Risikofaktoren für Vorhofflimmern und deren Einfluss auf die Entstehung eines Vorhofflimmerns: Eine Analyse mittels Cox-Regression

Bachelorarbeit zur Erlangung des akademischen Grades Bachelor of Science – Angewandte Mathematik

Sebastian Dohmen

Eingereicht im September 2020

Abstract

Diese Bachelorarbeit befasst sich mit der Identifikation von weiteren, bisher nicht oder nur wenig beachteten Risikofaktoren für Vorhofflimmern, indem bekannte Risikofaktoren, die überwiegend auf Grundlage von Primärdatenstudien erfasst wurden, anhand von Sekundärdaten in Form von Abrechnungsdaten der Gesetzlichen Krankenversicherung (GKV) evaluiert werden. Als methodischer Ansatz wurde hierfür eine Analyse mittels Cox-Regression durchgeführt. Zur Validierung der Ergebnisse erfolgte die Analyse anhand von zwei Modellen: den ICD-10-Codes und den hierarchisierten Morbiditätsgruppen des Risikostrukturausgleichs. Im Ergebnis zeigt sich, dass besonders bei kardiovaskulären Erkrankungen, aber auch bei anderen Risikofaktoren, weitere ähnliche Krankheitsbilder berücksichtigt werden sollten.

Keywords

Risikofaktoren · Vorhofflimmern · Sekundärdaten · Abrechnungsdaten · Cox-Regression · Überlebenszeitanalyse · Versorgungsforschung

Gliederung

Tabellenverzeichnis

Abkürzungsverzeichnis

BMG	Bundesministerium für Gesundheit
COPD	Chronische obstruktive Lungenkrankheit
DxG	Diagnosegruppe
FEV1	Forced Expiratory Pressure in 1 Second (Einsekundenkapazität)
GKV	Gesetzliche Krankenversicherung
HMG	hierarchisierte Morbiditätsgruppe
HR	Hazard Ratio
ICD-10	internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (10. Revision)
KI	Konfidenzintervall
Morbi-RSA	Morbiditätsorientierter Risikostrukturausgleich
OR	Odds Ratio
RR	Risk Ratio
STROSA 2	Standardisierte Berichtsroutine für Sekundärdaten Analysen, Version 2
VHF	Vorhofflimmern
WIG2	Wissenschaftliches Institut für Gesundheitsökonomie und Gesundheitssystemforschung

1 Einleitung

Vorhofflimmern ist eine Herzrhythmusstörung mit ungeordneter Tätigkeit der Herzvorhöfe und gilt als die in Europa meistverbreitete Herzrhythmusstörung. Es ist zu erwarten, bedingt durch die zunehmende Alterung der Gesellschaften, dass in Zukunft die Prävalenz weiter zunehmen wird, Schätzungen zufolge von 8,8 Millionen Erwachsenen (95 % Konfidenzintervall (KI): 6,5–12,3 Millionen) in 2010 bis auf 17,9 Millionen Erwachsene (95 % KI: 13,6–23,7 Millionen) in 2060 (Krijthe et al., 2013). Wenngleich Vorhofflimmern keine unmittelbaren Folgen für die Gesundheit und die Lebensqualität der Patient:innen hat, sorgt Vorhofflimmern für ein deutlich erhöhtes Risiko für Schlaganfälle und andere kardiovaskuläre Ereignisse. So haben Patient:innen mit Vorhofflimmern ein bis zu fünfmal höheres Risiko, ein Schlaganfallereignis zu erleiden (Wolf, Abbott & Kannel, 1991). Bei mittelalten Patient:innen mit Vorhofflimmern liegt die Wahrscheinlichkeit einer stationären Aufnahme oder des Todes aufgrund von kardiovaskulären Ereignissen in den folgenden 20 Jahren bei schätzungsweise 75 %, was einer 2- bis 3-fachen Erhöhung gegenüber Patient:innen ohne Vorhofflimmern entspricht (Stewart, Hart, Hole & McMurray, 2002). In vielen Fällen verläuft die Erkrankung symptomfrei, das heißt, die Patient:innen bemerken keine Veränderung des Herzrhythmus. Der Anteil der Patient:innen mit sogenanntem „stillen“ Vorhofflimmern an allen Patient:innen mit Vorhofflimmern liegt bei etwa 40 % (Passman & Bernstein, 2016). Die Identifikation möglicher Risikofaktoren und ein gezieltes Screening betroffener Patient:innen könnte diesen Anteil deutlich verringern und den Krankheitsverlauf positiv beeinflussen.

Die vorliegende Bachelorarbeit setzt an diesem Punkt an und soll einen Beitrag zur Abgrenzung möglicher weiterer Risikofaktoren leisten. Als thematische Einführung in das Thema Vorhofflimmern beschreibt Kapitel 2 das Krankheitsbild des Vorhofflimmerns und zeigt Möglichkeiten der Diagnostik und Therapieformen auf. Eine Betrachtung der epidemiologischen Situation in Deutschland verdeutlicht die Aktualität des Krankheitsbildes, da in einzelnen Studien Gesamtprävalenzen von bis zu 3,2 % festgestellt wurden (Schnabel, Wilde, Wild, Munzel & Blankenberg, 2012; Wilke et al., 2013). Weiterhin werden bereits bekannte Risikofaktoren aus der Leitlinie zur Therapie von Vorhofflimmern (Kirchhof et al., 2016) identifiziert. Von dieser Zusammenstellung ausgehend hat die Bachelorarbeit zwei zentrale Ziele:

Zunächst sollen die in der Leitlinie erfassten Risikofaktoren, die überwiegend auf Grundlage von Primärdatenstudien angeführt werden, anhand von Sekundärdaten in Form von Abrechnungsdaten der Gesetzlichen Krankenversicherung (GKV) evaluiert werden. In einem zweiten Schritt sollen im Rahmen eines explorativen Ansatzes bislang unberücksichtigte Risikofaktoren identifiziert werden.

Für beide Ansätze wird eine Cox-Regression als methodisches Werkzeug genutzt. Dieses Regressionsverfahren zählt zu den Methoden der Überlebenszeitanalyse (auch Ereigniszeitanalyse), in der die Zeit bis zum Eintreten eines interessierenden Ereignisses in verschiedenen Gruppen verglichen wird (Klein & Moeschberger, 2003). Die methodischen Hintergründe werden in Kapitel 3 dargelegt. Im speziellen Fall wird der Einfluss eines Sets von potenziellen Risikofaktoren auf den Eintritt eines Vorhofflimmerns als interessierendes Ereignis bewertet.

Zur Validierung der identifizierten Risikofaktoren werden diese in einem ersten Modell in die internationale statistische Klassifikation der Krankheiten und verwandter Gesundheitsprobleme (10. Revision; ICD-10) transformiert, um eine Zuordnung anhand von Diagnosedaten in den verwendeten Sekundärdaten zu ermöglichen. Diese Risikofaktoren werden anschließend als Kovariablen genutzt.

Ein zweites vergleichendes Modell nutzt das System der hierarchisierten Morbiditätsgruppen (HMG) des morbiditätsorientierten Risikostrukturausgleichs (Morbi-RSA). Dazu werden die Diagnosedaten von der ICD-10-Codierung in das System der HMGs übertragen. Dieses Modell wird schrittweise modifiziert, um signifikante Risikofaktoren zu finden und zu bewerten.

Auf diese Weise sollen weitere potenzielle Risikofaktoren identifiziert werden, die einen signifikanten Einfluss auf die Entwicklung eines Vorhofflimmerns haben, die bislang jedoch nicht oder nur unzureichend betrachtet werden. Kapitel 4 stellt einen detaillierten Analyseplan zur Beschreibung des Vorgehens auf. Ergebnis der Analyse sind dann die validierten Risikofaktoren, die in der Literatur bereits Erwähnung finden, ergänzt um weitere Risikofaktoren für ein Vorhofflimmern.

2 Vorhofflimmern

Für einen einführenden Überblick in das Thema Vorhofflimmern wird im ersten Unterkapitel zunächst Bezug auf die Klinik und die damit zusammenhängende Symptomatik des Vorhofflimmerns genommen. Eine kurze Darstellung der Behandlungssituation erfolgt anhand diagnostischer Methoden und möglicher Therapieansätze. Weiterhin soll die Betrachtung der Epidemiologie in Deutschland und der verschiedenen Risikofaktoren die Relevanz der Erkrankung beleuchten.

2.1 Symptomatik und Ursachen

Vorhofflimmern und Vorhofflattern sind Tachykardien im Bereich der Herzvorhöfe. Eine Tachykardie zeichnet sich durch eine kontinuierliche Beschleunigung des Herzschlages auf über 100 Schläge pro Minute aus (Lüderitz & Lewalter, 2010). Ursachen sowohl für Vorhofflimmern als auch für Vorhofflattern sind sogenannte kreisende Erregungen, deren genaue Entstehung im Folgenden beschrieben wird.

Das menschliche Herz besteht funktional betrachtet aus zwei getrennten Pumpensystemen. Das rechte und das linke Herz besteht jeweils aus einer Herzkammer und einem davor gelegenen Vorhof. Das rechte Herz transportiert venöses Blut in den Lungenkreislauf. Das linke Herz bringt mit Sauerstoff gesättigtes Blut vom Lungenkreislauf in den Körperkreislauf (Klinke, Pape, Kurtz & Silbernagl, 2010). Dabei stellt der Herzmuskel eine Funktionseinheit dar, die aus unterschiedlichen Zelltypen besteht und die elektrisch leitfähig und kontraktil ist. Es lassen sich im Wesentlichen Herzmuskelzellen und andere Zelltypen unterscheiden. Die Herzmuskelzellen sorgen für die elektrische Leitfähigkeit und die Kontraktionsfähigkeit des Gewebes. Demgegenüber bildet nicht-muskuläres Gewebe eine Barriere für die elektrische Erregungsleitung. Das Membranpotenzial einer Herzmuskelzelle wird von der Ionenkonzentration und dem Ladungsgradienten über der Zellmembran bestimmt, das Potenzial einer nichterregten Muskelzelle wird dementsprechend als Ruhemembranpotenzial bezeichnet. Wird eine Herzmuskelzelle durch einen adäquaten Reiz erregt, so ist das Aktionspotenzial die elektrophysiologische Antwort darauf. Dabei tritt eine Erregung durch eine Absenkung des Membranpotenzials unter das Schwellenpotenzial ein und löst das Aktionspotenzial aus. Zunächst steigt das Aktionspotenzial innerhalb von Millisekunden an, bleibt dann für Bruchteile von Sekunden auf einem hohen Wert und sinkt anschließend in der Refraktärzeit erneut auf das Ruhemembranpotenzial ab. Auf dem Plateau des Aktionspotenzials und am Beginn der Repolarisation ist die Muskelzelle absolut unempfindlich gegenüber Erregungen. Diese sogenannte Refraktärzeit unterteilt sich in eine relative Refraktärzeit, in der zur Erregung deutlich erhöhte Schwellenreize nötig sind, und in eine supernormale Phase, in der zur Erregung leicht erhöhte Schwellenreize nötig sind. Auf diese Weise wird eine unmittelbare Wiedererregung einer Muskelzelle verhindert und eine gerichtete Erregungsleitung ermöglicht (Lüderitz & Lewalter, 2010).

In Verbindung damit verlaufen Reizbildung und Erregungsleitung im Herzen bedingt durch die verschiedenen elektrophysiologischen Eigenschaften der teilnehmenden Strukturen nach einem bestimmten zeitlichen und räumlichen Muster. Eine Wiederregung von Muskelzellen durch die gleiche Erregungswelle wird verhindert, indem die Refraktärzeit länger als die Ausbreitungszeit der Erregung ist. Ist nun zum einen die Erregungsleitung unidirektional blockiert und dadurch verzögert, kommt es zu einer kreisenden Erregung (engl. Reentry). Hierbei erfolgt die Fortleitung der Erregung über eine alternative Bahn und macht so die Wiedererregung bereits erregter Areale möglich, die nicht refraktär sind. Ist die Wellenlänge der Erregung kürzer als die Kreisbahn, so erhält sich die kreisende Erregung.

Im Fall von Vorhofflimmern treten mehrere solcher kreisenden Erregungen im Bereich der Vorhöfe auf, sogenannte Mikro-Reentry-Erregungen. Andernfalls können auch Makro-Reentry-Mechanismen

ursächlich sein, dabei kreist eine einzige größere kreisende Erregung um ein funktionales oder anatomisches Hindernis. Auch sogenannte fokale Impulsbildungen können Vorhofflimmern verursachen, indem sie in den Depolarisations- und Repolarisationsvorgängen innerhalb der Zellmembran eine kreisende Erregung auslösen. Die Ursache für Vorhofflattern sind Makro-Reentry-Mechanismen (Lüderitz & Lewalter, 2010).

2.2 Diagnostik und Therapie

In der Diagnostik wird zwischen paroxysmalem, persistierendem und permanentem Vorhofflimmern unterschieden. Die Differenzierung erfolgt anhand der Dauer einzelner Episoden. Das paroxysmale Vorhofflimmern ist selbstterminierend und dauert in den meisten Fällen nicht länger als 48 Stunden. Einzelne Episoden können bis zu sieben Tage andauern. Persistierendes Vorhofflimmern bezeichnet ein länger als sieben Tage anhaltendes Vorhofflimmern. Bei permanentem Vorhofflimmern liegt das Vorhofflimmern ununterbrochen vor (Kirchhof et al., 2016).

Die Diagnose Vorhofflattern wird in ein isthmusabhängiges, typisches und ein nicht-isthmusabhängiges, atypisches Vorhofflattern kategorisiert. Dabei wird differenziert, ob das Vorhofflattern verursachende Makro-Reentry den cavotrikuspidalen Isthmus, ein spezifisches Herzareal, durchzieht (Lüderitz & Lewalter, 2010).

Die Diagnose von Vorhofflimmern und Vorhofflattern erfolgt über EKG-Aufzeichnungen. Eine Episode gilt als diagnostiziert, wenn sie länger als 30 Sekunden andauert. Besonders Episoden paroxysmalen Vorhofflimmerns werden so in vielen Fällen nicht erfasst. Liegen asymptomatische Episoden vor, wird von sogenanntem „stillen“ Vorhofflimmern gesprochen, welches nicht entdeckt wird. Solche Vorhofflimmer-Erkrankungen lassen sich nur durch regelmäßiges Screening diagnostizieren (Kirchhof et al., 2016).

Die Wiederherstellung des normalen Herzrhythmus wird als Kardioversion bezeichnet. Es werden die pharmakologische, also auf Medikamenten basierende, und die elektrische Kardioversion, ausgelöst durch gezielte Stromstöße, unterschieden. Eine dauerhafte Stabilisierung des Herzrhythmus kann außerdem durch einen operativen Eingriff, die Katheterablation, erfolgen. Dabei werden Herzareale, die das Vorhofflimmern auslösen, elektrisch isoliert und sozusagen stillgelegt. Da ein Vorhofflimmern die Gerinnselbildung des Blutes in den Vorhöfen begünstigt, ist weiterhin eine Antikoagulation sinnvoll. Dabei handelt es sich um den Einsatz von Blutverdünnern, die ebendiese Gerinnselbildung verhindern sollen. Hierbei spielt auch die Schlaganfallprävention eine entscheidende Rolle, da Gerinnsel Schlaganfälle auslösen, wenn sie zum Gehirn wandern und dort eine Arterie verstopfen (Lüderitz & Lewalter, 2010).

2.3 Epidemiologie in Deutschland

An dieser Stelle soll eine kurze Betrachtung der Epidemiologie von Vorhofflimmern stehen. Vorhofflimmern ist eine der am häufigsten auftretenden Herzrhythmusstörungen. Im Rahmen der Gutenberg-Gesundheitsstudie wurde eine Zufallsstichprobe aus der Bevölkerung der Region Mainz-Bingen auf kardiovaskuläre Erkrankungen untersucht. Bei 3,2 % von den 5.000 teilnehmenden Patient:innen wurde ein Vorhofflimmern diagnostiziert. Die Prävalenz betrug bei Männern 4,6 % gegenüber 1,9 % bei Frauen. Stratifiziert nach Altersgruppen nahm die Prävalenz mit dem Alter nicht-linear zu. So betrug die Prävalenz bei Männern in der Altersgruppe der 35- bis 44-Jährigen 0,7 % und stieg auf 10,6 % in der Altersgruppe der 65- bis 74-Jährigen. Bei Frauen ist eine ähnliche Zunahme zu beobachten. Hier stieg die Prävalenz in den beiden Altersgruppen von 0,3 % auf 4,9 % (Schnabel et al., 2012).

In einer Sekundärdatenstudie, beruhend auf Abrechnungsdaten zweier Krankenkassen der GKV in Deutschland mit zusammen 8,3 Millionen Versicherten, wurde die Population auf Prävalenz und Inzidenz im Jahr 2008 untersucht. Dabei wurde eine Gesamtprävalenz von 2,1 % festgestellt. Die Prävalenz in den Altersgruppen stieg bei Männern von 0,2 % in der Altersgruppe der 35- bis 39-Jährigen auf 9,1 % in der Altersgruppe der 70- bis 74-Jährigen. Bei Frauen stieg die Prävalenz in diesen Altersgruppen von 0,1 % auf 6,0 %. Weiterhin wurden auch ältere Altersgruppen analysiert. Dabei stieg die Prävalenz bei Männern mit dem Alter weiter bis auf 17,8 % in der Altersgruppe der 85- bis 89-Jährigen und sinkt dann auf 16,5 % bei den über 89-Jährigen. Bei den Frauen stieg die Prävalenz zunächst auf 14,0 % und sank dann auf 11,8 % (Wilke et al., 2013).

Die Unterschiede zwischen beiden Studien sind mit großer Wahrscheinlichkeit auf die Stichprobenbildung zurückzuführen. Die auf die Bevölkerung gewichtete Prävalenzrate der Gutenberg-Gesundheitsstudie beträgt 2,5 % und weicht so nur um 0,4 % von der entsprechenden Rate der Sekundärdatenstudie ab. Bedingt durch die Alterung der Gesellschaft und dem damit verbundenen Anstieg der Morbiditätslast ist eine weitere Zunahme der Prävalenz von Vorhofflimmern zu erwarten. In einer niederländischen Studie wurden die Zahlen der Prävalenz von Vorhofflimmern einer prospektiven Kohortenstudie mit Baseline-Periode in den Jahren 2000 und 2001 zur Schätzung der Entwicklung der Zahlen in der EU bis 2060 genutzt. Die Schätzungen für 2010 belaufen sich auf 8,8 Millionen Erwachsene (95 % KI: 6,5–12,3 Millionen), und prognostizieren einen Anstieg der Prävalenz bis auf 17,9 Millionen Erwachsene (95 % KI: 13,6–23,7 Millionen) im Jahr 2060. Dies entspricht Prävalenzraten von 1,8 % bzw. 3,5 % der Gesamtbevölkerung (Krijthe et al., 2013).

2.4 Risikofaktoren

Eine Vielzahl von Risikofaktoren ist mit der Entstehung eines Vorhofflimmerns verbunden. Die vorliegende Arbeit orientiert sich bei der Identifizierung der Risikofaktoren an der Behandlungsleitlinie für Vorhofflimmern der European Society of Cardiology (Kirchhof et al., 2016). Die dort genannten Risikofaktoren sind in **Tabelle 1** zusammengefasst. Dabei wird jeweils mittels Hazard Ratios (HR), Odds Ratios (OR) und Risk Ratios (RR) der statistische Zusammenhang zwischen den Risikofaktoren und dem Auftreten eines Vorhofflimmerns dargestellt. Es werden überwiegend Primärdatenstudien oder Metadatenstudien referenziert, die die Risikofaktoren nicht gesamthaft, sondern einzeln oder nur in Subgruppen betrachten. Diese Vorgehensweise hat den Nachteil, dass die komplette Menge der Risikofaktoren nicht als Ganzes bewertet wird. Die vorliegenden Werte beziehen sich, wenn nichts anderes spezifiziert ist, auf den Referenzfall, dass der betreffende Risikofaktor nicht vorliegt.

Zu den Risikofaktoren zählen genetische Prädispositionen und ein hohes Alter. Hierbei zeigt sich, dass ein erhöhtes Alter einen starken Einfluss auf die Entwicklung eines Vorhofflimmerns hat. So liegt schon in der Gruppe der 60- bis 69-Jährigen das Risiko eines Vorhofflimmerns fünfmal höher als in der Referenzgruppe der 50- bis 59-Jährigen. Dieses Risiko steigt bei den 80- bis 90-Jährigen bis auf das 9-fache gegenüber der Referenzgruppe an.

Herz-Kreislauf-Erkrankungen sind ebenfalls häufig mit der Entstehung eines Vorhofflimmerns assoziiert. Bei Bluthochdruckpatient:innen liegt das Risiko bei 32 %, bei Patient:innen mit diagnostizierter Herzinsuffizienz bei 43 %, bei Patient:innen mit stattgehabtem Herzinfarkt bei 46 % und bei Patient:innen mit einer Herzklappenerkrankung sogar um das 1,4-fache höher als bei Patient:innen ohne eine Diagnose der jeweiligen Komorbidität.

Weiterhin sorgen bestimmte endokrine Erkrankungen, Ernährungs- und Stoffwechselkrankheiten für ein erhöhtes Risiko, ein Vorhofflimmern zu entwickeln. Funktionsstörungen der Schilddrüse sind je nach Art der Störung mit unterschiedlichem Risiko behaftet. Eine Schilddrüsenunterfunktion (Hypothyreose) ist mit einem um 23 % erhöhten Risiko gegenüber einer normalen Schilddrüsenfunktion verbunden. Eine subklinische Überfunktion der Schilddrüse steigert das Risiko um 31 % und eine offene Überfunktion um 42 %. Übergewichtige sowie an Adipositas leidende Menschen weisen gegenüber normalgewichtigen Menschen ein um 13 % und 37 % höheres Risiko auf. Ein diagnostizierter Diabetes mellitus ist mit einem 25 % höheren Risiko einer Diagnose Vorhofflimmern assoziiert.

Patient:innen mit chronischer obstruktiver Lungenkrankheit (COPD) werden nach Forced Expiratory Pressure in 1 Second (Einsekundenkapazität; FEV1) differenziert. Dabei wird ein FEV1-Wert von über 80 % als Referenz genutzt. Liegt der FEV1-Wert nur zwischen 60 % und 80 %, liegt das Risiko bereits um 28 % höher. Ein FEV1-Wert von unter 60 % geht sogar mit einem 2,5-fachen Risiko für Vorhofflimmern einher.

Ebenso wirkt sich das Rauchen auf das Risiko eines Vorhofflimmerns aus. Patient:innen, die niemals geraucht haben, werden hier als Referenzwert betrachtet. Relativ dazu haben Ex-Rauchende ein um 32 % höheres Risiko und aktiv Rauchende sogar ein zweimal so hohes Risiko gegenüber Nichtrauchenden.

Patient:innen mit obstruktiver Schlafapnoe haben ein 2,2-mal so großes Risiko wie Patient:innen ohne diese Diagnose.

Bei chronisch Nierenerkrankten unterscheidet sich das Risiko je nach Stadium. Die Stadien werden dabei nach der glomerulären Filtrationsrate, einem Maß für die Filtrationsfähigkeit der Niere, spezifiziert. Die Stadien 1 und 2 bezeichnen dabei eine Rate von über 60 ml/min/1,73 m^2 Körperoberfläche und erhöhen das Risiko gegenüber Patient:innen ohne chronische Nierenerkrankung um das 2,67-fache. Patient:innen des dritten Stadiums, mit einer glomerulären Filtrationsrate von 30 bis unter 60 ml/min/1,73 m^2 Körperoberfläche, haben dagegen nur ein um 68 % erhöhtes Risiko. In den Stadien 4 und 5, hier liegt die Rate nur bei 30 ml/min/1,73 m^2 Körperoberfläche, bzw. 15 ml/min/1,73 m^2 Körperoberfläche, liegt eine präterminale bzw. terminale Niereninsuffizienz vor. Hier ist das Risiko um das 3,52-fache gegenüber Patient:innen ohne chronische Nierenerkrankung gesteigert.

Der Konsum von Alkohol steigert das Risiko eines Vorhofflimmerns gegenüber Patient:innen ohne Alkoholkonsum um 1 % (1–6 Getränke; je 12 g Alkohol) bis 39 % (> 21 Getränke; je 12 g Alkohol).

Bei der regelmäßigen Ausübung von Kraftsport senkt ein Trainingspensum von weniger als einem Tag pro Woche das Risiko gegenüber keinem Training um 10 %. Eine höhere Intensität sorgt für ein erhöhtes Risiko um bis zu 20 % bei 5 bis 7 Trainingstagen in der Woche.

Risikofaktor	Differenzierung	Assoziation mit VHF
Genetische Prädispositionen		HR range 0,4–3,2
Hohes Lebensalter	50–59 Jahre	HR 1,00 (Referenz)
	60–69 Jahre	HR 4,98 (95 % KI 3,49–7,10)
	70–79 Jahre	HR 7,35 (95 % KI 5,28–10,20)
	80–89 Jahre	HR 9,33 (95 % KI 6,68–13,00)
Bluthochdruck		HR 1,32 (95 % KI 1,08–1,60)
Herzinsuffizienz		HR 1,43 (95 % KI 0,85–2,40)
Herzklappenerkrankungen		RR 2,42 (95 % KI 1,62–3,60)
Herzinfarkt		HR 1,46 (95 % KI 1,07–1,98)
Funktionsstörungen der Schilddrüse	Hypothyreose	HR 1,23 (95 % KI 0,77–1,97)
	Subklinische Hypothyreose	RR 1,31 (95 % KI 1,19–1,44)
	Offene Hypothyreose	RR 1,42 (95 % KI 1,22–1,63)
Adipositas	Nicht vorhanden (BMI < 25 kg/m²)	HR 1,00 (Referenz)
	Übergewicht (BMI 25–30 kg/m²)	HR 1,13 (95 % KI 0,87–1,46)
	Fettleibigkeit (BMI ≥ 31 kg/m²)	HR 1,37 (95 % KI 1,05–1,78)
Diabetes mellitus		HR 1,25 (95 % KI 0,98–1,60)
COPD	FEV1 ≥ 80 %	RR 1,00 (Referenz)
	FEV1 60–80 %	RR 1,28 (95 % KI 0,79–2,06)
	FEV1 < 60 %	RR 2,53 (95 % KI 1,45–4,42)
Obstruktive Schlafapnoe		HR 2,18 (95 % KI 1,34–3,54)
Chronische Nierenkrankheit	Nicht vorhanden	OR 1,00 (Referenz)
	Stadium 1 oder 2	OR 2,67 (95 % KI 2,04–3,48)
	Stadium 3	OR 1,68 (95 % KI 1,26–2,24)
	Stadium 4 oder 5	OR 3,52 (95 % KI 1,73–7,15
Rauchen	Niemals geraucht	HR 1,00 (Referenz)
	Früher geraucht	HR 1,32 (95 % KI 1,10–1,57)
	Aktuell Rauchende:r	HR 2,05 (95 % KI 1,71 –2,47)
Alkoholkonsum	Kein Konsum	RR 1,00 (Referenz)
	1–6 Getränke (je 12 g Alkohol) pro Woche	RR 1,01 (95 % KI 0,94–1,09)
	7–14 Getränke (je 12 g Alkohol) pro Woche	RR 1,07 (95 % KI 0,98–1,17)
	15–21 Getränke (je 12 g Alkohol) pro Woche	RR 1,14 (95 % KI 1,01–1,28)
	> 21 Getränke (je 12 g Alkohol) pro Woche	RR 1,39 (95 % KI 1,22–1,58)
Regelmäßiger Kraftsport	Kein Training	RR 1,00 (Referenz)
	< 1 Tage pro Woche	RR 0,90 (95 % KI 0,68–1,20)
	1–2 Tage pro Woche	RR 1,09 (95 % KI 0,95–1,26)
	3–4 Tage pro Woche	RR 1,04 (95 % KI 0,91–1,19)
	5–7 Tage pro Woche	RR 1,20 (95 % KI 1,02–1,41)

***Tabelle 1**: Mit der Entwicklung eines Vorhofflimmerns assoziierte Komorbiditäten*[1]

[1] Quelle: Eigene Darstellung auf Basis von Kirchhof et al. (2016). Legende: VHF=Vorhofflimmern; BMI=Body Mass Index.

3 Die Cox-Regression

3.1 Grundlagen der Überlebenszeitanalyse

Bevor im dritten Kapitel als zentraler Aspekt des theoretischen Teils der Bachelorarbeit die Cox-Regression eingeführt wird, werden zunächst einige grundlegende Definitionen und Funktionen der Überlebenszeitanalyse vorgestellt. Dabei ist zu beachten, dass in der Überlebenszeitanalyse die Begriffe *Überlebenszeit* und *Todeszeitpunkt* definiert sind. Handelt es sich bei der Zielgröße nicht um den Tod eines Individuums, wird dementsprechend von der Zeit bis zum Eintreten des Ereignisses und analog dazu von Ereigniszeitpunkten gesprochen. Vereinfachend werden im folgenden Abschnitt dennoch die in der Literatur üblichen Begriffe der Überlebenszeit und des Todeszeitpunktes verwendet.

3.1.1 Überlebensfunktion und Hazard-Funktion

Folgende Beschreibungen beziehen sich auf Collett (2015). In der Überlebenszeitanalyse sind zwei Funktionen von grundlegender Bedeutung. Das ist zum einen die Überlebensfunktion. Die Überlebenszeit eines Individuums t wird durch die Zufallsvariable T dargestellt. Die Verteilungsfunktion von T ist folgendermaßen in Abhängigkeit der Wahrscheinlichkeitsdichtefunktion $f(t)$ definiert:

$$F(t) = P(T < t) = \int_0^t f(u)du \tag{3.1}$$

Es wird die Wahrscheinlichkeit dargestellt, dass T einen Wert kleiner t annimmt. Das heißt, dass das Individuum vor dem Zeitpunkt t stirbt. Dementsprechend lässt sich die Überlebensfunktion als die Wahrscheinlichkeit darstellen, dass T größer gleich t ist:

$$S(t) = P(T \geq t) = 1 - F(t) \tag{3.2}$$

Die zweite wichtige Funktion ist die Hazard-Funktion. Diese Funktion repräsentiert die Wahrscheinlichkeit, dass ein Individuum zum Zeitpunkt t stirbt, unter der Bedingung, dass es bis zu diesem Zeitpunkt t überlebt hat:

$$h(t) = \lim_{\delta t \to 0} \frac{P(t \leq T < t + \delta t | T \geq t)}{\delta t} \tag{3.3}$$

Sie beschreibt die Wahrscheinlichkeit, dass die Überlebenszeit T in einem Intervall $t \leq T < t + \delta t$ liegt. Der Grenzwert $\delta t \to 0$ beschreibt dann die Wahrscheinlichkeit, dass die Überlebenszeit t exakt T entspricht.

Nach dem Satz über bedingte Wahrscheinlichkeiten lässt sich die bedingte Wahrscheinlichkeit aus Gleichung (3.3) durch die Verteilungsfunktion $F(t)$ und die Überlebensfunktion $S(t)$ darstellen:

$$P(t \leq T < t + \delta t | T \geq t) = \frac{P(t \leq T < t + \delta t)}{P(T \geq t)} = \frac{F(t + \delta t) - F(t)}{S(t)} \tag{3.4}$$

Der erste Teil der Gleichung entspricht dabei genau der Ableitung der Verteilungsfunktion, also der Dichtefunktion $f(t)$, an der Stelle t. Für die Hazard-Funktion ergibt sich dann:

$$h(t) = \lim_{\delta t \to 0} \frac{F(t + \delta t) - F(t)}{\delta t} \frac{1}{S(t)} = \frac{f(t)}{S(t)} \tag{3.5}$$

Die Gleichung (3.2) liefert durch Umstellen $F(t) = 1 - S(t)$ und damit entsprechend $f(t) = -s(t)$, wobei $s(t)$ der Ableitung der Funktion $S(t)$ an der Stelle t entspricht. Diese Form ist identisch mit der negativen Ableitung der logarithmierten Überlebenszeitfunktion $S(t)$ an der Stelle t:

$$h(t) = \frac{-S'(t)}{S(t)} = -\frac{d}{dt}\big(logS(t)\big) \tag{3.6}$$

Integration nach t führt dann wiederum zu der Darstellung der kumulierten Hazard-Funktion:

$$H(t) = -logS(t) \tag{3.7}$$

3.1.2 Zensierte Überlebenszeiten

In vielen Fällen sind Überlebenszeiten zensiert, das heißt der Todeszeitpunkt für ein Individuum ist nicht bekannt. Dabei wird zwischen rechter und linker Zensierung sowie Intervall-Zensierung unterschieden. Rechte Zensierung liegt vor, wenn ein Individuum bei t_0 in die Studie aufgenommen wurde und der Todeszeitpunkt nicht bekannt ist. In diesem Fall ist $t_0 + c$ der letzte bekannte Überlebenszeitpunkt und c bezeichnet die zensierte Überlebenszeit.

In der Analyse der Bachelorarbeit wird nur der Fall der rechten Zensierung auftreten. Der Vollständigkeit halber werden die beiden anderen Arten der Zensierung an dieser Stelle kurz definiert: Linke Zensierung beschreibt den Fall einer kürzeren Überlebenszeit als der beobachteten Überlebenszeit. Eine solche Situation kann bei der Betrachtung von Todeszeitpunkten natürlich nicht auftreten. Als Beispiel lässt sich die Betrachtung von Krankheiten nennen. Dabei liegt der Todeszeitpunkt, bzw. Ereigniszeitpunkt, vor dem ersten Beobachtungszeitpunkt und lässt sich nicht genau feststellen. Bei der Intervall-Zensierung liegt der unbekannte, exakte Todeszeitpunkt im Intervall zwischen zwei Beobachtungszeitpunkten. Es handelt sich um das gleichzeitige Auftreten einer rechten und linken Zensierung (Collett, 2015).

3.1.3 Kaplan-Meier-Schätzer

Folgende Beschreibungen beziehen sich auf Collett (2015). Enthält ein Datensatz solche, im vorhergehenden Abschnitt definierte, zensierte Überlebenszeiten, bietet sich der Kaplan-Meier-Schätzer zur Schätzung der Überlebensfunktion an.

Gegeben seien n Individuen mit beobachteten Überlebenszeiten $t_1, t_2, \dots, t_n$, wobei $r \leq n$ verschiedene Todeszeitpunkte auftreten können, da rechte Zensierung möglich ist, der Todeszeitpunkt also unbekannt ist, oder mehrere Individuen dieselbe Überlebenszeit aufweisen können.

Weiterhin gegeben sind

- die in aufsteigender Reihenfolge geordneten Überlebenszeiten $t_j, j = 1,2, \dots, r$
- die Zahl der überlebenden Individuen bis unmittelbar vor t_j n_j
- und die Anzahl der am Zeitpunkt t_j verstorbenen Individuen $d_j, j = 1,2, \dots, r$

Es wird angenommen, dass im Falle von gleichzeitigem Auftreten eines Todeszeitpunktes und einer zensierten Überlebenszeit die zensierte Überlebenszeit unmittelbar nach dem Todeszeitpunkt auftrat. Daraus ergibt sich

- eine geschätzte Sterbewahrscheinlichkeit von $\frac{d_j}{n_j}$
- und dementsprechend eine Überlebenswahrscheinlichkeit von $\frac{n_j - d_j}{n_j}$
- im Zeitraum $\lim_{\delta \to 0}\big(t_{(j)} - \delta, t_{(j)}\big)$

Durch das Ordnen der Überlebenszeiten lässt sich folgern, dass es keine Tode im Intervall $\left[t_{(j)}, t_{(j+1)}\right)$ gibt und somit die Überlebenswahrscheinlichkeit gleich 1 ist. Damit beträgt die Wahrscheinlichkeit von $t_{(j)}$ bis $t_{(j+1)}$ zu überleben ebenfalls $\frac{n_j - d_j}{n_j}$.

Unter der Annahme, dass alle Todesfälle des Datensatzes statistisch unabhängig sind, lässt sich die Schätzung der Wahrscheinlichkeit über $t_{(k)}$ hinaus zu überleben, durch den Kaplan-Meier-Schätzer angeben:

$$\hat{S}(t) = \prod_{j=1}^{k} \left(\frac{n_j - d_j}{n_j}\right), t_{(k)} \leq t < t_{(k+1)}, k = 1,2,\dots,r \tag{3.8}$$

Per Definition gilt dabei für alle $t < t_{(1)} \hat{S}(t) = 1$, da vor dem ersten Todeszeitpunkt keine Todesfälle auftreten. Weiterhin ist zu bemerken, dass im Fall einer Zensierung der größten Überlebenszeit t^* $\hat{S}$ für $t > t^*$ undefiniert ist. Andernfalls gilt $\hat{S} = 0$ bei $t \geq t_{(r)}$, wenn die größte Überlebenszeit $t_{(r)}$ unzensiert ist.

Da Tode im Datensatz nur zu den Zeitpunkten $t_{(j)}, j = 1,2,\dots,r$ auftreten, wird die Hazard-Funktion zwischen zwei aufeinander folgenden Todeszeiten als konstant angenommen. Der geschätzte Hazard pro Zeiteinheit wird so durch eine Division des Zeitintervalls $\tau_j = t_{(j+1)} - t_{(j)}$ zwischen zwei Todeszeiten berechnet:

$$\hat{h}(t) = \frac{d_j}{n_j \tau_j}, t_{(j)} \leq t < t_{(j+1)}, \tag{3.9}$$

3.2 Cox-Regressionsmodell

3.2.1 Definition Cox-Regressionsmodell

Das „Proportional Hazards Model“, in der vorliegenden Arbeit kurz Cox-Regressionsmodel bezeichnet, wurde 1972 von Cox vorgestellt (Cox, 1972). Mithilfe dieses Modells lässt sich der Einfluss mehrerer Variablen auf die Hazard-Funktion $h(t)$ analysieren. Gegeben seien p erklärende Variablen $x_{1i}, x_{2i}, \dots, x_{pi}, i = 1,2,\dots,n$, deren Werte zum Zeitpunkt t_0 für jedes der n Individuen feststehen. Der Einfluss dieser Variablen wird über die Zeit als konstant angenommen. Die Funktion $h_0(t)$ beschreibt die Wahrscheinlichkeit, zum Zeitpunkt t zu sterben, im Fall, das alle erklärenden Variablen Null sind. Sie wird deshalb auch als Baseline-Hazard bezeichnet. Im Modell wird der Hazard für das Individuum i zum Zeitpunkt t durch folgende Gleichung definiert:

$$h_i(t) = exp\left(\beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{pi}\right) h_0(t) \tag{3.10}$$

Dabei stellen $\beta_1, \beta_2, \dots, \beta_p$ die Koeffizienten der erklärenden Variablen dar, deren Schätzung das Ziel ist. Die lineare Komponente, oder auch Risk Score, $\eta = \sum_{j=1}^{p} \beta_j x_{ji}$ bestimmt dabei das Verhältnis der Hazard-Funktion eines Individuums i zum Baseline-Hazard. Daraus folgt im Fall $\eta > 0$: $h_i(t) > h_0(t)$ und im Fall $\eta < 0$: $h_i(t) < h_0(t)$, aufgrund der Eigenschaften der Exponentialfunktion.

3.2.2 Partielle Likelihood-Funktion des Modells

Folgender Abschnitt wurde unter Hinzunahme von Klein und Moeschberger (2003) erstellt. Bei der Konstruktion der Likelihood-Funktion ist die Annahme, dass die Intervalle zwischen zwei aufeinander-

folgenden Todeszeitpunkten keine Information über den Einfluss der Kovariablen auf die Hazard-Funktion liefern, von grundlegender Bedeutung. Daraus folgt direkt, dass sowohl der Baseline-Hazard $h_0(t)$ als auch die Hazard-Funktion $h(t)$ Null in den Intervallen ohne Todesfälle sind. So ist also die Wahrscheinlichkeit, dass ein Individuum i zu einem Zeitpunkt $t_{(j)}$ stirbt, abhängig davon, dass $t_{(j)}$ zu den r beobachteten Todeszeitpunkten $t_{(1)}, t_{(2)}, \dots, t_{(r)}$ gehört. Grundsätzlich gilt im Cox-Regressionsmodell die Annahme einer kontinuierlichen Hazard-Funktion, sodass mehrere Todesfälle an einem Zeitpunkt nicht möglich sind. Daraus folgt die bedingte Wahrscheinlichkeit für den Tod des Individuums i zu einem Zeitpunkt $t_{(j)}$: (3.11)

$$\frac{h_i(t_{(j)})}{\sum_{l \in R(t_{(j)})} h_l(t_{(j)})}$$

Dabei beschreibt $R(t_{(j)})$ das Riskset zum Zeitpunkt $t_{(j)}$, d. h. alle Individuen, die bis unmittelbar vor den Zeitpunkt j überlebt haben.

Durch Einsetzen der Gleichung (3.10) kürzt sich der Baseline-Hazard und das Produkt dieser bedingten Wahrscheinlichkeiten über die r distinkten Todeszeitpunkte und bildet die Likelihood-Funktion des Cox-Regressionsmodell:

$$L(\beta) = \prod_{j=1}^{r} \frac{exp(\beta' x_{(j)})}{\sum_{l \in R(t_{(j)})} exp(\beta' x_l)} \tag{3.12}$$

Diese Funktion wird auch als partielle Likelihood-Funktion bezeichnet, da sie die zensierten und unzensierten Überlebenszeiten nicht direkt einbezieht.

Die Darstellung der partiellen Likelihood gilt so aber nur, wie oben erwähnt, unter der Annahme einer kontinuierlichen Hazard-Funktion. Um die Möglichkeit zuzulassen, dass Überlebenszeiten nicht distinkt sind, muss das Modell ergänzt werden. Diese Situation tritt sehr häufig auf, da Todeszeiten für gewöhnlich nur kategorisiert dokumentiert werden. Der Fall, dass Todeszeiten und zensierte Überlebenszeiten zusammen auftreten, lässt sich analog zum Vorgehen bei der Kaplan-Meier-Schätzung (Abschnitt 3.1.3) lösen. Dabei werden die Zensierungen nach allen Todesfällen durchgeführt, was zu einem eindeutig definierten Risk Set zum Zeitpunkt $t_{(j)}$ führt. Dieses Riskset enthält alle aufgetretenen Tode. So verbleiben nur Bindungen zwischen mehreren Todesfällen zum Zeitpunkt $t_{(j)}$. Die exakte Form für eine solche partielle Likelihood-Funktion existiert, es wird davon ausgegangen, dass Bindungen nur durch ungenaue Messungen entstehen und dass tatsächlich eine kontinuierliche Folge der Todeszeitpunkte vorliegt (Kalbfleisch & Prentice, 2002). Dabei werden alle möglichen Kombinationen von unabhängigen Todeszeitpunkten berechnet, um die Bindungen aufzulösen. Cox selbst schlug ein Vorgehen vor, um Bindungen innerhalb des Modells aufzulösen (Cox, 1972). Er verfolgt die Grundannahme, dass die Hazard-Funktion $h_i(t)$ nicht konitnuierlich, sondern diskret ist. Hier handelt es sich um ein logistisches Modell, indem das Cox-Regressionsmodell (Gleichung 3.10) angepasst wird. Es beschreibt die bedingte Wahrscheinlichkeit eines Individuums, im Intervall $(t, t+1)$ zu versterben, unter der Bedingung, bis t überlebt zu haben:

$$\frac{h_i(t)}{1 - h_i(t)} = \frac{h_0(t)}{1 - h_0(t)} exp(\beta' x_i) \tag{3.13}$$

Durch Einsetzen der Hazard-Funktion in Formel (3.11) analog zum kontinuierlichen Modell, ergibt sich die Likelihood-Funktion des diskreten Modells:

$$\prod_{j=1}^{D} \frac{exp(\beta' s_j)}{\sum_{l \in R(t_{(j)};d_j)} exp(\beta' s_l)} \tag{3.14}$$

Der Vektor s_j hat dabei die Länge p und enthält die Summen der p Kovariablen für alle Individuen, die zum Zeitpunkt $t_{(j)}$ verstorben sind. Die Notation $R(t_{(j)}; d_j)$ beschreibt ein Tupel von d_j Individuen aus dem Riskset zum Zeitpunkt $t_{(j)}$. Die Summe im Nenner der Formel ist damit definiert über alle möglichen Zusammensetzungen von d_j Individuen des Risksets zum Zeitpunkt $t_{(j)}$. Die Notation D beschreibt dabei die Gesamtzahl der zum Zeitpunkt i verstorbenen Individuen. Beide Ansätze sind jedoch eher komplex und rechenintensiv und sind bei Datensätzen mit einer großen Anzahl von Bindungen in den einzelnen Risk Sets nicht geeignet. Statistische Softwarepakete greifen aus diesem Grund in der Praxis auf verschiedene Approximationen zurück. Eine dieser Approximationen ist, neben der von Norman Breslow (1974) entwickelten, die von Bradley Efron (1977), welche auch in dieser Analyse verwendet wird:

$$\prod_{j=1}^{r} \frac{exp(\beta' s_j)}{\prod_{k=1}^{d_j} \left[\sum_{l \in R_{t_{(j)}}} exp(\beta' s_l) - \frac{k-1}{d_k} \sum_{l \in D(t_{(j)})} exp(\beta' x_l) \right]} \tag{3.15}$$

Dabei wird angenommen, dass die d_j Tode einzeln und sequenziell auftreten. s_j stellt den Vektor mit den Summen der p Kovariablen über alle Individuen, die am Zeitpunkt $t_{(j)}$ versterben, dar. Neben dem Riskset wird die Menge $D_{(J)}$ gebildet, die ebenfalls alle Individuen enthält, die zum Zeitpunkt $t_{(j)}$ versterben. Die Summe über diese Menge wird gemittelt und anteilig von der Summe der Kovariablen im Risk Set subtrahiert. Über die Multiplikation der einzelnen Risk Sets nach erfolgter Subtraktion der gewichteten Summen der Menge der Verstorbenen werden die möglichen Kombinationen der unabhängigen Todeszeitpunkte ohne Bindungen approximiert.

Im Fall, dass keine Bindungen vorliegen, resultieren alle Varianten der Likelihood-Funktion in der Standard-Formel (3.12).

3.2.3 Maximum-Likelihood-Schätzung

Zur Schätzung der β Koeffizienten wird die Maximum-Likelihood-Schätzung genutzt. Das Ziel ist die Maximierung der Likelihood-Funktion $L(\beta)$. Der erste Schritt der Maximierung ist die Berechnung der Log-Likelihood durch Logarithmieren:

$$l(\beta) = logL(\beta) \tag{3.16}$$

Aus diesem Term wird die Score-Funktion errechnet, welche der ersten Ableitung der Log-Likelihood nach dem Vektor der Koeffizienten, β, entspricht. Die Maximum-Likelihood-Gleichung wird durch Nullsetzen der Score-Funktion erhalten:

$$s(\beta) = \frac{\partial l(\beta)}{\partial \beta} = 0 \tag{3.17}$$

Die Lösung dieses nichtlinearen Gleichungssystems erfolgt mittels des Newton-Raphson-Algorithmus. Die Schätzung von β erfolgt iterativ:

$$\hat{\beta}^{k+1} = \hat{\beta}^k + \boldsymbol{F}^{-1}\left(\hat{\beta}^k\right) s\left(\hat{\beta}^k\right), f\ddot{u}r\ k = 0,1,2,\ldots \tag{3.18}$$

Dabei ist $\boldsymbol{F}(\beta) = \boldsymbol{H}(\beta) = \frac{\partial^2 l(\beta)}{\partial\beta\,\partial\beta'}$ mit den als zweite partielle Ableitungen $\partial^2 l(\beta)/\,\partial\beta_j\,\partial\beta_k$ definierten Elementen. Diese Matrix wird auch als beobachtete Informationsmatrix bezeichnet. Das Verfahren wird gestoppt, wenn ein vorgegebene Abbruchkriterium erfüllt ist, häufig wird $||\hat{\beta}^{k+1} - \hat{\beta}^k||/||\hat{\beta}^k|| \leq \epsilon$ gewählt. Bei Divergenz des Algorithmus existiert kein Maximum für ein endliches β.

3.2.4 Erweiterung um zeitabhängige Variablen

Das klassische Cox-Regressionsmodell, definiert in Abschnitt 3.2.1, geht von festen Kovariablen aus, die zum Zeitpunkt t $=$ 0 feststehen und konstant über die Zeit auf den Baseline-Hazard einwirken. Ist die Annahme dieser sogenannten Proportional-Hazards, Näheres zur Prüfung der Annahme in Abschnitt 3.2.5, ungültig, sind die betroffenen Kovariablen als zeitabhängig zu betrachten. Im Kontrast zur klassischen Definition (3.10) mit einem festen Wert des Vektors der Kovariablen x_i für das Individuum i, bildet

$$h_i(t) = exp\left(\beta_1 x_{1i}(t) + \beta_2 x_{2i}(t) + \cdots + \beta_p x_{pi}(t)\right) h_0(t) \tag{3.19}$$

die zeitabhängige Hazard-Funktion. Auch hier beschreibt $h_0(t)$ den Baseline-Hazard, im Fall, dass alle Kovariablen zu jedem Zeitpunkt Null sind. Die Likelihood-Funktion aus (3.12) wird dementsprechend bei der Schätzung der Koeffizienten um die zeitabhängigen Kovariablen ergänzt:

$$L(\beta) = \prod_{j=1}^{r} \frac{exp\left(\beta' x_{(j)}(t)\right)}{\sum_{l \in R(t_{(j)})} exp\left(\beta' x_l(t)\right)} \tag{3.20}$$

3.2.5 Auswertung des Cox-Regressionsmodells

Für die Aufnahme von Kovariablen in das Cox-Regressionsmodell müssen die betreffenden Kovariablen untereinander unkorreliert sein, um eine ungenaue Schätzung der Koeffizienten zu verhindern. Diese Forderung kann überprüft werden in dem die Multikorrelation des Modells gemessen wird. Ein Maß dafür ist der Varianzinflationsfaktor:

$$VIF_j = \frac{1}{1 - R_j^2} \tag{3.21}$$

$0 \leq R_j^2 \leq 1$ bezeichnet das Bestimmtheitsmaß der Regression der j-Kovariable auf alle anderen Variablen. Je größer die Streuung in der Regression ist, desto kleiner wird das Bestimmtheitsmaß. Daraus folgt für den VIF eine untere Grenze von 1, was auf eine unkorrelierte Variable schließen lässt. Je größer der Faktor, desto höher ist die Korrelation zwischen der Variable und den anderen Variablen. Als Grenze für eine problematische Multikorrelation wird ein Wert $VIF_j > 10$ angenommen (Fahrmeir, Kneib & Lang, 2009).

Die durch die Maximum-Likelihood-Schätzung erhaltenen Werte der Koeffizienten $\beta_1, \beta_2, \ldots, \beta_p$ werden auf Ihre statistische Signifikanz getestet. Dazu wird der Wald-Test verwendet. Die Teststatistik für eine Kovariable i ist definiert als $z = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$,

sie setzt also den geschätzten Koeffizienten der Variable in das Verhältnis des jeweiligen Standardfehlers. Dabei wird ausgewertet, ob der Koeffizient statistisch signifikant von 0 abweicht.

Die Annahme der proportionalen Hazards ist grundlegend im Cox-Regressionsmodell. Diese Annahme besagt, dass die Hazard-Funktion für eine bestimmte Kombination von Kovariablen proportional zur

Baseline-Hazard-Funktion ist. Das heißt im Umkehrschluss, dass der Einfluss zeitunabhängig ist. Ist diese Annahme verletzt, muss von einer zeitabhängigen Hazard-Funktion ausgegangen werden. Die Annahme lässt sich anhand der Schoenfeld-Residuen testen. Die Residuen $\hat{r}_i$ stellen die Abweichung der tatsächlichen Ausprägung einer Kovariablen eines zum Zeitpunkt verstorbenen Individuums $i \in D$ zum Erwartungswert dar (Moore, 2016):

$$\hat{r}_i = z_i - \sum_{k \in R_i} z_k \frac{e^{z_k \beta}}{\sum_{j \in R_k} e^{z_j \beta}} \tag{3.22}$$

Eine graphische Auswertung der Residuen aller Beobachtungen ist durch einen Plot gegen z_i möglich. Wenn diese um Null zentriert sind, ist die Proportional-Hazards-Annahme nicht verletzt. Für eine größere Anzahl von Variablen wird ein solches Vorgehen allerdings schnell unübersichtlich. Das R-Package *survival* enhält die Funktion *cox.zph*. Darin ist ein Test implementiert, der die Annahme prüft. Sollte die Null-Hypothese abgelehnt werden, so ist der Zusammenhang zwischen Baseline-Hazard und der Hazard-Funktion der Kovariable nicht linear. Die Annahme der Proportional-Hazards ist verletzt. Eine Möglichkeit diese Problematik in der Berechnung auszugleichen ist die Verwendung zeitabhängiger Kovariablen. Eine solche Erweiterung des Modells wird in Abschnitt 3.2.4 beschrieben.

4 Aufbau der Analyse

Nachdem in den ersten Kapiteln die theoretischen Grundlagen dargestellt und die Motivation und der Hintergrund der Bachelorarbeit näher beleuchtet wurden, erfolgt nun die Aufstellung eines detaillierten Analyseplans zur Umsetzung der Cox-Regression auf den vorhandenen Abrechnungsdaten der GKV. Die Arbeit wurde unter Berücksichtigung der Standardisierten Berichtsroutine für Sekundärdaten Analysen, Version 2 (STROSA 2; Swart et al., 2016) erstellt. Eine Übersicht der Kriterien findet sich im Anhang, **Tabelle 14**.

4.1 Datensatzbeschreibung

Datengrundlage der Versichertendaten innerhalb der Analysen ist die anonymisierte Forschungsdatenbank des Wissenschaftlichen Instituts für Gesundheitsökonomie und Gesundheitssystemforschung (WIG2). Die Versichertendaten beruhen hierbei auf den Abrechnungsdaten und Versichertendaten, die auf Grundlage des zehnten Kapitels des SGB V erhoben und dem WIG2 in anonymisierter Form zu Forschungszwecken bereitgestellt wurden. Zusätzlich wurden Referenzdaten des Bundesgesundheitsministeriums verwendet. An den Analysen waren ausschließlich Mitarbeiter:innen des WIG2 beteiligt. Ausschließlich aggregierte Ergebnisse wurden generiert und verteilt.

Da alle Daten im Rahmen des Sozialrechts für die Aufgaben der gesetzlichen Krankenkassen erforderlich sind und lediglich anonymisierte Daten verwendet wurden, entfällt die Verpflichtung zur Aufklärung der Versicherten und der Einholung einer Einwilligung für die Erhebung und Nutzung der Daten. Zur Wahrung der Anonymisierung wurden Ergebnisse lediglich aggregiert berichtet. Die Anonymisierung der Datenbank gewährleistet, dass einzelne Krankenkassen, Leistungserbringer:innen und Patient:innen nicht identifiziert werden können. Der Datenschutz wird weiterhin um die Vorgabe einer Mindeststichprobengröße von 100 Patient:innen für die Ausgangspopulation ergänzt.

In der Forschungsdatenbank sind Daten zur Inanspruchnahme und zum Ressourcenverbrauch innerhalb der GKV in Deutschland von ca. 4,5 Millionen anonymen Versicherten enthalten. Die Daten sind repräsentativ hinsichtlich der Alters- und Geschlechtsverteilung der deutschen Bevölkerung und eine longitudinale Beobachtung der Versicherten ist von 2010 bis 2019 möglich. In der Forschungsdatenbank sind demografische Daten (z. B. Alter, Geschlecht, Wohnregion), sowie Leistungs- und Kostendaten der ambulanten ärztlichen und stationären Versorgung, der Arzneimitteltherapie, der Heil- und Hilfsmitteltherapie und der häuslichen Krankenpflege gespeichert (WIG2, 2020). Medizinische Diagnosen werden durch das ICD-System dokumentiert.

Das ICD-System ist hierarchisch aufgebaut. Es besteht aus einer einstelligen, dreistelligen, einer detaillierten, vierstelligen und einer eventuellen, noch spezifischeren fünfstelligen Systematik. Der ICD-Code selbst ist alphanumerisch, bestehend aus einem führenden Buchstaben, der eine Einteilung in sogenannte Krankheitskapitel zulässt, und einer daran anschließenden Ziffernfolge. Die Ziffernfolge beruht dann der Systematik unterhalb der Krankheitskapitel. Die Selektierung von ICD-Codes erfolgt in dieser Analyse mittels Drei- und Vierstellern, in Einzelfällen werden auch spezifische Fünfsteller verwendet. Durch die Hierarchisierung werden auch zugehörige Codes einer tieferen Ebene aufgegriffen.

Die Analyse erfolgte auf Ebene von Versicherten. Für jeden Versicherten wurde individuell überprüft, ob dieser die Ein- und Ausschlusskriterien erfüllt und die entsprechenden Patient:innen wurden anschließend den einzelnen (nicht disjunkten) Kohorten zugeordnet. Auf Grundlage des Anteils der Patient:innen an allen Versicherten in den betrachteten Alters- und Geschlechtsgruppen, erfolgte anschließend eine Hochrechnung der beobachteten Patient:innenzahlen auf die GKV.

Der Aufgriff der Patient:innen mit entsprechenden ICD-Codierungen erfolgt über Diagnosedaten, die in den Leistungsdaten quartalsweise erfasst sind. Diese unterscheiden sich in ambulante und stationäre Diagnosen. Dabei werden stationäre Diagnosen als hinreichend gesichert betrachtet und fließen so direkt als Diagnose in die Analyse ein. Ambulante Diagnosen werden dagegen nur dann als Diagnose in die Analyse aufgenommen, wenn sie weiteren Kriterien genügen. So muss eine ambulante Diagnose zusätzlich mit dem Kennzeichen „gesichert“ versehen sein und das M2Q-Kriterium erfüllen, das heißt in mindestens zwei verschiedenen Quartalen eines Jahres dokumentiert sein, um als gesichert gewertet zu werden.

4.2 Studienpopulation

Im Folgenden werden konkrete Definitionen für die Bildung der Studienpopulation aufgestellt. Diese Studienpopulation bildet dann den Ausgangsdatensatz für die weiteren spezifischen Analysen in Form ein retrospektiven Kohortenstudie.

4.2.1 Ein- und Ausschlusskriterien

Der Zeithorizont der Daten reicht von 2010 bis 2018. Als Baseline-Periode werden die Kalenderjahre 2010 und 2011 festgelegt, um die Krankheitshistorie der Patient:innen abbilden zu können. Für die Studienpopulation wird als Indexjahr das Kalenderjahr 2012 definiert. Die Nachbetrachtung erfolgt dann anschließend im Follow-Up-Zeitraum von 2013 bis 2018. Ein Einschluss von Patient:innen in die Studienpopulation erfolgt, wenn folgende Kriterien erfüllt sind:

- Durchgängig versichert in der Baseline-Periode, das heißt mindestens 360 Versichertentage jeweils in den Jahren 2010 und 2011
- Mindestens 50 Jahre alt im Indexjahr 2012
- Durchgängig versichert, das heißt mindestens 360 Versichertentage in den Jahren 2013 bis 2018, oder verstorben im Follow-Up-Zeitraum

Ein Ausschluss aus der Studienpopulation ergibt sich, falls das folgende Kriterium erfüllt ist:

- Diagnostiziertes Vorhofflimmern in der Baseline-Periode oder dem Indexjahr, also in den Jahren 2010 bis 2012

Diese Kriterien dienen zum einen dazu, die Studienpopulation auf ältere Patient:innen mit einem höheren Risiko für Vorhofflimmern zu beschränken und zum anderen eine durchgehende Versicherung zu garantieren, damit sichergestellt wird, dass die vorliegenden Daten in der Datenbank das komplette Versorgungsgeschehen der Patient:innen darstellen. Weiterhin werden Patient:innen mit bereits diagnostiziertem Vorhofflimmern von der Analyse exkludiert, da die Zielgröße der Cox-Regression in dieser Analyse die erstmalige Diagnose eines Vorhofflimmerns ist. Die ICD-Codes, die ein diagnostiziertes Vorhofflimmern definieren, sind in **Tabelle 2** dargestellt.

ICD-10-GM	Beschreibung
I48.-	Vorhofflimmern und Vorhofflattern
I48.0	Vorhofflimmern, paroxysmal
I48.1	Vorhofflimmern, persistierend
I48.2	Vorhofflimmern, permanent
I48.3	Vorhofflattern, typisch inkl. Vorhofflattern, Typ 1
I48.4	Vorhofflattern, atypisch inkl. Vorhofflattern, Typ 2
I48.9	Vorhofflimmern und Vorhofflattern, nicht näher bezeichnet

***Tabelle 2**: Diagnosen nach ICD-10-GM zur Identifizierung von Vorhofflimmern, gültig ab 2013*[2]

Diese Codierung ist seit 2013 im Gebrauch. Bis einschließlich 2012 wurden Diagnosen für Vorhofflattern unter dem Viersteller I48.0- differenziert. Diagnosen für Vorhofflimmern wurden getrennt unter I48.1- dokumentiert. Innerhalb der vorliegenden Analysen der Bachelorarbeit werden alle Diagnosen im Rahmen von I48.- als Diagnosen für Vorhofflimmern betrachtet, dementsprechend wird auch Vorhofflattern betrachtet. Dies ist zum einen in der sehr ähnlichen Symptomatik begründet. Zum anderen lässt sich aber auch durch die Codierung im ICD-System nicht klar ein Vorhofflimmern von einem Vorhofflattern unterscheiden, da beide Erkrankungen bei unspezifischer Dokumentation als I48.9 notiert sind. Hinzu kommt die Änderung der Codierung von 2012 zu 2013, was ebenso einer Trennung ohne Informationsverlust entgegensteht. Die Verwendung des Begriffs Vorhofflimmern im weiteren Verlauf schließt damit die Diagnose eines Vorhofflatterns ein, äquivalent zum gesamten ICD-Dreisteller I48.-.

Die identifizierte Studienpopulation wird dann über einen Follow-Up Zeitraum von 2013 bis 2018 hinsichtlich der Entwicklung eines Vorhofflimmerns untersucht. In der Theorie des Cox-Regressionsmodells spricht man von den Todeszeitpunkten als interessierende Ereignisse. In dieser Analyse ist analog die erstmalige Diagnose eines Vorhofflimmerns das interessierende Ereignis. Dabei wird der Einfluss der Risikofaktoren durch zwei verschiedene Ansätze modelliert:

- Modellierung A: Abbildung der Risikofaktoren durch im ICD-System codierte Diagnosen
- Modellierung B: Abbildung der Risikofaktoren durch die HMGs des Morbi-RSA

4.2.2 Limitationen

Die Dokumentation von Diagnosen in den Versorgungsdaten erfolgt quartalsweise. Darauf basierend wird auch die verwendete Zeitachse im Modell in Quartalsschritten definiert. Weiterhin werden Todeszeitpunkte im Datensatz nur jährlich erfasst. Um eine quartalsweise Betrachtung zu ermöglichen werden die jeweiligen Todeszeitpunkte im letzten Quartal des betreffenden Jahres definiert.

4.3 Modellierung A

Im ersten Modellierungsansatz werden die in 2.4 herausgearbeiteten Risikofaktoren für Vorhofflimmern in das ICD-System transformiert, indem für jeden Risikofaktor und eventuelle Abstufungen eine Zuordnung zu einer Menge von ICD-Codierungen vorgenommen wird. Anhand dieser Kovariablen wird dann der Einfluss auf die Entwicklung eines Vorhofflimmerns untersucht. Die Zuordnungen der Krankheitsbilder zu den Risikofaktoren sind in **Tabelle 3** zusammengestellt.

Da die Behandlungsleitlinie für Vorhofflimmern überwiegend auf Primärdatenstudien oder auf Metaanalysen von solchen beruht (Kirchhof et al., 2016), lassen sich die gemessenen Studienendpunkte nicht deckungsgleich in Sekundärdaten, wie den Abrechnungsdaten, abbilden. Im Folgenden wird beschrieben,

[2] Quelle: Eigene Darstellung auf Basis der ICD-10-GM.

durch welche Annäherungen die Transformation der Daten in das ICD-System erfolgt ist. Gleichsam sind so die Limitationen der Modellierung A dargestellt, die aus den Annäherungen hervorgehen.

Das Lebensalter individueller Patient:innen kann über das ICD-System, welches nur Diagnosen dokumentiert, nicht dargestellt werden. Um das Alter zu bestimmen, wird auf die Versichertenstammdaten der Datenbank zurückgegriffen. Die Differenzierung der Patient:innen in 10-Jahres-Altersgruppen ab dem 50. Lebensjahr, die in der Leitlinie (Kirchhof et al., 2016) angeführt wird, wird übernommen und um die Altersgruppe der über 90-Jährigen ergänzt.

Die Leitlinie unterscheidet bei der Bewertung des Risikofaktors einer Schilddrüsenfehlfunktion Hypothyreose, subklinische Hyperthyreose und Hyperthyreose. Innerhalb des ICD-Systems ist nur eine gröbere Unterscheidung in Hypothyreose und Hyperthyreose möglich, die Spezifizierung subklinische Hypothyreose existiert nicht.

In der ICD-Systematik wird Übergewicht nicht dokumentiert, eine Dokumentation erfolgt nur bei einer vorliegenden Adipositas-Erkrankung. Es erfolgt die Definition von Adipositas ohne Abstufungen als Risikofaktor. Die verschiedenen Stadien einer chronischen Nierenkrankheit werden auch in der Codierung des Krankheitsbildes berücksichtigt. Ergänzend wird der Fall einer sonstigen chronischen Nierenkrankheit, im Sinne eines unbekannten Stadiums, betrachtet.

Rauchen und Alkoholkonsum werden im ICD-System lediglich bei der Identifikation von Suchtstadien berücksichtigt. Ein Selektieren von Rauchenden, die keine psychischen oder Verhaltensstörungen aufweisen, ist nicht möglich. Gleiches gilt für den Umfang des Alkoholkonsums. Dieser wird nicht dokumentiert, außer wenn es zu missbräuchlichem Konsum kommt. Die Auswirkung dieser Risikofaktoren können deshalb nur unter der Limitation betrachtet werden, dass es sich bereits um krankhaften Konsum handelt. Analog dazu wird die Intensität bzw. die Ausübung von Kraftsport nicht ärztlich dokumentiert und lässt sich nicht abbilden.

Risikofaktor	Differenzierung	Abbildung in ICD-10-GM	Beschreibung ICD-10-GM
Hohes Lebensalter	50–59 Jahre		Keine Abbildung über ICD: Abbildung über Versichertenstammdaten
	60–69 Jahre		Keine Abbildung über ICD: Abbildung über Versichertenstammdaten
	70–79 Jahre		Keine Abbildung über ICD: Abbildung über Versichertenstammdaten
	80–89 Jahre		Keine Abbildung über ICD: Abbildung über Versichertenstammdaten
	≥ 90 Jahre		Keine Abbildung über ICD: Abbildung über Versichertenstammdaten
Hypertonie		I10.-	Essenzielle (primäre) Hypertonie
		I11.- (ohne I11.0-)	Hypertensive Herzkrankheit
		I12.-	Hypertensive Nierenkrankheit
		I13.- (ohne I13.0-)	Hypertensive Herz- und Nierenkrankheit
		I15.-	Sekundäre Hypertonie
Herzinsuffizienz		I11.0-	Hypertensive Herzkrankheit mit (kongestiver) Herzinsuffizienz
		I13.0-	Hypertensive Herz- und Nierenkrankheit mit (kongestiver) Herzinsuffizienz
		I50.-	Herzinsuffizienz
Herzklappenerkrankungen		I05.-	Rheumatische Mitralklappenkrankheiten
		I06.-	Rheumatische Aortenklappenkrankheiten
		I07.-	Rheumatische Trikuspidalklappenkrankheiten

Risikofaktor	Differenzierung	Abbildung in ICD-10-GM	Beschreibung ICD-10-GM
		I08.-	Krankheiten mehrerer Herzklappen
		I09.-	Sonstige rheumatische Herzkrankheiten
		I34.-	Nichtrheumatische Mitralklappenkrankheiten
		I35.-	Nichtrheumatische Aortenklappenkrankheiten
		I36.-	Nichtrheumatische Trikuspidalklappenkrankheiten
		I37.-	Pulmonalklappenkrankheiten
		I38.-	Endokarditis, Herzklappe nicht näher bezeichnet
		Q22.-	Angeborene Fehlbildungen der Pulmonal- und der Trikuspidalklappe
		Q23.-	Angeborene Fehlbildungen der Aorten- und Mitralklappe
Herzinfarkt		I21.-	Akuter Myokardinfarkt
		I22.-	Rezidivierender Myokardinfarkt
Schilddrüsenfehlfunktion	Hypothyreose	E00.-	Angeborenes Jodmangelsyndrom
		E01.-	Jodmangelbedingte Schilddrüsenkrankheiten und verwandte Zustände
		E02.-	Subklinische Jodmangel-Hypothyreose
		E03.-	Sonstige Hypothyreose
	Hyperthyreose	E05.-	Hyperthyreose (Thyreotoxikose)
		E06.2-	Chronische Thyreoiditis mit transitorischer Hyperthyreose
Adipositas		E66.-	Adipositas
Diabetes Mellitus		E10.-	Diabetes mellitus, Typ 1
		E11.-	Diabetes mellitus, Typ 2
		E12.-	Diabetes mellitus in Verbindung mit Fehl- oder Mangelernährung (Malnutrition)
		E13.-	Sonstiger näher bezeichneter Diabetes mellitus
		E14.-	Nicht näher bezeichneter Diabetes mellitus
COPD		J44.-	Sonstige chronische obstruktive Lungenkrankheit
Obstruktive Schlafapnoe		G47.31	Obstruktives Schlafapnoe-Syndrom
Chronische Nierenkrankheit	Stadium 1/2	N18.1	Chronische Nierenkrankheit, Stadium 1
		N18.2	Chronische Nierenkrankheit, Stadium 2
	Stadium 3	N18.3	Chronische Nierenkrankheit, Stadium 3
	Stadium 4/5	N18.4	Chronische Nierenkrankheit, Stadium 4
		N18.5	Chronische Nierenkrankheit, Stadium 5
	Unbekannte Stufe	N18.8	Sonstige chronische Nierenkrankheit
Rauchen		F17.-	Psychische und Verhaltensstörung durch Tabak
Alkoholkonsum		F10.-	Psychische und Verhaltensstörungen durch Alkohol

***Tabelle 3**: Diagnosen nach ICD-10-GM zur Identifizierung von Risikofaktoren*[3]

[3] Quelle: Eigene Darstellung auf Basis von Kirchhof et al. (2016) und ICD-10-GM.

Die Variable Alter fließt, wie oben bereits erwähnt, in der Form von Altersgruppen als kategorische Variable mit mehreren Stufen in das Modell ein. Dazu werden Dummy-Variablen für die Altersgruppen der 60- bis 69-Jährigen, der 70- bis 79-Jährigen, der 80- bis 89-Jährigen und der mindestens 90-Jährigen gebildet. Dementsprechend wird die Altersgruppe der 50- bis 59-Jährigen als Referenzkategorie verwendet. Die Variable Geschlecht wird in Form der binären Eigenschaft *Männlich* in das Modell integriert, die Eigenschaft *Weiblich* bildet die Referenzkategorie. Die weiteren verwendeten Variablen fließen als binäre Variablen ein, sie werden 1 gesetzt, falls eine betreffende Diagnose entsprechend der Definition vorliegt, ansonsten nehmen sie den Wert 0 an. Basierend auf diesen Kovariablen wird ein Cox-Regressionsmodell mit der erstmaligen Diagnose eines Vorhofflimmerns als Zielgröße berechnet.

4.4 Modellierung B

Ein zweiter alternativer Ansatz nutzt ein Verfahren des Morbi-RSA der GKV in Deutschland. Dieser Ausgleich dient der Verteilung der finanziellen Belastung zwischen den gesetzlichen Krankenkassen. Um einen fairen Wettbewerb zwischen Krankenkassen mit unterschiedlichen Versichertenstrukturen hinsichtlich der Finanzkraft und der Morbiditätslast zu gewährleisten, erhalten die Krankenkassen Zuweisungen. Diese orientieren sich an der Morbidität der Versicherten und sollen eine Ausgabendeckung herbeiführen (BMG – Bundesministerium für Gesundheit, 2020). In diesem Verfahren werden die Versicherten HMGs zugeordnet. Diesen ist wiederum mindestens eine Diagnosegruppe (DxG) mit jeweils mindestens einem ICD-Code zugeordnet. Die Zuordnung erfolgt anhand der Diagnosen der Versicherten aus dem ambulanten und stationären Bereich, einzelne DxG sind außerdem an Altersgrenzen gebunden oder nur in Verbindung mit einem bestimmten Geschlechtskennzeichen gültig. Hierarchisierung bedeutet dabei, dass die verschiedenen Morbiditätsgruppen in einer Ordnung zueinander definiert sind. So werden Patient:innen in bestimmten Fällen, wenn zwei Morbiditätsgruppen besetzt sind, nur der höher eingestuften Gruppe zugeordnet (Bundesamt für soziale Sicherung, 2019). Dadurch entsteht eine Klassifizierung unterschiedlicher Schweregrade von Erkrankungen innerhalb des Systems. Für das jeweilige Ausgleichsjahr werden bei der Klassifikation jeweils die Daten aus dem vorangegangenen Kalenderjahr herangezogen. Für die Selektion der Population im Indexjahr 2012 ergibt sich daraus die Nutzung der Klassifikationsfestlegung für das Ausgleichsjahr 2013. Für das Ausgleichsjahr 2013 sind 155 verschiedene HMGs definiert, die wiederum 80 Krankheiten zugeordnet sind (BMG, 2020). Eine Übersicht aller HMGs ist unter den Festlegungen für das Ausgleichsjahr 2013 im Archiv des Bundesamts für soziale Sicherung (2012) zu finden.

Durch den Modellierungsansatz B sollen HMGs innerhalb des Morbi-RSA detektiert werden, die einen höheren Einfluss auf die Entwicklung eines Vorhofflimmerns haben und möglicherweise noch nicht hinreichend identifiziert sind. Dazu wird zunächst ein vollständiges Modell berechnet, welches alle 155 HMGs als Kovariablen der Cox-Regression nutzt. Anschließend wird die Menge der genutzten Kovariablen iterativ in Abhängigkeit der Signifikanz verringert. Das Modell wird so lange verkleinert, bis ein aussagekräftiges Modell vorliegt. Die Menge der verwendeten HMGs wird zunächst um die in **Tabelle 4** aufgezählten Gruppen reduziert, da diese per Definition entweder nur für jüngere Altersgruppen gelten oder aber angeborene Erkrankungen darstellen, die in dieser Form nicht mehr bei älteren Patient:innen diagnostiziert werden.

HMG	HMG Bezeichnung
87	Schwere angeborene Herzfehler (Alter < 18 Jahre)
88	Andere angeborene Herzfehler (Alter < 18 Jahre)
169	Schwere angeborene Fehlbildungen der Atmungs- und Verdauungsorgane
170	Sonstige angeborene Fehlbildungen des Zwerchfells und der Verdauungsorgane
218	Mukoviszidose (Alter < 12 Jahre)
233	Muskeldystrophie (Alter < 18 Jahre)
300	Angeborene Anomalien des Gefäßsystems inkl. Aorta und Herz

***Tabelle 4**: Ausgeschlossene HMGs, Modell 2*[4]

Die Codierung der Altersgruppen erfolgt analog zur Modellierung A in Form von Dummy-Variablen unter Verwendung der Altersgruppe der 50- bis 59-Jährigen als Referenz. Die Kovariable Geschlecht wird im Modell als die Eigenschaft *Männlich* binär definiert. Dementsprechend ist die Eigenschaft *Weiblich* im Modell die Referenzkategorie für das Geschlecht. Bei der Information zu den besetzten HMGs handelt es sich um binäre Variablen, die so in das Modell übernommen werden. Es wird erneut ein Cox-Regressionsmodell unter Verwendung des Zeitpunktes der erstmaligen Diagnose eines Vorhofflimmerns als Zielgröße durchgeführt. Für die Modellbildung wird dann eine Rückwärts-Selektion angewendet. Dazu wird zunächst ein vollständiges Modell berechnet. Dann werden in jedem Schritt der Selektion alle durch den Ward-Test bei einem Signifikanzniveau von 0,05 als insignifikant bewerteten Kovariablen entfernt. Es schließt sich eine Berechnung des Modells mit den verbliebenen Variablen an. Diese Schritte werden iterativ wiederholt, bis das finale Modell nur noch aus signifikanten Einflussfaktoren besteht. In diesem Modell werden dann die Einflüsse der unterschiedlichen HMGs analysiert.

4.5 Software

Die Daten werden mittels der Datenbankprogrammiersprache SQL aggregiert und aus der Datenbank ausgelesen. Zur Gruppierung der einzelnen Patient:innen in die HMGs wird eine Software zum Gruppieren von HMGs von 4K Analytics verwendet. Zur weiteren Analyse wird die Statistiksoftware R (Version 3.6.2) genutzt. Dabei wird bei allen Überlebenszeitanalysen auf das R-Package *Survival* zurückgegriffen. Für die Modelldefinition und -auswertung relevante Teile des R-Codes sind im Anhang angefügt.

[4] Quelle: Eigene Darstellung auf Basis der Festlegungen des Bundesamts für soziale Sicherungen (2012).

5 Ergebnisse

5.1 Studienpopulation

Bei der Selektion der Studienpopulation aus der Datenbank durch die in Abschnitt 4.2 definierten Ein- und Ausschlusskriterien ergab sich eine Selektierung der in **Tabelle 5** gegebenen Populationsgröße für die Analyse.

	Beschreibung	Anzahl Versicherte
1	Alle Versicherten im Indexjahr 2012	3.726.251
2	Alle Versicherten die im Indexjahr 2012 mindestens 50 Jahre alt und vollversichert waren	1.215.701
3	Alle Versicherten mit Vollversicherung in der Baseline-Periode und Vollversicherung oder Tod in Follow-Up-Periode	1.039.928
4	Alle Versicherten ohne eine VHF-Diagnose in der Baseline-Periode	964.894

***Tabelle 5**: Patient:innenzahlen nach Selektionsschritten für Studienpopulation*[5]

Die finale Studienpopulation hat also eine Größe von 964.894 Versicherten. Das Durchschnittsalter lag bei 64,2 Jahren, bei einer Standardabweichung von 10,0 Jahren. Der Anteil der männlichen Versicherten in der Population betrug 50,9 %.

5.2 Modellierung A

Zur Durchführung der Modellierung A erfolgte zunächst die Untersuchung der Studienpopulation auf das Vorliegen von Risikofaktoren für Vorhofflimmern im Indexjahr 2012. Die Zahlen und Anteile der Patient:innen mit dem jeweiligen Risikofaktor an der Studienpopulation zeigt **Tabelle 6**.

Risikofaktor	Anzahl Versicherte	Anteil Versicherte an Studienpopulation
Hypertonie	502.744	52,1 %
Herzinsuffizienz	54.935	5,7 %
Herzklappenerkrankungen	36.242	3,8 %
Herzinfarkt	21.604	2,2 %
Hypothyreose	71.495	7,4 %
Hyperthyreose	26.749	2,8 %
Adipositas	110.457	11,5 %
Diabetes mellitus	185.324	19,2 %
COPD	68.495	7,1 %
Obstruktive Schlafapnoe	10.544	1,1 %
Chronische Nierenkrankheit Stadium 1 oder 2	9.912	1,0 %
Chronische Nierenkrankheit Stadium 3	12.792	1,3 %
Chronische Nierenkrankheit Stadium 4 oder 5	4.856	0,5 %
Chronische Nierenkrankheit unbekannter Stufe	6.851	0,7 %
Rauchen	46.685	4,8 %
Alkoholkonsum	23.287	2,4 %

***Tabelle 6**: Absolute Patient:innenzahlen und relative Anteile der Patient:innen in der Studienpopulation*[6]

[5] Quelle: Eigene Darstellung.

[6] Quelle: Eigene Darstellung.

Die Hypertonie war mit 52,1 % der Patient:innen der am häufigsten diagnostizierte Risikofaktor in der Studienpopulation, gefolgt von Diabetes Mellitus und Adipositas, welche 19,2 % sowie 11,5 % der Patient:innen betrafen. 617.459 Patient:innen der Studienpopulation hatten mindestens einen diagnostizierten Risikofaktor, das entspricht einem Anteil von 64,0 %.

Alle im Modell enthaltenen Kovariablen wurden durch den jeweiligen Varianzinflationsfaktor auf eine mögliche Multikorrelation hin untersucht. Dieser Wert lag für alle Variablen nahe 1. Daraus folgt die Annahme, dass die Variablen unkorreliert sind und damit eine Modellierung durch das Cox-Regressionsmodell möglich ist (**Tabelle 7**).

Kovariable	VIF
Geschlecht	1,06
Altersgruppe 60–69 Jahre	1,29
Altersgruppe 70–79 Jahre	1,37
Altersgruppe 80–89 Jahre	1,20
Altersgruppe ab 90 Jahre	1,04
Hypertonie	1,23
Herzinsuffizienz	1,14
Herzklappenerkrankungen	1,05
Herzinfarkt	1,04
Hypothyreose	1,04
Hyperthyreose	1,01
Adipositas	1,12
Diabetes mellitus	1,16
COPD	1,08
Obstruktive Schlafapnoe	1,02
Chronische Nierenkrankheit Stadium 1/2	1,03
Chronische Nierenkrankheit Stadium 3	1,08
Chronische Nierenkrankheit Stadium 4/5	1,04
Chronische Nierenkrankheit unbekannter Stufe	1,04
Rauchen	1,09
Alkoholkonsum	1,04

***Tabelle 7**: Varianzinflationsfaktoren der Kovariablen in Modell A*[7]

In der vollständigen Modellierung wurde der Risikofaktor Hypothyreose als nichtsignifikant unter der Verwendung eines Signifikanzniveaus von 0,05 eingestuft. Die zugehörigen Koeffizienten, deren HR und deren Signifikanzen sind in **Tabelle 8** zu finden.

Kovariable	Coef	exp(coef)	Pr(>|z|)
Männlich	0,4013	1,4938	0
Altersgruppe 60–69 Jahre	0,8858	2,4248	0
Altersgruppe 70–79 Jahre	1,6022	4,9641	0
Altersgruppe 80–89 Jahre	2,1549	8,6267	0
Altersgruppe ab 90 Jahre	2,3761	10,7628	0
Hypertonie	0,3856	1,4706	0
Herzinsuffizienz	0,3622	1,4365	4,20E−263
Herzklappenerkrankungen	0,4700	1,6001	0

[7] Quelle: Eigene Darstellung.

Kovariable	Coef	exp(coef)	Pr(>|z|)
Herzinfarkt	0,1677	1,1826	8,29E−25
Hypothyreose	−0,0223	0,9780	0,0753
Hyperthyreose	0,0951	1,0998	8,89E−08
Adipositas	0,2004	1,2219	6,92E−101
Diabetes mellitus	0,1838	1,2018	1,19E−127
COPD	0,2744	1,3157	2,64E−148
Obstruktive Schlafapnoe	0,0963	1,1011	0,0002
Chronische Nierenkrankheit Stadium 1/2	0,0569	1,0586	0,0158
Chronische Nierenkrankheit Stadium 3	0,1497	1,1615	2,30E−14
Chronische Nierenkrankheit Stadium 4/5	0,4981	1,6456	2,56E−65
Chronische Nierenkrankheit Stufe unbekannt	0,1110	1,1174	2,89E−05
Rauchen	0,1041	1,1097	1,82E−11
Alkoholkonsum	0,2851	1,3299	1,61E−43

***Tabelle 8**: Koeffizienten der Kovariablen und deren Signifikanz*[8]

Aufgrund der fehlenden Signifikanz erfolgte eine weitere Modellierung unter Ausschluss des Risikofaktors Hypothyreose. In **Tabelle 9** sind die HR und die entsprechenden KI für die Kovariablen der zweiten Modellierung zusammengefasst. Die Ordnung erfolgt dabei absteigend nach der Größe des HR.

Kovariable	exp(coef)	lower95	upper95
Altersgruppe ab 90 Jahre	10,7684	10,2033	11,3648
Altersgruppe 80–89	8,6270	8,4111	8,8485
Altersgruppe 70–79	4,9636	4,8575	5,0720
Altersgruppe 60–69	2,4244	2,3698	2,4803
Chronische Nierenkrankheit Stadium 4/5	1,6448	1,5534	1,7416
Herzklappenerkrankungen	1,5996	1,5630	1,6370
Männlich	1,4966	1,4766	1,5170
Hypertonie	1,4699	1,4469	1,4933
Herzinsuffizienz	1,4359	1,4068	1,4656
Alkoholkonsum	1,3298	1,2772	1,3846
COPD	1,3151	1,2881	1,3426
Adipositas	1,2213	1,1990	1,2440
Diabetes mellitus	1,2016	1,1837	1,2198
Herzinfarkt	1,1824	1,1453	1,2208
Chronische Nierenkrankheit Stadium 3	1,1605	1,1167	1,2059
Chronische Nierenkrankheit Stufe unbekannt	1,1170	1,0604	1,1767
Rauchen	1,1093	1,0762	1,1435
Obstruktive Schlafapnoe	1,1007	1,0467	1,1575
Hyperthyreose	1,0977	1,0601	1,1365
Chronische Nierenkrankheit Stadium 1/2	1,0580	1,0102	1,1081

***Tabelle 9**: HR der Koeffizienten und die entsprechenden KI, Modell A*[9]

[8] Quelle: Eigene Darstellung.
[9] Quelle: Eigene Darstellung.

Bei der Auswertung des Modells ist zu beachten, dass nach Anwendung der Funktion *cox.zph* für die demografischen Kovariablen Männlich, Altersgruppe der 60- bis 69-Jährigen, Altersgruppe der 80- bis 89-Jährigen sowie Altersgruppe 90 Jahre und älter die Annahme der proportionalen Hazards verletzt ist. Gleiches gilt für die Risikofaktoren Hypertonie, Herzinsuffizienz, Herzklappenerkrankungen, Hyperthyreose, COPD und chronische Nierenkrankheit des ersten und zweiten Stadiums (**Tabelle 10**). Als Konsequenz daraus ist das Cox-Regressionsmodell in dieser Form ungültig.

Kovariable	chisq	df	p
Altersgruppe 60–69 Jahre	10,5283	1	0,0012
Altersgruppe 70–79 Jahre	0,0021	1	0,9635
Altersgruppe 80–89 Jahre	69,0680	1	9,51E−17
Altersgruppe ab 90 Jahre	43,2804	1	4,74E−11
Hypertonie	11,3931	1	0,0007
Herzinsuffizienz	64,5620	1	9,35E−16
Herzklappenerkrankungen	16,5139	1	4,83E−05
Herzinfarkt	1,5085	1	0,2194
Hyperthyreose	4,3345	1	0,0373
Adipositas	0,6959	1	0,4042
Diabetes mellitus	3,0637	1	0,0801
COPD	6,9751	1	0,0083
Obstruktive Schlafapnoe	0,0482	1	0,8262
Chronische Nierenkrankheit Stadium 1/2	4,4785	1	0,0343
Chronische Nierenkrankheit Stadium 3	1,1745	1	0,2785
Chronische Nierenkrankheit Stadium 4/5	0,7816	1	0,3767
Chronische Nierenkrankheit Stufe unbekannt	1,4884	1	0,2225
Rauchen	3,5259	1	0,0604
Alkoholkonsum	0,0184	1	0,8922
GLOBAL	210,6741	20	8,60E−34

***Tabelle 10**: Ausgabe der Funktion* cox.zph *zum Test auf die Proportional Hazard Annahme, Modell A*[10]

Um dennoch ein Cox-Regressionsmodell aufstellen zu können, wird der Ansatz der zeitabhängigen Kovariablen genutzt. Dabei werden die Variablen, für die die Bedingung der Proportionalität ungültig ist, über den gesamten Follow-Up-Zeitraum pro Quartal betrachtet. Dabei wird jeweils das erste Quartal dokumentiert, in dem der Risikofaktor diagnostiziert wurde. Das Cox-Regressionsmodell wird dann in den Intervallen, in denen eine Veränderung in der Konstellation der Risikofaktoren auftrat, separat geschätzt. Eine Ausnahme bei dieser Modellierung bilden die demografischen Kovariablen der Altersgruppen und des Geschlechts, die zum Indexzeitpunkt festgelegt werden und unter Limitation der Verletzung der Bedingung der proportionalen Hazards im Modell verbleiben.

In **Tabelle 11** sind die entsprechend geschätzten Koeffizienten dargestellt sowie die dazugehörigen KI. Den stärksten Einfluss verzeichnen dabei die verschiedenen Altersgruppen, dabei steigen die HR mit zunehmendem Alter von 2,32 in der Altersgruppe der 60- bis 69-Jährigen bis auf 9,29 in der Altersgruppe 90 Jahre und älter. Darauf folgen die kardiovaskulären Risikofaktoren Hypertonie, Herzklappenerkrankungen mit einem jeweils um ca. 66 % und Herzinsuffizienz mit einem um ca. 61 % erhöhten Risiko einer Vorhofflimmer-Diagnose gegenüber Versicherten im Baseline-Hazard ohne diagnostizierte Risikofaktoren. Bei Männern ist das Risiko um ca. 33 % gegenüber weiblichen Versicherten erhöht. Die

[10] Quelle: Eigene Darstellung.

weiteren betrachteten Risikofaktoren erhöhen allesamt das Risiko eines Vorhofflimmerns gegenüber der Baseline-Population, es existieren keine HR-Werte unter 1.

Kovariable	exp(coef)	lower95	upper95
Altersgruppe ab 90 Jahre	9,2870	8,7996	9,8015
Altersgruppe 80–89 Jahre	7,5622	7,3722	7,7570
Altersgruppe 70–79 Jahre	4,5415	4,4442	4,6409
Altersgruppe 60–69 Jahre	2,3196	2,2673	2,3731
Hypertonie	1,6639	1,6331	1,6953
Herzklappenerkrankungen	1,6634	1,6337	1,6937
Herzinsuffizienz	1,6159	1,5893	1,6428
Chronische Nierenkrankheit Stadium 4/5	1,5705	1,4834	1,6627
Männlich	1,4764	1,4566	1,4965
COPD	1,3282	1,3048	1,3521
Alkoholkonsum	1,3043	1,2527	1,3581
Adipositas	1,1959	1,1742	1,2180
Diabetes mellitus	1,1696	1,1523	1,1872
Hyperthyreose	1,1174	1,0867	1,1490
Chronische Nierenkrankheit Stadium 3	1,1162	1,0744	1,1596
Herzinfarkt	1,1079	1,0731	1,1437
Chronische Nierenkrankheit Stufe unbekannt	1,0940	1,0386	1,1523
Chronische Nierenkrankheit Stadium 1/2	1,0769	1,0485	1,1062
Rauchen	1,0681	1,0363	1,1010
Obstruktive Schlafapnoe	1,0595	1,0075	1,1141

***Tabelle 11**: HR der Koeffizienten und die entsprechenden KI, Modell A zeitabhängiger Ansatz*[11]

5.3 Modellierung B

Bei der Umsetzung von Modell B wurde ebenfalls zunächst der Varianzinflationsfaktor berechnet, um eine mögliche Multikorrelation auszuschließen. Auch hier befanden sich die Werte für alle Kovariablen nahe 1, sodass von einem unkorrelierten Datensatz ausgegangen werden kann. Anschließend wurde das im Abschnitt 4.4 dargestellte Verfahren zur Selektion der Variablen angewendet.

Auch in diesem Fall wurde die Annahme der proportionalen Hazards für die einzelnen HMG geprüft. Dabei wurde eine Verletzung für die in **Tabelle 12** erwähnten Gruppen mittels der Funktion *cox.zph* festgestellt. Da die Analyse nur die Gruppierung der Versicherten im Indexjahr vorsieht, lässt sich eine Modellierung der Kovariablen in Abhängigkeit der Zeit im Follow-Up-Zeitraum nicht umsetzen, sodass die Kovariablen mit der Limitation der verletzten Annahme im Modell verbleiben. Zu bemerken ist hierbei in jedem Fall, dass dies vorrangig Gruppen betrifft die zum einen kardiovaskuläre Krankheiten beinhalten, zum anderen sind Gruppen zeitabhängig, die Stoffwechselkrankheiten abbilden. HMG 92 der näher bezeichneten Arrhythmien ist hier tatsächlich außen vor zu lassen, da es in diesem Fall als Vorstufe von Vorhofflimmern zu betrachten ist.

[11] Quelle: Eigene Darstellung.

Kovariable	HMG-Bezeichnung	chisq	df	p
19	Diabetes ohne oder mit nicht näher bezeichneten Komplikationen	4,5154	1	0,0336
40	Osteoarthritis der Hüfte oder des Knies	5,2197	1	0,0223
46	Purpura und sonstige Gerinnungsstörungen	12,7828	1	0,0003
80	Herzinsuffizienz	83,5819	1	6,11E−20
84	Koronare Herzkrankheit/andere chronisch-ischämische Erkrankungen des Herzens	26,0579	1	3,31E−07
86	Erworbene Erkrankungen der Herzklappen und rheumatische Herzerkrankungen	8,1183	1	0,0044
91	Hypertonie, Hypertensive Herzerkrankung ohne Komplikationen und andere nicht näher bezeichnete Herzerkrankungen	44,8826	1	2,09E−11
92	Näher bezeichnete Arrhythmien	13,3641	1	0,0003
96	Zerebrale Ischämie oder nicht näher bezeichneter Schlaganfall	4,6621	1	0,0308
100	Hemiplegie/Hemiparese	7,0072	1	0,0081
103	Nicht näher spezifizierte Spätfolgen zerebrovaskulärer Erkrankungen	3,9082	1	0,0480
106	Atherosklerose, arterielles Aneurysma und sonstige, nicht näher bezeichnete Krankheiten der Arterien und Arteriolen	6,2950	1	0,0121
131	Nierenversagen	4,9945	1	0,0254
215	COPD oder Emphysem mit Dauermedikation, Bronchiektasen, sonstige interstitielle Lungenkrankheiten ohne Dauermedikation	5,0249	1	0,0250
251	Adipositas	5,8428	1	0,0156
290	Chronisch entzündliche Darmerkrankungen mit Dauermedikation	4,9203	1	0,0265

***Tabelle 12**: Ausgabe der Funktion* cox.zph *zum Test auf die Proportional Hazard Annahme, Modell B*[12]

Das finale Modell mit HRs und den zugehörigen KI ist in **Tabelle 13** abgebildet. Auch hier haben wiederum die einzelnen Altersgruppen nach Alter gestaffelt den stärksten Einfluss auf die Entwicklung eines Vorhofflimmerns im weiteren Zeitverlauf. Viele HMGs mit Bezug zu kardiovaskulären Diagnosen haben einen hohen Einfluss auf das Risiko der Entstehung eines Vorhofflimmerns. Besonders bereits vorliegende Arrhythmien gehen mit einem erhöhten Risiko der Entwicklung eines Vorhofflimmerns einher. In diesem Modell haben die HMGs 231 „Panikstörung, näher bezeichnete Phobien, sonstige anhaltende affektive Störungen“ und 58 „Depression“ sogar einen um 7 % verringernden Einfluss auf das Risiko ein Vorhofflimmern zu erleiden.

Kovariable	Beschreibung	exp(coef)	lower95	upper95
AG 90 plus	Versicherte älter als 90 Jahre	10,3873	9,8426	10,9621
AG 80 89	Versicherte zwischen 80 und 89 Jahren	8,2486	8,0418	8,4608
AG 70 79	Versicherte zwischen 70 und 79 Jahren	4,8506	4,7469	4,9565
AG 60 69	Versicherte zwischen 60 und 69 Jahren	2,4218	2,3673	2,4775
HMG 79	Herzstillstand/Schock	2,3744	2,0399	2,7638
HMG 92	Näher bezeichnete Arrhythmien	2,0161	1,8672	2,1767
HMG 77	Paroxysmale ventrikuläre Tachykardie	1,9249	1,7243	2,1488
HMG 130	Dialysestatus	1,8688	1,6770	2,0825
HMG 176	Transplantation der Niere	1,8003	1,5193	2,1334

[12] Quelle: Eigene Darstellung.

Kovariable	Beschreibung	exp(coef)	lower95	upper95
HMG 78	Pulmonale Herzkrankheit	1,7937	1,6851	1,9094
HMG 267	Morbus Hodgkin	1,7043	1,2547	2,3149
HMG 80	Herzinsuffizienz	1,6342	1,5983	1,6710
HMG 86	Erworbene Erkrankungen der Herzklappen und rheumatische Herzerkrankungen	1,6122	1,5732	1,6522
HMG 149	Hautulkus (ohne Dekubitalulzera)	1,5634	1,4763	1,6557
HMG 212	Erworbene hämolytische Anämie/Myelofibrose	1,5175	1,2096	1,9037
HMG 84	Koronare Herzkrankheit/andere chronisch-ischämische Erkrankungen des Herzens	1,4873	1,4573	1,5180
HMG 262	Akute myeloische Leukämie	1,4861	1,0502	2,1029
Männlich	Männliche Versicherte	1,4534	1,4333	1,4738
HMG 215	COPD oder Emphysem mit Dauermedikation, Bronchiektasen, sonstige interstitielle Lungenkrankheiten ohne Dauermedikation	1,4527	1,4133	1,4932
HMG 83	Angina pectoris/Z. n. altem Myokardinfarkt	1,4161	1,3769	1,4565
HMG 46	Purpura und sonstige Gerinnungsstörungen	1,4148	1,3471	1,4860
HMG 210	Hämophilie (Frauen) ohne Dauermedikation, sekundäre Thrombozytopenien, erworbener Mangel an Gerinnungsfaktoren	1,4022	1,2201	1,6115
HMG 89	Hypertensive Herz- und Nierenerkrankung, Enzephalopathie oder akutes Lungenödem	1,4011	1,3423	1,4625
HMG 175	Transplantation von Leber, Herz oder Pankreas	1,3953	1,0256	1,8982
HMG 104	Atherosklerose mit Ulkus oder Gangrän	1,3741	1,2538	1,5060
HMG 108	Sonstige interstitielle Lungenkrankheiten mit Dauermedikation	1,3666	1,1858	1,5750
HMG 91	Hypertonie, Hypertensive Herzerkrankung ohne Komplikationen und andere nicht näher bezeichnete Herzerkrankungen	1,3617	1,3400	1,3837
HMG 51	Alkohol- oder drogeninduzierte Psychose	1,3222	1,1978	1,4595
HMG 53	Schädlicher Gebrauch von Alkohol/Drogen ohne Abhängigkeitssyndrom	1,3091	1,2184	1,4064
HMG 265	Non-Hodgkin-Lymphom	1,3084	1,1964	1,4308
HMG 26	Leberzirrhose	1,3008	1,1952	1,4157
HMG 52	Alkohol- oder Drogenabhängigkeit	1,3005	1,2304	1,3745
HMG 16	Diabetes mit peripheren zirkulatorischen Manifestationen oder Ketoazidose	1,2716	1,1795	1,3710
HMG 134	Fortgeschrittene chronische Niereninsuffizienz	1,2622	1,1761	1,3546
HMG 251	Adipositas	1,2531	1,1887	1,3210
HMG 15	Diabetes mit renalen oder multiplen Manifestationen	1,2497	1,2062	1,2947
HMG 213	Myelodysplastische Syndrome	1,2292	1,0506	1,4381
HMG 216	Chronische respiratorische Insuffizienz, spezielle Pneumonien	1,2201	1,1242	1,3243
HMG 263	Akute lymphatische Leukämie, Erythroleukämie, Mastzellenleukämie, Multiples Myelom/Plasmozytom	1,2196	1,0489	1,4181
HMG 290	Chronisch entzündliche Darmerkrankungen mit Dauermedikation	1,2195	1,0894	1,3650
HMG 17	Diabetes mit sonstigen Komplikationen	1,2191	1,1798	1,2598
HMG 229	Rheumatoide Erkrankungen mit Dauermedikation	1,2159	1,1581	1,2766
HMG 40	Osteoarthritis der Hüfte oder des Knies	1,2105	1,1750	1,2471

Kovariable	Beschreibung	exp(coef)	lower95	upper95
HMG 81	Akuter Myokardinfarkt/instabile Angina Pectoris und andere akute ischämische Herzkrankheiten	1,1996	1,1295	1,2741
HMG 273	Bösartige Neubildung des Pharynx, Larynx, der Luftröhre, Bronchien, Lunge, Pleura, des Stütz- und Bindegewebes sowie der Mamma (Alter < 45 Jahre)	1,1972	1,1040	1,2982
HMG 99	Nicht näher bezeichnete Erkrankungen peripherer Gefäße	1,1792	1,1441	1,2155
HMG 96	Zerebrale Ischämie oder nicht näher bezeichneter Schlaganfall	1,1723	1,0897	1,2611
HMG 105	Lungenembolie/Periphere Gefäßerkrankungen (näher bezeichnet)	1,1677	1,1157	1,2222
HMG 19	Diabetes ohne oder mit nicht näher bezeichneten Komplikationen	1,1628	1,1429	1,1831
HMG 287	Ösophagusvarizen	1,1504	1,0035	1,3189
HMG 268	Lungenmetastasen, Metastasen der Verdauungsorgane, Tumorlyse-Syndrom	1,1460	1,0173	1,2910
HMG 266	Chronisch lymphatische Leukämie, Leukämie durch unspezifizierte Zellen, chronisch myeloproliferative Krankheit	1,1406	1,0235	1,2711
HMG 103	Nicht näher spezifizierte Spätfolgen zerebrovaskulärer Erkrankungen	1,1399	1,0883	1,1940
HMG 237	COPD oder Emphysem ohne Dauermedikation	1,1325	1,1019	1,1639
HMG 214	Polycythaemia Vera/sonstige näher bezeichnete Anämien	1,1205	1,0436	1,2032
HMG 22	Andere kostenintensive schwerwiegende endokrine oder metabolische Erkrankungen	1,1118	1,0380	1,1907
HMG 131	Nierenversagen	1,1030	1,0756	1,1311
HMG 112	Akute und nicht näher bezeichnete respiratorische Insuffizienz, Lungenabszess	1,0975	1,0388	1,1595
HMG 152	Schwerwiegende bakterielle Infektionen der Unterhaut und des Fettgewebes	1,0884	1,0445	1,1342
HMG 275	Bösartige Neubildung des Magens, der Nebennieren, weitere intrathorakale bösartige Neubildungen der Atemwege, Neubildung unklarer Dignität des ZNS	1,0848	1,0316	1,1408
HMG 74	Epilepsie	1,0831	1,0252	1,1443
HMG 157	Wirbelkörperfrakturen (inkl. pathologische)	1,0819	1,0174	1,1505
HMG 203	Sonstige pathologische Frakturen	1,0800	1,0118	1,1527
HMG 258	Schlafapnoe, Narkolepsie und Kataplexie	1,0713	1,0347	1,1091
HMG 100	Hemiplegie/Hemiparese	1,0702	1,0109	1,1331
HMG 106	Atherosklerose, arterielles Aneurysma und sonstige, nicht näher bezeichnete Krankheiten der Arterien und Arteriolen	1,0656	1,0329	1,0993
HMG 204	Osteoporose bei Frauen	1,0565	1,0088	1,1065
HMG 71	Polyneuropathie	1,0481	1,0192	1,0778
HMG 253	Chronischer Schmerz mit Dauermedikation	1,0447	1,0046	1,0865
HMG 220	Psoriasis und Parapsoriasis ohne Dauermedikation	1,0400	1,0005	1,0812
HMG 231	Panikstörung, näher bezeichnete Phobien, sonstige anhaltende affektive Störungen	0,9322	0,8859	0,9808
HMG 58	Depression	0,9319	0,9055	0,9590

***Tabelle 13**: HRs der Koeffizienten und die entsprechenden KI, Modell B*[13]

[13] Quelle: Eigene Darstellung.

6 Diskussion und Ausblick

6.1 Modellierung A

Die aus der finalen zeitabhängigen Modellierung A resultierenden Ergebnisse decken sich weitgehend mit denen, die in der Leitlinie aus verschiedenen Einzelstudien zusammengetragen wurden. Auch dort war das Alter der Risikofaktor mit den höchsten HR und ein höheres Alter war analog mit einem höheren Risiko von Vorhofflimmern verbunden. Die Werte der vorliegenden Analyse liegen in den einzelnen Altersgruppen etwas niedriger.

Auch für die überwiegende Mehrheit der weiteren Risikofaktoren bestätigen sich die Ergebnisse bei der Analyse auf den Abrechnungsdaten hinsichtlich der Stärke des Einflusses. Allgemein ist auch hier festzustellen, dass die errechneten HR leicht unter denen der Leitlinie liegen. Dieser Umstand ist sicherlich der Zusammensetzung der Studienpopulation geschuldet, da die gesamte Versichertenpopulation sehr viel diverser als eine Studienpopulation ist, die einem gewissen Selektionsbias unterliegt.

Auffälligkeiten ergeben sich bei der obstruktiven Schlafapnoe, die mit einem HR von 1,10 (95 % KI 1,01−1,11) deutlich unter dem Wert der Leitlinie von 2,18 (95 % KI 1,34−1,54) liegt. Der umgekehrte Effekt liegt bei der Hypertonie vor. Hier liefert die Analyse einen Wert von 1,66 (95 % KI 1,63−1,70) gegenüber 1,32 (95 % KI 1,08−1,60). Hier ist zusätzlich zu bemerken, dass dieser Risikofaktor bei 52,1 % der Versicherten der Studienpopulation vorliegt und so ohne Frage von einer Volkskrankheit gesprochen werden kann.

Die ebenfalls in der Bevölkerung sehr weit verbreiteten Risikofaktoren des Rauchens und des Alkoholkonsums konnten in der vorliegenden Analyse nur über die Diagnosen des Suchtstatus aufgegriffen werden. Im Fall von Alkohol deckt sich der errechnete HR mit der höchsten Stufe von mehr als 21 Getränken (je 12 g Alkohol) pro Woche (1,31 (95 % KI 1,25−1,36) und 1,39 (95 % KI 1,22−1,58)). Ein anderes Bild zeigt sich beim Rauchen. Rauchen mit diagnostiziertem Suchtstatus erhöht das Risiko eines Vorhofflimmerns nur um ca. 7 % (HR 1.07 (95 % KI 1,04−1,10)). Eine mögliche Erklärung ist, dass diese Diagnose selbst bei Rauchenden mit diesem Ausmaß an Tabakkonsum nicht unbedingt als ICD codiert wird.

Die demografische Kovariable Geschlecht wird in der Leitlinie nicht berücksichtigt. In der eigenen Modellierung zeigt sich dafür ein recht hoher Einfluss. So liegt ein HR von 1,48 (95 % KI 1,47−1,50)) für Männer vor, das heißt, es existiert in den Abrechnungsdaten ein um fast 50 % erhöhtes Risiko eines Vorhofflimmerns für Männer gegenüber Frauen der Baseline-Periode.

6.2 Modellierung B

Die Risikofaktoren der Modellierung A finden sich auch in Modellierung B in dazu korrespondierenden HMG wieder. Analog dazu werden Krankheitsbilder erweitert und zeigen ein noch höheres Risiko auf, so zum Beispiel im Fall der HMG 83, einer Angina Pectoris als Zustand nach einem stattgehabten Herzinfarkt, hier werden nicht nur Herzinfarkte im Indexjahr 2012 aufgenommen, sondern auch zurückliegende Ereignisse, sodass der HR bei 1,42 (95 % KI 1,38−1,46) gegenüber 1,11 (95 % KI 1,07−1,14) in Modellierung A liegt. Andere kardiovaskuläre Erkrankungen, die hier einen signifikanten Einfluss haben, fanden in der Modellierung A keine Berücksichtigung, hier sind exemplarisch die HMG 84 mit der koronaren Herzkrankheit mit einem HR von 1,49 (95 % KI 1,46−1,52), die HMG 78 der pulmonalen Herzkrankheit mit einem HR von 1,79 (95 % KI 1,69−1,91) und die HMG 79 Herzstillstand/Schock mit einem HR von 2,37 (95 % KI 2,04−2,76) zu nennen.

Bislang unberücksichtigte Krankheitsbilder wie Erkrankungen des Lymphsystems (z. B. HMG 263, HMG 265, HMG 267), respiratorische Erkrankungen (z. B. HMG 216) und verschiedene Krebserkrankungen sind ebenso mit einem erhöhten Risiko verbunden. Als auffällig ist noch die Tatsache zu bemerken, dass die HMG 231 und 58 mit den psychischen Erkrankungen Panikstörung, näher bezeichnete Phobien, sonstige anhaltende affektive Störungen und Depression, das Risiko eines Vorhofflimmerns sogar verringern. Die HR liegen bei 0,93 (95 % KI 0,86−0,98) und 0,93 (95 % KI 0,91−0,96).

6.3 Fazit

Die Bedeutung von Vorhofflimmern für das Gesundheitssystem in Deutschland zeigt sich bereits durch die Tatsache, dass knapp zwei Drittel der Versicherten in der Studienpopulation mindestens eine Diagnose eines Risikofaktors für Vorhofflimmern im Indexjahr 2012 haben. Die Anteile der einzelnen Risikofaktoren in Modellierung A (Tabelle 6) unterstreichen die Bedeutung der Vorsorge für Vorhofflimmern in Deutschland.

Das Ergebnis der vorliegenden Arbeit bestätigt dabei die bereits bekannten Risikofaktoren für Vorhofflimmern und zeigt gleichzeitig, dass Risikofaktoren auch in den Abrechnungsdaten der Krankenkassen der GKV identifizierbar sind. Dies kann sowohl über die ICD-Systematik, als auch über die HMGs des Morbi-RSA operationalisiert werden. Dabei zeigt letzteres Instrument insbesondere zusätzliche kardiovaskuläre Erkrankungen auf, die im Zusammenhang mit Vorhofflimmern stehen und die bislang nicht explizit in der Behandlungsleitlinie (Kirchhof et al., 2016) als Risikofaktor angeführt werden. Komplett neue Krankheitsbilder ohne jeglichen Zusammenhang zu den bisher bekannten Risikofaktoren, traten nur vereinzelt auf, sodass Modellierung B eher auf einen erweiterten Kreis an Risikofaktoren im Rahmen der HMG-Struktur hindeutet. Es lässt sich folgern, dass das gezielte Screening von potentiellen Patient:innen mit Vorhofflimmern durch die Identifikation anhand von Sekundärdaten unterstützt werden kann.

Die vorliegende Modellierung nutzt an einigen Stellen Approximationen, die durchaus Ansatzpunkte für weiterführende Forschung bieten. So existiert die Limitation, dass Diagnosen nur quartalsweise vorliegen. Diese Situation kann alternativ auch mittels Intervall-Zensierung modelliert werden, wobei angenommen wird, dass die Diagnose zwischen zwei Zeitpunkten liegt, anstatt per Definition auf einen exakten Zeitpunkt abgebildet zu werden. Eine auf diese Bachelorarbeit aufbauende Analyse kann weiterhin die Modellierung B um ein zeitabhängiges Modell erweitern und die Gruppierung der HMG jährlich vornehmen, so dass die Limitation der Verletzung der Annahme der proportionalen Hazards aufgelöst ist.

Die Ergebnisse lassen sich aufgrund der Stichprobengröße und der Repräsentativität der Datenbank hinsichtlich der deutschen Bevölkerungsstruktur innerhalb Deutschlands übertragen. Die Arbeit wurde unter Berücksichtigung der Standardisierten Berichtsroutine für Sekundärdaten Analysen, Version 2 (STROSA 2; Swart et al., 2016) erstellt. Eine Übersicht der Kriterien findet sich im Anhang, Tabelle 14.

Literaturverzeichnis

Breslow, N. (1974). Covariance Analysis of Censored Survival Data. *Biometrics, 30*(1), S. 89–99. DOI: 10.2307/2529620.

Bundesamt für soziale Sicherung (Hrsg.). (2019). *Festlegungen nach § 31 Absatz 4 RSAV für das Ausgleichsjahr 2020*. Online: <https://www.bundesamtsozialesicherung.de/de/themen/risikostrukturausgleich/festlegungen/> (abgerufen am 17.08.2020).

Bundesamt für soziale Sicherung (Hrsg.). (2012). *Festlegungen für das Ausgleichsjahr 2013. Festlegungen der zu berücksichtigenden Krankheiten für das Ausgleichsjahr 2020*. Online: <https://www.bundesamtsozialesicherung.de/de/themen/risikostrukturausgleich/festlegungen/archiv-festlegungen/> (abgerufen am 17.08.2020).

BMG – Bundesministerium für Gesundheit (Hrsg.). (2020). *Risikostrukturausgleich*. Online: <https://www.bundesgesundheitsministerium.de/risikostrukturausgleich.html> (abgerufen am 19.08.2020).

Collett, D. (2015). *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC. DOI: 10.1201/b18041.

Cox, D. R. (1972). Regression Models and Life-Tables. *Journal of the Royal Statistical Society*, *34*(2), S. 187–220.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the American Statistical Association, 72*(359), S. 557–565. DOI: 10.1080/01621459.1977.10480613.

Fahrmeir, L., Kneib, T. & Lang, S. (2009). *Regression*. Berlin, Heidelberg: Springer. DOI: 10.1007/978-3-642-01837-4.

Kalbfleisch, J. D. & Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Hoboken: John Wiley & Sons, Inc. DOI: 10.1002/9781118032985.

Kirchhof, P., Benussi, S., Kotecha, D., Ahlsson, A., Atar, D., Casadei, B., Castella, M., Diener, H.-C., Heidbuchel, H., Hendriks, J., Hindricks, G., Manolis, A. S., Oldgren, J., Popescu, B. A., Schotten, U., Van Putte, B. & Vardas, P. (2016). 2016 ESC Guidelines for the management of atrial fibrillation developed in collaboration with EACTS. *European Heart Journal, 37*(38), S. 2893–2962. ESC Scientific Document Group. DOI: 10.1093/eurheartj/ehw210.

Klein, J. P. & Moeschberger, M. L. (2003). *Survival Analysis*. New York: Springer New York. DOI: 10.1007/b97377.

Klinke, R., Pape, H.-C., Kurtz, A. & Silbernagl, S. (Hrsg.). (2010). *Physiologie*. Stuttgart: Georg Thieme Verlag. DOI: 10.1055/b-002-46974.

Krijthe, B. P., Kunst, A., Benjamin, E. J., Lip, G. Y. H., Franco, O. H., Hofman, A., Wittman, J. C. M., Stricker, B. H. & Heeringa, J. (2013). Projections on the number of individuals with atrial fibrillation in the European Union, from 2000 to 2060. *European Heart Journal, 34*(35), S. 2746–2751. DOI: 10.1093/eurheartj/eht280.

Lüderitz, B. & Lewalter, T. (2010). *Herzrhythmusstörungen*. Berlin [u. a.]: Springer. DOI: 10.1007/978-3-540-76755-8.

Moore, D. F. (2016). *Applied Survival Analysis Using R*. Cham: Springer International Publishing. DOI: 10.1007/978-3-319-31245-3.

Passman, R. & Bernstein, R. A. (2016). New Appraisal of Atrial Fibrillation Burden and Stroke Prevention. *Stroke, 47*(2), S. 570–576. DOI: 10.1161/STROKEAHA.115.009930.

Schnabel, R. B., Wilde, S., Wild, P. S., Munzel, T. & Blankenberg, S. (2012). Atrial Fibrillation. *Deutsches Ärzteblatt Online*. DOI: 10.3238/arztebl.2012.0293.

Stewart, S., Hart, C. L., Hole, D. J. & McMurray, J. J. (2002). A populationbased study of the long-term risks associated with atrial fibrillation: 20-year follow-up of the Renfrew/Paisley study. *The American Journal of Medicine, 113*(5), S. 359–364. DOI: 10.1016/S0002-9343(02)01236-6.

Swart, E., Bitzer, E., Gothe, H., Harling, M., Hoffmann, F., Horenkamp-Sonntag, D., Maier, B., March, S., Petzold, T., Röhrig, R., Rommel, A., Schink, T., Wagner, C., Wobbe, S. & Schmitt, J. (2016). STandardisierte BerichtsROutine für Sekundärdaten Analysen (STROSA) – ein konsentierter Berichtsstandard für Deutschland, Version 2. *Das Gesundheitswesen, 78*(Suppl. 1), e145–e160. DOI: 10.1055/s-0042-112008.

WIG2 – Wissenschaftliches Institut für Gesundheitsökonomie und Gesundheitssystemforschung (Hrsg.). (2020). *WIG2 Forschungsdatenbank*. Online: <https://www.wig2.de/analysetools/wig2-forschungsdatenbank.html> (abgerufen am 03.07.2020).

Wilke, T., Groth, A., Mueller, S., Pfannkuche, M., Verheyen, F., Linder, R., Linder, R., Maywald, U., Bauersachs, R., Breithardt, G. (2013). Incidence and prevalence of atrial fibrillation: an analysis based on 8.3 million patients. *Europace, 15*(4), S. 486–493. DOI: 10.1093/europace/eus333.

Wolf, P. A., Abbott, R. D. & Kannel, W. B. (1991). Atrial fibrillation as an independent risk factor for stroke: the Framingham Study. *Stroke, 22*(8), S. 983–988. DOI: 10.1161/01.STR.22.8.983.

Anhang

STROSA 2 Kriterien

Kriterium	Abschnitt
Titel, Abstract, Schlagworte	
Titel und Abstract	Abstract
Schlagworte	Keywords
Einleitung	
Hintergrund und Rationale	1
Zielsetzungen	1
Methoden	
Studiendesign	4.2
Datenquelle	4.1
Rechtsgrundlage	4.1
Datenschutz	4.1
Datenfluss	4.1
Analyseeinheit	4.1
Variablen	4.3 und 4.4
Studiengröße	4.1 und 5.1
Statistische Methoden	3.2
Ergebnisse	
Selektion der Studienpopulation	5.1
Deskriptive Ergebnisse	5.1
Hauptergebnisse	5.2 und 5.3
Weitere Ergebnisse	Keine
Diskussion	
Hauptergebnisse	6.3
Interne Validität und Risiko von Verzerrungen	6.1 und 6.2
Stärken und Schwächen	6.3
Interpretation	6.1, 6.2 und 6.3
Übertragbarkeit	6.3
Schlussfolgerungen	
Fazit	6.3
Interessenkonflikte	
Finanzierung	Keine
Rolle der Dateneigner	Keine
Sonstige Interessenkonflikte	Keine

***Tabelle 14**: Kriterien STROSA 2 und deren Position im Text*[14]

[14] Quelle: Eigene Darstellung auf Basis von Swart et al. (2016).

Modell A1 Berechnung

```
# string_cov_A String mit allen Covariablen des Modells A

result_cox_modell_A <- coxph(formula = as.formula(paste0('Surv(surv_time,
censored,type = "right") ~ ',string_cov_A)), data = data_A)

Modell_A_sum <- summary(result_cox_modell_A)
```

Modell A2 Berechnung

```
# string_cov_A String mit allen Covariablen des Modells A, ausgenommen
Hypothyreose

result_cox_modell_A_alt <- coxph(formula = as.formula(paste0('Surv
(surv_time, censored, type = "right") ~ ',string_cov_A_alt)), data =
data_A)

Modell_A_alt_sum <- summary(result_cox_modell_A_alt)
```

Modell A3 Berechnung

```
# zeitabhaengig modellierte Variablen sind mit dem Suffix "_interval"
versehen

result_cox_modell_A_time <- coxph(formula = Surv(tstart,tstop,VHF_Diag-
nose) ~ Maennlich + AG_60_69 + AG_70_79 + AG_80_89 + AG_90_plus + Hyper-
tonie_interval + Herzinsuffizienz_interval + Herzklappenerkrankungen_in-
terval + Herzinfarkt + Hyperthyreose_interval + Adipositas + Diabe-
tes_mellitus + COPD_interval + Obstruktive_Schlafapnoe + Chronische_Nie-
renkrankheit_Stadium_1_2_interval + Chronische_Nierenkrankheit_Stadium_3
+ Chronische_Nierenkrankheit_Stadium_4_5 + Chronische_Nierenkrank-
heit_Stufe_unbekannt + Rauchen + Alkoholkonsum, data = data_A_time_cp)

summary(result_cox_modell_A_time)
```

Modell B1 Berechnung

```
# Indikator-Variable, ob insignifikante Variablen vorliegen

cov_insignificant <- 1

# alle relevanten Kovariablen werden in das volle Modell aufgenommen

cov_loop <- paste(colnames(data_B)[!colnames(data_B) %in%
c('PID','Alter','Erstes_Quar-
tal_VHF','Verstorben','censored','surv_time')],collapse = ' + ')

# While-Loop entfernt insignifikante Variablen und speichert Modell mit
signifikanten Variablen

while(cov_insignificant > 0){

# Cox-Modell mit signikant betrachteten Kovariablen berechnen

result_cox <- coxph(formula = as.formula(paste0('Surv(surv_time, cen-
sored) ~ ',cov_loop)), data = data_B)

result_cox_sum <- summary(coxph(formula = as.for-
mula(paste0('Surv(surv_time, censored) ~ ',cov_loop)), data = data_B))

# Reduzierung des Strings der folgenden Iteration um insignifikante Kova-
riablen

cov_loop <- paste0(rownames(result_cox_sum$coefficients)[re-
sult_cox_sum$coefficients[,5] < 0.05],collapse = ' + ')}
```