendes (abnehmendes) Hazardverhältnis $h_{0k}(t)/h_{0K}(t)$ hin.

2.11.2 Überprüfung der Annahme konstanter Koeffizienten im Cox-Modell mit zeitabhängigen Kovariablen

Die PH-Annahme im Cox-Modell mit zeitunabhängigen Kovariablen impliziert die Annahme konstanter, zeitunabhängiger Koeffizienten. Jede Überprüfung der PH-Annahme kommt daher einer Überprüfung der Annahme der zeitliche Konstanz der Koeffizienten gleich. Generell ist im Cox-Modell mit zeitabhängigen Kovariablen die zeitunabhängige Konstanz der PH-Annahme zwischen zwei Patienten A und B nicht mehr erfüllt, denn sobald sich im Zeitverlauf der Wert einer Kovariablen bei einem der beiden Patienten verändert, ändert sich auch das Verhältnis der Hazardfunktionen. Die Annahme zeitlich unabhängiger Koeffizienten bleibt jedoch auch für das Cox-Modell mit zeitabhängigen Kovariablen (2.3) bestehen.

In Gegenwart zeitabhängiger Variablen wurde die Zeitunabhängigkeit der Koeffizienten - wie im voranstehenden Abschnitt - durch Berechnung des interessierenden Modells zu verschiedenen Zensierungszeitpunkten untersucht.[32] Dabei wurden die Veränderungen der Merkmalsausprägungen bei der zeitabhängigen Variablen bis zum jeweiligen Zensierungszeitpunkt berücksichtigt.

Für die einzelnen Kategorien jeder zeitunabhängigen Kovariablen des Modells ließ sich die Annahme der Konstanz der zugehörigen Koeffizienten wieder mit Hilfe der graphischen Methoden auf Basis des stratifizierten Modells nach Andersen et al. [7] überprüfen. Dazu musste nur bei den Merkmalsausprägungen jeder zeitabhängigen Kovariablen X_j mit $x_{ij}(t)$ statt x_{ij} der Zeit t in den Exponenten von (2.9)-(2.11) Rechnung getragen werden.

2.12 Untersuchung der Anpassung des prognostischen Modells an die Daten

Um die Modellanpassung an gegebene Daten zu untersuchen, werden Residuen betrachtet. Im klassischen multiplen linearen Modell errechnen sich die Residuen aus der Differenz zwischen dem beobachteten Wert der abhängigen Variablen und dem Wert den die Variable gemäß der Modellschätzung haben müsste. Im Rahmen der Überlebenszeitanalyse ist eine Herleitung von Residuen wegen der Zensierungen weit weniger offensichtlich und daher existieren auch viele verschiedene Definitionsvorschläge. Eine der umfassendsten Definitionen von Residuen wurde von Barlow und Prentice [14] veröffentlicht. Ihre Residuen sind sowohl für zeitunabhängige wie zeitabhängige multiple Cox-Modelle verwendbar. Die Residuen werden für jeden Patienten und jede Kovariable eines Modells gesondert berechnet. Sie drücken den Unterschied zwischen dem

[32] Vgl. Altman und De Stavola [5] sowie Sasieni [97].

Wert der Kovariablen zum maximalen Beobachtungszeitpunkt des Patienten und einem gewichteten Kovariablenmittel aller im Risikoset verbliebenen Patienten aus, welche durch Subtraktion einer Summe von Differenzen zu allen Ereigniszeitpunkten t_l bis hin zum maximalen Beobachtungszeitpunkt t_i des betrachteten Patienten i korrigiert wird [5, 14]. Die Definition von Barlow und Prentice [14] für das Residuum $\hat{e}_{ij}$ zu einem Patienten i und den für i beobachteten Werten zur Kovariablen X_j lautet wie folgt:

$$\hat{e}_{ij} = \underbrace{\left\{x_{ij}(t_i) - \hat{E}_j(\hat{\beta}, t_i)\right\}\delta_i}_{A} - \underbrace{\sum_{l=1}^{n}\left\{x_{ij}(t_l) - \hat{E}_j(\hat{\beta}, t_l)\right\}\delta_l \hat{p}_i(t_l)}_{B} \tag{2.12}$$

Dabei stehen

- $x_{ij}(t)$ für den Wert der Kovariablen j beim Patienten i zum Beobachtungszeitpunkt t
- $\hat{E}_j(\hat{\beta}, t)$ für ein gewichtetes Mittel der Kovariablen X_j bei den zum Zeitpunkt t noch im Risikoset befindlichen Patienten mit

  $$\hat{E}_j(\hat{\beta}, t) = \sum_{l=1}^{n} Y_l(t) x_{lj}(t) \hat{p}_l(t).$$

  Die Indikatorvariable $Y_l(t)$ hat den Wert 1, wenn Patient l zum Zeitpunkt t noch unter Risiko steht und 0, sonst. Die Gewichte $\hat{p}_l(t)$ berechnen sich aus

  $$\hat{p}_l(t) = \frac{Y_l(t)\, \exp\left\{\sum_{j=1}^{p} \hat{\beta}_j x_{lj}(t)\right\}}{\sum_{m=1}^{n} Y_m(t)\, \exp\left\{\sum_{j=1}^{p} \hat{\beta}_j x_{mj}(t)\right\}}\ .$$

  Damit entspricht für $t = t_l$ das Gewicht $\hat{p}_l(t)$ dem Faktor des Patienten l in der geschätzten Likelihood-Funktion $L(\hat{\beta})$, die man erhält, wenn man in (2.4) $\hat{\beta}$ einsetzt, wobei in $L(\hat{\beta})$ allerdings (wegen der Potenzierung mit δ_l) nur die Patienten zum Produkt beitragen, für welche ihr maximaler Beobachtungszeitpunkt der Todeszeitpunkt war.

Die Differenz von Teil A in (2.12) bezieht sich auf die maximale Beobachtungszeit t_i eines Patienten i. Je mehr zu diesem Zeitpunkt der Kovariablenwert $x_{ij}(t_i)$ vom gewichteten Mittel der Kovariablenwerte aller im Risikoset verbliebenen Patienten abweicht, desto höher ist der absolute Betrag der Differenz. Wegen δ_i, welchem dieselbe Bedeutung wie in (2.4) innewohnt, hat Teil A für alle zensierten Beobachtungszeiten den Wert 0. Dies korrespondiert mit dem gegenüber Ereignissen geringeren Einfluss von Zensierungen auf die Parameterschätzung im Cox-Modell.

In Teil B wird über alle Ereigniszeiten (Ereignis, $\Rightarrow \delta_l = 1$) die gewichtete Differenz zwischen dem Kovariablenwert des Patienten i zum Zeitpunkt t_l und $\hat{E}_j(\hat{\beta}, t_l)$, dem gewichteten Mittel der Kovariablen X_j bei den zum Zeitpunkt t_l noch im Risikoset befindlichen Patienten, aufsummiert. Die Gewichte $\hat{p}_i(t_l)$ geben für $t_l \leq t_i$ den Beitrag zum Produkt der geschätzten Likelihood-Funktion $L(\hat{\beta})$ an, wenn für Patient i zum Zeitpunkt t_l ein Ereignis beobachtet worden wäre. Im Gegensatz zu Teil A, kann Teil B für einen Patienten i, der zu seiner maximalen Beobachtungszeit t_i zensiert wurde, von 0 verschieden sein.

Die in Teil A berechnete Abweichung zwischen $x_{ij}(t_i)$ und $\hat{E}_j(\hat{\beta}, t_i)$ wird durch die Subtraktion von Teil B um ein gewichtetes Mittel an Abweichungen zu Ereigniszeitpunkten $t_l \leq t_i$

korrigiert. Die berechneten Residuen haben damit einen Erwartungswert von 0, sind asymptotisch unkorreliert und werden gegen die Ränge aller Ereignis-/Zensierungszeiten aufgetragen [5, 14]. Je weiter der Wert eines Residuums e_{ij} von 0 und der Masse der Residuen mit kleineren Beträgen entfernt liegt, desto schlechter ist die Modellanpassung an die für Patient i bis t_i beobachteten Kovariablenwerte von X_j.

Um den Einfluss von Patienten mit schlechter Modellanpassung auf die geschätzten Koeffizienten der Kovariablen zu untersuchen, berechnet man unter Weglassung dieser Patienten die Koeffizienten des interessierenden Modells noch einmal. Bei starker Abweichung zu den früheren Werten oder gar dem Rückgang der Signifikanz einer Variablen ist zu überlegen, ob man die Variablenselektion der multiplen Modellanalyse wiederholt, ohne die Patienten mit zuvor extremen e_{ij} zu berücksichtigen. Um ein willkürliches „Anpassen der Daten an das Modell“ zu vermeiden, wird das Ausschließen von solchen Patienten und die Neuberechnung des prognostischen Modells allerdings nur empfohlen, wenn die Variablenwerte der auszuschließenden Patienten auch unter klinischen Gesichtspunkten als ungewöhnliche Ausreißer zu betrachten sind, z.B. weil erkennbare Messfehler vorliegen. Ist ein Ausschluss von Patienten aus klinischer Sicht nicht gerechtfertigt, sollte bei gehäuftem Auftreten extremer Residuen die Variablenskalierung oder die Wahl des statistischen Modells überdacht werden.

2.13 Vom prognostischen Modell zum Prognosesystem

Der geschätzte, additive Teil im Exponenten der Hazardfunktion (2.3),

$$\mathrm{PI}(t) = \hat{\beta}_1 x_{i1}(t) + \hat{\beta}_2 x_{i2}(t) + \ldots + \hat{\beta}_p x_{ip}(t), \tag{2.13}$$

wird als „prognostischer Index“ bezeichnet [24]. In vielen Fällen, wie z.B. beim Sokal-Score [105] und dem New CML-Score [42], ist der Risikowert $\mathrm{RI}(t)$ eine monotone Transformation von $\mathrm{PI}(t)$. Allgemein gilt $\mathrm{RI}(t)=f(\mathrm{PI}(t))$, wobei für die Funktion f auch die identische Abbildung in Betracht kommt. Bei nur auf Baselinevariablen gestützten Prognosesystemen findet eine einmalige Berechnung von $\mathrm{RI}(t)$ zum Zeitpunkt $t = 0$ statt. Dieser Wert bleibt für nachfolgende Zeitpunkte t unveränderlich, ein Verweis auf die Zeit kann, mit RI anstelle von $\mathrm{RI}(t)$, unterbleiben. Enthält $\mathrm{PI}(t)$ zeitabhängige Kovariablen, erfordert deren Wertänderung bei den betroffenen Patienten die Neuberechnung von $\mathrm{RI}(t)$.

Durch Anwendung der „Minimal p-value“-Methode [6] auf $\mathrm{RI}(t)$ sollten zu jedem der fünf besonders interessierenden Verlaufszeitpunkte die Grenzen zwischen sich bzgl. der Überlebenswahrscheinlichkeiten maximal unterscheidenden Risikogruppen gefunden werden.[33] Durch die jeweilige Verwendung der aktualisierten Werte zur zytogenetischen Remission und daraus resultierenden Veränderungen der $\mathrm{RI}(t)$ konnten sich für die fünf Zeitpunkte unterschiedliche Gruppengrenzen ergeben. Zu allen Zeitpunkten wurde einheitlich die Definition von vier Risikogruppen angestrebt. Die Erfahrung hatte gezeigt, dass sich Überlebenswahrscheinlichkeiten medikamentös behandelter Patienten in der CML durch Prognosesysteme auf Basis von Baselineparametern gut in drei Gruppen trennen lassen [42, 105], ohne dass die Überlebenswahrscheinlichkeiten zweier Gruppen in den verschiedenen Patientenstichproben allzu nahe beieinander liegen [13, 18, 43, 64, 90]. Neben den drei nach dem New CML-Score zu erwartenden Gruppen

[33] Die Beurteilung erfolgte mittels des adjustierten p-Wertes p_{ad} zur Teststatistik des Logrank-Tests.

wurde mittels der Remissionsvariablen die Möglichkeit der Identifikation einer weiteren Gruppe mit besonders günstigen Überlebenswahrscheinlichkeiten vermutet. Dabei sollte immer die Mindestgruppengröße von 10% aller Patienten beachtet werden. Nach den rein statistisch ermittelten „Vorschlägen“, galt es, die Gruppengrenzen auch hinsichtlich medizinischer Aspekte zu überprüfen.

Mit dem gewählten Vorgehen werden zwar bei den zeitabhängigen Kovariablen nur die bis zum jeweiligen Zeitpunkt bekannten, aktuellsten Merkmalsausprägungen berücksichtigt, jedoch finden immer dieselben für (2.4) geschätzten Koeffizienten Verwendung. In die Schätzung dieser Koeffizienten gingen - ohne zeitliche Einschränkung - alle registrierten deutlichen ZR ein. Mit dem Ziel der Identifikation eines möglichst guten und zugleich verständlichen sowie leicht anwendbaren Prognosesystems erschien das hier gewählte Verfahren trotz des „zeitlichen Widerspruchs“ als das Optimale.
Die Gründe für Modell (2.3) und gegen ein Cox-Modell mit zu einer bestimmten Landmark „zeitunabhängigen“ Hazardraten wurden bereits in Abschnitt 2.7.1 erläutert. Man mochte einwenden, dass man statt nur zu einem, alternativ hätte zu jedem der fünf Zeitpunkte ein eigenes Cox-Modell berechnen können. Ein immenser Nachteil für die praktische Anwendung wäre dabei schon das Arbeiten mit zu jedem Zeitpunkt anderen Koeffizienten. Zu späteren Verlaufszeitpunkten würde beim Landmarkansatz eine Annährung an die Koeffizientenschätzer aus (2.4) zu erwarten sein. Zu den frühen Zeitpunkten aber stünde ein großer Unterschied zum Modell mit zeitabhängigen Hazardraten zu vermuten. Innerhalb der ersten neun Monate nach Therapiebeginn etwa, wurde erfahrungsgemäß nur ein Bruchteil aller insgesamt beobachtbaren ersten deutlichen ZR verzeichnet (vgl. z.B. [48, 57]). Entsprechend hat bei einem Cox-Modell zum Landmarkzeitpunkt „9 Monate“ ein Koeffizientenschätzer zur Variablen „zytogenetische Remission“ relativ zu den Schätzern für die Baselinevariablen ein erheblich geringeres Gewicht als im Falle der Schätzer aus (2.4). Obwohl die besonders günstigen Überlebenswahrscheinlichkeiten für Patienten mit früher wie mit später erster deutlicher ZR gelten, können so mit dem Cox-Modell zur Landmark „9 Monate“ - wegen ähnlicher RI(t) - Überlebenswahrscheinlichkeiten zwischen Patienten mit deutlicher ZR und Patienten ohne deutliche ZR nicht ausreichend unterschieden werden. Die Verwendung der Schätzer aus (2.4) verleiht dagegen dem günstigen Umstand einer frühen deutlichen ZR ein angemesseneres Gewicht und dürfte sehr wahrscheinlich zur Definition einer eigenen Risikogruppe führen, deren Zugehörigkeit durch günstige Baselinevariablenwerte mitbestimmt sein könnte.
Und schließlich, ist die Entwicklung eines Prognosesystems in der Lernstichprobe ein explorativ es Vorgehen. Insofern erschien es angebracht, mittels des zeitabhängigen Hazardmodells (2.3) alle verfügbaren Informationen bei der Entwicklung auszunutzen.

Ziel war die Identifikation eines möglichst guten und zugleich verständlichen sowie leicht anwendbaren Prognosesystems. Die Berücksichtigung der zytogenetischen Ergebnisse war das zentrale Element. Der hier gewählte Ansatz zur Entwicklung eines Prognosesystems unterstützt eine möglichst zeitnahe und angemessen gewichtete Verwendung der Remissionsergebnisse und erfordert dabei die Analyse mit nur einem Cox-Modell, welches hinsichtlich der zugelassenen Variablen und Ereignisse alle Informationen in der Lernstichprobe ausnutzt.

2.14 Beurteilung des Prognosesystems in der Lernstichprobe

Während die Aussagekraft z.B. eines „logistischen Prognosesystems“ anhand einer Kreuztabelle leicht beschrieben werden kann [53], gestaltet sich die Beurteilung der prognostischen Leistung

eines „Cox-Prognosesystems“ mathematisch schwieriger, insbesondere wenn zeitabhängige Kovariablen beteiligt sind.

Die prognostische Fähigkeit des identifizierten Cox-Modells läßt sich mittels vorausgesagter Überlebenswahrscheinlichkeiten beschreiben. Sind zeitabhängige Kovariablen beteiligt, muss die Aktualisierung der Risikowerte bedacht werden, auf welchen die Voraussagen fußen. Zwar sind die zukünftigen Werte der zeitabhängigen Kovariablen (formell) nicht bekannt, doch lässt sich u.U. eine approximative bedingte Überlebenswahrscheinlichkeit über ein kurzes Zeitintervall der Länge λ schätzen. Voraussetzung ist über einen Zeitabschnitt $(t, t+\lambda)$ die Annahme annähernd konstanter Kovariablenwerte x_{ij}, mit $i = 1, \ldots, n$ für n Patienten und $j = 1, \ldots, p$ für p Kovariablen. Ist diese Annahme vertretbar, so schätzt man für einen zum Zeitpunkt t lebenden Patienten i die Wahrscheinlichkeit bis zum Zeitpunkt $t+\lambda$ zu überleben [5, 24] durch

$$\hat{p}_i(t, t+\lambda) = \exp\left\{-(\hat{H}_0(t+\lambda) - \hat{H}_0(t)) \exp\left(\sum_{j=1}^{p} \hat{\beta}_j x_{ij}(t)\right)\right\}. \tag{2.14}$$

Die kumulierte Baselinehazardfunktion wird analog Abschnitt 2.11 nach Breslows Näherungsformel [22] berechnet:

$$\hat{H}_0(t) = \sum_{i=1}^{n} \frac{\delta_i \; I(t_i \leq t)}{\sum_{l \in R(t_i)} \exp\left\{\sum_{j=1}^{p} \hat{\beta}_j x_{lj}\right\}}.$$

Kann von einer konstanten Baselinehazardfunktion mit $h_0(t) = h^*$ ausgegangen werden, so vereinfacht sich (2.14) zu

$$\hat{p}_i(t, t+\lambda) = \exp\left\{-\hat{h}^* \lambda \exp\left(\sum_{j=1}^{p} \hat{\beta}_j x_{ij}(t)\right)\right\}, \tag{2.15}$$

wobei $\hat{h}^*$ für die geschätzte Steigung steht, wenn $\hat{H}_0(t)$ als lineare Funktion von t dargestellt wird. Die bedingte Wahrscheinlichkeit für den zum Zeitpunkt t noch lebenden Patienten i innerhalb des Zeitintervalls $(t, t+\lambda)$ ein Ereignis zu beobachten, wird dementsprechend durch

$$\hat{q}_i(t, t+\lambda) = 1 - \hat{p}_i(t, t+\lambda) \tag{2.16}$$

geschätzt. Diese nach den Risikowerten erwarteten Sterbewahrscheinlichkeiten können für verschiedene Zeitintervalle berechnet und mit dem tatsächlich beobachteten Anteil von Todesfällen innerhalb definierter Risikogruppen verglichen werden [5, 24]. Je näher erwartete und beobachtete Werte beieinander liegen, desto höher die prognostische Aussagekraft des Modells. Da das zeitunabhängige Modell immer konstante Kovariablenwerte besitzt, kann dort die Modellprognose $\hat{p}_i(t, t+\lambda)$ nach (2.14) exakt berechnet werden.

Nach Festlegung der Risikogruppen, sollte das neue Prognosesystem nach Maßgabe seiner prognostischen Fähigkeiten bei den Daten der Lernstichprobe, beurteilt durch die in Abschnitt 2.2 unter a) und b) angeführten Kriterien, auf seine Vorschlagsberechtigung untersucht werden. Da die p-Werte der nach Abschnitt 2.2 vorgenommenen Logrank-Tests der Lernstichprobe entstammen, können diese - wie auch die weiteren Ergebnisse in der Lernstichprobe - lediglich der

Stützung der Hypothese dienen, dass das untersuchte Prognosesystem für eine definierte Patientengruppe valide und verlässliche Vorhersagen liefert. Aussagekräftigere Ergebnisse können in unabhängigen Validierungsstichproben gewonnen werden.

2.15 Beurteilung des Prognosesystems in einer unabhängigen Validierungsstichprobe

Die Daten aus der Lernstichprobe, welche dem Prognosesystem Gestalt verliehen, liefern bei seiner Beurteilung ein tendenziell günstigeres Ergebnis als ein bei der Modellentwicklung nicht eingesetzter Datensatz. Idealerweise sollte die Leistung eines Prognosesystems daher innerhalb verschiedener, unabhängiger Validierungsstichproben untersucht werden [70, 85, 103, 121]. Auch der Einsatz ausgefeilter statistischer Überprüfungsmethoden in der Lernstichprobe kann die Überprüfung in einem neuen Datensatz nicht gleichwertig ersetzen.

Unter Verwendung der in der Lernstichprobe geschätzten Koeffizienten sollten zunächst die Risikowerte der Patienten der Validierungsstichprobe zu jedem der sieben Zeitpunkte berechnet und die Risikogruppeneinteilungen der Lernstichprobe auf die Validierungsstichprobe übertragen werden. Wie im Falle der Lernstichprobe, erfolgte dann die Beurteilung der prognostischen Leistung des Prognosesystems in der Validierungsstichprobe gemäß der in den Abschnitten 2.2 und 2.14 beschriebenen Kriterien.

Kapitel 3

Gewinnung und Aufbereitung der Patientendaten

Vor Beginn der Analyse prognostischer Faktoren beschreibt dieses Kapitel den Weg zum Erhalt der Analysestichprobe. Dabei wird dargestellt, wie Studien rekrutiert und welche Ein- und Ausschlusskriterien für Studien und Patienten festgelegt wurden. Insbesondere im Hinblick auf die zeitabhängige Variable werden Probleme erörtert, auf die vor der Bildung der Analysestichprobe aus Patientendaten verschiedener Studien eingegangen werden musste.

3.1 Identifikation und Rekrutierung relevanter Studien

Bereits zur Umsetzung des von Hasford et al. [42] realisierten ersten Vorhabens des C.P.F.P. wurden ab April 1995 Studienleiter mit relevanten Patientendaten kontaktiert. Um für das Projekt potenziell in Frage kommende Teilnehmer aus aller Welt zu identifizieren, wurden wiederholt MEDLINE-Recherchen durchgeführt, Abstracts und Konferenzberichte gelesen sowie Pharmafirmen und das „National Cancer Institute“ der USA angeschrieben. Auch dank des engen Kontaktes zur Gruppe der E.I.C.M.L. lagen Informationen über alle weltweit durchgeführten größeren Studien zur Verwendung von IFN-α bei CML-Patienten vor.

Jedem Studienleiter wurde ein das Forschungsprojekt skizzierender Brief, ein englischsprachiger Forschungsplan sowie eine Variablen- und Kodierungsliste zu den interessierenden Patientendaten zugesandt. Die zu erhebenden Variablen waren vorher mit den Mitgliedern der E.I.C.M.L. abgesprochen worden. Studienleiter, die nicht auf das Anschreiben reagierten, wurden wiederholt kontaktiert. Ziel war, zumindest eine Begründung für die Ablehnung einer Teilnahme zu erhalten.

Insgesamt wurden 48 Studienleiter angeschrieben. Nach mehrfachen Kontaktversuchen war von 40 eine Antwort zu bekommen. Davon erwiesen sich zehn Studien als nicht für das C.P.F.P relevant, zehn Studienleiter lehnten eine Teilnahme ab und weitere fünf beteuerten bei jedem Telefonat, sie würden ihre Daten in den nächsten Wochen weiterleiten, ohne dass die Daten jemals in München ankamen. Die Daten eines Studienleiters hatten einen zur Analyse ungenügenden Qualitätsstandard. Letztlich standen die Patientendaten von 14 Studien zur Durchführung des ersten Teils des C.P.F.P. zur Verfügung. Wenn auch einige Studienleiter kleinerer Studien sich nicht am C.P.F.P. beteiligten, so konnten doch die Patientendaten der meisten publizierten größeren, für das Projekt relevanten Studien in die Datenbank aufgenommen werden. Von den großen Studien zur Verwendung von IFN-α mit bereits einigen Jahren medianer Beobachtungs-

zeit wurden lediglich die Daten von Ozer et al. [83], der Gruppe um Kantarjian und Talpaz [60, 114] und der 1997 veröffentlichten Studie von Guilhot et al. [38] nicht bereitgestellt. Die erhaltenen Studiendaten kamen aus den Benelux-Ländern [15], Deutschland [47, 48, 65], Frankreich [37], Großbritannien [2], Italien [1, 57], Japan [80], Österreich [115, 116], Spanien [78, 107] und den USA [106]. Nicht immer entsprach die nach München versandte Patientenzahl der in den Publikationen angeführten: Manche Studienleiter von nicht randomisierten Studien hatten Daten zu zusätzlichen Patienten geschickt, andere hatten, in Übereinstimmung der ihnen mitgeteilten Ein- und Ausschlusskriterien, Daten zu nicht qualifizierten Patienten erst gar nicht weitergegeben.

Zur Beantwortung des Forschungsgegenstandes vorliegender Arbeit wurden die Studienleiter aller 14 Studien gebeten, ein Update ihrer Patientendaten und zusätzlich alle wesentlichen Details zur zytogenetischen Remission bereitzustellen. Mit Mahon et al. [74] war eine weitere Studiengruppe für das zweite Ziel des C.P.F.P. zu gewinnen. Dagegen konnten von den Verantwortlichen der Studien aus Rom [1], Essen [65], Castilla-León [78] und Ann Arbor [106] keine zufriedenstellenden Daten hinsichtlich eines Updates und v.a der zytogenetischen Remission gewonnen werden. Daher wurden diese vier Studien von der Auswertung ausgeschlossen. Mit der Studie aus Bordeaux [74] waren mehr potenziell auswertbare Patienten hinzugekommen (maximal n=141) als durch die vier kleineren Studien wegfielen (maximal n=125).[1]

3.2 Die Überprüfung der Datenqualität

Die eingetroffenen Datensätze wurden Qualitätssicherungsprozeduren mit ausführlichen Plausibilitätsprüfungen unterzogen. Alle Variablen wurden auf Vollständigkeit, atypische oder unklare Angaben überprüft. Datumsangaben wurden auf ihre logische Abfolge untersucht. Die Kodierung kategorialer Variablen musste klar definiert sein. Bei klinischen Variablen wurden, unter Berücksichtigung der Krankheit, die Überschreitung oberer und unterer Grenzwerte inspiziert. Inhaltlich zusammenhängende Variablen wurden miteinander verglichen und bei Widersprüchlichkeiten Rückfrage gehalten. Zur Überprüfung eines fehlerfreien Datentransfers wurden für alle Variablen einer Studie die Anzahl vorhandener und fehlender Angaben, für die kategorialen Variablen zusätzlich die Werteverteilung und für die metrischen Variablen Minimum, Maximum, Median und Mittelwert bestimmt. Für die Therapiearme jeder Studie wurden die Überlebenskurven der Patienten berechnet. Die ermittelten Ergebnisse wurden an die Studienleiter geschickt, mit der Bitte, alle Statistiken zu ihrer Studie anhand der Daten in der eigenen Datenbank gegenzuprüfen.

3.3 Die Ein- und Ausschlusskriterien

Studien

Von allen Studienleitern wurde das Protokoll ihrer Studien erbeten. Die Berücksichtigung von nur prospektiven Studien mit einheitlichem Protokoll sollte einen Grundstandard an Datenqualität garantieren. Jede Studie musste Patienten enthalten, die mit IFN-α allein oder in Kombination mit einer anderen Therapie behandelt worden waren.

Patienten

[1] Maximal, da durch fehlende oder nicht auswertbare Daten die Fallzahl bei der Analyse reduziert wurde.

Nur zum Zeitpunkt der Diagnose Ph-positive oder BCR-ABL-positive Patienten wurden für die Analyse zugelassen. Angaben zu den Variablen Alter, Geschlecht, Diagnosedatum, vorgesehene Therapien („intention-to-treat“ [ITT]) und tatsächliche Therapien, Therapiebeginn, Therapieende (ggf.), Überlebenszeit und Überlebensstatus mussten grundsätzlich vorhanden sein. Auch die direkt nach dem Diagnosedatum einsetzenden therapeutischen Maßnahmen (Vortherapien) mussten bekannt sein.

Zwischen Vortherapie und Kombinationstherapie wurde wie folgt unterschieden: Eine Chemotherapie, die bereits in der Zeit zwischen Diagnose und IFN-α-Therapiebeginn zur Anwendung kam, galt als Vortherapie. Jede Chemotherapie, die (noch) nach IFN-α-Therapiebeginn verabreicht wurde, bildete gemeinsam mit IFN-α einen als Kombinationstherapie bezeichneten Therapieansatz.

Das Diagnosedatum IFN-α-behandelter Patienten sollte nach dem 01.01.85 liegen, dem Jahr, ab welchem die ersten größeren Therapievergleiche zwischen IFN-α und einer Chemotherapie gestartet wurden. Die Dosierungen waren wegen häufig nicht konsequenter Dokumentation nicht erhoben worden, allerdings mussten die Patienten tatsächlich IFN-α erhalten haben, um für die Betrachtung einer durch IFN-α induzierten Remission von Interesse zu sein. Zwischen Diagnosedatum und Therapiebeginn mit IFN-α durfte nicht mehr als ein halbes Jahr liegen, um sicherzustellen, dass für die meisten Patienten der Analysestichprobe die Therapie noch in der frühen chronischen Phase begonnen hatte. In Anlehnung an die von der italienischen Studiengruppe aufgestellten Kriterien [57] wurde überprüft, ob sich die Patienten zum Zeitpunkt der Diagnose noch in chronischer Phase befanden. Dies war insofern erforderlich, als eine Therapie mit IFN-α im fortgeschrittenen Krankheitsstadium in den seltesten Fällen noch wirkt und damit ohnehin nicht empfohlen werden kann. Entsprechend wurden Patienten ausgeschlossen, die mindestens eine der nachfolgenden Bedingungen erfüllten und damit bei Diagnose bereits eher der akzelerierten Phase oder Blastenphase (A/BP) zugerechnet werden mussten:

- Mehr als 10% Blasten im peripheren Blut (p.B.)
- Zusammen mehr als 30% Blasten und Promyelozyten im p.B.
- Mehr als 15% Blasten im Knochenmark
- Zusammen mehr als 50% Blasten und Promyelozyten im Knochenmark
- Extramedulläre Manifestationen
- Mindestens eine der chromosomalen Aberrationen „Trisomie 8“, „zwei Ph-Chromosome“ oder „Isochromosom 17“

Ende 1999 wurde die Datenbank zu den 11 Studien geschlossen. Tabelle 3.1 gibt eine Übersicht bzgl. der erhaltenen Patientenzahlen. Von den insgesamt 3009 Patienten waren 2859 (95%) bei Diagnose Ph-positiv oder BCR-ABL-positiv, bei 106 (3,5%) waren Ph-Status und BCR-ABL-Status unbekannt und 44 (1,5%) wurden als „Ph-negativ“ diagnostiziert.[2] Von den 2859

[2]Der Anteil Ph-negativer CML-Patienten ist nicht repräsentativ; einige Studienleiter verschickten von vornherein keine Daten Ph-negativer Patienten.

Ph-positiven[3] Patienten sollten initial, überwiegend qua Randomisierungsergebnis, 1788 (62,5%) eine IFN-α-Therapie erhalten, die übrigen 1071 (37,5%) eine reine Chemotherapie. Von den 1788 Patienten, für die eine IFN-α-Therapie vorgesehen war, mussten entsprechend obiger Kriterien weitere Patienten von der Analyse ausgeschlossen werden. Nach Anwendung aller Ausschlusskriterien außer den sechs Kriterien für Akzeleration bzw. Blastenkrise und mit der Hinzunahme von einigen wenigen der 1071 „Chemotherapie-Patienten“, die außerplanmäßig (zusätzlich) IFN-α bekommen hatten und auch gemäß der anderen Charakteristika qualifiziert waren, umfasste die Analysestichprobe 1485 Patienten.

Aufgrund fehlender Daten konnten v.a. die letzten vier der Kriterien für Akzeleration / Blastenkrise nicht bei allen Patienten überprüft werden. Ein grundsätzlicher Ausschluss aller Patienten, bei welchen mindestens einer der zu untersuchenden Parameter fehlte, hätte die Reduktion der Patientenstichprobe auf 27% der Fälle bedeutet und damit die Aussagekraft nachfolgender Analyseresultate entscheidend eingeschränkt. Andererseits konnten, wie oben begründet, Patienten

Tabelle 3.1: Studien und Patienten

Studie	Anzahl der Studien-patienten	Für eine IFN-α-Therapie vorgesehene, Ph-positive Patienten (ITT)	Zur Analyse vorgesehene Pat. ohne Ausschluss wegen Akzeleration bzw. Blastenphase	Zur Analyse zugelassene Pat. nach Ausschluss wegen Akzeleration bzw. Blastenphase
	n	n	n^a	n^b
A - CML III [115]	52	52	46	45
A - CML V [116]	72	72	70	69
B/NL/LUX [15]	200	100	96	84
D - CML I [47]	605	134	121	96
D - CML II [48]	366	226	216	184
D - Essen [65]	51	51	0	0
E - C.-León [78]	50	29	0	0
E - Madrid [107]	131	131	100	99
F - Bordeaux [74]	164	156	141	141
F - Poitiers [37]	207	207	207	196
GB - CML III [2]	554	262	198	186
I - Bologna [57]	322	218	210	209
I - Rom [1]	55	49	0	0
J - Hamamatsu [80]	160	81	80	75
USA - Ann A. [106]	20	20	0	0
Gesamt	3009	1788	1485	1384

[a] Nach Anwendung aller Ein- und Ausschlusskriterien gemäß Abschnitt 3.3, jedoch ohne die sechs Kriterien für Akzeleration bzw. Blastenkrise.

[b] Nach Anwendung aller Ein- und Ausschlusskriterien gemäß Abschnitt 3.3, einschließlich der sechs Kriterien für Akzeleration bzw. Blastenkrise. Diese 1384 Patienten bildeten zunächst die Analysestichprobe.

[3] Zwischen „Ph-positiv“ und „Ph-negativ, aber BCR-ABL-positiv“ wird hier wie im folgenden nicht mehr unterschieden, der Zusatz „oder BCR-ABL-positiv“ wird weggelassen und nur noch von „Ph-positiven“ Patienten gesprochen.

mit Parameterwerten, die auf akzelerierte Phase oder Blastenkrise hinwiesen, nicht zur Analyse zugelassen werden. Ein Vorabvergleich der Patienten mit vollständigen und nicht vollständigen Daten zu den relevanten Kriterien (vgl. Tabelle 3.2) war erforderlich.[4]

Tabelle 3.2 enthält die Anzahl der auszuschließenden Patienten pro Ausschlusskriterium. So waren z.B. im Falle des Kriteriums „> 10% Blasten im p.B." 34 der 1463 Patienten mit vorliegenden Daten auszuschließen, ein Anteil von 0,0232. Diesem Anteil entsprechend, wurde die

Tabelle 3.2: Ausschluss wegen nicht chronischer Phase zum Diagnosezeitpunkt

Untersuchtes Ausschlusskriterium bei 1485 Patienten	Anzahl auszuschließender Patienten	Anzahl der Patienten ohne Zutreffen des Ausschlusskriteriums	Anzahl der Patienten mit fehlenden Werten	Geschätzte Anzahl auszuschließender Fälle unter Patienten mit fehlenden Werten
	n (Prozent)[a]	n (Prozent)[b]	n	n [95%-K.I.][c]
> 10% Blasten im p.B.	34 (2,32%)	1429 (97,68%)	22	1 [0; 1]
> 30% (Blasten + Promyelozyten) im p.B.	6 (0,45%)	1321 (99,55%)	158	1 [0; 2]
> 15% Blasten im Knochenmark	10 (1,11%)	888 (98,89%)	587	7 [3;13]
> 50% (Blasten + Promyelozyten) im Knochenmark	1 (0,17%)	574 (99,83%)	910	2 [0; 9]
Extramedulläre Manifestationen	48 (3,94%)	1170 (96,06%)	267	11 [7;14]
Bestimmte chromosomale Aberration[d]	15 (1,69%)	871 (98,31%)	599	11 [5;17]
Gesamt	114			33

[a]Die Prozentzahl der auszuschließenden Patienten bezieht sich auf die Gesamtzahl der Patienten (Anzahl Spalte 2 + Anzahl Spalte 3), für die das Kriterium überprüfbar war.

[b]Die Prozentzahl der Patienten ohne Zutreffen des Ausschlusskriteriums bezieht sich auf die Gesamtzahl der Patienten (Anzahl Spalte 2 + Anzahl Spalte 3), für die das Kriterium überprüfbar war.

[c]Anzahl Spalte 5 berechnete sich aus Anzahl Spalte 4 multipliziert mit dem geschätzten Anteil in Spalte 2 (Prozentzahl). Das 95%-Konfidenzintervall [95%-K.I.] wurde auf Basis einer Approximation der Binomialverteilung durch die Poissonverteilung geschätzt (vgl. Rüger [94]). Bei Bestimmung der unteren Grenze des 95%-K.I. wurde zur nächstkleineren ganzen Zahl abgerundet, bei Bestimmung der oberen Grenze des 95%-K.I. zur nächstgrößeren ganzen Zahl aufgerundet. Die tatsächlichen K.I. sind damit etwas kleiner.

[d]Mindestens eine der chromosomalen Aberrationen „Trisomie 8", „zwei Ph-Chromosome" oder „Isochromosom 17".

Anzahl auszuschließender Patienten unter den Patienten mit fehlenden Werten auf 1 geschätzt. Wegen der sechs Ausschlusskriterien mussten bei 1485 Patienten sechs Variablen mit insgesamt 8910 Werten überprüft werden. Davon waren 6367 Werte überprüfbar, 2543 Werte fehlten. Bei 11 Patienten lagen zwei Ausschlussgründe vor und bei einem Patienten drei, so dass die 114 verschiedenen Hinweise auf akzelerierte Phase / Blastenkrise zum Ausschluss von 101 Patienten

[4]Vor dem Zusammenfügen von Daten mit verschiedenen Therapieansätzen bedarf es im Hinblick auf den Zielparameter einer Untersuchung auf dessen Abhängigkeit vom Therapieansatz. Diese Untersuchung wird im nächsten Abschnitt nachgereicht.

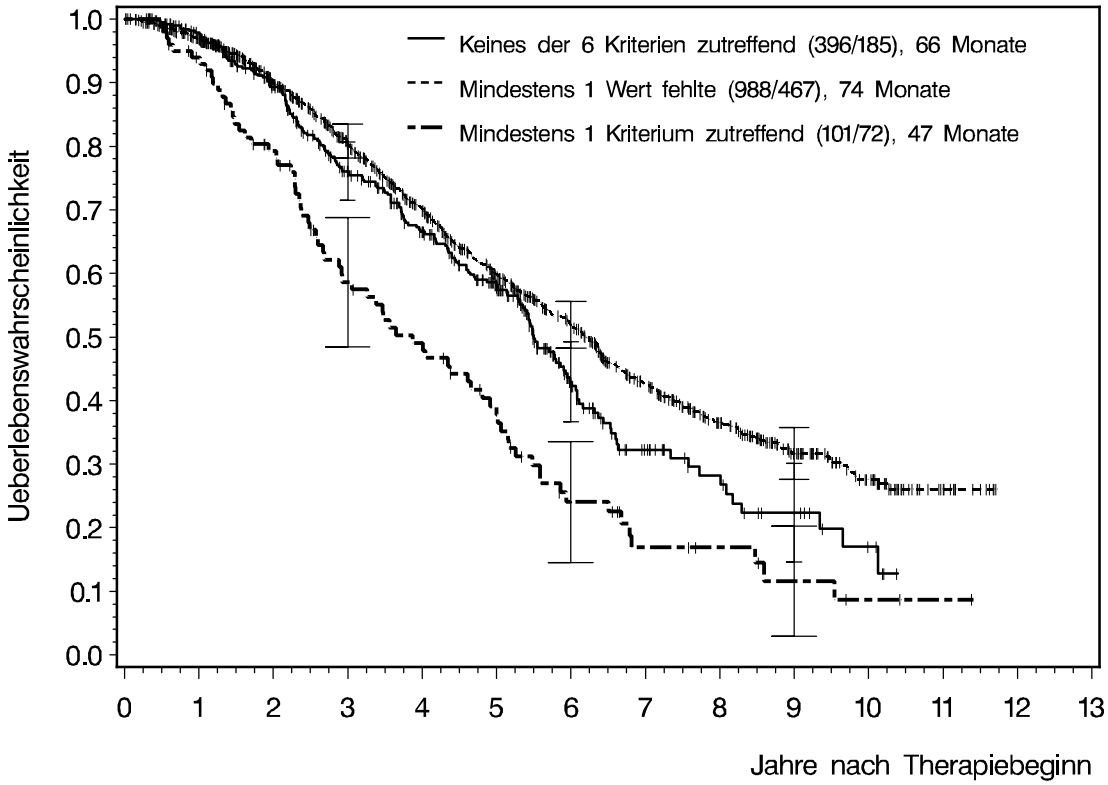

Abbildung 3.1: Überprüfung von 1485 Patienten auf chronische Phase zum Diagnosezeitpunkt. Kaplan-Meier-Kurven zur Schätzung der Überlebenswahrscheinlichkeiten dreier unterschiedlicher Patientengruppen. Mit Kriterien / Kriterium sind die Ausschlusskriterien gemeint. Die Legende „(396/185), 66 Monate“ bedeutet: Unter den 396 Patienten mit anhand der zugehörigen Kaplan-Meier-Kurve beschriebenen Überlebenswahrscheinlichkeiten wurden 185 Todesfälle beobachtet. Die mediane Überlebenszeit betrug 66 Monate. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Überlebenswahrscheinlichkeit mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Nennung in der Legende von oben nach unten.

führten. Mit derselben Quote wäre bei 33 verschiedenen Hinweisen eine geschätzte Anzahl von 29 Patienten auszuschließen.[5]
Da die Erfüllung eines Kriteriums zum Ausschluss genügte, war das Fehlen anderer Merkmalsausprägungen bei den 101 Patienten (7% von 1485) nicht weiter relevant - diese Patienten sind zu Recht ausgeschlossen. Bei 396 Patienten (27%) waren die Werte zu allen sechs Variablen vorhanden und kein Ausschlusskriterium erfüllt - diese Patienten sind zu Recht eingeschlossen. Bei den übrigen 988 Patienten (67%) fehlte mindestens eine der sechs Merkmalsausprägungen, jedoch führten die vorhandenen Werte nicht zu einem Ausschluss. Gemäß obiger Hochrechnung, hätten bei Vorliegen aller Merkmalsausprägungen 29 der 988 Patienten (3% von 988) ausgeschlossen werden müssen. Ungünstigenfalls, bei Addition der oberen Intervallgrenzen (Spalte 5) und bei Vernachlässigung, dass mehrere Ausschlussgründe jeweils nur einen Patienten betreffen

[5]Es ist zu beachten, dass bei diesen Schätzungen von einem zufälligen Fehlen der Werte ausgegangen wurde, d.h. das Fehlen der Werte galt als unabhängig von der unbekannten Merkmalsausprägung. Auch wurden keine Korrelationen zwischen den Variablen berücksichtigt.

könnten, hätte man ca. 56 Patienten (6%) ausschließen müssen.

Abbildung 3.1 zeigt deutlich niedrigere Überlebenswahrscheinlichkeiten für die 101 Patienten, deren Phase zum Diagnosezeitpunkt als „nicht chronisch“ bewertet wurde. Beim jeweiligen Vergleich mit einer der oberen Kurven lagen beide p-Werte der Logrank-Tests unter 0,0005. Zusammen mit der medianen Überlebenszeit von 47 Monaten bestätigten die Testergebnisse das Gelingen, mit Hilfe der sechs Ausschlusskriterien eine Gruppe von Patienten zu identifizieren, die sich zum Diagnosezeitpunkt offensichtlich bereits in fortgeschrittener Krankheitsphase befanden und deren Überlebenszeit durch Anwendung einer IFN-α-Therapie nur in Einzelfällen verlängert werden konnte. Wie erwartet, bedurfte es hier zur Erkennung der ungünstigen Prognose keines Modells; die Patienten wurden zu Recht von weiteren Analysen ausgeschlossen.
Die Überlebenskurve der 988 Patienten mit mindestens einem fehlenden Wert lag nach Ende des fünften Jahres erkennbar oberhalb der Kurve zu den 396 komplett überprüfbaren Patienten, bei welchen definitiv keines der Ausschlusskriterien zutraf. Der Logrank-Test zeigte einen statistisch signifikanten Überlebensvorteil zugunsten der 988 Patienten ($p = 0{,}0140$). Aus diesem Unterschied ließ sich keine medizinische Erkenntnis ableiten, wohl aber bekräftigte er das Ergebnis auf Basis der Hochrechnungen: Unter den 988 nicht komplett überprüfbaren Patienten dürften sich zum Diagnosezeitpunkt nur wenige bereits in fortgeschrittener Phase befunden haben. Die Zulassung der 988 Patienten beinhaltete die Verletzung der Ausschlusskriterien für geschätzte 29 Patienten, ein verschwindend geringer Nachteil im Vergleich zur Chance, dank der Daten zu ca. 950 zu Recht eingeschlossenen Patienten überhaupt ein verlässliches Prognosesystem identifizieren zu können.

3.4 Die Daten zum Hauptzielparameter Überlebenszeit

Nach Ausschluss der 101 Patienten, die sich bei Diagnose nicht mehr in chronischer Phase befunden hatten, waren 1384 Patienten für die Analysestichprobe qualifiziert. Der Weg dazu hatte über die Verwendung a priori festgelegter, von später erhaltenen Daten unabhängigen Definitionen sowohl der Überlebenszeit (Abschnitt 2.3) als auch der Ein- und Ausschlusskriterien (Abschnitt 3.3) geführt. Das Fehlen unerwünschter Einflüsse von Störvariablen auf die Überlebenszeit war damit aber nicht garantiert. Um bei der Identifizierung prognostischer Faktoren möglichst keinen Artefakten aufzusitzen, wurde die Daten der Analysestichprobe als nächstes hinsichtlich solch problematischer Zusammenhänge überprüft und Patienten ggf. auf Grund sich a posteriori ergebender Kriterien ausgeschlossen.

Damit Unregelmäßigkeiten wie z.B. ein Zusammenhang zwischen fortgeschrittener Krankheitsphase und einer Zensierung der Überlebenszeit erkennbar waren, mussten, zum Erreichen ausreichender Fallzahlen, Analysen im Datenpool aus allen elf Studien durchgeführt werden. Während die Überprüfung einer unverzerrten Überlebenszeitmessung meist den Datenpool voraussetzte, setzte umgekehrt, die Überprüfung der Vereinbarkeit der Überlebenszeiten aus verschiedenen Studien in einem Datenpool, unverzerrte Überlebenszeitmessungen voraus. Da nicht alle Sachverhalte auf einmal überprüfbar waren, wurden, nach Ausschluss von Patienten aufgrund einer bestimmten Datenlage, in der neu erhaltenen Analysestichprobe die zuvor schon untersuchten Zusammenhänge noch einmal betrachtet und ggf. weitere Patientenausschlüsse vorgenommen. Damit wurde angestrebt, dass in der bzgl. der Überlebenszeit endgültigen Analysestichprobe keine der überprüfbaren Verzerrungen oder Störparameter eine (erkennbare) Rolle spielte.

3.4.1 Verzerrungen und Störparameter innerhalb der einzelnen Studien

Im Rahmen des in Abschnitt 3.2 erwähnten Vorgehens wurden erkennbare Ungereimtheiten zunächst in den einzelnen Studien ermittelt. Wurden trotz kleiner Fallzahl unerwünschte Zusammenhänge bereits innerhalb einer Studie identifiziert, mussten die Gründe dafür geklärt oder Patienten von weiteren Analysen ausgeschlossen werden.

In einigen Studien war zwischen Diagnose und Therapiebeginn protokollgetreu grundsätzlich eine Chemotherapie vor der ersten IFN-α-Gabe verabreicht worden [2, 15], in anderen Studien oblag eine solche Vortherapie der individuellen Entscheidung des Arztes [57, 74, 107, 116].
Von den 69 Patienten der österreichischen Studie CML V [116] hatten 42 (61%) eine Vortherapie mit HU erhalten. Unabhängig vom Risikoprofil nach dem New CML-Score, zeigte sich für die 27 Patienten ohne Vortherapie ein statistisch signifikanter Überlebensvorteil gegenüber den 42 Vorbehandelten (Logrank-Test: $p = 0{,}0163$). Um von der Zeit der HU-Vortherapie herrührende, nichtzufällige Auswirkungen auf die Überlebenszeit ab IFN-α-Therapiebeginn auszuschließen, wurden die 42 vorbehandelten Patienten der Studie CML V aus der Analysestichprobe herausgenommen, zumal die Auswahlkriterien für die Vorbehandlung unbekannt waren.
Von den 209 Patienten aus Bologna [57] hatten 13 (6%) eine Vorbehandlung erfahren: 11 mit BU, einer mit Dibromomannitol und ein Patient mit anderer Therapie. Beim Vergleich der Überlebenswahrscheinlichkeiten stellte sich für die 13 vorbehandelten Patienten ein statistisch signifikanter Überlebensvorteil gegenüber den 196 ohne Vortherapie heraus (Logrank-Test: $p = 0{,}0219$). Auf die 13 vorbehandelten italienischen Patienten wurde bei der Analysestichprobe aus denselben Gründen verzichtet, wie bei den Österreichern. Bei allen übrigen Studien wurde kein statistisch signifikanter Zusammenhang zwischen Vortherapie und Überlebenszeit festgestellt.

In Bezug auf die Zeit zwischen Diagnose und Therapiebeginn war innerhalb der einzelnen Studien kein statistisch signifikanter Einfluss auf die Überlebenszeit feststellbar.

Abzüglich der 42 österreichischen und der 13 italienischen Patienten bestand die Analysestichprobe nunmehr aus 1329 Patienten.

3.4.2 Zusammenhänge zwischen Therapieverlauf, Zensierung und Follow-up der Überlebenszeit

Zum Datenbankschluss befanden sich 222 Patienten (16,7% von 1329) sowohl in chronischer Phase als auch noch unter IFN-α-Therapie. Die übrigen 1107 (83,3%) hatten zwar ihre IFN-α-Therapie aus verschiedenen Gründen beendet, waren z.T. aber auch noch in chronischer Phase. Die 1107 Gründe für den Abbruch der IFN-α-Therapie setzten sich zusammen aus 280 Therapieresistenzen (25,3%), 262mal unerwünschten Nebenwirkungenen (23,7%), 193mal A/BP (17,4%), 176 allogenen SZT (15,9%), 62 Verweigerungen weiterer IFN-α-Therapie (5,6%), 57 Todesfällen „unter Therapie“ in chronischer Phase (5,2%), 38 Therapieabbrüchen aus unbekannten Gründen (3,4%), 23 Therapieerfolgen (2,1%), 12 autologen SZT (1,1%), zweimal Zentrumswechsel (0,2%) und zweimal „loss to follow-up“ (0,2%). Mehrfachnennungen waren nicht möglich.

Therapieverlauf und Follow-up der Überlebenszeit - Rückschlüsse anhand von 193 Patienten mit IFN-α-Therapieabbruch wegen A/BP

Zur Einschränkung des Aufwandes für die Studiengruppen, war das Eintreten einer A/BP nach Therapieabbruch nur bei Patienten mit späterer SZT erfragt worden. Doch auch die 193

Patienten mit Therapieabbruchgrund „(A/BP)“ waren besonders geeignet für eine deskriptive Untersuchung auf einen möglichen Zusammenhang zwischen Therapieverlauf und fortgesetzter Dokumentation der Überlebenszeit, weil für diese Patienten bei ausreichender Beobachtungszeit ein relativ hoher Anteil an Verstorbenen und damit ein Indikator für eine gewissenhafte Dokumentation der Überlebenszeit erwartet werden konnte.
Mit 177 von 193 waren 91,7% der Patienten mit Therapieabbruchgrund „A/BP“ bereits verstorben. Von den übrigen 16 lag der Zensierungszeitpunkt bei 13 Patienten in Übereinstimmung mit dem letzten Update der jeweiligen Studie. Nur drei Patienten (1,6% von 193) waren als „lost to follow-up“ zu werten.
Abzüglich der 193 in fortgeschrittener Phase Transplantierten sowie der generell zensierten 244 Patienten mit SZT in CP, verblieben 670 Therapieabbrecher aus anderen Gründen. Davon waren 454 Patienten verstorben (67,8%). Von den 216 zensierten Patienten galten 23 (3,4% von 670) als „lost to follow-up“ bzgl. der Überlebenszeit. Zwischen den 193 und den 670 Patienten bestand weder ein statistisch signifikant unterschiedlicher Anteil an Patienten mit „loss to follow-up“ (χ^2-Test) noch hatten die beiden Gruppen ab IFN-Therapieabbruch unterschiedliche Zensierungsmuster (Logrank-Test mit zensierten Überlebenszeiten als Ereignis) oder waren die zensierten Patienten (16 vs. 216) unterschiedlich lange beobachtet worden (U-Test). Die Überlebenswahrscheinlichkeiten der 193 Patienten waren erwartungsgemäß statistisch signifikant geringer (Logrank-Test, Überlebenszeiten ab Therapiebeginn oder ab Therapieabbruch: jeweils $p < 0{,}0001$). Die Daten der 193 Patienten mit Therapieabbruchgrund „A/BP“ ließen für die Analysestichprobe insgesamt auf eine vom Therapieverlauf unabhängige Erfassung der Überlebenszeit schließen.[6]

Zusammenhänge zwischen Therapieverlauf, SZT und der Überlebenszeit

Von allen 1329 Patienten erhielten 365 Patienten (27,5%) eine allogene oder autologe SZT. Mit 244 stellten die Patienten mit allogener SZT in 1. CP den größten Anteil (18,4% von 1329). Die mediane Beobachtungszeit bis zum Zeitpunkt der SZT in 1. CP betrug 20 Monate. Nach SZT wurde die mediane Überlebenszeit noch nicht ereicht, 103 von 244 Patienten (42,2%) verstarben. Weitere 62 Patienten (4,7% von 1329) erhielten eine allogene SZT nach Ende der 1. CP (mediane Zeit bis zur SZT: 28 Monate). Die statistisch signifikant später durchgeführte SZT in fortgeschrittener Phase (U-Test: $p = 0{,}0092$) deutete das Verständnis der SZT als „Rettungstherapie“ an. Für die 62 Patienten lag die mediane Überlebenszeit nach SZT in (oder nach) A/BP bei 9 Monaten, 45 Patienten waren verstorben (72,6% von 62). Das vorherige Scheitern der IFN-α-Therapie führte nach der SZT damit zu ähnlich ungünstigen Überlebenswahrscheinlichkeiten wie man sie bei Fortsetzung der konservativen Therapie in A/BP beobachtete [52].
Dasselbe galt für die 18 Patienten (1,4%), die in fortgeschrittener Phase eine autologe Transplantation erhielten. Die mediane Zeit bis zur SZT betrug hier 35 Monate und die mediane Überlebenszeit danach 12 Monate (13 Todesfälle).
Eine autologe SZT in 1. CP erhielten weitere 41 Patienten (3,1%, 11 verstarben). Die mediane Zeit bis zur SZT lag bei 23 Monaten und die nachfolgende mediane Überlebenszeit bei 70 Monaten. Da die Daten der 59 autolog transplantierten Patienten aus einer Zeit stammten (medianes Transplantationsdatum: 05/04/95), in der diese Therapieform im Vergleich zu IFN-α oder einer allogenen SZT noch nicht etabliert war, sondern eher mangels vorherigem Therapieerfolg

[6] Bei den Analysen in diesem Teilabschnitt wurde angenommen, dass unter 32 Abbrechern „aus unbekannten Gründen“ kein so schwerwiegender wie „Abbruch wegen A/BP“ hätte übersehen werden können. Beschränkte man sich beim Vergleich mit den 193 Therapieabbrechern wegen A/BP auf die 638 Patienten mit bekannten Abbruchgründen (436 verstorben, 202 zensiert) erhielt man vergleichbare Ergebnisse.

verabreicht wurde[7], wurde auch zum Zeitpunkt der autologen SZT in CP nicht zensiert. Unter Vernachlässigung der Heterogenität der beiden Patientengruppen[8], zeigten Simon-Makuch-Kurven [104] und Mantel-Byar-Test [75] vergleichbare Überlebenswahrscheinlichkeiten bei den 41 autolog in CP transplantierten und den 1044 nicht oder in A/BP transplantierten Patienten. Damit sprachen auch die Überlebensdaten der in CP autolog transplantierten Patienten für eine Beibehaltung der Nichtzensierung zum SZT-Zeitpunkt, auch wenn unter Fortsetzung einer konservativen Therapie die Überlebenszeit in manchen Fällen anders ausgesehen haben mochte. Zusammengefasst ließ sich festhalten, dass die in Abschnitt 2.3 bzgl. der verschiedenen SZT-Arten angegebene Definition des Hauptzielparameters durch die Daten zu den 365 SZT-Patienten die erwartete nachträgliche Rechtfertigung erhielt.

Einflüsse des Diagnosedatums und des Datums der ersten IFN-α-Gabe auf die Überlebenszeit

Das früheste Diagnosedatum und zugleich erste Behandlungsdatum mit IFN-α war der 11.12.1985. Die letzte Diagnose einer CML bei einem der 1329 Patienten wurde am 22.06.1998 gestellt, der jüngste IFN-α-Therapiebeginn lag am 06.07.1998. Die Analysestichprobe wurde jeweils zum medianen Diagnosedatum (01.04.1990) und zum medianen IFN-α-Behandlungsbeginn (24.04.1990) in zwei Hälften geteilt. Die Überlebenswahrscheinlichkeiten der früher diagnostizierten und behandelten Patienten unterschieden sich, v.a. die ersten 6 Jahre, kaum von den Überlebenswahrscheinlichkeiten der später diagnostizierten und behandelten Patienten (Kaplan-Meier-Kurven und Logrank-Tests), doch waren die Beobachtungszeiten der letzteren Gruppe definitionsgemäß deutlich kürzer.

3.4.3 Die Überlebenszeit in Abhängigkeit vom vorgesehenen IFN-α-Therapieansatz

Für die 1329 Patienten der 11 verschiedenen Studien waren als Therapienansätze IFN-α-Monotherapie oder IFN-α in Kombination entweder mit Hydroxyurea oder Ara-C vorgesehen. Unter Betrachtung dieser drei Therapienansätze als Kategorien der Variablen „IFN-α-Therapieansatz“ sollte deren Einfluss auf den Hauptzielparameter Überlebenszeit untersucht werden, bevor eine gemeinsamen Analysestichprobe ohne Adjustierung für „IFN-α-Therapieansatz“ akzeptiert werden konnte. Diese Notwendigkeit wurde durch die Ergebnisse von Guilhot et al. [38] unterstrichen. Die Franzosen hatten einen statistisch signifikanten Überlebensvorteil der Kombinationstherapie IFN-α + Ara-C gegenüber einer IFN-α-Monotherapie nachweisen können. Schon wegen der durch die verschiedenen Studien bedingten unterschiedlichen Selektionsmechanismen, konnte dem hier durchgeführten, „Therapievergleich“ nur ein deskriptiver Charakter zugebilligt werden, dessen Hauptaufgabe das Erkennen eines in allen Analysen ggf. gesondert zu berücksichtigenden Überlebensunterschiedes zwischen den Therapieansätzen war.

Festlegung der Stichprobe zum Vergleich der vorgesehenen Therapieansätze

Um keine subjektiven Entscheidungen aus dem Therapieverlauf einfließen zu lassen, wurde auf Basis der 1329 für die Entwicklung des Prognosesystems relevanten Patienten das „ITT-Prinzip“ angewandt. Bei den sieben randomisierten Studien [2, 15, 37, 47, 48, 57, 80], denen 1017 (77%) der 1329 Patienten angehörten, war die vorgesehene Therapiekombination (ITT) mit dem Ran-

[7] Auch bei den in 1. CP transplantierten Patienten betrug die mediane Zeit bis zur SZT immerhin 23 Monate.
[8] Die autolog transplantierten Patienten waren tendenziell jünger, die Auswahl nicht zufällig etc.

domisationsergebnis festgelegt. Für die vier übrigen Studien [74, 107, 115, 116] mit zusammen 312 Patienten (23%) existierte jeweils ein per Studienprotokoll genau festgelegter Therapieansatz, welcher die vorgesehene Therapiekombination ebenfalls eindeutig definierte. Ohne 29 Patienten, die initial für HU randomisiert worden waren oder für die, abweichend vom Studienprotokoll, bereits initial andere Therapien vorgesehen waren, verblieben 1300 Patienten in der Stichprobe für die Therapievergleiche.

Unterschiedliche Selektionsmechanismen bei Daten verschiedener Studien

Für die Therapieansätze des vorliegenden Datensatzes waren die Patientenanteile, die einem bestimmten Selektionsmechanismus unterlagen, völlig unterschiedlich. Verschiedene Selektionsmechanismen tragen wesentlich zur natürlichen biologischen Heterogenität zwischen Patientenstichproben bei. Inwieweit Überlebensunterschiede zwischen Stichproben mit verschiedenen Selektionsmechanismen eher dieser Heterogenität oder unterschiedlichen Therapieansätzen zuzuschreiben ist, bleibt schwer beurteilbar. Durch die beschriebenen Ein- und Ausschlusskriterien wurde der Heterogenität aufgrund unterschiedlicher Selektionsmechanismen bereits entgegengewirkt. Indem man beim Vergleich verschiedener Patientengruppen hinsichtlich des Zielparameters eine nach validierten Risikogruppen stratifizierte Analyse vornimmt, kann der Einfluss der Heterogenität auf das Analyseergebnis wesentlich weiter reduziert werden. Dieser klassischen Aufgabe eines Prognosesystems (vgl. Abschnitt 1.3) werden im Falle des Zielparameters „Überlebenszeit IFN-α-behandelter Patienten“ die Risikogruppen des New CML-Scores von Hasford et al. [42] gerecht.

Überprüfung einer stichprobenunabhängigen Anwendbarkeit des New CML-Scores

Vor einer Anwendung des New CML-Scores [42] auf alle Patienten der Analysestichprobe galt es zu prüfen, ob das Prognosesystem bei den an seiner Entwicklung unbeteiligten Patienten von vergleichbarer Aussagekraft sein würde. Der New CML-Score war für 1279 (96%) der 1329 Patienten berechenbar. Die mediane Überlebenszeit der 1279 betrug 72 Monate, 611 Patienten (48%) waren verstorben.[9] Die Daten von 826 der 1279 Patienten (65%) waren als Teil der JNCI-Lernstichprobe [42] bei der Entwicklung des New CML-Scores beteiligt. Die mediane Überlebenszeit der 826 Patienten (412 verstorben (50%)) lag bei 69 Monaten und bei den 453 an der Score-Entwicklung unbeteiligten Patienten (199 verstorben (44%)) bei 76 Monaten. Abbildung 3.2 zeigt die Kaplan-Meier-Kurven zu den Risikogruppen des New CML-Scores in beiden Stichproben, zwischen welchen sich die Überlebenswahrscheinlichkeiten der jeweils selben Risikogruppe nicht statistisch signifikant unterschieden, während jeder paarweise Vergleich der Überlebenswahrscheinlichkeiten zweier verschiedener Risikogruppen zu einem p-Wert $< 0{,}0005$ führte (Logrank-Test). Mit der stichprobenunabhängigen, deutlichen Diskriminierung der drei Risikogruppen brauchte in der Analysestichprobe bei Verwendung des New CML-Scores im folgenden keine Unterscheidung für die Herkunft aus der JNCI-Lernstichprobe vorgenommen werden. Die hohen statistisch signifikanten Korrelationen zwischen Risikogruppe und Überlebenswahrscheinlichkeiten (Logrank-Test: $p < 0{,}0001$ in beiden Stichproben) sprachen für risikostratifizierte Vergleiche [86] zwischen Stichproben mit unterschiedlichen Verteilungen hinsichtlich

[9] Die zeitliche Differenz zwischen dem Erhebungszeitpunkt der Scorevariablen (bei Diagnose) und dem Beginn der Überlebenszeitrechnung (erster Therapietag mit IFN-α) war, wegen ihrer „Ereignislosigkeit“ und ihrer kurzen Dauer in Relation zu den medianen Beobachtungszeiten, für die Aussagekraft des New CML-Scores unerheblich.

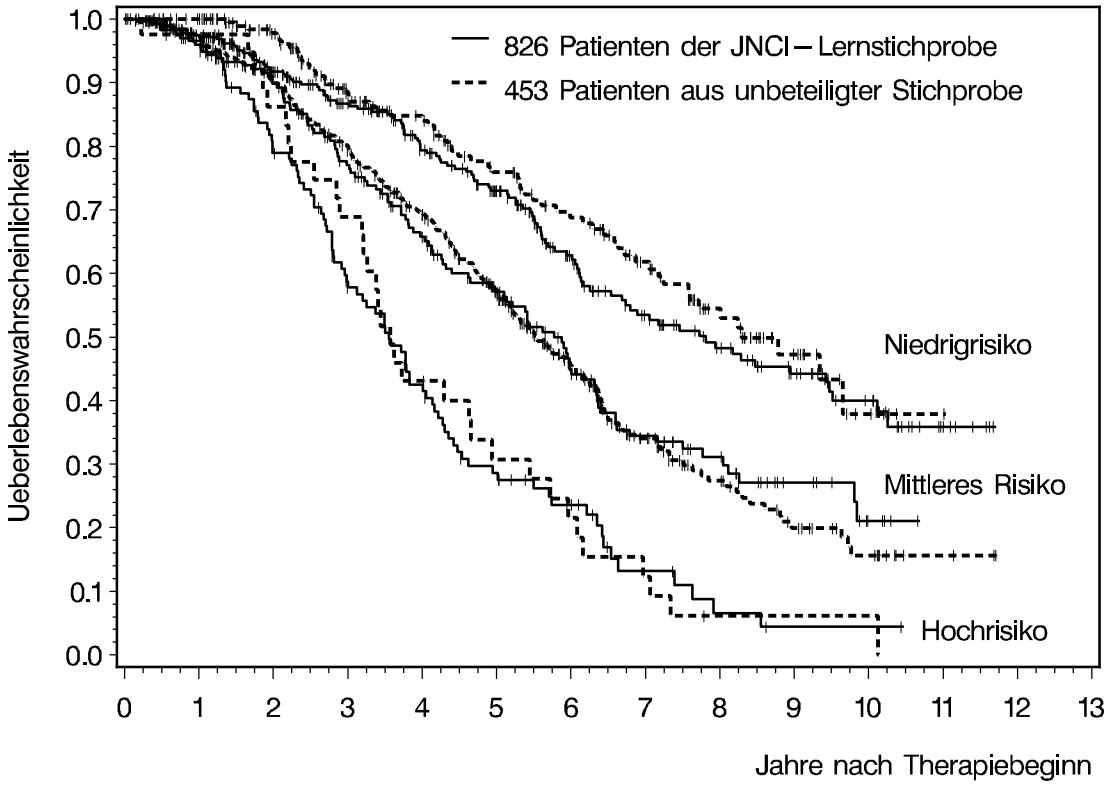

Abbildung 3.2: Kaplan-Meier-Kurven zur Schätzung der Überlebenswahrscheinlichkeiten in Abhängigkeit von der Risikogruppe des New CML-Scores und von der Stichprobenbeteiligung an seiner Entwicklung. In der JNCI-Lernstichprobe gehörten 41,0% der Patienten zur Niedrigrisikogruppe (n = 339, 110 verstorben, mediane Überlebenszeit: 94 Monate), 44,7% zur mittleren Risikogruppe (n = 369, 215 verstorben, mediane Überlebenszeit: 67 Monate) und 14,3% zur Hochrisikogruppe (n = 118, 87 verstorben, mediane Überlebenszeit: 43 Monate). In der unbeteiligten Stichprobe ergab sich folgende Verteilung: 45,5% zur Niedrigrisikogruppe (n = 206, 62 verstorben, mediane Überlebenszeit: 100 Monate), 45,5% zur mittleren Risikogruppe (n = 206, 104 verstorben, mediane Überlebenszeit: 69 Monate) und 9,1% zur Hochrisikogruppe (n = 41, 33 verstorben, mediane Überlebenszeit: 43 Monate).

der Risikogruppen des New CML-Scores.[10]

Ein prospektiver, randomisierter Vergleich zweier IFN-α-Therapieansätze

Die einzige der 11 Studien, bei welcher mit einem prospektiven, randomisierten Design zwei IFN-α-Therapienansätze verglichen wurden, war die Studie von Guilhot et al. aus dem Jahre 1988 [37]. Von der französischen Studie verblieben 196 Patienten in der Analysestichprobe (vgl. Tabelle 3.1). Davon waren 99 Patienten in den IFN-α-Monotherapiearm randomisiert worden (mediane Überlebenszeit 72 Monate, siehe Tabelle 3.3) und 97 in den IFN-α + Ara-C-Kombinationstherapiearm (mediane Überlebenszeit 80 Monate). Die Überlebenswahrscheinlichkeiten der beiden Therapien waren weder unstratifiziert noch risikogruppenstratifiziert sta-

[10] Die Überlebenswahrscheinlichkeiten der JNCI-Lernstichprobe waren niedriger als bei den 453 unbeteiligten Patienten (Logrank-Test: p = 0,0861). Allerdings hatte die JNCI-Lernstichprobe das ungünstigere Risikoprofil (vgl. Legende Abbildung 3.2). Ein nach den drei Risikogruppen stratifizierter Logrank-Test [86] führte zum p-Wert 0,4007 und konnte damit einen großen Teil der unterschiedlichen Überlebenswahrscheinlichkeiten erklären.

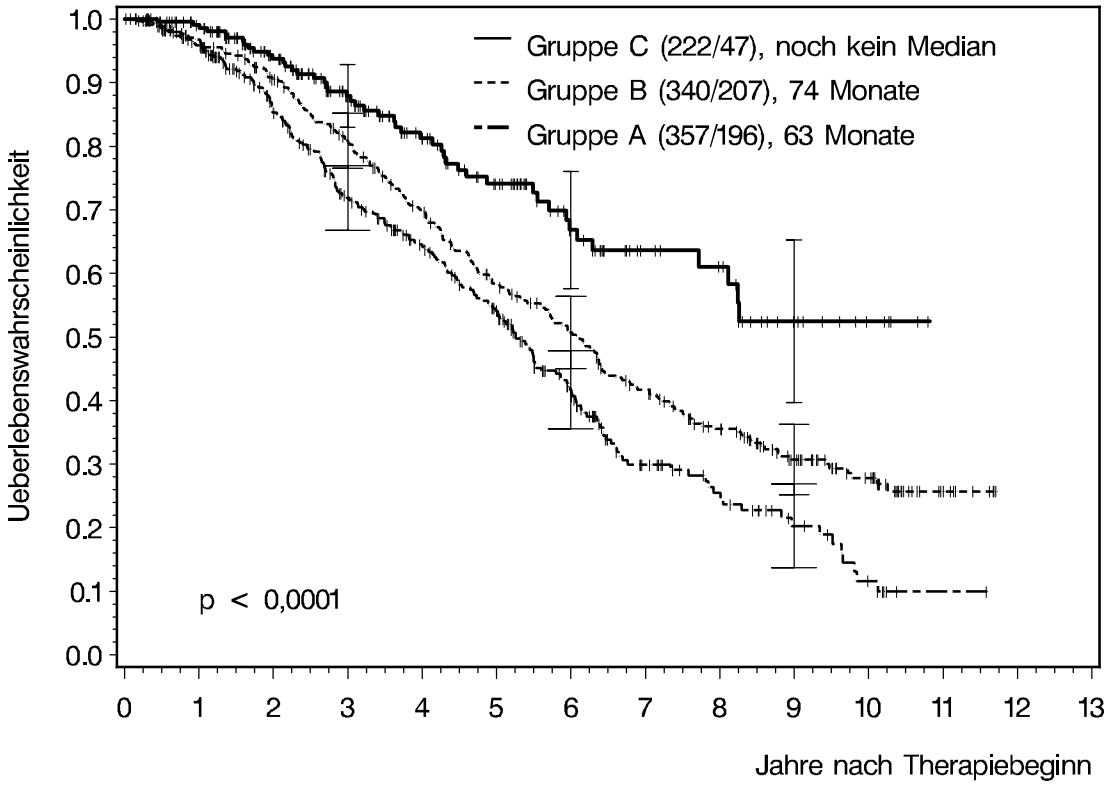

Abbildung 3.3: Einteilung der acht IFN-α-Monotherapiearme in drei Gruppen mit unterschiedlichen Überlebenswahrscheinlichkeiten. Kaplan-Meier Kurven zu den drei Gruppen. Die Legende „(222/47), noch kein Median“ bedeutet: Unter den 222 Patienten mit nach Kaplan-Meier geschätzten Überlebenswahrscheinlichkeiten wurden 47 Todesfälle beobachtet. Die mediane Überlebenszeit wurde noch nicht erreicht. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Überlebenswahrscheinlichkeit mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizentalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Nennung in der Legende von oben nach unten. Der p-Wert ist das Ergebnis des Logrank-Tests beim gemeinsamen Vergleich der drei Kurven.

tistisch signifikant unterschiedlich (p-Werte $> 0{,}5$).

Einteilung nach Therapieansatz und Überlebenswahrscheinlichkeiten

Bevor die 1300 Patienten zu Therapiearmen mit jeweils demselben vorgesehenen IFN-α-Therapieansatz zusammengefügt wurden, wurde die Heterogenität in den Überlebenswahrscheinlichkeiten innerhalb des jeweiligen IFN-α-Therapieansatzes betrachtet. Bei den ersten drei der in Tabelle 3.3 angeführten Studien beobachtete man IFN-α-Monotherapiearme mit einander sehr ähnlichen Überlebenswahrscheinlichkeiten; die Kaplan-Meier-Kurven waren kaum unterscheidbar. Da sich die drei Studien auch weder bzgl. der Risikogruppenverteilungen des New CML-Scores noch bei den entsprechend stratifizierten Logrank-Tests als statistisch signifikant unterschiedlich herausstellten, wurden die drei Studien in einer „Gruppe A“ zusammengefasst. Mit derselben Argumentation ließen sich in Tabelle 3.3 die fünf nachfolgenden Studienarme zu zwei Gruppen „B“ und „C“ zusammenfassen. Abbildung 3.3 zeigt die drei bzgl. der Überlebenswahrscheinlichkeiten statistisch signifikant unterschiedlichen Gruppen, in welche sich die acht IFN-α-Monotherapiearme

Tabelle 3.3: Vergleich der Überlebensdaten zwischen verschiedenen Studien und den vorgesehenen IFN-α-Therapieansätzen

Studie	Anzahl der Patienten	Anzahl der beobachteten Todesfälle	Mediane Beobachtungszeit noch unter Risiko stehender Patienten	Mediane Überlebenszeit	Einteilung nach Therapie und Überlebenswahrscheinlichkeiten
	n	n (%)	Monate (n^a)	Monate	Gruppe
D - CML I [47]	96	61 (64%)	110 (9)	65	IFN-α-
GB - CML III [2]	186	100 (54%)	78 (40)	63	Monotherapie
J - Hamamatsu [80]	75	35 (47%)	63 (31)	66	Gruppe A
A - CML III [115]	45	23 (51%)	101 (18)	80	IFN-α-
F - Poitiers [37]					Monotherapie
IFN-α	99	56 (57%)	94 (34)	72	Gruppe B
I - Bologna [57]	196	128 (65%)	122 (42)	74	
E - Madrid [107]	99	21 (21%)	49 (52)	n.e.[b]	IFN-α-
F - Bordeaux [74]	123	26 (21%)	51 (80)	n.e.	Monotherapie
					Gruppe C
B/NL/LUX [15]	79	49 (62%)	86 (14)	62	IFN-α + HU
D - CML II [48]	178	68 (38%)	64 (63)	68	Gruppe D
A - CML V [116]	27	5 (19%)	54 (22)	n.e.	IFN-α + Ara-C
					Gruppe E
F - Poitiers [37]					IFN-α + Ara-C
IFN-α + Ara-C	97	50 (52%)	92 (32)	80	Gruppe F
Gesamt[c]	1300	622 (48%)	74 (437)	72	

[a]Die Anzahl n der noch unter Risiko stehenden Patienten (der Datenbasis für die Berechnung der medianen Beobachtungszeit) beinhaltete alle noch lebenden Patienten, die aus anderen Gründen als „SZT in 1. chronischer Phase“ zensiert wurden.

[b]Das Kürzel „n.e.“ steht in dieser Tabelle für noch „nicht erreichte“ mediane Überlebenszeit.

[c]Zieht man von den 1300 Patienten die verstorbenen 622 und die noch unter Risiko stehenden 437 ab, verbleiben die 241 (19%), welche eine allogene SZT in erster chronischer Phase erhielten und für die damit kein Ereignis mehr unter IFN-α beobachtet werden konnte. Analog lassen sich aus der Tabelle die zensierten SZT-Patienten auch für die einzelnen Studien(arme) berechnen.

einteilen ließen (gemeinsamer Logrank-Test: $p < 0{,}0001$, paarweise Logrank-Tests: p jeweils $< 0{,}0025$). Außer studienspezifischen Heterogenitäten hatten die zwischen den Gruppen A, B und C statistisch signifikant unterschiedlichen Prognosegruppenverteilungen nach dem New CML-Score (paarweise χ^2-Tests: p jeweils $\leq 0{,}05$) einen Einfluss auf die gruppenspezifischen Überlebenswahrscheinlichkeiten. Gruppe A hatte das ungünstigste Risikogruppenprofil, Gruppe C das Günstigste. Die Überlebenswahrscheinlichkeiten bei nach Risikogruppen stratifizierter Analyse blieben für die Gruppenvergleiche A vs. C und B vs. C weiter statistisch signifikant unterschiedlich, während die Adjustierung für das unterschiedliche Risikogruppenprofil beim Vergleich A vs. B nun zu einem statistisch nicht signifikanten Ergebnis führte.

Von den vier Studien(armen) zu den Kombinationstherapien waren, analog obigen Vorgehens,

nur die zwei IFN-α + HU-Arme [15, 48] zur Gruppe D zusammenfassbar; die beiden Studien(arme) zu IFN-α + Ara-C [37, 116] mussten jeweils in einer eigenen Gruppe belassen werden: bei einander sehr ähnlicher Risikogruppenverteilung nach dem New CML-Score indizierten der unstratifizierte wie der risikogruppenstratifizierte Logrank-Test statistisch signifikant günstigere Überlebenswahrscheinlichkeiten für die österreichischen Patienten ($p = 0{,}0445$ und $p = 0{,}0308$).

Die Kombinationstherapiearme im Vergleich zu den Monotherapiegruppen

Im nächsten Schritt wurde untersucht, wo die drei verschiedenen Kombinationstherapiegruppen bzgl. der beobachteten Überlebenswahrscheinlichkeiten und Risikogruppenverteilungen im Vergleich zu den drei IFN-α-Monotherapie-Gruppen A, B und C einzuordnen waren. Die Überlebenswahrscheinlichkeiten der IFN-α + HU-Gruppe D (257 Patienten, 117 beobachtete Todesfälle, 65 Monate mediane Überlebenszeit) lagen zwischen denen der Gruppen A und B. Auch risikogruppenstratifziert waren keine statistisch signifikanten Unterschiede zwischen A und D oder zwischen B und D festzustellen. Ebenso wie die einander ähnlichen Überlebenswahrscheinlichkeiten der Gruppen C und E waren auch diejenigen der Gruppen B und F miteinander vergleichbar und führten beim stratifizierten Logrank-Test zu keinem statistisch signifikanten Ergebnis.
Zusammenfassend ließ sich feststellen, dass bei den vorliegenden Studien(armen) für keine der Kombinationstherapien Resultate vorlagen, die nicht derjenigen einer der drei großen Patientengruppen (A, B oder C), jeweils bestehend aus zwei oder drei IFN-α-Monotherapiearmen, entsprochen hätten. Die Risikogruppenzugehörigkeit bei Diagnose - nicht die vorgesehene Therapie - unterstrich ihre prognostische Bedeutung für die künftigen Überlebenswahrscheinlichkeiten. Da keine der betrachteten Studien für sich genommen außergewöhnliche Überlebenswahrscheinlichkeiten aufwies, wurden abschließend alle Studienarme zu Therapieansätzen zusammengefasst.

Einteilung der zwölf Studien(arme) in drei IFN-α-Therapieansätze

Die ersten 4-5 Jahre war keine Überlegenheit einer der drei Therapien erkennbar (Abbildung 3.4). Danach zeigte sich eine Schere zwischen den beiden Kombinationstherapien, in deren Mitte die Kaplan-Meier-Kurve der Monotherapie-Gruppe lag. Zwischen den Risikogruppenverteilungen der drei Therapieansätze existierten keine statistisch signifikanten Unterschiede (paarweise χ^2-Tests). Weder unadjustiert noch nach Risikogruppen adjustiert erwiesen sich der gemeinsame Logrank-Test über alle drei Therapieansätze oder die paarweisen Logrank-Tests IFN-α-Monotherapie vs. IFN-α + Ara-C und IFN-α-Monotherapie vs. IFN-α + HU als statistisch signifikant. Dagegen zeitigte der Paarvergleich zwischen den Kombinationstherapien, ob unadjustiert oder adjustiert, eine statistisch signifikante Überlegenheit zugunsten von IFN-α + Ara-C ($p = 0{,}0203$ bzw. $p = 0{,}0105$). Aufgrund des multiplen Testens ohne Adjustierung der Teststatistiken sind auch diese p-Werte als deskriptive, nur explorativ zur Hypothesenentwicklung verwertbare Ergebnisse anzusehen. Therapeutische Schlussfolgerungen verboten sich, weil die Therapievergleiche auf drei Patientengruppen basierten, die nicht innerhalb derselben Studie randomisiert wurden, sondern bei ihrer Rekrutierung völlig unterschiedlichen Selektionsmechanismen unterlagen. Inwieweit dieser Heterogenität durch retrospektiv festgelegte, gemeinsame Ein- und Ausschlusskriterien und den New CML-Score entgegengewirkt werden konnte, ist nicht bezifferbar. Festzuhalten blieb v.a., dass der „ITT-Therapievergleich“ keine außergewöhnlichen Überlebensunterschiede hervorbrachte, die signifikanten p-Werte aber vor der späteren Modellentwicklung ein besonderes Augenmerk auf die mit den vorgesehenen Therapien meist übereinstimmenden, tatsächlich verabreichten Therapien nahelegten.

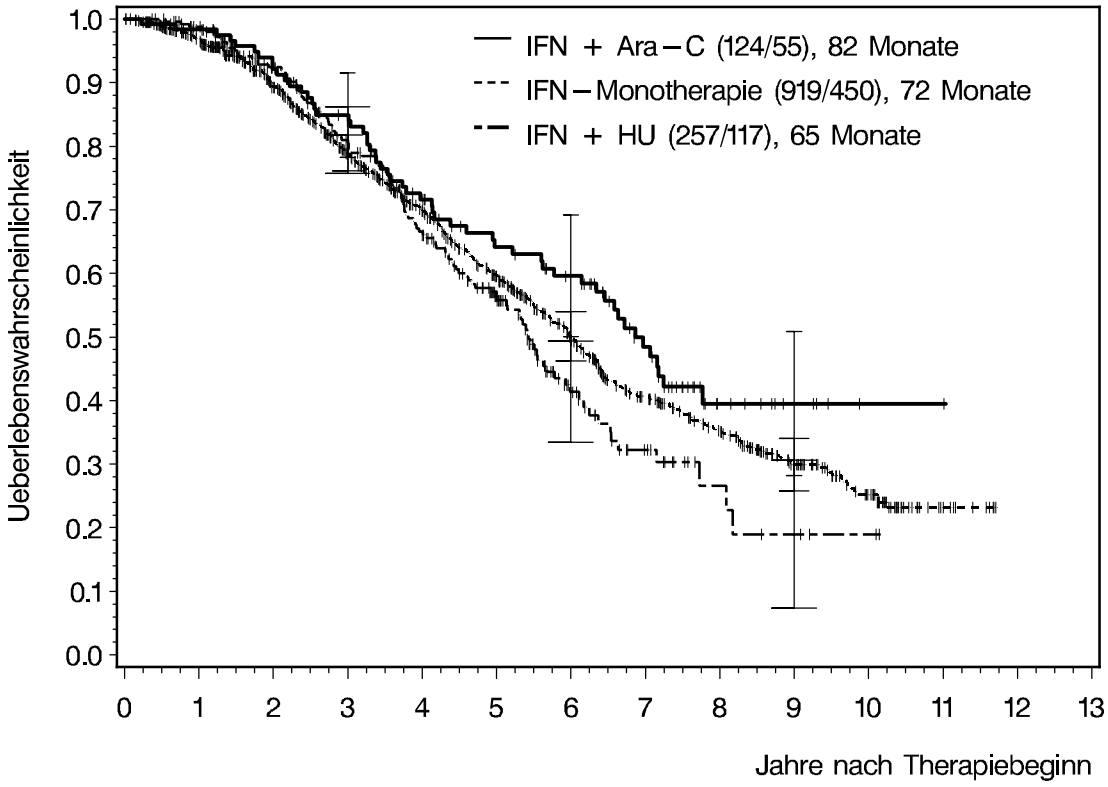

Abbildung 3.4: Kaplan-Meier-Kurven zu den drei IFN-α-Therapieansätzen gemäß der Studienprotokolle zu 1300 Patienten aus zwölf Studien(armen). Die Legende „(124/55), 82 Monate“ bedeutet: Unter den 124 Patienten mit anhand der zugehörigen Kaplan-Meier-Kurve geschätzten Überlebenswahrscheinlichkeiten wurden 55 Todesfälle beobachtet. Die mediane Überlebenszeit betrug 82 Monate. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzten Überlebenswahrscheinlichkeiten mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Nennung in der Legende von oben nach unten.

3.4.4 Die Überlebenszeit in Abhängigkeit vom applizierten Therapieansatz, der Vortherapie und der Zeit zwischen Diagnose und Therapiebeginn

Im Vergleich zu den 1300 Patienten, die entweder für einen IFN-α-Therapieansatz randomisiert wurden oder die im Protokoll vorgesehene Therapie erhielten (vgl. Tabelle 3.3), kamen bei Betrachtung der tatsächlich applizierten Therapieansätze 29 Patienten hinzu und bei 85 hatte sich eine Therapieänderung ergeben. Wenn statt der zur objektiveren Beurteilung gewählten 1300 Patienten mit den vorgesehenen IFN-α-Therapiekombinationen die 1329 Patienten der Analysestichprobe mit ihren tatsächlich verabreichten Therapien gewählt wurden, führten die voranstehend beschriebenen Analysen zu denselben Schlussfolgerungen - allerdings war der Paarvergleich IFN-α + Ara-C vs. IFN-α + HU nur beim risikogruppenstratifizierten Logrank-Test signifikant ($p = 0{,}0181$).

Von den 1329 Patienten der Analysestichprobe (631 verstorben, mediane Überlebenszeit 72 Monate) hatten 327 (24,6%) eine Vortherapie erhalten. Dabei wurde 243 Patienten HU verabreicht (74,3%), 78 Patienten BU (23,9%), 5 Patienten BU + Thioguanine (1,5%) und einem Patienten

HU + BU (0,3%). Es existierten keine unterschiedlichen Überlebenszeiten in Abhängigkeit von der Art der verabreichten Chemotherapie(n). Die 1002 Patienten ohne Vortherapie (464 verstorben, mediane Überlebenszeit 74 Monate) hatten gegenüber den 327 Patienten mit Vortherapie (167 verstorben, mediane Überlebenszeit 64 Monate) einen statistisch signifikanten Überlebensvorteil (Logrank-Test: $p = 0{,}0028$) aber auch eine statistisch signifikant günstigere Risikogruppenverteilung (χ^2-Test: $p = 0{,}0002$). Da der p-Wert bei der risikogruppenstratifizierten Analyse nicht statistisch signifikant war, konnten die unterschiedlichen Überlebenswahrscheinlichkeiten durch die unterschiedliche Risikogruppenverteilung erklärt werden. Das Ergebnis wurde in einem multiplen Cox-Modell bestätigt: Während der New CML-Score einen p-Wert $< 0{,}0001$ erzielte, lag der p-Wert zum Einfluss der Vortherapie bei 0,20. Auch ohne Cox-Modell blieb die Frage, inwieweit die Unterschiede in den Überlebenswahrscheinlichkeiten nicht eher biologischer Heterogenität zuzuschreiben waren: 82% der Patienten mit Vortherapie gehörten zur Benelux-Studie [15] oder zur britischen Studie [2], deren mediane Überlebenszeiten mit 64 bzw. 63 Monaten unter den medianen Überlebenszeiten aller Studien mit Patienten ohne Vortherapie lagen.

Die Zeitspanne zwischen Diagnose und IFN-α-Therapiebeginn war für die 1329 Patienten ohne statistisch signifikanten Einfluss auf die Überlebenszeit (Cox-Modell). Die Spannweite lag zwischen 0 und 182 Tagen, das 1. Quartil bei 2 Tagen, der Median bei 19 und das 3. Quartil bei 69 Tagen.

Da keine die weiteren Analysen einschränkenden Zusammenhänge zwischen der Überlebenszeit und den möglichen Störfaktoren „applizierte Therapiekombination“, „Vortherapie“ und „Zeit zwischen Diagnose und Therapiebeginn“ noch sonstige schwerwiegende Verzerrungen gemäß der in Abschnitt 3.4 untersuchten Gesichtspunkte zu erkennen waren, wurden alle 1329 Patienten für die Identifikation prognostischer Faktoren bzgl. der Überlebenszeit zugelassen.

3.5 Die Daten zur zytogenetischen Remission

3.5.1 Variablendefinition sowie Zusammenhänge zwischen erhobenen Studiendaten, medizinischen und methodischen Aspekten

Mit dem Einschlusskriterium „Ph-positiv“ war kein Patient der Gesamtstichprobe zum Diagnosezeitpunkt frei von der CML-typischen Aberration im Knochenmark. Wenn einzelne Patienten weniger als 96% Ph-positive Zellen in ihrem Knochenmarkaspirat aufwiesen, hatte dies zum Diagnosezeitpunkt keinen Einfluss auf den Beginn einer Therapie mit IFN-α. Zum Baselinezeitpunkt $t = 0$ wurde in den statistischen Modellen für alle Patienten vom Stadium „keine zytogentische Remission“ ausgegangen.

Wegen der als sehr hoch eingestuften prognostischen Bedeutung für die Überlebenszeit und weil sich eine erste komplette ZR manchmal erst nach Jahren einstellt[11], wurden die Studienleiter gebeten, die Ergebnisse aller ab Therapiebeginn vorgenommenen zytogenetischen Untersuchungen bereitzustellen.
Dass sich die Qualität der zytogenetischen Remission aus dem Anteil der Ph-positiven Metaphasen an n ausgezählten Metaphasen errechnet, barg methodische Probleme. Der Stichprobenumfang n lag häufig unter zehn, was für den Anteilsschätzer $\hat{p}$ zu einer großen Standardabweichung und damit möglicherweise ungenauen Schätzergebnissen führte. Mit dem Ziel, bei der Analyse prognostischer Faktoren eine gewisse „Verlässlichkeit“ der verwendeten zytogenetischen Daten

[11] Siehe z.B. bei der italienischen Studiengruppe [58].

garantieren zu können, wurden nur Ergebnisse benutzt, die auf der Basis von mindestens 20 ausgezählten Metaphasen beruhten. Mit $n = 20$ hat man bei einem geschätzten Anteil von $\hat{p} = 0\%$ Ph-positiven Metaphasen (komplette ZR) für den wahren Anteil P das 95%-K.I. [0%;17%] und bei $\hat{p} = 35\%$, der Obergrenze für eine partielle ZR, das 95%-K.I. [15%;58%], womit das Intervall weit in die Kategorie „geringe Remission“ hineinreicht. Legt man $n = 10$ als Mindeststichprobe zugrunde, dehnen sich die entsprechenden 95%-K.I. auf [0%;31%] und [9%;70%] aus, bei $\hat{p} = 35\%$ nun sogar bis in die Kategorie „minimale Remission“.[12]

Die Verlaufsdaten zur zytogenetischen Remission wurden nicht vollständig und exakt zum vorgesehenen Zeitpunkt erhoben. Für die mangelnde Protokolltreue gab es nicht zuletzt medizinische Gründe. Die Entnahme von Knochenmark bedeutet für den Patienten eine weitaus höhere Belastung als die Abnahme einer Blutprobe. Vor allem bei den bis Mitte der 80er Jahre in die Studien aufgenommenen Patienten wurde daher die Berechnung der zytogenetischen Remission noch nicht regelmäßig durchgeführt. Der Bedeutsamkeit der Verminderung der Ph-positiven Metaphasen durch IFN-α bei einer beachtenswerten Zahl von Patienten wurde man mit ihren Konsequenzen für die Überlebenszeit erst nach und nach gewahr. Mit dem Wissen um die Wichtigkeit zytogenetischer Evaluationen erhöhte sich im Laufe der Zeit die Bereitschaft zur Gewinnung eines Knochenmarkaspirates. Die Zahl der hier auswertbaren Datensätze wurde durch die von IFN-α verursachte Hemmung der für die Bestimmung der zytogenetischen Remission notwendigen Proliferation zusätzlich reduziert.

Ereignis und zulässiger Beobachtungszeitraum für die Variable „zytogenetische Remission“

Entsprechend der häufig üblichen Dichotomisierung sowie zur Förderung von Übersichtlichkeit und Verständlichkeit, wurde zur Datenqualitätsprüfung in Kapitel 3 nur zwischen „deutlicher ZR“ ($\leq$ 35% Ph-positive Metaphasen) und „keine deutliche ZR“ unterschieden. Soweit nicht ausdrücklich anders definiert, vergleichen daher im folgenden die im Zusammenhang mit ZR verwendeten Ausdrücke „höhere Remissionswahrscheinlichkeit“ oder „besseres Remissionsergebnis“ den Anteil an Beobachtungen des Ereignisses „erste deutliche zytogenetische Remission“ zwischen verschiedenen Patientengruppen. Die Ereigniszeit für die erste deutliche ZR wurde in Tagen ab IFN-α-Therapiebeginn gemessen. Keine Chemotherapie-Vorbehandlung hatte zu einer partiellen oder kompletten Remission geführt. Der zulässige Beobachtungszeitraum für die Variable „zytogenetische Remission“ endete mit dem Datum der letzten IFN-α-Gabe plus einer Nachwirkungszeit von 45 Tagen, soweit nicht bereits vor Ende dieser 45 Tage ein den zulässigen Beobachtungszeitraum weiter einschränkendes Ereignis festgestellt wurde: Ende der chronischen Phase, SZT oder kürzere Beobachtung der Überlebenszeit. Bei Patienten ohne deutliche ZR während der Therapiedauer mit IFN-α erfolgte die Zensierung der Remissionszeitvariablen mit dem Ende des zulässigen Beobachtungszeitraumes.

3.5.2 Verzerrungen und Störparameter innerhalb der einzelnen Studien

Die möglichen Verzerrungen bei der zytogenetischen Remissionsvariablen konnten nicht alle auf einmal überprüft werden. Wie schon beim Zielparameter Überlebenszeit, wurden bei jeder Reduzierung der Analysestichprobe die früher vorgenommenen Überprüfungen wiederholt.

Von den 209 italienischen Patienten [57] waren zwar ein für die Analyse der Baselinevariablen

[12] Die Konfidenzintervalle wurden nach Koller [68] berechnet.

wichtiges Update der Überlebenszeiten und alle Remissionsergebnisse, mangels ausreichender Dokumentation jedoch nicht die Stichprobenumfänge der ausgezählten Metaphasen verfügbar. Um dem oben vorgestellten „Verlässlichkeitsstandard“ gerecht werden zu können, wurden für die Analysestichprobe zur zytogenetischen Remission nur die zehn Studien mit Angaben zur ausgezählten Metaphasenanzahl zugelassen. Von den verbliebenen 1175 Patienten hatten während des zulässigen Beobachtungszeitraums 819 (70%) mindestens eine zytogenetische Evaluation auf Basis von ≥ 20 Metaphasen. Tabelle 3.4 bietet einen Überblick zu diesen 819 Patienten.

Einflüsse auf die zytogenetische Remission durch die Art des Therapieansatzes mit IFN-α, die Vortherapie und die Zeit zwischen Diagnosedatum und Therapiebeginn mit IFN-α

Da keine konfirmatorische Ablehnung der Hypothese „vergleichbare zytogenetische Remission trotz unterschiedlicher Therapieansätze“ vorlag, wurde, im Gegensatz zum Parameter „Überlebenszeit“, umgehend die Wirkung der tatsächlich angewandten Therapie betrachtet.
Bei den drei Studien aus Poitiers [37], Bordeaux [74] und Madrid [107] kamen unterschiedliche Therapieansätze mit IFN-α zur Anwendung (vgl. Tabelle 3.4). Statistisch signifikant unterschiedliche Ergebnisse zur Wahrscheinlichkeit einer ersten deutlichen ZR waren bei den französischen Studien nicht feststellbar (Logrank-Test).[13] Die acht mit IFN-α + HU behandelten spanischen Patienten schnitten in allen Logrank-Tests zu den Remissionsergebnissen statistisch signifikant schlechter ab als die Patienten mit IFN-α-Monotherapie, egal ob der Vergleich nur mit den 18, die ebenfalls ohne Vortherapie geblieben waren ($p = 0{,}0095$) oder mit allen 40 Monotherapie-Patienten ($p = 0{,}0427$) vorgenommen wurde. Weil aber die zusätzliche Administration von HU bei den acht „ITT“-Monotherapie-Patienten erst im Therapieverlauf stattfand und als mögliche Reaktion auf unbefriedigende Remissionsergebnisse gelten konnte, wurden die acht spanischen Patienten nicht von der Analysestichprobe ausgeschlossen.

Von den fünf Studien mit Chemo-Vortherapien, Großbritannien [2], Benelux [15], Bordeaux [74], Madrid [107] und die österreichische Studie CML V [116] (vgl. Tabelle 3.4), resultierte der Logrank-Test nur bei letzterer in statistisch signifikant unterschiedlichen Remissionswahrscheinlichkeiten ($p = 0{,}0246$). Demnach zeigte sich die HU-Vortherapie bei den 16 Patienten der CML V für das Erreichen einer ersten deutlichen ZR von Nachteil. Wie schon zuvor für die Analysen zur Überlebenszeit, wurden die österreichischen Patienten mit HU-Vortherapie auch von den weiteren Analysen zur zytogenetischen Remission ausgeschlossen.

Für die Zeit zwischen Diagnosedatum und Therapiebeginn mit IFN-α konnte innerhalb der einzelnen Studien kein medizinisch relevanter statistisch signifikanter Einfluss auf das zytogenetische Remissionsergebnis festgestellt werden.

3.5.3 Untersuchung der Konsequenzen aus der Minimalforderung nach 20 ausgezählten Metaphasen

Wäre man von der Forderung nach mindestens 20 ausgezählten Metaphasen abgerückt, wären, abzüglich der 16 Österreicher, nicht 803 sondern 907 Patienten nach den bisherigen Überprüfungen Teil der Analysestichprobe gewesen. Mithin stellte sich die Frage, ob die „Qualitätsgarantie“ mittels der Forderung einer minimalen Metaphasenanzahl nicht durch eine Verzerrung wegen

[13] Auch der Vergleich der vorgesehenen Studienarme IFN-α + Ara-C und IFN-α-Monotherapie nach dem ITT-Prinzip führte bei den 154 Patienten aus Poitiers zu keinem Trend zugunsten einer der Therapieansätze.

Tabelle 3.4: Vergleich des besten zytogenetischen Remissionsergebnisses zwischen verschiedenen Studien und tatsächlich verabreichten IFN-α-Therapien

Studie	Patienten-zahl	Patienten mit DZR[a]		Patienten ohne DZR	Patienten mit 1. DZR bis Ende Monat 18	Einteilung nach IFN-α-Therapie
		Patienten mit KZR[b]	Patienten mit PZR[c]			
	n	n (%)	n (%)	n (%)	%[d] (n[e])	
A - CML III [115]	24	3 (13%)	3 (13%)	18 (75%)	22% (0)	IFN-α-
D - CML I [47]	22	2 (9%)	0 (0%)	20 (91%)	5% (2)	Mono-
E - Madrid [107]						therapie
Vortherapie: HU[f]	22	4 (18%)	3 (14%)	15 (68%)	32% (3)	
E - Madrid [107]						
keine Vortherapie	18	8 (44%)	4 (22%)	6 (33%)	46% (1)	
F - Bordeaux [74]	75	25 (33%)	9 (12%)	41 (55%)	44% (11)	
F - Poitiers [37]	86	11 (13%)	16 (19%)	59 (69%)	35% (1)	
GB - CML III [2]						
Vortherapie: BU[g]	78	8 (10%)	7 (9%)	63 (81%)	17% (5)	
GB - CML III [2]						
Vortherapie: HU	94	11 (12%)	8 (9%)	75 (80%)	17% (5)	
J - Hamamatsu [80]	56	6 (11%)	6 (11%)	44 (79%)	26% (4)	
B/NL/LUX [15]	68	5 (7%)	8 (12%)	55 (81%)	13% (2)	IFN-α
D - CML II [48]	122	9 (7%)	20 (16%)	93 (76%)	16% (29)	+ HU
E - Madrid [107]						
keine Vortherapie	8	0 (0%)	0 (0%)	8 (100%)	0% (2)	
F - Bordeaux [74]	40	12 (30%)	5 (13%)	23 (58%)	43% (2)	
F - Poitiers [37]	2	0 (0%)	0 (0%)	2 (100%)	0% (0)	
A - CML V						IFN-α
keine Vortherapie	14	5 (36%)	4 (29%)	5 (36%)	59% (0)	+ Ara-C
A - CML V [116]						
Vortherapie: HU	16	1 (6%)	2 (13%)	13 (81%)	18% (0)	
F - Bordeaux [74]	8	1 (13%)	1 (13%)	6 (75%)	36% (1)	
F - Poitiers [37]	66	17 (26%)	11 (17%)	38 (58%)	41% (1)	
Gesamt[h]	819	128 (16%)	107 (13%)	584 (71%)	26% (69)	

[a] Deutliche zytogenetische Remission: $\leq$35% Ph-positive Metaphasen.
[b] Komplette zytogenetische Remission: 0% Ph-positive Metaphasen.
[c] Partielle zytogenetische Remission: 1-35% Ph-positive Metaphasen.
[d] Anteile geschätzt nach der Kaplan-Meier-Methode [63]
[e] Anzahl bis zuletzt mit IFN-α behandelter Patienten, für die bis Ende Monat 18 keine DZR vorlag.
[f] Ein Patient ohne DZR erhielt BU statt HU.
[g] Fünf Patienten, einer davon mit DZR, erhielten zusätzlich Thioguanine.
[h] In 356 Fällen (30% von 1175) waren keine Daten zur zytogenetischen Remission verfügbar.

„Verpassens“ deutlicher Remissionen, die aber nur anhand von 8 oder 12 Metaphasen festgestellt worden waren, entwertet wurde. Ohne die „Metaphasenzahlrestriktion“ war nicht nur die Gesamtpatientenzahl höher, sondern waren auch einige Zeitdauern bis zu einer ersten deutlichen

ZR kürzer. Wegen zu kleiner Fallzahlen waren keine aufschlussreichen Erkenntnisse durch die Daten einzelner Studien zu gewinnen. Mehr Aussagekraft versprach die studienunabhängige, gemeinsame Untersuchung der beiden 803 bzw. 907 Patienten umfassenden Stichproben, wobei das Spektrum der Remissionsergebnisse wieder dichotomisiert betrachtet wurde. Von den 907 Patienten hatten - ohne Metaphasenzahlrestriktion - 301 (33%) eine deutliche ZR, mit der Metaphasenzahlrestriktion verblieben 232 (29% von 803) Patienten. Bei 203 Patienten (67% von 301, Fall A) führte das Vorliegen von ≥ 20 Metaphasen in beiden Stichproben zum identischen Ergebnis für die erste deutliche Remission. Beschränkt auf Evaluationen mit ≥ 20 ausgewerteten Metaphasen, lag bei weiteren 29 Patienten (10% von 301, Fall B) die erste deutliche Remission zeitlich nach derjenigen auf Basis einer beliebiger Metaphasenzahl und hatten 45 andere (15%, Fall C) im gesamten Therapieverlauf keine deutliche ZR, während ohne Restriktion mindestens eine deutliche ZR registriert werden konnte. Von den 104 Patienten, deren Zytogenetikergebnisse sich ausschließlich auf < 20 Metaphasen stützten, hatten 24 Patienten eine deutliche Remission (8% von 301 bzw. 23% von 104, Fall D).

Mit dem Ziel, die Hypothese H_0 „Die Verlässlichkeit des Ergebnisses „deutliche Remission“ kann als von der ausgewerteten Metaphasenanzahl ≥ 20 oder < 20 unabhängig betrachtet werden“ einer Prüfung zu unterziehen, wurde „Rezidiv nach deutlicher Remission“ als zweites Ereignis zugelassen. Um durch die unterschiedliche Evaluationshäufigkeit verzerrte Resultate zu vermeiden, wurde wie folgt vorgegangen: Qualifiziert für den entscheidenden χ^2-Test waren alle 250 Patienten, zu welchen, nach ihrer ersten deutlichen ZR auf Basis einer - entsprechend der Nullhypothese H_0 - beliebigen Metaphasenzahl, mindestens eine weitere zytogenetische Untersuchung vorlag. Ein Patient wurde als „Rezidivfall“ betrachtet, wenn er gemäß der direkt auf die erste deutliche Remission folgenden Chromosomenanalyse seine deutliche ZR verloren hatte. Von den 250 Patienten gehörten 177 (71%) zu Fall A, der Gruppe mit ≥ 20 ausgewerteten Metaphasen bereits bei der ersten deutlichen Remission. Unter Fall C und D verblieben 35 bzw. 9 Patienten für die Gruppe „< 20 ausgezählte Metaphasen bei der ersten (und allen folgenden) deutlichen ZR“. Auch die 29 Patienten aus Fall B wurden letzterer Gruppe zugeschlagen, da trotz späterer deutlicher ZR auf Basis ≥ 20 Metaphasen, ihre erste deutliche ZR anhand < 20 Metaphasen evaluiert worden war.[14] Im Fall A musste bei 46 der 177 Patienten (26%) das zweite zytogenetische Ergebnis als Rezidiv gewertet werden, unter B, C und D trat diese Situation 6mal (bei 21% von 29 Patienten), 24mal (69% von 35) und 3mal (33% von 9) auf. Zusammengefasst ergab dies für die Gruppe „< 20 Metaphasen“ 33 Rezidive bei 73 Patienten (45%). Der χ^2-Test zum Vergleich der 177 versus der 73 Patienten hinsichtlich der Rezidivhäufigkeit führte zum p-Wert 0,0030 und damit zur Ablehnung der Nullhypothese.[15] Die deutlichen ZR, welche auf Basis der kleineren Metaphasenzahlen identifiziert worden waren, hatten bei der nachfolgenden Analyse statistisch signifikant häufiger ein Rezidiv gezeigt, was die Vermutung nahelegte, dass hier einige der „erkannten“ deutlichen ZR in Wirklichkeit auf einer Fehleinschätzung beruhten und dass der Anteil dieser Fehleinschätzungen höher war als bei den Analysen auf der Basis von mindestens 20 Metaphasen. Damit sprachen auch die empirischen Ergebnisse dafür, zugunsten eines sich u.a. auf die Datenqualität der zytogenetischen Evaluationen stützenden Prognosemodells auf Zytogenetikdaten zu verzichten, welchen weniger als 20 ausgezählte Metaphasen zugrundela-

[14] Das Verfahren mit den Patienten von Fall B führte zu einer Verzerrung zugunsten der Gruppe „< 20 Metaphasen“, da die Bestätigung der deutlichen ZR jeweils aus dem Ergebnis der direkt nachfolgenden Chromosomenanalyse bestand.

[15] Verglich man, nun Verzerrungen durch die Evaluationshäufigkeit hinnehmend, statt der Rezidivhäufigkeit bei der Folgeuntersuchung, die Anzahl der Patienten, die a) irgendwann oder b) zuletzt ein Rezidiv hatten, lagen die p-Werte ebenfalls unter 0,01. Noch niedriger waren alle p-Werte, ließ man die 29 Patienten von Fall B außen vor oder ordnete man sie der Gruppe „≥ 20 Metaphasen“ zu.

gen. Dieser Verzicht versprach die Vermeidung größerer Verzerrungen als durch ein „Verpassen“ deutlicher Remissionen entstehen konnten.

Wenn im folgenden vom Vorliegen zytogenetischer Daten gesprochen wird, so bezieht sich dies immer auf Evaluationen anhand ≥ 20 Metaphasen. Nur in Abschnitt 3.5.5 wurde von dieser Regel noch einmal abgewichen.

3.5.4 Überprüfung der „relativen“ Plausibilität der jeweiligen Studiendaten

Soweit möglich, galt es die Remissionsdaten jeder Studie auf Plausibiltiät und Eignung für eine Prognosemodellentwicklung zu prüfen. Mangels besserer Alternative wurden als Referenz für eine Studie die Daten der anderen Studien zu Rate gezogen. Durch die komplett verfügbaren Zytogenetikdaten aller an der Analysestichprobe beteiligten Studien waren wesentlich umfangreichere Untersuchungen möglich als simple Vergleiche mit den wenigen Angaben, die Veröffentlichungen üblicherweise zu entnehmen sind.

Die Anzahl zytogenetischer Evaluationen und deutlicher Remissionen in Abhängigkeit von Studie und Zeitraum nach Therapiebeginn

Zu den 803 Patienten standen die Daten von 2905 Chromosomenanalysen während der zulässigen Beobachtungszeiträume zur Verfügung. Zur Beantwortung der Frage, wieviele Evaluationen auf bestimmte Zeiträume nach Therapiebeginn fielen, mussten diese zunächst definiert werden. In Übereinstimmung mit Berechnungshäufigkeit und Bedeutung, erfolgte die ersten beiden Therapiejahre eine vierteljährliche Einteilung, im dritten noch eine halbjährliche und ab dem vierten Jahr eine jährliche. Nach dem zehnten Jahr unter IFN-α-Therapie lagen keine Ergebnisse mehr vor.[16] In 90% der Fälle entfiel auf einen Zeitraum höchstens eine Evaluation pro Patient. Waren es mehrere, so wurde, unter Annahme einer zufälligen Auswahl und nur für die Zwecke dieses Abschnitts 3.5.4, lediglich die letzte berücksichtigt. Damit lagen zur weiteren Untersuchung 2591 Evaluationen vor (vgl. Tabelle 3.5), die sich alle entweder im Hinblick auf den Patienten oder den Zeitraum unterschieden. Für jeden Zeitraum j und jede Studie s wurden Vierfeldertafeln erstellt, im Rahmen derer die Anzahl der in der Studie evaluierten Patienten n^e_{js} und die Anzahl der Evaluierbaren aber Nichtevaluierten n^{ne}_{js} mit den entsprechenden Zahlen bestehend aus der Summe über alle übrigen Studien, z.B. zum Zeitraum „1. Quartal, 1. Jahr“ $173 - n^e_{js}$ mit $946 - n^{ne}_{js}$ (siehe Tabelle 3.5), verglichen wurden. Dabei ergaben statistisch signifikante p-Werte zugehöriger χ^2-Tests bzw. exakter Fisher-Tests, dass, jeweils im Vergleich zu allen anderen Studien, in über der Hälfte aller Zeiträume bei den Studien D - CML I [47], D - CML II [48] und jener aus Madrid [107] unterdurchschnittlich viele der evaluierbaren Patienten tatsächlich evaluiert wurden, während es umgekehrt, bei den Studien GB - CML III [2] und Bordeaux [74] überdurchschnittlich viele waren.

Nach demselben Prinzip wurden, je Zeitraum und Studie, die Anzahl der Patienten mit und ohne deutliche ZR den beiden korrespondierenden Summen dieser Zahlen aus alle übrigen Studien gegenüber gestellt. Statistisch signifikant unter dem Durchschnitt aller anderen Studien lagen die Patientenzahlen mit deutlicher Remission in den Studien GB - CML III [2] und D - CML II [48] zu je fünf Zeiträumen und in der Studie aus B/NL/LUX [15] zu zwei Zeiträumen. Eine

[16] Das Jahr J nach Therapiebeginn endete nach auf die volle Tageszahl gerundeten $J \times 365{,}25$ Tagen, das Schaltjahr fiel also auf das 2., 6. und 10. Jahr. Das v-te Vierteljahr innerhalb eines Jahres J endete mit der gerundeten Tageszahl aus v $\times$ 3 $\times$ 30,4 Tagen. In den Schaltjahren kam im letzten Quartal ein Tag hinzu.

Tabelle 3.5: Die Anzahl deutlicher zytogenetischer Remissionen sowie der Evaluationen pro Zeitraum

Zeitraum	Patienten mit DZR[a]	Patienten ohne DZR	Im Zeitraum evaluierte Patienten[b]	Nicht im Zeitraum evaluierte Patienten	Im Zeitraum evaluierbare Patienten[c]
	n^{dzr} (% von n^e)	n^{odzr} (% von n^e)	n^e (% von n^{alle})	n^{ne} (% von n^{alle})	n^{alle}
1. Quartal, 1. Jahr	9 (5%)	164 (95%)	173 (15%)	946 (85%)	1119
2. Quartal, 1. Jahr	37 (10%)	318 (90%)	355 (35%)	669 (65%)	1024
3. Quartal, 1. Jahr	62 (20%)	243 (80%)	305 (33%)	616 (67%)	921
4. Quartal, 1. Jahr	59 (23%)	196 (77%)	255 (30%)	590 (70%)	845
1. Quartal, 2. Jahr	66 (29%)	162 (71%)	228 (30%)	533 (70%)	761
2. Quartal, 2. Jahr	60 (33%)	120 (67%)	180 (27%)	498 (73%)	678
3. Quartal, 2. Jahr	57 (34%)	112 (66%)	169 (27%)	456 (73%)	625
4. Quartal, 2. Jahr	53 (39%)	83 (61%)	136 (24%)	440 (76%)	576
1. Hälfte, 3. Jahr	74 (33%)	149 (67%)	223 (45%)	274 (55%)	497
2. Hälfte, 3. Jahr	59 (36%)	106 (64%)	165 (40%)	251 (60%)	416
4. Jahr	62 (36%)	108 (64%)	170 (52%)	159 (48%)	329
5. Jahr	39 (38%)	63 (62%)	102 (43%)	133 (57%)	235
6. Jahr	27 (39%)	42 (61%)	69 (46%)	81 (54%)	150
7. Jahr	19 (63%)	11 (37%)	30 (35%)	55 (65%)	85
8. Jahr	11 (61%)	7 (39%)	18 (41%)	26 (59%)	44
9. Jahr	4 (40%)	6 (60%)	10 (43%)	13 (57%)	23
10. Jahr	3 (100%)	0 (0%)	3 (25%)	9 (75%)	12
Gesamt	701 (27%)	1890 (73%)	2591 (31%)	5749 (69%)	8340

[a]Deutliche zytogenetische Remission: $\leq$35% Ph-positive Metaphasen.

[b]Bei mehreren Evaluationen pro Patient in einem Zeitraum wurde das Ergebnis der letzten (aktuellsten) Evaluation berücksichtigt.

[c]Der Beitrag eines Patienten i zu n^{alle} errechnete sich für den interessierenden Zeitraum j aus dem Quotienten (Anzahl der von i in j verbrachten Tage) / (Anzahl der Tage in j). Nach Therapiebeginn standen alle 1133 für die Analysestichprobe qualifizierten Patienten mindestens einen Tag unter Beobachtung.

statistisch signifikant überdurchschnittlich hohe Zahl deutlicher Remissionen zeigte sich in den Studien aus Poitiers [37], Bordeaux [74] und Madrid [107], zu vier, sieben und zwei Zeiträumen. Damit lagen, sowohl im Falle ungewöhnlich hoher wie auch ungewöhnlich niedriger Anteile an beobachteten deutlicher ZR jeweils Studien vor, in welchen im Vergleich zu den anderen Studien a) überwiegend statistisch signifikant weniger Patienten evaluiert worden waren, b) zu fast allen Zeiträumen keine statistisch signifikant unterschiedlichen Berechnungshäufigkeiten vorlagen und c) statistisch signifikant mehr Zytogenetikdaten gegeben waren.

Gestützt auf die Ergebnisse dieser Häufigkeitsvergleiche wurde gefolgert, dass in der Analysestichprobe insgesamt von keinem systematischen Zusammenhang zwischen der Anzahl deutlicher Remissionen und der Anzahl durchgeführter Evaluationen ausgegangen werden musste. Anders formuliert, schienen die Ärzte - gemäß der betrachteten 2591 Chromosomenanalysen - nicht in

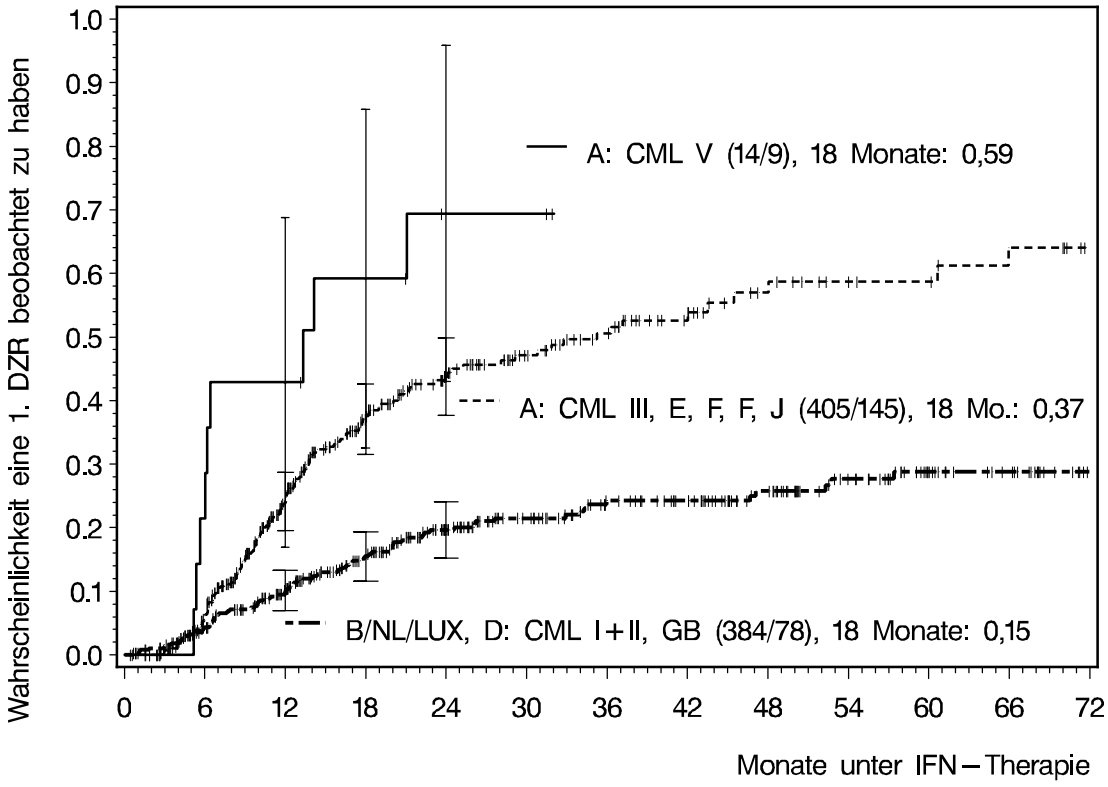

Abbildung 3.5: Kaplan-Meier-Kurven mit geschätzten Wahrscheinlichkeiten bei 803 Patienten in Abhängigkeit ihrer Studienherkunft eine erste deutliche zytogenetische Remission (DZR) beobachtet zu haben. Die Legende „B/NL/LUX, D - CML I, D - CML II, GB (384/78): 0,15“ beschreibt mit B/NL/LUX, D - CML I, D - CML II, GB Länderkennzeichen und Studien der in einer Gruppe zusammengefassten Patienten, vgl. Tabelle 3.4. Die Zahlen „(384/78): 0,15“ bedeuten: Unter 384 Patienten mit anhand der zugehörigen Kaplan-Meier-Kurve beschriebenen Wahrscheinlichkeiten wurden 78 DZR beobachtet. Die geschätzte Wahrscheinlichkeit, bis Ende des 18. Monats eine DZR beobachtet zu haben, lag bei 0,15. Zu den Zeitpunkten 12, 18 und 24 Monate wurden um die geschätzte Wahrscheinlichkeit mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizentalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Nennung in der Legende von oben nach unten.

Erwartung eines bestimmten Remissionsergebnisses eine Knochenmarkaspiration vorgenommen oder darauf verzichtet zu haben.

Wahrscheinlichkeiten nach bestimmter Therapiedauer eine erste deutliche ZR beobachtet zu haben

Mit der Modellannahme, dass sich ein Patient ab Therapiebeginn bis zum Erhalt eines gegenteiligen Ergebnisses in „keiner deutlichen Remission“ befand, beeinflussten die unterschiedlichen Evaluationshäufigkeiten das Ausmaß einer verzögerten Feststellung der ersten deutlichen Remission und in Einzelfällen vielleicht auch deren Verpassen. Abbildung 3.5 zeigt drei Patientengruppen, bei welchen fünf paarweise Logrank-Tests, zwei nach dem New CML-Score stratifizierte und die drei unstratifizierten, jeweils mit $p < 0{,}0500$ auf statistisch signifikant unterschiedliche

Wahrscheinlichkeiten hinwiesen, eine erste deutliche ZR beobachtet zu haben.[17] Nur der risikogruppenstratifizierte Vergleich der beiden Gruppen mit höheren Remissionswahrscheinlichkeiten war nicht statistisch signifikant. Wie in Abschnitt 3.4.3, waren zunächst Studien mit vergleichbaren Wahrscheinlichkeiten zusammengefasst worden.[18] Bei der Gruppenbildung in Abbildung 3.5 ergab sich eine hohe Übereinstimmung mit den voranstehend unterschiedenen Gruppen. So gehörten wiederum zu zwei großen Gruppen mit unterschiedlich günstigen Remissionsergebnissen Studien mit sehr verschiedenen Häufigkeiten einer Chromosomenanalyse. Da sich die 14 Patienten mit den besten Remissionsergebnissen nur beim unstratifizierten Test von der zweitbesten Gruppe unterschieden und deren Konfidenzintervalle nahezu komplett in den großen Konfidenzintervallen der kleinen Gruppe enthalten waren, wurden die 14 österreichischen Patienten für die weiteren Analysen beibehalten.

3.5.5 Zusammenhänge zwischen Therapieverlauf, der zytogenetischen Remission und der Überlebenszeit

Zusammenhang zwischen fehlenden Verlaufsdaten und der Überlebenszeit

Nach Ausschluss der 42 vorbehandelten Patienten der Studie A - CML V waren 1133 Patienten für die Analysestichprobe qualifiziert. In 330 Fällen (29%) lagen jedoch auf Basis von $\geq$ 20 Metaphasen keine Daten zur Zytogenetik im Therapieverlauf vor. Zu 104 der 330 Patienten waren zytogentische Verlaufsdaten übermittelt worden, doch lag die ausgewertete Metaphasenzahl immer unter 20. Um bezüglich der möglichen Ursachen für ein wirkliches Fehlen von Zytogenetikdaten keinen durch die selbst eingeführte Metaphasenzahlrestriktion verfälschten Ergebnissen aufzusitzen, musste dieses Qualitätskriterium hier noch einmal fallengelassen, die 104 den 803 Patienten hinzugefügt werden. Der Logrank-Test zum Überlebenszeitvergleich der 907 Patienten (381 verstorben, mediane Überlebenszeit: 75 Monate) versus der 226 Patienten ohne jegliche Daten (122 verstorben, mediane Überlebenszeit: 50 Monate) lieferte einen statistisch signifikanten p-Wert $< 0{,}0001$.
Die Behandlungszeit mit IFN-α war im Falle der 226 Patienten statistisch signifikant kürzer (U-Test: $p < 0{,}0001$), für 79 Patienten (35%) betrug sie höchstens drei Monate, was bei den 907 lediglich auf 27 Patienten (3%) zutraf. Mit 133 der 226 Patienten ohne Zytogenetikdaten bekamen 59% maximal neun Monate eine IFN-α-Therapie, wobei 112 (84% von 133 und 50% von 226) einen Abbruch wegen Therapieversagens vornehmen mussten. Dem standen von den 907 Patienten 144 Patienten mit maximaler Therapiedauer von einem Dreivierteljahr (16%) und dabei 95 Abbrüchen wegen Therapieversagens (66% von 144 und 10% von 907) gegenüber. Die χ^2-Tests aus den Vierfeldertafeln zum Vergleich von 59% versus 16% respektive 84% versus 66% resultierten in $p < 0{,}0001$ und $p = 0{,}0005$.
Stratifizierte man den Logrank-Test zum Vergleich der Überlebenswahrscheinlichkeiten zwischen den 907 und den 226 Patienten nach „Therapieversagen innerhalb der ersten neun Monate: ja oder nein“, ergab sich der nicht signifikante p-Wert 0,1142.[19]
Die vorliegenden Ergebnisse führten zur der Annahme, dass sich hinter dem Fehlen von jeglichen

[17] Bei 773 (von 803) bzgl. des New CML-Scores berechenbaren Patienten galt statistisch signifikant (Logrank-Test, $p < 0{,}0001$): je niedriger die Risikogruppe des New CML-Scores desto besser das Remissionsergebnis. Wenn die Risikogruppenverteilung zwischen zu vergleichenden Gruppen hinsichtlich einer unterschiedlichen Signifikanz der p-Werte zum unstratifizierten bzw. stratifizierten Logrank-Tests eine Rolle spielte, wurde darauf eingegangen.

[18] Die in Abbildung 3.5 zu einer Gruppe vereinigten Studien sind mit Hilfe des Nationenkennzeichens und Tabelle 3.4 identifizierbar.

[19] Unterschiedliche Risikogruppenverteilungen nach dem New CML-Score spielten bei diesem Vergleich keine statistisch signifikante Rolle.

zytogenetischen Daten ein frühes Therapieversagen als Ursache für die ungünstigen Überlebenswahrscheinlichkeiten der 226 Patienten verbarg. Unter den 803 Patienten mit mindestens 20 analysierten Metaphasen verzeichneten 9 Patienten ihre erste deutliche Remission innerhalb der ersten drei Monate (4% von 232 mit deutlicher ZR). Nach neun Monaten Therapie waren es 88 (38% von 232). Ein Prognosesystem, welches sich maßgeblich auf das zytogenetische Remissionsergebnis stützt, kann dementsprechend nach drei oder neun Monaten IFN-α-Therapie für die Patienten ohne deutliche ZR noch keine abschließende Therapiebeurteilung liefern. Weiteres Warten auf eine deutliche ZR kommt für Patienten mit frühem Therapieversagen nicht in Frage, sie sollten umgehend eine andere Therapie erhalten. Sie stellen damit keine zentrale Zielgruppe des zu entwickelnden Prognosesystems dar, weswegen die Nichtberücksichtigung der 226 Patienten ohne zytogenetische Daten keine wesentliche Beeinträchtigung für die Gültigkeit oder Verlässlichkeit des angestrebten Prognosesystems bedeutete.

Zusammenhänge zwischen Therapieverlauf und dem zytogenetischen Remissionsergebnis

Der in einer Nennung des Abbruchgrundes für IFN-α zusammengefasste Therapieverlauf wurde nun speziell für die Patienten ohne deutliche ZR betrachtet. Von den 803 Patienten der Analysestichprobe hatten 232 (29%) mindestens eine deutliche ZR. Von den restlichen 571 Patienten wurden 380 (67%) solange mit IFN-α behandelt, bis ein Abbruch wegen Therapieversagens beschlossen wurde. Die 571 Patienten ohne deutliche ZR erhielten statistisch signifikant länger IFN-α (Median: 643 Tage) und hatten ihre letzte zytogenetische Evaluation nach Therapiebeginn statistisch signifikant später (Median: 430 Tage), als die für die 232 Patienten bis zur ersten deutlichen Remission beobachteten Zeitdauern (Median: 350 Tage, U-Tests: $p < 0{,}0001$ und $p = 0{,}0037$). Dieselbe Tendenz und vier statistisch signifikante p-Werte $< 0{,}0250$ ergaben sich, wenn statt der 571 Patienten die beiden Untergruppen mit (n=380) und ohne Therapieversagen (n=191) mit den 232 verglichen wurden. Die signifikanten Zeitunterschiede ließen vermuten, dass i.A. auch bei den 571 Patienten ohne deutliche ZR „ernsthafte“ Therapieversuche zum Erreichen einer ZR unternommen worden waren und dass die Nichtfeststellung einer deutlichen ZR tendenziell nicht an einer frühzeitigen Einstellung der zytogenetischen Evaluationen gelegen hatte.
Beim Vergleich der Kategorien der jeweils besten erreichten Remission wurde festgestellt, dass von kompletter ZR (n=127 Patienten) über partielle (n=105), geringe (n=67) und minimale ZR (n=168) bis zu keiner ZR (n=336) der Anteil an Fällen mit Abbrüchen wegen Therapieversagens immer größer wurde und umgekehrt, sowohl die Zeit bis zur letzten zytogenetischen Evaluation als auch die IFN-α-Therapiedauer nach und nach abfiel. Die Daten zeigten im Hinblick auf Therapiedauer, Therapieabbruchgrund (Versagen: ja oder nein) und der Feststellung einer deutlichen ZR Zusammenhänge, wie man sie aus medizinischer Sicht erwartet hätte: Patienten ohne deutliche ZR erfuhren eher frühe Abbrüche wegen Therapieversagens, Patienten mit deutlicher Remission standen zumeist länger unter IFN-α-Therapie. Die Erkennbarkeit solcher Logik sprach für die Qualität der zytogenetischen Daten, welche ein wesentliches Fundament für die Verlässlichkeit eines entwickelten Prognosemodells bildet.

Einflüsse des Diagnosedatums und des Datums der ersten IFN-α-Gabe auf die zytogenetische Remission

Das mediane Diagnosedatum der 803 Patienten der Analysestichprobe zur zytogenetischen Remission war der 15.10.1990 und das mediane Datum der ersten IFN-α-Gabe der 19.11.1990.

Weder zwischen den bis zum medianen Diagnosedatum diagnostizierten und den später diagnostizierten Patienten noch zwischen den bis zum medianen Datum der ersten IFN-α-Gabe bereits mit IFN-α behandelten und den danach zum ersten Mal mit IFN-α behandelten Patienten gab es statistisch signifikant unterschiedliche Remissionswahrscheinlichkeiten (Kaplan-Meier-Kurven, nach dem New CML-Score stratifizierte und unstratifizierte Logrank-Tests).

3.5.6 Die zytogenetische Remission in Abhängigkeit von der Vortherapie, vom Therapieansatz und von der Zeit zwischen Diagnose und Therapiebeginn

Von den 803 Patienten wurden 263 (33%) mit einer Chemotherapie vorbehandelt, davon 184 Patienten mit HU (70%), 74 Patienten mit BU (28%) und 5 Patienten mit BU + Thioguanine (2%). Zwischen den verschiedenen Chemotherapien existierten keine statistisch signifkikanten Unterschiede in den Risikogruppen des New CML-Scores und in den späteren Remissionswahrscheinlichkeiten unter IFN-α (nach den Risikogruppen stratifizierter und unstratifizierter Logrank-Test). Wohl aber lagen statistisch signifikante ungünstigere Resultate für alle 263 Vorbehandelten im Vergleich mit den 540 nichtvorbehandelten Patienten vor, sowohl bzgl. der Risikogruppen (χ^2-Test: $p < 0{,}0001$) als auch in Bezug auf die Remissionswahrscheinlichkeiten (Logrank-Test, stratifiziert: $p < 0{,}0001$; unstratifiziert: $p = 0{,}0011$). Allerdings gehörten 239 (91%) der Patienten mit einer Vortherapie zu den Studien mit den geringsten Remissionswahrscheinlichkeiten (Abbildung 3.5, Gruppe A) und machten dort 62% der 384 Patienten aus. Die übrigen 24 vorbehandelten Patienten bildeten 9% der 405 Patienten der mittleren Gruppe B. Der nach den Gruppen A und B stratifizierte Logrank-Test zum Vergleich von vorbehandelten und nichtvorbehandelten Patienten führte einem p-Wert $> 0{,}6$.[20] Wurde für die zwischen den Studien bestehende Heterogenität der Patienten adjustiert, hatte eine Vorbehandlung mit Chemotherapie somit keine statistisch signifikante Bedeutung hinsichtlich der Wahrscheinlichkeit einer ersten deutlichen Remission.

Die Behandlung der 803 Patienten erfolgte in 476 Fällen mit IFN-α-Monotherapie (59%), 240 Patienten erhielten IFN-α + HU (30%) und 87 Patienten IFN-α + Ara-C (11%). Die drei Therapieansätze besaßen keine unterschiedlichen Risikogruppenverteilung hinsichtlich des New CML-Scores, aber jeweils statistisch signifikant unterschiedliche Wahrscheinlichkeiten eine erste deutliche ZR beobachtet zu haben (jeweils drei stratifizierte und unstratifizierte paarweise Logrank-Tests, alle $p < 0{,}0050$). Wenn man statt des New CML-Scores die Heterogenität zwischen den Studien ins Spiel brachte, verschwanden die Unterschiede in den Remissionswahrscheinlichkeiten zwischen den Therapieansätzen. Weder war innerhalb der Gruppe A der Vergleich zwischen den 190 (49%) Patienten mit IFN-α-Monotherapie und den 194 (51%) Patienten mit IFN-α + HU statistisch signifikant, noch die paarweisen Vergleiche in der Gruppe B, zwischen den 281 (69%) Patienten mit IFN-α-Monotherapie, den 50 (12%) Patienten mit IFN-α + HU und den 74 (18%) Patienten mit IFN-α + Ara-C - einerlei ob der Logrank-Test nach dem New CML-Score stratifiziert wurde oder nicht.[21] Wiederum konnten die signifikanten Remissionsunterschiede auf die Heterogenität v.a. zwischen den Gruppen A und B zurückgeführt werden. Beim Vergleich der vorgesehenen statt der tatsächlich angewandten IFN-α-Therapien ergaben sich im Wesentlichen keine anderen Resultate

[20] Ohne Gruppe C, da sie mit den 14 Österreichern ausschließlich aus Patienten ohne Vortherapie bestand.

[21] Innerhalb von Gruppe C war kein Therapievergleich möglich, da 13 von 14 Patienten IFN-α + Ara-C erhalten hatten und nur einer eine IFN-α-Monotherapie.

Die Zeit zwischen Diagnose und Therapiebeginn mit IFN-α belief sich für die 803 Patienten im Median auf 29 Tage (Minimum: 0 Tage, 1. Quartil 1 Tag, 3. Quartil 83 Tage, Maximum 182 Tage) und war bei Stratifikation nach den Gruppen A, B und C ohne statistisch signifikanten Einfluss auf die Wahrscheinlichkeiten einer ersten deutlichen ZR.

Weil statistisch signifikante Unterschiede in den Wahrscheinlichkeiten einer ersten deutlichen ZR nicht in Abhängigkeit von der Vortherapie, des Therapieansatzes oder der Zeit zwischen Diagnose und Therapiebeginn, sondern als Ergebnis der Heterogenität zwischen den zehn betrachteten Studien gesehen wurden, erfolgte die Zulassung aller 803 Patienten für die endgültige Analysestichprobe zur zytogenetischen Remission.

3.6 Lernstichprobe und Validierungsstichprobe

Prinzipiell empfiehlt es sich, vor der Entwicklung von Prognosesystemen die Analysestichprobe in eine Lern- und eine Validierungsstichprobe aufzuteilen (vgl. Abschnitt 2.5). Für den Hauptzielparameter Überlebenszeit waren nach den in den Abschnitten 3.3 und 3.4.3 beschriebenen Analysen zunächst 1384 Patienten aus 11 Studien qualifiziert. Mit dem besonderen Interesse an der zytogenetischen Remission kamen für Lern- und Validierungsstichprobe jedoch nur solche Studien in Frage, für welche die gewünschten Remissionsdaten geliefert werden konnten. Daher wurden nur zehn Studien mit 803 Patienten hinsichtlich der zytogenetischen Remission als auswertbar erachtet. Bei den multiplen Analysen fielen weitere Patienten weg, da einige nicht zu allen interessierenden Kovariablen vollständige Daten aufwiesen.

Beim Abzweigen weiterer Studien und Patienten zur Bereitstellung einer Validierungsstichprobe wäre zur Analyse prognostischer Faktoren unter Einschluss der zytogenetischen Remission eine so kleine Lernstichprobe verblieben, dass für ein daraus gewonnenes Prognosesystem die Reproduzierbarkeit seiner prognostischen Qualität in anderen, an ihrer Entwicklung unbeteiltigten Studien von vornherein gefährdet schien (vgl. Abschnitt 2.5). Aus diesem Grund wurde für die zunächst verfügbaren Daten auf die Abstellung einer Validierungsstichprobe verzichtet. Im Sommer 2005, als längere Beobachtungszeiten und höhere Ereigniszahlen für die deutschen Studien CML III [109] und CML IIIA [110] vorlagen, wurde nachträglich eine Validierungsstichprobe gefunden.

Kapitel 4

Die Entwicklung des Prognosesystems

4.1 Deskription des Hauptzielparameters und der Kovariablen

4.1.1 Der Hauptzielparameter Überlebenszeit

Die Analysen des Hauptzielparameters Überlebenszeit in Abschnitt 3.4 führten zur Lernstichprobe von 1329 Patienten. Deren mediane Überlebenszeit lag bei 72 Monaten; 631 Patienten waren

Tabelle 4.1: Beobachtete Überlebens- und Behandlungszeiten ab IFN-α-Therapiebeginn

Variable	Patientenzahl	Minimum	Median	Maximum	Mittelwert	Standardabweichung
	n	Tage	Mon.[a]	Mon.	Mon.	Mon.
Beobachtete Überlebenszeit						
Alle Patienten	1329	13	48	141	52	32
Patienten ohne allo. SZT in 1. CP[b]	1085	14	56	141	59	32
Alle Patienten mit zens. Zeiten	698	13	53	141	56	37
Zens. Zeiten ohne allo. SZT in 1. CP	454	14[c]	73	141	73	33
Zeitdauer der IFN-α-Gabe						
Alle Patienten	1329	1	24	141	33	30
Patienten, die kein IFN-α mehr erhielten	1107	1	18	123	25	22
Patienten, die weiter IFN-α erhielten	222	14	69	141	72	33

[a] Monate.

[b] „Allo. SZT in 1. CP“ steht für „allogene SZT in 1. chronischer Phase“.

[c] Nur sieben der 454 noch unter Risiko stehenden und vor 1998 Patienten mit zensierten Zeiten wurden kürzer als ein Jahr beobachtet. Der nur 14 Tage beobachtete Patient gehörte zu den 222 nach wie vor unter IFN-α-Therapie stehenden Patienten und galt als „lost to follow-up“.

verstorben (47%). Als Todesursache wurde bei 487 Patienten (77% der 631) „CML bezogen“, bei 79 (13%) „nicht CML bezogen“ und bei zwölf (2%) „SZT bezogen“ angegeben. Die zwölf zuletzt genannten hatten eine SZT erhalten, jedoch keine allogene SZT in erster chronischer

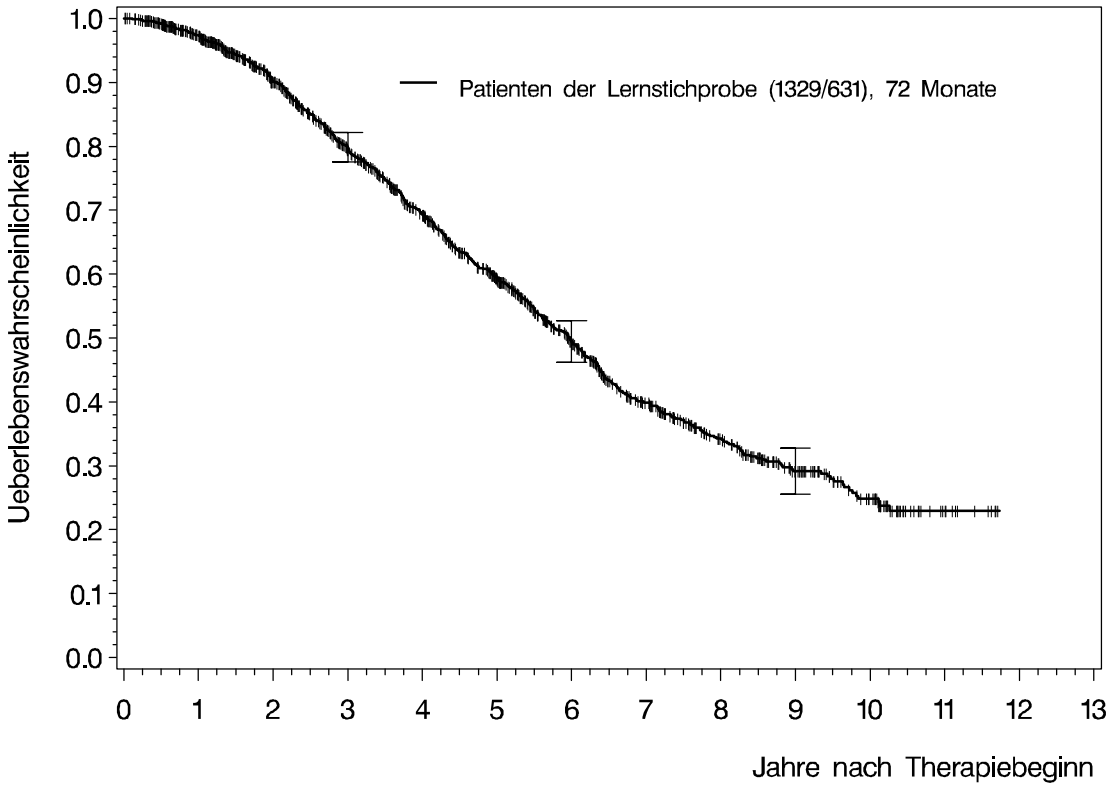

Abbildung 4.1: Kaplan-Meier-Kurve mit geschätzten Überlebenswahrscheinlichkeiten der 1329 Patienten der Lernstichprobe. Die Legende „(1329/631), 72 Monate“ bedeutet: Unter 1329 Patienten mit anhand der zugehörigen Kaplan-Meier-Kurve beschriebenen Überlebenswahrscheinlichkeiten wurden 631 Todesfälle beobachtet. Die mediane Überlebenszeit betrug 72 Monate. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Überlebenswahrscheinlichkeit mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet.

Phase. Von den verbliebenen 8% der 631 Verstorbenen lag in 27 Fällen (4%) die explizite Angabe „unbekannt“ vor und bei den übrigen 26 (4%) war keine Todesursache vor Datenbankschluss eruierbar. Von den 698 zensierten Patienten erhielten 244 (35% von 698) eine allogene SZT in 1. chronischer Phase, wobei wegen der Zensierung gemäß der Definition des Hauptzielparameters (Abschnitt 2.3) die danach Verstorbenen nicht zu den 631 zuvor genannten dazugezählt wurden. Die mediane Zeit, die alle 1329 Patienten unter Beobachtung standen, betrug 48 Monate (vgl. Tabelle 4.1). Die weiterhin unter Ereignisrisiko stehenden 454 Patienten wurden im Median 73 Monate beobachtet. Von diesen 454 erhielten bei Datenbankschluss mit 222 Patienten (49%) fast die Hälfte nach wie vor IFN-α. Weitere Details zum Therapieverlauf, zu Abbruchgründen und Transplantationen wurden in Abschnitt 3.4.2 beschrieben. Abbildung 4.1 zeigt die aus der Stichprobe geschätzten Überlebenswahrscheinlichkeiten.

4.1.2 Die Baselinevariablen

Wie in Kapitel 2 begründet[1], wurden später nur diejenigen Baselinevariablen zur multiple Analyse in der Lernstichprobe zugelassen, zu welchen für wenigstens 90% der Patienten Merkmalsausprägungen vorlagen. Diese Bedingung wurde von den sechs am New CML-Score beteiligten

[1] Siehe Abschnitt 2.6

[42] und drei weiteren Variablen erfüllt. Auf die übrigen, nicht zur multiplen Analyse zugelassenen Baselinevariablen, die aber bereits bei Entwicklung des New CML-Scores intensiv untersucht wurden, wurde in vorliegender Arbeit nicht noch einmal eingegangen. Die nachfolgende Tabelle 4.2 bietet einen Einblick in die Werteverteilung der für die multiple Analyse zugelassenen Variablen.

Tabelle 4.2: Initiale Charakteristika der für die multiple Analyse zugelassenen Baselinevariablen

Variable	Patientenzahl	Fehlende Werte	Minimum	Median	Maximum	Mittelwert	Standardabweichung
	n	n					
Metrische Skalierung							
Alter in vollen Jahren	1329	0	9	49	85	48	14
Hämoglobin in g/dl	1295	34	4,2	12,0	17,5	11,9	2,1
Hämoglobin in g/dl - Frauen	553	10	5,1	11,7	17,1	11,5	1,9
Hämoglobin in g/dl - Männer	742	24	4,2	12,3	17,5	12,2	2,1
Leukozytenzahl in 10^9/l	1290	39	11	113	626	137	104
Thrombozytenzahl in 10^9/l	1317	12	43	401	3490	503	362
Blasten im p.B.[a] in %	1315	14	0	1	10	1,7	2,1
Basophile im p.B. in %	1309	20	0	3	33	4,1	3,6
Eosinophile im p.B. in %	1291	38	0	2	20	2,5	2,4
Milzvergrößerung in cm[b]	1311	18	0	3	30	4,9	5,9
New CML-Score [42]	1279	50	0	888	2559	883	512

Variable			Anzahl (Prozentualer Anteil)
	n	n	n (%)
Nominale/ordinale Skalier.			
Alter in vollen Jahren	1329	0	jünger als 50 Jahre: 682 (51)
Geschlecht	1329	0	Männer: 766 (58)
Basophile im p.B. in %	1309	20	weniger als 3%: 512 (39)
Thrombozytenzahl in 10^9/l	1317	12	weniger als 1500 in 10^9/l: 1286 (98)
New CML-Score [42]	1279	50	Patienten mit Niedrigrisiko: 545 (43)
			Patienten mit mittlerem Risiko: 575 (45)
			Patienten mit Hochrisiko: 159 (12)

[a] Die Abkürzung „p.B.“ steht für „peripheres Blut“.
[b] Die Milzvergrößerung wurde in Zentimetern unter dem linken Rippenbogen gemessen.

4.1.3 Die zeitabhängige Kovariable zytogenetische Remission

Die 803 Patienten, welche sich mit ihren zytogenetischen Daten nach den Analysen von Kapitel 3 für die Lernstichprobe qualifiziert hatten, standen im Median 826 Tage (27 Monate) unter IFN-α-Therapie [Minimum: 16 Tage, Maximum: 10 Jahre und 3 Monate]. Das beste unter IFN-

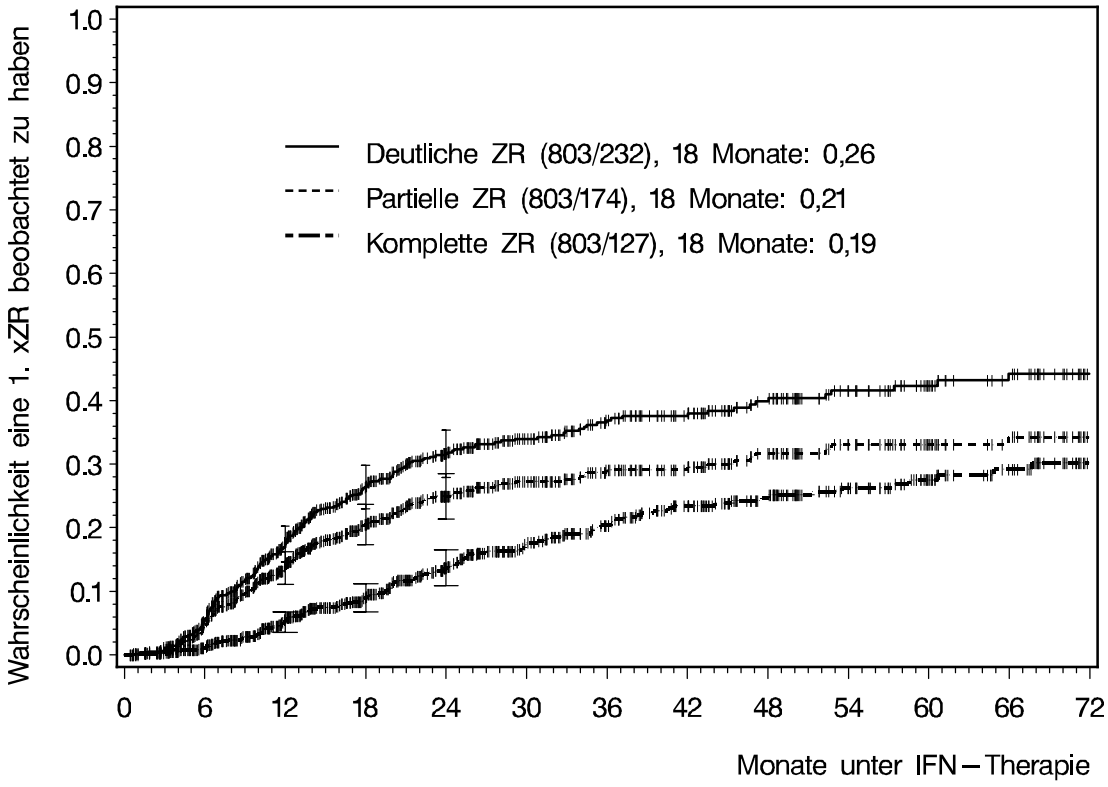

Abbildung 4.2: Kaplan-Meier-Kurven mit bei 803 Patienten der Lernstichprobe geschätzten Wahrscheinlichkeiten bis zu einer gewissen Therapiedauer eine deutliche, partielle oder komplette zytogenetische Remission (xZR) beobachtet zu haben. Das „x" fungiert als Platzhalter für „deutlich", „partiell" oder „komplett". Die Zahlen „(803/232), 18 Monate: 0,26" bedeuten: Unter 803 Patienten wurden 232 deutliche ZR beobachtet. Die geschätzte Wahrscheinlichkeit, bis Ende des 18. Monats eine deutliche ZR beobachtet zu haben, lag bei 0,26. Die beiden anderen Legenden sind analog zu verstehen. Patienten ohne entsprechende Remission wurden zum Zeitpunkt des Therapieabbruchs zensiert, bei der mittleren Kurve auch zum Erreichen einer kompletten ZR. Zu den Zeitpunkten 12, 18 und 24 Monate wurden um die geschätzte Wahrscheinlichkeit mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge von oben nach unten.

α-Therapie erreichte zytogenetische Remissionsergebnis war für 232 Patienten (29% von 803) eine deutliche ZR, wobei 127 (16%) eine komplette und 105 (13%) eine partielle ZR verzeichneten. Weiter hatten 67 Patienten (8%) eine geringe ZR als bestes Ergebnis und weitere 168 (21%) wenigstens eine minimale ZR. Für 336 Patienten (42%) wurde zu keinem Zeitpunkt eine ZR beobachtet. Vor der kompletten ZR als bestem Resultat, registrierte man im Falle von 69 Patienten (54% von 127) eine partielle ZR als erstes deutliches Remissionsergebnis. Die mediane Zeit bis zum Erreichen einer ersten partiellen ZR betrug unter den 174 (105+69) Patienten 11 Monate [Minimum: 34 Tage, Maximum: 9 Jahre und 6 Monate]. Eine erste komplette Remission wurde im Median nach 19 Monaten Therapiedauer beobachtet [Minimum: 23 Tage, Maximum: 6 Jahre und 4 Monate]. Bedingt durch die etwas längeren Zeiten bei den 58 Patienten, die als erste deutliche Remission direkt eine komplette Remission erzielten, belief sich die mediane Zeit bis zum Erreichen einer ersten deutlichen ZR (174+58 Patienten) auf 12 Monate IFN-α-Therapiedauer [Minimum: 23 Tage, Maximum: 9 Jahre und 6 Monate]. Abbildung

4.2 zeigt die Wahrscheinlichkeiten, eine erste partielle, komplette oder deutliche zytogenetische Remission nach einer bestimmten Zeit unter IFN-α-Therapie beobachtet zu haben. Patienten, deren IFN-α-Therapie ohne die Festellung einer mindestens partiellen ZR endete, wurden zum Therapieabbruch zensiert.[2] Bei der Kurve zur partiellen Remission wurden zudem 58 Patienten zum Zeitpunkt ihrer ersten kompletten Remission zensiert, da sie das Stadium einer partiellen Remission „übersprungen“ hatten.[3] Die Wahrscheinlichkeiten, zu den Zeitpunkten 12, 15, 18, 21 und 24 Monaten eine ZR beobachtet zu haben, lagen für die partielle Remission bei 0,14, 0,18, 0,21, 0,23 und 0,25, im Falle der kompletten Remission bei 0,05, 0,07, 0,09, 0,12 und 0,14 und für eine deutliche Remission bei 0,17, 0,23, 0,26, 0,30 und 0,32.

4.2 Die univariate Analyse des Einflusses auf die Überlebenszeit

4.2.1 Die Baselinevariablen

Mit Ausnahme des Geschlechts waren alle zum Diagnosezeitpunkt erhobenen Kovariablen metrisch skaliert. Über den gesamten Wertebereich einer jeden metrischen Kovariablen konnte bei den verschiedenen Übergängen von einer Merkmalsstufe auf die nächsthöhere nicht von derselben Änderung des relativen Überlebensrisikos ausgegangen werden. Daher sollte durch geeignete Kategorisierung nach ein oder mehreren zusätzlichen Variablendefinitionen gesucht werden. Außer aufschlussreicherer Erkenntnis über prognostische Unterschiede, boten die alternativen Unterteilungen des Wertebereiches in verschiedene Gruppen eine erhöhte Flexibilität bei der Zusammensetzung multipler Modelle. Bei Bildung der Kategorien wurde Sorge getragen, dass innerhalb jeder Kategorie von homogenen Überlebenswahrscheinlichkeiten ausgangen werden konnte. Überlebenswahrscheinlichkeiten wurden mit Hilfe von Kaplan-Meier-Kurven und Logrank-Test verglichen. Dabei erfolgten die Vergleiche von Patientengruppen verschiedener Wertebereiche immer paarweise. Von „statistischer Signifikanz“ wurde gesprochen, wenn ein p-Wert unter dem gewählten Signifikanzniveau von $\alpha = 0,05$ lag. Die Kategorienbildung wurde durch Anwendung der „Minimum p-value“-Methode unterstützt, wobei in diesem Fall bei der Logrank-Statistik die p-Werte nach Lausen und Schumacher [73] für die multiple Cutpointsuche adjustiert und zur Kenntlichmachung mit p_{ad} bezeichnet wurden. Weil bei einer Variablen auch die Übergänge von einer Kategorie zur nächsten nicht jeweils dieselbe Veränderung des relativen Risikos zufolge haben müssen, wurde im Cox-Modell eine Referenzkategorie gewählt und die übrigen Kategorien durch Dummyvariablen kodiert. Der p-Wert zur Beschreibung des prognostischen Einflusses auf die Überlebenswahrscheinlichkeiten ergab sich für Kovariablen mit $x \geq 3$ Kategorien aus der gemeinsamen Wald-Statistik zu den $x - 1$ Dummyvariablen.

Das Alter

Das Patientenalter zum Diagnosezeitpunkt wurde zunächst in Altersgruppen eingeteilt, die jeweils fünf Jahre umfassten. Eine Ausnahme bildeten die jüngsten und die ältesten Patienten, bei welchen wegen sonst zu geringer Fallzahlen alle Patienten bis einschließlich 20 Jahre ($n = 41$, 3% von 1329) bzw. alle ab 71 Jahre ($n = 24$, 2%) jeweils zu einer Gruppe zusammengelegt wurden. Verglich man die Überlebenswahrscheinlichkeiten der 41jährigen mit denjenigen der

[2] Vgl. Abschnitt 3.5.1.

[3] Das Erreichen einer ersten kompletten ZR vor Beobachtung der ersten partiellen ZR kann als „Competing risk“ gesehen werden [35, 59]. Die Wahrscheinlichkeiten, eine erste partielle ZR zu beobachten, müssten daher mit dem „Cumulative Incidence“-Schätzer [59] geschätzt werden. Die in der Graphik einheitliche Verwendung der Kaplan-Meier-Methode führte aber im vorliegenden Fall zu vernachlässigbaren Überschätzungen der Ereigniswahrscheinlichkeiten.

45jährigen und die der 56jährigen mit denjenigen der 60jährigen hatten die Jüngeren einen statistisch signifikanten Überlebensvorteil (Logrank-Test). Nach Änderung der Gruppengrenzen in 36-41 und 42-45 sowie 51-56 und 57-60 ergab sich in keiner der insgesamt zwölf Altersgruppen ein statistisch signifikanter Überlebensvorteil beim Vergleich der Jüngsten mit den Ältesten; zudem war nun innerhalb keiner der Gruppen eine strikte Anordnung der Kaplan-Meier-Kurven nach dem Alter erkennbar[4]. Die Altersgruppe der 36-41jährigen wies ebensowenig einen statistisch signifikanten Überlebensunterschied im Vergleich zu den vier jüngeren Altersgruppen auf, wie die 42-45jährigen im Vergleich zu den sechs nachfolgenden Altersgruppen. Im Gegensatz dazu, zeigte der Logrank-Test statistisch signifikant höhere Überlebenswahrscheinlichkeiten für die Jüngeren, wenn man die 36-41jährigen mit den 42-45jährigen verglich. All dies sprach dafür, dass das Alter in zwei Gruppen eingeteilt werden konnte, die intern jeweils einen sehr ähnlichen prognostischen Einfluss auf die Überlebenswahrscheinlichkeiten aufwiesen, sich voneinander bzgl. dieses Einflusses jedoch deutlich unterschieden. Entsprechendes war im Cox-Modell zu beobachten: Bezogen auf die Referenzkategorie „Patienten bis 20 Jahre“ stieg die relative Hazardfunktion von 1,090 (36-41 Jahre) auf 1,999 (42-45 Jahre) sprunghaft an und erreichte ihr Maximum mit 2,331 bei den 66-70jährigen; weder vor noch nach dem Sprung zeigte sich bzgl. der relativen Hazardfunktionen ein lineares Wachstum mit dem Alter. Mit Hilfe des „Minimum p-value“-Methode wurde als einziger Cutpoint „41 Jahre“ gefunden (Logrank-Test: $p_{ad} \leq 0{,}0001$). Die im Cox-Modell für „Alter in Jahren - ohne Gruppierung“, „Alter - zwei Gruppen: $\leq$ 41 Jahre (34% von 1329) und $>$ 41 Jahre (66%)“ und „Alter - zwei Gruppen (wie beim New CML-Score [42]): $\leq$ 49 Jahre (51%) und $>$ 49 Jahre (49%)“ erzielten p-Werte (Wald-Test) lagen alle unter 0,0001 (vgl. Tabellen 4.3 und 4.4).

Geschlecht

Von den 1329 Patienten waren 563 Frauen (42%) und 766 Männer (58%). Ein statstistisch signifikanter Überlebensunterschied zwischen beiden Geschlechtern existierte nicht (vgl. Tabelle 4.4).

Hämoglobin

Bei den Referenzbereichen für Hämoglobin wird zwischen Frauen (12-16 g/dl) und Männern (14-18 g/dl) unterschieden.[5] Entsprechend wurden Frauen und Männer zunächst getrennt bewertet. Durch Rundung auf ganze Zahlen wurden die bis auf eine Stelle nach dem Komma gegebenen Originalwerte in erste Gruppen zusammengefasst. Zur Altersvariablen analoge Betrachtungen ergaben für die Frauen zwei Gruppen mit intern vergleichbaren Überlebenswahrscheinlichkeiten, wenn man die gerundete Grenze 11,4 auf 11,3 g/dl heruntersetzte und für die Männer eine Gruppe, wenn man von 13,4 auf 13,5 g/dl erhöhte (vgl. Tabelle 4.4).[6] Die beiden neuen Grenzen wurden jeweils als Cutpoint bei Anwendung der „Minimum p-value“-Methode bestätigt ($p_{ad} \leq$ 0,0001, Logrank-Test). Männer mit Hämoglobinwerten $>$ 13,5 g/dl konnten über die Grenze 14,4 g/dl in zwei Gruppen mit unterschiedlichen Überlebenswahrscheinlichkeiten weiter unterteilt werden. Zur Bildung einer gemeinsamen kategorialen Kovariablen für die 2 + 3 Gruppen

[4] So lag z.B. die Kaplan-Meier-Kurve der 41jährigen fast komplett über den Kurven der 36, 39 und 40jährigen, während die Kurven der 37 und 38jährigen alle anderen Kurven schnitten. Dem beobachteten Ergebnis wurde vor der allgemeinen Neigung zu „Grenzen als Vielfacher von 10“ Vorrang gegeben.

[5] Siehe z.B. Pschyrembel Klinisches Wörterbuch [87].

[6] Patientinnen mit 11,3 g/dl wiesen im Vergleich zu jenen mit 11,4 g/dl statistisch signifikant schlechtere Überlebenswahrscheinlichkeiten auf (Logrank-Test, $p = 0{,}0007$), bei den Männern war der Unterschied zwischen 13,4 und 13,5 g/dl klar erkennbar, aber statistisch nicht signifikant.

Tabelle 4.3: Univariate Analysen im Cox-Modell: Einfluss der Baselinevariablen auf die Überlebenswahrscheinlichkeiten - metrische Skalierung

Variable[a]	Patientenzahl	Schätzung Koeffizient	Standardabweichung	Walds χ^2-Statistik		p-Wert	
	n/tot[b]	$\hat{\beta}$	$\hat{v}$	X^2	df[c]	p	RR[d]
Alter	1329/631	0,0158	0,0031	**25,5139**	**1**	**<0,0001**	1.016
Hämoglobin - Frauen	553/266	-0,1076	0,0326	**10,9040**	**1**	**0,0010**	0.898
Hämoglobin - Männer	742/346	-0,1212	0,0246	**24,3352**	**1**	**<0,0001**	0.886
Hämoglobin - alle[e]	1295/612	-0,1146	0,0196	**34,1272**	**1**	**<0,0001**	0.892
Leukozyten / 100	1290/618	0,1847	0,0366	**25,4908**	**1**	**<0,0001**	1.203
Thrombozyten / 100	1317/626	0,0365	0,0109	**11,3057**	**1**	**0,0008**	1.037
Blasten im p.B.	1315/624	0,1013	0,0169	**35,7824**	**1**	**<0,0001**	1.107
Basophile im p.B.	1309/621	0,0380	0,0104	**13,3682**	**1**	**0,0003**	1.039
Eosinophile im p.B.	1291/617	0,0802	0,0159	**25,3735**	**1**	**<0,0001**	1.084
Milzvergrößerung	1311/621	0,0455	0,0065	**49,8578**	**1**	**<0,0001**	1.047
New CML-Score / 100	1279/611	0,0860	0,0079	**119,8248**	**1**	**<0,0001**	1.090

[a]Einheiten und Messgenauigkeit wie in Tabelle 4.2, nur wurden Leukozyten und Thrombozyten auf die nächste Zehnerstelle gerundet. Ebenso wie der New CML-Score, wurden Leukozyten- und Thrombozytenzahl anschließend durch 100 dividiert, um dadurch eine mit den übrigen Variablen vergleichbare Größe der Koeffizienten und Standardabweichungen zu erhalten.

[b]n: Gesamtzahl der Patienten mit Daten, tot: die davon inzwischen Verstorbenen.

[c]Freiheitsgrade.

[d]Relatives Risiko: Verhältnis der Hazardfunktion zum Variablenwert $x+1$ zur Hazardfunktion zum Variablenwert x.

[e]Von den Hämoglobinwerten der Frauen wurde deren medianer Wert 11,7 g/dl, von den Werten der Männer deren medianer Wert 12,3 g/dl abgezogen.

der Frauen und Männer wurden Überlebenskurven sowie Koeffizienten im Cox-Modell (Referenzkategorie: „Männer, Hämoglobinwerte > 14,4 g/dl, 16% von 742 Männern“) miteinander verglichen. Wegen nahezu identischer Ergebnisse konnten dabei Frauen mit Hämoglobinwerten ≤ 11,3 g/dl (44% von 553 Frauen) und Männer mit Werten ≤ 13,5 g/dl (71%) ebenso zu einer Gruppe zusammengefasst werden wie Frauen mit Werten > 11,3 g/dl (56%) und Männer mit 13,6-14,4 g/dl (13%). Bei der so entstandenen gemeinsamen Variablen unterschieden sich die drei Kategorien statistisch signifikant (Logrank-Test bei paarweisen Gruppenvergleichen). Durch eine weitere Zusammenlegung wurde als zweiter Kategorisierungsvorschlag eine dichotome Kovariable bereitgestellt, die sich nur an den beiden Gruppengrenzen nach der „Minimum p-value“-Methode orientierte. Nachdem von den Hämoglobinwerten der Frauen und Männer jeweils die geschlechtsspezifischen Medianwerte 11,7 g/dl bzw. 12,3 g/dl abgezogen wurden, wurden die so adjustierten Werte zu einer gemeinsamen metrischen Kovariablen zusammengefügt (vgl. Tabelle 4.3). Alle drei Variablendefinitionen wiesen im univariaten Cox-Modell p-Werte ≤ 0,0001 auf (Wald-Test).

Tabelle 4.4: Univariate Analysen im Cox-Modell: Einfluss der Baselinevariablen auf die Überlebenswahrscheinlichkeiten - nominale/ordinale Skalierung

Variable[a]	Patientenzahl	Schätzung Koeffizient	Standardabweichung	Walds χ^2-Statistik		p-Wert		MÜ[b]
	n/tot[c]	$\hat{\beta}$[d]	$\hat{v}$	X^2	df[e]	p	RR[f]	Mo.[g]
Alter								
Zwei Gruppen	1329/631			**24,1721**	**1**	**<0,0001**		
≤41 Jahre	446/133	0	-				1	90
>41 Jahre	883/498	0,4803	0,0977				1,617	66
Zwei Gruppen[h]	1329/631			**20,5423**	**1**	**<0,0001**		
≤49 Jahre	682/240	0	-				1	80
>49 Jahre	647/391	0,3725	0,0822				1,451	65
Geschlecht	1329/631			**0,5600**	**1**	**0,4543**		
Frauen	563/272	0	-				1	74
Männer	766/359	0,0602	0,0804				1,062	70
Hämoglobin								
Drei Gruppen	1295/612			**43,1446**	**2**	**<0,0001**		
w[i], ≤11,3 g/dl & m, ≤13,5 g/dl	771/402	0,8386	0,1679				2,313	63
w, >11,3 g/dl & m, 13,6-14,4 g/dl	406/171	0,3781	0,1775				1,460	81
m, >14,4 g/dl	118/ 39	0	-				1	106
Zwei Gruppen	1295/612			**40,3018**	**1**	**<0,0001**		
w, ≤11,3 g/dl & m, ≤13,5 g/dl	771/402	0,5422	0,0854				1,720	63
w, >11,3 g/dl & m, >13,5 g/dl	524/210	0	-				1	86
Leukozyten								
Zwei Gruppen	1290/618			**32,8346**	**1**	**<0,0001**		
$\leq 50 \times 10^9$/l	315/112	0	-				1	99
$>50 \times 10^9$/l	975/506	0,5995	0,1046				1,821	67

[a] Einheiten und Messgenauigkeit wie in Tabelle 4.2.
[b] Mediane Überlebenszeit.
[c] n: Gesamtzahl der Patienten mit Daten, tot: die davon inzwischen Verstorbenen.
[d] Der Wert „0“ steht jeweils für die Referenzgruppe.
[e] Freiheitsgrade.
[f] Relatives Risiko: Verhältnis der geschätzten Hazardfunktion zur Hazardfunktion der Referenzkategorie.
[g] Monate.
[h] Dichotomisierung wie im New CML-Score
[i] w: weiblich; m: männlich.

Leukozyten

Die in 10^9/l gegebenen Leukozytenzahlen wurden auf die nächste Zehnerstelle gerundet und dann in Gruppen, die jeweils 50 Werte umfassten, unterteilt. Patienten mit Werten $\geq 310 \times 10^9$/l ($n =$

98, 8% von 1290) bildeten eine gemeinsame Gruppe. Alle Patienten mit Werten $> 100 \times 10^9$/l konnten zu einer Gruppe zusammengefasst werden, innerhalb welcher kaum unterscheidbare Kaplan-Meier-Kurven zu beobachten waren. Während die Überlebenswahrscheinlichkeiten der 315 Patienten mit Werten $\leq 50 \times 10^9$/l (24%) statistisch signifikant höher waren (Logrank-Test bei paarweisen Gruppenvergleichen) als die der Gruppen „60-100“ ($n = 296$, 23%) und „$> 100 \times 10^9$/l“ ($n = 679$, 53%), unterschieden sich zwar die Wahrscheinlichkeiten der verbliebenen zwei Gruppen mit höheren Leukozytenzahlen, jedoch nicht statistisch signifikant. Mit 50×10^9/l als zugleich einziger durch die Anwendung der „Minimum p-value“-Methode entdeckter Grenze (Logrank-Test, $p_{ad} \leq 0{,}0001$) erbot sich die Einführung der dichtomen Variablen „≤ 50“ vs. „$> 50 \times 10^9$/l“. Wie die metrische Kovariable mit gerundeten Leukozytenzahlen, führte die kategoriale Variablendefinition im univariaten Cox-Modell zu einem p-Wert $\leq 0{,}0001$ (Wald-Test, vgl. Tabellen 4.3 und 4.4).

Thrombozyten

Auch die Thrombozytenzahlen wurden zuerst auf die nächste Zehnerstelle gerundet und in Gruppen, die jeweils 50 Werte umfassten, unterteilt. Patienten mit Werten $\geq 1490 \times 10^9$/l ($n = 31$, 2% von 1317, entspricht der Grenze aus dem New CML-Score) bildeten eine gemeinsame Gruppe. Statistisch signifikant und zudem bestätigt durch die „Minimum p-value“-Methode konnten zwei Gruppen unterschieden werden (Logrank-Test, $p_{ad} = 0{,}0085$): „$\leq 1350 \times 10^9$/l“ (97%) vs. „$> 1350 \times 10^9$/l“ (3%). Auch bei einer Verschiebung zur Grenze 1490 (New CML-Score) wurden zwei Gruppen mit statistisch signifikant unterschiedlichen Überlebenszeiten beobachtet (Logrank-Test). Sowohl die gerundete metrische Kovariable als auch die Kategorisierungen nach zwei Gruppen erwiesen sich im univariaten Cox-Modell als statistisch signifikant (Wald-Test: alle p-Werte $< 0{,}0010$, vgl. Tabellen 4.3 und 4.5).

Blasten im peripheren Blut

Hinsichtlich statistisch signifikant unterschiedlicher Überlebenswahrscheinlichkeiten ließen sich aus den einzelnen Prozentangaben zwischen 0 und 10% drei Gruppen mit 0% ($n = 513$, 39% von 1315), 1-7% ($n = 766$, 58%) und $> 7\%$ ($n = 36$, 3%) Blasten im p.B. bilden (Kaplan-Meier-Kurven und Logrank-Test bei paarweisen Gruppenvergleichen). Die Patienten ohne Blasten hatten gegenüber den nachfolgenden zwei Gruppen einen Überlebensvorteil, danach kamen die Patienten mit 1-7% Blasten als zweitgünstigste Gruppe. Die Analyse mit der „Minimum p-value“-Methode erbrachte die Einteilung „0%“ vs. „$> 0\%$“ (Logrank-Test: $p_{ad} \leq 0{,}0001$). Die Variable mit den Originaldaten und die beiden kategorisierten Alternativen waren im univariaten Cox-Modell alle mit p-Werten $\leq 0{,}0001$ statistisch signifikant (Wald-Test, vgl. Tabellen 4.3 und 4.5).

Basophile im peripheren Blut

Für die Basophile wurden bei Analyse der 14 Patientengruppen mit Prozentwerten zwischen 0% und 12% und Werten $> 12\%$ ($n = 32$, 2% von 1309) drei Gruppen mit statistisch signifikant unterschiedlichen Überlebenswahrscheinlichkeiten ermittelt (Kaplan-Meier-Kurven und Logrank-Test bei paarweisen Gruppenvergleichen). Die Gruppe mit den höchsten Überlebenswahrscheinlichkeiten bestand aus den Patienten mit 0-2% Basophilen ($n = 512$, 39% von 1309), die mit den zweithöchsten war die Gruppe mit 3-11% Basophilen ($n = 743$, 57%) und die mit den niedrigsten die Gruppe mit $> 11\%$ Basophilen ($n = 54$, 4%). Wie schon bei der Entwicklung des New CML-Scores [42], wurde mit Hilfe der „Minimum p-value“-Methode die stärkste

Tabelle 4.5: Univariate Analysen im Cox-Modell: Einfluss der Baselinevariablen auf die Überlebenswahrscheinlichkeiten - nominale/ordinale Skalierung

Variable[a]	Patientenzahl	Schätzung Koeffizient	Standardabweichung	Walds χ^2-Statistik		p-Wert		MÜ[b]
	n/tot[c]	$\hat{\beta}$[d]	$\hat{v}$	X^2	df[e]	p	RR[f]	Mo.[g]
Thrombozyten								
Zwei Gruppen	1317/626			**13,7249**	**1**	**0,0002**		
$\leq 1350 \times 10^9$/l	1273/595	0	-				1	73
$>1350 \times 10^9$/l	44/ 31	0,6834	0,1845				1,981	45
Zwei Gruppen[h]	1317/626			**12,2515**	**1**	**0,0005**		
$\leq 1490 \times 10^9$/l	1286/604	0	-				1	72
$>1490 \times 10^9$/l	31/ 22	0,7616	0,2176				2,142	45
Blasten								
Drei Gruppen	1315/624			**39,8574**	**2**	**<0,0001**		
0%	513/205	0	-				1	83
1-7%	766/393	0,4419	0,0865				1,556	66
>7%	36/ 26	1,0350	0,2086				2,815	35
Zwei Gruppen	1315/624			**30,1680**	**1**	**<0,0001**		
0%	513/205	0	-				1	83
>0%	802/419	0,4700	0,0856				1,600	66

[a] Einheiten und Messgenauigkeit wie in Tabelle 4.2, nur Leukozyten und Thrombozyten wurden auf die nächste Zehnerstelle gerundet.
[b] Mediane Überlebenszeit.
[c] n: Gesamtzahl der Patienten mit Daten, tot: die davon inzwischen Verstorbenen.
[d] Der Wert „0“ steht jeweils für die Referenzgruppe.
[e] Freiheitsgrade.
[f] Relatives Risiko: Verhältnis der geschätzten Hazardfunktion zur Hazardfunktion der Referenzkategorie.
[g] Monate.
[h] Dichotomisierung wie im New CML-Score

Trennung der Überlebenswahrscheinlichkeiten zwischen den Gruppen „0-2%“ und „> 2%“ gefunden (Logrank-Test: $p_{ad} \leq 0{,}0001$). Die metrische Variable wie auch die beiden kategorialen Einteilungen wiesen im univariaten Cox-Modell p-Werte $\leq 0{,}0001$ auf (Wald-Test, vgl. Tabellen 4.3 und 4.6).

Eosinophile im peripheren Blut

Untersucht wurden die 11 Patientengruppen mit Werten zwischen 0 und 9% sowie > 9% Eosinophile ($n = 33$, 2% von 1291). Die Gruppeneinteilung in Bezug auf statistisch signifikant unterschiedliche Überlebenswahrscheinlichkeiten (Kaplan-Meier-Kurven und Logrank-Test bei paarweisen Gruppenvergleichen) führte zu denselben Resultaten wie bei den Basophilen, nur fand sich bei den Eosinophilen als Obergrenze der Gruppe mit den zweithöchsten Überlebenswahrscheinlichkeiten 8% anstatt 11% (vgl. Tabelle 4.6): 800 Patienten (62%) besaßen 0-2% Eosi-

Tabelle 4.6: Univariate Analysen im Cox-Modell: Einfluss der Baselinevariablen auf die Überlebenswahrscheinlichkeiten - nominale/ordinale Skalierung

Variable[a]	Patientenzahl	Schätzung Koeffizient	Standardabweichung	Walds χ^2-Statistik		p-Wert		MÜ[b]
	n/tot[c]	$\hat{\beta}$[d]	$\hat{v}$	X^2	df[e]	p	RR[f]	Mo.[g]
Basophile								
Drei Gruppen	1309/621			**24,2042**	**2**	**<0,0001**		
0-2%	512/222	0	-				1	81
3-11%	743/369	0,3136	0,0852				1,368	66
>11%	54/ 30	0,8110	0,1952				2,250	43
Zwei Gruppen[h]	1309/621			**16,6476**	**1**	**<0,0001**		
0-2%	512/222	0	-				1	81
>2%	797/399	0,3429	0,0841				1,409	65
Eosinophile								
Drei Gruppen	1291/617			**28,5945**	**2**	**<0,0001**		
0-2%	800/354	0	-				1	77
3-8%	456/241	0,3692	0,0837				1,447	63
>8%	35/ 22	0,8076	0,2199				2,243	38
Zwei Gruppen	1291/617			**23,9398**	**1**	**<0,0001**		
0-2%	800/354	0	-				1	77
>2%	491/263	0,3993	0,0861				1,491	62
Milzvergröß.								
Zwei Gruppen	1311/621			**44,1601**	**1**	**<0,0001**		
0-7cm	940/414	0	-				1	77
>7cm	371/207	0,5684	0,0855				1,765	53
New CML-Score								
Drei Gruppen	1279/611			**119,2284**	**2**	**<0,0001**		
Niedrigrisiko	545/172	0	-				1	98
Mittleres Risiko	575/319	0,6133	0,0950				1,847	67
Hochrisiko	159/120	1,3108	0,1205				3,709	43

[a] Einheiten und Messgenauigkeit wie in Tabelle 4.2.
[b] Mediane Überlebenszeit.
[c] n: Gesamtzahl der Patienten mit Daten, tot: die davon inzwischen Verstorbenen.
[d] Der Wert „0“ steht jeweils für die Referenzgruppe.
[e] Freiheitsgrade.
[f] Relatives Risiko: Verhältnis der geschätzten Hazardfunktion zur Hazardfunktion der Referenzkategorie.
[g] Monate.
[h] Dichotomisierung wie im New CML-Score

nophile, 456 (35%) 3-8% und 35 (3%) mehr als 8% Eosinophile. Bei Anwendung der „Minimum p-value“-Methode wurde die Gruppenteilung „0-2%“ vs. „> 2%“ identifiziert (Logrank-Test: $p_{ad} \leq 0{,}0001$). Die Analyse der damit drei Variablendefinitionen bei den Eosinophilen führte im univariaten Cox-Modell jeweils zu p-Werten $\leq 0{,}0001$ (Wald-Test, vgl. Tabellen 4.3 und 4.6).

Die Milzvergrößerung

Für die Milzvergrößerung wurden die zwölf Gruppen mit den Werten zwischen 0 und 10 cm sowie 11-30 cm ($n = 219$, 17% von 1311 Patienten) unterschieden. Innerhalb der Gruppen „0-2 cm“ ($n = 637$, 49%), „3-7 cm“ ($n = 303$, 23%) und „8-30 cm“ ($n = 371$, 28%) war für die verschiedenen Werte kein Trend hinsichtlich der Überlebenswahrscheinlichkeiten erkennbar. Bei den paarweisen Gruppenvergleichen mit dem Logrank-Test lagen statistisch signifikant unterschiedliche Überlebenswahrscheinlichkeiten vor, doch unterschieden sich die Überlebenswahrscheinlichkeiten der Gruppen „0-2 cm“ und „3-7“ über den zeitlichen Verlauf maximal mit 0,12 (mediane Überlebenszeiten: 79 und 72 Monate). Daher wurden beide Gruppen zusammengelegt. Mit der „Minimum p-value“-Methode wurde als einziger Cutpoint „7 cm“ identifiziert (Logrank-Test: $p_{ad} \leq 0{,}0001$). Im univariaten Cox-Modell erreichten die metrische Skalierung und die dichotome Einteilung „0-7 cm“ vs. „$>$ 7 cm“ p-Werte $\leq 0{,}0001$ (Wald-Test, vgl. Tabellen 4.3 und 4.6).

Der New CML-Score

Der New CML-Score war für 1279 der 1329 Patienten berechenbar. Von den 1279 gehörten 43% zur Niedrigrisikogruppe, 45% zur mittleren Risikogruppe und 12% zur Hochrisikogruppe. Die medianen Überlebenszeiten betrugen 98, 67 und 43 Monate. Weil der New CML-Score mit den von Hasford et al. [42] definierten Risikogruppengrenzen (vgl. Abschnitt 2.2) inzwischen mit Hilfe unabhängiger Stichproben validiert [13, 18, 43, 64, 90] wurde und zudem Bonifazi et al. [19] zeigten, dass innerhalb der Hochrisikogruppe Patienten durch das Erreichen einer kompletten ZR keinen Überlebensvorteil gegenüber Patienten ohne ZR gewinnen, sollte an der Score-Berechnung und bei den Risikogruppengrenzen nichts verändert werden. Wie man gemäß Abschnitt 3.4.3 erwarten konnte, lagen die p-Werte im univariaten Cox-Modell sowohl im Falle der metrischen Risikowerte als auch bei Einteilung in drei Risikogruppen unter 0,0001 (Wald-Test, vgl. Tabellen 4.3 und 4.6).

Die Überprüfung der Annahme proportionaler Hazardfunktionen

Zur Überprüfung der Annahme proportionaler Hazardfunktionen als Anwendungsvoraussetzung für das Cox-Modell, wurde das univariate Cox-Modell einer jeden kategorialen Baselinevariable X um den Wechselwirkungsterm $X \times \ln t$ erweitert (vgl. Abschnitt 2.11). Der Wechselwirkungsterm erwies sich bei keiner Kovariablen als statistisch signifikant (Wald-Test, $\alpha = 0,05$), weswegen die Annahme proportionaler Hazardfunktionen nicht verworfen wurde. Auch verliefen die Kurven $\ln(-\ln \hat{S}_k(t))$ versus $\ln t$ zu den $k = 1, \ldots, K$ Kategorien der untersuchten Variablen in allen Fällen annähernd parallel. Eine Untersuchung der Residuen nach Barlow und Prentice [14] (Formel 2.12) bei den metrischen Kovariablen aus Tabelle 4.3 ergab für das Alter keinen Ausreißer und für die übrigen Baselinevariablen 2-4 Ausreißer, d.h. Residuen deren Betrag im Vergleich zu den anderen Residuen deutlich weiter entfernt von 0 aber auch von den Residuen der benachbarten Ränge waren. Bei den Ausreißern handelte es sich vornehmlich um Patienten, die trotz hoher Parameterwerte bei Diagnose relativ lange lebten. Die Parameterwerte an sich waren jedoch nicht ungewöhnlich, so dass keine Patienten von der weiteren Analyse ausgeschlossen wurden. Insgesamt boten die Bilder der Residuenverteilungen keinen Anlass, die Annahme proportionaler Hazardfunktionen für eine der Variablen zu verwerfen.

Zusammenfassung der univariaten Analyse des Einflusses der Baselinevariablen auf die Überlebenszeit

Während hinsichtlich des Geschlechts kein statistisch signifikanter Einfluss auf die Überlebenszeit festgestellt wurde, beobachtete man zwischen dem Hauptzielparameter und jeder der acht metrischen Baselinevariablen ausnahmslos einen statistisch signifikanten Zusammenhang (Wald-Test, $\alpha = 0{,}05$). Weiter wurde mit Hilfe der „Minimal p-value"-Methode bei jeder der metrischen Variablen genau ein Cutpoint identifiziert, welcher für die jeweilige Variable zwei Gruppen definierte, die sich in Bezug auf die Überlebenswahrscheinlichkeiten am meisten statistisch signifikant unterschieden (Logrank-Test, p-Werte adjustiert, $\alpha = 0{,}05$). Bei Alter und Hämoglobin wurde bevorzugt, dem Cutpoint-Vorschlag gemäß der Daten anstatt einer Einteilung nach sog. „runden Werten" zu folgen. Durch die Betrachtung der Kaplan-Meier-Kurven einzelner Werte(bereiche) wurden - zusätzlich zum Cutpoint nach der „Minimal p-value"-Methode - weitere Unterteilungen der Variablen Hämoglobin, Blasten, Basophile und Eosinophile gefunden, die sich deutlich von den anderen beiden Gruppen unterschieden. In den Tabellen 4.4 - 4.6 ist dies anhand der Koeffizienten $\hat{\beta}$ und der Unterschiede in der medianen Überlebenszeit von ≥ 23 Monaten im Vergleich zur nächstgelegenen Gruppe erkennbar. Dass die Überlebensunterschiede in diesen vier Fällen nur mit dem Logrank-Test ohne die p-Wert-Adjustierung für multiples Testen statistisch signifikant war, dürfte nach Betrachtung der Kaplan-Meier-Kurven eher an der zu geringen Power der kleinen Gruppen als an einem nicht tatsächlich vorhandenen Überlebensunterschied gelegen haben, bleibt aber Spekulation. Es wurde entschieden, trotz kleiner Fallzahl in einer der Gruppen, die vier alternativen Skalierungen mit den drei Kategorien zu berücksichtigen.
Mit Ausnahme von Hämoglobin galt für alle metrischen Kovariablen die Tendenz, je höher der Variablenwert, desto ungünstiger die Überlebenswahrscheinlichkeiten. Bei der Definition der kategorialen Variablen wurde als Referenzkategorie im Falle des Hämoglobins die Gruppe mit den höchsten Werten gewählt, wodurch auch hier die Koeffizientenschätzer positiv wurden. Die bei Entwicklung des New CML-Scores gefundenen Cutpoints für Alter, Thrombozyten und Basophile sorgten auch in der vergrößerten Lernstichprobe für eine statistisch signifikante Differenzierung von Überlebenswahrscheinlichkeiten.

4.2.2 Die zeitabhängige Kovariable zytogenetische Remission

Bei der Untersuchung des besten Remissionsresultates unter IFN-α-Therapie wurde festgestellt, dass das Erreichen einer geringen oder minimalen Remission im Vergleich zum Ergebnis „keine Remission" nicht zu statistisch signifikant unterschiedlichen Überlebenswahrscheinlichkeiten führte (Mantel-Byar-Test, s. Abschnitt 2.8.2). Dagegen zeitigte das Erreichen einer partiellen ZR gegenüber den Patienten mit geringerem Remissionsgrad statistisch signifikant höhere Überlebenswahrscheinlichkeiten (Mantel-Byar-Test, $p \leq 0{,}0001$). Von den 174 Patienten, für die eine partielle ZR verzeichnet wurde und die ab dem Zeitpunkt ihrer partiellen ZR von der Gruppe „geringer als partielle Remission" in die Gruppe „partielle Remission" wechselten, verstarben 33 (19% von 174). Zusätzlich zu den früher beschriebenen Zensierungsgründen für die Überlebenszeit, wurde hier auch die Überlebenszeit der 69 Patienten mit kompletter ZR ab dem Zeitpunkt ihrer kompletten Remission zensiert (49% von 141 Zensierungen). Unter den 629 Patienten, bei welchen keine partielle ZR beobachtet wurde, gab es 292 Todesfälle (46%). Wegen des Erreichens einer kompletten Remission wurde die Überlebenszeit von 58 Patienten zensiert (17% von 337 Zensierungen). Abbildung 4.3 beschreibt mit Hilfe von Simon-Makuch-Kurven die sich deutlich unterscheidenden Überlebenswahrscheinlichkeiten der beiden Gruppen. Um der Kurve zu den Patienten mit partieller ZR durch eine gewisse Fahlzahl Stabilität zu verleihen, wurden die Überlebenswahrscheinlichkeiten der beiden Kurven erst ab der medianen Zeit bis

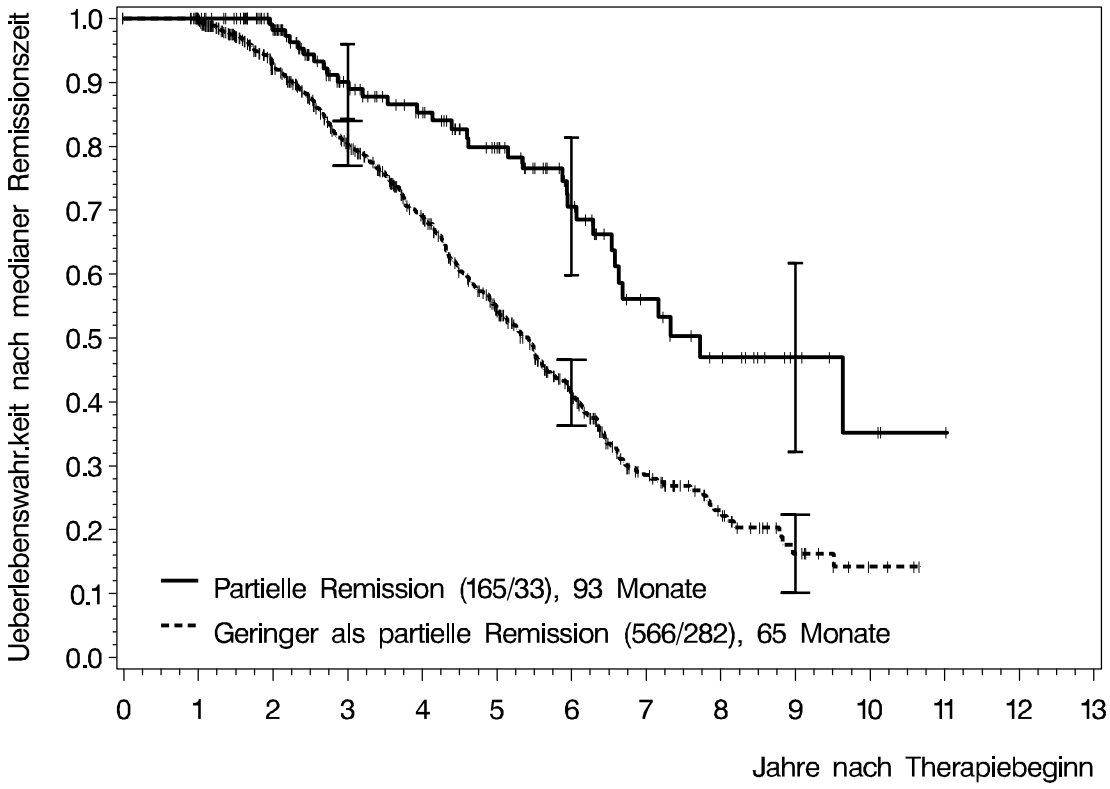

Abbildung 4.3: Simon-Makuch-Kurven mit ab der medianen Zeit bis zum Erreichen einer partiellen Remission (330 Tage) geschätzten Überlebenswahrscheinlichkeiten von Patienten mit partieller Remission im Vergleich zu Patienten mit geringerem Remissionsgrad. Von den 803 Patienten der Lernstichprobe wurden 731 länger als 330 Tage beobachtet. Eine partielle zytogenetische Remission erreichten 165 der 731 Patienten (23%). Die Legende „(165/33), 93 Monate" bedeutet: Zum Zeitpunkt des letzten Datenstandes hatten 165 Patienten eine partielle Remission erreicht, 33 waren unter diesem Status verstorben. Die mediane Überlebenszeit betrug 93 Monate. Analoges gilt für die Legende der anderen Gruppe. Bei Berechnung der Überlebenswahrscheinlichkeiten ab 330 Tagen zählten 78 Patienten, deren Remissionszeit unter und deren Überlebenszeit über 330 Tagen lag, bereits zu Beginn zu der Kurve „partielle Remission". Die 87 mit Remissionszeiten nach 330 Tagen trugen erst ab ihrem Remissionszeitpunkt zur Schätzung der Überlebenswahrscheinlichkeiten dieser Kurve bei, während die bis dahin beobachteten Überlebenszeiten in die Überlebenswahrscheinlichkeiten der Kurve „geringer als partielle Remission" eingingen. Die Überlebenszeiten der 566 Patienten ohne partielle ZR dienten ausschließlich der Berechnung der unteren Kurve. In beiden Kurven wurden die Überlebenszeiten von Patienten mit kompletter ZR ab dem Zeitpunkt der kompletten Remission zensiert. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Wahrscheinlichkeit 95%-K.I. [104] berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

zum Erreichen einer partiellen ZR (330 Tage) geschätzt. Unter den gewählten Bedingungen für die Berechnung der Simon-Makuch-Kurven lag die mediane Überlebenszeit der Patienten mit partieller Remission bei 93 Monaten (Überlebenswahrscheinlichkeit nach 9 Jahren: 0,47) und bei geringerem Remissionsgrad bei 65 Monaten (nach 9 Jahren: 0,20).

Der Vergleich der Überlebenswahrscheinlichkeiten von Patienten mit partieller Remission und

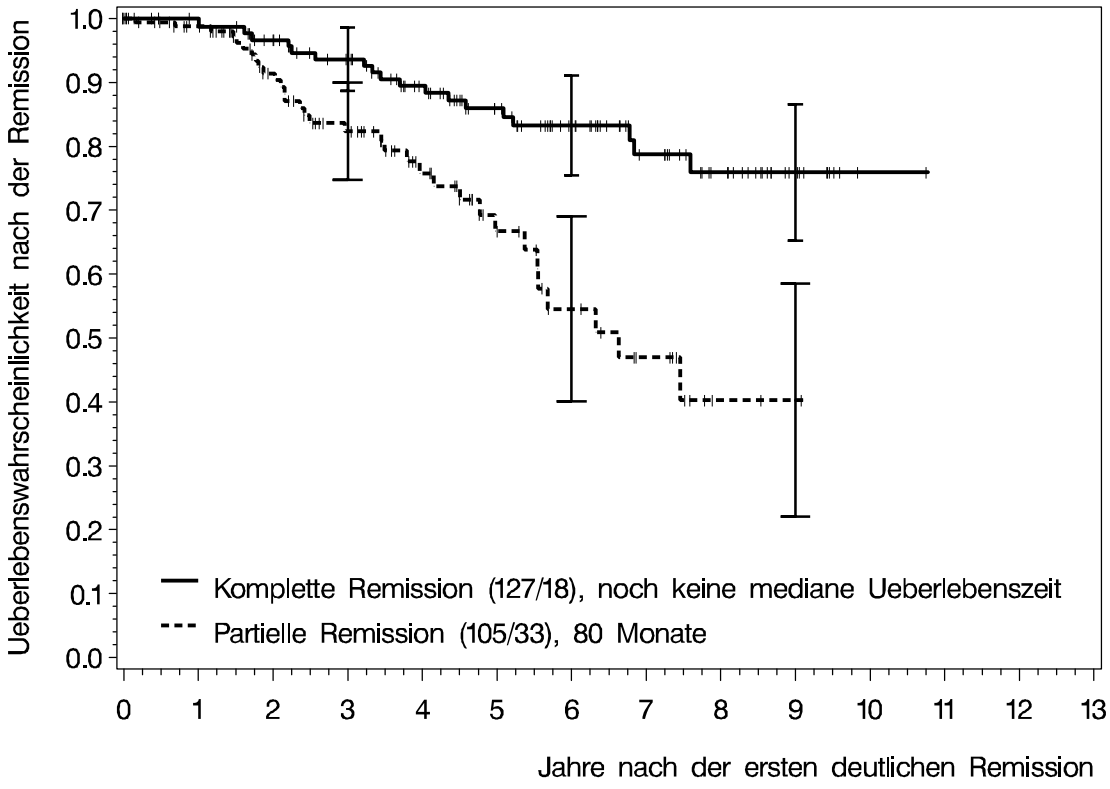

Abbildung 4.4: Simon-Makuch-Kurven mit ab dem Zeitpunkt der ersten deutlichen Remission geschätzten Überlebenswahrscheinlichkeiten von Patienten mit kompletter Remission im Vergleich zu Patienten mit partieller Remission. Von den 803 Patienten der Lernstichprobe erreichten 232 (29%) eine deutliche Remission. Eine komplette zytogenetische Remission erreichten 127 der 232 (55%). Die Legende „(127/18)“ bedeutet: Zum Zeitpunkt des letzten Datenstandes hatten 127 Patienten eine komplette Remission erreicht, 18 waren unter diesem Status verstorben. Analoges gilt für die Legende der anderen Gruppe. Bei Berechnung der Überlebenswahrscheinlichkeiten ab dem ersten Remissionstag zählten 58 Patienten, die vor ihrer ersten kompletten ZR nie im Stadium „partielle ZR“ beobachtet wurden, bereits von Beginn an zu der Kurve „komplette Remission“. Von 69 Patienten trugen die Überlebenszeiten vom Zeitpunkt ihrer partiellen Remission bis zum Zeitpunkt ihrer kompletten Remission zur Schätzung der Überlebenswahrscheinlichkeiten der Kurve „partielle Remission“ bei und seit der Beobachtung ihrer kompletten ZR zur Schätzung der erstgenannten Kurve. Die Überlebenszeiten der übrigen 105 Patienten mit partieller aber ohne kompletter ZR dienten ausschließlich der Berechnung der unteren Kurve (mediane Überlebenszeit: 80 Monate). Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Wahrscheinlichkeit 95%-K.I. [104] berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

Patienten mit kompletter Remission erbrachte für die letztgenannten einen statistisch signifikanten Überlebensvorteil (Mantel-Byar-Test, p =0,0004). Die Überlebenszeiten waren dabei ab Beginn der ersten deutlichen Remission berechnet worden. Gemäß dem Ergebnis ihrer ersten deutlichen Remission befanden sich zu Anfang 174 Patienten in der Gruppe „partielle Remission“ und 58 Patienten in der Gruppe „komplette Remission“. Ab dem Beobachtungszeitpunkt ihrer ersten kompletten Remission wechselten im Laufe der Zeit 69 der 174 Patienten von „partielle“ in die Gruppe „komplette Remission“. Von 127 Patienten, die schließlich das Stadium

„komplette ZR“ innehatten, verstarben 18 (14%). Unter den 105 Patienten mit partieller ZR als bester Remission ergaben die 33 Todesfälle einen Anteil von 31%. Die unterschiedlichen Überlebenswahrscheinlichkeiten der beiden Gruppen werden in Abbildung 4.4 jeweils ab Beginn der ersten deutlichen Remission anhand von Simon-Makuch-Kurven veranschaulicht. Nach partieller Remission betrug die mediane Überlebenszeit 80 Monate und die Überlebenswahrscheinlichkeit zum Zeitpunkt „9 Jahre“ 0,40.[7] Im Falle einer kompletten Remission wurde die mediane Überlebenszeit nach Remission nicht erreicht und die Überlebenswahrscheinlichkeit nach 9 Jahren belief sich auf 0,76.

Für die Festlegung auf eine mögliche Landmark zur Therapieentscheidung wurden die verschiedenen Zeitintervalle betrachtet, in welche die ersten partiellen zytogenetischen Remissionen der 174 Patienten fielen. Die ersten zwei Jahre wurden in acht Quartale unterteilt, das dritte Jahr halbjährlich und danach folgten Jahresintervalle. Bis zum Ende des ersten Jahres waren 96 der 174 ersten partiellen ZR (55%) eingetreten. Einen Anteil von jeweils mindestens 10% hatten das zweite Quartel (16%), das dritte (21%), das vierte (14%) und das fünfte Quartal (14%). Nach Ende des siebten Quartals (Monat 21) waren 145 (83%) erste partielle Remissionen gemeldet. Von den spätesten ersten partiellen Remissionen wurde im Jahr 5 ($n = 2$), Jahr 6 ($n = 1$) und Jahr 10 ($n = 1$) berichtet. Um zu untersuchen, ob in Abhängigkeit des Remissionszeitpunktes nach IFN-α-Therapiebeginn unterschiedliche Überlebenswahrscheinlichkeiten vorlagen, wurde für alle Patienten, deren erste partielle ZR in dasselbe Zeitintervall fiel, eine gemeinsame Kaplan-Meier-Kurve berechnet. Das Überleben ab der ersten partiellen ZR wurde bis zu den bekannten Endpunkten gemessen und bei einer kompletten ZR oder dem Ablauf von fünf Jahren zusätzlich zensiert. Die vier spätesten partiellen ZR (s.o.) wurden nicht berücksichtigt. Zwischen den elf Kaplan-Meier-Kurven zu den acht Quartalen, zwei Halbjahresintervallen und dem Jahresintervall waren weder statistisch signifikante Überlebensunterschiede (Logrank-Tests) noch ein zeitlicher Trend erkennbar. Bei den schon wegen der unterschiedlichen Intervallängen zu vergleichenden Zensierungsmustern ergaben sich für die Zeitintervalle keine statistisch relevanten Differenzen. Eine eventuelle Bedeutung des Zeitpunktes einer partiellen ZR war bei Einführung des Cutpoints „21 Monate nach Therapiebeginn“ zu erkennen. Bei partieller ZR bis zum Zeitpunkt „21 Monate“ verstarben danach 22 von 145 Patienten (15%) und bei partieller ZR nach 21 Monaten 11 von 29 Patienten (38%). Der Logrank-Test zum Vergleich der Überlebenswahrscheinlichkeiten zeitigte allerdings kein statistisch signifikantes Ergebnis ($p = 0{,}1213$). Inwieweit der Grenzzeitpunkt „21 Monate“ bei der partiellen ZR später in einem Prognosesystem für die Gesamtüberlebenswahrscheinlichkeiten bedeutsam sein könnte, wurde in Abschnitt 4.4.4 untersucht.

Bei den 127 kompletten Remissionen fielen 37 (29%) der ersten Feststellungen ins erste Therapiejahr. Ein Minimum von 10% aller ersten kompletten Remissionen wurden im dritten Quartal (10%), im vierten (13%), im fünften (10%) und im siebten Quartal (11%) beobachtet. Weitere 13% waren es im ersten Halbjahr des dritten Jahres und 10% im vierten Jahr. Verglichen mit der partiellen Remission, waren bis zum Ende des siebten Quartals 57% ($n = 72$) der ersten kompletten Remissionen eingetreten; 83% (n =106) waren mit dem Ende des dritten Jahres verzeichnet. Im fünften, sechsten und siebten Jahr lagen noch vier, drei und eine erste komplette Remission. Analog dem Vorgehen bei der partiellen Remission, wurden die Zeitintervalle zu den ersten kompletten ZR hinsichtlich Überlebenszeit und Zensierungsmuster verglichen. Die entsprechen-

[7] Die Unterschiede in den Überlebenswahrscheinlichkeiten unter partieller ZR im Vergleich zu Abbildung 4.3 ergaben sich aus dem prinzipiell verschiedenen Beobachtungsbeginn, wodurch u.a. im Verlauf unterschiedliche Patienten zeitgleich unter Risiko standen.

den Kaplan-Meier-Kurven unterschieden sich nicht statistisch signifikant (Logrank-Tests) und ihre Anordungen zeigten keinen erkennbaren Trend. Die acht Remissionen der letzten drei Jahre blieben bei den Vergleichen außen vor. Die Überlebenswahrscheinlichkeiten nach früherer oder späterer kompletter ZR blieben konstant.

Während die Landmark „6 Monate“ für die komplette ZR zu früh lag (nur sieben Patienten), besaßen 34 Patienten, die bis dahin ihre erste partielle Remission verzeichnet hatten, gegenüber den 749 Patienten (noch) ohne deutliche Remission statistisch signifikant höhere Überlebenswahrscheinlichkeiten (Kaplan-Meier-Kurven ab Ende Monat 6 und Logrank-Test, $p = 0{,}0033$). Ab Ende des neunten Therapiemonats zeigten die paarweisen Logrank-Tests im Vergleich mit den Patienten ohne deutliche ZR ($n = 682$) einen statistisch signifikanten Überlebensvorteil zugunsten von partieller ($n = 65$, $p \leq 0{,}0001$) bzw. von kompletter Remission ($n = 20$, $p = 0{,}0129$). Erst ab Ende des zweiten Jahres zeitigten die Überlebenswahrscheinlichkeiten der Kaplan-Meier-Kurve zu den nun 74 Patienten mit bis dahin erster kompletter ZR statistisch signifikant günstigere Werte auch gegenüber den Überlebenswahrscheinlichkeiten der jetzt 109 Patienten mit partieller ZR (Logrank-Test, $p = 0{,}0364$).

Als sinnvollster möglicher Entscheidungszeitpunkt für oder gegen Beibehaltung einer IFN-α-basierten Therapie in Abhängigkeit vom Ergebnis einer deutlichen Remission ergab sich auf Basis der vorliegenden Daten die Landmark „21 Monate“ (vgl. Abbildung 4.5). Von den insgesamt 232 Patienten mit deutlicher ZR hatten bis Ende Monat 21 192 Patienten (83%) ihre erste deutliche ZR erreicht. Während der Informationszugewinn über deutliche Remissionen von Quartal zu Quartal bis Ende Monat 21 jeweils mehr als 7% aller insgesamt beobachteten deutlichen Remissionen ausmachte, sank er danach auf 3% und weniger ab, obwohl ab Jahr 3 die Zeiträume auf Jahresintervalle anwuchsen. Zudem betrug bei 38 Verstorbenen die aus allen 803 Patienten geschätzte Überlebenswahrscheinlichkeit Ende des 21. Monats noch 0,95, war aber drei Monate später mit 18 zusätzlich Verstorbenen bereits auf 0,92 gefallen.
Bei 120 Patienten war die erste deutliche ZR bis Ende Monat 21 eine partielle (62,5% von 192) und bei 72 Patienten eine komplette ZR (37,5%). In beiden Gruppen war die Überlebenszeit von jeweils sechs Patienten kürzer als 21 Monate beobachtet worden, wobei jeweils drei Patienten wegen einer allogenen SZT in 1. chronischer Phase und die jeweils übrigen drei wegen anderer Gründe zensiert worden waren. Neununddreißig Patienten, die bis Ende Monat 21 eine partielle Remission aufwiesen, erzielten später eine komplette Remission. Die Kaplan-Meier-Kurve ab der Landmark „21 Monate“ schätzte für die 66 verbliebenen[8] Patienten mit kompletter ZR zum Zeitpunkt „neun Jahre“ eine Überlebenswahrscheinlichkeit von 0,72; 14 Patienten wurden länger beobachtet, wovon keiner danach verstarb. Im Falle der 114 Patienten mit partieller Remission lag nach neun Jahren die Überlebenswahrscheinlichkeit bei 0,57; sieben Patienten wurden länger beobachtet, keiner war danach verstorben. Keine deutliche ZR bis Ende des 21. Therapiemonats wurde bei 611 (76% der 803 Patienten) registriert. Für die Landmarkanalyse verblieben 501 Patienten, 110 Patienten standen weniger als 21 Monate unter Beobachtung. Von diesen 110 Patienten erfuhren 53 (48%) zuvor eine SZT, 19 (17%) wurden aus anderen Gründen zensiert und 38 (35%) waren bereits verstorben. Die mediane Überlebenszeit der 501 Patienten wurde nach 66 Monaten erreicht. Die Überlebenswahrscheinlichkeit nach neun Jahren betrug 0,20; 15 Patienten lebten länger, wovon noch zwei verstarben. Der Statistik zum Logrank-Test über alle drei Kurven entsprach ein p-Wert $< 0{,}0001$. Das signifikante Ergebnis beruhte auf dem Unterschied zwischen den Patienten ohne deutliche ZR und den Patienten mit partieller ZR bzw. mit

[8] Ohne die sechs kürzer Beobachteten. Analog bei den Patienten mit partieller ZR: 114 statt 120.

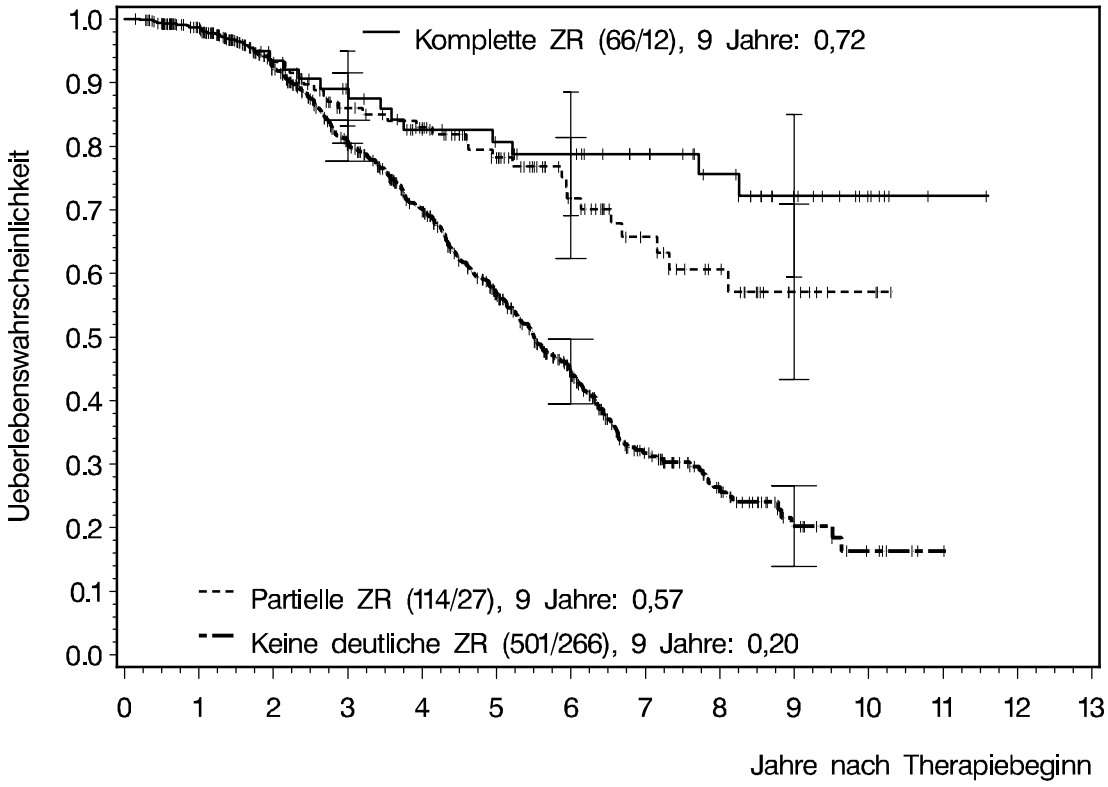

Abbildung 4.5: Kaplan-Meier-Kurven ab der Landmark „21 Monate“ mit geschätzten Überlebenswahrscheinlichkeiten in drei zytogenetischen Remissionsstufen. Bis zur Landmark „21 Monate“ wurden die Überlebenswahrscheinlichkeiten aller 803 Patienten der Lernstichprobe gemeinsam geschätzt. Ab Ende des 21. Therapiemonats wurden die verbliebenen 681 Patienten je nach Remissionsergebnis auf die drei Gruppen verteilt. Die Legende „(66/12), 9 Jahre: 0,72“ bedeutet: Von 66 Patienten sind 12 verstorben. Die geschätzte Wahrscheinlichkeit, bis Ende des 9. Jahres zu überleben, lag bei 0,72. Die beiden anderen Legenden sind analog zu verstehen. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Wahrscheinlichkeit mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

kompletter ZR. Beide paarweisen Vergleiche ergaben p-Werte $< 0{,}0001$. Der Vergleich „partielle ZR“ versus „komplette ZR“ war nicht statistisch signifikant.
Eine komplette wie auch schon eine partielle ZR innerhalb von 21 Monaten prognostizierte damit relativ hohe Überlebenswahrscheinlichkeiten bei mit IFN-α-behandelten CML-Patienten.

Die zytogenetische Remission als zeitabhängige Kovariable im univariaten Cox-Modell

Um den Einfluss der zytogenetischen Remission auf die Überlebenszeit mit dem Cox-Modell zu untersuchen, erhielten die partielle wie die komplette zytogenetische Remission jeder einen eigenen, zeitabhängigen Faktor, dessen Wert bei einem Patienten im Cox-Modell von 0 auf 1 gesetzt wurde, sobald die Überlebenszeiten der noch unter Beobachtung Stehenden zum ersten Mal über der jeweiligen Remissionszeit des Patienten lagen. Bei Patienten, die sich zunächst im Stadium „partielle ZR“ befanden, wurde zum Zeitpunkt der ersten beobachteten kompletten ZR

der Faktor zur partiellen Remission zurück auf 0 gesetzt.
Für den Faktor zur partiellen Remission ergab sich der Effektschätzer -0,8289 mit einer Standardabweichung von 0,1843. Bei Datenbankschluss hatten 105 Patienten eine partielle ZR als beste Remission erreicht, 33 Patienten waren verstorben und das geschätzte relative Risiko lag für einen Patienten mit partieller ZR im Vergleich zu einem Patienten ohne erste partielle ZR bei 0,437. Das Ergebnis zum Faktor für die komplette ZR zeigte einen Effektschätzer von -1.8601 mit einer Standardabweichung von 0,2463. Von 127 Patienten, für die eine erste komplette ZR registriert wurde, waren 18 verstorben; das relative Risiko im Vergleich zu den Patienten ohne eine komplette ZR wurde zuletzt mit 0,156 geschätzt. Die Wald-Statistik zu den beiden Faktoren hatte den Wert 71,8039, was einem p-Wert $< 0{,}0001$ entspricht (χ^2-Verteilung mit zwei Freiheitsgraden). Partielle und komplette ZR zeigten auch im Cox-Modell einen statistisch signifikanten Einfluss auf die Überlebenszeit. Die negativen Werte der Effektschätzer und die relativen Risiken unter 1 sprachen für höhere Überlebenswahrscheinlichkeiten bei Patienten mit Remission.
Mit den zeitabhängigen Faktoren konnte nicht mehr von über den gesamten Zeitraum gleichbleibenden, proportionalen Hazardfunktionen zwischen zwei Patienten ausgegangen werden. Wohl aber wird beim Cox-Modell mit zeitabhängigen Kovariablen für alle Kovariablen ein konstanter, zeitunabhängiger Effekt angenommen. Um die Konstanz der geschätzten Effekte zu untersuchen, wurden dem Cox-Modell die Wechselwirkungsterme der beiden Faktoren mit $\ln t$ beigefügt. Keiner der Wechselwirkungsterme erwies sich als statistisch signifikant: Die beobachteten Remissionszeiten einer ersten partiellen oder kompletten ZR waren im Verhältnis zu den beobachteten Überlebenszeiten eher kurz, was die Tatsache, dass die Effektschätzer erst mit der Zahl der sich nach und nach einstellenden ersten Remissionen an Bedeutung gewinnen können relativierte.
Bei beiden Faktoren lagen die Beträge der Residuen nach Barlow und Prentice [14] (Formel 2.12) unter 1. Höhere Werte zwischen 0,6 und 1 ergaben sich für 26 Patienten, die nach einer ersten partiellen ZR innerhalb von 5,5 Jahren verstarben bzw. für die 18 Patienten, die nach erster kompletter Remission verstarben. Die Bilder der Residuen gaben keinen Anlass, an den Werten einzelner Patienten oder an der Anpassung des Cox-Modells an die Daten zu zweifeln.

4.3 Zusammenhänge zwischen den Kovariablen

4.3.1 CART: Suche nach Zusammenhängen zwischen Werten verschiedener Baselinevariablen im Hinblick auf die Überlebenswahrscheinlichkeiten

Bereits im Rahmen der Entwicklung des New CML-Scores war ein auf Klassifikationsbäumen (CART, vgl. Abschnitt 2.9.1) beruhendes Modell als mögliche Alternative untersucht worden. Ziel der Anwendung von CART war nun kein neuer Modellvorschlag, sondern die Suche nach Zusammenhängen zwischen Wertekonstellationen verschiedener Baselinevariablen und unterschiedlichen Überlebenswahrscheinlichkeiten in der nun gegebenen Lernstichprobe. Die dabei gemachten Entdeckungen sollten später als Wechselwirkungsterme in multiplen Cox-Modellen geprüft werden.

In der Lernstichprobe besaßen 1231 Patienten vollständige Werte zu den neun Variablen, die bei der CART-Prozedur berücksichtigt werden sollten. Neben den acht metrischen Größen aus Tabelle 4.2 war „Geschlecht“ der einzige kategoriale Parameter.[9] Die Abbruchkriterien für jede weitere Partitionierung durch CART waren erfüllt, sobald entweder bei den Logrank-Tests keine

[9] Thrombozyten, Leukozyten und Hämoglobin wurden gerundet wie in Abschnitt 4.2.1 beschrieben. Bei Hämoglobin wurden die besonderen Grenzen 11,3 bei den Frauen und 13,5 g/dl bei den Männern beachtet, d.h. bei 11,4 bereits aufgerundet bzw. bei 13,5 abgerundet.

Unterteilung mit $p_{ad} \leq \alpha = 0{,}1$ oder keine zwei Gruppen mit $n \geq 36$ mehr gefunden werden konnten.[10]

An der Baumwurzel wurde die Unterteilung in Milzgröße ≤ 7 ($n = 878$) versus > 7 cm ($n = 353$) als Variable und Grenze mit der höchsten χ^2-Statistik identifiziert, $p_{ad} < 0{,}0001$. Vier Patienten mit einer Milzgröße > 7 cm und Thrombozyten $> 1350 \times 10^9$/l wurden zu den 37 mit Milzgröße ≤ 7 cm und Thrombozyten $> 1350 \times 10^9$/l hinzugefügt. Damit war die Gruppe
a) Thrombozyten $> 1350 \times 10^9$/l ($n = 41$, mediane Überlebenszeit: 42 Monate)
unabhängig von jeder anderen Kovariablen definiert.
Die übrigen 349 Patienten mit Milzgröße > 7 cm ließen sich auf drei hinsichtlich der Überlebenswahrscheinlichkeiten statistisch signifikant unterschiedliche Gruppen reduzieren:
b) Eosinophile $> 2\%$ im p.B. und Alter > 49 Jahre ($n = 66$, mediane Überlebenszeit: 34 Monate),
c) [Eosinophile $\leq 2\%$ im p.B. und Alter > 59 Jahre] ODER [Eosinophile $> 2\%$ im p.B. und Alter ≤ 49 Jahre] ($n = 129$, mediane Überlebenszeit: 53 Monate) und
d) Eosinophile $\leq 2\%$ im p.B. und Alter ≤ 59 Jahre ($n = 154$, mediane Überlebenszeit: 74 Monate).
Auch bei den 841 Patienten mit Milzgröße ≤ 7 cm und Thrombozyten $\leq 1350 \times 10^9$/l ergaben sich durch diese Definition mit Hilfe von Eosinophilen und Alter bzgl. der Überlebenswahrscheinlichkeiten drei paarweise statistisch signifikant unterschiedliche Gruppen (Logrank-Test, alle $p < 0{,}01$). Viel stärker aber diskriminierte die 841 die durch CART definierte Einteilung in die drei Gruppen
e) Hämoglobin > 13 g/dl und Leukozyten $\leq 50 \times 10^9$/l ($n = 135$, mediane Überlebenszeit noch nicht erreicht),
f) [Hämoglobin ≤ 13 g/dl und (Alter ≤ 49 Jahre ODER Leukozyten $\leq 50 \times 10^9$)] ODER [Leukozyten $> 50 \times 10^9$/l und Alter ≤ 49 Jahre] ($n = 397$, mediane Überlebenszeit: 91 Monate)
g) Leukozyten $> 50 \times 10^9$/l und Alter > 49 Jahre ($n = 309$, mediane Überlebenszeit: 67 Monate).
Unter den 349 mit Milzgröße > 7 cm offenbarte die Übertragung dieser Gruppendefinitionen interessanterweise, dass es nur zwei Patienten mit Hämoglobin > 13 g/dl und Leukozyten $\leq 50 \times 10^9$/l gab. Die noch relativ günstigen Blutwerte stehen für ein frühes Stadium der Krankheit und führen i.d.R. anscheinend noch nicht zu einem starken Anschwellen der Milz. Unter Hinzunahme der beiden Patienten mit Milzgröße > 7 cm könnte die Gruppe e) damit unabhängig von der Milzgröße definiert werden. Die Überlebenswahrscheinlichkeiten der Gruppen f) und g) unterschieden sich auch bei den 349 ($p = 0{,}0050$).
Die beiden Gruppen d) und g) zusammengelegt, blieben 1190 Patienten in fünf Gruppen unterteilt, deren Wahrscheinlichkeiten sich bei den Logrank-Tests mit p-Werten $< 0{,}0050$ unterschieden. An diesem Ergebnis änderte sich nichts, wenn man zur Einteilung aller 1231 die 41 Patienten mit Thrombozyten $> 1350 \times 10^9$/l der Gruppe c) hinzufügte. Mit den Definitionen der Gruppen a) bis g) standen, auf Basis von CART, Vorschläge zu im multiplen Cox-Modell beachtenswerten Wechselwirkungen aus zwei oder drei zeitunabhängigen Kovariablen zur Verfügung.

4.3.2 Korrelationen zwischen den Baselinevariablen

Die Normalverteilungsannahme für die metrischen Variablen war nicht erfüllt (Shapiro-Wilk-Test [93, 101], alle p-Werte $< 0{,}0001$). Dementsprechend wurden Zusammenhänge zwischen einer

[10] Die Wurzel aus 1231 lag knapp über 35.

kategorialen und einer metrischen Variablen mit Hilfe des U-Tests bzw. des Kruskal-Wallis-Tests untersucht und zur Korrelation zwischen zwei metrischen Variablen wurde Spearmans Korrelationskoeffizient berechnet. Den p-Wert zu Spearmans Korrelationskoeffizienten erhielt man über die dazugehörige, von SAS [96] bereitgestellte, t_{n-2}-verteilte Statistik. Korrelationen zwischen zwei kategorialen Variablen wurden durch den χ^2-Test beurteilt. Alle im folgenden erwähnten Zusammenhänge besaßen unter $\alpha = 0{,}05$ liegende p-Werte. Der Zusatz „statistisch signifkant“ wurde im folgenden Teil des Abschnitts den Komparativen und den Korrelationsangaben nicht mehr explizit hinzugefügt.

Die Frauen waren bei Diagnose älter als die Männer (Mediane: 51 vs. 48 Jahre). Bei den Patienten unter 50 Jahren betrug der Frauenanteil 37% und bei den Patienten ab 50 Jahren 48%. Ein starker Zusammenhang wurde zwischen Alter, Geschlecht und Milzvergrößerung beobachtet. Männer unter 50 Jahren hatten eine deutlichere Milzvergrößerung als jede der drei anderen Kombinationen aus Geschlecht und Altersgruppe (Mediane: 4 vs. 1 oder 2 cm). Die Frauen besaßen gegenüber den Männern höhere Thrombozytenzahlen (Mediane: 480 vs. 350 $\times$ 10^9/l). Bei Patienten bis 41 Jahre lag in 44% der Fälle der Basophilenanteil unter 3% im p.B., bei den älteren Patienten waren es 37%. Abgesehen von den Thrombozyten und Hämoglobin, war die Milzvergrößerung mit jedem der anderen Blutparameter positiv korreliert. Zwischen Hämoglobin und Milzvergrößerung bestand eine negative Korrelation; nur die Thrombozytenzahl schien keinen Einfluss auf die Milzvergrößerung zu nehmen. Eine negative Korrelation von Hämoglobin wurde auch zum einen mit den Leukozyten und zum anderen mit den Blasten im p.B. beobachtet, die negative Korrelation mit den Eosinophilen war schwächer ausgeprägt. Weiter waren die Leukozyten mit den Blasten und den Eosinophilen positiv korreliert. Bemerkenswerterweise hatten die Patienten mit extremer Thrombozytenzahl größer 1350 $\times$ 10^9/l relativ kleine Leukozytenzahlen (50% $\leq$ 50, nur 4 von 44 über 150 $\times$ 10^9/l). Die Thrombozyten waren zudem mit den Eosinophilen und den Basophilen positiv korreliert. Paarweise positive Korrelationen bestanden auch zwischen Blasten, Basophilen und Eosinophilen. Der New CML-Score war mit den Leukozyten positiv und mit Hämoglobin negativ korreliert. Frauen hatten höhere Riskowerte, die Verteilung auf die Risikogruppen unterschieden sich jedoch nicht von derjenigen der Männer.

4.3.3 Einfluss der Baselinevariablen auf die zytogenetische Remission

Zentraler Gedanke bei der Entwicklung und späteren Verwendung eines prognostischen Modells unter Einschluss einer für den Krankheitsverlauf einflussreichen zeitabhängigen Kovariablen ist das Abwarten eines bestimmten Zeit, um zu sehen, ob sich die interessierende Kovariable unter Therapie verändert und dann entsprechend dem aktualisierten prognostischen Resultat zu handeln. Theoretisch bestünde auch die Möglichkeit, Beobachtungswahrscheinlichkeiten eines Ereignisses bei der zeitabhängigen Variablen mit Hilfe eines aus Baselinevariablen entwickelten Prognosesystems statistisch zu unterscheiden. Da die zytogenetische Remission sich bereits als sehr wichtiger prognostischer Faktor herausstellte (vgl. Abschnitt 4.2.2), war es mit der im vorliegenden Fall relativ geringen Wahrscheinlichkeit, in den ersten Monaten zu versterben, jedoch erstrebenswerter, im Sinne von Verlässlichkeit und Genauigkeit der zu prognostizierenden Überlebenswahrscheinlichkeiten, das tatsächliche Remissionsergebnis bis zu einem bestimmten Entscheidungszeitpunkt abzuwarten. Daher wurde zwar - im Hinblick auf das für die Überlebenszeit zu suchende Prognosemodell - der multiple Einfluss der Baselinevariablen auf die zytogenetische Remission mit dem Cox-Modell untersucht, aber weder eine Risikogruppenbildung vorgenommen noch ein Prognosesystem entwickelt.

Gäbe es einen allgemein anerkannten Therapieentscheidungszeitpunkt in Abhängigkeit des Ergebnisses zur ZR, wäre die Modellierung der zytogenetischen Remission als abhängige Variable in einem logistischen Modell denkbar. Dieser Zeitpunkt existierte jedoch nicht. Der für den vorliegenden Datensatz „optimale“ Entscheidungszeitpunkt „21 Monate“ (vgl. Abschnitt 4.2.2) wird von den Ärzten in Anbetracht der therapeutischen Alternativen als „zu spät“ erachtet und ist auch das Ergebnis von z.T. geringer zytogenetischer Untersuchungshäufigkeit. In Anbetracht dieser Gegebenheiten wurde von der Definition einer festen Landmark abgesehen, alle Beobachtungzeiten und Ereignisse (gleichermaßen) berücksichtigt und die abhängige Variable „zytogenetische Remission“ als „Time-to-event“-Variable modelliert.

Wie die Abbildungen 4.3 und 4.5 und v.a. die Ergebnisse aus Abschnitt 4.2.2 nahelegen, kann bereits durch das Erreichen einer partiellen ZR prognostisch eine Überlebenszeitverlängerung erreicht werden. Ausgehend von mehr als 35% Ph-positiven Metaphasen, bildet die „partielle Remission“ für jene Patienten ein Zwischenstadium, welche später eine komplette ZR erreichen.[11] Insofern war für die Baselinevariablen schon per definitionem ein Zusammenhang zwischen ihrer prognostischen Relevanz hinsichtlich des Eintretens einer partiellen ZR und hinsichtlich des Eintretens einer kompletten ZR zu erwarten. Aus diesen Gründen und wegen der Höhe der Fallzahl bot es sich an, zur Untersuchung eines Zusammenhangs zwischen den Baselinevariablen und der zytogenetischen Remission zunächst die „Time-to-event“-Variable „erstes Beobachten einer deutlichen zytogenetischen Remission“ (803 Patienten, 232 Ereignisse, davon 174 partielle und 58 komplette ZR) zu wählen. Andererseits sollte das Ergebnis statistisch signifikant unterschiedlicher Überlebenswahrscheinlichkeiten von Patienten mit partieller ZR und solchen mit kompletter ZR nicht unberücksichtigt bleiben.[12] Daher wurde der Einfluss der Baselinevariablen später zusätzlich für „erstes Beobachten einer partiellen ZR“ (803 Patienten, 174 Ereignisse) und „erstes Beobachten einer kompletten ZR“ (803 Patienten, 127 Ereignisse) betrachtet. Im folgenden findet sich zunächst eine ausführliche Darstellung der prognostisch relevanten Einflüsse der Baselinevariablen auf das erste Beobachten einer deutlichen ZR.

Univariater Einfluss der Baselinevariablen

Von den Baselinevariablen offenbarten sich Alter, Hämoglobin, die Milzvergrößerung, Leukozyten, Blasten, Basophile und Thrombozyten als statistisch signifikante prognostische Faktoren für die Beobachtung einer ersten deutlichen Remission. Erwartungsgemäß gingen tendenziell höhere Variablenwerte mit kleineren Remissionswahrscheinlichkeiten einher, umgekehrt nur im Falle von Hämoglobin. Zur Bildung kategorialer Variablen als Alternative zur metrischen Skalierung, wurden mit Hilfe von Kaplan-Meier-Kurven die Beobachtungswahrscheinlichkeiten bei verschiedenen Baselinewerten untersucht. Variablenwerte mit vergleichbaren Beobachtungswahrscheinlichkeiten wurden zu Gruppen zusammengefasst. Die Suche nach Cutpoints mit Hilfe der „Minimal p-value“-Methode ergab für das Alter die Gruppierung ≤ 43 Jahre (21 Monate: 0,37)[13]

[11] Im Falle von 69 Patienten wurden zuerst Zytogenetiken mit dem Resultat „partielle ZR“ und danach mit dem Resultat „komplette ZR“ verzeichnet. Bei den übrigen 58 Patienten mit kompletter ZR wurde vermutlich während der Zeitspanne in partieller Remission keine Zytogenetik entnommen. Die Kaplan-Meier-Kurven der 69 versus der 58 Patienten wiesen ab dem Zeitpunkt der Feststellung der ersten kompletten ZR kaum unterscheidbare Überlebenswahrscheinlichkeiten auf. Ob vor der ersten kompletten ZR eine partielle ZR gemessen wurde, hatte erwartungsgemäß keinen statistischen Einfluss.

[12] Siehe Abschnitt 4.2.2.

[13] Die in Klammern angegebenen Wahrscheinlichkeiten beziehen sich auf die Beobachtungswahrscheinlichkeit einer deutlichen Remission nach 21 Therapiemonaten in der jeweils definierten Patientengruppe.

versus > 43 Jahre (0,26). Trotz unterschiedlicher Referenzbereiche konnte für Frauen wie Männer bei Hämoglobin der gemeinsame Cutpoint 12,8 g/dl gefunden werden ($\leq$ 12,8 g/dl: 0,23, > 12,8 g/dl: 0,39). Andere Gruppierungen waren: Milzvergrößerung $\leq$ 0 cm (0,41) versus > 0 cm (0,21), Leukozyten $\leq 50 \times 10^9$/l (0,47) versus $> 50 \times 10^9$/l (0,24), Blasten $\leq$ 1% (0,27) versus > 1% (0,20), Basophile $\leq$ 5% (0,34) versus > 5% (0,19) und Thrombozyten $\leq 800 \times 10^9$/l (0,34) versus $> 800 \times 10^9$/l (0,07). Vor Anwendung der „Minimal p-value"-Methode war Basophile die einzige Variable, für die über die Betrachtung von Kaplan-Meier-Kurven mit den drei Gruppen $\leq$ 5% (0,34), 6-11% (0,20) und > 11% (0,10) zusätzlich eine nicht dichotome Einteilung entdeckt wurde, deren Kategorien deutlich unterschiedliche Beobachtungswahrscheinlichkeiten aufwiesen. Mit seinen drei Risikogruppen definierte der New CML-Score auch für die Beobachtungswahrscheinlichkeiten einer ersten deutlichen ZR drei paarweise statistisch signifikant unterschiedliche Prognosegruppen (alle p-Werte der paarweisen Logrank-Tests < 0,01; Niedrigrisikogruppe: 0,36, mittlere Risikogruppe: 0,28, Hochrisikogruppe: 0,11). Eine auf die Anwendung der „Minimal p-value"-Methode gestützte Neudefinition der Risikogruppengrenzen führte zu dem Vorschlag, vier Risikogruppen hinsichtlich der Beobachtungswahrscheinlichkeiten zu unterscheiden.[14] Wie für alle hier gefundenen Risikogruppen, sind statistisch signifikante Unterschiede hinsichtlich der ZR Hypothesen kreierend und wären in unabhängigen Validierungsstichproben zu überprüfen.

Multipler Einfluss der Baselinevariablen

Analyse mit CART

Bei der Analyse mit CART ergab sich für die neun Baselinevariablen in Bezug auf die abhängige „Time-to-event"-Variable „Beobachtung einer ersten deutlichen Remission" die Einteilung Leukozyten $\leq 50 \times 10^9$/l versus $> 50 \times 10^9$/l als diejenige mit dem kleinsten p-Wert an der Baumwurzel. Insgesamt wurden sechs Gruppen gefunden, die sich in drei Prognosegruppen mit statistisch signifkant unterschiedlichen Beobachtungswahrscheinlichkeiten einteilen ließen (p-Werte der drei paarweisen Logrank-Tests $\leq$ 0,0005). Die geringsten Beobachtungswahrscheinlichkeiten (0,12 zum Zeitpunkt „21 Monate") und 38 deutliche Remissionen hatten 302 Patienten mit Leukozyten $> 50 \times 10^9$/l und ENTWEDER [Alter $\leq$ 41 Jahre aber Thrombozyten $> 700 \times 10^9$/l] ODER [Alter > 41 Jahre und Milzvergrößerung > 0 cm]. Für die mittlere Gruppe von 113 Patienten wurden 31 deutliche Remissionen beobachtet (Wahrscheinlichkeit nach 21 Monaten: 0,28). Sie definierte sich durch Leukozyten $> 50 \times 10^9$/l und zugleich Alter > 41 Jahre sowie Milzvergrößerung = 0 cm. Die Prognosegruppe mit den höchsten Beobachtungswahrscheinlichkeiten (0,44 nach 21 Monaten) und 142 deutlichen Remissionen bestand aus 328 Patienten mit Leukozytenzahlen $\leq 50 \times 10^9$/l ODER [Leukozyten $> 50 \times 10^9$/l aber Alter $\leq$ 41 Jahre und zudem Thrombozyten $\leq 700 \times 10^9$/l].

Analyse mit multiplem Cox-Modell

Neben den im vorliegenden Abschnitt beschriebenen Kategorisierungen der Baselinevariablen wurden die in den Tabellen 4.3 - 4.6 angeführten Skalierungen berücksichtigt. Das beste prognostische Modell wurde nach dem Selektionsverfahren aus Abschnitt 2.10 gewählt. Für die 743 Patienten (211 deutliche ZR) mit Daten zu allen neun Baselineparametern bestand das beste Modell aus dichotomen Variablen zu Alter, Thrombozyten, Leukozyten, Milzvergrößerung und

[14] Die Grenzen lagen bei $\leq$ 500, $\leq$ 1150 und $\leq$ 1610, die Beobachtungswahrscheinlichkeiten zum Zeitpunkt „21 Monate" waren: 0,42, 0,29, 0,19 und 0,02. Wäre die Prognose der zytogenetischen Remission von Relevanz, müßten diese Risikogruppen allerdings dem Vergleich mit einem speziell für die zytogenetische Remission entwickelten Prognosesystem standhalten.

Basophilen. In Tabelle 4.7 finden sich alle Informationen zum identifizierten besten Modell, welches schließlich für 768 Patienten (225 deutliche ZR) mit vollständigen Daten berechenbar war.

Tabelle 4.7: Multiple Analysen im Cox-Modell: Das beste Modell zum Einfluss der Baselinevariablen auf die Beobachtungswahrscheinlichkeiten einer ersten deutlichen zytogenetischen Remission bei den 768 Patienten mit vollständigen Daten

	Patien-tenzahl	Schätzung Koeffizient	Standard-abweichung	Walds χ^2-Statistik		p-Wert	
Variable[a]	n/dZR[b]	$\hat{\beta}$[c]	$\hat{v}$	X^2	df[d]	p	RR[e]
Alter							
Zwei Gruppen	768/225			**20,2425**	**1**	**<0,0001**	
≤43 Jahre	280/100	0	-				1
>43 Jahre	488/125	-0,6281	0,1396				0,534
Thrombozyten							
Zwei Gruppen	768/225			**19,0869**	**1**	**<0,0001**	
$\leq 800 \times 10^9$/l	651/214	0	-				1
$>800 \times 10^9$/l	117/ 11	-1,4309	0,3275				0,239
Leukozyten							
Zwei Gruppen	768/225			**16,2369**	**1**	**<0,0001**	
$\leq 50 \times 10^9$/l	193/ 90	0	-				1
$>50 \times 10^9$/l	575/135	-0,6431	0,1596				0,526
Milzvergröß.							
Zwei Gruppen	768/225			**6,8749**	**1**	**0,0087**	
0 cm	312/124	0	-				1
>0 cm	456/101	-0,4179	0,1593				0,658
Basophile							
Zwei Gruppen	768/225			**3,8754**	**1**	**0,0490**	
0-5%	546/182	0	-				1
>5%	222/ 43	-0,3443	0,1749				0,709

[a] Einheiten und Messgenauigkeit wie in Tabelle 4.2.
[b] n: Gesamtzahl der Patienten mit Daten, dZR: die Patienten mit erster deutlicher zytogenetischer Remission
[c] Der Wert „0“ steht für die mit „0“ kodierte Referenzgruppe. Die andere Gruppe wurde jeweils mit „1“ kodiert.
[d] Freiheitsgrade.
[e] Relatives Risiko: Verhältnis der geschätzten Hazardfunktion zur Hazardfunktion der Referenzkategorie.

Die negativen Koeffizienten und die relativen Risiken unter 1 stehen für die bei den höheren Variablenwerten verminderten Wahrscheinlichkeiten, eine deutliche Remission zu beobachten. Die Hinzunahme von Interaktionen zwischen den Variablen konnte die $-2\ln L(\hat{\beta})$-Statistik nicht statistisch signifkant reduzieren. Das beste Modell unter Einschluss des New CML-Scores statt seiner einzelnen Variablen verminderte die $-2\ln L(\hat{\beta})$-Statistik statistisch signifikant geringer.
Sämtliche Gruppengrenzen waren bei den univariaten Analysen bereits im Hinblick auf möglichst große Heterogenität bzgl. der Remissionswahrscheinlichkeiten definiert worden. Die Effekte so entstandener kategorialer Variablen werden im Cox-Modell meist überschätzt. Die Stärke der Überschätzung fällt in späteren Validierungsstichproben verschiedentlich aus und kann nicht

vorhergesagt werden.[15] Weil aber im vorliegenden Fall - außer bei der Milzvergrößerung - die Verwendung der kategorialen Skalierung der Variablen anstatt der metrischen Skalierung zu einer Reduktion der $-2 \ln L(\hat{\beta})$-Statistik um mehr als 4 führte, wurden die kategorisierten Variablen für das Endmodell bevorzugt.[16] Im Falle der Milzvergrößerung lag zwar im Vergleich zur Originalskalierung keine Reduktion um mehr als vier vor, doch galt ebenso wie bei den anderen metrischen Skalierungen, dass ein höherer Variablenwert nicht unbedingt ein höheres relatives Risiko zur Folge hatte.[17] Die dichotome Variable erschien auch hier sinnvoller.

Zur Überprüfung auf zeitunabhängige, konstante Proportionalität zwischen den Kategorien der einzelnen Variablen des Endmodells wurde jeweils das aus der zu untersuchenden Variablen X und dem Wechselwirkungsterm $X \times \ln t$ bestehende Cox-Modell berechnet (vgl. Abschnitt 2.11). Statistisch signifikante Wechselwirkungsterme bei Alter und Leukozyten (p-Werte $< 0{,}025$) wiesen auf eine Zeitabhängigkeit hin. Eine Betrachtung der Kaplan-Meier-Kurven und der Graphen $\ln(-\ln \hat{S}(t))$ versus $\ln t$ zu den beiden Kategorien der jeweiligen Variablen offenbarte die Ursache: In der Gruppe mit den höheren Werten (vgl. Tabelle 4.7) fanden nach etwa zwei Jahren - im Gegensatz zu vorher - im Falle der Variablen „Alter“ und „Leukozyten“ kaum noch erste deutliche Remissionen statt. Dagegen wurden jeweils in der Gruppe mit den niedrigeren Werten nach zwei Jahren noch immer einige erste deutliche Remissionen registriert.
Im multiplen Modell wurden entsprechend dem Vorschlag von Sasieni [97] (siehe Abschnitt 2.11) die beobachteten Remissionszeiten zu verschiedenen Verlaufszeitpunkten zwischen 12 und 60 Monaten zensiert. Der Absolutbetrag der geschätzten Koeffizienten zu Alter und Leukozyten erhöhte sich über die Zeit immer stärker und zeigte damit den wachsenden Unterschied zwischen den jeweiligen Kategorien der Variablen an. Während Alter im multiplen Cox-Modell von Anfang an statistisch signifikant war, stellte sich die Signifikanz bei den Leukozyten erst nach 18 Monaten ein. Auch die beiden graphischen Methoden nach Andersen [7] ließen für beide Variablen den mit der Zeit weiter zunehmenden Unterschied zwischen den Hazardfunktionen erkennen.
Die sechs Bilder der Barlow-Prentice-Residuen [14] (Formel 2.12) zu jeder der Variablen des Endmodells zeigten keine Werte mit einem über 1 liegenden Betrag. Vier Variablen hatten zwischen null und vier Ausreißern, bei den Thrombozyten waren es zehn. Die betroffenen Patienten hatten keine ungewöhnliche Wertekonstellation bei den Baselinevariablen, weswegen niemand ausgeschlossen wurde. Insgesamt durfte von einer guten Modellanpassung ausgegangen werden - trotz der Abweichung von der PH-Annahme für Alter und Leukozyten.

Für die „Time-to-event“-Variable „erste partielle ZR“ wurde bei den 743 Patienten (161 partielle ZR) hinsichtlich der Variablen exakt dasselbe Modell als bestes multiples Cox-Modell identifiziert. Wertete man nur die 117 kompletten ZR unter den 743 Patienten als Ereignis, so änderte sich das beste multiple Modell dahingehend, dass die Basophile und die Milzvergrößerung als nicht statistisch signifikant wegfielen und sattdessen die Blasten in nichtkategorisierter Originalskalierung hinzukamen.

[15] Ein Vergleich z.B. der sechs Koeffizienten des New CML-Scores [42] mit den geschätzten Koeffizienten bei 453 Patienten, die an der JNCI-Lernstichprobe unbeteiligt waren (siehe Abschnitt 3.4.3, u.a. Abbildung 3.2), ergab für die dichotomen Variablen uneinheitlich eine höhere und zwei niedrigere Schätzungen.

[16] Das Kriterium der Reduktion „um mehr als 4“ wurde gewählt, weil 4 die erste natürliche Zahl ist, ab der eine χ^2-verteilte Teststatistik mit Freiheitsgrad 1 einen p-Wert $< 0{,}05$ besitzt.

[17] Beim Versuch mit Hilfe der Betrachtung von Kaplan-Meier-Kurven eine Variable mit mehr als zwei Kategorien und statistisch signifikant unterschiedlichen Remissionswahrscheinlichkeiten zu finden, wurde im Falle der Milzgröße z.B. festgestellt, dass Patienten mit einer Milzvergrößerung von 1-4 cm keine statistisch signifkant günstigeren Remissionswahrscheinlichkeiten besaßen als Patienten mit mehr als 10 cm Vergrößerung ($p = 0{,}7799$).

Den identifizierten statistisch signifikanten Zusammenhängen zwischen den Baselinevariablen und der zytogenetischen Remission sollten beim späteren Cox-Modell für die Überlebenswahrscheinlichkeiten besondere Aufmerksamkeit zuteil werden.

4.4 Multiple Analyse und Entwicklung des Prognosesystems

4.4.1 Die Selektion des besten prognostischen Modells

Bei der Selektion des besten Modells für die Überlebenszeit wurden alle Baselinevariablen und ihre vorab beschriebenen, verschiedenen Skalierungen berücksichtigt.[18] Als Ergebnis signifikant unterschiedlicher Überlebenswahrscheinlichkeiten[19] erhielten sowohl die partielle wie auch die komplette zytogenetische Remission jeder einen eigenen, zeitabhängigen Faktor, dessen Wert wie zuvor von 0 auf 1 gesetzt wurde, sobald zum ersten Mal die entsprechende Remission erreicht worden war. Bei Patienten, die zuerst das Stadium „partielle ZR“ erreicht hatten, wurde zum Zeitpunkt der ersten beobachteten kompletten ZR der Faktor zur partiellen Remission zurück auf 0 gesetzt. Für 743 Patienten mit 324 Ereignissen waren vollständige Daten zu den neun Baselinevariablen und der zytogenetischen Remission vorhanden. Damit lag das Verhältnis von Ereignissen zu untersuchten Kovariablen über 30 zu 1. Die mediane Überlebenszeit der 743 Patienten betrug 74 Monate. Das beste prognostische Modell für die Überlebenswahrscheinlichkeiten der 743 Patienten enthielt dichotome Skalierungen zu Alter, Milzvergrößerung, Eosinophile, Hämoglobin und Basophile sowie die zeitabhängigen Faktoren zur partiellen und kompletten Remission.
Alle wichtigen Informationen zum besten prognostischen Modell, welches für die sechs Variablen bei 761 Patienten berechenbar war, sind in Tabelle 4.8 dargestellt. Positive Koeffizienten und relative Risiken über 1 stehen für die bei den höheren Baselinevariablenwerten erhöhten Sterbewahrscheinlichkeiten. Eine Ausnahme bildete wieder Hämoglobin, wo das höhere Risiko den kleineren Werten zuzuordnen war. Bei den Baselinevariablen waren nach Maßgabe der $-2\ln L(\hat{\beta})$-Statistik in allen Fällen die dichotomen Skalierungen den anderen Skalierungen vorzuziehen, was allerdings wiederum die in Abschnitt 4.3.3 erläuterte Gefahr der Effektüberschätzung in sich barg. Die partielle und v.a. komplette zytogenetische Remission verminderten die Sterbewahrscheinlichkeiten deutlich, die relativen Risiken bei den Hazardraten lagen zuletzt bei der Hälfte bzw. einem Fünftel. Eine Verbesserung des Modells durch den Einschluss von Interaktionen konnte nicht erreicht werden. Dies betraf insbesondere auch die in Abschnitt 4.3.3 identifizierten Zusammenhänge zwischen den Baselinevariablen und der zytogenetischen Remission. Die Verwendung des New CML-Scores statt seiner sechs Einzelvariablen führte immer zu einem Modell mit statistisch signifikant höherer $-2\ln L(\hat{\beta})$-Statistik als beim Modell in Tabelle 4.8 und damit zu einem schlechteren Ergebnis.

4.4.2 Überprüfung der Annahme konstanter Koeffizienten im Cox-Modell

Die Überprüfung auf konstante Effekte war für den univariaten Fall bereits Bestandteil der Abschnitte 4.2.1 und 4.2.2. Auch dem multiplen Modell der Tabelle 4.8 wurden zur Untersuchung eines zeitlichen Einflusses auf die Koeffizientenschätzer Interaktionen zwischen den Kovariablen und $\ln t$ beigefügt. Zusätzlich zu den sechs Kovariablen war jedoch keine der Interaktionen statistisch signifikant (Wald-Test, $\alpha = 0{,}05$).

[18] Vgl. Tabellen 4.3 - 4.6 und Abschnitt 4.3.3.
[19] Vgl. Abschnitt 4.2.2

Tabelle 4.8: Multiple Analysen im Cox-Modell: Das beste Modell zum Einfluss der Baselinevariablen und der zytogenetischen Remission auf die Überlebenswahrscheinlichkeiten der 761 Patienten mit vollständigen Daten

	Patientenzahl	Schätzung Koeffizient	Standardabweichung	Walds χ^2-Statistik		p-Wert	
Variable[a]	n/tot[b]	$\hat{\beta}$[c]	$\hat{v}$	X^2	df[d]	p	RR[e]
ZR[f]							
Drei Gruppen[g]	761/326			**50,3024**	**2**	**<0,0001**	
keine dZR	546/280	0	-				1
pZR, (noch) keine kZR	95/ 28	-0,6643	0,2023				0,515
kZR	120/ 18	-1,6351	0,2483				0,195
Alter							
Zwei Gruppen	761/326			**17,2990**	**1**	**<0,0001**	
≤41 Jahre	237/ 60	0	-				1
>41 Jahre	524/266	0,6228	0,1497				1,864
Milzvergröß.							
Zwei Gruppen	761/326			**16,3823**	**1**	**<0,0001**	
0-7 cm	563/221	0	-				1
>7 cm	198/105	0,5294	0,1308				1,698
Eosinophile							
Zwei Gruppen	761/326			**10,1109**	**1**	**0,0015**	
0-2%	473/185	0	-				1
>2%	288/141	0,3634	0,1143				1,438
Hämoglobin							
Zwei Gruppen	761/326			**8,3162**	**1**	**0,0039**	
w[h], ≤11,3 g/dl & m, ≤13,5 g/dl	311/111	0,3597	0,1247				1,433
w, >11,3 g/dl & m, >13,5 g/dl	450/215	0	-				1
Basophile							
Zwei Gruppen	761/326			**8,0362**	**1**	**0,0046**	
0-2%	282/100	0	-				1
>2%	479/226	0,3480	0,1228				1,416

[a] Einheiten und Messgenauigkeit bei den Baselinevariablen wie in Tabelle 4.2.

[b] n: Gesamtzahl der Patienten mit Daten, tot: die davon inzwischen Verstorbenen.

[c] Der Wert „0“ steht für die mit „0“ kodierte Referenzgruppe. Die andere Gruppe wurde jeweils mit „1“ kodiert.

[d] Freiheitsgrade.

[e] Relatives Risiko: Verhältnis der geschätzten Hazardfunktion zur Hazardfunktion der Referenzkategorie.

[f] ZR: Zytogenetische Remission; dZR, pZR, kZR: erste deutliche (d), partielle (p) oder komplette (k) ZR.

[g] Die Anzahl der Patienten und der Remissionsgrade in den drei Gruppen gibt den aktuellsten Datenstand zum Zeitpunkt des Datenbankschlusses wieder. Die Gruppenzugehörigkeit veränderte sich über die Zeit.

[h] w: weiblich; m: männlich.

Tabelle 4.9 zeigt die Koeffizientenschätzer $\hat{\beta}$ und die Standardabweichungen $\hat{v}$ zu allen Variablen des besten Modells, wenn die Überlebenszeit zu den Verlaufszeitpunkten 5, 6, 7, 8 und 9 Jahre zensiert wird. „Fünf Jahre“ wurde als erster Zensierungszeitpunkt gewählt, um eine aus-

Tabelle 4.9: Die Koeffizienten des Modells aus Tabelle 4.8 bei verschiedenen Zensierungszeitpunkten für die Überlebenszeit

Zensierungszeit (in Jahren):	5		6		7		8		9	
Anzahl der Verstorbenen:	222		268		305		316		324	
Variable[a]	$\hat{\beta}$[b]	$\hat{v}$	$\hat{\beta}$	$\hat{v}$	$\hat{\beta}$	$\hat{v}$	$\hat{\beta}$	$\hat{v}$	$\hat{\beta}$	$\hat{v}$
ZR[c]										
partielle ZR, (noch) keine komplette ZR	-0,83	0,28	-0,76	0,24	-0,66	0,21	-0,64	0,21	-0,69	0,21
komplette ZR	-1,30	0,33	-1,41	0,29	-1,61	0,27	-1,64	0,26	-1,61	0,25
Alter										
>41 Jahre	0,62	0,18	0,61	0,16	0,61	0,16	0,63	0,15	0,60	0,15
Milzvergrößerung										
>7 cm	0,63	0,15	0,58	0,14	0,54	0,13	0,56	0,13	0,53	0,13
Eosinophile										
>2%	0,37	0,14	0,37	0,13	0,36	0,12	0,33	0,12	0,36	0,11
Hämoglobin										
w[d], ≤11,3 g/dl & m, ≤13,5 g/dl	0,45	0,16	0,39	0,14	0,39	0,13	0,38	0,13	0,37	0,13
Basophile										
>2%	0,36	0,15	0,46	0,14	0,41	0,13	0,39	0,13	0,37	0,12

[a] Definitionen und Einheiten der Variablen wie in Tabelle 4.8.
[b] Es werden nur die Gruppen mit von „0“ verschiedenen Werten angeführt.
[c] Zytogenetische Remission.
[d] w: weiblich; m: männlich.

reichende Beobachtungsdauer von Überlebenszeiten vorliegen zu haben, welche eine zuverlässige Schätzung prognostischer Unterschiede zwischen den Kovariablen prinzipiell ermöglichte. Minimale Veränderungen beobachtete man bei den Koeffizientenschätzern zum Alter und zu den Eosinophilen (vgl. Tabelle 4.8). Bei den Basophilen brachte der Übergang vom 5. zum 6. Jahr als Zensierungszeitpunkt eine Zunahme des Schätzers von 28%, danach aber wieder die Annäherung an den Ausgangswert 0,36. Die Schätzer von Milzvergrößerung und Hämoglobin schrumpften vom ersten bis zum letzten Zensierungszeitpunkt um 16% bzw. 18% und hatten bis zum 3. bzw. 2. Zensierungszeitpunkt ihren endgültigen Wert (Tabelle 4.8) fast erreicht. Bei der zytogenetischen Remission nahm der Absolutbetrag des Koeffizientenschätzers zur partiellen ZR bis zum letzten Zensierungspunkt um 17% ab, während der Absolutbetrag des Schätzers zur kompletten Remission um 24% zunahm. Diese Zunahme unterstrich die wachsende prognostische Bedeutung der kompletten ZR hinsichtlich einer möglichst deutlichen Diskriminierung von Überlebenswahrscheinlichkeiten. Betrachtete man wegen ihres Zusammenhangs die Summen der

Absolutbeträge beider Schätzer, so lag die Gesamtzunahme zwischen erstem und letztem Zensierungszeitpunkt bei 13%. Ab dem 7. Jahr als Zensierungszeitpunkt lagen sechs der sieben Schätzer um maximal 0,03 von den endgültigen Schätzern entfernt, nur bei den Basophilen betrug die Differenz zunächst noch 0,06 (Vergleich von Tabelle 4.8 mit Tabelle 4.9). Die p-Werte zu den Wald-Statistiken waren für alle Kovariablen zu allen Zensierungszeitpunkten statistisch signifikant ($\alpha = 0{,}05$).

Für die Baselinevariablen bestand mit den in Abschnitt 2.11 vorgestellten graphischen Verfahren nach Andersen et al. [7] eine dritte Möglichkeit, die Koeffizientenschätzer aus einem multiplen Cox-Modell im Hinblick auf eine zeitunabhängige Konstanz zu überprüfen. Die partielle Likelihoodfunktion wurde nach den Kategorien einer zu untersuchenden Baselinevariablen stratifiziert und die Koeffizienten der übrigen Variablen des multiplen Modells geschätzt. Mit Hilfe der Schätzer wurden die logarithmierten kumulierten Baselinehazardfunktionen der einzelnen Kategorien berechnet und gegen den zeitlichen Verlauf aufgetragen (vgl. Abbildung 4.6). Dargestellt wurden exemplarisch die Ergebnisse zur Milzvergrößerung, wo bei der Zensierung zu obigen Verlaufszeitpunkten die größten absoluten Veränderungen für die Koeffizientenschätzer zu beobachten waren. Die generell höheren Werte zur Gruppe „$>$ 7 cm“ korrespondierten mit den ungünstigeren Überlebenswahrscheinlichkeiten. Der letzte Patient dieser Gruppe verstarb nach 8 Jahren, länger wurden noch 7 Patienten beobachtet. In der Gruppe „$\leq$ 7 cm“ lag der letzte Todesfall 9,5 Jahre nach Therapiebeginn, 19 Patienten besaßen eine längere Überlebenszeit. Analog zur Abnahme des Koeffizientenschätzers über die verschiedenen Zensierungszeitpunkte (Tabelle 4.9) näherten sich die beiden logarithmierten kumulierten Baselinehazardfunktionen mehr und mehr an. Die Abstände zwischen den Kurven entsprachen zu den Zeitpunkten 5 und 6 Jahre nach Therapiebeginn in etwa den Schätzern aus Tabelle 4.9, danach wurde die Übereinstimmung geringer. Bei den vier anderen Baselinevariablen zeigten die Graphen noch stärker einen parallelen Verlauf.

In einer zweiten Darstellungsweise nach Andersen et al. [7] werden die kumulierten Baselinehazardfunktionen direkt miteinander verglichen (Abbildung 4.7). Die Ergebnisse zu einer Variablenkategorie dienen dabei als Referenz. Im Falle der Milzvergrößerung lagen die Werte der Kategorie „$>$ 7 cm“ im zeitlichen Verlauf immer über den Werten der Kategorie „$\leq$ 7 cm“ (Steigung > 1), ein Hinweis auf die geringeren Überlebenswahrscheinlichkeiten in der höheren Kategorie. Die Steigung verringerte sich ab der Koordinate (0,1;0,21) und der resultierende leicht konkave Kurvenverlauf deutete eine Annäherung der Hazardfunktionen der beiden Kategorien bzw. eine Abnahme des Koeffizientenschätzers im zeitlichen Verlauf an. Die kumulierten Baselinehazardfunktionen der übrigen zeitunabhängigen Kovariablen waren im Vergleich zur kumulierten Baselinehazardfunktion ihrer jeweiligen Referenzkategorie einer Geraden ähnlicher als der Graph aus Abbildung 4.7.

Die Erhöhung des Anteils an partiellen bzw. kompletten ZR, und damit ein steigender prognostischer Einfluss auf die Überlebenswahrscheinlichkeiten, konnte sich erst nach und nach mit den beobachteten Remissionszeiten einstellen. Ein steigender prognostischer Einfluss bedeutete aber einen tendenziell wachsenden Betrag des Koeffizientenschäetzers und somit entgegen der Modellannahme dessen Zeitabhängigkeit.
Die Auswirkungen der durch die Variablendefinition impliziten Modellverletzung auf das beste multiple Modell blieben jedoch gering. Interaktionen der Koeffizienten mit der Zeit waren nicht signifikant, die Unterschiede zwischen den Koeffizientenschätzern bei der Wahl verschiedener Zensierungszeitpunkte (Tabelle 4.9) nicht erheblich und die verglichenen (logarithmierten)

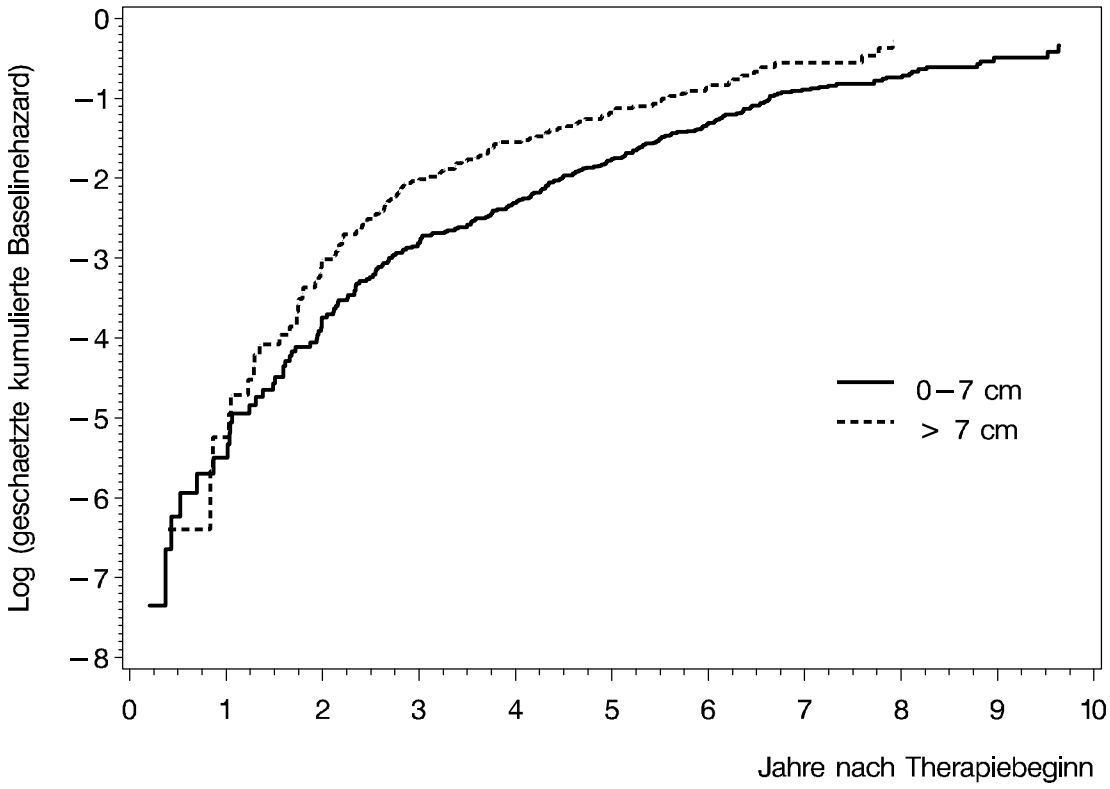

Abbildung 4.6: Die logarithmierten kumulierten Baselinehazardfunktionen (nach Andersen et al. [7]) für die beiden Kategorien der Kovariablen Milzvergrößerung. Mit jedem Ereignis (Todesfall) erfolgte eine Veränderung der zur betroffenen Patientengruppe gehörigen Funktion. Entsprechend wurden die logarithmierten kumulierten Baselinehazardfunktionen bis zum Zeitpunkt des letzten Todesfalles in ihrer Gruppe dargestellt.

kumulierten Baselinehazardfunktionen verliefen annähernd (parallel) gerade (Abbildungen 4.6 und 4.7). Die Robustheit des multiplen Modells gegen die in Kauf genommene Verletzung der Modellannahme wurde durch die in Relation zu den Überlebenszeiten kurzen Remissionszeiten unterstützt. Während zum Zeitpunkt „drei Jahre nach Therapiebeginn“ bereits 95% aller beobachteten partiellen ZR und 83% aller kompletten ZR erfolgt waren, lag die Überlebenswahrscheinlichkeit der 761 Patienten noch bei 0,83; die mediane Überlebenszeit wurde nach 75 Monaten erreicht. Demzufolge beobachtete man für den überwiegenden Teil an beobachteten Lebensjahren keine Veränderung des Remissionsstatus.
Die Verletzung der Annahme konstanter Koeffizienten wurde als nicht gravierend beurteilt und stand der Verwendung des Cox-Modells für die gegebene Datensituation nicht entgegen.

4.4.3 Überprüfung der Anpassung des besten multiplen Modells an die Daten

Zur Prüfung der Modellanpassung wurden die Residuen nach Barlow-Prentice [14] (Formel 2.12) berechnet. Fünfzehn der 761 Patienten mit vollständigen Daten zum besten multiplen Cox-Modell hatten mindestens einmal ein Residuum mit einem Betrag > 1. Zwei Patienten hatten sowohl zum Alter als auch zur Milzvergrößerung Residuen mit einem Betrag > 2. Abbildung 4.8 zeigt die Residuen zum Alter, der Kovariablen, für welche die beiden größten Abweichungen von 0 und den übrigen Residuen zu beobachten waren. Die Residuen der beiden betroffenen Patien-

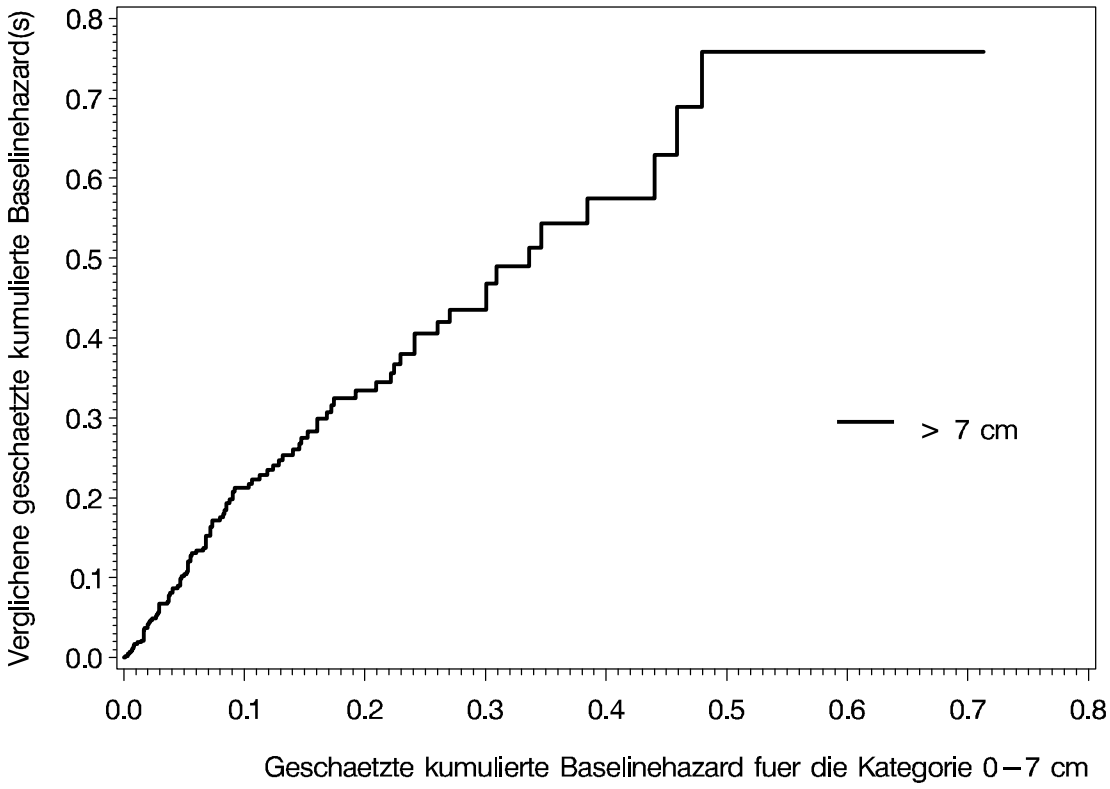

Abbildung 4.7: Vergleich der kumulierten Baselinehazardfunktion (nach Andersen et al. [7]) der Kategorie „> 7 cm Milzvergrößerung“ mit den Referenzwerten der kumulierten Baselinehazardfunktion der Kategorie „≤ 7 cm Milzvergrößerung“.

ten sind in der Abbildung oben rechts zu erkennen. Die Überlebenszeiten der beiden Patienten wurden nach 9,5 und 10,25 Jahren zensiert und waren damit relativ lang, obwohl - außer beim Alter - zu allen Kovariablen des besten multiplen Modells (vgl. Tabelle 4.8) Werte der prognostisch ungünstigsten Gruppe vorlagen, also auch von keiner zytogenetischen Remission berichtet worden war. Gerade weil das Alter $\leq$ 41 Jahre im Cox-Modell die einzige „Erklärung“ für die lange Überlebenszeit bot, lieferten die beiden Patienten einen vergleichsweise ungewöhnlich großen Beitrag zur Erhöhung des zugehörigen Koeffizientenschätzers respektive zur Verstärkung des prognostischen Unterschiedes zwischen den beiden Altersgruppen. Der überdurchschnittliche Einfluss der beiden Patienten auf den Koeffizientenschätzer drückte sich in den stark positiven Residuen > 2 aus. Umgekehrt war es unter Berücksichtigung der bis auf das Alter ungünstigen Kovariablenwerte speziell für einen Patienten mit Milzgröße > 7 cm außergewöhnlich, eine so lange Überlebenszeit aufzuweisen. Unter den Patienten mit den 30 längsten Überlebenszeiten, von denen nur zwei verstorben waren, stellten die beiden mit ihren extremen Residuen die Einzigen, die überhaupt zur Diagnose eine Milzgröße > 7 cm besessen hatten.[20] Die Konsequenz war ein überdurchschnittlich starker negativer Einfluss auf die Größe des Koeffizientenschätzers, d.h. der prognostische Unterschied zwischen den beiden Gruppen zur Milzgröße wurde durch die beiden Patienten abgeschwächt. Seine Entsprechung fand dies in zwei negativen Residuen $<$ -2.

[20] Im übrigen lagen weder das Alter der betroffenen Patienten, noch deren Milzvergrößerung nahe am jeweiligen Cutpoint.

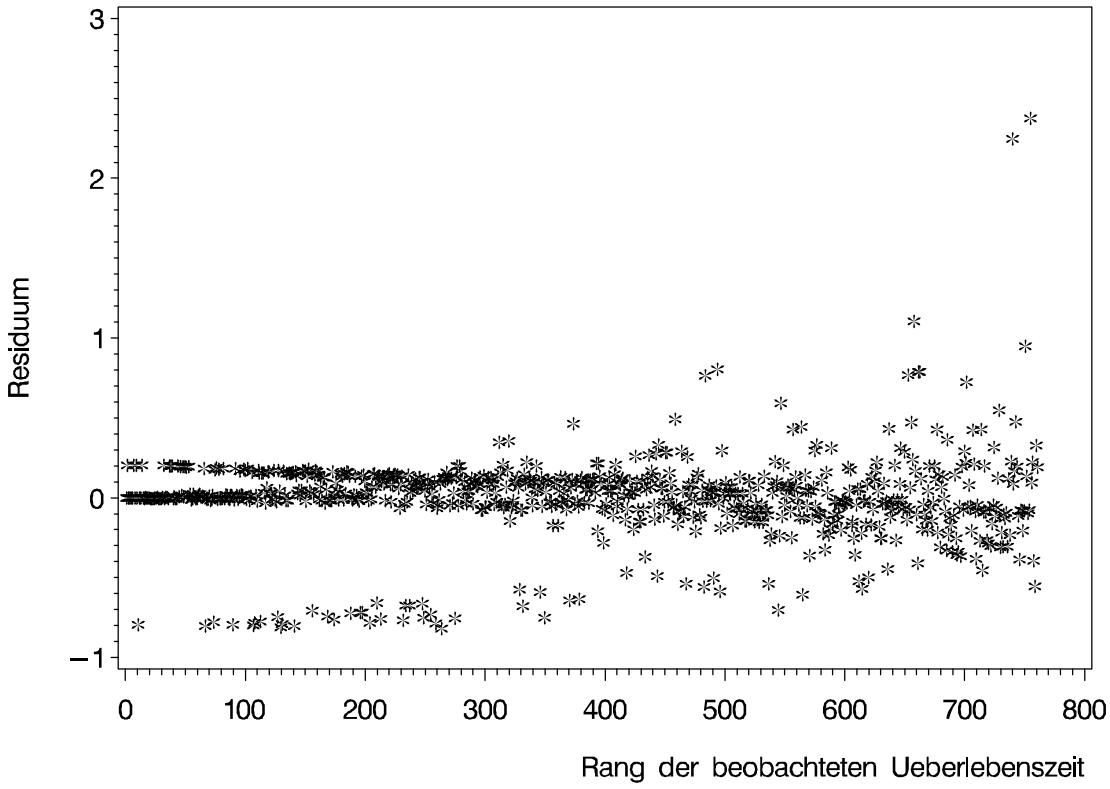

Abbildung 4.8: Die Barlow-Prentice-Residuen (nach Formel (2.12) aus Abschnitt 2.12) zur Kovariablen Alter aus dem besten multiplen Cox-Modell für die Überlebenszeit zu den 761 Patienten mit vollständigen Daten.

Die 15 Patienten mit einem Residuenbetrag > 1 wurden aus dem Datensatz entfernt und das beste multiple Cox-Modell erneut gesucht. Dabei fanden sich genau dieselben Kovariablen, nun alle mit p-Werten $< 0{,}001$ (Wald-Test). Jeder Koeffizientenschätzer hatte sich erhöht, nur der Schätzer zum Alter war in etwa gleich geblieben.
Weil einerseits nichts gegen die Korrektheit der zu den 15 Patienten vorliegenden Daten sprach und sich andererseits das Verhältnis der auf Basis von 746 Patienten geschätzten Koeffizienten zueinander so wenig veränderte, dass die spätere Risikogruppendefinition für jeden Patienten in exakt derselben Zuteilung resultierte wie die Definition auf Basis der Schätzer zu allen 761 Patienten, wurde das Modell aus Tabelle 4.8 beibehalten. Kraft der offensichtlichen Modellstabilität, wurde auf die nachträgliche Anwendung von „Bootstrap resampling“ und die Berechnung von „Shrinkage“ verzichtet.

4.4.4 Vom prognostischen Modell zum Prognosesystem

Der in Abschnitt 2.13 vorgestellte prognostische Index[21] enthält alle relevanten Informationen des prognostischen Modells, um die prognostischen Unterschiede zwischen verschiedenen Patienten zu beschreiben.

[21] Siehe Formel 2.13.

Berechnung von Risikowerten

Als eine monotone Transformation des prognostischen Indizes wurde für einen Patienten zum Zeitpunkt t nach dem Modell aus Tabelle 4.8 folgender Risikowert berechnet:

$$
\begin{aligned}
\mathrm{RI}(t) = \\
1000 \times (\quad & 0,6228 \times \text{Alter [1, falls Alter in vollendeten Jahren} > 41 \text{ Jahre; 0, sonst]} \\
+ \; & 0,5294 \times \text{Milzgröße [1, falls Milzgröße} > 7 \text{ cm unter dem Rippenbogen; 0, sonst]} \\
+ \; & 0,3634 \times \text{Eosinophile [1, falls Eosinophile} > 2\%\text{; 0, sonst]} \\
+ \; & 0,3597 \times \text{Hämoglobin bei Frauen [1, falls Hämoglobin} < 11{,}4 \text{ g/dl; 0, sonst] bzw.} \\
& \qquad \text{Hämoglobin bei Männern [1, falls Hämoglobin} < 13{,}6 \text{ g/dl; 0, sonst]} \\
+ \; & 0,3480 \times \text{Basophile [1, falls Basophile} > 2\%\text{; 0, sonst])} \\
- \; & 0,6643 \times \text{Partielle ZR [1, sobald } t \geq \text{ der Zeit des Eintretens und} \\
& \qquad \text{solange keine komplette ZR; 0, sonst]} \\
- \; & 1,6351 \times \text{Komplette ZR [1, sobald } t \geq \text{ der Zeit des Eintretens; 0, sonst].} \qquad (4.1)
\end{aligned}
$$

Das Alter wurde auf die Anzahl der vollendeten Lebensjahre abgerundet. Der Hämoglobinwert wurde auf die erste Nachkommastelle gerundet, die anderen Variablenwerte und der Risikowert $\mathrm{RI}(t)$ auf die nächste ganze Zahl.

Prognostizierte versus aus den Überlebenszeiten geschätzte Überlebenswahrscheinlichkeiten

Nach dem Vorschlag von Christensen et al. [24] wurden für die 761 Patienten auf Basis von Formel (2.14) Überlebenswahrscheinlichkeiten $\hat{p}_i(t, t+\lambda)$ aus dem besten identifizierten Modell prognostiziert. Da die Baselinehazardfunktion - als Bestandteil von (2.14) - wegen der gesunkenen Fallzahl nach acht Jahren begann, anstatt konstante, sprunghafte Anstiege zu verzeichnen, wurden die Wahrscheinlichkeiten nur in der Zeit zwischen Therapiebeginn bis zum Ende des achten Beobachtungsjahres betrachtet. Für λ wurde die Zeitspanne „1 Jahr“ gewählt, womit von einem Patient bis zu acht Schätzungen eingehen konnten. Kürzere Zeitspannen mit entsprechend weniger Ereignissen hätten eine zu geringe Differenzierung verschiedener Überlebenswahrscheinlichkeiten erlaubt. Dieser Differenzierungsnachteil wurde stärker gewichtet als der Vorteil einer geringeren Verletzung der Annahme konstanter Kovariablenwerte über das Zeitintervall $(t, t+\lambda)$, welchen kürzere Zeitintervalle gegenüber dem Einjahresintervall v.a. während der ersten beiden Therapiejahre besessen hätten. Abbildung 4.9 zeigt, in Abhängigkeit vom Risikowert zum Intervallbeginn t, den Vergleich der rein aus den beobachteten Zeiten geschätzten Überlebenswahrscheinlichkeiten versus den nach dem Cox-Modell unter obigen Annahmen zu erwartenden Wahrscheinlichkeiten, das nächste Jahr zu überleben. Zur Sicherstellung einer minimalen Schätzgenauigkeit wurden nur Ergebnisse zu Risikowerten betrachtet, für die die Überlebenswahrscheinlichkeit aus den beobachteten Überlebenszeiten zu wenigstens zehn Intervallen geschätzt werden konnte. Diese Bedingung erfüllten 55 von 75 Risikowerten mit 3550 von 3642 Intervallen. Der niedrigst mögliche Risikowert -1683 ($n = 35$ Intervalle) war durch Werte in der jeweils günstigeren Gruppe bei allen fünf Baselinevariablen und der späteren Beobachtung einer kompletten Remission zu erreichen. Patienten, die zu allen Baselinevariablen Werte der ungünstigeren Risikogruppe und im Therapieverlauf (zunächst) keine deutliche Remission aufwiesen, nahmen mit 2223 ($n = 123$) den höchsten möglichen Risikowert ein. Die Vergleichskurven beider Überlebenswahrscheinlichkeiten in Abbildung 4.9 folgten einem gemeinsamen Trend und

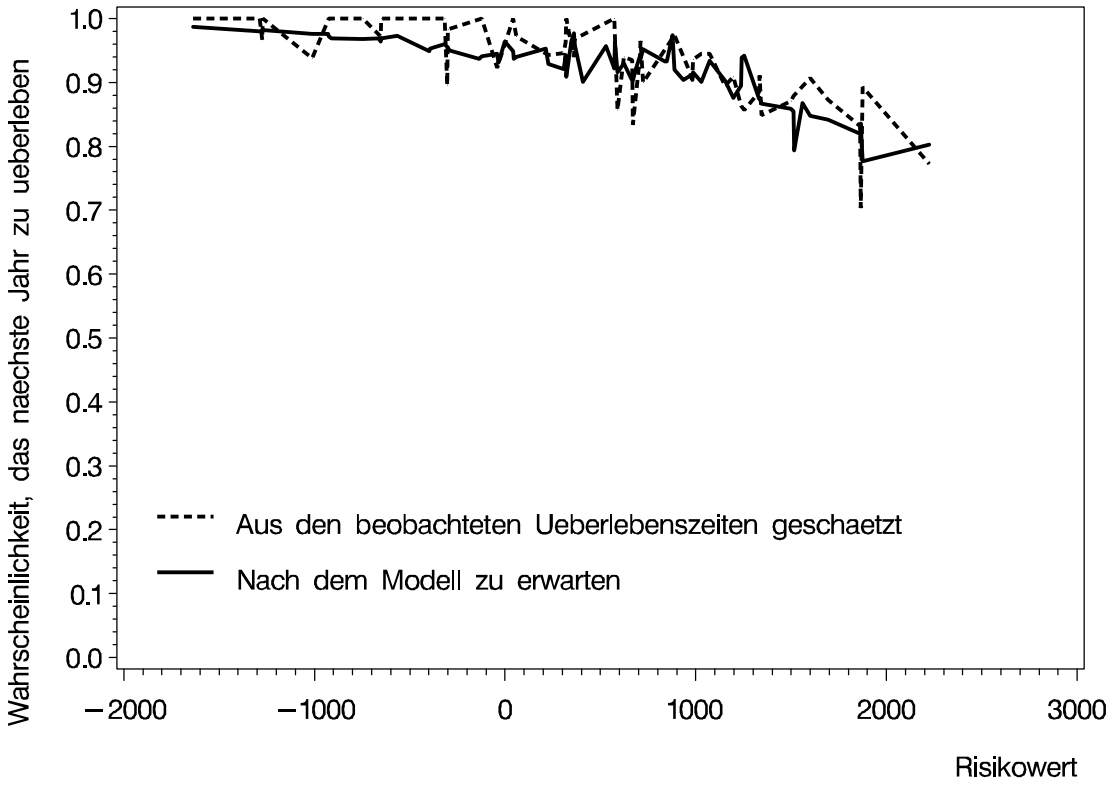

Abbildung 4.9: Für dem Cox-Modell aus Tabelle 4.8 nach Christensen et al. [24] zu erwartende Wahrscheinlichkeiten $\hat{p}_i(t, t+\lambda)$ (Formel (2.14)), das nächste Jahr zu überleben, verglichen mit den aus den beobachteten Überlebenszeiten während der Intervalle $(t, t+\lambda)$ geschätzten Wahrscheinlichkeiten. Beide Überlebenswahrscheinlichkeiten wurden in Abhängigkeit vom Risikowert zum Intervallbeginn geschätzt. Die aus den Beobachtungen geschätzte Überlebenswahrscheinlichkeit zu einem bestimmten Risikowert erhielt man, indem die Anzahl aller dem Wert zugehörigen Zeitintervalle *ohne* Ereignis durch die Summe der zugehörigen Zeitintervalle *mit und ohne* Ereignis geteilt wurde. Auch wenn von einem Patienten zwischen ein und acht Intervalle stammen konnten, wurden alle Intervalle als unabhängige Beobachtungseinheiten behandelt. Die Überlebenswahrscheinlichkeiten sollten formal alleine vom Risikowert zum Intervallbeginn abhängen. Zur Sicherstellungen einer minimalen Schätzgenauigkeit der Überlebenswahrscheinlichkeiten aus den beobachteten Daten wurden nur Risikowerte berücksichtigt, zu welchen wenigstens zehn Intervalle vorlagen.

nahmen mit steigendem Risikowert ab. Für die 55 Risikowerte lagen die aus den Überlebenszeiten geschätzten Überlebenswahrscheinlichkeiten zwischen 0,70 und 1, die nach dem Modell erwarteten Überlebenswahrscheinlichkeiten zwischen 0,78 und 0,99. Nur in sechs von 55 Fällen überschritt der Absolutbetrag der Differenz beider Wahrscheinlichkeiten die Grenze von 0,06. Die größten Unteschiede zwischen den Überlebenswahrscheinlichkeiten wurden für die beiden nebeneinander liegenden Risikowerte 1864 ($n = 27$) und 1875 ($n = 28$) registriert. Im ersten Fall lag die prognostizierte Überlebenswahrscheinlichkeit (0,836) mit 0,132 über der direkt aus den Überlebenszeiten geschätzten (0,704) und im zweiten Fall um 0,116 darunter (0,777 vs. 0,893). Insgesamt wurde geschlussfolgert, dass die aus den beobachteten Überlebenszeiten geschätzten Überlebenswahrscheinlichkeiten gut mit den nach dem Modell erwarteten Überlebenswahrscheinlichkeiten prognostiziert werden konnten.

Die Definition von Risikogruppen

Die Risikowerte der Patienten wurden zu den Verlaufszeitpunkten 6, 9, 12, 15, 18, 21, 24, 30 und 36 Monate nach Therapiebeginn mit IFN-α berechnet. Bis zum Ende des 36. Monats waren 154 der 163 vorliegenden partiellen Remissionen sowie 100 der 120 kompletten Remissionen beobachtet, wodurch nur noch bei 28 Patienten eine Änderung des Risikowertes erfolgte.[22] Gemäß dem New CML-Score können für mit IFN-α behandelte Patienten anhand von Baselinevariablenwerten drei Prognosegruppen gefunden werden, welche bei ausreichender Patientenzahl und Beobachtungszeit zuverlässig statistisch signifikant unterschiedliche Überlebenswahrscheinlichkeiten aufweisen. Die Hinzunahme der zeitabhängigen Variablen „zytogenetische Remission“ ließ auf Basis der zusätzlichen Information und der bisherigen Ergebnisse die Möglichkeit erwarten, eine weitere Prognosegruppe definieren zu können, so dass im Therapieverlauf Überlebenswahrscheinlichkeiten von vier Prognosegruppen valide und zuverlässig statistisch signifikant unterscheidbar sein müssten. Über die „Minimal p-value“-Methode wurde für die Risikowerte zu den einzelnen Verlaufszeitpunkten nach den drei Grenzen gesucht, die vier Gruppen definierten, welche die Überlebenswahrscheinlichkeiten der zum jeweiligen Verlaufszeitpunkt noch unter Beobachtung stehenden Patienten mit maximalem Wert der Logrank-Statistik diskriminierten.[23] Nach dem Prinzip der Landmarkanalyse [8] wurden bis zum jeweils interessierenden Zeitpunkt registrierte partielle oder komplette Remissionen bei der Berechnung der Risikowerte berücksichtigt. Mit der Verwendung der Effektschätzer des endgültigen Modells zu allen Berechnungszeitpunkten wurden nicht nur alle Überlebenszeitinformationen, sondern auch aller zugelassenen Remissionsinformationen ausgenutzt. Selbst wenn z.B. zum Zeitpunkt „6 Monate“ erst wenige komplette Remissionen vorlagen, konnte so der auch auf den späteren Remissionen beruhende große Effektschätzer bei der Risikogruppendefinition zu einer klaren Zuordnung in die vorteilhafteste Risikogruppe führen. Möglichst viel Information aus der Lernstichprobe bei der Entwicklung eines Prognosesystems einfließen zu lassen war nicht nur methodisch sinnvoll, sondern durch die vom Remissionszeitpunkt unabhängig hohen Überlebenswahrscheinlichkeiten der Patienten mit kompletter zytogenetischer Remission auch aus klinischer Sicht absolut gerechtfertigt.

Die Veränderung der Risikowerte über die Zeit und die Abnahme der Zahl unter Beobachtung stehender Patienten führten bei Anwendung der „Minimal p-value“-Methode zu den verschiedenen Zeitpunkten z.T. nicht zu denselben drei Gruppengrenzen. Dennoch ließen sich, im Sinne einer im medizinischen Alltag praktikablen, nicht alle drei Monate zu ändernden Risikogruppendefinition, drei Gruppengrenzen finden, welche die vier Risikogruppen sowohl zum Baselinezeitpunkt als auch zu allen Verlaufspunkten ohne relevante Veränderung der p-Werte paarweise statistisch signifikant unterschieden und die prognostische Differenzierung des Prognosesystems kaum veränderten. Die drei Gruppengrenzen besaßen die Werte 348, 1196 und 1860.
In Abschnitt 4.2.2 hatte sich auf Basis der vorliegenden Daten das Ende des 21. Therapiemonats als sinnvollster Entscheidungszeitpunkt herauskristallisiert. Abbildung 4.10 beschreibt das Ergebnis der Landmarkanalyse nach 21 Monaten für die noch unter Beobachtung stehenden 646 Patienten, aufgeteilt in die durch die Gruppengrenzen entstandenen Risikogruppen.[24] Die

[22] Ein Patient hatte zunächst eine partielle und dann eine komplette Remission.

[23] Optimal wäre ein Verfahren, welches die gleichzeitige (gleichberechtigte) Suche nach den drei Gruppengrenzen unter der Berücksichtigung des multiplen Testens erlauben würde. Solch ein Verfahren ist bisher jedoch nicht bekannt. Beim hier angewandten rekursiven Partionierungsverfahren hängen - wie bei CART - die Suchergebnisse zu nachfolgenden Gruppengrenzen vom Ergebnis der ersten Gruppengrenze an der Baumwurzel ab.

[24] Aus Gründen späterer Vergleichbarkeit wurde ein Patient (keine deutliche Remission, zensierte Überlebenszeit 1764 Tage (58 Monate) mit fehlendem Wert zum New CML-Score ausgeschlossen. Damit bildeten ab Therapiebeginn nun mehr 760 anstatt 761 Patienten die Basis für alle weiteren Analysen.

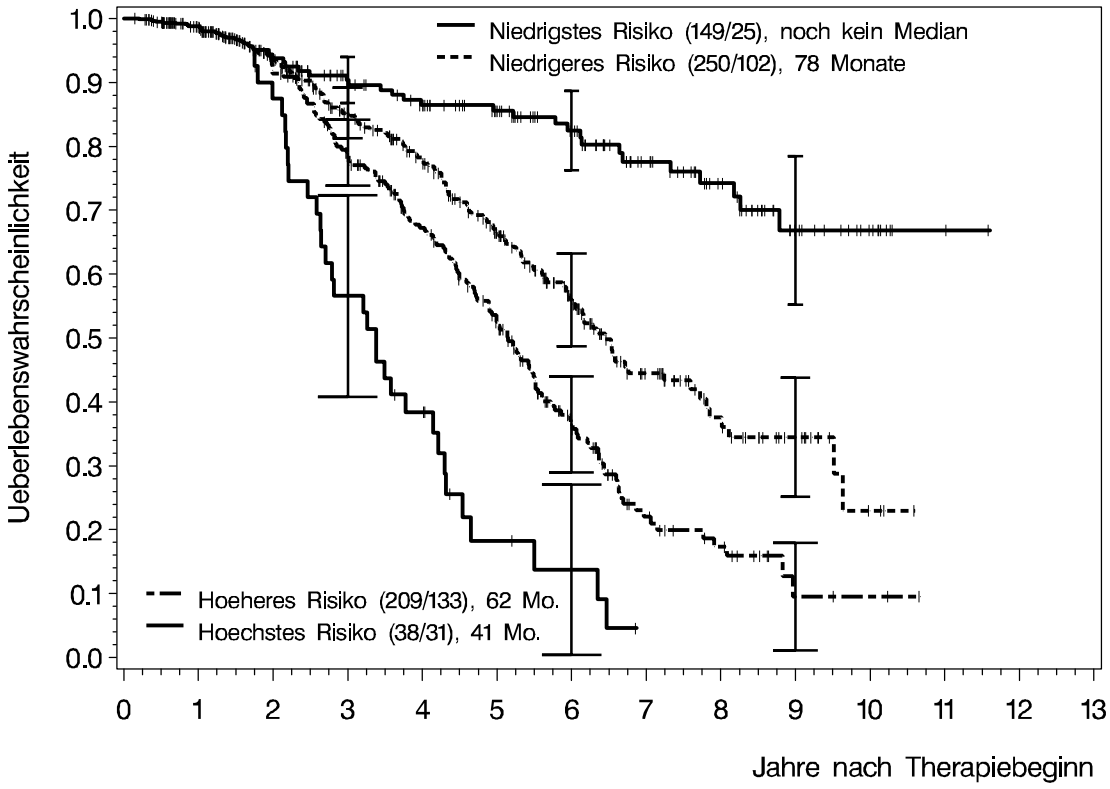

Abbildung 4.10: Kaplan-Meier-Kurven ab der Landmark „21 Monate“ mit geschätzten Überlebenswahrscheinlichkeiten in den vier Risikogruppen des neuen Prognosesystems. Bis zur Landmark „21 Monate“ wurden die Überlebenswahrscheinlichkeiten aller 760 Patienten der Lernstichprobe gemeinsam geschätzt. Ab Ende des 21. Therapiemonats wurden die verbliebenen 646 Patienten je nach Risikowert auf die vier Gruppen verteilt. Die Legende „(149/25), noch kein Median“ bedeutet: Unter den 149 Patienten wurden 25 Todesfälle beobachtet. Die mediane Überlebenszeit wurde nicht erreicht. Die drei anderen Legenden sind analog zu verstehen, wobei hier die medianen Überlebenszeiten, mit z.B. 62 Monaten in der Gruppe „höheres Risiko“, vorlagen. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzten Wahrscheinlichkeiten mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

Überlebenswahrscheinlichkeit am Ende des 21. Monats betrug 0,9507. Von den 114 Patienten, die nicht mehr unter Beobachtung standen, waren 35 (31%) verstorben, 56 transplantiert (49%) und 23 (21%) zensiert. Die χ^2-Statistik des Logrank-Tests zum Vergleich der vier Kurven besaß den Wert 142,0629 (3 Freiheitsgrade) mit einem zugehörigen $p < 0{,}0001$. Unter dieser Grenze lagen auch alle sechs p-Werte zu den paarweisen Vergleichen von je zwei der Risikogruppen. Während die mediane Überlebenszeit in der Niedrigstrisikogruppe ($n = 149$, 23,1% von 646) nicht erreicht wurde, betrug sie in den Gruppen „niedrigeres Risiko“ ($n = 250$, 38,7%), „höheres Risiko“ ($n = 209$, 32,4%) und „höchstes Risiko“ ($n = 38$, 5,9%) 78, 62 und 41 Monate. Die 9-Jahresüberlebenswahrscheinlichkeiten beliefen sich auf 0,6684 [95%-K.I.: 0,5522; 0,7846] (niedrigstes Risiko), 0,3446 [0,2511; 0,4381] (niedrigeres Risiko) und 0,0954 [0,0116; 0,1791] (höheres Risiko). In der Gruppe „höchstes Risiko“ wurde kein Patient länger als sieben Jahre beobachtet; die 6-Jahresüberlebenswahrscheinlichkeit war 0,1371 [0,0038; 0,2704].
Zu Therapiebeginn befanden sich 40 der 646 Patienten (6,2%) in der Höchstrisikogruppe. Zwei

Patienten mit partieller bzw. mit kompletter Remission wechselten vor der Landmark „21 Monate“ durch die Reduktion ihrer Risikowerte in die Gruppen „höheres Risiko“ bzw. „niedrigeres Risiko“. Von den 268 Patienten (41,5% von 646), welchen zu Therapiebeginn höheres Risiko attestiert wurde, verblieben 208 ohne Remission in dieser Gruppe. Alle 42 Patienten mit partieller Remission wechselten in die nächst günstigere Risikogruppe und alle 18 Patienten mit kompletter Remission in die Niedrigstrisikogruppe. Zur Gruppe „niedrigeres Risiko“ gehörten zu Anfang 290 Patienten (44,9%), was für 207 (davon einer mit partieller Remission) auch nach 21 Monaten der Fall war. Die übrigen 83 Patienten, 49 mit partieller und 34 mit kompletter Remission, kamen in die Niedrigstrisikogruppe. Die Niedrigstrisikogruppe umfasste bereits zu Therapiebeginn 48 Patienten (7,4%), wovon 11 bis zur Landmark eine partielle und 8 eine komplette Remission erfuhren. Insgesamt verbesserten sich innerhalb der 21 Monate 92 der 104 Patienten mit partieller Remission um eine Risikogruppe und 52 der 61 Patienten mit kompletter Remission erreichten neu die Niedrigstrisikogruppe.
Zu Therapiebeginn und zu den Landmarkzeitpunkten bis zu 36 Monaten danach lagen die p-Werte der Logrank-Tests über alle vier Risikogruppen des Prognosesystems unter 0,0001. Dies galt ebenso für 58 der 66 paarweisen Vergleiche. Die übrigen acht p-Werte lagen alle unter 0,0250.

Das identifizierte Prognosesystem bot in der Lernstichprobe zu jeder Verlaufszeit eine gute Diskriminierung von vier Risikogruppen mit statistisch signifikant unterschiedlichen Überlebenswahrscheinlichkeiten. Dieses erste Prognosesystem für die Unterscheidung der Patientenrisiken stützte sich nur auf den linearen Prädiktor des Cox-Modells. Nun aber galt es zu prüfen, inwieweit mit diesem Prognosesystem bereits frühere wichtige prognostische Erkenntnisse mitberücksichtigt wurden und falls nicht, wie man es als Ausgangsbasis für Korrekturen zur Verbesserung der Prognose nutzen könnte. Zuletzt erschien die Suche nach einer möglichst anwendungerfreundlichen, einfachen Methode zur Berechnung der Risikogruppen sinnvoll, ohne dabei jedoch die Qualität der prognostischen Diskriminierung der verschiedenen Risikogruppen wesentlich zu beeinträchtigen.

Die deutlichen Remissionen in der Hochrisikogruppe des New CML-Scores

In der aktualisierten und um neue Patienten erweiterten Validierungsstichprobe von Hasford et al. [43] und in einer unabhängigen Patientenstichprobe von Bonifazi et al. [18, 90] lagen in der Hochrisikogruppe Fallzahlen und Beobachtungszeiten vor, die eine stabile Schätzung der Überlebenswahrscheinlichkeiten dieser Patientengruppe innerhalb der jeweiligen Stichprobe erlaubten und in beiden Fällen betrug die mediane Überlebenswahrscheinlichkeit 45 Monate, nahe den 42 Monaten in der Hochrisikogruppe der zur Entwicklung benutzten ursprünglichen Lernstichprobe [42]. Der New CML-Score vermag offensichtlich Patienten mit hohem Risiko bei IFN-α basierter Therapie verlässlich zu erkennen; diese Patienten sollten nicht hauptsächlich mit IFN-α therapiert werden. Auf Basis dieser Ergebnisse erschien es daher medizinisch und methodisch sinnvoll, bei der Hochrisikogruppe auf den mehrfach validierten und bereits etablierten New CML-Score [13, 18, 43, 64, 90] zu bauen und alle Patienten mit Hochrisiko nach dem New CML-Score bei dem neu entwickelten Prognosesystem von Anfang an der Höchstrisikogruppe mit zuzuordnen. Weiter hatten Pfirrmann und Hasford für die Patienten innerhalb der Hochrisikogruppe des New CML-Scores festgestellt, dass Patienten mit deutlicher Remission im Therapieverlauf gegenüber Patienten ohne deutliche Remission hinsichtlich der Überlebenswahrscheinlichkeiten nicht profitieren [46, 89]. Unter den 93 Hochrisikopatienten der Lernstichprobe erreichten zehn eine deutliche Remission, wovon acht verstarben (mediane Überlebenszeit: 41 Monate, p-Wert des Mantel-Byar-Testes zum Vergleich mit den 83 Patienten ohne Remission: $> 0{,}9$). Ähnli-

ches berichteten Bonifazi et al. [19], welche Daten zu Patienten mit kompletter zytogenetischer Remission sammelten. Sie registrierten unter ihren 16 Hochrisikopatienten (New CML-Score) sechs Todesfälle, wobei jedoch nur drei der 16 Patienten länger als drei Jahre nach Remissionsbeginn beobachtet wurden. Zwischen drei und vier Jahren nach Remissionsbeginn wurde die Überlebenswahrscheinlichkeit 0,51 erreicht. Von den zum Diagnosezeitpunkt 52 Patienten der Höchstrisikogruppe gehörten 37 zur Hochrisikogruppe nach dem New CML-Score. Unter den übrigen 15 waren zehn verstorben, bei den anderen fünf wurde keine deutliche Remission dokumentiert. Mit diesen Ergebnissen zur Höchstrisikogruppe und mangels verbesserter Überlebensprognose bei Remissionserfolgen in der Hochrisikogruppe des New CML-Scores war die Konsequenz, das Eintreten von partieller oder kompletter Remission im Verlauf einer IFN-α-Therapie bei der Risikowertberechnung zu ignorieren.[25] Entsprechend der verzeichneten Resultate, sollte die Höchstrisikogruppe zeitunveränderlich die Patienten anzeigen, für welche eine Therapie mit IFN-α geringste Erfolgsaussichten hat.
Zu initial 52 Patienten der Höchstrisikogruppe kamen weitere 56 mit Hochrisiko nach dem New CML-Score (sechs wurden gemäß dem neuen Prognosesystem ursprünglich zur niedrigeren, 50 zur höheren Risikogruppe gezählt), so dass die Höchstrisikogruppe zu Therapiebeginn nun 108 Patienten umfasste.

Die deutlichen Remissionen in der Niedrigstrisikogruppe des neuen Prognosesystems

Der Risikowert „348“ ergab sich über die verschiedenen Verlaufszeitpunkte als ideale Grenze, um Patienten mit inzwischen beobachteter deutlicher Remission - und damit oft langen Überlebenszeiten - über die entsprechende Risikowertverringerung von ungünstigeren Risikogruppen in die Niedrigstrisikogruppe zu verschieben. Die Wahl dieser einheitlichen Grenze bedeutete aber auch, dass Patienten, die bei keiner der Baselinevariablen oder nur bei den Basophilen die höhere Risikogruppe innehatten[26] unabhängig von späteren Remissionsergebnissen schon zu Therapiebeginn der Niedrigstrisikogruppe angehörten. In der Lernstichprobe ergaben sich zum Diagnosezeitpunkt für 53 Patienten die Risikowerte 0 oder 348.[27]. Während aber von den 27 Patienten mit einer deutlichen Remission im späteren Therapieverlauf nur einer verstarb, gab es unter den 26 Patienten ohne deutliche Remission 9 Todesfälle (Mantel-Byar-Test: $p = 0{,}0025$)
Nach der Erweiterung der Höchstrisikogruppe im voranstehenden Abschnitt, verblieben beim neuen Prognosesystem zum Landmarkzeitpunkt „21 Monate“ 148 Patienten in der Niedrigstrisikogruppe und 237 in der Gruppe mit niedrigerem Risiko. Auch beim Vergleich der Überlebenswahrscheinlichkeiten ab der Landmark zeigte sich innerhalb der Niedrigstrisikogruppe ein statistisch signifikanter Unterschied zwischen den 119 Patienten mit deutlicher Remission innerhalb der 21 Monate (16 verstarben, 9-Jahresüberlebenswahrscheinlichkeit: 0,79) und den 29 ohne deutliche ZR (8 verstarben, die mediane Überlebenszeit wurde nach 106 Monaten erreicht), Logrank-Test: $p = 0{,}0295$. Dagegen war der Unterschied zwischen den Überlebenswahrscheinlichkeiten der 29 ohne deutliche ZR und der 237 der niedrigeren Risikogruppe (92 verstorben, mediane Überlebenszeit 81 Monate) nicht statistisch signifikant (Logrank-Test: $p = 0{,}1481$). Damit sprachen sowohl der vom Verlaufszeitpunkt unabhängige Mantel-Byar-Test als auch die Analyse zur Landmark „21 Monate“ statistisch für die Aufteilung von Patienten mit Niedrigstri-

[25] Eine einfache Veränderung der Risikogruppengrenze beim neu entwickelten Prognosesystem konnte keine ausreichend große gemeinsame Schnittmenge zwischen Patienten der höchsten Risikogruppen beider Systeme definieren.

[26] Vgl. zu den Risikogruppen Tabelle 4.8. Wegen seiner Vorläufigkeit, wurden für das erste hier vorgestellte Prognosesystem keine Patientencharakteristika der einzelnen Risikogruppen beschrieben.

[27] Patienten mit diesen beiden Werten besaßen vergleichbare Überlebenswahrscheinlichkeiten.

siko in zwei Gruppen, mit und ohne deutliche Remission.
Noch stärker wogen die medizinischen Argumente. Vom Alter her waren alle 53 Patienten für eine allogene SZT geeignet. Wenn nun zu einem gewählten Entscheidungszeitpunkt keine deutliche Remission vorliegt, wird angesichts der möglichen Alternativen, wie auch Imatinib, ein Beibehalten der bisherigen Therapie kaum in Frage kommen. Das hier zu entwickelnde Prognosesystem soll aber eine Niedrigstrisikogruppe definieren, die nur Patienten enthält, welchen unter IFN ein besonders vielversprechender Therapieverlauf prognostiziert werden kann.
Es wurde entschieden, die Niedrigstrisikogruppe prinzipiell nur aus Patienten mit deutlicher Remission unter IFN-Therapie bestehen zu lassen. Damit konnte zu IFN-Therapiebeginn kein Patient zur Niedrigstrisikogruppe gehören. Patienten, die aufgrund ihrer Baselinevariablen die Risikowerte 0 oder 348 besaßen aber noch keine deutliche Remission verzeichneten, wurden ab sofort der Gruppe „niedrigeres Risiko“ hinzugefügt.

Partielle Remissionen nach 21 Monaten

In Abschnitt 4.2.2 hatten sich mögliche Überlebensunterschiede nach partieller Remission zwischen Patienten mit erster partieller Remission vor und Patienten mit erster partieller Remission nach 21 Monaten angedeutet. Um zu untersuchen, inwieweit partielle Remissionen nach 21 Monaten im neuen Prognosesystem Berücksichtigung finden sollten, wurden die dafür relevanten Prognosegruppen „höheres Risiko“ und „niedrigeres Risiko“ in Augenschein genommen. Nach den zuletzt vorgenommenen Änderungen zum neuen Prognosesytem befanden sich zur Landmark „21 Monate“ nun 266 Patienten in der Gruppe mit niedrigerem Risiko. Nach den 21 Monaten erfuhren noch 21 (8%) eine partielle ZR, wovon anschließend 7 (33% von 21) verstarben und 2 weitere später eine komplette ZR hatten. Unter den 245 der 266 Patienten, für die nie eine partielle ZR beobachtet wurde, verstarben 91 (37% von 245) und weitere 23 erreichten nach mehr als 21 Monaten noch eine komplette ZR. Der Mantel-Byar-Test ergab keinen Unterschied in den Überlebenswahrscheinlichkeiten zwischen Patienten mit und ohne partielle ZR ($p = 0{,}1520$).[28]
Unter den 173 Patienten mit höherem Risiko hatten nur 6 Patienten eine erste partielle Remission nach mehr als 21 Monaten, 3 verstarben und 2 hatten später eine komplette ZR. Von den 167 Patienten ohne partielle Remission verstarben 103 (62% von 167), 2 weitere hatten noch eine komplette ZR. Der Mantel-Byar-Test war nicht signifikant, die Power war allerdings minimal.[29] Die Testergebnisse und die Kurven stützten den Entschluss, Patienten mit erster partieller Remission nach 21 Monaten trotz des Remissionserfolgs in ihrer bisherigen Risikogruppe zu belassen.
Weil sich beim Überleben nach erster kompletter ZR in den beiden untersuchten Gruppen kein Einfluss des Remissionszeitpunktes nach Therapiebeginn zeigte, wurde ein dadurch indizierter Gruppenwechsel beim neuen Prognosesystem weiterhin jederzeit durchgeführt.

Der Vorschlag eines vereinfachten Prognosesystems

Nach den in drei Schritten vorgenommenen Veränderungen befanden sich nun zu Therapiebeginn 389 Patienten in der Gruppe „niedrigeres Risiko“, 263 Patienten in der Gruppe „höheres Risiko“ und 108 Patienten in der Gruppe „Höchstrisiko“. Abgesehen von 64 Patienten mit Hochrisiko

[28] Die Überlebenszeiten ab einer kompletten ZR wurden wegen des Wechsels in die Niedrigstrisikogruppe zensiert.

[29] Nach den Gruppenadjustierungen in den letzen Abschnitten hatten sich die signifikanten Überlebensunterschiede zwischen den Gruppen „niedrigeres Risiko“ und „höheres Risiko“ (vgl. Abbildung 4.10) nicht geändert. Es zeigte sich für die Patienten mit erster partieller Remission nach 21 Monaten auch dann kein signifikanter Überlebensvorteil, wenn man die beiden Risikogruppen gemeinsam analysierte.

nach dem New CML-Score, konnte man nun zu Therapiebeginn für 673 der übrigen 696 Patienten (97%) eine identische Risikogruppenzuteilung erhalten, wenn man, anstatt den Risikowert nach (4.1) zu berechnen und die Grenzen 1196 und 1860 zu berücksichtigen, einfach die Anzahl der in (4.1) auf den Wert 1 zu setzenden Baselinefaktoren notierte und dann folgende Einteilung vornahm: Patienten, welchen in (4.1) bei keinem ($n = 34$), einem ($n = 95$) oder zwei Baselinefaktoren ($n = 243$) der Wert 1 zuzuordnen war, kamen in die Gruppe „niedrigeres Risiko“, Patienten mit drei ($n = 204$) oder vier ungünstigen Baselinefaktorwerten ($n = 82$) kamen in die Gruppe „höheres Risiko“ und Patienten mit Wert 1 zu allen fünf Baselinefaktoren ($n = 38$) zur Gruppe „Höchstrisiko“. Unter Beachtung oben beschriebener Erkenntnisse, wurden auch die 64 Patienten mit Hochrisiko nach dem New CML-Score (aber weniger als fünf Baselinefaktoren mit Wert 1 in (4.1)) wieder zur Höchstrisikogruppe gezählt.
Im Vergleich zur früheren Gruppeneinteilung, hatten durch die Vereinfachung nur 23 von 760 (3%) einen Risikogruppenwechsel zu Therapiebeginn zu verzeichnen. Zum einen handelte es sich dabei um 17 Patienten, die sowohl hinsichtlich der Basophile und der Eosinophile als auch beim Hämoglobinwert jeweils zur ungünstigeren Gruppe gehörten, deren aus (4.1) resultierender Risikowert 1071 nach früherer Einteilung aber im Bereich der Gruppe „niedrigeres Risiko“ lag. Nach der neuen Risikogruppenzuteilung werden die 17 Patienten, wegen dreier Baselinefaktorwerte in der jeweils ungünstigeren Risikogruppe, der Kategorie „höheres Risiko“ zugerechnet. Weiter waren 6 Patienten betroffen, welche nach Alter, Milzvergrößerung, Eosinophilen und ENTWEDER Basophilen ODER Hämoglobin den Wert 1 innehatten. Gemäß der alten Einteilung gehörten die Patienten mit den Risikowerten 1864 bzw. 1875 bereits zur Höchstrisikogruppe, nun fielen sie mit vier höheren Baselinefaktorwerten in die Gruppe „höheres Risiko“. Da keiner der sechs Patienten (vier verstarben) eine deutliche Remission verzeichnete, könnte sich die neue Risikogruppe - bei Festlegung einer Wartezeit bzgl. des Eintretens einer deutlichen ZR unter IFN-Therapie anstatt der sofortigen Wahl einer Therapiealternative - für solche Patienten als nachteilig erweisen. Schlösse man nur die Höchstrisikogruppe von einer IFN-Behandlung aus, würde sich dagegen bei den 17 vorherigen Patienten an der therapeutischen Konsequenz nichts ändern.

Der zweite Schritt zur Vereinfachung des Prognosesystems widmete sich der deutlichen ZR im Therapieverlauf. Als neue Regel für die Risikogruppenzugehörigkeit nach einer partiellen ZR innerhalb von 21 Monaten verbesserten sich alle Patienten zum Zeitpunkt ihrer ersten partiellen ZR jeweils um eine Risikogruppe in die Nächstgünstigere. Sobald für einen Patienten eine komplette ZR beobachtet wurde, erfolgte seine sofortige Versetzung in die Niedrigstrisikogruppe. Wie zuvor, bewirkten weder erste partielle ZR später als 21 Monaten nach Therapiebeginn noch deutliche Remissionen in der Höchstrisikogruppe eine Änderung der Risikogruppenzugehörigkeit.
Die vereinfachte Risikogruppenzuteilung unter Berücksichtigung deutlicher ZR führte bei allen vorliegenden Patienten zu exakt derselben Risikogruppe wie über die Berechnung des Risikowertes gemäß (4.1) und die Verwendung der drei Risikogruppengrenzen. Dies galt auch für die drei Patienten mit deutlicher ZR unter den 17 mit anfänglichem Risikowert 1071. Nach der neuen Vereinfachung ging die Verbesserung ihrer Prognose durch eine partielle ZR mit dem Wechsel in die Gruppe „niedrigeres Risiko“ einher, während sie nach der alten Einteilung zu den Patienten der Gruppe „niedrigeres Risiko“ gehörten, die auch bei partieller ZR in dieser Gruppe verblieben ($1071 - 664 = 407 > 348$). Ein weiterer Unterschied betraf wiederum die 6 Patienten mit den Risikowerten 1864 und 1875. Gemäß der neuen Einteilung kämen sie nach einer ersten partiellen ZR in die Gruppe „niedrigeres Risiko“, im Falle einer ersten kompletten ZR sogar in die Niedrigstrisikogruppe, anstatt wie zuvor unveränderlich der Höchstrisikogruppe anzugehören.

Die neue Einteilung gewichtete die fünf Baselinefaktoren gleich und stellte sowohl zum Therapiebeginn als auch im Therapieverlauf eine wesentliche und anschauliche Vereinfachung bei der Bestimmung der Prognosegruppe dar. Dieses Vorgehen wurde durch die ausreichende Vergleichbarkeit der Effektschätzer im Zusammenspiel mit den zuvor dichotomisierten Baselinevariablen ermöglicht. Wie Berechnungen von Cox-Modellen selbst unter Herausnahme von Patienten mit extremem Einfluss auf die Koeffizientenschätzer zeigten[30], wurden stets dieselben prognostischen Faktoren als statistisch signifikant identifiziert. Nun auch unabhängig vom exakten Wert der geschätzten Koeffizienten und damit ebenso von möglichen (meist kleinen) Korrekturen durch „Bootstrap resampling"-Verfahren [4] oder „Shrinkage"-Faktoren [118], führten das einfache Abzählen der Baselinefaktorwerte mit ungünstigerer Prognose gemeinsam mit den Regeln für spätere Gruppenwechsel im neuen Prognosesystem zur fast identischen Prognosegruppeneinteilung wie zuvor. Einfachheit der Risikogruppenbestimmung, Anschaulichkeit und Verständlichkeit bei zugleich hoher Übereinstimmung mit der Risikogruppenzugehörigkeit nach dem früheren System gaben den Ausschlag, für die weitere Anwendung das neue, vereinfachte Prognosesystem vorzuziehen. Der Algorithmus zur Risikogruppenbestimmung nach dem neuen Prognosesystem wird nachfolgend zusammengefasst:

1. Bestimme ob nach dem New CML-Score [42] gemäß der bei Diagnose erhobenen Variablenwerte die Zugehörigkeit zur Hochrisikogruppe vorliegt[31]
2. Weise allen Patienten, die nach dem New CML-Score zur Hochrisikogruppe gehören, beim neuen Prognosesystem die Höchstrisikogruppe zu
3. Bei Patienten, die nicht zur Hochrisikogruppe nach dem New CML-Score gehören, bestimme zu Therapiebeginn die Risikogruppe nach dem neuen Prognosesystem wie folgt:
   (a) Zähle die Anzahl n der Baselinefaktorwerte, die zum Diagnosezeitpunkt die Bedingungen i. bis v. erfüllen
      i. Das Alter in vollendeten Jahren liegt über 41
      ii. Die Milzgröße unter dem Rippenbogen beträgt mehr als 7 cm
      iii. Die Eosinophile im p.B. betragen mehr als 2%
      iv. Die Basophile im p.B. betragen mehr als 2%
      v. Der Hämoglobinwert liegt unter 11,4 g/dl bei einer Frau bzw. unter 13,6 g/dl bei einem Mann
   (b) Nehme entsprechend der Anzahl n folgende Einteilung vor:
      i. Falls $n = 0$, 1 oder 2 kommt der Patient in die Gruppe „niedrigeres Risiko"
      ii. Falls $n = 3$ oder 4 kommt der Patient in die Gruppe „höheres Risiko"
      iii. Falls $n = 5$ kommt der Patient in die Gruppe „Höchstrisiko"
4. Versetze Patienten der Gruppen „niedrigeres Risiko" oder „höheres Risiko" im Falle einer partiellen ZR innerhalb von 21 Monaten nach Therapiebeginn ab dem Zeitpunkt der ersten Remissionsfeststellung in die nächstgünstigere Risikogruppe:
   (a) Patienten der Gruppe „höheres Risiko" zu Therapiebeginn kommen in die Gruppe „niedrigeres Risiko"

[30] Siehe Abschnitt 4.4.3.

[31] Eine Berechnung ist via Internet über die Homepage http://www.pharmacoepi.de/cgi-bin/pharmacoepi/cmlscore.cgi möglich.

   (b) Patienten der Gruppe „niedrigeres Risiko“ zu Therapiebeginn kommen in die Gruppe „Niedrigstrisiko“

5. Versetze Patienten der Gruppen „niedrigeres Risiko“ oder „höheres Risiko“ im Falle einer kompletten ZR nach Therapiebeginn ab dem Zeitpunkt der Remissionsfeststellung in die Gruppe „Niedrigstrisiko“

Zu beachten ist, dass das neue Prognosesystem für Patienten entwickelt wurde, welche die in Kapitel 3 angeführten Ein- und Ausschlusskriterien erfüllen. Die Zeitpunkte der Diagnosestellung und des Therapiebeginns sollten möglichst nahe beieinander liegen. Eine deutliche Remission wird erst nach Therapiebeginn mit IFN-α berücksichtigt.

4.4.5 Die Risikogruppen des neuen Prognosesystems

Nach dem neuen Prognosesystem befanden sich zu Therapiebeginn 372 Patienten (49% von 760) in der Gruppe „niedrigeres Risiko“, 286 Patienten (38%) in der Gruppe „höheres Risiko“ und 102 Patienten (13%) in der Gruppe „Höchstrisiko“. Tabelle 4.10 beschreibt die Verteilung der Baselinevariablenwerte innerhalb der drei Gruppen. Mit steigendem Risiko nahmen in allen Fällen die Anteile ungünstigerer Werte bei den Baselineprognosefaktoren zu. In der Gruppe „niedrigeres Risiko“ lagen 56% der Patienten über dem Alterscutpoint „41 Jahre“, 22% bzw. 35% weniger als in den anderen beiden Gruppen (78% und 91%). Nur 8% der Patienten der günstigsten Prognosegruppe hatten eine stärkere Milzvergrößerung als 7 cm; die Mediane der drei Risikogruppen stiegen von 0, über 3, auf 10 cm. Wie bei der Milzvergrößerung, gab es auch bei den Eosinophilen zwischen allen drei Gruppen erkennbare prozentuale Unterschiede hinsichtlich der Werte in der ungünstigeren Baselinegruppe (8%, 31% und 75% bzw. 16%, 55% und 69%), während sich ein solcher prozentualer Unterschied bei den Basophilen und beim Hämoglobin v.a. auf die Gruppe „niedrigeres Risiko“ (jeweils 38%) vs. der beiden anderen Gruppen (86% und 89% sowie 79% und 81%) beschränkte. Insgesamt besaßen 449 Patienten (59%) einen Hämoglobinwert der ungünstigeren Gruppe, jedoch waren die geschlechtsspezifischen Anteile mit 41% Frauen $<$ 11,4 g/dl (133 von 323) und 72% Männer $<$ 13,6 g/dl (316 von 437) deutlich verschieden.[32] Die Gruppe „niedrigeres Risiko“ bestand aus 66% Niedrigrisikopatienten nach dem New CML-Score (n = 247) und 34% Patienten (n = 125) mit mittlerem Risiko. In der Gruppe „höheres Risiko“ lagen die entsprechenden Anteile bei 28% (n = 79) und 72% (n = 207). Definitionsgemäß befanden sich alle 93 Patienten mit Hochrisiko nach dem New CML-Score in der Höchstrisikogruppe des neuen Prognosesystems und machten 91% der 102 Patienten aus. Die weitere Zusammensetzung war 1% mit Niedrigrisiko (n = 1) und 8% mit mittlerem Risiko (n = 8).

Beim neuen Prognosesystem wurden Informationen aus dem Therapieverlauf berücksichtigt; die Koeffizientenschätzer der Baselinevariablen im multiplen Cox-Modell wurden durch die zytogenetische Remission mitbeeinflusst. Mit seinem Bezug zum Therapieverlauf kann der zeitabhängige Faktor zu Therapiebeginn allerdings noch nicht zur prognostischen Diskriminierung beitragen. Bei der Entwicklung des New CML-Scores wurde - durch die alleinige Konzentration auf den Baselinezeitpunkt - das prognostische Potenzial der Baselinefaktorwerte voll ausgeschöpft - ohne „einschränkende“ Korrektur durch Verlaufsinformationen. Daher bleibt für Diagnosezeitpunkt und Therapiebeginn der New CML-Score das Prognosesystem der Wahl. Speziell die Hochrisikogruppe behält eine besondere Bedeutung. Im Therapieverlauf besitzt das neue Prognosesystem

[32] Analysen hatten ergeben, dass mit der geschlechtsspezifischen Definition zweier Hämoglobingruppen der Faktor „Geschlecht“ keinerlei prognostische Bedeutung besaß. Auch univariat zeigte „Geschlecht“ im Cox-Modell keine statistische Signifikanz (Wald-Test).

Tabelle 4.10: Werte der prognostischen Baselinefaktoren zu Therapiebeginn

Risikogruppe Variable	Minimum	Median	Maximum	Mittelwert	Standardabweichung
„Niedrigeres Risiko“ ($n = 372$)					
Alter in vollen Jahren	11	44	77	45	14
Milzvergrößerung (cm)	0	0	30	2,4	4,1
Eosinophile im p.B.[a] (%)	0	1	11	1,5	1,5
Basophile im p.B. (%)	0	2	15	2,7	2,6
Hämoglobin (g/dl) - F[b] ($n = 174$)	6,4	12,6	15,5	12,2	1,6
Hämoglobin (g/dl) - M[c] ($n = 198$)	5,5	13,3	17,5	12,9	2,2
„Höheres Risiko“ ($n = 286$)					
Alter in vollen Jahren	16	51	83	50	13
Milzvergrößerung (cm)	0	3	24	5,0	5,5
Eosinophile im p.B. (%)	0	3	14	2,9	2,2
Basophile im p.B. (%)	0	5	17	5,1	2,9
Hämoglobin (g/dl) - F ($n = 100$)	5,1	11,0	16,0	11,0	2,1
Hämoglobin (g/dl) - M ($n = 186$)	4,2	12,0	15,7	11,7	1,9
„Höchstrisiko“ ($n = 102$)					
Alter in vollen Jahren	18	56	69	55	11
Milzvergrößerung (cm)	0	10	30	11,2	6,6
Eosinophile im p.B. (%)	0	4	20	4,4	3,5
Basophile im p.B. (%)	0	6	33	6,9	4,9
Hämoglobin (g/dl) - F ($n = 49$)	5,5	10,1	14,6	10,1	2,3
Hämoglobin (g/dl) - M ($n = 53$)	7,4	10,8	14,8	10,7	1,8

Werte in ungünstigerer Gruppe des jeweiligen Baselinefaktors	Alter > 41 J.	Milzver. > 7 cm	Eosino. $> 2\%$	Baso. $> 2\%$	Hämo.[d] $< 11{,}4$ oder $< 13{,}6$ g/dl
Risikogruppe	n (%)	n (%)	n (%)	n (%)	n (%)
„Niedrigeres Risiko“ ($n = 372$)	208 (56)	31 (8)	59 (16)	142 (38)	141 (38)
„Höheres Risiko“ ($n = 286$)	222 (78)	90 (31)	158 (55)	245 (86)	225 (79)
„Höchstrisiko“ ($n = 102$)	93 (91)	77 (75)	70 (69)	91 (89)	83 (81)
Alle Patienten ($n = 760$)	523 (69)	198 (26)	287 (38)	478 (63)	449 (59)

[a] Peripheres Blut.
[b] Frauen.
[c] Männer.
[d] Die erste Grenze gilt für die Frauen, die zweite Grenze für die Männer.

gegenüber dem New CML-Score mit fortschreitender Zeit dann mehr und mehr den Vorteil aktualisierter, zusätzlicher Information und damit verbundener Prognoseverbesserung.

Zum Baselinezeitpunkt befanden sich nach dem New CML-Score 327 Patienten in der Niedrigrisikogruppe (43% von 760, 92 Patienten verstorben), 340 in der mittleren Risikogruppe (45% von 760, 166 verstorben) und 93 in der Hochrisikogruppe (12% von 760, 68 verstorben). Die medianen Überlebenszeiten betrugen 98, 72 und 43 Monate. Die 9-Jahresüberlebenswahrschein-

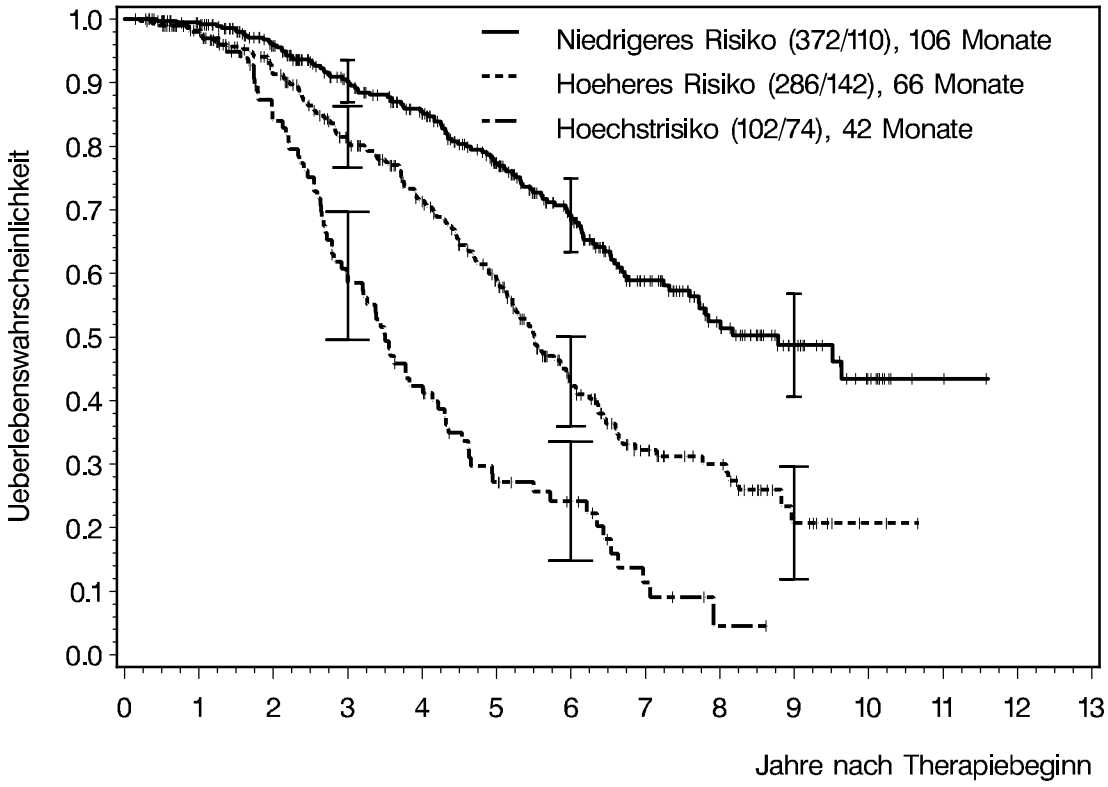

Abbildung 4.11: Kaplan-Meier-Kurven ab dem Baselinezeitpunkt „Therapiebeginn" mit geschätzten Überlebenswahrscheinlichkeiten in den drei Baselinerisikogruppen des vereinfachten neuen Prognosesystems. Die Legende „Niedrigeres Risiko (372/110), 106 Monate" bedeutet: Unter den 372 Patienten der Gruppe „Niedrigeres Risiko" wurden 110 Todesfälle beobachtet. Die mediane Überlebenszeit betrug 106 Monate. Die zwei anderen Legenden sind analog zu verstehen. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzten Wahrscheinlichkeiten mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

lichkeiten lagen in den ersten beiden Gruppen bei 0,4667 [95%-K.I.: 0,3759; 0,5576] und 0,2748 [95%-K.I.: 0,1952; 0,3544]. In der Hochrisikogruppe wurde die längste Überlebenszeit nach 8 Jahren und 8 Monaten zensiert. Die 6-Jahresüberlebenswahrscheinlichkeit war 0,2503 [95%-K.I.: 0,1526; 0,3480]. Die χ^2-Statistik des Logrank-Tests zum Vergleich der drei Überlebenskurven hatte den Wert 80,8348, was bei zwei Freiheitsgraden mit einem $p < 0{,}0001$ korrespondierte.

Abbildung 4.11 zeigt die drei Prognosegruppen nach dem neuen Prognosesystem. Von den 372 Patienten, die zu Therapiebeginn der Gruppe „niedrigeres Risiko" angehörten, verstarben 110. Die mediane Überlebenszeit betrug 106 Monate und die 9-Jahresüberlebenswahrscheinlichkeit 0,4871 [95%-K.I.: 0,4061; 0,5681]. In den beiden anderen Prognosegruppen, „höheres Risiko" und „Höchstrisiko", verstarben von 286 Patienten 142 bzw. von 102 Patienten 74. Die medianen Überlebenswahrscheinlichkeiten lagen bei 66 und 42 Monaten, die 9-Jahresüberlebenswahrscheinlichkeit der Gruppe „höheres Risiko" war 0,2075 [95%-K.I.: 0,1187; 0,2964]. Wegen der hohen Übereinstimmung mit der Hochrisikogruppe nach dem New CML-Score, änderte sich die nach 6 Jahren beobachtete Überlebenswahrscheinlichkeit bei Höchstrisiko lediglich um 0,0088

auf 0,2415 [95%-K.I.: 0,1479; 0,3350]. Die χ^2-Statistik des Logrank-Tests zum Vergleich der drei Kurven erhöhte sich auf 105,0158. Wie beim New CML-Score, ergaben sich für die Logrank-Tests zu den drei paarweisen Vergleichen von je zwei Kurven ausnahmslos p-Werte $< 0{,}0001$.

Mit mehr Patienten sowohl in der ungünstigsten als auch in der günstigsten Risikogruppe und (trotzdem) einer größeren Spanne hinsichtlich der medianen Überlebenszeiten (42 und 106 Monate vs. 43 und 98 Monate) sowie einer um ca. 30% erhöhten Logrank-Statistik schien das neue Prognosesystem dem New CML-Score auch zum Baselinezeitpunkt prognostisch überlegen zu sein. Tatsächlich liegt kein fairer Vergleich vor. Erstens profitierte die Höchstrisikogruppe des neuen Prognosesystems definitionsgemäß entscheidend von der prognostischen Stärke des New CML-Scores. Zweitens handelte es sich hier um die Lernstichprobe und die Gruppe „niedrigeres Risiko“ stützte sich über Koeffizientenschätzer und Risikogruppendefinition indirekt auf die Zusatzinformationen zur zytogenetischen Remission eben dieser Patienten.[33] Drittens gehörten alle Patienten zur Lernstichprobe für die Entwicklung des neuen Prognosesystems, jedoch nicht alle zur Lernstichprobe für die Entwicklung des New CML-Scores. Nichtsdestoweniger war im Hinblick auf das neue Prognosesystem die klar signifikante Diskriminierung dreier Risikogruppen mit ab dem Baselinezeitpunkt deutlich verschiedenen Überlebenswahrscheinlichkeiten bemerkenswert.

Abbildung 4.12 illustriert für die Landmarkanalyse ab Monat 21 die Überlebenswahrscheinlichkeiten der vier Prognosegruppen des neuen Prognosesystems. Im Vergleich zum früheren Ergebnis (vgl. Abbildung 4.10) reduzierte sich die Niedrigstrisikogruppe um 30 Patienten auf 119 Patienten (18% von 646), durch die gleichzeitige Verringerung um 9 Verstorbene von 25 auf 16 besaß sie dafür aber eine um mehr als 0,08 erhöhte 9-Jahresüberlebenswahrscheinlichkeit von 0,7494 [95%-K.I.: 0,6473; 0,8515]. Die Gruppe „niedrigeres Risiko“ setzte sich aus 256 Patienten (40%, 99 Verstorbene) zusammen. Ihre mediane Überlebenszeit lag bei 81 Monaten, die 9-Jahresüberlebenswahrscheinlichkeit bei 0,3530 [95%-K.I.: 0,2551; 0,4509]. Ein „höheres Risiko“ bestand für 187 Patienten (29%, 111 Verstorbene), mit einer medianen Überlebenszeit von 63 Monaten und einer 9-Jahresüberlebenswahrscheinlichkeit von 0,1071 [95%-K.I.: 0,0130; 0,2013]. Mit der grundsätzlichen Hinzunahme aller Hochrisikopatienten nach dem New CML-Score wuchs die Gruppe der zur Landmark verbliebenen Höchstrisikopatienten von 38 auf 84 um mehr als das Doppelte an (13%, 65 Verstorbene). Die mediane Überlebenszeit stieg kaum merklich auf 43 Monate, doch wurden ungefähr vom Ende des vierten Jahres an etwas günstigere Überlebenswahrscheinlichkeiten als in Abbildung 4.10 beobachtet. So wuchs die 6-Jahresüberlebenswahrscheinlichkeit um über 0,11 auf 0,2533 [95%-K.I.: 0,1515; 0,3552] an. Der χ^2-Wert des Logrank-Tests zum gemeinsamen Vergleich der Überlebenswahrscheinlichkeiten aller vier Risikogruppen erreichte mit 138,4122 einen ähnlichen Betrag wie zuvor (3 Freiheitsgrade, $p < 0{,}0001$).

In Tabelle 4.11 werden die Verteilungen der Baselinewerte der zeitunabhängigen prognostischen Faktoren für die nach 21 Monaten Beobachtungszeit nun vier Risikogruppen beschrieben.
Weil für Patienten der Höchstrisikogruppe kein Gruppenwechsel möglich war, wurde für die Baselinefaktoren der nach 21 Monaten verbliebenen 84 Patienten (3 wurden vorher zensiert, 6 in 1. CP allogen transplantiert, 9 verstarben) nur der jeweilige Werteanteil in der ungünstigeren Gruppe eines Baselinefaktors angegeben. Ein vergleichender Blick auf die 102 Höchstrisikopati-

[33] Auch wenn durch obige Vereinfachung beides für die Risikogruppenbestimmung des neuen Prognosesystems nicht mehr gebraucht wird, so erfolgte die Vereinfachung doch in Anlehnung an die Koeffizientenschätzer des „ursprünglichen“ Prognosesystem.

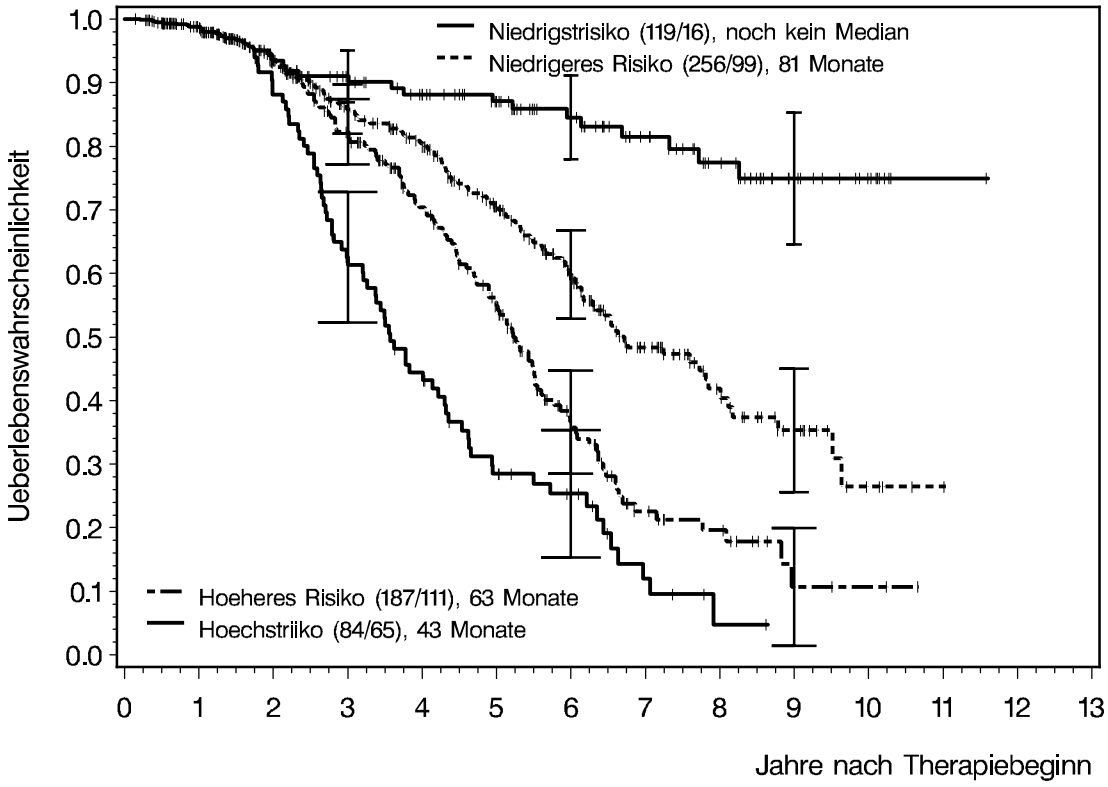

Abbildung 4.12: Kaplan-Meier-Kurven ab der Landmark „21 Monate“ mit geschätzten Überlebenswahrscheinlichkeiten in den vier Risikogruppen des neuen Prognosesystems. Bis zur Landmark „21 Monate“ wurden die Überlebenswahrscheinlichkeiten aller 760 Patienten der Lernstichprobe gemeinsam geschätzt. Ab Ende des 21. Therapiemonats wurden die verbliebenen 646 Patienten nach dem beschriebenen Algorithmus auf die vier Gruppen verteilt. Die Legende „(119/16), noch kein Median“ bedeutet: Unter den 119 Patienten wurden 16 Todesfälle beobachtet. Die mediane Überlebenszeit wurde nicht erreicht. Die drei anderen Legenden sind analog zu verstehen, wobei hier die medianen Überlebenszeiten, mit z.B. 63 Monaten in der Gruppe „höheres Risiko“, vorlagen. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzten Wahrscheinlichkeiten mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

enten in Tabelle 4.10 zeigte in allen fünf Fällen Veränderungen um maximal vier Prozentpunkte. Die Gruppe „höheres Risiko“ nahm um 35% von 286 auf 187 ab. Von den 99 „verschwundenen“ Patienten waren 44 kürzer als 21 Monate beobachtet worden (11 zensiert, 17 in 1. CP allogen transplantiert, 16 verstorben), 37 Patienten mit partieller ZR in die Gruppe „niedrigeres Risiko“ und 18 mit kompletter ZR in die Niedrigstrisikogruppe gewechselt. Paarweise Vergleiche der Baselinevariablenwerte der 187 vs. der 37 vs. der 18 Patienten ergab nur bei den 18 Patienten mit kompletter ZR vs. den 187 ohne deutliche ZR statistisch signifikant niedrigere Milzvergrößerungen (U-Test [wie in den folgenden Vergleichen]: $p = 0{,}0322$). In der Gruppe „höheres Risiko“ selbst waren nach 21 Monaten nur unwesentliche Veränderungen der Baselinewerteverteilungen feststellbar (vgl. Tabellen 4.10 und 4.11).
Die Patientenzahl der Gruppe „niedrigeres Risiko“ reduzierte sich von 372 auf 256 um 31%. Dabei standen 52 Patienten kürzer als 21 Monate unter Beobachtung (9 zensiert, 33 in 1. CP allogen

Tabelle 4.11: Verteilungen der Baselinewerte der zeitunabhängigen prognostischen Faktoren nach 21 Monaten Beobachtungszeit

Risikogruppe Variable	Minimum	Median	Maximum	Mittelwert	Standardabweichung
„Niedrigstrisiko“ ($n = 119$)					
Alter in vollen Jahren	20	43	77	46	13
Milzvergrößerung (cm)	0	0	16	1,7	3,6
Eosinophile im p.B.[a] (%)	0	2	14	1,8	1,9
Basophile im p.B. (%)	0	2	13	2,8	2,6
Hämoglobin (g/dl) - F[b] ($n = 45$)	8,1	12,7	14,7	12,3	1,5
Hämoglobin (g/dl) - M[c] ($n = 74$)	5,5	13,6	17,5	13,3	2,1
„Niedrigeres Risiko“ ($n = 256$)					
Alter in vollen Jahren	18	49	73	48	14
Milzvergrößerung (cm)	0	0	30	2,8	4,5
Eosinophile im p.B. (%)	0	1	11	1,7	1,8
Basophile im p.B. (%)	0	2	15	2,9	2,7
Hämoglobin (g/dl) - F ($n = 123$)	6,2	12,5	15,5	12,0	1,7
Hämoglobin (g/dl) - M ($n = 133$)	7,3	12,9	17,0	12,6	2,2
„Höheres Risiko“ ($n = 187$)					
Alter in vollen Jahren	16	52	75	50	13
Milzvergrößerung (cm)	0	4	24	5,2	5,4
Eosinophile im p.B. (%)	0	3	13	2,9	2,1
Basophile im p.B. (%)	0	5	17	5,3	3,0
Hämoglobin (g/dl) - F ($n = 65$)	6,4	11,0	16,0	11,2	2,2
Hämoglobin (g/dl) - M ($n = 122$)	6,6	11,9	15,3	11,6	1,8

Werte in ungünstigerer Gruppe des jeweiligen Baselinefaktors	Alter > 41 J.	Milzver. > 7 cm	Eosino. $> 2\%$	Baso. $> 2\%$	Hämo.[d] $< 11{,}4$ oder $< 13{,}6$ g/dl
Risikogruppe	n (%)	n (%)	n (%)	n (%)	n (%)
„Niedrigstrisiko“ ($n = 119$)	66 (55)	11 (9)	28 (24)	53 (45)	48 (40)
partielle ZR ($n = 60$)	28 (47)	7 (12)	10 (17)	21 (35)	23 (38)
komplette ZR ($n = 59$)	38 (64)	4 (7)	18 (31)	32 (54)	25 (42)
ehem. „nied. Risiko“ ($n = 41$)	21 (51)	2 (5)	7 (17)	18 (44)	12 (29)
ehem. „höh. Risiko“ ($n = 18$)	17 (94)	2 (11)	11 (61)	14 (78)	13 (72)
„Niedrigeres Risiko“ ($n = 256$)	165 (64)	27 (11)	53 (21)	111 (43)	110 (43)
keine ZR ($n = 219$)	137 (63)	16 (7)	31 (14)	81 (37)	81 (37)
partielle ZR ($n = 37$)	28 (76)	11 (30)	22 (59)	30 (81)	29 (78)
„Höheres Risiko“ ($n = 187$)	146 (78)	61 (33)	101 (54)	163 (87)	147 (79)
„Höchstrisiko“ ($n = 84$)	79 (94)	60 (71)	55 (65)	74 (88)	68 (81)
Alle Patienten ($n = 646$)	456 (71)	159 (25)	237 (37)	401 (62)	373 (58)

[a] Peripheres Blut.
[b] Frauen.
[c] Männer.
[d] Die erste Grenze gilt für die Frauen, die zweite Grenze für die Männer.

transplantiert, 10 verstorben). Weitere 60 hatten eine partielle und 41 eine komplette ZR und kamen daher zur Niedrigstrisikogruppe. Den insgesamt 153 Abgängen standen die 37 aus der Gruppe „höheres Risiko“ hinzugekommenen Patienten gegenüber. Die Baselinevariablenwerte der Gruppe „niedrigeres Risiko“ wurden mit Blick auf die 372 Patienten zu Therapiebeginn etwas ungünstiger. Zwischen den 60 und 41 Patienten der beiden deutlichen Remissionskategorien existierten hinsichtlich der Baselinevariablenwerte keine statistisch signifikanten Unterschiede. Im Vergleich zu den in der Gruppe „niedrigeres Risiko“ verbliebenen 219 Patienten, hatten jedoch sowohl die 60 Patienten mit partieller ZR als auch die 41 Patienten mit kompletter ZR geringere Milzvergrößerungen ($p = 0{,}0170$ und $p = 0{,}0078$). Die Patienten mit deutlicher Remission waren etwas jünger als die 219 ohne deutliche ZR, ein signifikantes Ergebnis zeitigte aber nur der Vergleich zu den 60 Patienten mit partieller ZR ($p = 0{,}0390$).
Neben den 101 früheren Patienten der Gruppe „niedrigeres Risiko“, bestand die Niedrigstrisikogruppe aus 18 früheren Patienten der Gruppe „höheres Risiko“. Diese 18 unterschieden sich von den anderen 41 Patienten mit kompletter ZR durch statistisch signifikant höhere Werte in Bezug auf Alter ($p = 0{,}0029$), Eosinophile ($p = 0{,}0011$) und Basophile ($p = 0{,}0312$). Dieselben signifikanten Variablen (mit etwas kleineren p-Werten) ergaben sich, wurden die 18 Patienten mit allen 101 ehemaligen Patienten der Gruppe „niedrigeres Risiko“ verglichen. Wegen der ungünstigeren Baselinewerte der 18 aus der Gruppe „höheres Risiko“ stammenden Patienten blieb der statistisch signifikante Unterschied im Alter erhalten, wenn man alle 59 Patienten mit kompletter ZR den 60 Patienten mit partieller ZR gegenüber stellte ($p = 0{,}0352$).
Unabhängig von der ursprünglichen Gruppe, waren im Hinblick auf alle acht näher untersuchten metrischen Baselinevariablen[34] die Werte der 119 Niedrigstrisikopatienten statistisch signifikant günstiger als bei den 187 Patienten der Gruppe „höheres Risiko“ (beim Alter: $p = 0{,}0023$, sonst: $p < 0{,}0001$). Im Vergleich zu den 256 Patienten der Gruppe „niedrigeres Risiko“ zeigten die günstigeren Werte der Niedrigstrisikogruppe noch im Falle der Variablen Milzvergrößerung, Hämoglobin bei den Männern, Thrombozyten, Leukozyten und Blasten statistisch signifikante Unterschiede (alle $p \leq 0{,}035$). Die Gruppe „höheres Risiko“ besaß in allen Belangen ungünstigere Werte als die Gruppe „niedrigeres Risiko“, außer beim Alter lag dabei immer statistische Signifikanz zugrunde (alle $p \leq 0{,}005$).
In der Regel erzielte man dieselben statistisch signifikanten oder nicht signifikanten Ergebnisse, wenn für zwei Patientengruppen, statt der metrischen Variablen, die Werte der entsprechenden, in Tabelle 4.11 angeführten dichotomen Faktoren mit dem exakten Test von Fisher verglichen wurden. Im Falle der dichtotomen Aufteilung nicht mehr statistisch signifikant waren die Milzvergrößerungen bei den Vergleichen der 18 vs. der 187, der 219 vs. der 60 und der 219 vs. der 41.[35] Beim Vergleich der 18 vs. der 41 ($p = 0{,}0037$) und der 18 vs. der 101 kamen nun statistisch signifikante Unterschiede bei Hämoglobin hinzu ($p = 0{,}0039$). Anstatt beim Alter gab es bei Betrachtung der 59 vs. der 60 den einzigen statistisch signifikanten Unterschied bei den Basophilen ($p = 0{,}0431$). Die dichotomen Definitionen nivellierten alle signifikanten Unterschiede zwischen den 119 Patienten der Gruppe „niedrigeres Risiko“ und den 256 Patienten der Niedrigstrisikogruppe. Dagegen kam beim Vergleich der Gruppen „niedrigeres Risiko“ vs. „höheres Risiko“ nun eine statistische Signifikanz für den Altersunterschied hinzu. Die p-Werte aller fünf Vergleiche der Niedrigstrisikogruppe mit der Gruppe „höheres Risiko“ hinsichtlich der Baselinefaktoren des neuen Prognosesystems lagen unter 0,0001.
Die Anteile der Niedrigrisikopatienten nach dem New CML-Score lagen in den Gruppen „Niedrigstrisiko“, „niedrigeres Risiko“ und „höheres Risiko“ bei 66%, 55% und 26%. Während sich die Anteile der ersten beiden Gruppen nach dem Fisher-Test nicht statistisch signifikant un-

[34] Es handelte sich um die acht Baselinevariablen aus Tabelle 4.2.
[35] Zur Beschreibung der hinter den Fallzahlen stehenden Patientengruppen siehe oben und Tabelle 4.11.

terschieden ($p = 0{,}0559$), führten ihre beiden Vergleiche zur Gruppe „höheres Risiko“ jeweils zu einem $p < 0{,}0001$. Hinsichtlich der Geschlechterverteilung in den Gruppen - mit steigendem Risiko wurden Männeranteile von 62%, 52% und 65% beobachtet - waren Unterschiede, jedoch kein Trend erkennbar.
Unterschiede in den Werten der prognostischen Baselinevariablen und den Anteilen an deutlichen ZR waren zwischen den Risikogruppen definitionsgemäß zu erwarten. Interessanter war die relativ große Ähnlichkeit in den Baselinevariablenwerten zwischen den Patienten mit partieller und kompletter Remission und den in ihrer Ursprungsgruppe verbliebenen Patienten ohne deutliche ZR. Dies indizierte für das Prognosesystem einen von den Baselinefaktoren weitgehend unabhängigen, zusätzlichen Informationsbeitrag durch den zytogenetischen Remissionsstatus nach 21 Monaten. In Übereinstimmung damit hatten sich bei der Modellbildung keine statistisch signifikanten Interaktionen zur unterstützenden Erklärung der Überlebenswahrscheinlichkeiten gefunden - die Einflüsse der in Abschnitt 4.3.3 identifizierten Zusammenhänge zwischen den Baselinevariablen und der zytogenetischen Remission waren durch die Hauptfaktoren und Risikogruppen offensichtlich bereits in ausreichendem Maße vertreten.
Die Überlebenswahrscheinlichkeiten der 37 neu in die Gruppe „niedrigeres Risiko“ zugeordneten Patienten mit partieller ZR (10 verstorben, 9-Jahresüberlebenswahrscheinlichkeit[36]: 0,5402) unterschieden sich nicht statistisch signifikant (Logrank-Test: $p = 0{,}1376$) von den Überlebenswahrscheinlichkeiten der in der Risikogruppe verbliebenen 219 Patienten ohne ZR (89 verstorben, mediane Überlebenszeit: 79 Monate, 9-Jahresüberlebenswahrscheinlichkeit: 0,3253). Einen statistisch signifikanten Unterschied ($p = 0{,}0276$) zeitigte aber der Vergleich der 37 mit den 60 Patienten mit partieller ZR, die von der Gruppe „niedrigeres Risiko“ in die Niedrigstrisikogruppe gewechselt waren (6 verstorben, 9-Jahresüberlebenswahrscheinlichkeit: 0,7626). Damit manifestierten sich die günstigeren Überlebenswahrscheinlichkeiten der 60 gegenüber den 37 zurecht in einer unterschiedlichen Risikogruppenzugehörigkeit. Dagegen besaßen in der Niedrigstrisikogruppe die aus der Gruppe „niedrigeres Risiko“ stammenden 60 Patienten mit partieller und 41 Patienten mit kompletter ZR (5 verstorben, 9-Jahresüberlebenswahrscheinlichkeit: 0,7941) einander sehr ähnliche Überlebenswahrscheinlichkeiten. Verglich man die Überlebenswahrscheinlichkeiten der aus der Gruppe „höheres Risiko“ kommenden 18 Patienten mit kompletter ZR (5 verstorben, 9-Jahresüberlebenswahrscheinlichkeit: 0,6025) mit jenen der obigen 41 oder mit jenen aller 101 früheren Patienten der Gruppe „niedrigeres Risiko“ (11 verstorben, 9-Jahresüberlebenswahrscheinlichkeit: 0,7811), so erhielt man keine signifikanten Ergebnisse.

Vergleich zwischen der Risikogruppenklassifikation des neuen Prognosesystems und derjenigen eines auf der Landmark 21 Monate aufgebauten Prognosesystems

Anstatt ein Cox-Modell mit zeitabhängigen Faktoren mit dem Ziel zu wählen, in der Lernstichprobe alle relevant erscheinenden Informationen auszunutzen und ein von einem vorher festgesetzten Entscheidungszeitpunkt unabhängiges Prognosesystem zu entwickeln, wurde oben die Alternative diskutiert, sich von vorneherein einen definitiven Entscheidungszeitpunkt zu überlegen und auf Basis der bis dahin erhaltenen Informationen über ein zeitunabhängiges Cox-Modell zu einem vielleicht „besseren“ Prognosesystem zu kommen. Um die Informationen zur zytogenetischen Remission zum größten Teil einzubringen, sollte in einem solchen Alternativmodell - wie zuvor bei den Kaplan-Meier-Kurven - kein allzu früher Entscheidungszeitpunkt ausgesucht werden. Der früheren Argumentation folgend, wurde die Landmark „21 Monate“ gewählt. Im

[36]Hier - wie im übrigen Abschnitt - wurden für alle Patientengruppen bis zur Landmark Ende Monat 21 die gemeinsamen Überlebenswahrscheinlichkeiten aller 760 Patienten zugrundegelegt, vgl. Abbildungen 4.10 und 4.12. Die Überlebenswahrscheinlichkeit am Ende des 21. Monats lag bei 0,9507.

Gegensatz zum Vorgehen beim neuen Prognosesystem, standen so aber nicht mehr 743 Patienten mit vollständigen Daten zu allen interessierenden Variablen zur Verfügung, sondern nur noch die Daten von 630 Patienten, die mindestens 21 Monate überlebten. Neben dem Verzicht auf Überlebenszeiten und Ereignisse von 113 Patienten, konnten auch keine unterschiedlichen Zeiten bis zur ersten deutlichen ZR innerhalb der ersten 21 Monate und v.a. keine Informationen zu deutlichen ZR danach berücksichtigt werden. Das beste prognostische Modell auf Basis der Kenntnisse Ende des 21. Monats beinhaltete dieselben Baselinevariablen mit denselben Cutpoints wie das Endmodell aus Tabelle 4.8 und auch die jeweiligen Koeffizientenschätzer wichen mit unerheblichen Veränderungen für die spätere Risikogruppeneinteilung maximal 0,0529 voneinander ab.[37] Den Unterschied machten die ebenfalls ins Modell aufgenommenen Faktoren zur zytogenetischen Remission. Während der Betrag des Koeffizienten zur kompletten Remission mit -1,2586 um fast 0,4 niedriger lag als in Tabelle 4.8, hatte sich umgekehrt, der Betrag des Koeffizienten zur partiellen Remission mit -1,0578 um fast 0,4 erhöht. Die Erklärung liegt hauptsächlich im Ignorieren der deutlichen Remissionen nach der Landmark. Im Gegensatz zum früheren Ansatz, schlugen im Cox-Modell die hohen Überlebenswahrscheinlichkeiten der Patienten, die nach 21 Monaten eine komplette Remission erfuhren, nun nicht zugunsten des Faktors zur kompletten Remission zu Buche. Dafür vergrößerten einerseits die Patienten mit partieller ZR vor und kompletter ZR nach 21 Monaten mit ihren ebenfalls hohen Überlebenswahrscheinlichkeiten das Gewicht des Koeffizientenschätzers zur partiellen ZR und andererseits nahmen sich die niedrigeren Überlebenswahrscheinlichkeiten von Patienten mit erster partieller ZR zu einem Zeitpunkt nach 21 Monaten für sein Gewicht nicht verringernd aus. Analog dem Vorgehen und den Überlegungen bei der Entwicklung des neuen Prognosesystems, wurde auch für das Landmark-Prognosesystem ein finaler Algorithmus für die Risikogruppeneinteilung gefunden. Ein Vergleich zum Zeitpunkt „21 Monate“ ergab, dass das Landmark-Prognosesystem entsprechend seiner höheren Gewichtung der partiellen ZR, 28 Patienten mit drei ungünstigeren Baselinevariablenwerten aber mit partieller ZR bis Ende Monat 21 der Niedrigstrisikogruppe zuteilte. Ihre Überlebenswahrscheinlichkeiten sprachen jedoch eher für die Zuteilung zur Gruppe „niedrigeres Risiko“, wie es das neue Prognosesystem vorschlug. Die Risikogruppeneinteilung der übrigen Patienten war zwischen beiden Systemen vergleichbar. Auf Basis der jeweils geschätzten Koeffizienten, führten dieselben Überlegungen und Vereinfachungen beim neuen Prognosesystem aber zu einer überzeugenderen prognostischen Trennung als dies beim zwar speziell auf den Zeitpunkt „21 Monate“ ausgerichtete, jedoch auf Information verzichtende Landmark-Prognosesystem der Fall war.

Illustration des neuen Prognosesystems anhand von Simon-Makuch-Kurven

Risikounterschiede frei von der Wahl eines Therapieentscheidungszeitpunktes und Überlebenswahrscheinlichkeiten unter nahezu vollständiger Ausnutzung der in der Lernstichprobe enthaltenen Information bieten die Simon-Makuch-Kurven aus Abbildung 4.13. Mit ihrer Hilfe wurden für insgesamt 758 Patienten die Überlebenswahrscheinlichkeiten in den vier Risikogruppen ab Ende des 3. Therapiemonats geschätzt. Lediglich die Daten von zwei Patienten mit einer Beobachtungszeit von weniger als drei Monaten konnten nicht berücksichtigt werden. Sobald für einen Patienten eine deutliche Remission beobachtet wurde und dadurch nach dem beschriebenen Algorithmus ein Gruppenwechsel indiziert war, erfolgte umgehend die neue Zuordnung in die günstigere Risikogruppe. Damit wurde das Wissen um das Erreichen einer deutlichen Remission zum Tag der Kenntnisnahme unverzüglich ausgenutzt und ein Patient stand immer nur dort unter Risiko, wo er entsprechend dem beschriebenen Algorithmus hingehörte. Obwohl

[37] Die Koeffizientenschätzer zu den in Tabelle 4.8 aufgeführten Baselinevariablen änderten sich zu 0,6742 (Alter), 0,5822 (Milzvergrößerung), 0,3849 (Eosinophile), 0,3880 (Hämoglobin) und 0,3419 (Basophile).

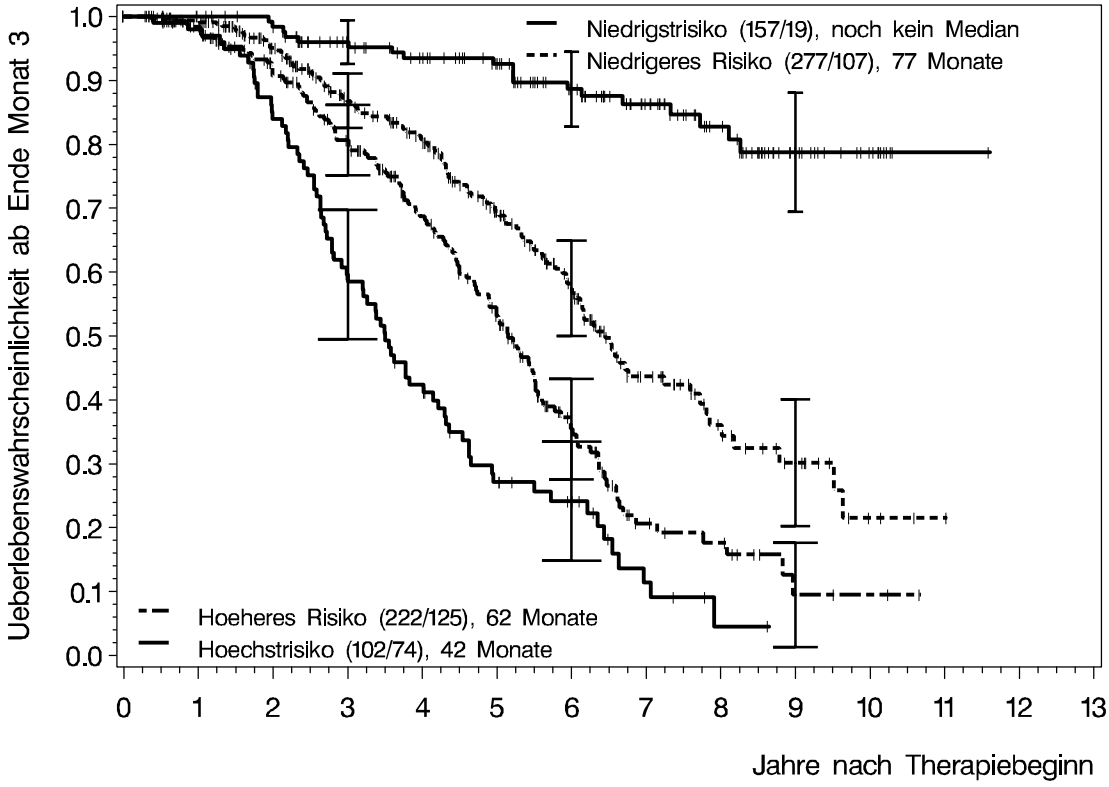

Abbildung 4.13: Simon-Makuch-Kurven zu 758 Patienten mit ab Ende des 3. Monats geschätzten Überlebenswahrscheinlichkeiten in den vier Risikogruppen des neuen Prognosesystems. Zwei Patienten wurden kürzer als 3 Monate beobachtet und mussten für die Kurvendarstellung unberücksichtigt bleiben. Bei entsprechender Indizierung durch den beschriebenen Algorithmus wechselte ein Patient - unabhängig von einer Landmark - am Tag der Beobachtung einer deutliche Remission sofort in die günstigere Risikogruppe. Daher konnten zu Therapiebeginn - außer im Falle des Höchstrisikos - für die Prognosegruppen auch keine Maximalzahlen von ausschließlich in einer Gruppe unter Risiko stehenden Patienten angeführt werden. Die Legende „(157/19)“ bedeutet: Zum Zeitpunkt des letzten Datenstandes hatten 157 Patienten die Niedrigstrisikogruppe erreicht, 19 waren danach verstorben. Analoges gilt für die anderen Gruppen. Die Monate stehen für die mediane Überlebenszeit in der jeweiligen Gruppe. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Wahrscheinlichkeit 95%-K.I. [104] berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangabe von oben nach unten.

im 1. Quartal nur 6 Patienten in die Niedrigstrisikogruppe gewechselt waren, konnten mangels Ereignisse die Überlebenswahrscheinlichkeiten der Niedrigstrisikogruppe bereits ab Ende des 3. Therapiemonats mit dem selben Ergebnis geschätzt werden, wie bei höherer Fallzahl zu späteren Zeitpunkten.

Zum Zeitpunkt des letzten Datenstandes waren 157 Patienten (21% von 760) in die Niedrigstrisikogruppe gewechselt. Nach ihrem Wechsel sind 19 davon verstorben. Die 9-Jahresüberlebenswahrscheinlichkeit lag bei 0,7872 und alle 37 Patienten, die mindestens 100 Monate (8,3 Jahre) beobachtet worden waren, befanden sich zuletzt am Leben. Von der Gruppe „niedrigeres Risiko“ stammten 118 Patienten (75% von 157) und von der Gruppe „höheres Risiko“ 39 (25%) Patienten. Die Gruppe „niedrigeres Risiko“ hatte von der Gruppe „höheres Risiko“ eine Zuwachs von 24

Patienten erhalten, so dass am Ende 278 Patienten dazugehörten (37% von 760). Es wurden 107 Todesfälle gezählt, die mediane Überlebenszeit betrug 78 Monate und die 9-Jahresüberlebenswahrscheinlichkeit 0,3015. Wegen der prognostischen Verbesserung von 63 Patienten verblieben noch 223 Patienten mit „höherem Risiko“ (29% von 760), wovon 126 verstarben. Die mediane Überlebenszeit sank auf 62 Monate und die 9-Jahresüberlebenswahrscheinlichkeit auf 0,0951.[38] Von den 102 Höchstrisiko-Patienten (13% von 760) verschieden 74. Kein Patient stand mindestens 9 Jahre unter Beobachtung. Die mediane Überlebenszeit lag bei 42 Monaten und die 6-Jahresüberlebenswahrscheinlichkeit bei 0,2415. Der Mantel-Byar-Test, mit welchem die Überlebenswahrscheinlichkeiten der vier Risikogruppen unter Berücksichtigung der Gruppenwechsel verglichen wurden, nahm im vorliegenden Fall den Wert 159,5196 an, was bei drei Freiheitsgraden einem p-Wert $< 0{,}0001$ entsprach.

Vergleich zwischen Simon-Makuch-Kurven und der Landmarkanalyse mit Kaplan-Meier-Kurven

Beide Methoden berücksichtigen bei der Berechnung von Überlebenswahrscheinlichkeiten ab Therapiebeginn zwischenzeitlich gewonnene Informationen zu zeitabhängigen Kovariablen. Im Gegensatz zur Landmarkanalyse, muss bei Simon-Makuch-Kurven keine Landmark in Abwägung zwischen Informationsgewinn und adäquater Beachtung des Sterberisikos festgelegt werden. Es genügt, einen (möglichst frühen) Startzeitpunkt zu wählen, der für alle Gruppen ausreichende Fallzahlen gewährleistet, um zu - im vorliegenden Fall vom Startzeitpunkt weitestgehend unabhängigen - Schätzungen der Überlebenswahrscheinlichkeiten zu kommen. In Abbildung 4.13 wurde dies durch den Startzeitpunkt „3 Monate“ sichergestellt. Weil bei der Landmarkanalyse nach der Landmark keine Informationen zu zeitabhängigen Kovariablen mehr berücksichtigt werden können, liegt die Landmark fast immer nach dem Startzeitpunkt der Simon-Makuch-Methode. Einen möglichst hohen Anteil an (erwartbarer) Information nutzen zu können ist bei der Landmarkanalyse wichtig, denn Nichtberücksichtigung bedeutet z.B. für alle Patienten mit erster kompletter ZR *nach der Landmark*, trotz des Therapieerfolges weiter der ungünstigeren Prognosegruppe zugerechnet zu werden, was dort i.d.R. zu einer Überschätzung von Überlebenswahrscheinlichkeiten führt. Die Simon-Makuch-Methode hat den Vorteil, dass für jeden Patienten nach dem Startzeitpunkt jederzeit die aktuelle Information zu einer zeitabhängigen Kovariablen in die Berechnung der Risikogruppen und ihrer Überlebenswahrscheinlichkeiten miteinbezogen werden kann. Alle Informationen nach dem Startzeitpunkt können so voll ausgeschöpft werden und die Risikogruppenzugehörigkeit spiegelt den neusten Datenstand wider. Damit ist auch kein „Kompromiss“ zwischen „Informationsgewinnung“ und „Patientenrisiko“ erforderlich. Simon-Makuch-Kurven erscheinen in Fällen wie dem vorliegenden als die optimalere Darstellungsform der Überlebenswahrscheinlichkeiten, jedoch korrespondiert die Landmarksituation eher mit der Praxis im Umgang mit prognostischen Modellen: Ein Arzt wählt mit einem Patienten entsprechend einer Landmark einen maximalen Zeitraum, den beide bis zu einem Therapieerfolg abzuwarten willens sind und trifft dann aufgrund der bis dahin gewonnenen Information seine Therapieentscheidung. Auch ergeben sich für die Simon-Makuch-Kurven praktische Nachteile durch die umständliche Berechnung der Überlebenswahrscheinlichkeiten. Während Kaplan-Meier-Kurven mit Hilfe einer SAS-Prozedur leicht berechenbar sind, musste für die Simon-Makuch-Kurven des Prognosesystems eigens ein SAS IML-Macro konzipiert werden (s. Anhang A.2).[39]

[38] Die beiden kürzer als drei Monate beobachteten Patienten gingen nicht in die Schätzung der Überlebenswahrscheinlichkeiten ein.

[39] Als Basis konnte dabei auf ein Macro von Clemens Biller zurückgegriffen werden, welches den Wechsel von einer Gruppe in die andere berücksichtigte. Dieses Macro wurde nun auf die vier Kurven des Prognosesystems

Ein Vergleich der Abbildungen 4.12 und 4.13 zeigt die Vorteile der Simon-Makuch-Kurven hinsichtlich der geschätzten Überlebenswahrscheinlichkeiten. Bereits zwischen den Monaten 3 und 21 war ohne Verzicht auf spätere Remissionsinformationen eine differenzierte Betrachtung der 34 Ereignisse in diesem Zeitraum und damit die Darstellung unterschiedlicher Überlebenswahrscheinlichkeiten der vier Risikogruppen möglich. Die Berücksichtigung aller 29 erstmaligen kompletten ZR nach Monat 21 ließ im Vergleich zur Landmarkanalyse die Überlebenswahrscheinlichkeiten der Gruppe „niedrigeres Risiko“ nach und nach bis auf eine um 0,05 niedrigere 9-Jahresmarke absinken (vgl. Abbildung 4.12 vs. 4.13). Bei der Landmarkanalyse waren die hohen Überlebenswahrscheinlichkeiten von 24 der 29 Patienten mit kompletter ZR statt der Niedrigstrisikogruppe der Gruppe „niedrigeres Risiko“ zugeschlagen worden. Die 9-Jahresüberlebenswahrscheinlichkeit der Niedrigstrisikogruppe lag im Falle der Simon-Makuch-Methode nicht nur wegen der 29 Patienten um 0,04 über dem Ergebnis der Landmark-Darstellung, sondern auch weil bei letzterer die Überlebenswahrscheinlichkeiten bis Monat 21 für alle vier Risikogruppen gemeinsam berechnet wurden. Da durch die Wahl dieser Landmark die Information von 83% der Patienten mit deutlicher Remission berücksichtigt werden konnte, unterschieden sich die Gesamtbilder beider Darstellungsformen nicht wesentlich, die Umständlichkeit der Überlebenswahrscheinlichkeitsberechnung bei der Simon-Makuch-Methode wurde im vorliegenden Fall nicht durch ein erheblich genaueres Ergebnis wettgemacht.

erweitert; für die Kurven selbst wurde die Darstellung von Zensierungsstrichen und von Konfidenzintervallen eingeführt.

Kapitel 5

Das neue Prognosesystem in Lern- und Validierungsstichprobe

5.1 Beurteilung des neuen Prognosesystems in der Lernstichprobe

Die p-Werte der Logrank-Statistiken zum gemeinsamen Vergleich aller vier Risikogruppen ab den Landmarkzeitpunkten 3, 6, 9, 12, 15, 18, 21 und 24 Monate nach Therapiebeginn lagen alle unter 0,0001. Zu Therapiebeginn befanden sich per definitionem keine Patienten in der Niedrigstrisikogruppe. Nach 3, 6, 9 und 12 Monaten hatten sich durch deutliche ZR die Besetzungszahlen auf 6, 24, 53 und 75 erhöht.
Maßgeblich für eine Überprüfung der Kriterien zum Vorschlag eines neuen Prognosesystems waren die Zeitpunkte zwischen 12 und 24 Monaten. Die insgesamt 30 paarweisen Vergleiche der Überlebenswahrscheinlichkeiten der vier Risikogruppen zu den fünf Landmarkzeitpunkten bis zum Ende des zweiten Jahres resultierten alle in p-Werten $< 0{,}001$. Sämtliche Anforderungen an die Kriterien zum Vorschlag eines neuen Prognosesystems wurden erfüllt.[1]

5.1.1 Prognostizierte und tatsächliche Ereigniszahlen in den Risikogruppen

Zunächst wurden die drei Koeffizienten des zeitabhängigen Cox-Modells ermittelt, bei welchem die höhere, die niedrigere und die niedrigste Risikogruppe des neuen Prognosesystems durch Dummyvariablen repräsentiert wurden. Die Höchstrisikogruppe diente als Baselinegruppe. Auf Basis der geschätzten, statistisch signifikanten Koeffizienten wurde dann die kumulierte Baselinehazardfunktion $\hat{H}_0(t)$ nach Breslow berechnet. Wie schon zuvor in Abschnitt 4.4.4, konnte $\hat{H}_0(t)$ für die ersten acht Jahre nach Therapiebeginn stabil geschätzt werden. Als Berechnungszeitraum für die Überlebenswahrscheinlichkeiten $\hat{p}_i(t, t + \lambda)$ nach dem Vorschlag (2.14) von Christensen et al. [24] bot sich wieder die Zeitspanne $\lambda = 1$ Jahr an. Mit den Dummyvariablen bestand der prognostische Index (2.13) eines Patienten zum Zeitpunkt t lediglich aus dem Koeffizientenschätzer zu der Risikogruppe, zu welcher er zum Zeitpunkt t gehörte. Geordnet nach ansteigendem Risiko, besaßen die vier Indizes die Werte -2,4133, -1,0825, -0,5341 und im Falle der Höchstrisikogruppe 0. Anstatt 55 Risikowerte mit zum Teil geringer Anzahl beobachteter Intervalle wie bei Abbildung 4.9, lagen zu den vier Indizes nun 664, 1490, 1078 und 405 betrachtete Intervalle vor. Die für die vier Risikogruppen aus der beobachteten Zahl der Ereignisse berechneten Wahrscheinlichkeiten, das nächste Jahr zu überleben, betrugen 0,9744, 0,9315, 0,8859 und

[1] Vgl. Abschnitt 2.2, a) Das neue Prognosesystem in der Lernstichprobe.

0,8173. Die auf Basis des neuen Prognosesystems mittels Cox-Modell und Baselinehazardfunktion geschätzten $\hat{p}_i(t, t + 365 \text{ Tage})$ ergaben die Überlebenswahrscheinlichkeiten 0,9657, 0,9231, 0,8789 und 0,8127. Damit unterschieden sich die für die Intervalle beobachteten und die aus dem neuen Prognosesystem geschätzten Wahrscheinlichkeiten, das nächste Jahr zu überleben, um maximal 0,0087, was für eine hohe Vorhersagekraft seiner Risikogruppen spricht.
Um alternativ die Anzahl der beobachteten mit der Anzahl der prognostizierten Ereignisse zu vergleichen, wurden über die Gegenwahrscheinlichkeit die Sterbewahrscheinlichkeiten zu obigen vier aus dem Modell geschätzten Überlebenswahrscheinlichkeiten $\hat{p}_i(t, t + 365 \text{ Tage})$ berechnet (2.16). Danach waren die Sterbewahrscheinlichkeiten mit der Anzahl der beobachteten Einjahresintervalle zu multiplizieren. Somit erhielt man als gerundete Anzahl prognostizierter Todesfälle für die Niedrigstrisikogruppe $n = 23$, für die Gruppe „niedrigeres Risiko“ $n = 115$, für die Gruppe „höheres Risiko“ $n = 131$ und für die Höchstrisikogruppe $n = 76$. Dem standen mit 17, 102, 123 und 74 die innerhalb der ersten acht Jahre tatsächlich beobachteten Zahlen an Todesfällen gegenüber. Analog dem direkten Zusammenhang zu den zuvor beschriebenen Überlebenswahrscheinlichkeiten, sind auch die Differenzen von 6, 13, 8 und 2 Todesfällen bei 664, 1490, 1078 und 405 betrachteten Intervallen als gering zu erachten.

5.1.2 Das neue Prognosesystem im Vergleich mit dem New CML-Score

Zum Vergleich der prognostischen Fähigkeiten des neuen Prognosesystems mit jenen des New CML-Scores wurden - zugunsten des New CML-Scores - nur Patienten gewählt, die schon bei der früheren Entwicklung des New CML-Scores Teil der damaligen Lernstichprobe waren. Die Schnittmenge dieser 908 Patienten [42] mit den Patienten, für welche auch ausreichend Daten zum neuen Prognosesystem vorlagen, führte zu Therapiebeginn zu 463 evaluierbaren Patienten. Wiederum mit Rücksicht auf den New CML-Score, wurden die damals vorliegenden Überlebensdaten herangezogen. Von den 463 Patienten waren 145 verstorben. Die mediane Überlebenszeit lag bei 62 Monaten und die 6-Jahresüberlebenswahrscheinlichkeit[2] bei 0,40. Tabelle 5.1 bietet für Prognosezeitpunkte innerhalb der ersten beiden Therapiejahre eine Übersicht der Ergebnisse zu den Prognosegruppen der beiden Prognosesysteme. Über alle Zeitpunkte gelang dem neuen Prognosesystem eine stärkere Trennung der 6-Jahresüberlebenswahrscheinlichkeiten. Es identifizierte Niedrigstrisiko-Patienten mit besonders hohen Überlebenswahrscheinlichkeiten, aber gleichzeitig auch mehr Patienten für die jeweils höchste Risikogruppe, letzteres allerdings dank der per definitionem entscheidenden Unterstützung durch den New CML-Score. Die Gruppen „niedrigeres Risiko“, „höheres Risiko“ und „Höchstrisiko“ entsprachen in ihren Überlebenswahrscheinlichkeiten den drei Gruppen des New CML-Scores, mit einer beim neueren System minimal stärkeren Differenzierung zum 6-Jahreszeitpunkt.[3] Dass, abgesehen von den letzten beiden Zeitpunkten, die mit der Niedrigrisikogruppe vergleichbare Gruppe „niedrigeres Risiko“ mehr Patienten umfasste und die Gruppe „höheres Risiko“, welche am ehesten der „indifferenten“ mittleren Risikogruppe entsprach, wesentlich weniger Patienten enthielt, war als ein weiterer Pluspunkt für die prognostische Differenzierungskraft des neuen Prognosesystems zu werten. Die Logrank-Statistik zum neuen Prognosesystems lag zwischen 8,76 (3 Monate) und 21,35 (24 Mo-

[2] Die Wahl der 6-Jahresüberlebenswahrscheinlichkeiten erfolgte in Konkordanz mit der bisherigen Hervorhebung dieses Zeitpunktes durch die Konfidenzintervalle in den Kurven, hier und bei Hasford et al. [42]. Zudem wurden in der Lernstichprobe des New CML-Scores nur wenige Patienten neun Jahre beobachtet; für die Gruppe mit der jeweils ungünstigsten Prognose lag zum 9-Jahreszeitpunkt bei beiden Prognosesystemen kein Patient mehr vor.

[3] Man vergleiche die Differenz zwischen Niedrig- und Hochrisikogruppe (New CML-Score) sowie zwischen der Gruppe „niedrigeres Risiko“ und der Höchstrisikogruppe (neues Prognosesystem). In 11 von 18 Fällen war beim neuen Prognosesystem außerdem die Differenz der dawischenliegenden Risikogruppe zu einer benachbarten Risikogruppe größer.

Tabelle 5.1: Vergleich des neuen Prognosesystems mit dem New CML-Score bei allen evaluierbaren Patienten aus der Lernstichprobe des New CML-Scores

Zeitpunkt ab Prognosesystem	6-Jahresüberlebenswahrscheinlichkeiten[a]	Risikogruppengrößen[b]	Logrank χ^2-Statistik		p-Wert
	$\hat{p}$	n	X^2	df[c]	p
Therapiebeginn		$n = 463$, 145 tot[d]			
New CML-Score	0,58/0,34/0,21	183/208/72	43,23	2	<0,0001
Neues Prognosesystem	n.e.[e]/0,58/0,35/0,18	n.e./204/180/79	52,01	2	<0,0001
3 Monate n.T.[f]		$n = 456$, 144 tot			
New CML-Score	0,58/0,34/0,21	179/205/72	43,57	2	<0,0001
Neues Prognosesystem	n.b.[g]/0,56/0,35/0,18	4/198/175/79	52,33	3	<0,0001
6 Monate n.T.		$n = 441$, 141 tot			
New CML-Score	0,58/0,35/0,22	169/202/70	43,15	2	<0,0001
Neues Prognosesystem	n.b./0,58/0,34/0,18	9/193/162/77	52,40	3	<0,0001
9 Monate n.T.		$n = 426$, 138 tot			
New CML-Score	0,58/0,35/0,22	164/194/68	44,35	2	<0,0001
Neues Prognosesystem	0,94/0,55/0,33/0,18	22/182/147/75	57,98	3	<0,0001
12 Monate n.T.		$n = 407$, 133 tot			
New CML-Score	0,59/0,36/0,22	155/186/66	44,93	2	<0,0001
Neues Prognosesystem	0,73/0,59/0,32/0,18	28/171/136/72	60,61	3	<0,0001
15 Monate n.T.		$n = 387$, 128 tot			
New CML-Score	0,59/0,36/0,22	145/180/62	45,64	2	<0,0001
Neues Prognosesystem	0,72/0,59/0,32/0,18	34/157/128/68	62,72	3	<0,0001
18 Monate n.T.		$n = 364$, 122 tot			
New CML-Score	0,60/0,36/0,24	136/170/58	42,03	2	<0,0001
Neues Prognosesystem	0,70/0,60/0,33/0,19	39/139/123/63	57,38	3	<0,0001
21 Monate n.T.		$n = 344$, 113 tot			
New CML-Score	0,61/0,37/0,25	128/165/51	38,19	2	<0,0001
Neues Prognosesystem	0,72/0,59/0,32/0,21	44/127/117/56	54,39	3	<0,0001
24 Monate n.T.		$n = 319$, 117 tot			
New CML-Score	0,62/0,39/0,26	117/153/49	39,78	2	<0,0001
Neues Prognosesystem	0,76/0,60/0,34/0,21	42/117/106/54	61,13	3	<0,0001

[a] Berechnet nach der Kaplan-Meier-Methode ab der jeweiligen Landmark. Die Nennung erfolgt für den New CML-Score in der Gruppenreihenfolge „Niedrigrisiko“, „mittleres Risiko“, „Hochrisiko“ und beim neuen Prognosesystem in der Gruppenreihenfolge „Niedrigstrisiko“, „niedrigeres Risiko“, „höheres Risiko“, „Höchstrisiko“.

[b] Reihenfolge der Nennung wie bei den 6-Jahresüberlebenswahrscheinlichkeiten.

[c] Freiheitsgrade.

[d] 463 Patienten mit Daten, wovon 145 verstarben. Analoge Angaben zu den übrigen Zeitpunkten.

[e] Die Abkürzung „n.e.“ steht für „nicht existent“.

[f] Abkürzung „n.T.“ steht für „nach Therapiebeginn“.

[g] Die Abkürzung „n.b.“ steht für „nicht beobachtet“. Die Überlebenszeit des letzten zensierten Patienten war kürzer als sechs Jahre.

nate) über jener des New CML-Scores. Diese Zahlen widerspiegeln ein weiteres Mal die größere prognostische Differenzierungskraft des neuen Prognosesystems im Therapieverlauf, auch wenn man den um 1 erhöhten Freiheitsgrad bedenken muss.

5.2 Beurteilung des neuen Prognosesystems in einer unabhängigen Validierungsstichprobe

Zu Beginn dieser Arbeit lagen für die 1995 und 1997 gestarteten deutschen Studien CML III [109] und CML IIIA [110] noch keine ausreichenden Beobachtungszeiten vor, um ihre Daten für Analysen in der Lernstichprobe nutzen zu können. Die seither vergangene Zeit ermöglichte es nun, die Patientendaten für eine Validierungsstichprobe in Betracht zu ziehen.
Startzeitpunkt für alle Verlaufszeiten war wieder der Beginn der IFN-α-Therapie. Im Gegensatz zu den Daten der Lernstichprobe galt für die beiden Studien zu überlegen, wie mit den Beobachtungszeiträumen unter Imatinib-Therapie umgegangen werden sollte. Die zytogenetischen Daten durften selbstverständlich nur Behandlungsintervallen mit IFN-α entstammen. Die meisten Patienten, die noch in erster chronischer Phase Imatinib erhielten, standen vorher seit geraumer Zeit unter IFN-α-Therapie (Median: 29 Monate, vgl. Tabelle 5.2) und befanden sich beim Therapiewechsel bereits in sog. „später chronischer Phase“.[4] Kantarjian et al. [62] zeigten bei Patienten in später chronischer Phase statistisch signifikant überlegene Überlebenswahrscheinlichkeiten von Imatinib-Therapie nach IFN-α-Versagen im historischen Kontrollvergleich zu Patienten mit einem IFN-α-Beginn in ebenso später Phase.[5] Auch unter Beachtung aufgetretener Verzerrungen[6] durch diesen Vergleich mit historischer Kontrolle, war ein signifikanter Überlebensvorteil bei Imatinib-Behandlung zu vermuten. Um einem verzerrenden Einfluss der neuen Therapie auf das Überleben unter IFN-α in der Validierungsstichprobe entgegenzuwirken, wurden daher die Überlebenszeiten ab Beginn einer Imatinib-Behandlung in chronischer Phase zensiert.[7] Überlebenszeiten von Patienten, die nach Versagen von IFN-α in akzelerierter Phase oder Blastenkrise eine allogene SZT erhielten, wurden ebensowenig zensiert wie die Überlebenszeiten von Patienten, deren Behandlungsbeginn mit Imatinib bereits in eine fortgeschrittene Phase fiel.

5.2.1 Die Daten der Validierungsstichprobe

Im Juni 2005 lagen zu den beiden Studien Daten zu insgesamt 1299 Patienten vor, welche die Ein- und Ausschlusskriterien der Studienprotokolle [109, 110] erfüllt hatten. Davon besaßen 524 Patienten sowohl Variablenwerte, die zusätzlich im Rahmen der in Kapitel 3 beschriebenen Ein- und Ausschlusskriterien lagen, als auch Daten zu allen Variablen, die zur Berechnung des neuen Prognosesystems erforderlich waren.

Beobachtungs- und Überlebenszeiten

Während des oben für die IFN-α-Therapie definierten Beobachtungszeitraumes waren 110 von 524 Patienten verstorben (21%), weitere 110 standen noch unter Risiko (21%). Die mediane Überlebenszeit betrug 101 Monate, die 9-Jahresüberlebenswahrscheinlichkeit lag bei 0,4803 [95%-K.I.: 0,3892; 0,5714]. Damit waren die Überlebenswahrscheinlichkeiten in der Validierungsstichpro-

[4] Während 129 von 179 (72%) unter IFN-α keine deutliche Remission erreicht hatten, wurde bei den anderen 50 die Therapie wegen eines zytogenetischen Rezidivs oder Nebenwirkungen abgesetzt.

[5] Nach ihrer Erkenntnis sind bei Patienten in später chronischer Phase die weiteren Ergebnisse der therapeutischen Zielparameter unabhängig davon, ob vorher IFN-α gegeben wurde oder nicht.

[6] Es gab große Unterschiede in den Beobachtungszeiten und Zensierungsmustern, vgl. Fig. 3 bei Kantarjian et al. [62]

[7] Ein Nichtzensieren der 179 Patienten mit Imatinib-Start in 1. CP würde für die vorliegenden 524 Patienten allerdings zu ähnlichen Überlebenswahrscheinlichkeiten führen, 9-Jahresüberlebenswahrscheinlichkeit: 0,5015 [95%-K.I.: 0,4313; 0,5717] vs. 0,4803 [95%-K.I.: 0,3892; 0,5714] bei Zensierung.

be deutlich günstiger als in der Lernstichprobe. Wegen einer SZT in erster chronischer Phase wurden 125 Patienten zensiert (24%) und wegen des Erhaltes von Imatinib in erster chronischer Phase 179 Patienten (34%). Der früheste Zeitpunkt einer Imatinib-Gabe in erster chronischer

Tabelle 5.2: Validierungsstichprobe: Beobachtete Überlebens- und Behandlungszeiten ab IFN-α-Therapiebeginn

Variable	Patientenzahl	Minimum	Median	Maximum	Mittelwert	Standardabweichung
	n	Tage	Mon.[a]	Mon.	Mon.	Mon.
Beobachtete Überlebenszeit						
Alle Patienten	524	70	30	122	38	29
Patienten noch unter Risiko[b]	110	113	67	122	63	34
Alle Patienten mit zensierten Zeiten	414	70	27	122	37	30
Zeiten unter Risiko bis allo. SZT in 1. CP[c]	125	70	14	67	17	12
Zeiten unter Risiko bis Imatinib in 1. CP	179	77	29	98	35	24
Zeiten ohne SZT oder Imatinib in 1. CP	220	113	45	122	52	30

[a] Monate.
[b] Unter Risiko auf Basis einer IFN-α-Therapie.
[c] „Allo. SZT in 1. CP“ steht für „allogene Stammzelltransplantation in 1. chronischer Phase“.

Phase war der 15.12.1999. Die mediane Beobachtungszeit betrug 30 Monate bei allen 524 Patienten und 67 Monate bei den 110 noch unter Risiko stehenden Patienten (vgl. Tabelle 5.2). Unter den 125 Transplantierten betrug die mediane Zeit bis zur SZT in erster chronischer Phase 14 Monate und bei den 179 die mediane Zeit bis zum Start von Imatinib 29 Monate. Die mediane Behandlungsdauer mit IFN-α lag für die 524 Patienten bei 18 Monaten.

Die Baselinevariablen

Tabelle 5.3 zeigt die Werteverteilung der Baselinevariablen. Vergleicht man Tabelle 5.3 mit Tabelle 4.2, so stellt man keine auffälligen Unterschiede fest, in der Validierungsstichprobe waren nur die Milzwerte etwas kleiner. Die Verteilung der Risikogruppen des New CML-Scores war weder im Vergleich mit den 1279 für die Lernstichprobe qualifizierten Patienten aus Tabelle 4.2 noch im Vergleich zu den 760 mit vollständigen Daten zur Berechnung des neuen Prognosesystems statistisch signifikant unterschiedlich.

Die Risikogruppen nach dem neuen Prognosesystem werden für die Baselinefaktoren in Tabelle 5.4 dargestellt. Weder hinsichtlich der einzelnen Faktoren noch hinsichtlich der drei Risikogruppen des neuen Prognosesystems gab es statistisch signifikante Unterschiede im Vergleich zur Lernstichprobe.[8]

[8] Vgl. Tabelle 5.4 mit Tabelle 4.10.

Tabelle 5.3: Initiale Charakteristika der Baselinevariablen in der Validierungsstichprobe

Variable	Patientenzahl	Fehlende Werte	Minimum	Median	Maximum	Mittelwert	Standardabweichung
	n	n					
Metrische Skalierung							
Alter in vollen Jahren	524	0	11	53	83	50	14
Hämoglobin in g/dl	524	0	6,4	12,2	18,8	12,1	2,2
Hämoglobin in g/dl - Frauen	190	0	6,5	11,8	18,8	11,7	2,0
Hämoglobin in g/dl - Männer	334	0	6,4	12,6	17,7	12,3	2,2
Leukozytenzahl in 10^9/l	523	1	4	93	604	128	109
Thrombozytenzahl in 10^9/l	524	0	46	395	4535	503	416
Blasten im p.B.[a] in %	524	0	0	1	10	1,6	2,2
Basophile im p.B. in %	524	0	0	3	29	3,9	3,7
Eosinophile im p.B. in %	524	0	0	2	50	2,7	3,1
Milzvergrößerung in cm[b]	524	0	0	2	30	3,8	5,3

Variable			Anzahl (Prozentualer Anteil)
	n	n	n (%)
Nominale/ordinale Skalier.			
Geschlecht	524	0	Männer: 334 (64)
New CML-Score [42]	524	0	Patienten mit Niedrigrisiko: 200 (38)
			Patienten mit mittlerem Risiko: 262 (50)
			Patienten mit Hochrisiko: 62 (12)

[a]Die Abkürzung „p.B." steht für „peripheres Blut".
[b]Die Milzvergrößerung wurde in Zentimetern unter dem linken Rippenbogen gemessen.

Tabelle 5.4: Validierungsstichprobe: Werte der prognostischen Baselinefaktoren zu Therapiebeginn

Werte in ungünstigerer Gruppe des jeweiligen Baselinefaktors	Alter > 41 J.	Milzver. > 7 cm	Eosino. $> 2\%$	Baso. $> 2\%$	Hämo.[a] $< 11{,}4$ oder $< 13{,}6$ g/dl
Risikogruppe	n (%)	n (%)	n (%)	n (%)	n (%)
„Niedrigeres Risiko" ($n = 273$)	176 (64)	11 (4)	42 (15)	95 (35)	100 (37)
„Höheres Risiko" ($n = 182$)	147 (81)	52 (29)	119 (65)	149 (82)	138 (76)
„Höchstrisiko" ($n = 69$)	59 (86)	49 (71)	55 (80)	62 (90)	54 (78)
Alle Patienten ($n = 524$)	382 (73)	112 (21)	216 (41)	306 (58)	292 (56)

[a]Die erste Grenze gilt für die Frauen, die zweite Grenze für die Männer.

5.2.2 Die Risikogruppen des neuen Prognosesystems in der Validierungstichprobe

Risikogruppenstratifizierte Überlebenswahrscheinlichkeiten ab IFN-α-Therapiebeginn

Von den 200 Patienten mit Niedrigrisiko nach dem New CML-Score waren 23 verstorben. Die mediane Überlebenszeit wurde nicht beobachtet und die 9-Jahresüberlebenswahrscheinlichkeit lag bei 0,5795 [95%-K.I.: 0,3875; 0,7714]. Für die 262 Patienten mit mittlerem Risiko wurden 64 Todesfälle beobachtet, die mediane Überlebenszeit war 89 Monate, die 9-Jahresüberlebenswahrscheinlichkeit 0,4703 [95%-K.I.: 0,3596; 0,5810]. Die Hochrisikogruppe verzeichnete 23 Verstorbene. Die mediane Überlebenszeit der 62 Patienten war nach 59 Monaten erreicht, neun Jahre wurde niemand beobachtet. Als 6-Jahresüberlebenswahrscheinlichkeit wurde 0,2848 [95%-K.I.: 0,0877; 0,4820] notiert. Die Statistik zum Logrank-Test besaß bei zwei Freiheitsgraden exakt den p-Wert 0,0001. Damit erwies sich der New CML-Score zum wiederholten Male als zuverlässig diskriminierendes Prognosesystem.

Von den 273 Patienten, die nach dem neuen Prognosesystem zu Therapiebeginn der Gruppe „niedrigeres Risiko“ angehörten, verstarben 44. Während die mediane Überlebenszeit nicht erreicht wurde, betrug die 9-Jahresüberlebenswahrscheinlichkeit 0,5276 [95%-K.I.: 0,3697; 0,6855]. In den Prognosegruppen „höheres Risiko“ und „Höchstrisiko“ verstarben 39 von 182 Patienten bzw. 27 von 69 Patienten.[9] Die medianen Überlebenswahrscheinlichkeiten zeitigten 88 und 55 Monate. Als 9-Jahresüberlebenswahrscheinlichkeit wies die Gruppe „höheres Risiko“ 0,4992 [95%-K.I.: 0,3672; 0,6312] auf. Für die Gruppe „Höchstrisiko“ konnte mit 0,2310 [95%-K.I.: 0,0556; 0,4063] nur die 6-Jahresüberlebenswahrscheinlichkeit beobachtet werden. Die χ^2-Statistik zum Logrank-Test hatte zwar einen p-Wert $< 0{,}0001$, doch deuten es die 9-Jahresüberlebenswahrscheinlichkeit an: Die Kaplan-Meier-Kurven zu den Gruppen „niedrigeres Risiko“ und „höheres Risiko“ lagen über den ganzen Beobachtungszeitraum nahe beieinander, erheblich näher als bei die beiden besten Gruppen des New CML-Scores. Letzterer unterstrich damit, zum Baselinezeitpunkt das Prognosesystem der Wahl zu stellen. Der hohe Wert der Teststatistik beim neuen Prognosesystem ging auf das Konto der ungünstigen Höchstrisikogruppe. Zum Baselinezeitpunkt ist es die Hauptaufgabe des neuen Prognosesystems, Patienten der Höchstrisikogruppe klar zu identifizieren. Dieser Aufgabe wurde auch in der Validierungsstichprobe erfüllt.[10]

Die zeitabhängige Kovariable zytogenetische Remission

Von den 524 Patienten erreichten 80 eine komplette ZR (15%) und 88 eine partielle ZR (17%). Maximal eine geringe ZR wurde bei 54 (10%) der Patienten beobachtet und eine nur minimale ZR bei 110 (21%). Keine ZR erzielten 192 Patienten (37%). Für 24 der 80 Patienten mit kompletter ZR wurde zuvor eine partielle ZR registriert (30% von 80). Die mediane Zeit bis zur Beobachtung des Eintritts einer partiellen ZR lag für die 112 (88+24) Patienten bei 11 Monaten [Maximum: 5 Jahre und 2 Monate]. Eine erste komplette Remission wurde im Median nach 15 Monaten Therapiedauer beobachtet [Maximum: 8 Jahre und 10 Monate]. Die Wahrscheinlich-

[9] Innerhalb der Hochrisikogruppe des New CML-Scores (62 Patienten) war auch in der Validierungsstichprobe kein möglicher Überlebensvorteil für die Patienten mit deutlicher Remission zu erkennen (vier von neun verstorben, mediane Überlebenszeit nach der Simon-Makuch-Kurve ab 7 Monaten: 43 Monate, vgl. Abschnitt 4.4.4.)

[10] Ohne einen bis drei Monate nach Therapiebeginn zensierten, für den Zielparameter einflusslos gebliebenen Patienten, sind die Überlebenswahrscheinlichkeiten der Patienten der Höchstrisikogruppe in Abbildung 5.2 beschrieben.

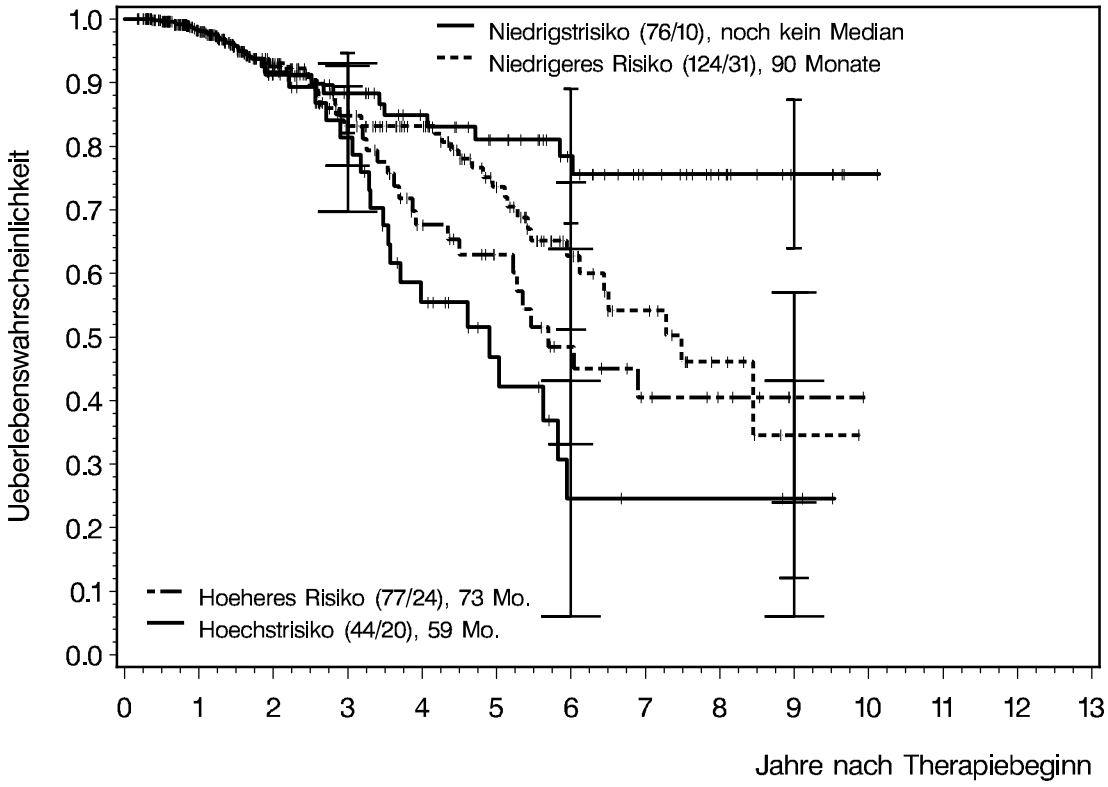

Abbildung 5.1: Kaplan-Meier-Kurven ab der Landmark „21 Monate“ mit geschätzten Überlebenswahrscheinlichkeiten in den vier Risikogruppen des neuen Prognosesystems. Bis zur Landmark „21 Monate“ wurden die Überlebenswahrscheinlichkeiten aller 524 Patienten der Validierungsstichprobe gemeinsam geschätzt. Ab Ende des 21. Therapiemonats wurden die verbliebenen 321 Patienten nach dem zuvor beschriebenen Algorithmus auf die vier Gruppen verteilt. Die Legende „(76/10), noch kein Median“ bedeutet: Unter den 76 Patienten wurden 10 Todesfälle beobachtet. Die mediane Überlebenszeit wurde nicht erreicht. Die drei anderen Legenden sind analog zu verstehen, wobei hier die medianen Überlebenszeiten, mit z.B. 90 Monaten in der Gruppe „niedrigeres Risiko“, vorlagen. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzten Wahrscheinlichkeiten mit Hilfe der Greenwood-Formel [36, 40] 95%-K.I. berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangaben von oben nach unten.

keiten, zu den Zeitpunkten 12, 15, 18, 21 und 24 Monaten eine ZR beobachtet zu haben, lagen für die partielle Remission bei 0,16, 0,20, 0,23, 0,25 und 0,26, im Falle der kompletten Remission bei 0,07, 0,11 0,15, 0,17 und 0,17.

Risikogruppenstratifizierte Überlebenswahrscheinlichkeiten ab dem Landmarkzeitpunkt 21 Monate

Die Landmark „21 Monate nach Therapiebeginn“ hatte sich nach den Daten der Lernstichprobe als optimaler Entscheidungszeitpunkt empfohlen. Bei der Validierungsstichprobe fielen 97 der 112 registierten partiellen ZR (87%) und 61 der 80 kompletten ZR (76%) in den Zeitraum der ersten 21 Monate. Die Überlebenswahrscheinlichkeiten ab Monat 21 werden für die vier Prognosegruppen des neuen Prognosesystems in Abbildung 5.1 dargestellt. In der Niedrigstrisi-

kogruppe, welche für 76 Patienten (24% von 324) verzeichnet wurde, verstarben 10 Patienten. Die 9-Jahresüberlebenswahrscheinlichkeit lag bei 0,7562 [95%-K.I.: 0,6425; 0,8699]. Die Gruppe „niedrigeres Risiko“ bestand aus 124 Patienten (39% von 324). Es wurden 31 Todesfälle, eine mediane Überlebenszeit von 90 Monaten und eine 9-Jahresüberlebenswahrscheinlichkeit von 0,3456 [95%-K.I.: 0,1044; 0,5868] beobachtet. Zur Gruppe „höheres Risiko“ gehörten 77 Patienten (24%), wobei 24 verstarben und eine mediane Überlebenszeit von 73 Monaten sowie eine 9-Jahresüberlebenswahrscheinlichkeit von 0,4052 [95%-K.I.: 0,2349; 0,5755] erreicht wurden.[11] In der Höchstrisikogruppe verblieben 44 Patienten (14%). Zwanzig Patienten verstarben. Die mediane Überlebenszeit betrug 59 Monate und die 6-Jahresüberlebenswahrscheinlichkeit 0,2459 [95%-K.I.: 0,0484; 0,4434]. Der Wert der χ^2-Statistik zum Logrank-Test korrespondierte mit einem $p < 0{,}0001$.
Abgesehen vom ähnlichen Resultat bei den Niedrigstrisikopatienten, bei welchen wie in der Lernstichprobe in den späteren Jahren ein Plateau zu beobachten war (Abbildung 5.1), widerspiegeln die Ergebnisse in den Risikogruppen, dass die Überlebenswahrscheinlichkeiten in der Validierungsstichprobe (mit ansteigendem Gruppenrisiko: 35% und 41% nach neun sowie 25% nach sechs Jahren, mediane Überlebenszeiten: 90, 73 und 59 Monate) insgesamt höher waren als in der Lernstichprobe (35% und 11% nach neun sowie 25% nach sechs Jahren, mediane Überlebenszeiten: 81, 63 und 43 Monate). Ein Grund für die insgesamt günstigeren Überlebenswahrscheinlichkeiten der Validierungsstichprobe könnten die in den späteren Studien optimalere (konsequentere) Dosierung und Komedikation sein. Obwohl aufgrund der geringen Ereigniszahl sehr große Konfidenzintervalle vorlagen, zeigt Abbildung 5.1 außerdem, dass sich die Überlebenswahrscheinlichkeiten der Niedrigstrisikogruppe ab dem Zeitpunkt „vier Jahre“ deutlich von jenen der anderen Gruppen unterschieden. Die gemeinsame Beschreibung aller Überlebenswahrscheinlichkeiten bis zur Landmark „21 Monate“ verhinderte eine frühere Trennung, wie sie die quasi „landmarkunabhängige“ Simon-Makuch-Berechnung ab drei Monaten (ohne vorheriges Ereignis) in Abbildung 5.2 klar erkennen läßt. In den ersten Jahren fand hier keine Kurvenüberkreuzung, sondern nur mehr eine Berührung statt. Beiden Abbildungen 5.1 und 5.2 gemein blieb die Überkreuzung der Kurven der mittleren Risikogruppen nach acht Jahren. Der Sprung in der Überlebenswahrscheinlichkeit wegen des einen Ereignisses nach 8,5 Jahren und die großen Konfidenzintervalle verdeutlichen es: zum Zeitpunkt des letzten Ereignisse standen nur noch wenige Patienten unter Beobachtung, der Kurvenüberkreuzung war somit wenig Aussagekraft beizumessen. Bedenkt man die relativ geringe Ereigniszahl, kann man in der Validierungsstichprobe insgesamt von einer guten Diskriminierung des neuen Prognosessystems in vier Risikogruppen sprechen.

Die Verteilungen der Baselinewerte der zeitunabhängigen prognostischen Faktoren nach 21 Monaten Beobachtungszeit werden für die vier Risikogruppen in Tabelle 5.5 beschrieben. Hinsichtlich der Unterschiede bei den Baselinevariablen verzeichnete man ähnliche Ergebnisse wie zuvor in der Lernstichprobe.[12] Außer beim Alter, besaß die Niedrigstrisikogruppe im Vergleich zur Gruppe „höheres Risiko“ signifikant höhere Anteile an Werten in den günstigeren Gruppen der Baselinefaktoren (alle $p \leq 0{,}0001$). Dasselbe signifikante Ergebnis war beim Vergleich zwischen den Gruppen „niedrigeres“ und „höheres Risiko“ zu finden, nur dass in der günstigeren Ri-

[11] Bei 10 der 124 und bei 3 der 77 der beiden mittleren Risikogruppen wurde nach 21 Monaten noch eine erste partielle ZR verzeichnet. Zwei der 10 verstarben danach, was wie in der Lernstichprobe eher dafür sprach, dass eine erste partielle ZR nach 21 Monaten später keinen Überlebensvorteil bietet. Allerdings war die Power, mittels eines Mantel-Byar-Testes einen tatsächlich vorhandenen Unterschied erkennen zu können, wiederum sehr gering (vgl. Abschnitt 4.4.4).

[12] Vgl. Tabelle 5.5 mit Tabelle 4.11.

Tabelle 5.5: Werte in ungünstigerer Gruppe des jeweiligen prognostischen Baselinefaktors nach 21 Monaten Beobachtungszeit in der Validierungsstichprobe

	Alter	Milzver.	Eosino.	Baso.	Hämo.[a]
					$<$ 11,4 oder
	$>$ 41 J.	$>$ 7 cm	$>$ 2%	$>$ 2%	$<$ 13,6 g/dl
Risikogruppe	n (%)	n (%)	n (%)	n (%)	n (%)
„Niedrigstrisiko“ ($n = 76$)	63 (83)	3 (4)	17 (22)	40 (53)	27 (36)
partielle ZR ($n = 28$)	21 (75)	0 (0)	7 (25)	12 (43)	7 (25)
komplette ZR ($n = 48$)	42 (88)	3 (6)	10 (21)	28 (58)	20 (42)
ehem. „nied. Risiko“ ($n = 29$)	23 (79)	1 (3)	3 (10)	10 (34)	5 (17)
ehem. „höh. Risiko“ ($n = 19$)	19 (100)	2 (11)	7 (37)	18 (95)	15 (79)
„Niedrigeres Risiko“ ($n = 124$)	96 (77)	9 (7)	24 (19)	49 (40)	49 (40)
keine ZR ($n = 102$)	77 (75)	5 (5)	10 (10)	29 (28)	35 (34)
partielle ZR ($n = 22$)	19 (86)	4 (18)	14 (64)	20 (91)	14 (64)
„Höheres Risiko“ ($n = 77$)	69 (90)	22 (29)	46 (60)	63 (82)	54 (70)
„Höchstrisiko“ ($n = 44$)	41 (93)	34 (77)	35 (80)	40 (91)	35 (80)
Alle Patienten ($n = 321$)	269 (84)	68 (21)	122 (38)	192 (60)	165 (51)

[a]Die erste Grenze gilt für die Frauen, die zweite Grenze für die Männer.

sikogruppe auch noch das Alter signifikant geringer war ($p = 0{,}0367$). Die Werteverteilungen der Niedrigstrisikogruppe und der Gruppe „niedrigeres Risiko“ waren für keinen der Baselinefaktoren statistisch signifikant unterschiedlich. Neben den definitionsgemäß erwartbaren Unterschieden, wurden dagegen auch in der Validierungsstichprobe vergleichbare Werteverteilungen zwischen Patienten mit deutlicher ZR und den in den ursprünglichen Gruppen Verbliebenen beobachtet. Innerhalb der beiden besten Risikogruppen konnten zwischen den Patienten mit den unterschiedenen Stati zytogenetischer Remission keine signifikanten Unterschiede zwischen den Überlebenswahrscheinlichkeiten festgestellt werden, allerdings war in den Untergruppen die Zahl der Ereignisse sehr klein.

Risikogruppenstratifizierte Überlebenswahrscheinlichkeiten - die Landmarkzeitpunkte zwischen 12 und 24 Monaten

Die Besetzungszahlen der Niedrigstrisikogruppe lagen mit 55 Patienten erst ab dem Monat 12 über 10%. Zuvor waren zu den Landmarkzeitpunkten 6 und 9 Monate 23 (4,6% von 497) und 43 (9,2% von 465) Patienten beobachtet worden. Der gemeinsame Vergleich der Überlebenswahrscheinlichkeiten aller vier Risikogruppen resultierte immer in p-Werten $< 0{,}0001$. Die vier Kaplan-Meier-Kurven zu den Überlebenswahrscheinlichkeiten der vier Risikogruppen befanden sich generell immer in der richtigen Reihenfolge - je günstiger die Risikogruppe, desto höher waren die Überlebenswahrscheinlichkeiten.
Für die insgesamt 30 paarweisen Überlebenszeitvergleiche der Risikogruppen zu den Zeitpunkten 12, 15, 18, 21 und 24 Monate nach Therapiebeginn wurden in 21 Fällen p-Werte $< 0{,}05$ beobachtet. Die Niedrigstrisikogruppe unterschied sich von der Höchstrisikogruppe immer mit p-Werten $\leq 0{,}0001$. Die Vergleiche der Niedrigstrisikogruppe vs. der Gruppe „höheres Risiko“ führte nach 12 Monaten zu p=0,0059 und danach zu p-Werten $< 0{,}001$. Die beste Risikogruppe verglichen mit der zweitbesten Risikogruppe ergab p-Werte $< 0{,}05$. Alle Vergleiche zwischen

den Gruppen „niedrigeres Risiko“ und „Höchstrisiko“ resultierten in p-Werten $< 0{,}01$. Dagegen waren die Vergleiche „niedrigeres Risiko“ vs. „höheres Risiko“ nie signifikant. Der p-Wert im Vergleich der beiden ungünstigsten Gruppen lag zum Monat 12 bei 0,0396, die vier p-Werte ab Monat 15 waren nicht signifikant.

Das neue Prognosesystem hat in der Validierungsstichprobe somit gezeigt, dass es prinzipiell vier Risikogruppen zu unterscheiden vermag. Die wichtigste Unterscheidung, die Identifikation einer Niedrigstrisikogruppe von Patienten mit deutlicher ZR unter IFN-α, hat sehr gut funktioniert. Die Trennung zwischen Patienten der Gruppen „niedrigeres Risiko“ und „höheres Risiko“ war jedoch nicht zufriedenstellend. Ein entscheidender Grund dafür waren sicher zu niedrige Fallzahlen und zu kurze Beobachtungszeiten, was zu einer sehr geringen Power führte. Um die Power mit Hilfe des Programmes „PS Power and Sample Size Calculations“[13] [32, 33, 100] abzuschätzen, wurde zur Landmark „21 Monate“ von folgenden Gegebenheiten ausgegangen: Als „Control group“ wurden in „PS Power“ die 77 Patienten der Gruppe „höheres Risiko“ gewählt (vgl. Abbildung 5.1), für die eine mediane Überlebenszeit von 63 Monaten (vgl. Abbildung 4.12) erwartet wurde. Entsprechend bildeten die 124 Patienten der Gruppe „niedrigres Risiko“ mit der medianen Überlebenszeit von 81 Monaten die „Experimental group“. Das Verhältnis der kleineren zur größeren Gruppe lag somit bei 0,62. Als Rekrutierungszeit wurden 84 Monate[14] angenommen und als zusätzliches Follow-up 36 Monate[15]. Unter den gegebenen Annahmen wurde für $\alpha = 0{,}05$ die Power von 0,2371 errechnet.
Unter den 760 Patienten der Lernstichprobe, für die das neue Prognosesystem berechenbar war, betrug die mediane Beobachtungszeit der 434 zensierten Patienten 58 Monate. Der Anteil der Patienten, die wegen SZT in erster chronischer Phase zensiert wurden, lag mit 130 Patienten bei 17%. Dagegen wurden die 414 zensierten Patienten der Validierungsstichprobe im Median nur 27 Monate beobachtet (vgl. Tabelle 5.2) und der Anteil der 125 Patienten mit SZT in erster chronischer Phase machte 24% aus (χ^2-Test zum Anteilsvergleich: $p = 0{,}0029$).[16] Nach einem Update, v.a. der Studie CML IIIA, und weiteren Daten zu IFN-α-behandelten Patienten der italienischen Studiengruppe sollten in naher Zukunft die Power-Probleme überwunden und definitivere Aussagen über den Vergleich der beiden mittleren Risikogruppen des neuen Prognosesytems möglich sein.

Risikogruppenstratifizierte Überlebenswahrscheinlichkeiten nach Simon-Makuch

Mittels Simon-Makuch-Kurven bietet Abbildung 5.2 für 521 Patienten eine Beschreibung der Überlebenswahrscheinlichkeiten in den vier Risikogruppen ab Ende des 3. Therapiemonats. Die Daten zu drei Patienten mit einer Beobachtungszeit von weniger als drei Monaten konnten nicht berücksichtigt werden. Zum Zeitpunkt des letzten Datenstandes befanden sich 120 Patienten (23% von 524) in der Niedrigstrisikogruppe. Elf von ihnen sind danach verstorben. Die 9-Jahresüberlebenswahrscheinlichkeit lag bei 0,7937. Von der Gruppe „niedrigeres Risiko“ stammten 93 Patienten (78% von 120) und von der Gruppe „höheres Risiko“ 27 (23%) Patienten. Die Gruppe „niedrigeres Risiko“ hatte von der Gruppe „höheres Risiko“ einen Zuwachs von 24 Patienten erhalten, so dass am Ende 203 Patienten dazugehörten (39% von 524). Es wurden 38

[13] Das Programm ist verfügbar unter: http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize.
[14] Bis auf vier „Ausreißer“ wurden die 201 Patienten zwischen Januar 1995 und Januar 2002 rekrutiert.
[15] Die längste Beobachtung eines im Januar 1995 rekrutierten Paienten war 120 Monate.
[16] Der Verzicht auf eine Zensierung der 179 mit Imatinib behandelten Patienten hätte übrigens mit dann 40 Monaten medianer Beobachtungszeit unter 389 zensierten Patienten keine wesentlichen Auswirkungen auf die Ergebnisse in der Validierungsstichprobe gehabt.

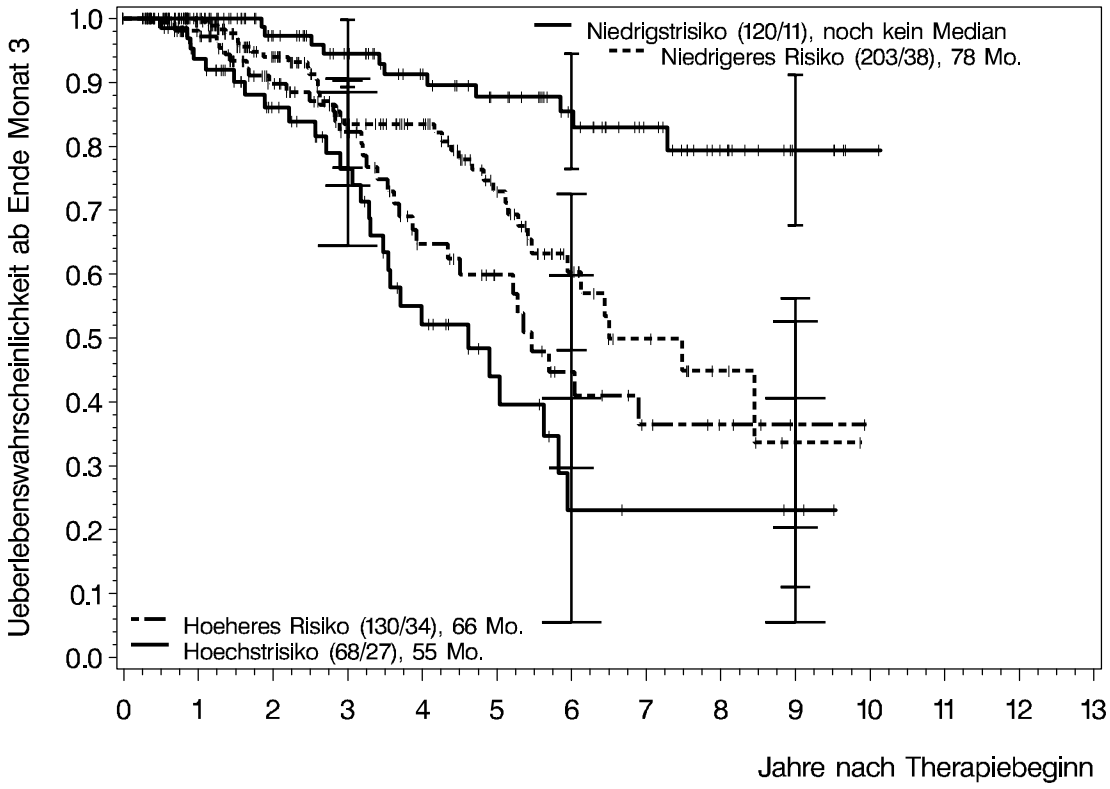

Abbildung 5.2: Simon-Makuch-Kurven zu 521 Patienten mit ab Ende des 3. Monats geschätzten Überlebenswahrscheinlichkeiten in den vier Risikogruppen des neuen Prognosesystems. Drei Patienten wurden kürzer als 3 Monate beobachtet und mussten für die Kurvendarstellung unberücksichtigt bleiben. Bei entsprechender Indizierung durch den Algorithmus des Prognosesystems wechselte ein Patient - unabhängig von einer Landmark - am Tag der Beobachtung einer deutliche Remission sofort in die günstigere Risikogruppe. Daher konnten zu Therapiebeginn - außer im Falle des Höchstrisikos - für die Prognosegruppen auch keine Maximalzahlen von ausschließlich in einer Gruppe unter Risiko stehenden Patienten angeführt werden. Die Legende „(120/11)“ bedeutet: Zum Zeitpunkt des letzten Datenstandes hatten 120 Patienten die Niedrigstrisikogruppe erreicht, 11 waren danach verstorben. Analoges gilt für die anderen Gruppen. Die Monate stehen für die mediane Überlebenszeit in der jeweiligen Gruppe. Zu den Zeitpunkten 3, 6 und 9 Jahre wurden um die geschätzte Wahrscheinlichkeit 95%-K.I. [104] berechnet. Die Länge der horizontalen Abschlusslinien für die in die Kurven eingezeichneten 95%-K.I. wächst mit der Reihenfolge der Legendenangaben von oben nach unten.

Todesfälle gezählt, die mediane Überlebenszeit betrug 78 Monate und die 9-Jahresüberlebenswahrscheinlichkeit 0,3365.[17] Wegen der prognostischen Verbesserung von 51 Patienten verblieben noch 130 Patienten mit „höherem Risiko“ (25% von 524), wovon 34 verstarben. Die mediane Überlebenszeit war 66 Monate und die 9-Jahresüberlebenswahrscheinlichkeit 0,3645. Von den 68 Höchstrisiko-Patienten (13% von 524) verschieden 27. Die mediane Überlebenszeit lag bei 55 Monaten und die 9-Jahresüberlebenswahrscheinlichkeit bei 0,2310. Beim Mantel-Byar-Test [75] nahm die χ^2-verteilte Statistik den Wert 37,7284 an, was bei drei Freiheitsgraden einem p-Wert $<$ 0,0001 entsprach. Die zwei günstigeren Risikogruppen waren sich im Vergleich von Lern- (Abbil-

[17] Die drei kürzer als drei Monate beobachteten Patienten gingen nicht in die Schätzung der Überlebenswahrscheinlichkeiten ein, hätten aber auch keinerlei Einfluss gehabt, da sie ohne deutliche Remission erreicht zu haben in erster CP transplantiert wurden.

dung 4.13) und Validierungsstichprobe (Abbildung 5.2) in ihrem Ergebnis sehr ähnlich, dagegen hatten die zwei schlechteren Prognosegruppen nun etwas günstigere Überlebenswahrscheinlichkeiten. Gemessen an der wesentlich geringeren Ereigniszahl (110 verstorbene Patienten) und den kürzeren Beobachtungszeiten, war das in der Validierungsstichprobe beobachtete Ergebnis zur Trennung der Überlebenswahrscheinlichkeiten im Vergleich zur Lernstichprobe (325 verstorbene Patienten) durchaus zufriedenstellend.

5.2.3 Prognostizierte und tatsächliche Ereigniszahlen in den Risikogruppen

Zunächst wurden die drei Koeffizienten des zeitabhängigen Cox-Modells ermittelt, bei welchem die höhere, die niedrigere und die niedrigste Risikogruppe des neuen Prognosesystems durch Dummyvariablen repräsentiert wurden. Die Höchstrisikogruppe diente als Baselinegruppe.
Auf Basis der in der Lernstichprobe geschätzten Koeffizienten wurde in der Validierungsstichprobe die kumulierte Baselinehazardfunktion $\hat{H}_0(t)$ nach Breslow berechnet. Entsprechend der Beobachtungszeit, konnte $\hat{H}_0(t)$ für die ersten sechs Jahre nach Therapiebeginn stabil geschätzt werden. Als Berechnungszeitraum für die Überlebenswahrscheinlichkeiten $\hat{p}_i(t, t+\lambda)$ wurde wieder die Zeitspanne $\lambda = 1$ Jahr gewählt. Zu den vier Risikogruppen lagen, geordnet nach ansteigendem Risiko, 290, 774, 468 und 223 Beobachtungsintervalle vor. Von der Niedrigstrisikogruppe bis hin zur Höchstrisikogruppe besaßen die aus den beobachteten Ereigniszahlen geschätzten Überlebenswahrscheinlichkeiten die Werte 0,9724, 0,9561, 0,9316 und 0,8789. Demgegenüber standen die über das Prognosesystem geschätzten $\hat{p}_i(t, t+365$ Tage) 0,9764, 0,9485, 0,9196 und 0,8540. Für die beiden niedrigeren Risikogruppen waren die Unterschiede zwischen den Wahrscheinlichkeiten mit 0,0040 und 0,0076 noch etwas geringer als in der Lernstichprobe. Dafür drückten sich die tatsächlich günstigeren Überlebenswahrscheinlichkeiten der höheren Risikogruppen in den Wahrscheinlichkeitsunterschieden 0,0120 und 0,0249 aus.
Die Anzahl der beobachteten Einjahresintervalle multipliziert mit den Sterbewahrscheinlichkeiten ergab als gerundete Zahl prognostizierter Todesfälle für die Niedrigstrisikogruppe $n = 7$, für die Gruppe „niedrigeres Risiko“ $n = 40$, für die Gruppe „höheres Risiko“ $n = 38$ und für die Höchstrisikogruppe $n = 32$. Die in den ersten sechs Jahren tatsächlich beobachteten Zahlen an Todesfällen lagen bei 8, 34, 32 und 27, womit die Differenzen 1, 6, 6 und 5 betrugen. Bedenkt man die angesprochene Problematik der relativ kurzen Beobachtungszeiten, sind dies akzeptable Ergebnisse.

5.2.4 Das neue Prognosesystem im Vergleich mit dem New CML-Score

Wie im Falle der gemeinsamen Teilmenge der Lernstichprobe, wurden 6-Jahresüberlebenswahrscheinlichkeiten miteinander verglichen. Die 6-Jahresüberlebenswahrscheinlichkeit der 524 Patienten der Validierungsstichprobe betrug 0,59.

Tabelle 5.6: Vergleich des neuen Prognosesystems mit dem New CML-Score bei allen evaluierbaren Patienten aus der Validierungstichprobe

Zeitpunkt ab Prognosesystem	6-Jahresüberlebenswahrscheinlichkeiten[a]	Risikogruppengrößen[b]	Logrank χ^2-Statistik		p-Wert
	$\hat{p}$	n	X^2	df[c]	p
Therapiebeginn		$n = 524$, 110 tot[d]			
New CML-Score	0,75/0,57/0,28	200/262/62	18,40	2	0,0001
Neues Prognosesystem	n.e.[e]/0,68/0,60/0,23	n.e./273/182/69	19,70	2	<0,0001
3 Monate n.T.[f]		$n = 521$, 110 tot			
New CML-Score	0,75/0,57/0,28	199/261/61	18,40	2	0,0001
Neues Prognosesystem	n.b.[g]/0,68/0,60/0,23	8/266/179/68	19,83	3	0,0002
6 Monate n.T.		$n = 497$, 109 tot			
New CML-Score	0,75/0,57/0,28	186/251/60	18,56	2	<0,0001
Neues Prognosesystem	0,85/0,66/0,61/0,23	23/244/163/67	21,28	3	<0,0001
9 Monate n.T.		$n = 465$, 106 tot			
New CML-Score	0,76/0,57/0,29	170/240/55	17,68	2	0,0001
Neues Prognosesystem	0,86/0,65/0,58/0,23	43/222/138/62	22,61	3	<0,0001
12 Monate n.T.		$n = 420$, 102 tot			
New CML-Score	0,76/0,57/0,31	149/222/49	12,34	2	0,0021
Neues Prognosesystem	0,88/0,64/0,56/0,25	55/190/119/56	19,80	3	0,0002
15 Monate n.T.		$n = 391$, 98 tot			
New CML-Score	0,76/0,58/0,31	133/211/47	10,91	2	0,0043
Neues Prognosesystem	0,88/0,62/0,54/0,25	72/162/103/54	23,65	3	<0,0001
18 Monate n.T.		$n = 349$, 92 tot			
New CML-Score	0,77/0,59/0,32	117/191/41	11,21	2	0,0037
Neues Prognosesystem	0,85/0,64/0,52/0,28	77/140/85/47	25,10	3	<0,0001
21 Monate n.T.		$n = 321$, 85 tot			
New CML-Score	0,78/0,60/0,33	105/178/38	11,67	2	0,0029
Neues Prognosesystem	0,84/0,67/0,52/0,26	76/124/77/44	24,28	3	<0,0001
24 Monate n.T.		$n = 300$, 80 tot			
New CML-Score	0,80/0,61/0,34	95/170/35	12,09	2	0,0024
Neues Prognosesystem	0,86/0,67/0,52/0,27	69/120/71/40	26,32	3	<0,0001

[a]Berechnet nach der Kaplan-Meier-Methode ab der jeweiligen Landmark. Die Nennung erfolgt für den New CML-Score in der Gruppenreihenfolge „Niedrigrisiko“, „mittleres Risiko“, „Hochrisiko“ und beim neuen Prognosesystem in der Gruppenreihenfolge „Niedrigstrisiko“, „niedrigeres Risiko“, „höheres Risiko“, „Höchstrisiko“.

[b]Reihenfolge der Nennung wie bei den 6-Jahresüberlebenswahrscheinlichkeiten.

[c]Freiheitsgrade.

[d]524 Patienten mit Daten, wovon 110 verstarben. Analoge Angaben zu den übrigen Zeitpunkten.

[e]Die Abkürzung „n.e.“ steht für „nicht existent“.

[f]Abkürzung „n.T.“ steht für „nach Therapiebeginn“.

[g]Die Abkürzung „n.b.“ steht für „nicht beobachtet“. Die Überlebenszeit des letzten zensierten Patienten war kürzer als sechs Jahre.

Tabelle 5.6 bietet eine Übersicht zu den Prognosezeitpunkten der ersten beiden Therapiejahre. Wie in Abschnitt 5.1.2, wurde zu den Verlaufszeitpunkten ab Monat 6 auf Basis der Niedrigst- und Höchstrisikogruppe eine stärkere Differenzierung der Überlebenswahrscheinlich-

keiten beobachtet. Ab Monat 12 besaß die Logrank-Statistik im Vergleich zum New CML-Score zwischen 7,46 (Monat 12) und 14,23 (Monat 24) erhöhte Werte. Im Gegensatz zur gemeinsamen Lernstichprobe war ein Vorteil des neuen Prognosesystems sogar im Vergleich „niedrigeres Risiko“ vs. Niedrigrisko und „höheres Risiko“ vs. mittlere Risikogruppe nicht mehr zu sehen: Die Überlebenswahrscheinlichkeiten der beiden mittleren Gruppen des neuen Prognosesystems lagen erheblich näher beieinander als die Niedrigrisikogruppe und die mittlere Risikogruppe des New CML-Scores.

5.3 Das neue Prognosesystem in Lern- und Validierungsstichprobe - Resümee

Das neue Prognosesystem erfüllte in der Lernstichprobe alle Kriterien, die gemäß Abschnitt 2.2 a) zu bewältigen waren: Bereits ab dem Zeitpunkt „3 Monate“ entsprachen die Werte der beobachteten Überlebenswahrscheinlichkeiten der logischen Reihenfolge der Risikogruppen und die Kaplan-Meier-Kurven überschnitten sich nicht. Untersucht bis Monat 24, waren die Logrank-Tests über alle vier Risikogruppen statistisch signifikant (alle $p < 0{,}0001$). Zu den maßgeblichen Zeitpunkten zwischen 12 und 24 Monaten repräsentierten alle Risikogruppen mindestens 10% aller nach dem jeweiligen Landmarkzeitpunkt unter Risiko stehenden Patienten und die paarweisen Vergleiche der Überlebenswahrscheinlichkeiten der Risikogruppen führten immer zu $p < 0{,}001$. Die nach Christensen et al. [24] auf Basis des Prognosesystems geschätzten Einjahresüberlebenswahrscheinlichkeiten lagen mit der maximalen Differenz von 0,0087 sehr nahe an den in den Risikogruppen beobachteten und entsprechendes galt damit auch für prognostizierte und beobachtete Todesfälle.
Der Vergleich mit dem New CML-Score in der gemeinsamen Teilmenge der Lernstichprobe (463 Patienten, Datenstand: Zeitpunkt der Entwicklung des New CML- Scores) nach Abschnitt 2.2 b) zeitigte für das neue Prognosesystem - bereits vor dem Einjahreszeitpunkt - die für Überlebensprognosen im Therapieverlauf berechtigte Etablierung einer zusätzlichen Risikogruppe bei gleichzeitig stärkerer Diskriminierung der Überlebenswahrscheinlichkeiten über alle vier Gruppen (vgl. Tabelle 5.1).
In der Validierungsstichprobe waren ab dem Zeitpunkt „6 Monate“ mit $n = 23$ ausreichend Patienten in der Niedrigstrisikogruppe, um für diesen und spätere Landmarkzeitpunkte die Einhaltung der logischen Reihenfolge der Risikogruppen bzgl. der Überlebenswahrscheinlichkeiten sowie über alle vier Risikogruppen statistisch signifikante Logrank-Tests (jeweils mit $p < 0{,}0001$) zu dokumentieren. Während zu den entscheidenden Zeitpunkten zwischen 12 und 24 Monaten die Risikogruppengrößen immer mindestens 10% aller Patienten betrugen, waren, im Gegensatz zu den anderen 20 paarweisen Vergleiche, für die paarweisen Vergleiche „niedrigeres Risiko“ vs. „höheres Risiko“ und „höheres Risiko“ vs. „Höchstrisiko“ in neun von zehn Fällen keine statistisch signifikant unterschiedlichen Überlebenswahrscheinlichkeiten zu verzeichnen. Sowohl die Gegenüberstellungen der Abbildungen 4.12 vs. 5.1 und 4.13 vs. 5.2 als auch die Betrachtung der Einjahresüberlebenswahrscheinlichkeiten und der Zahl der Todesfälle dokumentierten für die Validierungstichprobe v.a. im Falle der beiden ungünstigeren Risikogruppen höhere Überlebenswahrscheinlichkeiten als nach der Lernstichprobe und dem neuen Prognosesystem zu erwarten standen. Auch im Vergleich mit dem New CML-Score zeigte sich beim neuen Prognosesystem das nahe Beieinanderliegen der beiden mittleren Risikogruppen.

Neben den erwartbar positiven Ergebnissen in der Lernstichprobe aus 760 Patienten, bleibt als Resümee festzuhalten, dass auch eine Überlegenheit des neuen Prognosesystems über den

New CML-Score in der gemeinsamen, 463 Patienten umfassenden Teilmenge der Lernstichprobe (Abschnitt 5.1.2) beobachtet werden konnte. Auf Basis der vorläufigen Ergebnisse zur Validierungsstichprobe, kann man für das neue Prognosesystem von einer zuverlässigen, statistisch signifikanten Trennung der Überlebenswahrscheinlichkeiten in die drei Gruppen „Niedrigstrisiko", „niedrigeres oder höheres Risiko" und „Höchstrisiko" sprechen.[18] Der Nachweis einer zufriedenstellenden Diskriminierung auch der Überlebenswahrscheinlichkeiten der Gruppen „niedrigeres Risiko" und „höheres Risiko" sowie gleichzeitig der Gruppen „höheres Risiko" und „Höchstrisiko" oder die auf ausreichender statistischer Power fußende Feststellung eines diesbezüglichen Scheiterns konnte mit der vorliegenden Validierungsstichprobe nicht erbracht werden, aber mit einer Erhöhung von Fallzahl und Beobachtungszeiten sollte eine baldige Beurteilung möglich sein.
Zur Förderung der Glaubwürdigkeit und Verallgemeinerbarkeit sollten für die Validierungsstichprobe zusätzlich Daten zu IFN-α aus anderen Ländern und Studien berücksichtigt werden können, deren Erhalt aufgrund der Fokussierung auf Imatinib allerdings erschwert ist. Am Überzeugendsten wäre eine Überprüfung des neuen Prognosesystems anhand einer internationalen Validierungsstichprobe durch eine an seiner Entwicklung unbeteiligte ausländische Institution.

[18] Für die jeweils fünf Vergleiche zwischen 12 und 24 Monaten nach Therapiebeginn lagen die p-Werte im Falle „Niedrigstrisiko" vs. „niedrigeres oder höheres Risiko" unter 0,01, im Falle „niedrigeres oder höheres Risiko" vs. Höchstrisiko unter 0,025.

Kapitel 6

Die Bedeutung des neuen Prognosesystems in der Imatinib-Ära

Das Prognosesystem wird verschiedenen Aufgaben, wie der individuelleren Vorhersage des Krankheitsverlaufes mittels der Risikogruppe, die theoretische Ermöglichung der Entwicklung einer risikoadaptierten Therapie und Erklärungen von Abweichungen im Krankheitsverlauf gerecht. Ein für die Risikogruppen des Prognosesystems stratifizierter Vergleich des Überlebens einer historischen Patientenstichprobe von IFN-α-behandelten mit Imatinib-behandelten Patienten wird den gegenwärtigen Erfolg von Imatinib wohl nur bestätigen und somit keinen Einfluss auf das zukünftige Therapieverhalten haben. Unter der Prämisse, dass eine deutliche zytogenetische Remission unter Imatinib, risikostratifiziert nach dem New CML-Score für Patienten ohne Hochrisiko, zu ebenso guten Überlebenswahrscheinlichkeiten führt wie bei IFN-α, prognostizierten Hasford und Pfirrmann [45, 91] wegen eines um ca. 60% Prozentpunkte erhöhten Anteils an deutlichen Remissionen [79] für Imatinib langfristig wesentlich günstigere Überlebenswahrscheinlichkeiten. Einzig die Patienten der Niedrigstrisikogruppe stellen mit ihren Überlebenswahrscheinlichkeiten eine Herausforderung für Imatinb und neuere Therapien, aber gerade diese Patienten werden wohl auch unter Imatinib sehr gute Überlebenswahrscheinlichkeiten aufzuweisen haben.
Nach wie vor gibt es Patienten, die weiterhin erfolgreich mit IFN-α behandelt werden. Für das Studientreffen der deutschen CML-Studiengruppe im November 2005 wurden Daten der Studien CML III [109] und CML IV [111] analysiert. Von 324 lebenden Patienten der ersten Studie mit Rekrutierungsende im Dezember 2001 erhielten zuletzt 49 (15%) der Patienen IFN-α aber kein Imatinib ebenso wie 46 der 90 Patienten (51%), die seit der im Juli 2002 gestarteten Pilotphase der Studie CML IV [111] in den Arm mit IFN-α als Primärtherapie randomisiert wurden und für die mindestens ein Jahr Beobachtungszeit vorlag. Für diese Patienten sind die prognostizierten Überlebenswahrscheinlichkeiten der Niedrigstrisikogruppe hochinteressant.

In Anbetracht seiner Entwicklung für IFN-α-behandelte Patienten, welche Bedeutung könnte das neue Prognosesystem nach den Remissionserfolgen von Imatinib [79] heute noch für neudiagnostizierte Patienten haben? Zum aktuellen Erkenntnisstand (2006) ist, außer bei Höchstrisiko-Patienten, sicherlich auch bei „höherem Risiko“ der sofortige Beginn mit Imatinib oder einer neueren Therapie unstrittig. Für die Daten der zehn Studien der endgültigen Lernstichprobe lag das mediane Datum des IFN-α-Therapiebeginns im Jahr 1990. Bei einer konsequenteren Verabreichung von IFN stünde speziell für die Gruppe „niedrigeres Risiko“ eine Erhöhung des Anteils an deutlichen Remissionen und zugleich eine Verringerung von Fällen mit Progression zu erwarten, als dies in der Lernstichprobe der Fall war. Dazu würde man mit häufigeren zy-

togenetischen Untersuchungen vermutlich schon vor der Landmark „21 Monate“ den Zeitpunkt erreichten, zu welchem man über 80% aller Patienten mit zu erwartender deutlicher Remission erfasst hätte. Bei regelmäßiger ärztlicher Evaluation und einem jederzeit möglichen Wechsel zu einer Imatinib-Therapie wäre ein durch einen bestimmten Entscheidungszeitpunkt begrenzter Beginn mit IFN-α-basierter Therapie bei Patienten mit niedrigerem Risiko zumindest solange begründbar, bis auch für die mit Imatinib behandelten Patienten vergleichbar hohe Langzeit-Überlebenswahrscheinlichkeiten beobachtet wurden, wie für die Patienten der Niedrigstrisikogruppe. Mit dem Wissen um die hohen Remissionserfolge und die bessere Verträglichkeit von Imatinib [79] bevorzugt heutzutage allerdings der überwiegende Teil der Ärzte eine Imatinib-Therapie. Dies zeigte sich auch in der dreijährigen Pilotphase der deutschen Studie CML IV [17, 111], mit der Konsequenz, dass für die Hauptphase ab Juli 2005 der Arm mit IFN-α als Primärtherapie geschlossen wurde.

Stratifiziert nach den Risikogruppen bei IFN-α-Therapie, kann man das Überleben Imatinib und IFN-α-behandelter Patienten miteinander vergleichen, doch mit einer Rate von über 80% deutlichen Remissionen bereits im ersten Therapiejahr [79] ergibt sich für Imatinib eine völlig andere Verteilung der Patienten auf die Risikogruppen. Man wird für Imatinib und künftige Therapien eigene Prognosemodelle entwickeln müssen.

Kapitel 7

Zusammenfassung

Ziel vorliegender Arbeit war die Entwicklung eines Prognosesystems für Überlebenswahrscheinlichkeiten von Patienten, deren Primärtherapie auf Interferon-α (IFN-α) basiert. In Erweiterung des bereits existierenden, validierten New CML-Scores [42], welcher sich auf ausschließlich zum Diagnosezeitpunkt erhobene Baselinevariablen stützt, sollten dabei Therapieverlaufsdaten zur zytogenetischen Remission die Prognoseergebnisse weiter verfeinern.

Alle in der „IFN-α-Ära“ üblicherweise (d.h. zu 90%) erfassten und bereits von der Entwicklung des New CML-Scores als (potenziell) prognostisch relevant bekannten Baselineparameter [42] wurden bei der Modellentwicklung berücksichtigt: Alter, Geschlecht, Hämoglobin, Leukozytenzahl, Blasten, Basophile, Eosinophile (alle drei aus dem peripheren Blut), Thrombozytenzahl und Milzvergrößerung. Die zytogenetische Remission (ZR) wurde mit Hilfe der beiden dichotomen Ereignisvariablen „Erreichen einer ersten partiellen ZR (1-35% Ph-positive Metaphasen)“ und / oder „Erreichen einer ersten kompletten ZR (0% Ph-positive Metaphasen)“ modelliert, da beide Resultate zu signifikant günstigeren Überlebenswahrscheinlichkeiten geführt hatten [2, 34, 57, 60, 74, 107].

Um von den unterschiedlichen Evaluationshäufigkeiten möglichst unabhängig zu sein, wurde zu einem Therapieverlaufszeitpunkt immer das bisher günstigste beobachtete Remissionsstadium angenommen und ein Rezidiv in ein ungünstigeres Stadium nicht berücksichtigt. Zur Entwicklung des Prognosemodells wurde die Cox-Regression mit Zeitabhängigkeit für die beiden zytogenetischen Ereignisse gewählt, damit für sie alle zugelassenen Informationen ausgeschöpft werden können. Alternativ ein Cox-Modell ab einer bestimmten Landmark zu rechnen, hätte nur bis zur Landmark beobachtete Patienten eingeschlossen und Remissionen nach der Landmark nicht berücksichtigt. Mit der Wahl einer Landmark wäre das Modell für diesen Zeitpunkt festgelegt gewesen.

Es gelang in der Lernstichprobe gemäß der in Kapitel 2 angeführten Kriterien ein neues Prognosesystem zu identifizieren, welches insbesondere zu den medizinisch interessanten, möglichen Therapieentscheidungszeitpunkten 12, 15, 18, 21 und 24 Monate nach Erstverabreichung von IFN-α vier Risikogruppen statistisch signifikant diskriminierte.
Die vier Risikogruppen wurden mit Hilfe der „Minimal p-value“-Methode [6] auf Basis des über die Cox-Regression identifizierten endgültigen Prognosemodells definiert. Sie wiesen zu allen, in den Studienprotokollen vorgesehenen, maximal vierteljährlichen Evaluationszeitpunkten der ersten beiden Jahre nach IFN-α-Therapiebeginn statistisch signifikant unterschiedliche Überlebenswahrscheinlichkeiten auf. Die Verwendung der Effektschätzer des endgültigen Modells zu

allen Berechnungszeitpunkten - mit ihrer impliziten Ausnutzung sowohl der Überlebenszeitinformation als auch der Information zu allen zugelassenen Remissionen - erlaubte auch in einer unabhängigen Validierungsstichprobe schon ab Landmarkzeitpunkten innerhalb des ersten Jahres die günstigen Überlebenswahrscheinlichkeiten der oft noch wenigen Patienten der Niedrigstrisikogruppe von den übrigen signifikant zu unterscheiden.

Wegen zu potenziellen Berechnungszeitpunkten einheitlicher Risikogruppengrenzen sowie ausschließlich dichotomer Baselinevariablen im finalen Modell und dabei ausreichend vergleichbarer Effektschätzer konnte ein vereinfachtes Prognosesystem definiert werden, ohne die prognostische Diskriminierung der Risikogruppen wesentlich zu beeinträchtigen. Anstatt der entsprechenden Effektgröße, wurde pro Merkmalsausprägung in einer der höheren Risikogruppen der dichotomen Baselinevariablen jeweils einheitlich der Risikowert 1 vergeben. Die höheren Risikogruppen bestanden aus Alter > 41 Jahre, Milzvergrößerung > 7 cm, Eosinophile $> 2\%$, Basophile $> 2\%$, bei Frauen: Hämoglobin $< 11{,}4$ g/dl und bei Männern: Hämoglobin $< 13{,}6$ g/dl.
Patienten mit allen fünf Merkmalsausprägungen in den ungünstigeren Risikogruppen oder Patienten mit Hochrisiko nach dem New CML-Score verbleiben ungeachtet jedes späteren Remissionserfolges unveränderlich in der Höchstrisikogruppe des neuen Prognosesystems. Gründe hierfür waren das von Anfang an relativ hohe Sterberisiko und der bei Hochrisikopatienten nach dem New CML Score nicht erkennbare Überlebensvorteil durch Erreichen einer partiellen oder kompletten Remission [19, 46, 89]. Zu Therapiebeginn werden im neuen Prognosesystem Patienten mit 3 oder 4 Werten in einer der höheren Risikogruppen der Gruppe „höheres Risiko“ und mit ≤ 2 Werten der Gruppe „niedrigeres Risiko“ zugeordnet. Beobachtet man für einen Patienten der beiden letztgenannten Gruppen eine partielle Remission innerhalb von 21 Therapiemonaten, so verbessert sich der Patient um eine Risikogruppe, von der Gruppe „höheres Risiko“ in die Gruppe „niedrigeres Risiko“ bzw. von der Gruppe „niedrigeres Risiko“ in die im Therapieverlauf entstehende Gruppe „Niedrigstrisiko“. Außer im Falle der Höchstrisikogruppe, werden Patienten mit kompletter Remission direkt in die Niedrigstrisikogruppe versetzt.

Den gemäß der vorliegenden gültigen Zytogenetikdaten zu 803 Patienten „optimalen Entscheidungszeitpunkt“ für oder gegen IFN-α erreichte man nach 21 Monaten, als 83% aller verzeichneten deutlichen Remissionen ($\leq 35\%$ Ph-positive Metaphasen) bereits berichtet waren. Von 803 Patienten war für 760 Patienten die Risikogruppe des neuen Prognosesystems berechenbar. Gemäß der Risikogruppenzugehörigkeit zur Landmark „21 Monate“ konnten für 646 weiterhin unter Beobachtung stehende Patienten folgende Ergebnisse berichtet werden: 119 Patienten der Niedrigstrisikogruppe (18% von 646, 16 Verstorbene) besaßen eine 9-Jahresüberlebenswahrscheinlichkeit von 0,7494. Zur Gruppe „niedrigeres Risiko“ gehörten 256 Patienten (40%, 99 Verstorbene) mit einer medianen Überlebenszeit von 81 Monaten und einer 9-Jahresüberlebenswahrscheinlichkeit von 0,3530. „Höheres Risiko“ verzeichneten 187 Patienten (29%, 111 Verstorbene). Ihre mediane Überlebenszeit betrug 63 Monaten und die 9-Jahresüberlebenswahrscheinlichkeit 0,1071. Von 84 Höchstrisikopatienten (13%) verstarben 65. Bei einer medianen Überlebenszeit von 43 Monaten, wurden keine Überlebensdauer von 9 Jahren beobachtet. Die 6-Jahresüberlebenswahrscheinlichkeit war 0,2533.

Einschränkungen

Die Dichotomisierung der metrischen Baselinevariablen bedeutet Informationsverlust und speziell für Patienten mit Werten im Grenzbereich ist das Setzen „harter Grenzen“ eingedenk (kleiner?) biologischer Unterschiede und von Messungenauigkeiten nicht optimal. Nur wenn festge-

stellt wurde, dass die Überlebenswahrscheinlichkeiten eher ab gewissen Grenzen sprunghaft als über den gesamten Wertebereich kontinuierlich abfallen, wurden kategoriale Skalierungen der Variablen als Alternative für das multiple Modell zugelassen.

Die durch Tasten ermittelte Milzvergrößerung und die Schätzung der zytogenetischen Remission auf Basis von selten mehr als 25 Metaphasen schränken die Reliabiltät der Messergebnisse zu den Prognosefaktoren ein. Hinzu kommen im Falle der Zytogenetik die Zeitverzögerung zwischen dem Eintreten einer deutlichen ZR und dem Feststellen derselben. Doch trotz dieser Ungenauigkeiten boten die Daten zu beiden Faktoren prognostisch hochrelevante Informationen und waren daher für das Prognosesystem unverzichtbar.

Erschwert durch eine geringe Power, konnte mit einer Validierungsstichprobe aus 524 Patienten keine ausreichende Diskriminierung der Überlebenswahrscheinlichkeiten zwischen allen vier Risikogruppen gezeigt werden, wohl aber über alle Zeitpunkte eine auch paarweise statistisch signifikante Unterscheidung der drei Gruppen „Niedrigstrisiko“, „niedrigeres oder höheres Risiko“ und „Höchstrisiko“. Längere Beobachtungszeiten lassen für die Zukunft eine fundiertere Beurteilung erhoffen. Daten aus Studien anderer Institutionen wären zur besseren Einschätzung einer allgemeinen Anwendbarkeit des neuen Prognosesystems wünschenswert.

Leistungen des neuen Prognosesystems

Angesichts der hohen Überlebenswahrscheinlichkeiten der Niedrigstrisikogruppe wurde das Ziel, mit einem neuen Prognosesystem im Therapieverlauf eine Patientengruppe zu finden, die von einem Behandlungsbeginn mit IFN-α profitieren könnte - ohne dabei die Patienten mit initialem Höchstrisiko außer Acht zu lassen - erreicht. Die deutliche, statistisch signifikante Trennung dieser unterschiedlichsten Risikogruppen wurde durch eine von der Modellentwicklung unabhängige Validierungsstichprobe bestätigt.

Das neue Prognosesystem gibt ein methodisches Beispiel für die Entwicklung und Validierung eines Prognosesystems unter Berücksichtigung von Informationen aus dem Therapieverlauf. Dabei war es insbesondere möglich, die Gewinnung eines von einer festen Landmark unabhängigen Prognosesystem aufzuzeigen, dessen Risikogruppen über die ersten beiden Therapiejahre, während derer die zytogenetische Remission im Brennpunkt steht, zu jedem frei wählbaren Entscheidungszeitpunkt auf dieselbe Weise leicht berechnet werden können.[1] Die maximal zwei Risikogruppenwechsel in ausschließlich günstigere Stadien unterstützen die einfache Anwendbarkeit des Prognosesystems und die Interpretierbarkeit der Überlebenswahrscheinlichkeiten seiner Risikogruppen. Für die Patienten ist das Ausschließen einer Risikogruppenverschlechterung psychologisch von positiver Bedeutung.

Nach (früher) Stabilisierung der Überlebenskurve zur Niedrigstrisikogruppe können mit Hilfe von Simon-Makuch-Kurven die Überlebenswahrscheinlichkeiten aller vier Risikogruppen - ohne einschränkende Landmark - jederzeit nach dem aktuellsten Informationsstand berechnet werden.

Für immer noch viele mit IFN-α als Primärtherapie behandelte Patienten bleiben die prognostizierten Überlebenswahrscheinlichkeiten der Niedrigstrisikogruppe von Bedeutung.

[1] Die zum Diagnosezeitpunkt erforderlich Identifikation von Hochrisikopatienten nach dem New CML-Score kann via Internet (www.pharmacoepi.de/cgi-bin/pharmacoepi/cmlscore.cgi) sehr schnell, per Taschenrechner in wenigen Minuten und zur Not auch mit Papier und Bleistift vorgenommen werden.

Danksagung

Bei allen Personen und Institutionen, die die Realisierung der vorliegenden Arbeit förderten, möchte ich mich herzlich bedanken.

Insbesondere danke ich Herrn Professor Dr. Joerg Hasford für die Ermutigung zu dieser Arbeit, seine wertvollen Anregungen und all' seine sonstige Unterstützung dabei. Ihm und den anderen beteiligten Wissenschaftlern des „Collaborative Prognostic Factors Project“ gebührt mein außerordentlicher Dank für die Zurverfügungstellung der Patientendaten und damit die prinzipielle Ermöglichung des Vorhabens.

Der Geschäftsführerin der GIS e.V., Frau Brigitte E. Weber, danke ich für ihre freundschaftliche Zuneigung, welche stets das Fundament für das großartige und motivierende Arbeitsklima in der GIS bildeten. Ebenso sei allen anderen Kollegen und Mitarbeitern der GIS e.V. gedankt, die mich über die Jahre begleiteten.

Ich möchte mich bei allen meinen Freunden für ihr Interesse und ihre Anteilnahme an meiner Arbeit bedanken und dabei stellvertretend die beiden erwähnen, die mir in München die meiste Zeit über zur Seite standen: Christiane M. Knopp und, na klar, der Meyer Olli.

Meiner Tante Waltraud E. Schraffenberger danke ich dafür, dass sie immer für mich da war und weiterhin ist.

Anhang A

SAS Programme

Die beiden Programme zur Berechnung der Barlow-Prentice-Residuen sowie von Simon-Makuch-Kurven und des Mantel-Byar-Tests für das neue Prognosesystem wurden der Arbeit angehängt. Sie werden auf Anfrage (E-Mail: pfi@ibe.med.uni-muenchen.de) gerne zur Verfügung gestellt.

A.1 Programm zur Berechnung der Barlow-Prentice-Residuen

```
/*****************************************************************/
/* NAME OF THE PROGRAMME: T:\GISeV\epfp\response\residue2.mac     */
/*                                                               */
/* PURPOSE: This macro calculates residuals to assess a Cox      */
/*          relative risk regression model in accordance with    */
/*          Barlow and Prentice, Biometrika (1988), 75:65-74     */
/*          formula (11)                                         */
/*                                                               */
/* OUTPUT:  - A residual plot for each variable, displaying      */
/*            the residual for each patient in the model         */
/*            For each variable "varname", the plots are saved   */
/*            under T:\GISeV\epfp\response\varname.eis.ps (*.cgm) */
/*          - A working data set "info" including patient id,    */
/*            survival times, survival status, all explantory    */
/*            variables, event times of the time-dependent       */
/*            variables, and the residuals to each patient for   */
/*            each variable                                      */
/*                                                               */
/* DATE:    02/11/2000           Author: Markus Pfirrmann        */
/*                                                               */
/*****************************************************************/

/*****************************************************************/
/* Variables:    indata  : data set with all important variables  */
/*               id      : patient's identification number        */
/*               sutime  : survival time                          */
/*               sustatus: survival status                        */
/*               censval : values of survival status indicating   */
/*                         censoring of the surival time          */
/*               basevar : baseline variables with fixed,         */
/*                         time-independent baseline values       */
/*                         (time-independent explanatory          */
/*                         variables)                             */
/*               timevar : (dummy) variables with a possible      */
/*                         change in their values in dependence   */
```

```
/*                          on the observation of certain          */
/*                          events over time                       */
/*                          (time-dependent explanatory variables) */
/*               times   : times when events were recorded        */
/*                          leading to changes in the values of    */
/*                          the time-dependent explanatory         */
/*                          variables                              */
/*                          (one time variable for each "timevar") */
/*               events  : the values indicating the              */
/*                          observation of an event for the        */
/*                          time-dependent variables               */
/*                          (one value for each "timevar")         */
/*               chanvar : indicating the place of the variable   */
/*                          in the list of "timevar" whose change  */
/*                          also leads to a change of values in    */
/*                          the other (dummy) variable(s)          */
/*               labels  : the labels firstly, of the "basevar"   */
/*                          variables, secondly, of the "timevar"  */
/*                          variables in the same order as the     */
/*                          variables were listed before           */
/*                          (labels should be separated by "/")    */
/*               tit4    : possibility to provide a title in      */
/*                          line 4 of the output                   */
/*****************************************************************/
**** Please note: this programme was written for the special case *
**** where status of (cytogenetic) remission was expressed by two *
**** time-dependent dummy variables. As soon as either - partial  *
**** or complete remission - was observed, the according variable *
**** value changed from 0 to 1. A change back to 0 was only       *
**** accepted for the dummy variable indicating partial remission *
**** and only in the case when complete remission was observed.   *
**** Thus, this programme needs a few changes, if time-dependent  *
**** events are definded in a different way                       *;

%macro resi    (indata   = ,
                id       = number ,
                sutime   = sutime ,
                sustatus = sustatus ,
                censval  = 0 1 ,
                basevar  = ,
                timevar  = ,
                times    = ,
                events   = ,
                chanvar  = ,
                labels   = ,
                tit4     =
               );

options mprint linesize=111 pagesize=69 nodate pageno=1 mtrace;

title4 "&tit4";

****** Reading all important information into data set "base" *********;
data base;
  set &indata (keep=&id &sutime &sustatus &basevar &timevar &times);
run;


proc iml;
```

```
****** Module to create matrix with survival data and covariables *****;
start matrix;             **** remember: no arguments => local=global *;

 use base;
 read all var {&id &sutime &sustatus &basevar &timevar} into covar;

 nr=nrow(covar);       **** x serves the identification of missing ****;
 x=j(nr,1,0);          **** values in any of the variables         ****;
 nc=ncol(covar);
 do i=1 to nc by 1;
   x=x+covar[,i];
 end;

 use base;
 read all var {&times} into times;

 numtvar=ncol(times);      **** number of time-dependent variables *****;
 numvar=ncol(covar)-3;     **** number of explanatory variables    *****;
 numvar=numvar||numtvar;   **** storing both information           *****;
 covar=covar||times||x;

 create comatrix var {&id &sutime &sustatus &basevar &timevar &times help};
   append from covar;
 close comatrix;

 ***** observations with missing values are removed ****;
 edit comatrix;
   delete all where(help=.);
   purge;
 close comatrix;
 sort comatrix by &sutime;

 create numvars var {numvar numtvar};
   append from numvar;
 close numvars;

finish matrix;
run matrix;
quit;


************ internal names for time-dependent variables *************;
data numvars;
 set numvars;
 call symput('numtvar',numtvar); ** assigns no. of time-dependent   **;
                                 ** variables to macro variable     **;
 call symput('numvar',numvar);   ** assigns no. of all              **;
                                 ** variables to macro variable     **;
run;

** Macro NAMES creates macro internal names for time-dependent      **;
** variables                                                        **;
%macro names;
 %do i=1 %to &numtvar;
   %global time&i eve&i;
   %let time&i=%scan(&times,&i,%str( ));     ** response times      **;
   %let eve&i=%scan(&events,&i,%str( ));     ** response values     **;
 %end;
```

```
%mend names;
%names;

** Macro AUXRES creates auxiliary response variables needed for     **;
** procedure PHREG                                                  **;
%macro auxres;
  %do i=1 %to &numtvar;
    resp&i
  %end;
%mend auxres;

** Macro EIS creates names for the residuals of all varibales       **;
%macro eis;
  %do i=1 %to &numvar;
    ei&i
  %end;
%mend eis;

********************** estimation of coefficients  ******************;
*************** for the special situation described above ************;
proc phreg data=comatrix outest=koeff;  ** application of Cox model **;
  model &sutime*&sustatus(&censval)= &basevar %auxres;
  %do i=1 %to &numtvar;
    %if &i ne &chanvar %then %do;
      resp&i=%eval(&&eve&i-1);
      if &sutime >= &&time&i and &&time&i ^= . then resp&i=&&eve&i;
    %end;
    %if &i eq &chanvar %then %do;
      resp&i=%eval(&&eve&i-1); %let y=%eval(&i-1);
      if &sutime >= &&time&i and &&time&i ^= . then do;
         resp&i=&&eve&i;
         resp&y=%eval(&&eve&y-1);
      end;
    %end;
  %end;
run;

* tkoeff: contains variable "&uezeit" with the estimated coefficients *;
proc transpose data=koeff out=tkoeff;
  var &basevar %auxres;
run;

** all censored values are set to the common value 0, events to 1 **;
data newvalue;
  set comatrix;
  if &sustatus in (&censval) then &sustatus= -1;
  else &sustatus = 1;
  if &sustatus eq -1 then &sustatus=0;
  drop help;
run;


proc iml;

******************** Modul to evaluate p(i,t) ************************;
************* p(i,t): see formula (11) of the paper ******************;
start pit;

  ********** only patients still at risk at time t_risk **************;
```

```
  status=j(nr,1,1);
  if t_risk>1 then status[loc(covar[,2] < covar[t_risk,2])]=0;
  yit=sum(status);              *** no. patients at risk at t_risk ***;
  z=j(yit,&numtvar,99);
  do j=tvarbeg to tvarend;
    ij=j-tvarbeg+1;        *** to address right column of covar    ***;
    itimes=j+&numtvar;     *** to address corresponding t to event ***;
    helptime=covar[(nr-yit+1):nr,itimes];       *** times to event ***;
    z[,ij]=j(yit,1,(events[ij]-1));      *** vector with no events ***;
    ichange=(&chanvar-1);

    ****** only patients with better/worse status (e.g. partial    ***;
    ****** response) but not yet best / worst status have event    ***;
    if ij=ichange then do;

      z[loc(covar[t_risk,2]>=helptime & helptime>0),ij]=events[ij];

      helpchan=covar[(nr-yit+1):nr,(itimes+1)]; *** times to event ***;
                                                *** best status    ***;

      ** best status not yet/never recorded: auxiliary value sets    *;
      ** event into far future, time not yet observed for any pat.   *;
      helpchan[loc(helpchan=.),]=9999;

      * best status already recorded, less response set back to none:*;
      z[loc(covar[t_risk,2]>=helpchan),ij]=(events[ij]-1);
    end;

    *** z is set to event where survival time >= event time **********;
    else if ij^=ichange then
      z[loc(covar[t_risk,2]>=helptime & helptime>0),ij]=events[ij];
  end;

  *** now z contains the covariable value at risk time t_risk ********;
  *** if-condition to include the values of baseline variables *******;
  if &numvar <> &numtvar then z=covar[(nr-yit+1):nr,4:(tvarbeg-1)]||z;
  ********* pits contains all pi_t at risk time t=t_risk *************;
  pits=exp(z*esti)/sum(exp(z*esti));

finish pit;
************************** end of pit ********************************;


******************* Modul to evaluate E(beta,t) ***********************;
************* E(beta,t): see formula (11) of the paper ***************;
start ebetat;

  run pit;
  ****** with pits, ebeta_ti is instantly calculated at t=t_risk *****;
  ebeta_ti=T(z)*pits;

finish ebetat;
********************** end of modul ebetat ****************************;


******* Main part: The Barlow-Prentice residuals for individual i *****;
******** (Various variable names in accordance with the paper) ********;
use newvalue var {&id &sutime &sustatus &basevar &timevar &times};
read all into covar;
```

```
use tkoeff var {sutime};
read all into esti;

events={&events};          *** value of events for time-dep. covar. ***;
nc=ncol(covar);

tvarbeg=(nc-2*&numtvar+1); *** place of first tdp var. in comatrix  ***;
tvarend=(nc-&numtvar);     *** place of last tdp var. in comatrix   ***;



nr=nrow(covar);            *** no. of obs with complete data        ***;

ebet_tls=j(1,&numvar,0);   *** for ebeta values at risk time tl     ***;
e_is=j(1,&numvar,0);       *** for residuals of individual i        ***;
do i=1 to nr;
  xi_tls=j(1,&numvar,0);   *** covariable values at time tl         ***;
  pi_tls=j(1,1,0);         *** pi values at time tl                 ***;
  l=1;
  *** i could only be at risk before own last recorded survival time **;
  do while (covar[l,2] <= covar[i,2]);
    if covar[l,3]=1 then do;    *** death observed for individual l ***;
      t_risk=l;

      if l=i then do;
        run ebetat;
        ebeta_ti=T(ebeta_ti);    *** ebeta at time ti for patient i ***;
        e_i1=covar[i,4:tvarend]-ebeta_ti;     *** xi(ti)-E(beta,ti) ***;
      end;
      else if l^=i then run pit; *** different pi at tl-times <= ti ***;

      xi_tl=z[(nrow(z)-(nr-i)),]; *** covar. values at different tl ***;
      pi_tl=pits[(nrow(pits)-(nr-i)),];
    end;
    else if covar[l,3]=0 then do;        *** patient l was censored ***;

      if l=i then do;
        ebeta_ti=ebet_tls[1,]; *** ebeta is set to 0 as delta_i is 0 **;
        e_i1=ebet_tls[1,];     *** e_i1 is set to 0 as delta_i is 0 ***;
      end;

      xi_tl=xi_tls[1,];        *** set to 0 as delta_l is 0         ***;
      pi_tl=pi_tls[1,];        *** set to 0 as delta_l is 0         ***;
    end;

    xi_tls=xi_tls//xi_tl;
    pi_tls=pi_tls//pi_tl;
    if l=nr then goto lastobs;
    else l=l+1;
  end;

  lastobs:
  sametime=nrow(xi_tls)-nrow(ebet_tls); *** if > 1 pat. with same tl ***;
  do repeat=1 to sametime;
    ebet_tls=ebet_tls//ebeta_ti;    *** matrix of ebeta at times tl ***;
  end;

  e_i2=T(xi_tls-ebet_tls)*pi_tls;   *** sum: functional cov. - risk ***
```

```
                                    *** set mean over times tl      ***;
  e_i=e_i1-T(e_i2);                 *** Barlow-Prentice residual i  ***;
  e_is=e_is//e_i;

  sametime=sametime-1;              *** remove additional rows,     ***;
  nebe_tls=nrow(ebet_tls)-sametime; *** only appended if > 1 pat.   ***;
                                    *** with same tl                ***;
  ebet_tls=ebet_tls[1:nebe_tls,];   *** matrix of ebeta at times tl ***;
end;

nroweis=nrow(e_is);
e_is=e_is[2:nroweis,];              *** matrix of all residuals     ***;
r=rank(covar[,2]);                  *** ranks of survival times     ***;
e_is=covar[,1]||r||e_is;            *** covar[,1] contains id       ***;

names={&id ranks %eis};
create eisrank var names;           *** final data set eisrank      ***;
  append from e_is;
close eisrank;

quit;


***************** graphical representation of results ***************;
***
*** postscript
***;
*goptions reset=all
         rotate=landscape
         lfactor=2.5
         gsfname=grafout
         gsfmode=replace
         hsize=9 vsize=7
         hpos=73 vpos=35
         ftext=swiss
         device=pslmono;

***
*** word
***;
*goptions reset=all gsfname=grafout nodash gsfmode=replace
         gaccess=sasgastd gwait=5 targetdevice=pslmono
         hsize=9 vsize=7 hpos=73 vpos=35 dashscale=0.4
         ftext=swiss device=cgmofml lfactor=2.5 rotate=landscape;

***
*** Tex
***;
goptions reset=all gsfname=grafout gsfmode=replace horigin=3.0cm vorigin=15.5cm
         ftext=swiss hsize=15.0cm vsize=10.5cm hpos=80 vpos=35
         device=pslmono lfactor=2.5;


%do i=1 %to &numvar;
  %let var=%scan(&labels,&i,%str(/));

  data plot;
    set eisrank (keep=&id ei&i ranks);
```

```
    *** creation of possible output files ***;
    *filename grafout "T:\GISeV\epfp\response\&var..eis.cgm";
    filename grafout "T:\GISeV\epfp\response\&var..eis.ps";

    *** possible titles ***;
    *title1 j=c h=1.5 'Collaborative Prognostic Factors Project';
    *title2 'Residuals for relative risk regression';
    *title3 'according to Barlow and Prentice';
    *title4 "for variable &var";
    title;
    *footnote j=l h=0.5 "Munich UPDATE: &sysdate (T:\GISeV\epfp\response\residue2.mac)";
    footnote;

    symbol1 font=marker v=T width=3;

  proc gplot data=plot;
    axis1 label=(j=c h=1.2 "Rank of survival time") value=(h=1.2);
    axis2 label=(j=c h=1.2 a=90 rotate=360 "Residual") value=(h=1.2);
    plot ei&i*ranks /
        vaxis=axis2
        haxis=axis1
        overlay;
  run;
  quit;

%end;

proc sort data=eisrank; by &id; run;
proc sort data=base; by &id; run;

data info;
  merge base eisrank; by &id;
run;
proc sort data=info; by ranks; run;

proc print data=info n;
  var &id ranks &sutime &sustatus &basevar &timevar &times %eis;
run;

%mend resi;
*** end of macro;

*** The following example should be part of a programme CALLING the MACROS CCRSCORE;
** %resi (indata   = test1 ,
          id       = number ,
          sutime   = sutime ,
          sustatus = sustatus ,
          censval  = 0 1 ,
          basevar  = spleen ,
          timevar  = fpr fcr ,
          times    = fprtime fcrtime ,
          events   = 1 1 ,
          chanvar  = 2 ,
          labels   = spleen / pcr / ccr ,
          tit4     = "Model with spleen, partial and complete cytogenetic remission"
         );

** Example:
** sutime  : survival time;
```

```
** sustatus: survival status;
** censval : values of "survival status" which are to be censored;
** basevar : spleen=spleen enlargement in cm;
** timevar : fpr=first partial remission (at start of therapy
             "no remission" coded "0" and coded "1" after observation of fpr)
             fcr=first complete remission (at start of therapy
             "no remission" coded "0" and coded "1" after observation of fcr);
** times   : fprtime and fcrtime contain the times in days after start of therapy
             when partial or complete remissiom were observed for the first time;
** events  : as stated above, actual first observations (=events)
             of partial or complete remission were coded with "1";
** chanvar : the "2" stands for the place of the time-dependent variable in the
             "timevar list" (fcr) for which the observation of an event leads to
             the change of the other variable back to "0" (fpr);
** labels  : separated by "/";
```

A.2 Programm zur Berechnung von Simon-Makuch-Kurven und Mantel-Byar-Test für das neue Prognosesystem

```
/*****************************************************************/
/* NAME OF THE PROGRAMME: T:\GISeV\epfp\response\ccrscore.mac     */
/*                                                               */
/* PURPOSE: Simon-Makuch curves for the New prognostic system    */
/*          considering major (i.e. partial and complete)        */
/*          cytogenetic remission                                */
/*           - four different risk groups,                       */
/*             called "MCR risk groups" -                        */
/*                                                               */
/*          Graphical representation according to Simon R. and   */
/*          Makuch W., Statistics in Medicine (1984), 3:35-44    */
/*                                                               */
/*          Test according to Mantel N. and Byar D.P.,           */
/*          JASA, (1974), 69: 81-86                              */
/*                                                               */
/* OUTPUT:  - Cross tabulation of surv. status at last follow-up */
/*            by MCR score risk group at baseline (time=0)       */
/*          - Chi-squared test statistic and continueous         */
/*            continueous-corrected chi-squared test statistic   */
/*            of the Mantel-Byar test together with the          */
/*            corresponding p-values                             */
/*          - The starting point x of the Simon-Makuch curves    */
/*            (allowing for a certain time to response)          */
/*          - Cross tabulation of surv. status at last follow-up */
/*            by MCR risk group at last follow-up                */
/*          - A plot of survival probabilities after starting    */
/*            point x versus time after start of therapy         */
/*            The plot could be saved under                      */
/*            T:\GISeV\epfp\response\ccrscore.ps (*.cgm)         */
/*                                                               */
/* DATE:    07/09/2004           Author: Markus Pfirrmann        */
/*                                                               */
/*****************************************************************/

/*****************************************************************/
/* Variables:    indata  : data set with all important variables */
/*               cond    : case selection criteria for input data */
/*               sutime  : survival time                         */
/*               sustatus: survival status                       */
/*               censor  : values of survival status indicating  */
/*                         censoring of the surival time         */
/*               hasscore: New CML (Hasford) score               */
/*               timezero: MCR score at baseline                 */
/*               time0cat: MCR risk group at baseline            */
/*               times   : The two variables containing          */
/*                         time to PCR and time to CCR,          */
/*                         (please place time to PCR at first)   */
/*               score   : -1 -2 (MCR score values for PCR CCR)  */
/*               chanvar : indicating the place of the variable  */
/*                         in the list of "times" whose change   */
/*                         also leads to a change of values in   */
/*                         the other (dummy) variable(s)         */
/*               title1  : possibility to provide a title in     */
/*                         line 1 of the output                  */
/*               title2  : possibility to provide a title in     */
```

```
/*                          line 2 of the output                    */
/*              title3  : possibility to provide a title in         */
/*                          line 3 of the output                    */
/*              x       : starting time for Simon-Makuch curves     */
/*              time    : choose survival time unit for             */
/*                          graphical representation                */
/*              tlabel  : choose name for time axis x               */
/*              plabel  : choose name for surv. prob. axis y        */
/*              torder  : range for axis x in accordance with       */
/*                          variable time                           */
/*              tsmall  : number of ticks drawn between major       */
/*                          tick of time axis x                     */
/*              legend1 : legend for group with lowest risk         */
/*              legend2 : legend for group with 2nd lowest risk     */
/*              legend3 : legend for group with 2nd highest risk    */
/*              legend4 : legend for group with highest risk        */
/*                          variables                               */
/*              pt1     : choose time for 1st Greenwood             */
/*                          95% confidence interval                 */
/*              pt2     : choose time for 2nd Greenwood             */
/*                          95% confidence interval                 */
/*              pt3     : choose time for 3rd Greenwood             */
/*                          95% confidence interval                 */
/*              print   : if print=1 then MCR risk groups at        */
/*                          baseline will also be printed           */
/*              footnote: choose footnote                           */
/*              outprn  : choose name for graphical output file     */
/*****************************************************************/

%macro ccrscore(indata   =  ,
                cond     =  ,
                sutime   = survival ,
                sustatus = status ,
                censor   = 0 1 ,
                hasscore = ,
                timezero = zeroscor ,
                time0cat = zeroscor ,
                times    = ,
                score    = -1 -2 ,
                chanvar  = 2 ,
                title1   = ,
                title2   = ,
                title3   = ,
                x        = 91 ,
                time     = 365.25 ,
                tlabel   = Time ,
                plabel   = Survival probability after three months of therapy ,
                torder   = ,
                tsmall   = ,
                legend1  = "Lowest risk" ,
                legend2  = "Lower risk" ,
                legend3  = "Higher risk" ,
                legend4  = "Highest risk" ,
                pt1      = 3 ,
                pt2      = 6 ,
                pt3      = 9 ,
                print    = 1 ,
                footnote = ,
                outprn   = );
```

```
options mprint linesize=111 pagesize=69 nodate pageno=1 mtrace;

**** This file is to define the symbol for censored patients within the GIS e.V.;
**** Please substitute the handling of your own institution with regard to this matter;
libname gfont0 'T:\GISeV\cml\bztfont';

****** Reading all important information into data set "base" *********;
data base;
  set &indata (keep=&sutime &sustatus &hasscore &timezero &time0cat &times);
  &cond;
  **** re-defining survival status in accordance with the programme ***;
  if &sustatus in (&censor) then &sustatus= -1;
  else &sustatus=1;
  if &sustatus eq -1 then &sustatus=0;
run;

proc sort data=base;
  by &sutime;
run;

**** Formats for the four risk groups and survival status;
proc format;
  value score 0="Lowest"
              1="Lower"
              2="Higher"
              3="Highest";

  value uesta 0="Censored"
              1="Event";

title1 &title1;
title2 &title2;
title3 &title3;
footnote &footnote;

*** table of survival status at last follow-up by MCR risk group at baseline (time=0);
%if &print = 1 %then %do;
proc freq data=base;
  tables &sustatus*&time0cat;
  format &time0cat score. &sustatus uesta.;
  label  &sustatus ='Survival Status'
         &time0cat ='MCR score: Risk group at baseline';
run;
%end;

proc iml;

** Module surv to calculate Simon-Makuch curves and Mantel-Byar test;
start surv;                          *** no arguments => local=global;

 use base;
 read all var {&times} into times;  *** matrix times: first column for #(days after start of therapy
                                    *** when PCR was observed). Second column for #(days after start
                                    *** of therapy when CCR was observed).;
 read all var {&sutime &sustatus &hasscore &timezero &time0cat} into covar; **** matrix covar: first
                                    *** column: survival time, second column: survival status, third
                                    *** column: New CML (Hasford) risk group, forth column: MCR score
                                    *** at baseline (time=0), fifth column: MCR risk group at baseline
```

```
                                *** (can be identical to &timezero, depends on definition);
close base;

numtvar=ncol(times); *** #(columns of times);
help=j(1,numtvar,0);
times=help//times;   *** adding a first row with zeros to times;

nc=ncol(covar);      *** #(columns of covar);
help=j(1,nc,0);
covar=help//covar;   *** adding a first row with zeros to covar;

nr=nrow(covar);      *** #(rows of covar) (= #(patients)+1);
score={&score};      *** score of the time-dependent (dummy) covariables;

lastobs=j(1,8,0);    *** providing the start for the matrix finally containing survival status
                     *** and risk group of each patients at last follow-up;

hv1=j(nr,8,0);       *** providing the matrix hv1 containing the survival status of each patients
                     *** AT LAST FOLLOW-UP (=covar[,2]) BUT the MCR risk group AT BASELINE
                     *** (time=0, =covar[,5]) of course: no patients in lowest risk group with MCR
                     *** (hv1[,1] and hv1[,2]), yet;
hv1[,3] = (covar[,2]=1 & covar[,5]=1); *** dead lower risk group **;
hv1[,4] = (covar[,2]=0 & covar[,5]=1); *** cens lower risk group **;
hv1[,5] = (covar[,2]=1 & covar[,5]=2); *** dead higher risk group **;
hv1[,6] = (covar[,2]=0 & covar[,5]=2); *** cens higher risk group **;
hv1[,7] = (covar[,2]=1 & covar[,5]=3); *** dead highest risk group **;
hv1[,8] = (covar[,2]=0 & covar[,5]=3); *** cens highest risk group **;

nrisk=j(1,4,0);      *** #(patients at risk) at time=0
                     *** of course: no patients at risk in lowest risk group with MCR,
                     *** yet => (nrisk[1,1]=0);
nrisk[1,2]=sum(hv1[,3])+sum(hv1[,4]);
nrisk[1,3]=sum(hv1[,5])+sum(hv1[,6]);
nrisk[1,4]=sum(hv1[,7])+sum(hv1[,8]);

k=1;                 *** counter for #(patients with same survival time);
change=(times[,(&chanvar-1)]>639);        *** Partial cyto. remission after 21 mo. is neglected;
                                          *** - the corresponding times are set to missing;
if sum(change>0) then times[loc(change>0),(&chanvar-1)]=.;

do i=2 to nr;              *** loop from 2 to nr: counter from first to last patient,
                               *** patients are ordered by ascending survival time;
  riskhelp=nrisk[(i-1),1:4];   *** #(patients at risk in the four risk groups at the time when patient
                               *** (i-1) was censored or died);

  if covar[i,1]=covar[(i-1),1] then k=k+1;  *** if pat. i has same survival time as pat. (i-1): k=k+1;

  else if covar[i,1]>covar[(i-1),1] then do;
                           *** loop if pat. i does not have same survival time as pat. (i-1)
                                        *** the #(patients at risk in the four risk groups) needs
                                        *** a new calculation;

    do l=1 to k;                        *** all pat. with the previous (lower) survival time
                                        *** are removed from "at risk" matrix hv1;
      hv1[(i-l),1:8]=j(1,8,0);
    end;
    k=1;                       *** now, only patient i with "same survival time" is left: k=1;

    do j=1 to numtvar;         *** loop for counter from first to last time-dependent covariable;
```

```
*** the changes up to the previous survival time of patient (i-1) were already taken into account
*** the event times are thus set to zero;
change=(times[,j]<=covar[(i-1),1] & times[,j]>0);

if sum(change>0) then times[loc(change>0),j]=0;

*** changes in the time-dependent covariables between the previous survival time (i-1) and the
*** currently considered one of patient i need to be looked at;
change=(times[,j]<=covar[i,1] & times[,j]>0);
if sum(change>0) then do;  *** loop for changes in time-dependent variable j;
  timechan=covar[,4]+change*score[1,j];     *** calculate new MCR score for covariable j;
  covar[,4]=timechan;                       *** changed MCR score in forth column of matrix covar;

  timechan=covar[,4]#change;                *** only patients with changes in covariable j between
                                            *** previous survival time (i-1) and current survival
                                            *** time i need to be considered for an MCR risk group
                                            *** change;
  timecat=covar[,5];                        *** current MCR risk group;
  *** to avoid errors due to value zero (usually) standing for "no change";
  if min(change)=0 then timechan[loc(timechan=0 & change=0),1]=9999;

  if min(timechan)<2 then timecat[loc(timechan<2 & change>0),1]=1; ** new MCR risk group: lower;
  if min(timechan)<1 then timecat[loc(timechan<1 & change>0),1]=0; ** new MCR risk group: lowest;
  if max(covar[,5])=3 then timecat[loc(covar[,5]=3),1]=3;    *** patients remain in highest risk;
  change=covar[,5]-timecat;                     *** #(risk groups improved);

  if sum(change>0) then do;  *** loop for risk group changes due to covariable j;
    covar[,5]=timecat;               *** changed MCR risk group in fifth column of matrix covar;
    hv2=hv1#change;                  *** matrix hv2 indicates places in "at risk" matrix hv1
                                     *** where an MCR risk group change was observed;
    *** From higher risk to lowest risk: two risk groups improved => hv2[,5] or hv2[,6]=2;
     *** "at risk" matrix hv1 is appropriately changed for patients with survival status "death";
     if max(hv2[,5])=2 then do;
       hv1[loc(hv2[,5]=2),1]=1; hv1[loc(hv2[,5]=2),5]=0; hv2[loc(hv2[,5]=2),5]=0;
     end;
     *** "at risk" matrix hv1 is appropriately changed for pat. with survival status "censored";
     if max(hv2[,6])=2 then do;
       hv1[loc(hv2[,6]=2),2]=1; hv1[loc(hv2[,6]=2),6]=0; hv2[loc(hv2[,6]=2),6]=0;
     end;
    *** From higher risk to lower risk: one risk group improved => hv2[,5] or hv2[,6]=1;
     *** "at risk" matrix hv1 is appropriately changed for patients with survival status "death";
     if max(hv2[,5])=1 then do;
       hv1[loc(hv2[,5]=1),3]=1; hv1[loc(hv2[,5]=1),5]=0;
     end;
     *** "at risk" matrix hv1 is appropriately changed for pat. with survival status "censored";
     if max(hv2[,6])=1 then do;
       hv1[loc(hv2[,6]=1),4]=1; hv1[loc(hv2[,6]=1),6]=0;
     end;
    *** From lower risk to lowest risk: one risk group improved => hv2[,3] or hv2[,4]=1;
     *** "at risk" matrix hv1 is appropriately changed for pat. with survival status "death";
     if max(hv2[,3])=1 then do;
       hv1[loc(hv2[,3]=1),1]=1; hv1[loc(hv2[,3]=1),3]=0;
     end;
     *** "at risk" matrix hv1 is appropriately changed for pat. with survival status "censored";
     if max(hv2[,4])=1 then do;
       hv1[loc(hv2[,4]=1),2]=1; hv1[loc(hv2[,4]=1),4]=0;
     end;
  end;          *** end of loop for risk group changes due to cov. j: "if sum(change>0) then do";
```

```
      end;          *** end of loop for changes in time-dep. covariable j: "if sum(change>0) then do";
    end;          *** end of loop for counter from first to last time-dep. covar.: "j=1 to numtvar";

    *** lowest risk group: #(pat. at risk when pat. i was censored/died);
    riskhelp[1,1]=sum(hv1[,1])+sum(hv1[,2]);
    *** lower risk group: #(pat. at risk when pat. i was censored/died);
    riskhelp[1,2]=sum(hv1[,3])+sum(hv1[,4]);
    *** higher risk group: #(pat. at risk when pat. i was censored/died);
    riskhelp[1,3]=sum(hv1[,5])+sum(hv1[,6]);
    *** highest risk group: #(pat. at risk when pat. i was censored/died);
    riskhelp[1,4]=sum(hv1[,7])+sum(hv1[,8]);

  end;                   *** end of loop if pat. i does not have same survival time as pat. (i-1):
                         *** "else if covar[i,1]>covar[(i-1),1] then do";

  lastobs=lastobs//hv1[i,1:8]; *** survival status and risk group of patients i at last follow-up;
  nrisk=nrisk//riskhelp;       *** #(patients at risk in the four risk groups at the time when
                               *** patient i was censored or died);
end;                        *** end of loop for counter from first to last patient: "i=2 to nr";

lastobs=lastobs[2:(nrow(lastobs)),]; *** removal of first row with zeros;
nrisk=nrisk[2:(nrow(nrisk)),];       *** removal of first row with #(patients at risk at baseline);
covar=covar[2:(nrow(covar)),];       *** removal of first row with zeros;

*** preparation for table of MCR risk group at last follow-up by survival status;
timeXcat={0 0 1 1 2 2 3 3};       *** MCR risk group;
sustatus={1 0 1 0 1 0 1 0};       *** survival status;
finnum=j(1,8,0);
do i=1 to 8;
  finnum[1,i]=sum(lastobs[,i]);   *** #(patients in each category);
end;

hv1=covar[,1]||lastobs; *** matrix hv1: first column: survival time, columns 2-9: MCR risk groups
                        *** at last follow-up - always 1st column with dead, 2nd with cens. pat.;

*** preparation for table of MCR risk group at last follow-up by survival status AND who are
*** part of the survival probability estimation by Simon-Makuch, since survival times are > x;
help=loc(hv1[,1]>&x);
hv2=hv1[(help[1,1]):(nrow(hv1)),2:9];
curvnum=j(1,8,0);
do i=1 to 8;
  curvnum[1,i]=sum(hv2[,i]);   *** #(patients in each category);
end;

create riskchan var{timeXcat sustatus finnum curvnum}; *** creation of data set "riskchan";
  append;
close riskchan;

*** by matrix m_help: creation of vectors for the MCR risk groups (always 1st dead, 2nd cens. pat.)
*** patients with same survival time are summarized;
create m_help var {survival dead_0 cens_0 dead_1 cens_1 dead_2 cens_2 dead_3 cens_3};
  append from hv1;
summary class {survival} stat {sum}
  var {dead_0 cens_0 dead_1 cens_1 dead_2 cens_2 dead_3 cens_3} opt {noprint save};
close m_help;

hv1=covar[,1]||nrisk; *** matrix hv1: first column: survival time, columns 2-5: #(patients
                      *** at risk in the corresponding MCR risk groups at the given survival times);
```

```
*** by matrix m_help: creation of vectors for the #(patients at risk in their corresponding
*** MCR risk group at each DIFFERENT survival time);
create m_help var {survival nrisk0 nrisk1 nrisk2 nrisk3};
  append from hv1;
summary class {survival} stat {mean}
  var {nrisk0 nrisk1 nrisk2 nrisk3} opt {noprint save};
close m_help;

**** calculation of Mantel-Byar test ****;
hv1=dead_0+dead_1+dead_2+dead_3; *** total #(deaths) at different survival time;
hv1=loc(hv1>0);                  *** locating survival times with total #(deaths)>0;
dead0=dead_0[hv1];  *** #(deaths in lowest risk group) at survival times with total #(deaths)>0;
dead1=dead_1[hv1];  *** #(deaths in lower risk group) at survival times with total #(deaths)>0;
dead2=dead_2[hv1];  *** #(deaths in higher risk group) at survival times with total #(deaths)>0;
dead3=dead_3[hv1];  *** #(deaths in highest risk group) at survival times with total #(deaths)>0;
n0=nrisk0[hv1];     *** #(pat. at risk in lowest risk group) at above survival times;
n1=nrisk1[hv1];     *** #(pat. at risk in lower risk risk group) at above survival times;
n2=nrisk2[hv1];     *** #(pat. at risk in higher group) at above survival times;
n3=nrisk3[hv1];     *** #(pat. at risk in highest risk group) at above survival times;

hv1=dead0+dead1+dead2+dead3; ** Di: total #(deaths) at diff. surv. times i with total #(deaths)>0;
hv2=n0+n1+n2+n3;             ** Ni: total #(pat. at risk) at different survival times i with Di>0;

E=j(1,3,0);               *** vector for observed - expected #(deaths) **;
V=j(3,3,0);               *** matrix for covariance of observed #(deaths) **;
e0=hv1#n0/hv2;
E[1,1]=dead0[+]-e0[+];    *** observed #(deaths)-expected #(deaths) in lowest risk group;
var0=e0#(hv2-n0)#(hv2-hv1)/(hv2#(hv2-1)); *** at survival times i;
V[1,1]=var0[+];           *** variance of observed #(deaths) in lowest risk group;
cov01=-e0#n1#(hv2-hv1)/(hv2#(hv2-1));
V[1,2]=cov01[+];          *** covariance of observed #(deaths) between lowest and lower risk group;
V[2,1]=cov01[+];          *** covariance of observed #(deaths) between lowest and lower risk group;
cov02=-e0#n2#(hv2-hv1)/(hv2#(hv2-1));
V[1,3]=cov02[+];          *** covariance of observed #(deaths) between lowest and higher risk group;
V[3,1]=cov02[+];          *** covariance of observed #(deaths) between lowest and higher risk group;
e1=hv1#n1/hv2;
E[1,2]=dead1[+]-e1[+];    *** observed #(deaths)-expected #(deaths) in lower risk group;
var1=e1#(hv2-n1)#(hv2-hv1)/(hv2#(hv2-1));
V[2,2]=var1[+];           *** variance of observed #(deaths) in lower risk group;
cov12=-e1#n2#(hv2-hv1)/(hv2#(hv2-1));
V[2,3]=cov12[+];          *** covariance of observed #(deaths) between lower and higher risk group;
V[3,2]=cov12[+];          *** covariance of observed #(deaths) between lower and higher risk group;
e2=hv1#n2/hv2;
E[1,3]=dead2[+]-e2[+];    *** observed #(deaths)-expected #(deaths) in higher risk group;
var2=e2#(hv2-n2)#(hv2-hv1)/(hv2#(hv2-1));
V[3,3]=var2[+];           *** variance of observed #(deaths) in higher risk group;

chi=E*INV(V)*T(E);        *** test statistic for Mantel-Byar test **;
pchi=1-probchi(chi,3);    *** and its p-value **;

**** output of results for calculated Mantel-Byar test ****;
print "            Chi-Square:" chi     "   p-value:" pchi;

*** calculation of the surv. funct. S0, S1, S2, S3 - plus cens. times cens0, cens1, cens2, cens3;
*** selection of the times >= x (starting time for Simon-Makuch curves) ****;
hv1=loc(survival>&x);
survival=survival[hv1];
cens_0=cens_0[hv1];
dead_0=dead_0[hv1];
```

```
cens_1=cens_1[hv1];
dead_1=dead_1[hv1];
cens_2=cens_2[hv1];
dead_2=dead_2[hv1];
cens_3=cens_3[hv1];
dead_3=dead_3[hv1];
nrisk0=nrisk0[hv1];
nrisk1=nrisk1[hv1];
nrisk2=nrisk2[hv1];
nrisk3=nrisk3[hv1];

survival=survival/&time;  *** macro variable &time defines the time unit;
nt=nrow(survival);        *** #(different survival times > x);

*** initialisation of survival functions and censored times;
s0=1; s1=1; s2=1; s3=1; *** start of the vectors for the surv. prob. - one for each risk group;

*** variables for the current survival prob. at survival time t - one per group;
surv0=1; surv1=1; surv2=1; surv3=1;
*** start of the vectors for the stand. deviation of the surv. prob. - one for each risk group;
sdv0=0; sdv1=0; sdv2=0; sdv3=0;
*** variables helping to calculate stand. deviation of the surv. prob. - one for each risk group;
hvsdv0=0; hvsdv1=0; hvsdv2=0; hvsdv3=0;
*** start of vectors for the survival times in the corresponding risk groups;
time0=0; time1=0; time2=0; time3=0;
*** start of the vectors for the survival probabilities at censored times one for each risk group;
cens0=1; cens1=1; cens2=1; cens3=1;
*** start of vectors for the censored times in the corresponding risk groups;
ti0cens=0; ti1cens=0; ti2cens=0; ti3cens=0;
*** variables to always save the last survival prob. at the last censored time in each risk group;
lcensu0=1; lcensu1=1; lcensu2=1; lcensu3=1;


*** calculation of survival probabilies and vectors for observed survival and censored times;
do t=1 to nt;          *** loop for counter from first to last survival time: "t=1 to nt";

  if dead_0[t]^=0 then do;               *** if a death at time t was observed then;
    surv0=surv0*(1-dead_0[t]/nrisk0[t]); *** calculate the current survival probability by;
    s0=s0//surv0;                        *** multiplying the previous survival probability (surv0);
    time0=time0//survival[t];            *** with the current one(1-#(deaths at t)/#(at risk at t).;
                                         *** Add the new survival probability to vector s0 and the
                                         *** survival time t to vector time0;
    if (nrisk0[t]-dead_0[t])^=0 then do;
      hv1=dead_0[t]/(nrisk0[t]*(nrisk0[t]-dead_0[t]));
      hvsdv0=hvsdv0//hv1;
      hv1=surv0*sqrt(hvsdv0[+]);  *** the standard deviation of the survival probability at time t;
      sdv0=sdv0//hv1;             *** vector of standard deviations at event times;
    end;
  end;
  if cens_0[t]^=0 then do;        *** if a patient was censored at time t then add the last survival;
    cens0=cens0//surv0; lcensu0=surv0; *** prob. surv0 to vector cens0 and also save it in lcensu0.;
    *** Add the censored time t to vector ti0cens and also save it in lcenti0;
    ti0cens=ti0cens//survival[t]; lcenti0=survival[t];
  end;

  if dead_1[t]^=0 then do;        *** lower risk: as described above for the lowest risk group;
    surv1=surv1*(1-dead_1[t]/nrisk1[t]);
    s1=s1//surv1;
    time1=time1//survival[t];
```

```
    if (nrisk1[t]-dead_1[t])^=0 then do;
      hv1=dead_1[t]/(nrisk1[t]*(nrisk1[t]-dead_1[t]));
      hvsdv1=hvsdv1//hv1;
      hv1=surv1*sqrt(hvsdv1[+]);  *** the standard deviation of the survival probability at time t;
      sdv1=sdv1//hv1;             *** vector of standard deviations at event times;
    end;
  end;
  if cens_1[t]^=0 then do;        *** lower risk: as described above for the lowest risk group;
    cens1=cens1//surv1; lcensu1=surv1;
    ti1cens=ti1cens//survival[t]; lcenti1=survival[t];
  end;

  if dead_2[t]^=0 then do;        *** higher risk: as described above for the lowest risk group;
    surv2=surv2*(1-dead_2[t]/nrisk2[t]);
    s2=s2//surv2;
    time2=time2//survival[t];
    if (nrisk2[t]-dead_2[t])^=0 then do;
      hv1=dead_2[t]/(nrisk2[t]*(nrisk2[t]-dead_2[t]));
      hvsdv2=hvsdv2//hv1;
      hv1=surv2*sqrt(hvsdv2[+]);  *** the standard deviation of the survival probability at time t;
      sdv2=sdv2//hv1;             *** vector of standard deviations at event times;
    end;
  end;
  if cens_2[t]^=0 then do;        *** higher risk: as described above for the lowest risk group;
    cens2=cens2//surv2; lcensu2=surv2;
    ti2cens=ti2cens//survival[t]; lcenti2=survival[t];
  end;

  if dead_3[t]^=0 then do;        *** highest risk: as described above for the lowest risk group;
    surv3=surv3*(1-dead_3[t]/nrisk3[t]);
    s3=s3//surv3;
    time3=time3//survival[t];
    if (nrisk3[t]-dead_3[t])^=0 then do;
      hv1=dead_3[t]/(nrisk3[t]*(nrisk3[t]-dead_3[t]));
      hvsdv3=hvsdv3//hv1;
      hv1=surv3*sqrt(hvsdv3[+]);  *** the standard deviation of the survival probability at time t;
      sdv3=sdv3//hv1;             *** vector of standard deviations at event times;
    end;
  end;
  if cens_3[t]^=0 then do;        *** highest risk: as described above for the lowest risk group;
    cens3=cens3//surv3; lcensu3=surv3;
    ti3cens=ti3cens//survival[t]; lcenti3=survival[t];
  end;

end;               *** end of loop for counter from first to last survival time: "t=1 to nt";

if lcensu0=surv0 then do;  *** if last obs. in the lowest risk group was censored, a line has to be;
  s0=s0//surv0;            *** drawn to that obs. That is why vars lcensu0 and lcenti0 are needed;
  time0=time0//lcenti0;
  sdv0=sdv0//sdv0[(nrow(sdv0)),];
end;
if lcensu1=surv1 then do;    *** as described above for the lowest risk group;
  s1=s1//surv1;
  time1=time1//lcenti1;
  sdv1=sdv1//sdv1[(nrow(sdv1)),];
end;
if lcensu2=surv2 then do;    *** as described above for the lowest risk group;
  s2=s2//surv2;
  time2=time2//lcenti2;
```

```
   sdv2=sdv2//sdv2[(nrow(sdv2)),];
 end;
 if lcensu3=surv3 then do;    *** as described above for the lowest risk group;
   s3=s3//surv3;
   time3=time3//lcenti3;
   sdv3=sdv3//sdv3[(nrow(sdv3)),];
 end;

 print "Simon-Makuch curves from time to response:" &x;

 *** output files for the Simon-Makuch curves for the four risk groups;
 create m_outd0 var{time0 s0 sdv0}; *** m_outd0 with surv. times and surv. prob. of lowest risk group;
   append;
 close m_outd0;
 create m_outc0 var{ti0cens cens0}; *** m_outc0 with cens. times and surv. prob. of lowest risk group;
   append;
 close m_outc0;
 create m_outd1 var{time1 s1 sdv1}; *** lower risk: as described above for the lowest risk group;
   append;
 close m_outd1;
 create m_outc1 var{ti1cens cens1}; *** lower risk: as described above for the lowest risk group;
   append;
 close m_outc1;
 create m_outd2 var{time2 s2 sdv2}; *** higher risk: as described above for the lowest risk group;
   append;
 close m_outd2;
 create m_outc2 var{ti2cens cens2}; *** higher risk: as described above for the lowest risk group;
   append;
 close m_outc2;
 create m_outd3 var{time3 s3 sdv3}; *** highest risk: as described above for the lowest risk group;
   append;
 close m_outd3;
 create m_outc3 var{ti3cens cens3}; *** highest risk: as described above for the lowest risk group;
   append;
 close m_outc3;

 *** end of modul surv to calculate Simon-Makuch curves and Mantel-Byar test;
 finish surv;

 *** start modul surv to calculate Simon-Makuch curves and Mantel-Byar test;
 run surv;
quit;

data m_outd0;  *** surv. prob., sdv, and survival times of lowest risk group: group='s0' are kept;
  set m_outd0;
  group='s0';
  sdf=s0;
  time=time0;
  sdv=sdv0;
  keep group time sdf sdv;

data m_outc0;  *** surv. prob. and censored times of lowest risk group: group='z0' are kept;
  set m_outc0;
  group='z0';
  sdf=cens0;
  time=ti0cens;
  keep group time sdf;

data m_outd1;  *** surv. prob., sdv, and survival times of lower risk group: group='s1' are kept;
```

```
  set m_outd1;
  group='s1';
  sdf=s1;
  time=time1;
  sdv=sdv1;
  keep group time sdf sdv;

data m_outc1;  *** surv. prob. and censored times of lower risk group: group='z1' are kept;
  set m_outc1;
  group='z1';
  sdf=cens1;
  time=ti1cens;
  keep group time sdf;

data m_outd2;  *** surv. prob., sdv, and survival times of higher risk group: group='s2' are kept;
  set m_outd2;
  group='s2';
  sdf=s2;
  time=time2;
  sdv=sdv2;
  keep group time sdf sdv;

data m_outc2;  *** surv. prob. and censored times of higher risk group: group='z2' are kept;
  set m_outc2;
  group='z2';
  sdf=cens2;
  time=ti2cens;
  keep group time sdf;

data m_outd3;  *** surv. prob., sdv, and survival times of highest risk group: group='s3' are kept;
  set m_outd3;
  group='s3';
  sdf=s3;
  time=time3;
  sdv=sdv3;
  keep group time sdf sdv;

data m_outc3;  *** surv. prob. and censored times of highest risk group: group='z3' are kept;
  set m_outc3;
  group='z3';
  sdf=cens3;
  time=ti3cens;
  keep group time sdf;

data outdata;
  merge m_outd0 m_outc0 m_outd1 m_outc1 m_outd2 m_outc2 m_outd3 m_outc3;
  by group;
  lower=sdf-1.96*sdv; if . lt lower lt 0 then lower=0;  *** limit of lower confidence interval ***;
  upper=sdf+1.96*sdv; if upper gt 1 then upper=1;       *** limit of upper confidence interval ***;

proc sort data=outdata out=plot;
  by group descending time;
  where group le 's3';

**************** graphical representation of results ****************;
***
*** Word ***;
goptions reset=all gsfname=grafout gsfmode=replace gaccess=sasgastd dashscale=0.4
          targetdevice=pslmono hsize=10 vsize=7 hpos=80 vpos=35 ftext=swiss
```

```
          device=cgmofml lfactor=2.5 rotate=landscape;

*** Postscript ***;
*goptions reset=all rotate=landscape lfactor=2.5 gsfmode=replace gsfname=grafout
         hsize=10 vsize=7 hpos=80 vpos=35 dashscale=0.4 ftext=swiss
         device=pslmono;

*** Tex;
*goptions reset=all gsfname=grafout gsfmode=replace horigin=3.0cm vorigin=15.5cm
         ftext=swiss hsize=15.0cm vsize=10.5cm hpos=80 vpos=35
         device=pslmono lfactor=2.5;

*** Preparations to allow for Greenwood 95% confidence intervals at times &pt1, &pt2 and &pt3;
data plot;
  set plot;
  by group descending time;
  retain lastzeit;
  if first.group then lastzeit=time;

proc sort data=plot;
  by group descending sdf time;

data uez1 uez2 uez3;
  set plot;
  by group descending sdf;

  if time le &pt1 and lastzeit ge &pt1 then output uez1;
  if time le &pt2 and lastzeit ge &pt2 then output uez2;
  if time le &pt3 and lastzeit ge &pt3 then output uez3;
run;

data uez1;
  set uez1; by group; time=&pt1;
  if last.group and time ne . then output;
run;

data uez2;
  set uez2; by group; time=&pt2;
  if last.group and time ne . then output;
run;

data uez3;
  set uez3; by group; time=&pt3;
  if last.group and time ne . then output;
run;

data inval;
  set uez1 uez2 uez3;

  length function $8;
  xsys='2'; ysys='2';

  **** Greenwood 95% confidence interval for lowest risk group;
  if group = 's0' then do;
    function='move'; x=time; y=sdf; output;
    function='symbol'; size=0.2; text='dot'; output;
    function='draw'; l=1; size=2.0; x=time; y=lower; output;
    function='draw'; l=1; size=2.0; x=time-0.1; y=lower; output;
    function='draw'; l=1; size=2.0; x=time+0.1; y=lower; output;
```

```
    function='move'; x=time; y=sdf; output;
    function='draw'; l=1; size=2.0; x=time; y=upper; output;
    function='draw'; l=1; size=2.0; x=time-0.1; y=upper; output;
    function='draw'; l=1; size=2.0; x=time+0.1; y=upper; output;
  end;

  **** Greenwood 95% confidence interval for lower risk group;
  if group = 's1' then do;
    function='move'; x=time; y=sdf; output;
    function='symbol'; size=0.2; text='dot'; output;
    function='draw'; l=1; size=2.0; x=time; y=lower; output;
    function='draw'; l=1; size=2.0; x=time-0.2; y=lower; output;
    function='draw'; l=1; size=2.0; x=time+0.2; y=lower; output;
    function='move'; x=time; y=sdf; output;
    function='draw'; l=1; size=2.0; x=time; y=upper; output;
    function='draw'; l=1; size=2.0; x=time-0.2; y=upper; output;
    function='draw'; l=1; size=2.0; x=time+0.2; y=upper; output;
  end;

  **** Greenwood 95% confidence interval for higher risk group;
  if group = 's2' then do;
    function='move'; x=time; y=sdf; output;
    function='symbol'; size=0.2; text='dot'; output;
    function='draw'; l=1; size=2.0; x=time; y=lower; output;
    function='draw'; l=1; size=2.0; x=time-0.3; y=lower; output;
    function='draw'; l=1; size=2.0; x=time+0.3; y=lower; output;
    function='move'; x=time; y=sdf; output;
    function='draw'; l=1; size=2.0; x=time; y=upper; output;
    function='draw'; l=1; size=2.0; x=time-0.3; y=upper; output;
    function='draw'; l=1; size=2.0; x=time+0.3; y=upper; output;
  end;

  **** Greenwood 95% confidence interval for highest risk group;
  if group = 's3' then do;
    function='move'; x=time; y=sdf; output;
    function='symbol'; size=0.2; text='dot'; output;
    function='draw'; l=1; size=2.0; x=time; y=lower; output;
    function='draw'; l=1; size=2.0; x=time-0.4; y=lower; output;
    function='draw'; l=1; size=2.0; x=time+0.4; y=lower; output;
    function='move'; x=time; y=sdf; output;
    function='draw'; l=1; size=2.0; x=time; y=upper; output;
    function='draw'; l=1; size=2.0; x=time-0.4; y=upper; output;
    function='draw'; l=1; size=2.0; x=time+0.4; y=upper; output;
  end;

run;

*** Preparing the legends;
data label;
  length function $8;
  length text $70;
  length color $8;
  xsys='2'; ysys='2';

  symbol4 v=point i=stepj line=1 width=3 c=green;
  symbol3 v=point i=stepj line=8 width=3 c=lilac;
  symbol2 v=point i=stepj line=2 width=3 c=blue;
  symbol1 v=point i=stepj line=1 width=3 c=black;
```

```
  color='green';
  function = 'move';  x=0.2; y=0.02; line=1; size=3.0; output;
  function = 'draw';  x=0.6; y=0.02;         output;
  function = 'label'; x=1.0; y=0.02; size=1.0; position='6'; text=&legend4; output;

  color='lilac';
  function = 'move';  x=0.2; y=0.07;  line=8; size=3.0; output;
  function = 'draw';  x=0.6; y=0.07;          output;
  function = 'label'; x=1.0; y=0.07;  size=1.0; position='6'; text=&legend3; output;

  color='blue';
  function = 'move';  x=5.4; y=0.94; line=2; size=3.0; output;
  function = 'draw';  x=5.8; y=0.94;         output;
  function = 'label'; x=6.2; y=0.94; size=1.0; position='6';  text=&legend2; output;

  color='black';
  function = 'move';  x=5.4; y=0.99; line=1; size=3.0; output;
  function = 'draw';  x=5.8; y=0.99;         output;
  function = 'label'; x=6.2; y=0.99; size=1.0; position='6'; text=&legend1; output;

data label;
  set label inval;

  *** saving the graphical representation of the Simon-Makuch curves in file &outprn;
  filename grafout &outprn;

  title1 j=c h=1.5 &title1;
  title2 j=c h=1   &title2;
  title3 j=c h=1   &title3;
  footnote j=l h=0.5 &footnote;

  symbol4 v=point i=stepj line=1 width=3 c=green;
  symbol3 v=point i=stepj line=8 width=3 c=lilac;
  symbol2 v=point i=stepj line=2 width=3 c=blue;
  symbol1 v=point i=stepj line=1 width=3 c=black;
  symbol8 f=bzt v=h i=none c=green;
  symbol7 f=bzt v=h i=none c=lilac;
  symbol6 f=bzt v=h i=none c=blue;
  symbol5 f=bzt v=h i=none c=black;

  axis1 label=(j=r h=1.2 "&tlabel")
      %if %length(&tsmall) > 0 %then %do; minor=(number=&tsmall) %end; value=(h=1.2)
      %if %length(&torder) > 0 %then %do; order=(&torder) %end; width=2
      ;
  axis2 label=(j=c h=1.2 rotate=360 a=90 "&plabel") value=(h=1.2)
      order=0.0 to 1.0 by 0.1 width=2;

*** plotting the Simon-Makuch curves;
proc gplot data=outdata annotate=label;
  plot sdf*time=group / nolegend haxis=axis1 vaxis=axis2;
run;

quit;

*** table of survival status at last follow-up by MCR risk group at last follow-up;
proc freq data=riskchan;
  tables sustatus*timeXcat;
  weight finnum;
  format timeXcat score. sustatus uesta.;
```

```
  label sustatus ='Survival Status'
        timeXcat ='MCR score: Risk group at last follow-up';
run;

*** table of survival status at last follow-up by MCR risk group at last follow-up;
*** BUT ONLY with patients who are part of the survival probability estimation by Simon-Makuch,
*** i.e. where survival times are > x;
proc freq data=riskchan;
  tables sustatus*timeXcat;
  weight curvnum;
  format timeXcat score. sustatus uesta.;
  label sustatus ='Survival Status'
        timeXcat ="MCR score: Risk group at last follow-up with surv. times > &x";
run;

*** deletion of auxiliary datasets;
proc datasets;
  delete riskchan m_help m_outd0 m_outc0 m_outd1 m_outc1 m_outd2 m_outc2 m_outd3 m_outc3
         plot uez1 uez2 uez3;
run;
quit;
%mend ccrscore;
*** end of macro;

*** The following example should be part of a programme CALLING the MACROS CCRSCORE;
*%ccrscore (indata   = data ,
           cond     = where augie07 ne . and zeroscor ne .,
           sutime   = survival ,
           sustatus = status ,
           censor   = 0 1 ,
           hasscore = augie07 ,
           timezero = zeroscor ,
           time0cat = zeroscor ,
           times    = pcrtime ccrtime ,
           score    = -1 -2 ,
           chanvar  = 2 ,
           title1   = "Simon-Makuch curves" ,
           title2   = "from end of the 3rd months" ,
           title3   = "for 758 patients treated with IFN" ,
           x        = 91 ,
           time     = 365.25 ,
           tlabel   = Years after start of therapy,
           plabel   = Survival probability after 3 months ,
           torder   = 0 to 13 ,
           tsmall   = 3 ,
           legend1  = "Lowest risk (157/19), 10-yr surv. prob.: 0.78",
           legend2  = "Lower risk (277/107), median surv.: 77 months",
           legend3  = "Higher risk (222/125), median surv.: 62 months",
           legend4  = "Highest risk (102/74), median surv.: 42 months",
           pt1      = 3 ,
           pt2      = 6 ,
           pt3      = . ,
           print    = 1 ,
           footnote = "Munich, &sysdate (T:\GISeV\epfp\response\ccrscore.mac)",
           outprn   = "T:\GISeV\epfp\response\ccrscore91.cgm");

/****
*** explanation of macro keyword parameters and examples for input
%ccrscore
```

```
  (indata   = data ,                                  *** input data set with relevant variables
   cond     = where augie07 ne .  and zeroscor ne ., *** selection criteria regarding input data
   sutime   = survival ,                              *** survival times (e.g. in days after time=0)
   sustatus = status ,                                *** survival status
   censor   = 0 1 ,                                   *** values of survival status to be censored
   hasscore = augie07 ,                               *** variable with New CML (Hasford) score,
                                                      *** needed to identify highest risk and
                                                      *** thus used as selection criterium
   timezero = zeroscor ,                              *** MCR score at baseline (time=0)
   time0cat = zeroscor ,                              *** MCR risk group at baseline (time=0)
                                                      *** (if timezero and time0cat are different
                                                      *** which used to be the case at an earlier
                                                      *** stage of model development)
   times    = pcrtime ccrtime ,                       *** times to PCR and times to CCR
                                                      *** by matters of programming: "pcrtime" should
                                                      *** be placed in front of "ccrtime".
   score    = -1, -2 ,                                *** MCR score values for PCR and CCR
   chanvar  = 2 ,                                     *** indicates the place of time-dependent variable
                                                      *** "ccrtime". In this macro always choose "2"
   title1   = "Simon-Makuch curves" ,                 *** possibility to choose title1
   title2   = "from end of the 3rd months" ,          *** possibility to choose title2
   title3   = "for 758 patients treated with IFN" ,   *** possibility to choose title3
   x        = 91 ,                                    *** starting time for Simon-Makuch curves
                                                      *** (here: in days after time=0)
   time     = 365.25 ,                                *** possibility to change survival time unit for
                                                      *** graphical representation
   tlabel   = Years after start of therapy ,          *** possibility to choose name for time-axis
   plabel   = Survival after 3 mo. response time ,    *** possib. to choose name for surv. prob.-axis
   torder   = 0 to 13 ,                               *** range for time-axis (after unit change: years)
   tsmall   = 3 ,                                     *** number of minor tick marks drawn between each
                                                      *** major tick mark
   legend1  = Lowest risk  ,                          *** possibility to choose legend for curve 1
   legend2  = Lower risk ,                            *** possibility to choose legend for curve 2
   legend3  = Higher risk ,                           *** possibility to choose legend for curve 3
   legend4  = Highest risk ,                          *** possibility to choose legend for curve 4
   pt1      = 3 ,                                     *** first possibility to choose time for
                                                      *** Greenwood 95% confidence interval
   pt2      = 6 ,                                     *** second possibility to choose time for
                                                      *** Greenwood 95% confidence interval
   pt3      = 9 ,                                     *** third possibility to choose time for
                                                      *** Greenwood 95% confidence interval
   print    = 1 ,                                     *** if print=1 then MCR risk groups at baseline
                                                      *** will also be printed
                                                      *** possibility to choose footnote
   footnote = "Munich, &sysdate (T:\GISeV\epfp\response\ccrscore.mac)",
   outprn   = "T:\GISeV\epfp\response\ccrscore.cgm"); *** possibility to choose name for graph
*** end of explanation of macro keyword parameters and examples for input;
****/
```

Literaturverzeichnis

[1] G. Alimena, E. Morra, M. Lazzarino, A.M. Liberati, E. Montefusco, D. Inveradi, P. Bernasconi, M. Mancini, E. Donti, F. Grignani, C. Bernasconi, F. Dianzani, and F. Mandelli. Interferon alpha-2b therapy for Ph'-positive chronic myelogenous leukemia: a study of 82 patients treated with intermittent or daily administration. *Blood*, 72:642-647, 1988.

[2] N.C. Allan, S.M. Richards, and P.C.A. Shepherd, on behalf of the UK Medical Research Council's Working Parties for Therapeutic Trials in Adult Leukemia. UK Medical Research Council randomised, multicentre trial of interferon-αn1 for chronic myeloid leukemia: improved survival irrespective of cytogenetic response. *The Lancet*, 345:1392-1397, 1995.

[3] D.G. Altman. Systematic reviews of evaluations of prognostic variables. *British Medical Journal*, 323:224-228, 2001.

[4] D.G. Altman and P.K. Andersen. Bootstrap investigation of the stability of a Cox regression model. *Statistics in Medicine*, 8:771-783, 1989.

[5] D.G. Altman and B.L. De Stavola. Practical problems in fitting a proportional hazards model to data with updated measurements of the covariates. *Statistics in Medicine*, 13:301-341, 1994.

[6] D.G. Altman, B. Lausen, W. Sauerbrei, and M. Schumacher. Commentary: Dangers of using „optimal" cutpoints in the evaluation of prognostic factors. *Journal of National Cancer Institute*, 86:829-835, 1994.

[7] P.K. Andersen, O. Borgan, R.D. Gill, and N. Keiding. *Statistical models based on counting processes.* Springer-Verlag, New York, 1993.

[8] J.R. Anderson, K.C. Cain, and R.D. Gelber. Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1:710-719, 1983.

[9] G.R. Angstreich, B.D. Smith, and R.J. Jones. Treatment options for chronic myeloid leukemia: imatinib versus interferon versus allogeneic transplant. *Current Opinion in Oncology*, 16(2):95-99, 2004.

[10] H. Ansari, Ü. Aydemir, P. Dirschedl, J. Hasford, and R. Hehlmann. Analyse von prognostischen Faktoren bei Patienten mit chronisch-myeloischer Leukämie (CML) - ein Vergleich von CART-Technik und Cox-Modell. In: J. Michaelis, G. Hommel und S. Wellek (Hrsg.), *Europäische Perspektiven der Medizinischen Informatik, Biometrie und Epidemiologie. Medizinische Informatik, Biometrie und Epidemiologie 76*, pp. 168-172. MMV Medizin Verlag, München, 1993.

[11] H. Ansari, J. Hasford, R. Hehlmann, and the German CML Study Group. Fallacies of the intention-to-treat analysis. *Journal of Molecular Medicine*, 75:B243-B244, 1997.

[12] V. Apgar. A proposal for a new method of evaluation of the newborn infant. *Current Researches in Anasthesia and Analgesia*, 32:260-267, 1953.

[13] M. Baccarani, G. Rosti, A. de Vivo, F. Bonifazi, D. Russo, G. Martinelli, N. Testoni, M. Amabile, M. Fiacchini, E. Montefusco, G. Saglio, and S. Tura, for the Italian Cooperative Group on Chronic Myeloid Leukemia. A randomized study of interferon-α versus interferon-α and low-dose arabinosyl cytosine in chronic myeloid leukemia. *Blood*, 99:1527-1535, 2002.

[14] W.E. Barlow and R.L. Prentice. Residuals for relative risk regression. *Biometrika*, 75:65-74, 1988.

[15] The Benelux CML Study Group. Randomized study on hydroxyurea alone versus hydroxyurea combined with low-dose interferon-α2b for chronic myeloid leukemia. *Blood*, 91:2713-2721, 1998.

[16] U. Berger, G. Engelich, A. Reiter, A. Hochhaus, and R. Hehlmann. Imatinib and beyond - the new CML study IV. *Annals of Hematology*, 83:258-264, 2004.

[17] U. Berger, A. Hochhaus, M. Pfirrmann, C. Schoch, A. Reiter, G. Ehninger, T. Fischer, A. Gratwohl, J. Hasford, H. Heimpel, D.K. Hossfeld, H.-J. Kolb, S. Krause, C. Nerl, H. Pralle, A. Tobler, R. Hehlmann, and the German CML-Study Group. Concept, feasibility and results of the randomized comparison of imatinib combination therapies for chronic myeloid leukemia: The German CML-Study IV. *Blood*, 106:315a-316a, 2005.

[18] F. Bonifazi, A. de Vivo, G. Rosti, M. Tiribelli, D. Russo, E. Trabacchi, M. Fiacchini, E. Montefusco, and M. Baccarani. Testing Sokal's and the new prognostic score for chronic myeloid leukaemia treated with α-interferon. *British Journal of Haematology*, 111:587-595, 2000.

[19] F. Bonifazi, A. de Vivo, G. Rosti, F. Guilhot, J. Guilhot, E. Trabacchi, R. Hehlmann, A. Hochhaus, P.C.A. Shepherd, J.L. Steegmann, H.C. Kluin-Nelemans, J. Thaler, B. Simonsson, A. Louwagie, J. Reiffers, F.X. Mahon, E. Montefusco, G. Alimena, J. Hasford, S. Richards, G. Saglio, N. Testoni, G. Martinelli, S. Tura, and M. Baccarani, for the European Study Group on Interferon in Chronic Myeloid Leukemia. Chronic myeloid leukemia and interferon-alpha: a study of complete cytogenetic responders. *Blood*, 98:3074-3081, 2001.

[20] R. Brandmaier, H. Ansari, and D. Dickson. A general SAS macro for construction of classification and regression trees. In: *Proceedings of the 18th annual SAS User Group International (SUGI) Conference*, pp. 1032-1035. SAS Institute Inc., Cary, NC, 1993.

[21] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. *Classification and regression trees (CART).* Wadsworth, Belmont, 1984.

[22] N. Breslow. Covariance analysis of censored survival data. *Biometrics*, 30:89-99, 1974.

[23] D.P. Byar. Identification of prognostic factors. In: M.E. Buyse, M.J. Staquet, R.J. Sylvester (eds.), *Cancer clinical trial. Methods and practice*, pp. 423-443. Oxford University Press, Oxford, 1984.

[24] E. Christensen, P. Schlichting, P.K. Andersen, L. Fauerholdt, G. Schou, B. Vestergaard Pedersen, E. Juhl, H. Poulsen, N. Tygstrup, and Copenhagen Study Group for Liver Diseases. Updating prognosis and therapeutic effect evaluation in cirrhosis with Cox's multiple regression model for time-dependent variables. *Scandinavian Journal of Gastroenterology*, 21:163-174, 1986.

[25] CML-Studiengruppe (CML-SG), Süddeutsche Hämoblastosegruppe (SHG) e.V. und Schweizerische Arbeitsgruppe für Klinische Krebsforschung (SAKK). *Qualitätssicherungsprotokoll zur Therapieoptimierung bei chronischer myeloischer Leukämie (CML). Randomisierter kontrollierter Vergleich von Imatinib vs. Imatinib und Interferon-α vs. Imatinib 800 mg mit Prüfung des Stellenwertes der allogenen Stammzelltransplantation bei neu diagnostizierter CML in chronischer Phase. Kurzformel: CML-Studie IV.* R. Hehlmann, III. Medizinische Klinik, Klinikum Mannheim (Hrsg.). Fassung Mai 2005.

[26] Chronic Myeloid Leukemia Trialists' Collaborative Group. Interferon alfa versus chemotherapy for chronic myeloid leukemia: a meta-analysis of seven randomized trials. *Journal of the National Cancer Institute*, 89:1616-1620, 1997.

[27] D. Collett. *Modelling survival data in medical research.* Chapman & Hall, London, 1994.

[28] D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society, Series B*, 34:187-220, 1972.

[29] P. Dirschedl. Klassifikationsbäume - Grundlagen und Neuerungen. In: Fleischer et al. (Hrsg.), *Interaktive Datenanalyse*, pp. 15-30. Westarp, Essen, 1992.

[30] M. Deininger, E. Buchdunger, and B.J. Druker. The development of imatinib as a therapeutic agent for chronic myeloid leukemia. *Blood* 105:2640-2653, 2005.

[31] B.J. Druker, M. Talpaz, D.J. Resta, B. Peng, E. Buchdunger, J.M. Ford, N.B. Lydon, H. Kantarjian, R. Capdeville, S. Ohno-Jones, and C.L. Sawyers. Efficacy and safety of a specific inhibitor of the bcr-abl tyrosine kinase in chronic myeloid leukemia. *The New England Journal of Medicine*, 344:1031-1037, 2001.

[32] W.D. Dupont and W.D. Plummer. Power and sample size calculations: a review and computer programm. *Controlled Clinical Trials*, 11:116-128, 1990.

[33] W.D. Dupont and W.D. Plummer. Power and sample size calculations for studies involving linear regression. *Controlled Clinical Trials*, 19:589-601, 1998.

[34] S. Faderl, M. Talpaz, Z. Estrov, and H.M. Kantarjian. Chronic myelogenous leukemia: biology and therapy *Annals of Internal Medicine*, 131:207-219, 1999.

[35] T.A. Gooley, W. Leisenring, J. Crowley, and B.E. Storer. Estimation of failure probabilities in the presence of competing risks: New representations of old estimators *Statistics in Medicine*, 18:695-706, 1999.

[36] M. Greenwood. A report on the natural duration of cancer. *Reports on Public Health and Medical Subjects 33*, pp. 1-26. H.M. Stationary Office, London, 1926.

[37] F. Guilhot, A. Guerci, D. Fiere, J.-L. Harousseau, F. Maloisel, R. Bouabdallah, D. Guyotat, H. Rochant, A. Najman, F. Nicolini, P. Colombat, J.-F. Abgrall, N. Ifrah, J. Brière, F. Bauters, M. Navarro, P. Morice, D. Bordessoule, J.P. Vilque, B. Desablens, G. Tertian,

M. Blanc, C. Chastang, and J. Tanzer, on behalf of the French CML study group. The treatment of chronic myelogenous leukemia by interferon and cytosine-arabinoside: rational and design of the French trials. *Bone Marrow Transplantation*, 17 (Suppl. 3):S29-S31, 1996.

[38] F. Guilhot, C. Chastang, M. Michallet, A. Guerci, J.-L. Harousseau, F. Maloisel, R. Bouabdallah, D. Guyotat, N. Cheron, F. Nicolini, J.-F. Abgrall, and J. Tanzer, for the French chronic myeloid leukemia study group. Interferon alfa-2b combined with cytarabine versus interferon alone in chronic myelogenous leukemia. *The New England Journal of Medicine*, 337:223-229, 1997.

[39] J.A. Hansen, T.A. Gooley, P.J. Martin, F. Appelbaum, T.R. Chauncey, R.A. Clift, F.I.M.L.S., E.W. Petersdorf, J. Radich, J.E. Sanders, R.F. Storb, K.M. Sullivan, and C. Anasetti. Bone marrow transplants from unrelated donors for patients with chronic myeloid leukemia. *The New England Journal of Medicine*, 338:962-968, 1998.

[40] E.K. Harris and A. Albert *Survivorship analysis for clinical studies.* Marcel Dekker, New York, 1991.

[41] J. Hasford, H. Ansari, M. Pfirrmann, and R. Hehlmann. Analysis and validation of prognostic factors for CML. *Bone Marrow Transplantation*, 17 (Suppl. 3):S49-S54, 1996.

[42] J. Hasford, M. Pfirrmann, R. Hehlmann, N.C. Allan, M. Baccarani, J.C. Kluin-Nelemans, G. Alimena, J.L. Steegmann, and H. Ansari. A new prognostic score for survival of patients with chronic myeloid leukemia treated with interferon alfa. *Journal of the National Cancer Institute*, 90:850-858, 1998.

[43] J. Hasford, M. Pfirrmann, R. Hehlmann, P. Shepherd, F. Guilhot, F.X. Mahon, J. Thaler, J.L. Steegmann, H.C. Kluin-Nelemans, A. Louwagie, K. Ohnishi, and O. Kloke. Prognostic factors. In: A. Carella, G. Daley, C. Eaves, J. Goldman, and R. Hehlmann (Hrsg.), *Chronic myeloid leukemia - biology and treatment*, pp. 205-223. Martin Dunitz, London, 2001.

[44] J. Hasford, M. Pfirrmann, R. Hehlmann, M. Baccarani, F. Guilhot, F.X. Mahon, H.C. Kluin-Nelemans, K. Ohnishi, J. Thaler, and J.L. Steegmann for the Collaborative CML Prognostic Factors Group. Prognosis and prognostic factors for patients with chronic myeloid leukemia: nontransplant therapy. *Seminars in Hematology*, 40(1):4-12, 2003.

[45] J. Hasford, M. Pfirrmann, A. Hochhaus. How long will chronic myeloid leukemia patients treated with imatinib mesylate live? *Leukemia*, 19:497-499, 2005.

[46] J. Hasford, M. Pfirrmann, P. Shepherd, J. Guilhot, R. Hehlmann, F.X. Mahon, H.C. Kluin-Nelemans, K. Ohnishi, J.L. Steegmann, and J. Thaler. The impact of the combination of baseline risk group and cytogenetic response on the survival of patients with chronic myeloid leukemia treated with interferon-alpha. *Haematologica*, 90:335-340, 2005.

[47] R. Hehlmann, H. Heimpel, J. Hasford, H.-J. Kolb, H. Pralle, D.K. Hossfeld, W. Queißer, H. Löffler, A. Hochhaus, B. Heinze, A. Georgii, C.R. Bartram, M. Grießhammer, L. Bergmann, U. Essers, C. Falge, U. Queißer, P. Meyer, N. Schmitz, H. Eimermacher, F. Walther, W. Fett, U.R. Kleeberg, A. Käbisch, C. Nerl, R. Zimmermann, G. Meuret, A. Tichelli, L. Kanz, F.-J. Tigges, L. Schmid, W. Brockhaus, A. Tobler, A. Reiter, M. Perker, B. Emmerich, K. Verpoort, R. Zankovich, P. v. Wussow, O. Prümmer, J. Thiele, T. Buhr, F. Carbonell, H. Ansari, and the German CML Study Group. Randomized comparison of interferon-α with busulfan and hydroxyurea in chronic myelogenous Leukemia. *Blood*, 84:4064-4077, 1994.

[48] R. Hehlmann, U. Berger, M. Pfirrmann, A. Hochhaus, G. Metzgeroth, O. Maywald, J. Hasford, A. Reiter, D.K. Hossfeld, H.-J. Kolb, H. Löffler, H. Pralle, W. Queißer, M. Grießhammer, C. Nerl, R. Kuse, A. Tobler, H. Eimermacher, A. Tichelli, C. Aul, M. Wilhelm, J.T. Fischer, M. Perker, C. Scheid, M. Schenk, J. Weiß, C.R. Meier, S. Kremers, L. Labedzki, T. Schmeiser, H.-P. Lohrmann, H. Heimpel, and the German CML Study Group. Randomized comparison of interferon-α and hydroxyurea with hydroxyurea monotherapy in chronic myeloid leukemia (CML-Study II): prolongation of survival by the combination of interferon-α and hydroxyurea. *Leukemia*, 17:1529-1537, 2003.

[49] R. Hehlmann. A chance for cure for every patient with chronic myeloid Leukemia? *The New England Journal of Medicine*, 338:980-982, 1998.

[50] R. Hehlmann, A. Hochhaus, U. Berger, and A. Reiter. Current trends in the management of chronic myelogenous leukemia. *Annals of Hematology*, 79:345-354, 2000.

[51] A. Hochhaus und R. Hehlmann. Chronische myeloische Leukämie. Aktuelle Strategien zur Diagnostik, Therapie und Verlaufskontrolle. *Internist*, 37:1013-1021, 1996.

[52] A. Hochhaus und R. Hehlmann. *Chronische myeloische Leukämie (CML).* UNI-MED, Bremen, 2001.

[53] D.W. Hosmer and S. Lemeshow. *Applied logistic regression.* John Wiley & Sons, New York, 1989.

[54] D.W. Hosmer and S. Lemeshow. *Applied survival analysis. Regression modeling of time to event data.* John Wiley & Sons, New York, 1998.

[55] B.J.P. Huntly, A.G. Reid, A.J. Bench, L.J. Campbell, N. Telford, P. Shepherd, J. Szer, H.M. Prince, P. Turner, C. Grace, E.P. Nacheva, and A.R. Green. Deletions of the derivate chromosome 9 occur at the time of the Philadelphia translocation and provide a powerful and independent prognostic indicator in chronic leukemia. *Blood*, 98:1732-1738, 2001.

[56] The International Non-Hodgkin's Lymphoma Prognostic Factors Project. A predictive model for aggressive Non-Hodgkin's lymphoma. *The New England Journal of Medicine*, 329:987-994, 1993.

[57] The Italian Cooperative Study Group on Chronic Myeloid Leukemia. Interferon alfa-2a as compared with conventional chemotherapy for the treatment of chronic myeloid leukemia. *The New England Journal of Medicine*, 330:820-825, 1994.

[58] The Italian Cooperative Study Group on Chronic Myeloid Leukemia. Long-term follow-up of the Italian trial of interferon-α versus conventional chemotherapy in chronic myeloid leukemia. *Blood*, 92:1541-1548, 1998.

[59] J.D. Kalbfleisch and R.L. Prentice. *The statistical analysis of failure time data.* John Wiley & Sons, New York, 1980.

[60] H.M. Kantarjian, T.L. Smith, S. O'Brien, M. Beran, S. Pierce, M. Talpaz, and the Leukemia Service. Prolonged survival in chronic myelogenous leukaemia after cytogenetic response to interferon-α therapy.*Annals of Internal Medicine*, 122:254-261, 1995.

[61] H. Kantarjian, C. Sawyers, A. Hochhaus, F. Guilhot, C. Schiffer, C. Gambacorti-Passerini, D. Niederwieser, D. Resta, R. Capdeville, U. Zoellner, M. Talpaz, and B. Druker. Hematologic and cytogenetic responsese to imatinib mesylate in chronic myelogenous leukemia. *The New England Journal of Medicine*, 346:645-652, 2002.

[62] H. Kantarjian, S. O'Brien, J. Cortes, F. Giles, J. Shan, M.B. Rios, S. Faderl, S. Verstovsek, G. Garcia-Manero, W. Wierda, S. Kornblau, A. Ferrajoli, S. Giralt, M. Keating, and M. Talpaz. Survival advantage with imatinib mesylate therapy in chronic-phase chronic myelogenous leukaemia (CML-CP) after IFN-α failure and in late CML-CP, comparison with historical controls. *Clinical Cancer Research*, 10:68-75, 2004.

[63] E.L. Kaplan and P. Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457-481, 1958.

[64] H.C. Kluin-Nelemans, G. Buck, S. le Cessie, S. Richards, H.B. Beverloo, J.H.F. Falkenburg, T. Littlewood, P. Muss, D. Bareford, H. van der Lelie, A.R. Green, K.J. Roozendaal, A.E. Milne, C.S. Chapman, and P. Shepherd, for the UK CML Working Group of NCRI and the HOVON Trials Group. Randomized comparison of low-dose versus high-dose interferon-alfa in chronic myeloid leukemia: prospective collaboration of 3 joint trials by MRC and HOVON groups. *Blood*, 103:4408-4415, 2004.

[65] O. Kloke, N. Niederle, J.Y. Qiu, U. Wandl, T. Moritz, M. Nagel-Hiemke, I. Hawig, B. Opalka, S. Seeber, and R. Becher. Impact of interferon alpha-induced cytogenetic improvement on survival in chronic myelogenous leukaemia. *British Journal of Haematology*, 83:399-403, 1993.

[66] O. Kloke, N. Niederle, B. Opalka, I. Hawig, and R. Becher. Prognostic impact of interferon alpha-induced cytogenetic remission in chronic myelogenous leukaemia: long-term follow-up. *European Journal of Haematology*, 56:78-81, 1996.

[67] W.A. Knaus, D.P. Wagner, and J. Lynn. Short-term mortality predictions for critically ill hospitalized adults: science and ethics. *Science*, 254:389-394, 1991.

[68] S. Koller. *Graphische Tafeln zur Beurteilung statistischer Zahlen.* Darmstadt, 1969.

[69] R. Kurzrock, J.U. Gutterman, and M. Talpaz. The molecular genetics of Philadelphia chromosome-positive leukemias. *The New England Journal of Medicine*, 319:990-998, 1988.

[70] A. Laupacis, N. Sekar, and I.G. Stiell. Clinical prediction rules: a review and suggested modifications of methodological standards. *Journal of American Medical Association*, 277:488-494, 1997.

[71] B. Lausen, M. Kersting, and G. Schöch. The regression tree method and its application in nutritional epidemiology. *Informatik, Biometrie und Epidemiologie in Medizin und Biologie*, 28(1):1-13, 1997.

[72] B. Lausen, W. Sauerbrei, and M. Schumacher. Classification and regression trees (CART) used for the exploration of prognostic factors measured on different scales. In: P. Dirschedl and Rüdiger Ostermann (Hrsg.), *Computational Statistics. Papers collected on the occasion of the 25th conference on statistical computing at Schloß Reisensburg*, pp. 483-496. Physica-Verlag, Heidelberg, 1994.

[73] B. Lausen and M. Schumacher. Maximally selected rank statistics. *Biometrics*, 48:73-85, 1992.

[74] F.X. Mahon, C. Fabères, S. Pueyo, P. Cony-Makhoul, R. Salmi, J.M. Boiron, G. Marit, C. Bilhou-Nabera, A. Carrère, M. Montastruc, A. Pigneux, Ph. Bernard, and J. Reiffers. Response at three months is a good predictive factor for newly diagnosed chronic myeloid leukemia patients. *Blood*, 92:4059-4065, 1998.

[75] N. Mantel and D.P. Byar. Evaluation of response-time data involving transient states: an illustration using heart-transplant data. *Journal of the American Statistical Association*, 69:81-86, 1974.

[76] N. Mantel and W. Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22:719-748, 1959.

[77] R. Miller and D. Siegmund. Maximally selected chi square statistics. *Biometrics*, 38:1011-1016, 1982.

[78] M.J. Moro, S. Gil, C. Cañizo, J.F. Clemente, L. Guerras, R.M. Fisac, R. Jimenez-Galindo, and P. Fisac. The treatment of chronic myelogenous leukemia with interferon alfa-2b plus hydroxyurea versus hydroxyurea alone. *Haematologica*, 17 (Suppl. 4):117, 1991.

[79] S.G O'Brien, F. Guilhot, R.A. Larson, I. Gathmann, M. Baccarani, F. Cervantes, J.J. Cornelissen, T. Fischer, A. Hochhaus, T. Hughes, K. Lechner, J.L. Nielsen, P. Rousselot, J. Reiffers, G. Saglio, J. Shepherd, B. Simonsson, A. Gratwohl, J.M. Goldman, H. Kantarjian, K. Taylor, G. Verhoef, A.E. Bolton, R. Capdeville, and B.J. Druker for the IRIS Investigators. Imatinib compared with interferon and low-dose cytarabine for newly diagnosed chronic-phase chronic myeloid leukemia. *New England Journal of Medicine*, 348:994-1004, 2003.

[80] K. Ohnishi, R. Ohno, M. Tomonaga, N. Kamada, K. Onozawa, A. Kuramoto, H. Dohy, H. Mizoguchi, S. Miyawaki, K. Tsubaki, Y. Miura, M. Omine, T. Kobayashi, T. Naoe, T. Ohshima, K. Hirashima, S. Ohtake, I. Takahashi, Y. Morishima, K. Naito, N. Asou, M. Tanimoto, A. Sakuma, K. Yamada, and the Kouseisho Leukemia Study Group. A randomized trial comparing interferon-α with busulfan for newly diagnosed chronic myelogenous leukemia in chronic phase. *Blood*, 86:906-916, 1995.

[81] T. O'Hare, D.K. Walters, E.P. Stoffregen, T. Jia, P.W. Manley, J. Mestan, S.W. Cowan-Jacob, F.Y. Lee, M.C. Heinrich, M.W.N. Deininger, and B.J. Druker. In vitro activity of bcr-abl inhibitors AMN107 and BMS-354825 against clinically relevant imatinib-resistant abl kinase domain mutants. *Cancer Research*, 65:4500-4505, 2005.

[82] B. Otto. Aufbau und Wechselwirkung der Zytokine. In: N. Niederle (Hrsg.), *Zytokine: Präklinik und Klinik*, pp. 15-20. Gustav Fischer Verlag, Jena, 1996.

[83] H. Ozer, S.L. George, C.A. Schiffer, K. Rao, P.N. Rao, D.H. Wurster-Hill, D.D. Arthur, B. Powell, A. Gottlieb, B.A. Peterson, K. Rai, J.R. Testa, M. LeBeau, R. Tantravahi, and C.D. Bloomfield. Prolonged subcutaneous administration of recombinant α2b interferon in patients with previously untreated Philadelphia chromosome-positive chronic-phase chronic myelogenous leukemia: effect on remission duration and survival: Cancer and Leukemia Group B Study 8583. *Blood*, 82:2975-2984, 1993.

[84] G. Pasternak, A. Hochhaus, B. Schultheis, and R. Hehlmann. Chronic myelogenous leukemia: molecular and cellular aspects *Journal of Cancer Research and Clinical Oncology*, 124:643-660, 1998.

[85] P.N. Peduzzi, K.M. Detre, Y.K.Chan, A. Oberman, and G.R. Cutter. Validation of a risk function to predict mortality in a VA population with coronary artery disease. *Controlled Clinical Trials*, 3:47-60, 1982.

[86] R. Peto, M.C. Pike, P. Armitage, N.E. Breslow, D.R. Cox, S.V. Howard, N. Mantel, K. McPherson, J. Peto, and P.G. Smith. Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *British Journal of Cancer*, 35:1-39, 1977.

[87] *Pschyrembel Klinisches Wörterbuch.* Walter de Gruyter, Berlin, 258. Auflage, 1998.

[88] M. Pfirrmann and J. Hasford. Methods for the evaluation and validation of a prognostic score in a data pool of individual data from patients with chronic myeloid leukaemia. In: E. Greiser und M. Wischnewsky (Hrsg.), *Methoden der Medizinischen Informatik, Biometrie und Epidemiologie in der modernen Informationsgesellschaft. Medizinische Informatik, Biometrie und Epidemiologie 83*, pp. 360-363. MMV Medien und Medizin Verlag, München, 1998.

[89] M. Pfirrmann and J. Hasford for the Collaborative CML Prognostic Factors Project (C.P.F.P.). Impact of time-dependent response variables on the prognosis of chronic myeloid leukemia. *Haematologica*, 85, The Jubilee Meeting on CML Supplement:22, 2000.

[90] M. Pfirrmann and J. Hasford. Testing Sokal's and the new prognostic score for chronic myeloid leukaemia treated with α-interferon: Comments. *British Journal of Haematology*, 114:241-242, 2001.

[91] M. Pfirrmann, J. Hasford. A simulation study using validated prognostic factors to assess expected long-term survival. *Methods of Information in Medicine*, 44:577-583, 2005.

[92] A. Reiter, R. Hehlmann, U. Berger, A. Hochhaus, M. Pfirrmann, J. Hasford, H. Heimpel, D.K. Hossfeld, C. Huber, H.-J. Kolb, H. Löffler, H. Pralle, W. Queisser, A. Gratwohl, A. Tobler, and the German CML Study Group. Randomized comparison of early allogeneic related stem cell transplantation vs. IFN-based therapy in newly diagnosed chronic myeloid leukemia (CML): results of the German CML study III. *Onkologie*, 25 (Suppl. 4):177, 2002.

[93] J.P. Royston. An extension of Shapiro and Wilk's W for normality to large samples. *Applied Statistics*, 31:115-124, 1982.

[94] B. Rüger. *Induktive Statistik. Einführung für Wirtschafts- und Sozialwissenschaftler.* Oldenbourg-Verlag, München, Wien, 2. Auflage 1988.

[95] L. Sachs. *Angewandte Statistik. Anwendung statistischer Methoden.* Springer-Verlag, Berlin, Heidelberg, 8. Auflage 1997.

[96] SAS Institute Inc. *SAS OnlineDoc.* SAS Institute Inc., Cary, NC, Version 8, 1999.

[97] P. Sasieni. Maximum weighted partial likelihood estimators for the Cox model. *Journal of the American Statistical Association*, 88:144-152, 1993.

[98] H.N. Sather. The use of prognostic factors in clinical trials. *Cancer*, 58:461-467, 1986.

[99] C.L. Sawyers. Chronic myeloid leukemia. *The New England Journal of Medicine*, 340:1330-1340, 1999.

[100] D.A. Schoenfeld and J.R. Richter. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics*, 38:163-170, 1982.

[101] S.S. Shapiro and M.B Wilk. An analysis of variance test for normality (complete samples). *Biometrika*, 52:591-611, 1965.

[102] R. Simon. Use of regression models: statistical aspects. In: M.E. Buyse, M.J. Staquet, R.J. Sylvester (eds.), *Cancer clinical trial. Methods and practice*, pp. 444-466. Oxford University Press, Oxford, 1984.

[103] R. Simon and D.G. Altman. Statistical aspects of prognostic factor studies in oncology. *British Journal of Cancer*, 69:979-985, 1994.

[104] R. Simon and R.W. Makuch. A non-parametric graphical representation of the relationship between survival and the occurence of an event: application to responder versus non-responder bias. *Statistics in Medicine*, 3:35-44, 1984.

[105] J.E. Sokal, E.B. Cox, M. Baccarani, S. Tura, G.A. Gomez, J.E. Robertson, C.Y. Tso, T.J. Braun, B.D. Clarkson, F. Cervantes, C. Rozman, and the Italian Cooperative CML Study Group. Prognostic discrimination in "good-risk" chronic granulocytic leukemia. *Blood*, 63:789-799, 1984.

[106] Southwest Oncology Group. *A phase II study of recombinant human interferon-alfa and recombinant human interferon-gamma in previously untreated patients with chronic myelogenous leukemia.* Southwest Oncology Group, San Antonio, Texas, SWOG-8735, 1988.

[107] J.L. Steegmann, J. Odriozola, F. Rodriguez-Salvans, P. Giraldo, J. Garca-Laraña, M.T. Ferro, E. Bentez, C. Prez-Pons, M. Giralt, L. Escribano, E. Lavilla, A. Miguel, C. Areal, M. Prez-Encinas, A. Abad, J. Maldonado, I. Massague, and J.M. Fernndez-Rañada. Stage, percentage of basophils at diagnosis, hematologic response within six months, cytogenetic response in the first year: the main prognostic variables affecting outcome in patients with chronic myeloid leukemia in chronic phase treated with interferon-α. Results of the CML89 trial of the Spanish Collaborative Group on interferon-α2a and CML. *Haematologica*, 84:978-987, 1999.

[108] R.M. Stone. Optimizing treatment of chronic myeloid leukemia: a rational approach. *The Oncologist*, 9:259-270, 2004.

[109] Süddeutsche Hämoblastosegruppe (SHG) e.V. *Prospektive kontrollierte Studie zur Therapieoptimierung bei chronischer myeloischer Leukämie (CML). Multizentrische Studie zur Prüfung von Interferon alpha vs. allogene Knochenmarktransplantation vs. Interferon alpha mit nachfolgender intensiver Chemotherapie und Interferonerhaltungstherapie während der frühen chronischen Phase. Kurzformel: CML-Studie III.* R. Hehlmann, III. Medizinische Klinik, Klinikum Mannheim (Hrsg.). Endfassung Januar 1995.

[110] Süddeutsche Hämoblastosegruppe (SHG) e.V. *Prospektiver kontrollierter Therapieoptimierungsvergleich bei chronischer myeloischer Leukämie (CML). Multizentrischer Vergleich von Interferon alpha mit low-dose-Ara-C vs. allogene Stammzell-Knochenmarktransplantation vs. Hochdosischemotherapie mit Autotransplantation und Interferon-Erhaltungstherapie*

während der frühen chronischen Phase. Kurzformel: CML-Studie III A. R. Hehlmann, III. Medizinische Klinik, Klinikum Mannheim (Hrsg.). Endfassung Oktober 1997.

[111] CML-Studiengruppe (CML-SG) *Qualitätssicherungsprotokoll zur Therapieoptimierung bei chronischer myeloischer Leukämie (CML). Randomisierter kontrollierter Vergleich von Imatinib vs. Imatinib und Interferon-α vs. Imatinib und niedrig dosiertes Ara-C vs. Imatinib nach Interferon-α-Versagen mit Prüfung des Stellenwertes der allogenen Stammzelltransplantation bei neu diagnostizierter CML in chronischer Phase. Kurzformel: CML-Studie IV. Pilotphase* R. Hehlmann, III. Medizinische Klinik, Klinikum Mannheim (Hrsg.). Fassung November 2003.

[112] CML-Studiengruppe (CML-SG) *Qualitätssicherungsprotokoll zur Therapieoptimierung bei chronischer myeloischer Leukämie (CML). Randomisierter kontrollierter Vergleich von Imatinib vs. Imatinib und Interferon-α vs. Imatinib 800 mg mit Prüfung des Stellenwertes der allogenen Stammzelltransplantation bei neu diagnostizierter CML in chronischer Phase. Kurzformel: CML-Studie IV.* R. Hehlmann, III. Medizinische Klinik, Klinikum Mannheim (Hrsg.). Fassung Mai 2005.

[113] M. Talpaz, H.M. Kantarjian, K.B. McCredie, M.J. Keating, J. Trujillo, and J. Gutterman. Clinical investigation of human alpha interferon in chronic myelogenous leukemia. *Blood*, 69:1280-1288, 1987.

[114] M. Talpaz, H. Kantarjian, R. Kurzrock, J.M. Trujillo, and J.U. Gutterman. Interferon-alpha produces sustained cytogenetic responses in chronic myelogenous leukemia. *Annals of Internal Medicine*, 114:532-538, 1991.

[115] J. Thaler, G. Gastl, T. Fluckinger, D. Niederwieser, H. Huber, H. Seewann, H. Silly, A. Lang, C. Abbrederis, H. Gadner, W. Fereberger, L. Schiller, L. Köck, M. Fridik, C. Duba, M. Falk, M. Berger, T. Kühr, and C. Huber for the Austrian Biological Response Modifier (BRM) Study Group. Treatment of chronic myelogenous leukemia with interferon Alfa-2c: response rate and toxicity in a phase II multicenter study. *Seminars in Hematology*, 30 (Suppl. 3):17-19, 1993.

[116] J. Thaler and W. Hilbe for the Austrian CML Study Group. Comparative analysis of two consecutive phase II studies with IFN-α and IFN-α + Ara-C in untreated chronic-phase CML patients. *Bone Marrow Transplantation*, 17 (Suppl. 3): S25-S28, 1996.

[117] M.J. Thomas, J.A.E. Irving, A.L. Lennard, S.J. Proctor, P.R.A. Taylor, on behalf of the Northern Region Haematology Group. Validation of the Hasford score in a demographic study in chronic granulocytic leukaemia. *Journal of Clinical Pathology*, 54:491-493, 2001.

[118] P.J.M. Verweij and H.C. van Houwelingen. Cross-validation in survival analysis. *Statistics in Medicine*, 12:2305-2314, 1993.

[119] M. Warmuth, S. Danhauser-Riedl, and M. Hallek. Molecular pathogenesis of chronic myeloid leukemia: implications for new therapeutic strategies. *Annals of Hematology*, 78:49-64, 1999.

[120] B. Weidmann und N. Niederle. Ergebnisse der Interferontherapie bei myeloproliferativen Syndromen. In: N. Niederle (Hrsg.), *Zytokine: Präklinik und Klinik*, pp. 70-95. Gustav Fischer Verlag, Jena, 1996.

[121] J.C. Wyatt and D.G. Altman. Commentary: Prognostic models: clinically useful or quickly forgotten? *British Medical Journal*, 311:1539-1541, 1995.

Lebenslauf

Name:		Markus Pfirrmann
Geburtsdatum:		19. April 1967
Geburtsort:		Landau in der Pfalz
Adresse:		Setzbergstraße 11, 81539 München
Schulbildung:	1973 - 1977	Grundschule Wollmesheimer Höhe Landau
	1977 - 1986	Otto-Hahn-Gymnasium Landau Abschluss: Abitur
Wehrdienst:	1986 - 1987	Wehrdienstleistender in Koblenz
Studium:	1987 - 1993	Studium der Statistik an der Ludwig-Maximilians-Universität München
	Mai 1993	Abschluss: Diplom
	1993 - 1994	Master-Studiengang am University College London
	November 1994	Abschluss: Master of Science in Statistics „Applied Stochastic Systems“
Beruf:	1995 - 1997	Wissenschaftlicher Mitarbeiter am Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie der Ludwig-Maximilians-Universität München
	1997 - 1998	Wissenschaftlicher Mitarbeiter beim Biometrischen Zentrum für Therapiestudien BZT in München
	1998 - 2000	Wissenschaftlicher Mitarbeiter am Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie der Ludwig-Maximilians-Universität München
	2000 - 2007	Wissenschaftlicher Mitarbeiter bei der Gesellschaft für Informationsverarbeitung und Statistik in der Medizin e.V. in München
	seit März 2007	Wissenschaftlicher Mitarbeiter am Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie der Ludwig-Maximilians-Universität München